Lecture Notes on Missing Data and Measurement Error in Medical Studies

Ching–Yun (C. Y.) Wang

Division of Public Health Sciences, Fred Hutchinson Cancer Research Center,
P.O. Box 19024 , Seattle, WA 98109-1024, USA
Affiliate faculty, Biostatistics, University of Washington

November 8, 1999

Outline:

4. Logistic Regression with covariate Measurement Error:
   - 4.1 Logistic regression and regression calibration;
   - 4.2 Maximum likelihood estimator;
   - 4.3 Functional methods via sufficiency and conditional scores.
5. Design and Analysis on Nutritional Data:
   - 5.1 Design and modeling;
   - 5.2 Methods for stratified sampling;
   - 5.3 Methods for restricted sampling.
6. Cox regression with missing covariates:
   - 6.1 Introduction;
   - 6.2 Regression calibration estimator;
   - 6.3 Estimated partial likelihood estimator.
7. Cox regression with covariate measurement error:
   - 7.1 Introduction;
   - 7.2 Regression calibration estimator;
   - 7.3 EM algorithm.

1 Introduction and Examples

Abstract

Missing data and measurement error are two important problems in medical studies and other scientific fields. Sometimes they arise simultaneously in a real problem. In regression analysis, a significant difference may occur if covariate variables are measured imprecisely or due to missingness. Measurement error may occur even there is no instrumental error. In some real problems, a usual definition of a covariate variable may be a subject's long-term average of some measurements. This leads to measurement error too because it is impossible to measure an infinite number of observations. Some interesting examples will be described. A missing covariate regression model and a covariate measurement error model will be introduced.

1.1 Introduction

Data measured with error are very common in epidemiological studies. Such data can cause significant difficulties in deriving correct results and interpretation. For example, some commonly used instruments in nutritional data analysis include food frequency questionnaires, four-day food records, and 24-hour food recalls. Interest in these variables usually centers around a regression analysis, for example, the regression of the nutrient variables on disease incidences. Food intake variables like these usually involve subject recall errors, and these errors can cause the analysis difficulties associated with the measurement error problem. Errors in variables can also arise even when the measurement process per se is accurate. Consider, for example, when the variable of interest is the long-term average of a quantity which varies over time. Because in reality only a finite number of observations can be made, only an estimate of the variable of interest is obtained, which therefore has error associated with it. For example, long-term fat intake may be considered a an important risk factor for cancer and heart disease. As another example, systolic blood pressure may be related to the development of coronary heart disease. In the literature, the term measurement error is used synonymously with errors in variables.

The measurement error problem may be considered as an extreme case of the missing data problem in which all observations have missing data on the true value of the covariate. When the true variable is impossible to measure, one approach is to measure a surrogate replicatedly. A surrogate variable does not have an effect on the outcome variable given the true covariates of interest. On the other hand, the results of an analysis might be highly sensitive to the surrogate modeling assumptions. In some studies, a validation sample is collected on the true variables. Using the validation set, one may verify the surrogate modeling approach. This allows an assessment of the impact on results of different surrogate models. Modern statistical methods such as these have been reviewed in Carroll, Ruppert, and Stefanski (1995), Fuller (1987), and Little and Rubin (1987). New methodologies are under development.

A naive approach to analysis in the presence of missing data is to only use observations with complete data, that is called the complete–case (CC) analysis. This approach can give rise to bias and reduced statistical efficiency. A naive approach to data analysis when covariates have measurement error is to ignore it and run standard analyses using the mismeasured covariates. This approach can also give rise to serious bias. Therefore, naive approaches are not acceptable and developing appropriate statistical methods which accommodate missing and mismeasured data is important.

1.2 Motivating examples

The first motivating example for this problem is from the Women's Health Initiative (WHI). This is a large multicenter randomized trial and observational study involving 161,000 postmenopausal women sponsored by the National Institutes of Health and coordinated at Fred Hutchinson Cancer Research Center, Seattle, USA. An objective of the WHI observational study is to determine if bone mineral density predicts risk of bone fracture. We consider the response variable Y as binary indicating incidence of hip fracture and let the covariate of interest X be the bone mineral density which is available for all women scanned at three WHI clinical centers (Birmingham, Tucson, Pittsburgh). Data from these three centers will constitute the validation set for WHI women at all 40 clinical centers. A validation set is the subsample of the study cohort for which all variables are available. Surrogates for X in this setting might be race or hormone use, where a surrogate is a variable which has no effect on the response variable given true covariates of interest.

The second example we consider is the case that a validation sample is not available. Instead of measuring X, repeated surrogate measurements $W_{ij} = X_i + U_{ij}, j = 1, \ldots, k_i;\ i = 1, \ldots, n$ with $E(U_{ij}|X_i) = 0$, are available for all subjects of the study cohort. Here n is the total number of subjects and k_i is the number of repeated measurements of W for subject i. For example, a goal of the WHI observational study is to determine how dietary intake is related to breast cancer. We may consider X, the covariate of interest, as the long–term percentage energy from fat which can not be accurately ascertained. The measurement of % energy from fat obtained from a food frequency questionnaire would constitute the surrogate W. The food frequency questionnaire (FFQ) is commonly used to assess the diet of individuals because it is a cost efficient method of obtaining dietary information (Willett, 1990). However, it is well known that the questionnaire is an imperfect instrument for measuring diet. It may be subject to errors related to long term recall, choice of food items, portion size estimation, and other self–report biases. Note that the measurement errors in $\{W_{ij}, j = 1, \ldots, k_i\}$ for measuring long–term percent energy from fat may not be independent. In logistic regression, Wang, Carroll and Liang (1996) showed that ignoring correlated errors can cause serious bias in relative risk parameter estimation. Similar considerations are relevant in Cox regression (1972) analysis.

The third example we consider here is the estimation of the correlation between nutrient intake measures. As described above, the FFQ is cost–effective but may be associated with various errors. To evaluate the questionnaire's validity, the usual approach is to estimate the correlation coefficient between questionnaires and food records for nutrients of interest (Willett, et al., 1985). We note that eligibility criteria for the WHI exclude low fat intake participants in order to enhance study power to detect an intervention effect on clinical outcomes. To check the recruitment criteria, a baseline FFQ is given at a screening visit. At a later point in the eligibility screening process, 4 day food records are given to a subsample of eligible screenees. The relationship between assessments with these instruments in the sampled population is of interest. For example, one goal of the WHI Minority Dietary Assessment Study is to study and compare the validity of dietary instruments in different ethnic and socioeconomic groups. Based on food questionnaires, approximately 30% African American women are ineligible due to the recruitment requirement that a woman needs to have at least 32% energy from fat as assessed by questionnaire. Approximately 27% and 45% of Hispanic and white women, respectively, are not eligible. Note that only a random subsample of women meeting the questionnaire eligibility criteria were further given 4 day food records. In this setting, estimation of means, variances and correlation coefficients between questionnaire measured intakes and food records (or other methods) measured intakes based only on data from enrolled participants will be biased. The amount of bias depends on the critical value for the questionnaire which determines eligibility and on the distributions of fat assessments in the population screened. Note that if the relationship between questionnaire and food records is approximately linear, then a CC analysis will possibly underestimate the correlation coefficient since the range of data becomes smaller and the association is therefore weakened. If one is interested in comparing the correlation coefficients between subpopulations, such as ethnic groups, a comparison of correlation coefficients based on complete cases may lead to incorrect conclusions because subpopulations are likely to have different fat intakes and hence different exclusion rates.

We now consider an example of a child growth study with subjects identified from Group Health Cooperative, a health maintenance organization in Seattle. Overweight is of considerable public concern because it is related to cardiovascular disease, hypertension, and diabetes in adulthood and these are leading causes of death in the United States. We consider a subject's adult obesity as the outcome variable (Y). Obesity, a dichotomous variable, is defined as exceeding the 85th percentile of body mass index (BMI) for gender in adults using data on 21–29 year olds as a reference (NIH Consensus Development Conference Statement, 1995). Twenty-two percent of the adults in this study for whom data have been abstracted have been classified as obese at some point in their adult lives. In our illustration, we let X be a subject's long–term average of BMI's of age 2–7. Instead of observing X_i of subject i, BMI measurements W_{ij} are available at time t_{ij} for $j = 1, \ldots, k_i$ and $t_{ij} \in [2, 7]$. In this problem, we may assume that Y given X follows a logistic regression model and the primary interest is to estimate the parameters β_0 and β_1. Certainly, a naive estimator may be obtained from replacing X_i by $\overline{W}_{i\cdot}$, where $\overline{W}_{i\cdot} = \sum_{j=1}^{k_i} W_{ij}/k_i$. But, this naive estimator may cause a serious bias problem.

1.3 Missing covariate regression model

The analysis of missing covariate data is a practical and interesting problem in the regression analysis of a univariate response, Y, on a vector of covariates (X, Z), where the covariate X may be missing partially on a subset of records. Three missing mechanisms have been defined: *Missing Completely at Random* (MCAR) refers to missing data completely independent of (Y, X, Z); *Missing at Random* (MAR) implies that the missingness is associated with only those observed data (Rubin, 1976); and *Non–Ignorable Missing* implies that the missingness depends on the missing data. We consider the regression model $Y = g(Z, X, \beta) + \epsilon$ for a specified function g and $E(\epsilon|Z, X) = 0$. One objective of the regression analysis is to estimate regression coefficients β. For example, we often choose a linear regression $g(Z, X, \beta) = \beta_0 + \beta_1' Z + \beta_2' X$ to model the linear dependence of a continuous response on a vector of covariates, where $\beta = (\beta_0, \beta_1', \beta_2')'$, and choose a logistic regression $g(Z, X, \beta) = \{1 + \exp(-\beta_0 - \beta_1' Z - \beta_2' X)\}^{-1}$ to investigate the dependence of a binary response on a vector of covariates.

Complete-case (CC) analysis is based on the data $\{(Y_i, Z_i, X_i) : \eta_i = 1, i = 1, \ldots, n\}$, ignoring cases with one or more missing variables. Specifically, let $\widehat{\beta}$ denote the solution to the following estimating equation based on the complete cases:

$$\sum_{i=1}^{n} \eta_i \mathcal{X}_i \{Y_i - g(Z_i, X_i, \beta)\} = 0, \tag{1}$$

where $\mathcal{X}_i = (1, Z_i', X_i')'$ is a convenient choice, n is the total number of independent observations (including missing data), η_i is a binary indicator such that $\eta_i = 1$ if the covariate X_i is observed, and $\eta_i = 0$ otherwise, and $\pi_i = \mathrm{pr}(\eta_i = 1|Y_i, Z_i)$ denotes the probability of observing X_i (or the selection probability of X_i), which is assumed to be strictly positive. The CC analysis is simple but has two potential disadvantages: (i) loss of efficiency due to discarding data from incomplete cases; (ii) it may yield inconsistent estimates when the validation data is not a random subsample of the original cases. Nevertheless, as reviewed in Little (1992), CC analysis gives valid inference when the missingness depends only on the regressors. In addition, for logistic case-control sampling based on a binary outcome Y, CC analysis is valid for the slope estimates (Prentice and Pyke, 1979; Wang and Carroll, 1993).

1.4 Covariate measurement error model

In an epidemiological study, a gold standard measurement may be too expensive or impractical to measure. Instead, repeated measurements of a covariate variable may be available. A simple form of repeated measurement error model is

$$W_{ij} = X_i + U_{ij}, \qquad i = 1, \ldots, n, \qquad j = 1, \ldots, k_i,$$

where X_i is the "precise" covariate and $U_{ij}(i = 1, \cdots, k_i)$ are independent errors, and X_i is independent of U_{ij}. A classical additive measurement error model assumes that U_{ij} is normal with mean 0 and variance Σ_{uu}. A naive estimate of the regression coefficient can be obtained applying an ordinary regression with covariate data $(\overline{W}_{i\cdot}, Z_i)$, where $\overline{W}_{i\cdot} = \sum_{j=1}^{k_i} W_{ij}/k_i$. However, this naive estimator of β is in general biased (Hughes, 1993; Carroll, et al., 1995, Chapter 2).

As a special case, if $Y_i = \beta_0 + \beta_1 X_i + e_i$, where X_i has mean μ_x and variance σ_x^2, and e_i is independent of X_i, has mean zero and variance σ_e^2. If $\widehat{\beta}_N$ is obtained from an ordianry least square regression of Y_i on $\overline{W}_{i\cdot}$, then $\widehat{\beta}_N$ is not consistent for β. Under a multivariate normal assumption, when $k_i = k$ for $i = 1, \ldots, n$, the slope estimate $\widehat{\beta}_{N1}$ converges to $\lambda\beta_1$, where

$$\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2/k} < 1.$$

The attenuating factor, λ, is also called *reliability ratio.*

Exercise 1.3.1 Consider simple linear regression $Y_i = \beta_0 + \beta_1 X_i + e_i$, where $E(e_i|X_i) = 0$. Instead of observing X_i, $W_{ij} = X_i + U_{ij}$, $j = 1, \ldots, k$, are available for subject i, where both X and U are from independent normal distributions. Calculate $E(Y_i|\overline{W}_i)$ as a function of $\overline{W}_i$ and β.

1.5 Visualizing Measurement Error Effect

In Figure 1, we show the bias problem when the measurement error problem is ignored. We first generated 500 X values from a normal distribution with mean 1 and variance 1, each Y_i was generated by $1+X_i+e_i$, where e_i was from a normal distribution with mean 0 and variance 0.25. The surrogate variable was generated by $W_i = X_i + U_i$, where U_i was from a normal distribution with mean 0 and variance 1. The scatter plot of (W_i, Y_i) and three smoothing lowess (Splus function) lines are presented. Some simple Splus codes are given in the Appendix. It is seen that the solid line uses covariates X_i, the dotted line uses surrogates W_i and the broken line uses $E(X_i|W_i)$. The attenuation effect of the naive estimator can be seen since the dotted line is not as steep as the underlying solid line. We note also that the two lines have different intercepts; implying the effect on the estimation of β_0. The broken line was obtained using covariate $E(X_i|W_i)$. Note that when $(\mu_x, \sigma_x, \sigma_u)$ are known, then regressing Y_i on $E(X_i|W_i)$ leads to a consistent estimator. This can be seen from the figure that the broken line is very close to the solid line. Appendix A.1 provides some simple S-plus codes for the figure.

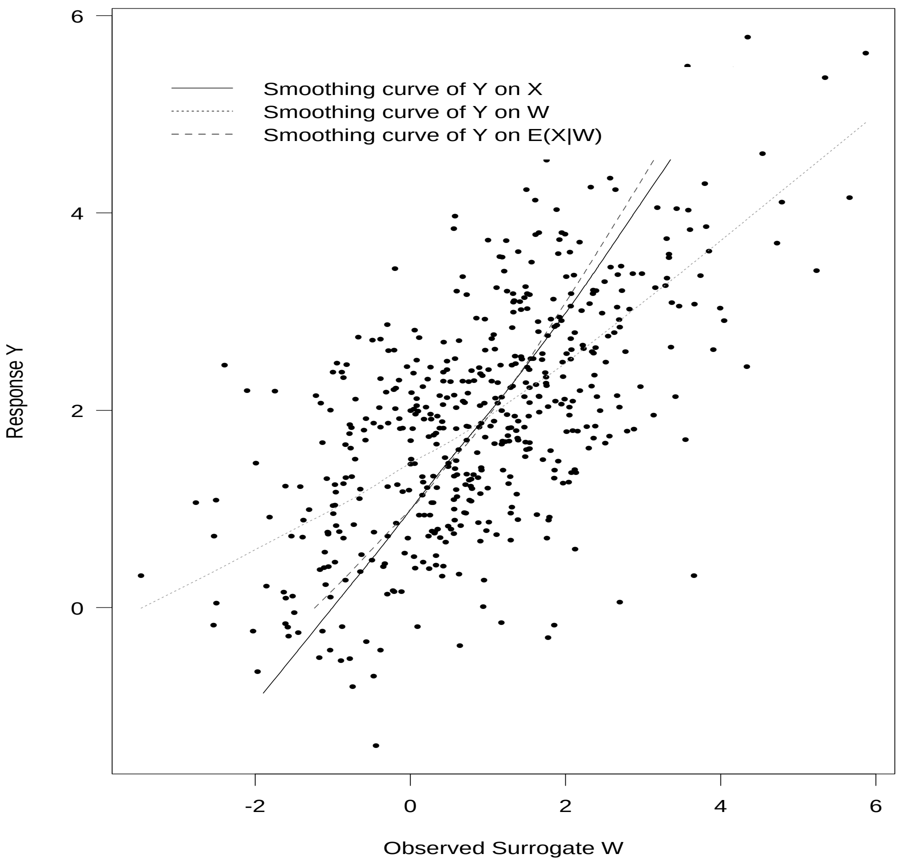

Figure 1: *Visualizing Attenuation Effect*

Appendix A.1: Splus codes for Figure 1

```
attenu <- function(n = 500){
    beta <- as.matrix(c(1,1));
    mux <- 1; sigmax <- 1;
    sigmau <- 1;
    xx <- 1 + (sigmax * rnorm(n));
    yy <- (cbind(as.matrix(rep(1,n)),xx)%*%beta) + (0.5 * rnorm(n));
    ww <- xx + (sigmau * rnorm(n));
    lambda <- (sigmax^2)/(sigmax^2 + sigmau^2);
    exgivenw <- (lambda * ww) + ((1 - lambda) * mux);
    xnew <- as.matrix(lowess(xx,yy)$x);
    y1hat <- as.matrix(lowess(xx,yy)$y);
    wnew <- as.matrix(lowess(ww,yy)$x);
    y2hat <- as.matrix(lowess(ww,yy)$y);
    exnew <- as.matrix(lowess(exgivenw,yy)$x);
    y3hat <- as.matrix(lowess(exgivenw,yy)$y);
    xmat <- cbind(xnew,wnew,exnew);
    ymat <- cbind(y1hat,y2hat,y3hat);
    motif();
    plot(ww,yy,las = 1,xlab = "Observed Surrogate W",ylab = "Response Y");
    matlines(xmat,ymat,type = "lll"); legend(min(ww),max(y1hat),bty = "n",
    legend = c("Smoothing curve of Y on X","Smoothing curve of Y on W",
    "Smoothing curve of Y on E(X|W)"),lty = 1 : 3)
    }
```

2 Regression Calibration

Abstract

Regression calibration a practical method in missing data and measurement error problems. In some sense, it is a type of *imputation*, or a replacement of a missing data. It is consistent when the primary regression model is linear. Even though it is not consistent, say in logistic regression, it is considered as a useful method because the magnitude of bias is much smaller than that of the relative risk parameter associated with the missing or mismeasured covariate variables. When measurement error is concerned in a medical study, say a type of plasma hormone concentration as a risk factor, this method has been well quoted. We will introduce it in the simplest models first, and then consider some straight-forward extensions.

2.1 General idea

The regression calibration (RC) method is a useful tool in the missing data and measurement error problem (Carroll, Ruppert and Stefanski, 1995, Chapter 3). In Cox regression, Prentice (1982) proposed the idea of estimating the radiation exposure due to atomic bombs. It has the advantages of (i) being conceptually simple; (ii) a simple version of it can be easily implemented; (iii) being efficient and it tends to have small bias as long as the relative risk is not too large (Wang, Hsu, Feng and Prentice, 1997; Wang and Pepe, 2000). It has been shown to have good properties in the noncensored data regression literature (*e.g.*, Rosner, Willett and Spiegelman, 1989; Carroll and Stefanski, 1990; Pierce, et al., 1992).

The regression calibration method may be implemented by the following algorithm:

(i) Using replication or validation data to estimate $E(X|Z,W) \equiv m(Z,W)$. For parametric calibration, we denote the resulting estimator of $E(X|Z,W)$ by $m(Z,W,\alpha)$, where α is an assumed parameter.

(ii) Replace an unobserved X by its estimate $m(Z,W,\widehat{\alpha})$ and run a standard analysis to estimate the parameter of interest, say β.

(iii) Estimate the standard error of the parameter estimate by a sandwich estimate or bootstrap.

A simple form of RC is the correction for attenuation (deattenuation). If the covariate measurement error is ignored, then the covariate effect can be much attenuated in (i) the simple linear regression model; and (ii) the classical additive error model (Carroll, Ruppert and Stefanski, 1995, Chapter 2). Without (i)-(ii), the effects of covariate measurement error can be complex. A similar consequence occurs when a sample average from finite repeated measurements is used while the long–term average is the true covariate.

The following is a deattenuation algorithm for the parameter of the effect of X, say β_x.

(i) Let $(\widehat{\beta}_N)_x$ be the naive estimator which is obtained by ignoring measurement error.

(ii) Let $\widehat{\sigma}^2_{W|Z}$ be the regression mean squared error from a linear regression of W on Z. This is the sample variance of W, $\widehat{\sigma}^2_w$, if there is no covariate Z.

(iii) Estimate β_x by $(\widehat{\beta}_N)_x\widehat{\sigma}^2_{W|Z}/(\widehat{\sigma}^2_{W|Z} - \widehat{\sigma}^2_u)$.

2.2 RC in the missing data problem

We first illustrate the method by a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + e_i,$$

where e_i is independent of X_i. We again use notation η_i for the indicator of observing X_i. We consider the case that X_i is missing completely at random, or the case that the selection probabilities do not depend on neither the response variable nor the missing data.

Recall that the CC analysis is based on the data $\{(Y_i, X_i) : \eta_i = 1, i = 1, \ldots, n\}$, ignoring cases with one or more missing variables. Some characteristics of it were described in Section 1.2. To illustrate the RC estimator, we assume that relationship between X and W is linear and that $E(X|W) = \alpha_0 + \alpha_1 W_i$ for some unknown α. To estimate α, we use the validation data by solving the estimating equation

$$\sum_{i=1}^{n} \eta_i \begin{pmatrix} 1 \\ W_i \end{pmatrix} (X_i - \alpha_0 - \alpha_1 W_i) = 0.$$

We denote the solution by $\widehat{\alpha}$ and let $\widehat{X}_i = \widehat{\alpha}_0 + \widehat{\alpha}_1 W_i$. Then the RC estimate of β solves

$$\sum_{i=1}^{n} \eta_i \begin{pmatrix} 1 \\ X_i \end{pmatrix} (Y_i - \beta_0 - \beta_1 X_i) + \sum_{i=1}^{n} (1-\eta_i) \begin{pmatrix} 1 \\ \widehat{X}_i \end{pmatrix} (Y_i - \beta_0 - \beta_1 \widehat{X}_i) = 0. \tag{2}$$

We note that: (i) solving (2) is equivalent to the estimator by regressing Y_i on $X_i^{\eta_i}\{\widehat{\alpha}_0 + \widehat{\alpha}_1 W_i\}^{1-\eta_i}$; (ii) if we write (2) as $U_n(\beta, \widehat{\alpha})$, then $U_n(\beta, \alpha)$ can be shown to be an unbiased estimating equation.

We now give a sketch of the proof of deriving the asymptotic variance of $\widehat{\beta}$.

(i) Write $0 = n^{-1/2} U_n(\beta, \alpha) - G(\beta, \alpha) n^{1/2}(\widehat{\beta} - \beta) + Q(\beta, \alpha) n^{1/2}(\widehat{\alpha} - \alpha) + o_p(1)$.

(ii) Write $n^{1/2}(\widehat{\alpha} - \alpha) = D^{-1}(\alpha) B_n(\alpha) + o_p(1)$. Then

$$n^{1/2}(\widehat{\beta} - \beta) = G^{-1}(\beta, \alpha)\{n^{-1/2} U_n(\beta, \alpha) + Q(\beta, \alpha) D^{-1}(\alpha) B_n(\alpha)\} + o_p(1).$$

(iii) $n^{1/2}(\widehat{\beta} - \beta)$ is asymptotically normally distributed with mean 0 and asymptotic variance

$$G^{-1}(\beta, \alpha) \mathrm{cov}\{n^{-1/2} U_n(\beta, \alpha) + Q(\beta, \alpha) D^{-1}(\alpha) B_n(\alpha)\}\{G^{-1}(\beta, \alpha)\}'.$$

Exercise 2.2.1: If Y_i given X_i is linear, as described in this subsection, what is the regression model of Y_i given W_i? If $\mathrm{pr}(\eta_i = 1 | Y_i, X_i, W_i)$ is not a function of Y_i, show that $U_n(\beta, \alpha)$ is unbiased, *i.e.*, it has expectation 0.

Exercise 2.2.2: In the proof of the asymptotic distribution of the RC estimator, find the expressions for $G(\beta, \alpha)$, $Q(\beta, \alpha)$, $D(\alpha)$ and $B_n(\alpha)$.

2.3 RC in the covariate measurement error problem

We consider the simple linear regression and the additive measurement error model that $W_{ij} = X_i + U_{ij}$ for $j = 1, \ldots, k$. We first illustrate the case with independent U_{ij}. Later in this subsection, we will demonstrate how to model the case when $U_{ij} = U_i(t_{ij})$, where t_{ij} are time points, $j = 1, \ldots, k_i$, for subject i.

Let $\Theta = (\beta_0, \beta_1, \mu_x, \sigma_x, \sigma_u)$, where the β's are the primary parameters and the rest are the nuisance parameters.

It was stated earlier that in this simpliest case the deattenuation technique leads to the RC estimator. We now consider an estimating equation approach for the purpose of easier extending to more complicated cases and of providing a simple method to obtain the asymptotic covariance of the resulting estimator. We note that under the bivariate normal assumption of $(\overline{W}_{i\cdot}, X_i)$, we have that $E(X_i|\overline{W}_{i\cdot}) = \lambda\overline{W}_{i\cdot} + (1-\lambda)\mu_x \equiv \mu_{X_i|\overline{W}_{i\cdot}}$, where $\lambda = \sigma_x^2/(\sigma_x^2 + \sigma_u^2/k)$.

$$\left\{\begin{array}{l} \sum_{i=1}^n \begin{pmatrix} 1 \\ \mu_{X_i|\overline{w}_{i\cdot}} \end{pmatrix} (Y_i - \beta_0 - \beta_1\mu_{X_i|\overline{w}_{i\cdot}}) = 0; \\ \sum_{i=1}^n (\overline{W}_{i\cdot} - \mu_x) = 0; \\ \sum_{i=1}^n \left\{\sum_{j=1}^k (W_{ij} - \overline{W}_{i\cdot})^2/(k-1) - \sigma_u^2\right\} = 0; \\ \sum_{i=1}^n \left\{(\overline{W}_{i\cdot} - \mu_x)^2 - \sigma_x^2 - \sigma_u^2/k\right\}. \end{array}\right. \tag{3}$$

Let $\widehat{\Theta}$ solve $\sum_{i=1}^n g_i(\Theta) = 0$, where $g_i(\Theta)$ is the summand in (3). Assume that $n^{-1}\sum_{i=1}^n(\partial/\partial\Theta')g_i(\theta)$ converges in probability to a negative definite matrix $G(\Theta)$. As in Reilly and Pepe (1995), using the result of Foutz (1977), there exists a solution $\widehat{\Theta}$ to (3) with probability converging to 1 as $n \to \infty$, and $\widehat{\Theta} \to \Theta$ in probability.

By some standard calculations,

$$n^{1/2}(\widehat{\Theta} - \Theta) = -G^{-1}(\Theta)n^{-1/2}\sum_{i=1}^n g_i(\Theta) + o_p(1).$$

Hence, it can be shown that $n^{1/2}(\widehat{\Theta} - \Theta)$ is asymptotically normally distributed with mean 0 and variance $G^{-1}(\Theta)V(\Theta)\{G^{-1}(\Theta)\}'$. The asymptotic variance of $n^{1/2}(\widehat{\Theta} - \Theta)$ can be consistently estimated by $G_n^{-1}(\widehat{\Theta})V_n(\widehat{\Theta})\{G_n^{-1}(\widehat{\Theta})\}'$, where $G_n(\Theta) = n^{-1}\sum_{i=1}^n(\partial/\partial\Theta')g_i(\Theta)$ and $V_n(\Theta) = n^{-1}\sum_{i=1}^n g_i(\Theta)g_i'(\Theta)$.

We now extend the method to the case with correlated measurement errors. Here X is a univariate covariate which is measured with error and the error follows an order–1 autoregressive model (AR(1)). We also consider unequally spaced times, $\{t_{ij}; i = 1, \ldots, n, j = 1, \ldots, k_i\}$, for the repeated measurements. For each subject i, instead of observing the true covariate X we observe $\tilde{W}_i$, where

$$W_i(t_{ij}) = X_i + U(t_{ij}); \qquad j = 1, \ldots, k_i.$$

We assume that X_i is normal with mean μ_x and variance σ_x^2, $E(\tilde{U}_i|X_i) = 0$ and $\mathrm{var}(\tilde{U}_i) = \Sigma_{U_i}$ and the (j,j')th element of Σ_{U_i} is equal to $\sigma_u^2 \rho^{|t_{ij}-t_{ij'}|}$ for $j, j' \in \{1, \ldots, k_i\}$ and ρ is the autocorrelation parameter. Under this model, $\Theta = (\beta', \mu_x, \sigma_x, \sigma_u, \rho)'$.

Now consider estimating equations for the nuisance parameters; the estimating equations for the primary parameters are the same as the first equation of (3). Note that for equal number of replicates k_i, a general idea of transforming $\tilde{W}_i$ to estimate η was given in Wang, et al. (1996). Because X_i is not observable, to estimate the parameter η_w. Wang and Pepe (2000) proposed a simple method which takes a first–order difference of W_{ij}. For $j = 1, \ldots, (k_i - 1)$, let $W_{ij}^* = W_{i(j+1)} - W_{ij}$, which has mean 0 and variance $2\sigma_u^2(1 - \rho^{|t_{i(j+1)}-t_{ij}|})$. Further, $\mathrm{cov}(W_{ij}^*, W_{i(j+1)}^*) = \sigma_u^2(\rho^{|t_{i(j+2)}-t_{i(j+1)}|} - 1 - \rho^{|t_{i(j+2)}-t_{ij}|} + \rho^{|t_{i(j+1)}-t_{ij}|})$. Therefore, to estimate η, the corresponding scores for the method of moments are

$$\begin{cases} \sum_{i=1}^n \left[\sum_{j=1}^{k_i-1} \left\{W_{ij}^{*2} - 2\sigma_u^2(1 - \rho^{|t_{i(j+1)}-t_{ij}|})\right\}\right] = 0; \\ \sum_{i=1}^n \left[\sum_{j=1}^{k_i-2} \left\{W_{ij}^* W_{i(j+1)}^* - \sigma_u^2(\rho^{|t_{i(j+2)}-t_{i(j+1)}|} - 1 - \rho^{|t_{i(j+2)}-t_{ij}|} + \rho^{|t_{i(j+1)}-t_{ij}|})\right\}\right] = 0; \\ \sum_{i=1}^n (\overline{W}_{i\cdot} - \mu_x) = 0; \\ \sum_{i=1}^n \left\{(\overline{W}_{i\cdot} - \mu_x)^2 - \sigma_x^2 - (1_{k_i}' \Sigma_{U_i} 1_{k_i})/k_i^2\right\} = 0. \end{cases}$$

Note that the estimating equations above do not require the normality of neither X_i nor U_{ij}. The limit distribution of the RC estimator of Θ can be derived by the same technique as we sketched earlier in the subsection.

Exercise 2.3.1: Show that (3) is indeed unbiased, *i.e.,* it has expectation 0.

2.4 *Refined RC Estimator for Binary Outcome*

Here we describe how to reduce the bias for logistic regression when the relative risk parameter is large. Consider the additive measurement error model described earlier. For binary outcome regression, the RC estimator may have a bias problem when either β_1 or $\mathrm{var}(X|\overline{W})$ is large, especially the former. A refined analysis may be used to reduce this bias problem. First, we consider probit regression such that $\mathrm{pr}(Y = 1|X) = \Phi(\beta_0 + \beta_1 X)$ where Φ is a standard normal distribution function. As in Carroll, et al. (1984) and Liang and Liu (1991),

$$\begin{aligned}\mathrm{pr}\{Y_i = 1|\overline{W}_i\} &= \int \Phi(\beta_0 + \beta_1 x) f(x|\overline{W}_i) dx \\ &= \int \Phi(\beta_0 + \beta_1 x) d\Phi\left(\frac{x - \mu_{x|\overline{W}_i}}{\sigma_{x|\overline{W}_i}}\right),\end{aligned}$$

where $\mu_{x|\overline{W}_i} = E(X|\overline{W}_i)$ and $\sigma^2_{x|\overline{W}_i} = (\sigma_x^2\sigma_u^2/k)/\{\sigma_x^2 + (\sigma_u^2/k)\}$. Let $\Phi(\frac{x - \mu_{x|\overline{W}_i}}{\sigma_{x|\overline{W}}}) = s$, then it can be shown that

$$\begin{aligned}\mathrm{pr}(Y_i = 1|\overline{W}_i) &= \int_0^1 \Phi(\beta_0 + \beta_1\Phi^{-1}(s)\sigma_{x|\overline{W}_i} + \beta_1\mu_{x|\overline{W}_i}) ds \\ &= \Phi\left((\beta_0 + \beta_1\mu_{x|\overline{W}_i})/(1 + \beta_1^2\sigma^2_{x|\overline{W}_i})^{1/2}\right). \qquad (4)\end{aligned}$$

Therefore, for probit regression, a refined analysis may be made based on (4).

For logistic regression, as in Liang and Liu (1991) that $\Phi(x) \approx H(1.7x)$, it can be shown that

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

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

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

The procedure above can be easily extended to the case with unequal number of replicates, k_i, for subject i. To estimate the standard error, a robust sandwich covariance estimator based on (2) may be applied. See Wang (2000) for details.

2.5 *A Simulation Study*

We now demonstrate the performance of the RC and the refined RC estimators via some numerical results. The results of Table 2.1 were obtained from logistic regression with covariate measured with error. We generated $n = 200$ X's independently from a standard normal distribution, and the corresponding Y's were obtained from a logistic regression model, with parameters given in

Table 2.1 Simulation Study on Logistic Regression with Covariate Measurement Error

		$\beta_1 = \ln(2)$			$\beta_1 = \ln(4)$			$\beta_1 = \ln(6)$		
		Naive	RC	RRC	Naive	RC	RRC	Naive	RC	RRC
$\sigma_u = 1$										
β_0	bias	-0.003	-0.002	-0.002	0.003	0.003	0.003	-0.001	-0.002	-0.002
	s.d.	0.145	0.145	0.149	0.162	0.165	0.178	0.165	0.168	0.188
β_1	bias	-0.171	0.006	0.024	-0.417	-0.086	0.013	-0.596	-0.186	-0.002
	s.d.	0.143	0.194	0.209	0.168	0.235	0.288	0.205	0.285	0.382
$\sigma_u = \sqrt{2}$										
β_0	bias	0.000	0.001	0.001	0.001	0.001	0.001	-0.001	0.000	0.000
	s.d.	0.148	0.152	0.157	0.157	0.166	0.185	0.160	0.174	0.203
β_1	bias	-0.287	-0.007	0.021	-0.652	-0.143	-0.003	-0.894	-0.277	-0.033
	s.d.	0.126	0.226	0.253	0.150	0.275	0.362	0.162	0.310	0.450

Note: Naive estimator uses $\overline{W}$ as the covariate, RC is the regression calibration using $E(X|\overline{W})$ as the covariate, RRC is the refined RC estimator. Results were from 1000 replicates.

the table. Instead of observing X_i, three repeated W_{ij} were available. We have also computed the naive estimator which used $\overline{W}$ as the covariate variable. We consider the magnitude of measurement error $\sigma_u = 1$ and $\sqrt{2}$, respectively. The biases and standard deviations of the three estimators were obtained from 1000 replicates. The RRC estimator in this simple study was obtained by multiplying the RC estimator by a coefficient of $(1 + \widehat{\beta}_1^2 \sigma^2_{X|\overline{W}}/2.89)^{1/2}$, with $\widehat{\beta}_1$ being that obtained from the RC estimator. It can be seen that the regression calibration significantly reduced the bias from the naive estimator and it performs very well when $\beta_1 = \ln(2)$. The refined RC estimator reduces the bias when β_1 is large, with the cost of lossing efficiency.

Exercise 2.5.1: In Table 2.1, it appears that the naive estimator does not have a bias problem on estimating β_0. This is probably because $\mu_x = 0$ was used in the simulation. Assume $Y|X$ is linear, and if $E(Y|\overline{W}) = \beta_0^* + \beta_1^* \overline{W}$, then write β as a function of β^*.

3 Two–stage Logistic Regression

3.1 Introduction

Logistic regression is a common tool to investigate factors related to disease incidence. Due to certain reasons such as high cost, some covariate data can only be collected for a small subsample. For example, LDL cholesterol may be related to the risk of heart diseases. However, to reduce the cost, it may be a plausible way to measure the total cholesterol of all study participants and then select a subset to further measure the LDL level. Here the total cholesterol can be treated as a surrogate for LDL. In general, let Y be the binary response, Z be a covariate which is always observed and X be another covariate that may be missing. Assume that W is a surrogate variable for X. We consider the logistic regression model

$$\mathrm{pr}(Y=1|X,Z)=H(\theta_0+\beta_1'Z+\beta_2'X), \tag{5}$$

where $H(u)=\{1+\exp(-u)\}^{-1}$ is the logistic distribution function and $\Theta=(\theta_0,\beta_1',\beta_2')'$ is a vector of parameters, Z is a covariate vector which is always observable and X is the vector of exposure variables, which is available for a subset (validation set) of individuals only. Let η_i be a random indicator of X_i being observed. The nonvalidation data set ($\eta_i=0$) consists of (Y_i,Z_i,W_i) and the validation data set ($\eta_i=1$) consists of (Y_i,X_i,Z_i,W_i). In this section we consider the case where X_i is assumed to be missing at random (MAR, Rubin (1976)) such that the probability of X_i being observed (selection probability) $\mathrm{pr}(\eta_i=1|Y_i,X_i,Z_i,W_i)=\pi(Y_i,Z_i,W_i)$ depends on (Y_i,Z_i,W_i) but not on X_i. In some cases, the data are obtained in two stages. At the first stage, $(Y_i,Z_i,W_i), i=1,\ldots,n$ are obtained for all subjects, and at the second stage X_i are measured in the validation data set.

Logistic regression with covariate measurement error models has been an active research area in recent years. Breslow and Cain (1988) proposed a pseudo conditional likelihood method for a two-stage case-control study such that at the second stage some X's are observed on each stratum classified by (Y,W) where W is a categorical surrogate. When the missingness of X does not depend on both outcomes and the missing values, Carroll and Wand (1991) and Pepe and Fleming

(1991) proposed semiparametric estimation which approximate the likelihood without modeling the distribution of X given (Z, W). Little (1992) reviewed related methods in this field. A mean-score method was proposed by Reilly and Pepe (1995) for discrete covariates when X is MAR. Robins, Rotnitzky and Zhao (1994) proposed efficient estimation by computing an optimal score function in semiparametric models. They showed that the optimal estimator attains the semiparametric variance bound in the sense of Begun, et al.(1983). Chen and Chen (1999) proposed a method which makes use of the multivariate-normality of the CC estimator and another nuisance estimator.

A case-control (retrospective) study is different from a cohort (prospective) study in the sampling design. Breslow and Cain (1988) investigated a two-stage case-control study when data are categorical. When all the X's are observed, Prentice and Pyke (1979) showed that the retrospective density of (Z, X) given $Y = y$ is

$$dF_{Z,X|Y=y}(z, x) = \frac{n}{n_y} H^y(\beta_0^* + \beta_1 z + \beta_2 x)\{1 - H(\beta_0^* + \beta_1 z + \beta_2 x)\}^{1-y} dQ(z, x),$$

where θ_0^* is a new intercept and $Q(z, x)$ is the marginal distribution function of (Z, X) under the case-control sampling scheme. Note that θ_0 is unidentifiable. They also showed that the usual logistic regression programs may be used for the estimation of the slopes (β_1, β_2), treating Y as a binary outcome. Hence, the prospective analysis of the case-control data gives consistent estimates of the slopes and the prospective covariance formulas for the slopes are correct asymptotically. Wang and Carroll (1993, 1995) extended the work to the literature of robust estimation. See also Carroll, Wang and Wang (1995) and Wang and Carroll (1994) for two stage case–controls studies.

Overall, there are various approaches to this type of problems – likelihood-based (including EM), multiple imputation (Rubin, 1987), semiparametric approaches (weighting, conditioning, etc). some of them will be described in detail next.

3.2 Likelihood Based on Conditional Distribution of Exposure

Satten and Kupper (1993) proposed methods using information on the probability-of-exposure when the missing exposure variable is categorical. Their methods are applicable to logistic regression with discrete exposure variables, and involve modeling the conditional probability of the exposure variable X given $(Z, W, Y = 0)$. We note that if validation is only on controls, then the class of Horvitz and Thompson (1952) type weighted estimators (Robins, et al., 1994) can not be applied because selection probabilities for cases will be zero. Wang, Wang and Carroll (1997) studied a prospective estimating equation method for two–stage case–control studies when X is either continuous or discrete. The methods of Satten and Kupper and Wang, Wang and Carroll (1997) are likelihood–based.

Consider now how to obtain the MLE for β. This approach was given in Wang, Wang and Carroll (1997) for case–control data using retrospective distributional calculations. We now consider the prospective likelihood function. The idea is to use observations (X, Z, W, Y)'s to link the relationship between the exposure X and (Z, W, Y), Following Satten and Kupper (1993), if X given $(Z, W, Y = 0)$ is modeled up to an unknown parameter α, then $\mathrm{pr}(Y = 1|Z, W) = H\{\beta_0 + \beta_1' Z + R(Z, W, \alpha, \beta)\}$, where $R(Z, W, \alpha, \beta) = \ln[E\{\exp(\beta_2' X)|Z, W, Y = 0\}]$. As noted in Wang, et al. (1997), modeling the conditional distribution of exposure in the controls (Y=0) as a member of the exponential family of distributions implies that X given (Z,W) is a mixture of two distributions from the same family. Specially, Wang, et al. (1997) showed that if X given (Z,W,Y=0) is a member of the exponential family of distributions (McCullagh and Nelder, 1989),

$$f_{X|Z,W,Y=0}(x) = C(x, \phi)\exp\{\frac{x\gamma + c(\eta)}{\phi}\}, \tag{6}$$

for the canonical parameter γ being a function of (Z,W) and scale ϕ, then X given $(Z, W, Y = 1)$ is from the same family but with canonical parameter $\gamma + \beta_2\phi$. Firstly we note that if we use $\mathcal{L}(Y|Z, W)$ to denote the likehood of Y given (Z,W), then $\mathcal{L}(Y, \eta X|Z, W) = \{\mathcal{L}(Y, X|Z, W)\}^{\eta}\{\mathcal{L}(Y|Z, W)\}^{1-\eta}$. Also $\mathcal{L}(Y, X|Z, W) = \mathcal{L}(Y|Z, W)\mathcal{L}(X|Z, W, Y)$. We note that $\mathcal{L}[X|Z, W, Y] = \mathcal{L}[X|Z, W, Y, \eta = 1]$ because X is MAR. If the missingness mechanism is not related to β, then we have that

$$\mathcal{L}[Y, \eta X|Z, W] = \mathcal{L}(Y|Z, W)\{\mathcal{L}(X|Z, W, Y, \eta = 1)\}^{\eta}.$$

The above is the unconditional likelihood of Satten and Kupper (1993) which involves parameters α and β. The MLE may be obtained by the method of scoring since $R(Z, W, \alpha, \beta)$ has an explicit form as long as X given (Z,W,Y) follows (6). The MLE under this setup is tractable and does not require complicated calculations such as the EM algorithm (Dempster, et al., 1977).

Wang and Huang (1999) proposed to relax modeling the conditional distribution of X given $(Z, W, Y = 0)$. Note that if $X|(Z, W, Y = 0)$ is normal, then $X|(Z, W)$ is a mixture of normals. On the other hand, if $X|(Z, W)$ is normal then the model assumption of $X|(Z, W, Y = 0)$ does not have a simple form. In this case, assuming normality of $(X|Z, W, Y = 0)$ may lead to bias estimation of β. Wang and Huang's estimation is to approximate the partial likelihood $\{P(Y|X, Z)\}^{\delta}\{P(Y|Z, W)\}^{1-\delta}$; discussed in Vach and Schumacher (1993). They observed that if the missingness does not depend on Y, then it is the same as $\{P(Y|X, Z, \delta = 1\}^{\delta}\{P(Y|Z, W, \delta = 0)\}^{1-\delta}$. Furthermore, when the missingness depends on (Y, Z, W), it is proportional to the likelihood $\{P(Y, X|Z, W)\}^{\delta}\{P(Y|Z, W)\}^{1-\delta}$ if $P(X|Z, W)$ does not depend on β. This is closely related to Carroll and Wand (1991) and Pepe and Fleming (1991).

3.3 Weighted estimator

Flanders and Greenland (1991) and Zhao and Lipsitz (1992) suggested a weighted estimator. In the spirit of Horvitz and Thompson (1952), the weighted estimating equation is

$$U_n(\beta, \pi) = n^{-1/2} \sum_{i=1}^{n} \frac{\eta_i}{\pi_i} \mathcal{X}_i\{Y_i - H(\beta_0 + \beta_1 + \beta_1' Z_i + \beta_2' X_i)\} = 0, \tag{7}$$

where $\mathcal{X}_i = (1, Z_i', X_i')'$ is a convenient choice, n is the total number of independent observations (including missing data), η_i is a binary indicator such that $\eta_i = 1$ if the covariate X_i is observed, and $\eta_i = 0$ otherwise, and $\pi_i = \text{pr}(\eta_i = 1|Y_i, Z_i)$ denotes the probability of observing X_i (or the selection probability of X_i), which is assumed to be strictly positive. It is easy to see that when the data are complete, the solution to (7) is equivalent to the maximum likelihood estimator of β in the cases of logistic or linear regression with normal error. It can be easily shown as in Zhao and Lipsitz (1992) that estimating equation (7) results in consistent and normally distributed estimates of the regression coefficients provided that the probability function π_i is correct. Note that the selection probability π_i is independent of the unobserved covariate X_i given (Y_i, Z_i, W_i) because we assume In our problem, π_i is a nuisance component and it remains to be estimated, although it may be prespecified at design stage in some applications. To estimate π_i, we may model the dependence of η_i on (Y_i, Z_i) via a logistic regression with a vector of nuisance parameters to be estimated from the data. This approach yields consistent estimates of the regression coefficients in the model provided that π_i is correctly specified.

In equation (7), the missing probabilities π_i $(i = 1, \ldots, n)$ are generally unknown. One approach is to assume a parametric function of π_i, as discussed in Robins et al. (1994). Let $V_i = (Y_i, Z_i, W_i)$. As an illustration, we assume that

$$\pi_i = \pi(Y_i, Z_i) = \mathrm{pr}(\eta_i = 1|Y_i, Z_i) = \{1 + \exp(-\alpha_0 - \alpha_1' V_i)\}^{-1}, \tag{8}$$

where α's are some unknown parameters. One interesting finding regarding this weighted estimator is about the efficiency gain due to the estimation of the nuisance parameter α.

Lemma 4.1. Assume that $\pi(w)$ follows model (8) with unknown nuisance parameters α. Let $\widehat{\beta}(\pi)$ and $\widehat{\beta}(\widehat{\pi})$ solve $U_n(\beta, \pi) = 0$ and $U_n(\beta, \widehat{\pi}) = 0$, respectively, where $\widehat{\pi}_i$ is the maximum likelihood estimate of π_i by fitting the model (7). Then $\mathrm{var}[n^{1/2}\{\widehat{\beta}(\widehat{\pi}) - \beta\}] \leq \mathrm{var}[n^{1/2}\{\widehat{\beta}(\pi) - \beta\}]$

The proof of the result is given in Appeneix (3.5). This phenomenon was also noted in the literature of survey sampling. Wang, Wang, et al. (1997) consider estimation when π is estimated nonparametrically by kernel smoothers and they noted the same phenomenon.

A more general class of weighted estimators considered in Robins, et al. (1994) is based on the following estimating equation:

$$n^{-1/2} \sum_{i=1}^{n} \frac{\eta_i}{\pi_i} \phi(Y_i, Z_i, X_i) + (1 - \frac{\eta_i}{\pi_i}) \psi(Y_i, Z_i, W_i) = 0, \tag{9}$$

where $\phi(Y_i, Z_i, X_i)$ is any estimating score such that $E\{\phi(Y_i, Z_i, X_i)\} = 0$ and $\psi(Y_i, Z_i, W_i)$ is any function. Robins, et al. discussed computations regarding the choice of ϕ and ψ in terms of efficiency. If semiefficiency is not a major concern, then one may consider a convenient choice that $\phi(Y_i, Z_i, X_i) = \mathcal{X}_i\{Y_i - H(\beta_0 + \beta_1' Z_i + \beta_2' X_i)\}$ and $\psi(Y_i, Z_i, W_i) = E\{\phi(Y_i, Z_i, X_i)|Y_i, Z_i, W_i\}$.

3.4 Mean score estimator

Recall that we used the notation V for $(Y, Z', W')'$ and now we denote its dimension by d. When the missingness process does not depend on outcome, Carroll and Wand (1991) and Pepe and

Fleming (1991) proposed semiparametric estimates of β which estimate the likelihood function without modeling the conditional distribution of X given V. Their estimates may be inconsistent when the selection of X depends on the response Y. When V is discrete, Reilly and Pepe (1995) generalized Pepe and Fleming's method and proposed a mean-score method for the case that the probability of $\eta_i = 1$ depends on both Y_i and V_i, i.e., X is MAR. To see their method, first we assume that the conditional density of X given (Y, V) in the validation is known, then an unbiased estimating equation is

$$\Phi_n(\beta) = n^{-1/2} \sum_{i=1}^{n} [\eta_i \phi(Y_i, X_i, Z_i) + (1 - \eta_i) E\{\phi(Y_i, X_i, Z_i) | Y_i, V_i, \eta_i = 1\}] = 0, \tag{10}$$

where ϕ is the estimating score when all data are observed. In this subsection, we concentrate on the case of logistic regression and hence $\phi(Y_i, X_i, V_i) = \mathcal{X}_i\{Y_i - H(\beta' \mathcal{X}_i)\}$. The method can be easily extended to other situations. The unbiasedness of $\Phi_n(\Theta)$ follows because $E\{\eta_i \phi(Y_i, X_i, Z_i)\} = E[\pi(Y_i, V_i) E\{\phi(Y_i, X_i, Z_i) | Y_i, V_i, \eta_i = 1\}]$ and $E[E\{\phi(Y_i, X_i, Z_i) | Y_i, V_i, \eta_i = 1\}] = 0$. Note that $E\{\phi(Y_i, X_i, Z_i) | Y_i, V_i, \eta_i = 1\} = E\{\phi(Y_i, X_i, Z_i) | Y_i, V_i\}$ since X_i is independent of η_i given (Y_i, V_i) for X_i is MAR. Because the conditional density of X given (Y, V) is unknown in general, when V is discrete, Reilly and Pepe suggested using $\widehat{E}\{\phi(Y_i, X_i, Z_i) | Y_i, V_i, \eta_i = 1\}$ which takes averages of $\phi(Y_k, X_k, Z_k)$ on the validation data such that $Y_k = Y_i$ and $V_k = V_i$.

When V is continuous, Wang and Wang (1997) suggested estimating the conditionally expected estimating score $E\{\phi(Y_i, X_i, Z_i) | Y_i, V_i\}$ in the nonvalidation data set by using a kernel smoother based on the validation data. Therefore, the corresponding estimating equation of the mean-score method is

$$\widehat{\Phi}_n(\beta) = n^{-1/2} \sum_{i=1}^{n} \left[\eta_i \phi(Y_i, X_i, Z_i) + (1 - \eta_i) \frac{\sum_{k \in N(Y_i)} \eta_k \phi(Y_k, X_k, Z_k) K_h(V_k - V_i)}{\sum_{k \in N(Y_i)} \eta_k K_h(V_k - V_i)} \right] = 0$$

with the solution $\widehat{\beta}$ as the mean-score estimate. The basic idea of the method is to use the score $\phi(Y_i, X_i, Z_i)$ when the data are complete, and use the estimated score when X_i is missing or mismeasured. Similar idea and calculations may be extended to surrogate endpoint problems (Pepe, Reilly and Fleming (1994)).

Under some regular conditions, the estimating score $\widehat{\Psi}(\beta)$ has the following linearization:

$$\widehat{\Phi}_n(\beta) = n^{-1/2}\sum_{i=1}^{n}[\eta_i\phi_i + (1-\eta_i)\{E(\phi_i|Y_i,V_i) + T_{ni}\}] + o_p(1),$$

where $T_{ni} = \{n_{VY_i}h^d f_{V|Y=Y_i,\eta=1}(V_i)\}^{-1}\sum_{k\in N(Y_i)}\eta_k R_{ki}K_h(V_k - V_i)$. The asymptotic distribution can be found in Wang and Wang (1997).

The mean-score estimator seems to have appealing motivation for discrete data. However, for continuous V, it may encounter the curse of dimensionality even the asymptotic normality is justified under *regular* conditions. Perhaps some further work on the implementation of the method to continuous V is useful for multivariate and continuous V. Finally, for discrete (Z,W), the mean-score estimator can be seen to be equal to the weighted estimator using estimated weights. Both estimators share the same spirit that for discrete (Y_i,V_i), if X_i is not observed, then the average score $E\{\phi(Y,Z_i,X_i|Y_i,Z_i,W_i\}$ obtained from the validation set is used for the estimation of β. For continuous data, Wang and Wang (1998) showed that they are asymptotically equivalent.

3.5 Appendix

Proof of Lemma 3.1: We first derive the asymptotic variance of $n^{1/2}\{\widehat{\beta}(\widehat{\pi}) - \beta\}$. Let $U_n(\beta,\alpha)$ be the same as in (7) and

$$G_n(\beta,\alpha) = n^{-1}\sum_{i=1}^{n}\frac{\eta_i}{\pi_i}\left[-\frac{\partial}{\partial\beta'}\mathcal{X}_i\{Y_i - H(\beta_0 + \beta_1' Z_i + \beta_2' X_i)\}\right].$$

Recall that we defined $\mathcal{X}_i = (1, Z_i', X_i')'$. Let $\mathcal{T}_i = (1, V_i')'$ and $D_n(\Theta,\alpha) = n^{-1}\sum_{i=1}^{n}\eta_i(1 - \frac{1}{\pi_i})\mathcal{X}_i\mathcal{T}_i'\{Y_i - g(Z_i,X_i,\beta)\}$. The consistency of $\widehat{\beta}(\widehat{\pi})$ for β can be shown by standard techniques as in Fortz (1977). Hence, by a Taylor series expansion, we have

$$n^{1/2}\{\widehat{\beta}(\widehat{\pi}) - \beta\} = G_n^{-1}(\beta,\alpha)\{U_n(\beta,\alpha) + D_n(\beta,\alpha)n^{1/2}(\widehat{\alpha} - \alpha)\} + o_p(1).$$

Also, $n^{1/2}(\widehat{\alpha} - \alpha) = A_n^{-1}(\alpha)B_n(\alpha) + o_p(1)$, where $A_n(\alpha) = n^{-1}\sum_{i=1}^{n}\mathcal{T}_i\mathcal{T}_i'\pi_i(1-\pi_i)$ and $B_n(\alpha) = n^{-1/2}\sum_{i=1}^{n}\mathcal{T}_i\{\eta_i - \pi_i\}$. It can be shown that $n^{1/2}(\widehat{\alpha} - \alpha)$ is asymptotically normally distributed with mean 0 and variance $A_n^{-1}(\alpha)$. The limit of $\mathrm{cov}(U_n, B_n)$ can be shown to be

$$E_n(\beta,\alpha) \quad = \quad n^{-1}\sum_{i=1}^{n}\mathcal{X}_i\mathcal{T}_i'(\frac{\eta_i}{\widehat{\pi}_i} - \eta_i)\{Y_i - g(Z_i,X_i,\beta)\}.$$

Furthermore, the asymptotic variance of $U_n(\Theta, \pi)$ is easily seen to be

$$V_n(\beta, \alpha) \quad = \quad n^{-1} \sum_{i=1}^{n} \frac{\eta_i}{\pi_i^2} \mathcal{X}_i \mathcal{X}_i' \{Y_i - g(Z_i, X_i, \beta)\}^2.$$

Notice that $E_n(\beta, \alpha) = -D_n(\beta, \alpha)$. Hence, by the central limit theorem $n^{1/2}(\widehat{\beta} - \beta)$ is asymptotically normally distributed with asymptotic covariance

$$G_n^{-1}(\beta, \alpha)\{V_n(\beta, \alpha) - E_n(\beta, \alpha) A_n^{-1}(\alpha) E_n'(\beta, \alpha)\} G_n^{-t}(\beta, \alpha).$$

We note that $E_n(\beta, \alpha) A_n^{-1}(\alpha) E_n'(\beta$ is positive definite. The proof of Lema 4.1 is completed by noting that the asymptotic variance of $n^{1/2}\{\widehat{\beta}(\alpha) - \beta\}$ is $G_n(\beta, \alpha) V_n(\beta, \alpha) \{G_n(\beta, \alpha)\}'$.

4 Logistic Regression with Covariate Measurement Error

Abstract

Logistic regression is an important tool in medical studies. Some variables such as radiation exposures, nutrient intakes, blood hormones, may be measured with error. In some situations a validation set may be available, in which all covariate variables are measured without error. Methods for missing data can often be applied to a measurement error problem when a validation set is available. However, often a gold standard measurement may not be available, and it leads to a pure measurement error problem. Replicated surrogate measurements are used as a reliability sample. Methods for this problem may be based on likelihood, conditional score or regression calibration approximation. The simulation extrapolation provides another approximation approach. Some methods will be described and examined.

4.1 *Logistic Regression and Regression Calibration*

In a binary regression problem, we are interested in the effects of some risk factors for a certain disease outcome variable. Let Y be the binary outocme variable, For covariate X, the odds of failure $\mathrm{pr}(Y=1|X)/\mathrm{pr}(Y=0|X)$ determines the likelihood function. Assume that

$$\frac{\mathrm{pr}(Y=1|X)}{\mathrm{pr}(Y=1|X)} = \exp(\beta_0 + \beta_1' X),$$

where β is the vector of paramters of interest. The exponential form for the odds has the advantage of being always positive. Define the logit transform of $\mathrm{pr}(Y=1|X)$ as the logarithm of the odds. The logit space of the likelihood consists of linear functions. Then it is easily seen that

$$\mathrm{pr}(Y=1|X) = H(\beta_0 + \beta_1' X_i), \quad \text{where } H(u) = \{1+\exp(-u)\}^{-1}.$$

The relative risk parameter β_1 has the advantage of being easy to interpret. For example, if $\beta_1 = \ln(2)$, then the relative risk (odds ratio ϕ) for the exposure at $X = x_0 + 1$ versus $X = x_0$ is 2.

If X_i were observed, then the maximum likelihood estimator of β solves

$$\sum_{i=1}^{n} (1, X_i')' \{Y_i - H(\beta_0 + \beta_1' X_i)\} = 0.$$

Solving the above equation can be done by a Newton-Raphson algorithm. Let the estimating score given above be $g(\beta)$, and $G(\beta) = \sum_{i=1}^n (1, X_i')'(1, X_i')H^{(1)}(\beta_0 + \beta_1' X_i)$. Then the iterative algorithm for solving β follows that $\widehat{\beta}_{(k+1)} = \widehat{\beta}_{(k)} + \{G(\widehat{\beta}_{(k)})\}^{-1} g(\widehat{\beta}_{(k)})$, where $\widehat{\beta}_{(k)}$ is the solution for iteration k and the algorithm stops if converges.

Suppose that in the logistic regression, X is the covariate variable that may be measured with error, while Z is measured without error. Write $\text{pr}(Y=1|X,Z) = H(\beta_0 + \beta_1' X_i + \beta_2' Z_i)$. Assume that observed surrogates $W_{ij} = X_i + U_{ij}, j = 1, \ldots, k$. Here k is the number of replicates for each subject. Let μ_x be the mean of a random variable X, Σ_x be the variance–covariance matrix of a random vector X and Σ_{xz} be the covariance matrix of (X, Z).

If both X_i and U_i are normal, then

$$E(X|\overline{W}, Z) = \mu_x + (\Sigma_x \; \Sigma_{xz}) \begin{bmatrix} \Sigma_x + \sigma_u^2/k_i & \Sigma_{xz} \\ \Sigma_{xz}' & \Sigma_z \end{bmatrix}^{-1} \begin{pmatrix} \overline{W} - \mu_x \\ Z - \mu_z \end{pmatrix}.$$

As described in Chapter 2, regression calibration (RC) provides a simple approximation for the measurement error problem. The RC estimator is simply to run a usual logistic regression with X_i replaced by $E(X_i|\overline{W}_i, Z_i)$ and hence it solves

$$\sum_{i=1}^n (1, E(X_i|\overline{W}_i)', Z_i')'[Y_i - H\{\beta_0 + \beta_1' E(X_i|W_i, Z_i) - \beta_2' Z_i\}] = 0.$$

The RC estimator is an approximation approach. The approximation error can be seen from a related work of Carroll and Stefanski (1990, Section 3). In addition, as in Carroll *et al.* (1995, page 65),

$$\text{pr}(Y_i = 1|\overline{W}_i, Z_i) \approx H\left(\frac{\beta_0 + \beta_1' E(X_i|\overline{W}_i) + \beta_2' Z_i}{(1 + \beta_1' \text{var}(X|\overline{W}_i, Z_i)\beta_1/2.89)^{1/2}}\right).$$

The above is related to the idea of the refined RC estimator. Hence it is clearly seen that approximately the bias of the RC estimator is an increasing function of the relative risk parameter $|\beta_1|$ and of the variance of the measurement error. For moderate sample sizes, the bias of the regression calibration estimator may be large if $|\beta_1|$ is large. Further, for any $|\beta_1|$, when the sample size n is sufficiently large, the standard error of the RC estimator will be dominated by the bias. However, when the relative risk is moderate, bias is often not a problem in realistic sample sizes.

Exercise: Let $\Psi(\cdot)$ be the standard normal cdf. For any real a, b, show that $\int_0^1 \Psi(a + b\Psi^{-1}(t))dt = \Psi(a/\sqrt{1+b^2})$. (Hint: this is used in the derivation of the refined RC estimator. May use convolution for the proof.)

4.2 *Maximum Likelihood Estimator*

Schafer and Purdy (1996) investigated likelihood-based estimation. The maximum likelihood (ML) estimator can be seen as an approach using conditioning. One simple way for the calculations is the expected estimating equation (EEE) approach of Wang and Pepe (2000). The ML estimator is based on an additional assumption on the distribution of X and U. Under the reasonable assumption that the distribution of $(\tilde{W}, X)$ does not involve β, then

$$\begin{aligned} \frac{\partial}{\partial \beta} \log P(Y|\tilde{W}) &= \frac{(\partial/\partial\beta) \int P(Y|X)P(\tilde{W}|X)P(X)dX}{P(Y, \tilde{W})} \\ &= E_\Theta\{\frac{\partial}{\partial\beta}\log P(Y|X)|Y, \tilde{W}\}. \end{aligned}$$

This indicates that the likelihood-based score of the observed data can be obtained by the conditional expectation of the likelihood-based score of the correctly measured true data given observed data. To calculate the conditional expectation, one needs to model the distribution of X_i and U_i. Let η be the set of nuisance parameters, and $\Theta = (\beta', \eta')'$. Let $\phi_\beta(Y_i, X_i) = (\partial/\partial\beta)\log P(Y|X)$, $\phi_{\eta_w}(\tilde{W}, X) = (\partial/\partial\eta_w)\log P(\tilde{W}|X)$ and $\phi_{\eta_x}(X) = (\partial/\partial\eta_x)\log P(X)$. The the ML estimate of Θ solves

$$\left\{ \begin{array}{l} \sum_{i=1}^n E_\Theta\{\phi_\beta(Y_i, X_i)|Y_i, \tilde{W}_i\} = 0; \\ \sum_{i=1}^n \Big[E_\Theta\{\phi_\eta(\tilde{W}_i, X_i)|Y_i, \tilde{W}_i\} + E_\eta\{\phi_\eta(\tilde{W}_i, X_i)|\tilde{W}_i\}\Big] = 0. \end{array} \right. \tag{11}$$

Calculations of the EEEs given above can be done by numerical integration or by the approximation of Crouch and Spiegelman (1990). For example, it is easy to see that

$$E_\Theta\{S_\beta(Y, X)|Y, \tilde{W}\} = \frac{\int S_\beta(Y, x)P(Y|x)P(\tilde{W}|x)P(x)dx}{\int P(Y|x)P(\tilde{W}|x)P(x)dx}.$$

The second equation of (11) is for the estimation of nuisance parameters, which can be done by some methods of moments. This leads to a pseudo ML estimator. It in general is slightly inefficient as the full ML estimator. One main motivation of the EEE approach is that it can be considered as an extension of a marginal regression analysis to longitudinal data. Further details can be seen from Wang and Pepe (2000).

Exercise: Show that

$$\frac{\partial}{\partial \eta_w} \log P(Y, \tilde{W}) = E_\Theta\{\frac{\partial}{\partial \eta_w} \log P(\tilde{W}|X)|Y, \tilde{W}\} + E_\eta\{\frac{\partial}{\partial \eta_w} \log P(\tilde{W}|X)|\tilde{W}\}.$$

4.3 *Functional Methods via Sufficiency and Conditional Scores*

In general, the RC estimator and ML estimator are not robust to the distributional assumption of X. Stefanski and Carroll (1985, 1987) proposed functional methods, which does not need the model assumption of X, by using sufficiency and conditional scores. Suppose the measurement error has a normal distribution and we treat X_i as a parameter, then the statistic $\Delta_i = \overline{W}_i + Y_i \Sigma_u \beta_1$ is complete and sufficient for X_i. As a result, the distribution of Y_i given Δ_i depends only on $Y_i, \overline{W}_i$ and β, but not on X_i. Further, $E(Y|\Delta) = H(\beta_0 + \beta_1 \Delta^* + \beta_2' Z)$, where $\Delta_i^* = \Delta_i - \Sigma_u \beta_1/(2k)$. The estimating equation for the sufficiency score estimator is

$$\sum_{i=1}^n (1, Z_i, \Delta^*)\{Y_i - H(\beta_o + \beta_1' \Delta^* + \beta_2 Z_i)\} = 0.$$

Of course, one may treat the sufficiency score estimator as a type of RC estimator by replacing X_i by Δ_i^*, but note that Δ_i^* is a function of Y_i and β_1. Stefanski and Carroll (1985) showed that the sufficiency score is optimal for a particular structural model. They also proposed a conditional score based on a very similar estimating equation:

$$\sum_{i=1}^n (1, Z_i, \Delta^{**})\{Y_i - H(\beta_o + \beta_1' \Delta^* + \beta_2 Z_i)\} = 0,$$

where $\Delta^{**} = \Delta_i - H(\beta_0 + \beta_1' \Delta_i + \beta_2' Z_i)\Sigma_u \beta_1/k$. Regarding the sufficiency and conditional scores, they stated that (Stefanski and Carroll, 1087, page 709) "the practical significance of the difference has not yet been investigated". The simulation results to be presented later show that the difference is in general very little, but it may be larger under some situations.

Implementing the sufficiency socre and conditional score can be done similarly to that of the RC estimator, see equation (3). Note that the estimating scores are independent over subjects and hence a sandwich covariance estimator can be easily applied. In general, one needs to calculate the estimating scores in a subroutine and the outer matrix of the sandwich estimator can be obtained by numerical derivatives.

4.4 *Simulation Study*

The results from the simulation study is to characterize the performance of the estimators. The unknown covariate X's were generated from a normal and a chi-squared with 1 degree of freedom, respectively, with mean 1 and variance 1. Then, outcome Y's were generated from the logistic regression model based on moderate and large relative risks, where the parameters are given in Table 4.1. For each subject, 3 replicated W_{ij} were generated. Measurement errors were generated from a normal distibution. The within/between subjects variation ratio $\sigma_u/\sigma_x = 1$. Sample sizes $n = 150$ and $n = 1000$ are considered. The distribution of X included a normal and a chi-squared with 1 degree of freedom, with mean 1 and variance 1. For RC, ML, sufficiency score and conditional score estimates, nuisance parameters were estimated by the method of moments, and hence the ML in this simulation is indeed a pseudo-ML estimator. The ML estimator was calculated based on normal X and normal error.

For the RC estimates, it performs well when $\beta_1 = \ln(2)$. When X is normal, its performance may be considered the best for the sample sizes we considered. But, the RC estimates deteriorate when the relative risk is large ($\beta_1 = \ln(4)$). As expected, the pseudo-ML estimator performs well if X is normal. Note that, even when X is normal, the pseudo-ML estimates may not have a small bias if the sample size is small. When X has a chi-squared distribution, the normal-based pseudo-ML is inconsistent and the resulting biases may be large when $\beta_1 = \ln(4)$, but not when $\beta_1 = \ln(2)$. Both the sufficiency and conditional score estimators are robust to the distributional assumption for X if the sample size is large (n=1000), but they may not have small biases when the sample size is small. When X is normal, the sufficiency and conditional score estimates are very close. But, the difference becomes larger when X is chi-squared distributed, n= 150, and the relative risk is large.

Table 4.1 Logistic regression with measurement error

		n=150					n=1000				
		Naive	RC	ML	SS	CS	Naive	RC	ML	SS	CS
	— Normal X, β_0 = -ln(2), β_1 = ln(2) —										
β_0	bias	0.166	-0.014	-0.042	-0.041	-0.041	0.186	0.016	-0.004	-0.004	-0.004
	s.d.	0.240	0.286	0.310	0.311	0.311	0.087	0.102	0.109	0.109	0.109
	mean(se)	0.236	0.279	0.300	0.301	0.301	0.090	0.106	0.113	0.113	0.113
	95% cov.	0.874	0.948	0.952	0.955	0.956	0.446	0.960	0.959	0.958	0.958
β_1	bias	-0.165	0.015	0.043	0.043	0.043	001841	-0.014	0.005	0.006	0.006
	s.d.	0.172	0.233	0.257	0.258	0.258	0.061	0.082	0.089	0.089	0.089
	mean(se)	0.162	0.220	0.243	0.244	0.244	0.062	0.083	0.090	0.091	0.091
	95% cov.	0.780	0.945	0.953	0.952	0.952	0.157	0.947	0.954	0.955	0.955
	— Normal X, β_0 = -ln(4), β_1 = ln(4) —										
β_0	bias	0.406	0.069	-0.089	-0.106	-0.116	0.437	0.119	-0.005	-0.005	-0.005
	s.d.	0.276	0.342	0.448	0.488	0.572	0.105	0.127	0.157	0.161	0.161
	mean(se)	0.272	0.332	0.428	0.469	0.462	0.103	0.125	0.155	0.157	0.157
	95% cov.	0.638	0.934	0.964	0.965	0.960	0.014	0.819	0.948	0.940	0.940
β_1	bias	-0.409	-0.073	0.085	0.101	0.108	-0.435	-0.117	0.007	0.007	0.007
	s.d.	0.202	0.279	0.388	0.433	0.472	0.074	0.102	0.133	0.136	0.136
	mean(se)	0.199	0.273	0.373	0.414	0.406	0.075	0.103	0.134	0.136	0.136
	95% cov.	0.448	0.910	0.964	0.962	0.959	0.001	0.774	0.946	0.946	0.945
	— Chi-squared X, β_0 = -ln(2), β_1 = ln(2) —										
β_0	bias	0.201	0.030	0.010	-0.061	-0.061	0.213	0.060	0.046	-0.008	-0.008
	s.d.	0.236	0.288	0.312	0.354	0.355	0.096	0.114	0.121	0.132	0.132
	mean(se)	0.234	0.284	0.307	0.360	0.358	0.089	0.107	0.113	0.125	0.125
	95% cov.	0.843	0.948	0.952	0.969	0.970	0.345	0.891	0.914	0.937	0.937
β_1	bias	-0.222	-0.049	-0.030	0.086	0.086	-0.241	-0.081	-0.073	0.010	0.010
	s.d.	0.164	0.235	0.262	0.374	0.376	0.065	0.088	0.095	0.124	0.124
	mean(se)	0.174	0.231	0.258	0.371	0.374	0.066	0.085	0.093	0.122	0.122
	95% cov.	0.719	0.932	0.949	0.967	0.967	0.054	0.799	0.855	0.955	0.955
	— Chi-squared X, β_0 = -ln(4), β_1 = ln(4) —										
β_0	bias	0.475	0.188	0.093	-0.111	-0.074	0.491	0.027	0.149	-0.020	-0.021
	s.d.	0.256	0.318	0.381	0.895	0.837	0.105	0.128	0.149	0.185	0.185
	mean(se)	0.256	0.328	0.394	0.576	0.541	0.098	0.124	0.144	0.180	0.179
	95% cov.	0.510	0.886	0.928	0.967	0.953	0.003	0.540	0.788	0.945	0.945
β_1	bias	-0.597	-0.306	-0.218	0.158	0.112	-0.608	-0.344	-0.272	0.026	0.026
	s.d.	0.179	0.257	0.320	0.865	0.830	0.075	0.103	0.124	0.211	0.211
	mean(se)	0.205	0.275	0.342	0.695	0.638	0.078	0.103	0.124	0.208	0.208
	95% cov.	0.180	0.722	0.832	0.951	0.939	0.000	0.114	0.399	0.954	0.954

NOTE: Nuisance parameters are $(\mu_x, \sigma_x, \sigma_u) = (1, 1, 1)$. There were 3 replicates for each subject. The calculations of the β part of pseudo-ML estimates were from normal assumptions of X, using numerical integration over 100 partitions. Results were from 1000 replicates.

5 Design and Analysis on Nutritional Data

Abstract

Nutritional data analysis is important in epidemiologic research. Some commonly used instruments include food frequency questionnaires, four day food records, 24 hour food recalls. However, measurement error is a main issue in this problem. Statistical modeling and design issues will be considered. We will described some estimating methods to this problem. The parameter of interest includes the correlation coefficient of two nutrient instruments and regression parameters of nutrient variables on disease incidences.

5.1 *Design and Modeling*

Reducing cancer incidence by dietary modification in the form of a low-fat eating pattern is a primary hypothesis tested in some randomized clinical trials such as the dietary component of the Women's Health Initiative (WHI), a large disease prevention trial and observational study involving 164,500 women sponsored by the National Institutes of Health and coordinated at the Fred Hutchinson Cancer Research Center. Among other instruments, the food frequency questionnaire (FFQ) is commonly used to assess the diet of individuals because its analysis can be automated and its use is cost efficient (Willett, 1990). For example, the FFQ used in screening eligibility for the dietary modification component of WHI, is one of the tools for assessing changes resulting from the WHI dietary intervention, and it is used to assess nutrient intakes in the WHI observational study. However, it may be subject to errors related to the limitations of long term recall, the choice of food items on the FFQ, the ability of respondents to accurately assess portion sizes, and other self–report biases. Additional more expensive instruments, such as repeated food records, usually are conducted in only a subsample of the study cohort. This subsample is sometimes called a *calibration sample*. Because this sample usually involves repeated measurements, such as food records, it is also called *reliability sample*. If the underlying intake is defined as the average intake reported over a long series of food records, it is of interest to estimate the correlation between the dietary intake measured from the FFQ and the corresponding long-term average intake (Willett, 1990).

A standard design for a calibration study is the stratified design. One illustrative example for this design is the American Association of Retired People (AARP) study. As discussed in Carroll, et al. (1997), the design of the main AARP study involves a two-stage sampling. At the first stage, a large number of randomly selected members are sent a FFQ to complete. At the second stage, a group of respondents are selected using stratified random sampling on the basis of the FFQ reported intake on a selected macronutrient of interest, such as fat. The stratified sampling ensures that subjects with extremely high or low reported intakes have a high probability of being selected. Some other instruments such as food records are then given. In this case, estimation of the correlation coefficient between those measured from the FFQ and food records from the calibration sample usually does not represent that for the study cohort due to the biased sampling scheme.

Let the total sample size of the study be denoted by n. For subject $i, i = 1, \ldots, n$, let Q_i be the FFQ measured intake, and W_{ij}, $j = 1, \ldots, k_i$, be the j-th food record measured intake, where k_i is the number of measurements. Further, denote the long-term average intake of interest for the i-th subject by X_i. A classical additive measurement error model assumes that $W_{ij} = X_i + U_{ij}$, where U_{ij} is the error associated with estimating X_i. Assume that U_{ij} has mean 0, variance σ_u^2; X_i has mean μ_x and variance σ_x^2. A standard approach to analyzing the variance components is to assume that U_{ij} is independent of $U_{ij'}$ if $j \neq j'$. As in Liu, et al. (1978), Freedman, et al., (1991), there is an attenuation effect if the sample correlation coefficient of $(Q_i, \overline{W}_{i\cdot})$ is applied to estimate the correlation coefficient of Q and X, denoted by ρ_{qx}. If $k_i = k$ and the errors U_{ij} are independent, then

$$\rho_{qx} = \rho_{q\overline{w}}\sqrt{1 + \sigma_u^2/(k\sigma_x^2)}, \tag{12}$$

where $\rho_{q\overline{w}}$ is the correlation coefficient of Q_i and $\overline{W}_{i\cdot}$.

The estimation of ρ_{qx} based on a sample correlation may lead to a larger bias problem if the calibration sample is not a simple random sample of the study cohort. Let the indicator of the subsample be denoted by η_i, *i.e.*, W_{ij} $(j = 1, \ldots, k_i)$ values are available if $\eta_i = 1$. Assume $\mathrm{pr}(\eta_i = 1|Q_i) = \pi(Q_i)$, for some selection probability function $\pi(\cdot) > 0$. For $\eta_i = 1$, observable measurements W_{ij} follow

$$W_{ij} = X_i + U_{ij};\ j = 1, \ldots, k_i; \quad \text{and} \quad \text{var}(\tilde{U}_i) = \Sigma_{ui},$$

where the j, j'th element of Σ_{ui} is the covariance of U_{ij} and $U_{ij'}$, $j, j' = 1, \ldots, k_i$. Wang (1999a) considered the situation in which the error process $U_{ij}, j = 1, \ldots, k_i$ follows an autoregressive model such that $\text{cov}(U_{ij}, U_{ij'}) = \sigma_u^2 \rho_u^{|j-j'|}$, for some $\rho_u \in (-1, 1)$.

5.2 *Mothods for Stratified Sampling*

One naive estimator is based on the deattenuation formula (12) from the calibration sample. Let $\overline{W}_{i\cdot} = k_i^{-1} \sum_{j=1}^{k_i} W_{ij}$,

$$\begin{aligned} \widehat{\mu}_x &= \sum_{i=1}^{n} \eta_i k_i \overline{W}_{i\cdot} / \sum_{i=1}^{n} \eta_i k_i, \quad \widehat{\sigma}_u^2 = \frac{\sum_{i=1}^{n} \eta_i \sum_{j=1}^{k_i} (W_{ij} - \overline{W}_{i\cdot})^2}{\sum_{i=1}^{n} \eta_i (k_i - 1)}, \\ \widehat{\sigma}_x^2 &= \frac{\left[\left\{\sum_{i=1}^{n} \eta_i k_i (\overline{W}_{i\cdot} - \widehat{\mu}_x)^2\right\} - (n_c - 1)\widehat{\sigma}_u^2\right]}{\{\sum_{i=1}^{n} \eta_i k_i - (\sum_{i=1}^{n} \eta_i k_i^2 / \sum_{i=1}^{n} \eta_i k_i)\}}, \end{aligned}$$

where $n_c = \sum_{i=1}^{n} \eta_i$. See Carroll, et al. (1995, Chapter 3) for the nuisance parameter estimation above. The naive estimator may be obtained by $\widehat{\rho}_{qx} = \widehat{\rho}_{q\overline{w}} n_c^{-1} \sum_{i=1}^{n} \eta_i \sqrt{1 + \widehat{\sigma}_u^2 / (k_i \widehat{\sigma}_x^2)}$.

However, due to the following two reasons the naive estimator may lead to a serious bias problem: (i) the selection bias on the calibration sample; (ii) the bias due to assuming independent errors, when in fact U_{ij} and $U_{ij'}$ may be dependent.

5.2.1 *Weighted Method of Moments*

When there is no missing data, estimation of the parameters modeling W_{ij} has been discussed in Wang, et al. (1996). Let $\tilde{W}_i$ be the vector of the repeated k_i measurements of the i-th subject. One approach is based on the method of moments of a transformation of $\tilde{W}_i$, such as $W_{ij}^* = W_{i(j+1)} - W_{ij}, j = 1, \ldots, k_i - 1$. To obtain the moment equations for σ_u and ρ_u, one can calculate the variance of W_{ij}^* $(= 2\sigma_u^2(1 - \rho_u))$ and the covariance of W_{ij} and $W_{ij'}$. Taking into consideration the Horvitz and Thompson (1952) inverse selection probability weights, joint estimating equations for $\Theta = (\rho_{qx}, \mu_q, \sigma_q, \mu_x, \sigma_x, \sigma_u, \rho_u)$ are:

$$\left\{\begin{array}{l}
\sum_{i=1}^{n} \frac{k_i\eta_i}{\pi(Q_i)} \left\{ \frac{(Q_i - \mu_q)(\overline{W}_{i\cdot} - \mu_x)}{\sigma_q\sigma_x} - \rho_{qx} \right\} = 0; \qquad \sum_{i=1}^{n}(Q_i - \mu_q) = 0; \\
\sum_{i=1}^{n} \left\{ (Q_i - \mu_q)^2 - \sigma_q^2 \right\} = 0; \qquad \sum_{i=1}^{n} \frac{k_i\eta_i}{\pi(Q_i)} (\overline{W}_{i\cdot} - \mu_x) = 0; \\
\sum_{i=1}^{n} \frac{k_i\eta_i}{\pi(Q_i)} \left\{ (\overline{W}_{i\cdot} - \mu_x)^2 - \sigma_x^2 - (1_{k_i}^t \Sigma_{ui} 1_{k_i})/k_i^2 \right\} = 0; \\
\sum_{i=1}^{n} \frac{\eta_i}{\pi(Q_i)} \left[\sum_{j=1}^{k_i-1} \left\{ W_{ij}^{*2} - 2\sigma_u^2(1-\rho_u) \right\} \right] = 0; \\
\sum_{i=1}^{n} \frac{\eta_i}{\pi(Q_i)} \left[\sum_{j=1}^{k_i-2} \left\{ W_{ij}^{*} W_{i(j+1)}^{*} - \sigma_u^2(2\rho_u - 1 - \rho_u^2) \right\} \right] = 0.
\end{array}\right. \tag{13}$$

The last 2 equations of (13) were derived from the autoregressive error process. The estimation of μ_q and σ_q is simply the sample mean and standard deviation from all subjects of the study cohort. The unbiasedness of the rest of the estimating equations (13) can be obtained from straight–forward calculations since for any $\phi(\tilde{W}_i, Q_i)$, $E\{\eta_i\phi(\tilde{W}_i, Q_i)/\pi(Q_i)\} = E[E\{\eta_i\phi(\tilde{W}_i, Q_i)/\pi(Q_i)|\tilde{W}_i, Q_i\}] = 0$ if $E\{\phi(\tilde{W}_i, Q_i)\} = 0$. By the unbiasedness of the estimating equations, the weighted method of moments (MoM) estimator can be shown to be a consistent estimator of Θ.

5.2.2 *Weighted Normality-based Estimator*

The method here makes the assumptions that X is normal with mean μ_x and variance σ_x^2; and that $\tilde{W}|X$ is multivariate normal with mean $X1_{k_i}$ and variance Σ_{ui}. As will become clear later, an estimator based on these two model assumptions is still consistent even if these normality assumptions are violated. To estimate the nuisance parameters, $(\mu_q, \sigma_q, \mu_x, \sigma_x, \sigma_u, \rho_u)$, related to the observed data $\tilde{W}_i$, one may consider its marginal distribution. Alternatively, to have the observed data estimating score, one may consider the conditional expectation of the full data estimating score (as a function of X also) given the observed data. Let $P(\tilde{W}), P(X)$ be the marginal distributions of $\tilde{W}, X$, respectively, and $P(\tilde{W}|X)$ be the conditional distribution of $\tilde{W}$ given X. It is easy to see that

$$\begin{aligned}
\frac{\partial}{\partial\mu_x}\log P(\tilde{W}) &= E\left\{\frac{\partial}{\partial\mu_x}\log P(X)|\tilde{W}\right\}; \quad \frac{\partial}{\partial\sigma_x}\log P(\tilde{W}) = E\left\{\frac{\partial}{\partial\sigma_x}\log P(X)|\tilde{W}\right\}; \\
\frac{\partial}{\partial\sigma_u}\log P(\tilde{W}) &= E\left\{\frac{\partial}{\partial\sigma_u}\log P(\tilde{W}|X)|\tilde{W}\right\}; \quad \frac{\partial}{\partial\rho_u}\log P(\tilde{W}) = E\left\{\frac{\partial}{\partial\rho_u}\log P(\tilde{W}|X)|\tilde{W}\right\}.
\end{aligned}$$

Taking into consideration the selection probabilities, Wang (1999a) the joint estimating equations for the normality–based estimator are

$$\left\{\begin{array}{l}
\sum_{i=1}^{n}\frac{k_i\eta_i}{\pi(Q_i)}\left\{\frac{(Q_i-\mu_q)(\overline{W}_{i\cdot}-\mu_x)}{\sigma_q\sigma_x}-\rho_{qx}\right\}=0; \qquad \sum_{i=1}^{n}(Q_i-\mu_q)=0;\\
\sum_{i=1}^{n}\left\{(Q_i-\mu_q)^2-\sigma_q^2\right\}=0; \qquad \sum_{i=1}^{n}\frac{\eta_i}{\pi(Q_i)}\Lambda_i(\tilde{W}_i-\mu_x)=0;\\
\sum_{i=1}^{n}\frac{\eta_i}{\pi(Q_i)}\left[\{\Lambda_i(\tilde{W}_i-\mu_x)\}^2-\Lambda_i(\sigma_x^2 1_k)\right]=0;\\
\sum_{i=1}^{n}\frac{\eta_i}{\pi(Q_i)}\mathrm{trace}\left[(\Sigma_{ui})^{-1}\{(\tilde{W}_i-\mu_{x|\tilde{W}_i})(\tilde{W}_i-\mu_{x|\tilde{W}_i})^t-\Sigma_{ui}+\sigma^2_{x|\tilde{W}}J_{k_i}\right]=0;\\
\sum_{i=1}^{n}\frac{\eta_i}{\pi(Q_i)}\mathrm{trace}\left[\Sigma_{u_i}^{-1}\{(\tilde{W}_i-\mu_{x|\tilde{W}_i})(\tilde{W}_i-\mu_{x|\tilde{W}_i})^t-\Sigma_{ui}+\sigma^2_{x|\tilde{W}}J_{k_i}\right]\Sigma_{u_i}^{-1}(\frac{\partial}{\partial\rho_u}\Sigma_{ui})=0.
\end{array}\right. \tag{14}$$

Wang, et al. (1996, page 404) demonstrated that the normality-based estimating equations for $(\mu_x, \sigma_x, \sigma_u, \rho_u)$ are unbiased. Wang (1999a) extended the unbiasedness result of robustness to a more general class of distributions to the missing data problem.

Proposition 1: Write (14) as $\sum_{i=1}^{n} S_i(\eta_i, Q_i, \tilde{W}_i, \Theta) = 0$ and let the resulting estimator be denoted by $\widehat{\Theta}$. If Q_i is from any distribution with mean μ_q and variance σ_q^2, and $\tilde{W}_i$ is from any multivariate distribution with mean $\mu_x 1_{k_i}$ and variance $\sigma_x^2 J_{k_i} + \Sigma_{ui}$. Assume that $G_n(\Theta) = n^{-1}\sum_{i=1}^{n}(\partial/\partial\Theta^t)S_i(Q_i, \tilde{W}_i, \Theta)$ converges in probability to a negative definite matrix $G(\Theta)$, and $V_n(\Theta) = n^{-1}\sum_{i=1}^{n} S_i(\eta_i, Q_i, \tilde{W}_i, \Theta)S_i^t(\eta_i, Q_i, \tilde{W}_i, \Theta)$ converges in probability to a positive definite matrix $V(\Theta)$. Then $n^{1/2}(\widehat{\Theta} - \Theta)$ is asymptotically normally distributed with mean 0 and variance $G^{-1}(\Theta)V(\Theta)\{G^{-1}(\Theta)\}^t$.

5.3 *Methods for Restricted Sampling*

This section will introduce methods for correlation coefficient estimation for bivariate data (Q, R) under the restricted sampling that R of a subject is observable only when the corresponding Q value is larger than a known cut-off point. For example, the WHI dietary modification clinical trial is a disease prevention trial and this restricted sampling is aimed for testing the dietary intervention effect on disease. This sampling design is not optimal for the purpose of estimating the correlation coefficient. To estimate the correlation coefficient, a standard design is to take 4DFR measurements from the whole range of FFQ measurements, with possibly different weights stratified by the FFQ values, the the previous section.

5.4 *Maximum Likelihood Estimator*

Let the observed data be $(Q_1, R_1), \ldots, (Q_{n_v}, R_{n_v})$ and $Q_{n_v+1}, \ldots, Q_n$, for some $n_v < n$. The R value is not available if the corresponding Q value is less than a truncation point κ. We assume (Q, R) has a bivariate normal distribution with parameter $\Theta = (\mu_Q, \mu_R, \sigma_Q, \sigma_R, \rho)'$, where the μ's represent the means, σ's the standard deviations and ρ the correlation coefficient. We say that the n_v pairs constitute the validation set. Note that $Q_{n_v+1}, \ldots, Q_n$ are less than the truncation point. The CC analysis simply uses the validation set without adjusting for the missingness process. The CC estimator, $\widehat{\Theta}_{CC}$, based on the sample mean, variance, and correlation coefficient is not valid because the R values are not missing completely at random.

Let $L(Q, R)$ and $L(Q)$ be the joint density of (Q, R) and the marginal density of Q, respectively, then the likelihood is $\prod_{i=1}^{n_v} L(Q_i, R_i) \prod_{i=n_v+1}^{n} L(Q_i)$. The likelihood has already taken the ignorable missingness mechanism into consideration. The likelihood of all the available data is proportional to

$$\prod_{i=1}^{n_v} |\Sigma|^{-1/2} \exp\{-\frac{1}{2}(Q_i - \mu_Q, R_i - \mu_R)\Sigma^{-1}(Q_i - \mu_Q, R_i - \mu_R)^t\} \prod_{i=n_v+1}^{n} \sigma_Q^{-1} \exp\{-\frac{(Q_i - \mu_Q)^2}{2\sigma_Q^2}\}. \quad (15)$$

The ML estimator is actually the most efficient estimator within the class of consistent and asymptotically normally distributed estimators under the bivariate normal sampling assumption.

Instead of solving estimating equations based on (15), one may reparametrize Θ and apply a convenient factorization of the likelihood. Let $\Theta^* = (\mu_Z, \sigma_Q^2, \beta_{R0\cdot Q}, \beta_{R1\cdot Q}, \sigma_{RR\cdot Q})$, where $\beta_{R1\cdot Q} = \sigma_{QR}/\sigma_Q^2$, $\beta_{R0\cdot Q} = \mu_R - \beta_{R1\cdot Q}\mu_Q$, $\sigma_{RR\cdot Q} = \sigma_R^2 - \sigma_{QR}^2/\sigma_Q^2$. Then the first part of (15) can be rewritten as a conditional normal likelihood $\prod_{i=1}^{n} f(R_i|Q_i, \Theta^*)$. The details of this approach were given in Little and Rubin (1987, Section 6). Because Θ^* is a one–one mapping of Θ, the uniqueness of the ML estimator of Θ^* implies that of Θ. This uniqueness of the ML estimator is interesting and practically important, especially when (Q, R) has a bimodal distribution.

Wang, et al. (1999) showed that the ML estimator based on a bivariate-normal model of (Q, R) has the following properties:

(i) $\widehat{\mu}_Q$ and $\widehat{\sigma}_Q$ are consistent for any random variable Q with finite second moment.

(ii) $\widehat{\mu}_R, \widehat{\sigma}_R$ and $\widehat{\rho}$ are consistent for any random variable R with finite second moment and such that R and Q are linearly associated so that $R_i = a + bQ_i + e_i$, where e_i satisfies $E(e_i|Q_i) = 0$ and $\mathrm{var}(R_i|Q_i) = \sigma_e^2$.

5.4.1 *Multiple Imputation*

In this section, a data augmentation method, multiple imputation, will be described. Instead of the likelihood function, assume the mean function of R given Q can be estimated by either the validation subset of the study data, or from an additional external validation set. Assume that R given Q is linear, $R_i = \alpha_0 + \alpha_1 Q_i + e_i^*$, where e_i^* is normal with mean 0 and variance σ^2. Let $\alpha = (\alpha_0, \alpha_1)^t$. The normal linear regression model imputation method (Rubin, 1987) involves drawing the parameters and data from their posterior distributions given the data. The normal linear regression model imputation has the following algorithm.

(i) Draw a value (α^*, σ^{*2}) from the posterior distribution of (α, σ^2) where $\sigma^{*2} = (n_v - 2)s^2/\chi^2_{(n_v-2)}$ and α^* is normal with mean $\widehat{\alpha}_v$, variance $\{\sum_{i=1}^{n_v}(1, Q_i)^t(1, Q_i)\}^{-1}\ \sigma^{*2}$.

(ii) Draw R_i from a normal distribution with mean $(1, Q_i)\alpha^*$ and variance σ^{*2} when R_i is not available, i.e. $i = n_v + 1, \ldots, n$. We denote it by R_i^*.

(iii) Repeat (i) and (ii) M times. For $m = 1, \ldots, M$, let $\widehat{\Theta}^{(m)}$ be the ML estimator of Θ using the mth created data set $\{(Q_1, R_1), \ldots, (Q_{n_v}, R_{n_v}), (Q_{n_v+1}, R^*_{n_v+1}), \ldots, (Q_n, R_n^*)\}$. Let V_m^w be the variance of the mth created data set.

(iv) Calculate $\widehat{\Theta}_{MI} = \sum_{m=1}^M \widehat{\Theta}^{(m)}/M$, $V = \overline{V}^w + (1 + 1/M)V^b$, where $\overline{V}^w = \sum_{m=1}^M V_m^w/M$ and $V^b = \sum_{m=1}^M(\widehat{\Theta}^{(m)} - \widehat{\Theta}_{MI})(\widehat{\Theta}^{(m)} - \widehat{\Theta}_{MI})^t/(M-1)$.

6 Cox Regression with Missing Covariates

Abstract

This chapter studies Cox (1972) regression with missing covariates. In this problem, regression calibration is a very practical approach although it may suffer from a bias problem under some situations. When covariate variables are discrete, the estimated partial likelihood estimator is a consistent estimator. Alternatively, EM is a standard approach although it may involve intensive calculations in addition to some technical complications when the missing covariates are continuous. These three estimators will be studied.

6.1 Introduction

We consider analyses of failure time regression with missing covariate data. In a sample of n subjects, let $T_i^0, i = 1, \ldots, n$, be the failure time of the ith subject, C_i be the censoring variable. Let $T_i = \min(T_i^0, C_i)$ be the observed time, and $\delta_i = I[T_i^0 \le C_i]$ be the noncensoring indicator for the ith subject. We consider the Cox (1972) hazard regression model

$$\lambda(t; X_i, Z_i) = \lambda_0(t)\exp(\beta_1' X_i + \beta_2' Z_i), \tag{16}$$

where $\lambda_0(\cdot) \ge 0$ is an unspecified baseline hazard function, X_i is a covariate vector of dimension p which might be missing in some study subjects, Z_i is a vector of the covariates that is always observed and $\beta = (\beta_1', \beta_2')'$ are the regression coefficients to be estimated. One advantage of the Cox model is that a consistent estimator may be obtained by maximizing the partial likelihood $\prod\{r_i(\beta, T_i)/\sum_{j=1}^n I[T_j \ge T_i] r_j(\beta, T_i)\}^{\delta_i}$, where $r_i(\beta, t) = \exp\{\beta_1' X_i(t) + \beta_2' Z_i(t)\}$. The partial likelihood estimator solves

$$n^{-1/2} \sum_{i=1}^n \delta_i \left\{ \begin{pmatrix} X_i \\ Z_i \end{pmatrix} - \frac{S^{(1)}(\beta, T_i)}{S^{(0)}(\beta, T_i)} \right\} = 0, \tag{17}$$

where

$$S^{(0)}(\beta, T_i) = n^{-1} \sum_{j=1}^n I[T_j \ge T_i] r(\beta, T_i), \qquad S^{(1)}(\beta, T_i) = n^{-1} \sum_{j=1}^n I[T_j \ge T_i] \begin{pmatrix} X_j \\ Z_j \end{pmatrix} r(\beta, T_i).$$

Solving (17) can be done by a Newton-Raphson algorithm, which is computationally simple and does not need to estimate $\lambda_0(t)$. This might be an important reason that the proportional hazards model has been well applied in censored data problems. The problem we will study is the case when some study subjects do not have their X values available, and the major interest is the estimation of β.

Let W be a surrogate measure for X, so that the hazard function for T is independent of W given (X, Z). Assume that a validation set is available where all variables (T, δ, X, Z, W) are observed. The nonvalidation set includes data only on (T, δ, Z, W), but not X. Assume that the selection probabilities of the validation data set depend possibly on (Z, W), but not on (T, δ, X). Prentice (1982) suggested the regression calibration approach in a measurement error problem, and Wang, Hsu, et al. (1997) studied the method when a validation exists. Lin and Ying (1993) proposed an approximated partial likelihood estimator when the selection probabilities do not depend on (Z, W). An estimated partial likelihood method was proposed by Zhou and Pepe (1995). Robins, et al. (1994) proposed a class of weighted estimators using inverse selection probability weights. *Exercise: In Splus, if "obtime" is an* $(n \times 1)$ *matrix for the observed time, "delta" is an* $(n \times 1)$ *matrix for the noncensored indicator, "X" is an* $(n \times m)$ *matrix for the covariates, and "eta" is an* $(n \times 1)$ *matrix for the validation indicator. Then "coxph(Surv(obtime,delta)∼ X)" is the function to obtain the partial likelihood estimates of* β. *How do you obtain the complete-case estimator which ignores subjects with missing covariates?*

6.2 Regression calibration

Prentice (1982) showed that under the independent censoring assumption, that given (X, Z, W) the hazard function is independent of the censorship, the induced hazard function given (Z, W) is

$$\lambda(t; Z, W) = \lambda_0(t)\exp(\beta_2' Z) E\{\exp(\beta_1' X) | Y(t) = 1, Z, W\}, \tag{18}$$

where $Y(t) = I[T \geq t]$ is an at risk indicator. Denote the relative risk functions for subjects with incomplete data as $\overline{r}_i(\beta, t) = \exp\{\beta_2' Z_i(t)\} E[\exp\{\beta_1' X_i(t)\} | S_i \geq t, Z_i(t), W_i(t)]$. From (18), it is easily seen that under normal measurement error and rare disease assumptions, the induced relative risk

$$\begin{aligned}&\exp\{\beta_2' Z\}E[\exp\{\beta_1' X\}|Y(t)=1,Z,W]\\&\approx \exp\{\beta_2' Z\}\exp\Big[\beta_1' E\{X|Z,W\}+\frac{1}{2}\beta_1'\mathrm{var}\{X|Z,W\}\beta_1\Big].\end{aligned}$$

From the equation given above, if the variance of X given (Z,W) does not depend on Z, then the induced relative risk is approximately equal to $c(t)\exp\{\beta_1' E(X|Z,W)+\beta_2' Z\}$, where the $c(t)$ is the resulting term from the conditional variance. Note that ignoring the conditioning on $Y(t)=1$ is equivalent to the situation when most subjects are still at risk, *i.e.* very few failures, which somewhat indicates a higher censoring rate. This argument suggests that if the relative risk parameter β_1 and $\mathrm{var}(X|Z,W)$ are not large, and the censoring rate is high, then the replacing a missing X by $E(X|Z,W)$ in a partial likelihood analysis will lead to a good estimation of β. Therefore, a standard Cox (1972) regression program based on the partial-likelihood estimating equation (such as the *coxreg* function in S-plus) may be applied to the replacement data for parameter estimation.

Wang, et al. (1997b) studied the implementation of the RC method in detail. The method they considered is to parametrically estimate $E(X|Z,W)$, assumed to be $g(Z,W,\mathcal{A})$ for some nuisance parameter $\mathcal{A}$, from the validation set and then replace a missing X by $g(Z,W,\widehat{\mathcal{A}})$. As a result, the RC estimator for estimating β can be carried out by the following two steps.

(i) Estimate the missing X by $g(Z,W,\widehat{\mathcal{A}})$, where $\widehat{\mathcal{A}}$ is estimated from the validation data.

(ii) Apply a Cox regression procedure by using covariate data (X,Z) in the validation set and $\{g(Z,W,\widehat{\mathcal{A}}),Z\}$ in the nonvalidation set. Equivalently, we use $\{\eta_i X_i+(1-\eta_i)g\{Z_i,W_i,\widehat{\mathcal{A}}\},Z_i\}$ as covariate data, where η_i indicates whether or not X_i is observed, $i=1,\ldots,n$.

If the relative risk is not too large, Wang, et al. found that this method has limited biases and is highly efficient in various situations. Their method is useful for both discrete and continuous covariate/surrogate variables. While not reported in their simulations, it is generally true that the biases will be greater if the censoring rate is lower and the conditional variance of $X|(Z,W)$ is larger.

Exercise: Under the assumption that the hazard function given (X, Z) is (16), show that the induced hazard is (18). Suppose further the failure time and censor random variable are independent given (X, Z, W), show that

$$\lambda(t, Z, W)h \approx E\{N((t+h)-) - N(t-)|T \geq t\},$$

where $h \to 0$. (Hint: using the calculations of Fleming and Harrington (1991, page 19), and the definition of surrogate)

6.3 Estimated partial likelihood estimator

This section considers discrete observed covariate Z and surrogate W. We can write the relative risk function given the observed covariates for an individual i as

$$R_i(\beta, t) = r_i(\beta, t)\eta_i + \overline{r}_i(\beta, t)(1 - \eta_i). \tag{19}$$

In (19), the induced relative risk given (Z, W), $\overline{r}_i(\beta, t)$, needs to be estimated. If the conditional distribution $f(X_i(t)|T_i \geq t, Z_i(t), W_i(t))$ is a known function up to a parameter θ, then the induced hazard function $R_i(\beta, t)$ is a known function up to parameters β and θ. Inference about β and θ therefore can be drawn from the usual partial likelihood based on $R_i(\beta, t)$. Since the conditional distribution is not readily available in practice, one could either make a parametric assumption or estimate the conditional distribution in order to incorporate the information in the incomplete cases. The condition $[T_i \geq t, Z_i(t), W_i(t)]$ generally implies that $\overline{r}(\beta, t)$ depends on the unspecified baseline hazard and underlying distribution of $X_i(t)$ and $Z_i(t)$. A parametric approach has the undesirable features that in addition to the marginal law of covariates, the nuisance baseline hazard function has to be specified. Zhou and Pepe (1995) used a non-parametric mean relative risk for a subject with missing $X_i(t)$ and proposed estimating β from the resultant estimated partial likelihood:

$$EPL(\beta) = \prod_{i=1}^{n} \Big[\frac{\widehat{R}_i(\beta, T_i)}{\sum_{j \in \mathcal{G}(T_i)} \widehat{R}_j(\beta, T_i)}\Big]^{\delta_i}, \tag{20}$$

where $\widehat{R}_i(\beta, t) = r_i(\beta, t)\eta_i + \widehat{\overline{r}}_i(\beta, t)(1 - \eta_i)$ is the estimated relative risk function for a subject i and $\mathcal{G}(S_i)$ is the risk set at time S_i. Zhou and Pepe (1995) investigated the case where the surrogate variable is discrete. They used $\widehat{r}_i(\beta,t)$, the empirical estimator of $r_i(\beta,t)$ based on the complete cases, that is,

$$\begin{aligned}\widehat{\widehat{r}}_i(\beta,t) &= \hat{E}[r_i(\beta)|T_i \geq t, Z_i(t), W_i(t)] \\ &= \frac{\sum_{j=1}^n \eta_j I_{[T_j \geq t, Z_j(t)=Z_i(t), W_j(t)=W_i(t)]} \exp\{\beta_1' X_j(t)\}}{\sum_{j=1}^n \eta_j I_{[T_j \geq t, Z_j(t)=Z_i(t), W_j(t)=W_i(t)]}} \exp\{\beta_2' Z_i(t)\}. \qquad (21)\end{aligned}$$

The empirical relative risk estimator requires $Z_i(t)$ and $W_i(t)$ to be discrete.

By (21), it is easily seen that the estimating equation corresponding to the estimated partial likelihood function (20) is

$$U_n(\beta) = n^{-1/2} \sum_{i=1}^n \delta_i \left\{ \frac{\widehat{R}_i^{(1)}(\beta, T_i)}{\widehat{R}_i(\beta, T_i)} - \frac{S^{(1)}(\beta, T_i)}{S^{(0)}(\beta, T_i)} \right\} = 0,$$

where

$$S^{(0)}(\beta, T_i) = n^{-1} \sum_{j=1}^n Y_j(T_i) \widehat{R}_j^{(0)}(\beta, T_i); \qquad S^{(1)}(\beta, T_i) = n^{-1} \sum_{j=1}^n Y_j(T_i) \widehat{R}_j^{(1)}(\beta, T_i).$$

One advantage of the EPL estimator is that it does not require the estimation of Λ_0. The work of Zhou and Pepe (1995) is for discrete covariate and surrogate variables. A similar approach was considered in Paik and Tsai (1997). The extension of this method to continuous covariate and surrogate variables using kernel estimation procedures was further developed by Zhou and Wang (1999). They noted that this approach may encounter the *curse of dimensionality* when (Z, W) has a higher dimension.

6.4 Nonparametric ML Estimator

Dempster, Laird and Rubin (1977) introduced the EM algorithm for computing maximum likelihood estimates when data are incomplete. Briefly speaking, the E-step is to calculate some conditional expectations, and the M-step is to maximize the complete data likelihood function. Wu (1983) provided some further theoretical justification of the convergence properties. The EM algorithm has been well applied to some practical incomplete data problems.

In this problem, the EM type Nonparametric maximumm likelihood (NPML) estimator has been proposed in Zhong, et al. (1996) and Chen and Little (1999). Instead of EM, the computations follow the expectation and conditional maximization (ECM), see Meng and Lubin (1993). The basic idea is to estimate the regression parameters, baseline cumulative hazard function $\Lambda_0(t)$ along with other related parameters iteratively. We consider discrete X, Z, W for the rest of the section to ease the presentation.

The likelihood function of the observed data can be written as $\prod_{i=1}^n L(T_i, \delta_i, Z_i, W_i, \beta, \Lambda_0, \gamma) = \prod_{i=1}^n E\{L(T_i, \delta_i | X_i, Z_i, \beta, \Lambda) | Z_i, W_i\} L(Z_i, W_i)$, where $L(Z_i, W_i)$ is the likelihood for (Z_i, W_i) and $L(T_i, \delta_i | X_i, Z_i, \beta, \Lambda) = L(T_i, \delta_i | X_i, Z_i)$ is the full likelihood function for the complete data

$$L(T_i, \delta_i | X_i, Z_i) = \{\lambda_0(T_i) \exp(\beta_1' X_i + \beta_2' Z_i)\}^{\delta_i} \exp\{-\Lambda_0 e^{\beta_1' X_i + \beta_2' Z_i}\}.$$

If $L(Z_i, W_i)$ and Λ_0 were known, then it can be easily shown that

$$\frac{\partial}{\partial \beta} \log L(T_i, \delta_i | Z_i, W_i) = E\{\frac{\partial}{\partial \beta} \log L(T_i, \delta_i | X_i, Z_i) | T_i, \delta_i, Z_i, W_i\}.$$

The equation above was also used for the idea of the expected estimating equation (Wang and Pepe, 2000). If X_i is observed, then given Λ_0, the estimating score for the full data for subject i is $S_i(T_i, \delta_i, X_i, Z_i) = (X_i, Z_i)' \{\delta_i - \Lambda_0 \exp(\beta_1' X_i + \beta_2' Z_i)\}$. When the X value for subject i is not observed, the induced score for β given observed data is

$$\phi_i(T_i, \delta_i, Z_i, W_i) = \delta_i \begin{pmatrix} E(X_i | T_i, \delta_i, Z_i, W_i) \\ Z_i \end{pmatrix} - \Lambda_0(T_i) e^{\beta_2' Z_i} \begin{pmatrix} E(X_i e^{\beta_1' X_i} | T_i, \delta_i, Z_i, W_i) \\ Z_i E(e^{\beta_1' X_i} | T_i, \delta_i, Z_i, W_i) \end{pmatrix}. \quad (22)$$

The condition expectation in (22) can be carried out by noting

$$\mathrm{pr}(X_i | T_i, \delta_i, Z_i, W_i) = \frac{L(T_i, \delta_i | X_i, Z_i) \mathrm{pr}(X_i | Z_i, W_i)}{\sum_{x \in \mathrm{supp}(X)} L(T_i, \delta_i | X_i = x, Z_i, W_i) \mathrm{pr}(X_i = x | Z_i, W_i)}. \quad (23)$$

Therefore, the estimation of the regression coefficients can be obtained using either $S_i(T_i, \delta_i, X_i, Z_i)$ or $\phi(T_i, \delta_i, Z_i, W_i)$ if Λ_0 and the other nuisance parameter γ involved in the distribution of $X|(Z, W)$ were known. To estimate the cumulative baseline hazard, we consider the Breslow (1974) estimator

$$\widehat{\Lambda}_0(t) = \sum_{j=1}^n \frac{I[T_j \le t] \delta_j}{\sum_{k=1}^n I[T_k \ge T_j] \exp(\beta_2' Z_k) E\{\exp(\beta_1' X_k | T_k, \delta_k, Z_k, W_i)\}}. \quad (24)$$

Again, the conditional expectation in (24) can be carried out by using (23). What is needed now is the estimation of the vector of parameters γ associated with $\mathrm{pr}(X|Z, W)$. A consistent estimator for $\mathrm{pr}(X_i = x|Z_i = z, W_i = w)$ without involving the iterative calculations is $\sum_{j=1}^n \eta_j I[X_i = x, Z_i = z, W_i = w]/\sum_{i=1}^n \eta_i I[Z_i = z, W_i = w]$. However, to be fully efficient the estimation of γ should be obtained by calculating the resulting conditional score which involves Λ_0 and β. Let the likelihood-based score for X_i given (Z_i, W_i) be $\psi_i(X_i, Z_i, W_i)$. Given fixed parameters for the conditional expectation of $(X|Z, T, \delta)$, the iterative estimation for γ is based on

$$\frac{\partial}{\partial\gamma} E[\log\{\mathrm{pr}(X_i|Z_i, W_i)] = E\{\psi_i(X_i, Z_i, W_i)|Z_i, W_i\}. \tag{25}$$

From the arguments given above, the NPML estimator can be described by the following. Let $\widehat{\beta}^{(0)}$, $\widehat{\Lambda}_0^{(0)}$ and $\widehat{\gamma}^{(0)}$ be starting values for β, Λ_0 and γ, respectively. Given $\widehat{\beta}^{(k)}$, $\widehat{\Lambda}_0^{(k)}$ and $\widehat{\gamma}^{(k)}$ $(k = 0, 1, \ldots)$, at the $(k+1)$th iteration,

(i) update $\widehat{\Lambda}_0^{(k+1)}$ by (24);

(ii) update $\widehat{\beta}^{(k+1)}$ by solving $U_n(\beta, \widehat{\Lambda}_0^{(k)}, \widehat{\gamma}^{(k)}) = 0$, where

$$U_n(\beta, \Lambda, \gamma) = \sum_{i=1}^n \delta_i \left[\begin{pmatrix} X_i^* \\ Z_i \end{pmatrix} - \frac{S^{(1)}(\beta, T_i)}{S^{(0)}(\beta, T_i)} \right];$$

$$X_i^* = \eta_i X_i + (1 - \eta_i) E(X_i|T_i, \delta_i, Z_i, W_i);$$

$$S^{(0)}(\beta, T_i) = n^{-1} \sum_{j=1}^n I[T_j \geq T_i] \widehat{r}(\beta);\ S^{(1)}(\beta, T_i) = n^{-1} \sum_{j=1}^n I[T_j \geq T_i] \widehat{r}^{(1)}(\beta);$$

$$\widehat{r}(\beta) = \eta_i r(\beta, X_i Z_i) + (1 - \eta_i)\exp(\beta_2' Zi) E\{\exp(\beta_1' X_i)|T_i, \delta_i, Z_i, W_i)\};$$

$$\widehat{r}^{(1)}(\beta) = \eta_i \begin{pmatrix} X_i \\ Z_i \end{pmatrix} r(\beta, X_i Z_i) + (1 - \eta_i)\exp(\beta_2' Z_i) \begin{pmatrix} E\{X_i \exp(\beta_1' X_i)|T_i, \delta_i, Z_i, W_i\} \\ Z_i E\{\exp(\beta_1' X_i)|T_i, \delta_i, Z_i, W_i\} \end{pmatrix}.$$

(iii) update $\widehat{\gamma}^{(k+1)}$ by (25).

The iterative procedure stops if $\widehat{\beta}^{(k+1)} \approx \widehat{\beta}^{(k)}$. Note that the estimation for γ may not be needed in the iterative process above. The standard error of the NPML estimator may follow the profile likelihood theory, or it can be obtained by bootstrap. Lipsitz and Ibrahim (1998) have also developed a similar method based on estimating equations.

6.5 *A simulation study*

This section demonstrates the finite sample performance of various estimator. The following methods were compared in this illustration. (i) CC estimator: The complete-case analysis which uses validation data only. (ii) Weighted estimator: The weighted complete-case analysis which applies the Horvitz-Thompson inverse-selection probability weights. (iii) APL estimator: The approximate partial-likelihood estimator proposed in Lin and Ying (1993). (iv) EPL estimator: The estimated partial likelihood method, proposed in Zhou and Pepe (1995) for discrete surrogates. (v) RC estimator: The regression calibration (RC) estimator (Wang, et al., 1997). (vi) NPML algorithm: The estimator described in the previous section.

Assume that the hazard function is $\lambda(t; X, Z) = \exp(\beta_1 X + \beta_2 Z)$. The data generating device involves first generating bivariate covariates (X, Z)'s such that $\mathrm{pr}(X = 0) = \mathrm{pr}(X = 1) = 0.5$; $\mathrm{pr}(Z = 0) = \mathrm{pr}(Z = 1) = 0.5$; $\mathrm{corr}(X, Z) = 0$. Note that the variable Z is not a surrogate because it has an effect on the hazard function given X. The censoring percentage was 50%. A total of 500 replicates were generated in each simulation configuration. In Table 6.1, "bias" means the average of $\widehat{\beta} - \beta$, where $\widehat{\beta}$ is an estimator of β. The "s.d." denotes the square root of the sample variance of the 1000 estimates. The "mean($\widehat{\text{s.e.}}$)" denotes the average of the 500 standard error estimates.

From the results, it can be seen that the performance of the EPL estimator is somewhat close to EM. The RC estimator performs well when the relative risk is not large. The APL estimator may be slightly better than the CC and the weighted estimator, but it is still quite inefficient. More detailed comparisons, including under various missignness mechanism, are given in the manuscript of Wang (1999).

Table 6.1: Bivariate covariate without surrogate. There were 50% missing X variables and 50% censoring percentage.

β		CC	WT	APL	EPL	RC	NPML
β_1=ln(2)	bias	0.005	0.010	-0.023	0.003	-0.009	-0.003
	s.d.	0.298	0.299	0.303	0.298	0.287	0.292
	mean($\widehat{\text{s.e.}}$)	0.297	0.297	0.292	0.296	0.289	0.296
	95 % cov.	0.955	0.965	0.945	0.965	0.955	0.940
β_2=-ln(2)	bias	0.001	0.018	-0.039	0.020	0.036	0.022
	s.d.	0.334	0.309	0.322	.232	0.222	0.230
	mean($\widehat{\text{s.e.}}$)	0.297	0.269	0.286	.210	0.209	0.222
	95 % cov.	0.930	0.930	0.930	.940	0.955	0.930
β_1=ln(4)	bias	0.042	0.046	-0.040	-0.034	-0.035	0.025
	s.d.	0.344	0.349	0.333	0.338	0.300	0.324
	mean($\widehat{\text{s.e.}}$)	0.325	0.328	0.316	0.321	0.299	0.308
	95 % cov.	0.945	0.945	0.950	0.950	0.960	0.935
β_2=-ln(4)	bias	0.060	-0.054	-0.054	-0.026	0.090	-0.019
	s.d.	0.374	0.323	0.357	0.280	0.242	0.282
	mean($\widehat{\text{s.e.}}$)	0.325	0.300	0.309	0.229	0.222	0.255
	95 % cov.	0.920	0.920	0.920	.920	0.895	0.955

Note: There were 50% missing X variables and 50% censoring percentage. Covariates X and Z are independent, and $\mathrm{pr}(X = 1) = 0.5 = \mathrm{pr}(Z = 1)$.

7 Cox Regression with Covariate Measurement Error

Abstract

We consider Cox regression when a covariate variable may not be measured precisely for all study subjects. Instead, repeated reliability measurements for the covariate of interest are available. We will consider some estimation methods in this problem, including regression calibration, corrected score estimator, EM algorithm.

7.1 *Introduction*

We consider analyses of failure time regression with missing covariate data. Let $(T_i, \delta_i), i = 1, \ldots, n$, be the observed times and noncensoring indicators for the ith study subject, where T_i is the minimum of the failure time and the potential censoring time. We consider the Cox (1972) hazard regression model

$$\lambda(t; X_i, Z_i) = \lambda_0(t)\exp(\beta_1' X_i + \beta_2' Z_i),$$

where $\lambda_0(\cdot) \geq 0$ is an unspecified baseline hazard function, X_i is a covariate vector of dimension p which is not available in the whole study subjects, Z_i is a vector of the covariates that is always observed (*e.g.* demographic information), and $\beta = (\beta_1', \beta_2')'$ are the regression coefficients to be estimated.

Instead of measuring X_i, we observed repeated surrogate variables W_{ij}, $j = 1, \ldots, k_i$,

$$W_{ij} = X_i + U_{ij}, \qquad i = 1, \ldots, n, \qquad j = 1, \ldots, k_i,$$

where k_i is the number of the repeated measurements for X_i and U_{ij} are the errors. A classical additive measurement error model assumes that U_{ij} is normal with mean 0 and variance σ_u^2. A naive estimate of the regression coefficient can be obtained by applying ordinary Cox regression with covariate data $(\overline{W}_{i\cdot}, Z_i)$, where $\overline{W}_{i\cdot} = \sum_{j=1}^{k_i} W_{ij}/k_i$. However, this naive estimator is biased in general (Hughes, 1993).

7.2 *Regression Calibration*

Prentice (1982) showed that under the independent censoring assumption that given (Z, W) the hazard function is independent of the censorship, the induced intensity process given (Z, W) is $\lambda\{t; Z, W\} = \lambda_0(t)\exp\{\beta_2' Z(t)\}E[\exp\{\beta_1 X(t)\}|Y(t) = 1, Z(u), W(u), u \leq t]$, where $Y(t) = I[T \geq t]$ is an at risk indicator. He then justified that under normal measurement error and high censoring rate assumptions, the induced relative risk

$$
\begin{aligned}
&\exp(\beta_2' Z)E\{\exp(\beta_1 X)|Y(t) = 1, Z, W\} \\
&\approx \exp(\beta_2' Z)\exp\Big\{\beta_1 E(X|Z, W) + \frac{1}{2}\beta_1' \mathrm{var}(X|Z, W)\beta_1\Big\}.
\end{aligned}
$$

This suggests that if the variance of X given (Z, W) does not depend on (Z, W), then replacing X_i by $E(X_i|Z_i, W_i)$ may work well. This replacement strategy, treating $E(X_i|Z_i, W_i)$ as X_i, is often called the regression calibration (RC) estimator (Carroll, et al., 1995, Chapter 3). This method is important in practice since it is easy to implement, and inference based on it is valid for testing $\beta_1 = 0$. See Tsiatis, et al. (1995) for an application of this method to the survival of AIDS patients. The rest of this section will introduce the robust sandwich covariance estimation of Wang (1999b).

The RC estimator estimates the unobservable X by $E(X|\overline{W}, Z)$. Let μ_x be the mean of a random variable X, Σ_{xx} be the variance–covariance matrix of a random vector X and Σ_{xz} be the covariance matrix of (X, Z). Assume that measurement errors U_{ij}, $j = 1, \ldots, k_i$, are independent with mean 0 and variance σ_u^2. Under a multivariate normal model,

$$
E(X|\overline{W}, Z) = \mu_x + (\Sigma_{xx}\ \Sigma_{xz}) \begin{bmatrix} \Sigma_{xx} + \sigma_u^2/k_i & \Sigma_{xz} \\ \Sigma_{xz}' & \Sigma_{zz} \end{bmatrix}^{-1} \begin{pmatrix} \overline{W} - \mu_x \\ Z - \mu_z \end{pmatrix}. \qquad (26)
$$

As mentioned in Carroll, et al. (1995, Chapter 2), even without the multivariate normal model assumption, (26) is still the best linear approximate to X given $(\overline{W}, Z)$. Let X_i^* be $E(X_i|\overline{W}_i, Z_i)$ and $\widehat{X}_i^*$ be a consistent estimator of $E(X_i|Z_i, W_i)$, using consistent estimates of nuisance parameters $(\mu_x, \sigma_x, \sigma_u, \mu_z, \sigma_z, \sigma_{xz})$. The RC estimator of β is to run a usual Cox regression program with covariates $(\widehat{X}_i^*, Z_i)$. The estimate of the covariance of the RC estimator of β can be obtained from the inverse of the information matrix, but as will be shown later in the simulation studies, this is a bias estimator. To correctly estimate the standard error of the RC estimate, one simple approach is to apply a sandwich estimate. For simplicity, we consider univariate X, Z and W. Define $\Theta = (\beta', \mu_x, \sigma_x, \sigma_u, \mu_z, \sigma_z, \sigma_{xz})'$, where σ_x is the standard deviation of a random variable X and σ_{xz} is the covariance of X and Z. Note that if $(\mu_x, \sigma_x, \sigma_u, \mu_z, \sigma_z, \sigma_{xz})$ were known, then the RC estimate of β can be obtained by solving the partial-likelihood estimating equation $\sum_{i=1}^n U_i(\beta) = 0$, where

$$U_i(\beta) = \delta_i \left\{ \begin{pmatrix} X_i^* \\ Z_i \end{pmatrix} - \frac{S^{(1)}(\beta, T_i)}{S^{(0)}(\beta, T_i)} \right\}, \tag{27}$$

where

$$S^{(0)}(\beta, T_i) = n^{-1} \sum_{j=1}^n Y_j(T_i) r(X_j^*, Z_j, \beta),$$
$$S^{(1)}(\beta, T_i) = n^{-1} \sum_{j=1}^n Y_j(T_i)(X_j^{*\prime}, Z_j')' r(X_j^*, Z_j, \beta),$$

and $r(X_i^*, Z_i, \beta) = \exp(\beta_1' X_i^* + \beta_2' Z_i)$. To estimate nuisance parameters $\phi \equiv (\mu_x, \sigma_x, \sigma_u, \mu_z, \sigma_z, \sigma_{xz})$, we consider solving $\sum_{i=1}^n \Psi_i(\phi) = 0$, where

$$\Psi_i(\phi) = \begin{pmatrix} k_i(\overline{W}_{i\cdot} - \mu_x) \\ k_i(\overline{W}_{i\cdot} - \mu_x)^2 - \sigma_u^2 - k_i\sigma_x^2 \\ \sum_{j=1}^{k_i}(W_{ij} - \overline{W}_{i\cdot})^2 - (k_i - 1)\sigma_u^2 \\ Z_i - \mu_z \\ (Z_i - \mu_z)^2 - \sigma_z^2 \\ (\overline{W}_{i\cdot} - \mu_x)(Z_i - \mu_z) - \sigma_{xz} \end{pmatrix}. \tag{28}$$

By the idea of *joint estimating equations*, Θ may be estimated by solving $\sum_{i=1}^n g_i(\Theta) = 0$ to give $\widehat{\Theta}$, where $g_i(\Theta) = \{U_i(\beta)', \Psi_i(\phi)'\}'$. Assume that $G_n(\Theta) \equiv n^{-1}\sum_{i=1}^n (\partial/\partial\Theta')g_i(\theta)$ converges in probability to a negative definite matrix $G(\Theta)$. As in Carroll, et al. (1995, Chapter 3), it is well-known that $\widehat{\Theta}$ in general does not converge to the true Θ, and we denote the limit by Θ_*. Note that $\Theta_* = (\beta_*', \phi')'$, where β_* solves $E\{\sum_{i=1}^n U_i(\beta)\} = 0$. Under some regular conditions, including the moment conditions imposed on $G_n(\Theta)$, it can be shown that $G_n(\Theta) \to G(\Theta)$ in probability uniformly in a neighborhood of Θ_*. Further by some standard techniques as these in the M-estimates (Huber, 1981, Chapter 3), $\widehat{\Theta}$ is consistent for Θ_* and is unique in a neighborhood of Θ_* with probability converging to 1. Therefore, by a Taylor series expansion,

$$n^{1/2}(\widehat{\Theta} - \Theta_*) = -G^{-1}(\Theta_*) n^{-1/2} \sum_{i=1}^n g_i(\Theta_*) + o_p(1).$$

The $\Psi(\phi)$ part of $g(\Theta)$ is the sum of n iid random variables, but which is not the case for the $U(\beta)$ part of $g(\Theta)$. Let $N_i(t) = I[T_i \le t, \delta_i = 1]$ be the failure time counting process. Similar to Wang, et al. (1997), let $M_i(t) = N_i(t) - \int_0^t Y_i(t)\lambda(u, X_i^*, Z_i)du$, $i = 1, \ldots, n$, which can be shown to be independent local square integrable martingales on [0,1] under some regular conditions as in Anderson and Gill (1981). See also Fleming and Harrington (1991). Let $s^{(j)}(\beta,t) = E\{S^{(j)}(\beta,t)\}$ for $j = 0, 1$. Then it can be shown that there exists a neighborhood $\mathcal{B}$ of β_* such that $\sup_{t\in[0,1],\beta\in\mathcal{B}}\|S^{(j)}(\beta,t) - s^{(j)}(\beta,t)\| \to 0$ in probability. Therefore, it can be easily seen that $U_i(\beta)$ can be approximated by an iid sum

$$U_i(\beta) = \int \left\{ \begin{pmatrix} X_i^* \\ Z_i \end{pmatrix} - \frac{s^{(1)}(\beta,u)}{s^{(0)}(\beta,u)} \right\} dM_i(u) + o_p(1).$$

Indeed, Lin and Wei (1989) showed that even if the proportional hazards model is misspecified, we have $U_i(\beta) = U_i^*(\beta) + o_p(1)$, where

$$\begin{aligned} U_i^*(\beta) &= \int \left\{ \begin{pmatrix} X_i^* \\ Z_i \end{pmatrix} - \frac{s^{(1)}(\beta,t)}{s^{(0)}(\beta,t)} \right\} dN_i(t) \\ &\quad - \int \frac{Y_i(t) r(X_i^*, Z_i, \beta)}{s^{(0)}(\beta,t)} \left\{ \begin{pmatrix} X_i^* \\ Z_i \end{pmatrix} - \frac{s^{(1)}(\beta,t)}{s^{(0)}(\beta,t)} \right\} d\tilde{F}(t), \end{aligned}$$

where $\tilde{F}(t) = E\{\sum_{i=1}^n N_i(t)/n\}$. Notice that $U_i^*(\beta) (i = 1, \cdots, n)$ are iid. Thus the asymptotic normality of $n^{1/2}\sum_{i=1}^n g_i(\Theta)$ follows from the multivariate central limit theorem. Define $a^{\otimes 2} = aa'$ for any vector a. Then, $n^{1/2}(\widehat{\Theta} - \Theta_*)$ has an asymptotic normal distribution with mean 0 and variance $G^{-1}(\Theta_*)V(\Theta_*)\{G^{-1}(\Theta_*)\}'$, where $V(\Theta_*) = n^{-1}\sum_{i=1}^n\{g_i^*(\Theta_*)\}^{\otimes 2}$, $g_i^*(\Theta) = \{U_i^*(\beta)', \Psi_i(\phi)'\}'$. Let

$$\begin{aligned} \widehat{U}_i^*(\beta) &= \delta_i \left\{ \begin{pmatrix} X_i^* \\ Z_i \end{pmatrix} - \frac{S^{(1)}(\beta,T_i)}{S^{(0)}(\beta,T_i)} \right\} \\ &\quad - n^{-1}\sum_{j=1}^n \frac{\delta_j Y_i(T_j) r(X_i^*, Z_i, \beta)}{S^{(0)}(\beta,T_j)} \left\{ \begin{pmatrix} X_i^* \\ Z_i \end{pmatrix} - \frac{S^{(1)}(\beta,T_j)}{S^{(0)}(\beta,T_j)} \right\}, \end{aligned}$$

and $\widehat{g}_i^*(\Theta) = \{\widehat{U}_i^*(\beta)', \Psi(\phi)'\}'$. Then $V(\Theta) = \widehat{V}(\Theta) + o_p(1)$, where $\widehat{V}(\Theta) = n^{-1}\sum_{i=1}^n\{\widehat{g}_i^*(\Theta)\}^{\otimes 2}$, $\widehat{g}_i^*(\Theta) = \{\widehat{U}_i^*(\beta)', \Psi(\phi)'\}'$. Therefore, under some regulat conditions it can be shown that

Proposition 7.1: For any underlying hazard function $\lambda(t)$ and any random variables X, Z, $n^{1/2}(\widehat{\Theta} - \Theta_*)$ is asymptotically normally distributed with mean 0 and a covariance matrix which can be consistently estimated by $G_n^{-1}(\widehat{\Theta})\ \widehat{V}(\widehat{\Theta})\{G_n^{-1}(\widehat{\Theta})\}'$.

Extension of the covariance estimation of $\widehat{\Theta}$ to the situation when X, Z and W are vectors requires mainly more complicated notation. Briefly, we may define the parameter vector Θ as $(\beta', \mu_x', \mathrm{vec}(\Sigma_{xx})', \mathrm{vec}(\Sigma_{uu})', \mu_z', \mathrm{vec}(\Sigma_{zz})', \mathrm{vec}(\Sigma_{xz})')'$. The joint estimating equations for ϕ are similar to (28) except that the last six equations corresponding to the estimation of Σ_{uu}, Σ_{xx}, Σ_{zz} and Σ_{xz} need to take their vector equations.

A refined analysis to RC is to calibrate within a risk set. Clayton (1991) proposed a form of risk set calibration, which is restricted in that the slope of the regression line $E\{X|Z, W, Y(t) = 1\}$ was held constant across risk set (Carroll, et al., 1996, page 256). Xie, et al. (1998) proposed a more relaxed risk set calibration in which $E\{X|Z, W, Y(t) = 1\}$ is a linear function of (Z, W) and both of the slope and the intercept vary across risk sets.

7.3 An EM Algorithm

Zhong, et al. (1996) proposed to estimate the regression parameters and baseline cumulative hazard function $\Lambda_0(t)$ simultaneously by an EM algorithm. They considered the likelihood function $\prod_{i=1}^n L(\beta, Z, W, \Lambda_0) = \prod_{i=1}^n E\{L(\beta, X, Z, \Lambda_0)|Z, W\}$, where $L(\beta, X, Z, \Lambda_0)$ is the full likelihood function for complete data

$$L(\beta, X_i, Z_i, \Lambda_0) = \prod_{t\in[0,1]} \left[Y_i(t) e^{\beta_1^t X_i + \beta_2^t Z_i} \Lambda_0(dt)\right]^{dN_i(t)} \exp\left[-\int_0^1 Y_i(t) e^{\beta_1^t X_i + \beta_2^t Z_i} \Lambda_0(dt)\right].$$

The E–step of the rth iteration of the EM algorithm is to estimate the expected log likelihood and this involves the estimation of $E\{X|\mathcal{G}^{(r)}\}$ and $E\{e^{\beta_1^t X}|\mathcal{G}^{(r)}\}$, where $\mathcal{G}^{(r)} = \{T_i, \delta_i, Z_i, W_i, \beta^{(r)}, \Lambda_0^{(r)}(t)\}$ and $\beta^{(r)}, \Lambda_0^{(r)}$ are the estimates of β, Λ_0 at the rth iteration. The conditional density function of X given $\mathcal{G}^{(r)}$ is

$$f_{X|\mathcal{G}^{(r)}}(x|\mathcal{G}^{(r)}) = \frac{\exp\{\delta_i(\beta_1^{(r)t} x + \beta_2^{(r)t} Z_i) - \Lambda_0^{(r)}(T_i) e^{\beta_1^{(r)t} x + \beta_2^{(r)t} Z_i}\} f_{X|Z_i,W_i}(x|Z_i, W_i)}{\int_u \exp\{\delta_i(\beta_1^{(r)t} u + \beta_2^{(r)t} Z_i) - \Lambda_0^{(r)}(T_i) e^{\beta_1^{(r)t} u + \beta_2^{(r)t} Z_i}\} f_{X|Z_i,W_i}(u|Z_i, W_i) du}.$$

Note that the log likelihood involves the unknown baseline hazard function. Zhong, et al. applied the Nelson–Aalen estimator

$$\widehat{\Lambda}_0(t) = \int_0^t \frac{1}{\sum_{i=1}^n Y_i(u) E(e^{\beta_1^t X_i}|\mathcal{G}) e^{\beta_2^t Z_i}} dN(u).$$

In the M–step, given $\widehat{\beta}^{(r)}, \widehat{\Lambda}_0^{(r)}$, the log–likelihood is maximized and this gives the resulting estimator $\widehat{\beta}^{(r+1)}$ and $\widehat{\Lambda}_0^{(r+1)}$. They noted that solving the full log–likelihood estimator is the same form as the Cox partial likelihood

$$L(\beta, Z_i, W_i) = \prod_{i=1}^{n} \prod_{t \in [0,1]} \left[\frac{Y_i(t) e^{\beta_1^t E(X|\mathcal{G}) + \beta_2^t Z_i}}{\sum_{j=1}^{n} Y_j(t_i) E(\beta_1^t X_j | \mathcal{G}) e^{\beta_2^t Z_j}} \right]^{dN_i(t)},$$

where $\mathcal{G} = \{T_i, \delta_i, Z_i, W_i, \beta, \Lambda_0\}$.

7.4 A Simulation Study

In this section, results from a simulation study will be presented. The main comparison is to compare a plug-in approach and the proposed sandwich covariance estimation. To understand the improvement that the RC estimator has made, naive estimates obtained from replacing X_i by $\overline{W}_{i\cdot}$ is also reported in the tables. The plug-in approach is to apply the regression calibration estimator by simply replacing X with an estimate of $E(X|\overline{W}, Z)$ obtained by using estimates for $(\mu_x, \sigma_x, \sigma_u, \mu_z, \sigma_z, \sigma_{xz})$. This is based on a usual Cox regression program and hence the covariance estimator is from the inverse of the information matrix.

In Table 7.1 we consider the case that the underlying hazards are proportional with $\lambda(t) = 1$, $\beta_1 = \ln(2)$ and $\beta_2 = \ln(2)$. Covariates X and Z were generated from a bivariate normal distribution such that both X and Z are standard normal with $\text{corr}(X, Z) = 0.33$. Common censoring times were applied and the censoring percentage was 50%. There were three repeated measurements W_{ij} for the unobserved X_i. A total of 1000 replicates were generated in each simulation configuration.

Observe that the naive estimator has a severe bias problem. The regression calibration performs quite well with limited biases. The plug-in inference may underestimate the covariance of $\widehat{\beta}$ when $\sigma_u/\sigma_x = \sqrt{2}$. The robust inference does improve the performance of coverage probabilities although most of them still did not reach the nominal level 95% due to the bias associated with the regression calibration estimator. Note that since the regression calibration is an approximation method, it will lead to larger biases on parameter estimation if in case the covariate effect is larger.

Table 7.1: Simulation with $\lambda(t|X,Z) = \exp\{\ln(2)X + \ln(2)Z\}$, and 50% censoring rate

			$\sigma_u/\sigma_x = 1$			$\sigma_u/\sigma_x = \sqrt{2}$		
			Naive	RC plug-in	RC sandwich	Naive	RC plug-in	RC sandwich
n = 100	β_1	bias	-0.207	-0.012	-0.012	-0.321	-0.005	-0.005
		s.d.	0.144	0.210	0.210	0.126	0.269	0.269
		mean($\widehat{\text{s.e.}}$)	0.141	0.200	0.202	0.123	0.235	0.265
		95% cov.	0.654	0.937	0.928	0.276	0.919	0.932
	β_2	bias	0.058	-0.006	-0.006	0.075	-0.027	-0.027
		s.d.	0.177	0.181	0.181	0.177	0.190	0.190
		mean($\widehat{\text{s.e.}}$)	0.170	0.174	0.173	0.169	0.178	0.186
		95% cov.	0.943	0.936	0.931	0.939	0.934	0.942
n = 300	β_1	bias	-0.208	-0.021	-0.021	-0.321	-0.031	-0.031
		s.d.	0.081	0.115	0.115	0.071	0.139	0.139
		mean($\widehat{\text{s.e.}}$)	0.079	0.110	0.112	0.069	0.124	0.135
		95% cov.	0.263	0.928	0.928	0.007	0.900	0.924
	β_2	bias	0.039	-0.023	-0.023	-0.056	-0.040	-0.040
		s.d.	0.094	0.099	0.096	0.094	0.101	0.101
		mean($\widehat{\text{s.e.}}$)	0.094	0.096	0.097	0.094	0.098	0.102
		95% cov.	0.946	0.944	0.945	0.919	0.925	0.934

Note: X and Z were sampled from a bivariate normal distribution such that $\text{corr}(X,Z) = 0.33$. There were 3 repeated measurements W_{ij} for unobserved covariate X_i. Results were from 1000 replications.

REFERENCES

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics* **10**, 1100-1120.

Begun, J. M., Hall, W. J., Huang, W. M., and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *The Annals of Statistics* **11**, 432-452.

Breslow, N. E. (1974). Covariance Analysis of Censored Survival Data. *Biometrics***30**, 89-99.

Breslow, N. E. and Cain, K. C. (1988). Logistic regression for two–stage case–control data. *Biometrika* **75**, 11–20.

Carroll, R. J. and Stefanski, L. A. (1990). Approximate quasilikelihood estimation in models with surrogate predictor. *Journal of the American Statistical Association* **85**, 652–663.

Carroll, R. J. and Wand, M. P. (1991). Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society*, Ser. B **53**, 573–585.

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

Carroll, R. J., Wang, S. and Wang, C. Y. (1995). Prospective analysis of stratified logistic case–control studies, *Journal of the American Statistical Association* **90**, 157-169.

Carroll, R. J., Freedman, L. and Pee, D. (1997). Design aspects of calibration studies in nutrition, with analysis of missing data in linear measurement error models. *Biometrics* **53**, 1440-1457.

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

Chen, H. Y. and Little, R. J. A. (1999). Proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, **94**, 896-908.

Chen, Y. H. and Chen H. (1999). A unified approach to regression under double sampling design. *Journal of the Royal Statistical Society*, Ser. B, revision.

Clayton, D. G. (1991). Models for the analysis of cohort and case–control studies with inaccurately measured exposure. In: *Statistical Models for Longitudinal Studies of Health.* Dwyer, J. H., Feinleib, M., Lipsert, P., et al. eds., pp. 301–333. Oford University Press, New York.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, Ser. B **34**, 187-220.

Crouch, E. A. C. and Spiegelman, D. (1990). The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(t)\exp(-t^2)dt$: Application to logistic-Normal Models. *Journal of the American Statistical Association* **85**, 464-469.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, Ser. B **39**, 1-38.

Flander, W. D. and Greenland, S. (1991). Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine* **10**, 739-747.

Foutz, R. V. (1977). On the unique consistent solution to the likelihood equations. *Journal of the American Statistical Association* 72, 147-148.

Freedman, L. S., Carroll, R. J. and Wax, Y. (1991). Estimating the relationship between dietary intake obtained from a food frequency questionnaire and true average intake. *American Journal of Epidemiology* **134**, 310–320.

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

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

Huber, P. J. (1981). *Robust Statistics.* New York: Wiley.

Hughes, M. D. (1993). Regression dilution in the proportional hazards model. *Biometrics* **49**, 1056-1066.

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

Lin, D. Y. and Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association* **84**, 1074-1978.

Lin, D. Y. and Ying, Z. (1993). Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association*, **88**, 1341-1349.

Lipsitz, S. R. and Ibrahim, J. G. (1998). Estimating equations with incomplete categorical covariates in the Cox Model. *Biometrics* **54**, 1002-1013.

Little, R. J. A. (1992). Regression with missing X's: A Review. *Journal of the American Statistical Association* **87**, 1227–1237.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, New York: John Wiley & Sons.

Liu, K, Stamler, J, Dyer, A., McKeever, J. and McKeever, P. (1978). Statistical methods to assess and minimize the role of intra-individual variability in obscuring the relationship between dietary lipids and serum cholesterol. *Journal of Chronic Disease* **31**, 399-418.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, Second Edition, London: Chapman and Hall.

Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm. *Biometrika***82**, 267-278.

Paik, M. C. and Tsai, W. Y. (1997). On using the Cox proportional hazards model with missing covariates. *Biometrika*, **84**, 579–593.

Pepe, M. S. and Fleming, T. R. (1991). A general nonparametric method for dealing with errors in missing or surrogate data. *Journal of the American Statistical Association* **86**, 108–113.

Pepe, M. S., Reilly, M. and Fleming, T. R. (1994). Auxiliary outcome data and the mean score method. *Journal of Statistical Planning and Inference* **42**, 137-160.

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. *Journal of the American Statistical Association*, **87**, 351–359.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69**, 331-342.

Prentice, R. L. (1996). Dietary fat and breast cancer: measurement error and results from analytic epidemiology. *Journal of the National Cancer Institute* **88**, 1738–1747.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case–control studies. *Biometrika* **66**, 403–411.

Prentice, R. L., Wang, C. Y. and Xie, X. (1997). Measurement error in failure time analysis. *Encyclopedia of Biostatistics*, ed: Andersen, P. K. and Keiding, N., to appear.

Reilly, M. and Pepe, M. S. (1995). A mean-score method for missing and auxiliary covariate data in regression models. *Biometrika* **82**, 299-314.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.

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. *Statistics in Medicine* **8**, 1051-1070.

Rubin, D.B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New Rork: John Wiley & Sons.

Satten, G.A. and L.L. Kupper (1993). Inferences about exposure–disease association using probability of exposure information. *Journal of the American Statistical Association* **88**, 200-208.

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

Stefanski, L. A. and Carroll, R. J. (1985). Covariate measurement error in logistic regression. *The Annals of Statistics*, **13**, 1335-1351.

Stefanski, L. A. and Carroll, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika*, **74**, 703-716.

Tsiatis, A. A., DeGruttola, V. and Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 count in patients with AIDS. *Journal of the American Statistical Association* **90**, 27-37.

Vach, W. and Schumacher, M. (1993). Logistic regression with incomplete observed categorical covariates: A comparison of three approaches. *Biometrika*, **80**, 353-362.

Wang, C. Y. (1999a). Weighted normality-based estimator in correcting correlation coefficient estimation between incomplete nutrient measurements. *Biometrics*, in press.

Wang, C. Y. (1999b). Robust sandwich covariance estimation for regression calibration estimator in Cox regression with measurement error. *Statistics and Probability Letters*, **45**, 371-378.

Wang, C. Y. (2000). Flexible regression calibration for covariate measurement error with longitudinal surrogate variables. *Statistica Sinica*, in press.

Wang, C. Y. and Carroll, R. J. (1993). Robust estimation in case–control studies, *Biometrika* **80**, 237-241.

Wang, C. Y. and Carroll, R. J. (1994). On robust estimates in case–control studies with errors in covariates, in: J. O. Berger and S. S. Gupta, eds., Statistical Decision Theory and Related Topics V, Vol. 5 (Springer–Verlag, New York) 107-120.

Wang, C. Y. and Carroll, R. J. (1995). On robust logistic case–control studies with response–dependent weights, *Journal of Statistical Planning and Inference* **43**, 331–340.

Wang, C. Y. and H. Y. Chen (1999). Sensitivity analysis on methods for Cox missing covariate regression. Preprint.

Wang, C. Y. and Huang, W. T. (1999). Conditional-cumulant-of-exposure method in logistic missing covariate regression. *Biometrics*, in press.

Wang, C. Y. and Pepe, M. S. (2000). Expected estimating equations to accommodate covariate measurement error. *Journal of the Royal Statistical Society*, Ser. B, in press.

Wang, C. Y. and Wang, S. (1997). Semiparametric methods in logistic regression with measurement error. Statistica Sinica, **7**, 1103–1120.

Wang, C. Y., Anderson, G. and Prentice, R. L. (1999). Estimation of the correlation between nutrient intake measures under restricted sampling. *Biometrics*, **55**, 711-717.

Wang, C.Y. ,Wang, S. and Carroll, R. J. (1997). Estimation in choice-based sampling with measurement error and bootstrap analysis. *Journal of Econometrics*, **77**, 65–86.

Wang, C. Y., Hsu, L., Feng, Z. D. and R. L. Prentice (1997). Regression calibration in failure time regression. *Biometrics*, **53**, 131–145.

Wang, C. Y., Wang, S., Zhao, L. P. and Ou, S. T. (1997). Weighted semiparametric estimation in regression analysis with missing covariate data. *Journal of the American Statistical Association*, **92**, 512-525.

Wang, C. Y., Chen, J., Lee, S. and Ou, S. (1999). Joint conditional likelihood estimator in logistic regression with missing covariate data. Preprint.

Wang, N., Carroll, R. J. and Liang, K. Y. (1996). Quasilikelihood estimation in measurement error models with correlated replicates. *Biometrics*, **52**, 381–400.

Wang, S. and Wang, C. Y. (1998). Asymptotic comparisons of kernel assisted estimators in missing covariate regression. Preprint.

Willett, W. C. (1990). *Nutritional Epidemiology.* New York, NY: Oxford University.

Willett, W. C., Sampson, L., Stamfer, M. L., et al. (1985). Reproducibility and validity of a semiquantitative food frequence questionnaire. *American Journal of Epidemiology* **122**, 51–65.

Wu, J. C. E. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics* **11**, 95-103.

Xie, S. X., Wang, C. Y. and Prentice, R. L. (1998). A risk regression calibration in failure time regression with covariate measurement error. Preprint.

Zhao, L. P. and Lipsitz, S. (1992). Designs and Analysis of Two-Stage Studies. *Statistics in Medicine* **11**, 769–782.

Zhong, M., Sen, P. K. and Cai, J. (1996). Cox regression model with mismeasured covariates or missing covariate data. *ASA Proceedings of the Biometrics Section*, 323-328.

Zhou, H and Pepe, M. S. (1995). Auxiliary covariate data in failure time regression analysis. *Biometrika*, **82**, 139-149.

Zhou, H and Wang, C. Y. (1999). Failure time regression analysis with measurement error in covariates. *Journal of the Royal Statistical Society*, Ser. B, revised.