MULTIPLE IMPUTATION FOR MEASUREMENT ERROR CORRECTION BASED ON A CALIBRATION SAMPLE

by

Ying Guo

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2010

Doctoral Committee:

Professor Roderick J. Little, Chair
Professor MaryFran R. Sowers
Associate Professor Bhramar Mukherjee
Research Assistant Professor Wen Ye

© Ying Guo 2010
All Rights Reserved

To my husband, my parents and my brother

ACKNOWLEDGEMENTS

I would like to first gratefully and sincerely thank my advisor, Dr. Roderick J. Little. After working with him for five years, I still find it difficult to believe that he is so knowledgeable in his field. I have been amazingly fortunate to have an advisor who gave me insightful guidance at each stage of my research, and at the same time taught me how to question thoughts and express ideas. It was truly a once-in-a-lifetime experience to receive his support and encouragement. I am so proud to be one of his students.

I wish to express my sincerest thanks to the members of my committee, MaryFran R. Sowers, Bhramar Mukherjee, and Wen Ye, for their insightful comments and suggestions on my work.

I would like to thank all people in the Department of Biostatistics at University of Michigan, especially those friends of mine for their inputs, valuable discussions and accessibility.

Finally, I am forever indebted to my husband, Jianhui Wu, for his understanding, endless patience, and encouragement when it was most required. His unwavering love helped me through each step of my doctoral study. I am also grateful to my parents, Jianwei Guo and Tingting Luo, who have always been devoted and supportive. It was under their watchful eyes that I gained so much drive and an ability to tackle challenges.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

Introduction

1.1 Background

An omnipresent issue in real-world statistical analysis problems is that obtaining accurate measurements may be difficult or impractical. Clinical and epidemiological studies are subject to measurement error from imprecise assays, inaccurate instruments, misreported questionnaires, biological variations, high cost of gold standard methods, and other sources (Racine-Poon et al., 1991; Hebert et al., 2002; Subar et al., 2003). As a specific example, to assess the relationship between diet and disease risk, it is important to quantify dietary intake of particular nutrients. The self-administered food frequency questionnaire (FFQ) is often used to assess a person's usual dietary intake of common foods. It has been recognized for decades that while dietary intake levels reported from FFQs are correlated with true values, they are usually measured with error, given that the assessment depends on the participant's ability to recall accurately usual frequency of each food consumed over a time period (Willett, 1989; Carroll et al., 1998; Spiegelman et al., 2005). As another example, in the investigation of the development of coronary heart disease, the primary predictor is systolic blood pressure. Since it is in general impractical to continuously track long-term systolic blood pressure, the blood pressure measured during a clinic

visit is often used as a surrogate. Kannel et al. (1989) point out that this simplified measurement is subject to biological variation since the long-term blood pressure and the single-visit blood pressure are likely to be different (i.e., the blood pressure has major daily, and plausibly seasonal variation).

Similar measurement error problems can be observed in a wide variety of other research fields. Sociology often needs to take into account human characteristics, such as students' motivation to study, personality and self-confidence, which cannot be directly measured but depend on investigators' personal experience. In surveys, participants' responses to interview questions may not accurately reflect the truth due to their subjective judgement (e.g., many people tend to misreport their health conditions if they have not seen a doctor for a long period of time).

The detection limit (DL) of an experimental instrument can also cause measurement error problems. The detection limit is a certain threshold below which measurements are considered unreliably quantified. Laboratory assays to measure biomarkers are commonly subject to DLs if a laboratory instrument cannot detect low levels. For example, a common design for estimating concentrations of biomarkers in biological samples is the *serial dilution assay*, in which measurements are taken at several different dilutions of a sample (Gelman et al., 2004). In serial dilution assays, the concentration of each sample is quantified by an automated optical reading of a color change and there is a certain limited range of concentrations for which the color change is informative. In particular, at low values, the color change is imperceptible and a standard computer program for analyzing such data may not give a close estimate and may simply indicate "below the detection limit". In many studies, the values below the DL are discarded with the consideration that these values are measured with significant errors. However, there are also numerous ob-

servational studies, such as human health risk and environmental impact studies carried out in environmental epidemiology, where measurements below the DL are required to evaluate potential effects of exposure on risk of disease. For example, in a study to evaluate the effect of exposure to "environmental estrogens" in pesticides and industrial chemicals on breast cancer, Cooper et al. (2002) find up to 99% of study participants have levels below the DL for some toxicants such as $1,1$-*dichloro*-$2,2$-*bis(p-chlorophenyl)ethylene*. Commonly used strategies are to substitute values below the DL with a simple constant value, e.g., $DL/2$, or use a random "fill-in" value selected from a predefined distribution. It is clear that, unless the proportion of observations below the DL is small, these substitution methods are prone to provide biased estimates of parameters in a regression analysis (Newman et al., 2007; Lubin et al., 2004).

Since statistical analysis that takes into account measurement error is often much more complicated than ordinary regression analysis, and major statistical analysis packages do not provide standard programs for correcting the effect of measurement error, a widely used approach is to ignore any observational error, implicitly assuming that the values are measured precisely. Statistical analysis ignoring such inherent error is commonly referred to "naïve analysis" in research literature.

The covariate measurement error problem has generated major research interest since the late 1970*s* (Prentice, 1982; Carroll and Spiegelman, 1986; Fuller, 1987; Willett, 1989; Dafni and Tsiatis, 1998; Carroll et al., 1999, 2006). It is well known that measurement error in covariates can cause bias in parameter estimation for regression models. Particularly, it attenuates the regression coefficient towards the null in comparison with the result computed from a regression on the same variable measured without error. Also, the estimates of regression coefficients of other covariates are

biased, even if these covariates are measured without error (Richardson and Gilks, 1993). Therefore, it is important to evaluate the magnitude of measurement error and consider appropriate procedures to correct for the effect of measurement error.

1.2 Measurement Error Model

The fundamental prerequisite to adjust for measurement error is to clearly define the structure of the error. We start by introducing a basic type of measurement error, *classical measurement error*, where the true value is measured with additive error. A typical example of the classic measurement error can be found in nutritional studies where investigators use a protein biomarker, namely urinary nitrogen, as a surrogate measurement of unobserved dietary protein intake.

In detail, let X denote the true variable that are not observed, and let W denote the observed measurement of X (which is sometimes also named the surrogate variable). The classical measurement error model states that

$$W = X + \xi$$

where the measurement error ξ is an independent variable with mean zero and usually constant variance.

Sometimes it may even be necessary to allow for more complex association between W and X. For example, if W is not an unbiased measure of X, the classical measurement error model clearly does not hold, and a more general model, such as,

$$W = \beta_0 + \beta_1 X + \xi$$

is needed. The above model is one variant of the classical measurement error model. If $\beta_0 = 0$ and $\beta_1 = 1$, then it is exactly the same to the classical model.

1.3 Calibration and Main Study Data

Considerable efforts have been devoted to developing methods to adjust for the adverse effect of the error on the statistical analysis. Many proposed methods to correct for measurement error require supplemental information available on the relationship between the error-free variable X and mismeasured variable W, which can then be used to correct the response-covariate association. Such extra information is often called *calibration data*, which can be replicate measurements on a subset of the study subjects (*replication data*), or a collection of error-free data in a validation sample (*validation data*). Examples of calibration data include 24-hour recalls to examine food frequency questionnaires as a measure of usual food intake, medical record review to validate self-report health conditions, home measurements of fine particles to verify cigarette use as a measure of air pollution, and concentrations of standard hormone samples to calibrate instrument readings.

It should be clarified that a distinction is often drawn between calibration data and main study data in measurement error studies. The main study data is used to investigate the association between the explanatory variables (covariates) and the response variable (disease) through a suitable regression model. Such main/calibration study designs have given rise to a large literature on methods to handle measurement error. Fuller (1987) and Carroll et al. (2006) summary methodologies for dealing with measurement error in linear regression and nonlinear regression, respectively.

The main/calibration design is also referred to *two-stage* or *two-phase* study design by some researchers, if the calibration data is a subset of the main study data (Breslow and Cain, 1988; Zhao and Lipsitz, 1992; Dahm et al., 1995). In their studies, the response, and some surrogate for primary covariate of interest are collected for

all subjects (*stage*-1), and then information about covariate of interest is collected from a subset of subjects (*stage*-2). Thereafter, the information gathered in both stages of measurements is combined for the entire analysis.

1.4 Measurement Error Correction

In general, existing methods for dealing with measurement error can be categorized as functional approach and structural approach, depending on whether or not a parametric distribution is assumed for the unobserved covariate.

1.4.1 Functional Approach

Regression calibration (RC) and simulation extrapolation (SIMEX) are common functional approaches (Carroll and Stefanski, 1990; Cook and Stefanski, 1994; Carroll et al., 1996; Spiegelman et al., 1997; Lin and Carroll, 1999). Functional approaches are often favored because no assumptions need to be made regarding the distribution for the unobserved covariate (and so it consequently avoids the risk of making incorrect assumptions). On the other hand, functional approaches may be limited in complicated problems with longitudinal studies and nonlinear models.

The RC method is implemented by substituting the unobserved X with its expectation given the surrogate W and then performing standard analysis. This approach is simple, but it has some inherent limitations. Carroll and Stefanski (1990) show that regression calibration produces unbiased estimates in the case of a simple linear regression, but its estimate is only approximately unbiased for a nonlinear regression. Ko and Davidian (2000) argue that regression calibration may eliminate bias in fixed effects but not in covariance parameters for nonlinear mixed effect models.

SIMEX is a simulation-based method aimed at estimating and reducing bias caused by measurement error. This method is originally presented by Cook and

Stefanski (1994) and further improved by Carroll et al. (1996). The key procedure of the SIMEX method is to add artificial measurement error to the original data, create a simple bivariate plot (error induced-bias versus the variance of artificially generated errors) via a simulation study, and extrapolate the plot to the error-free case. Since SIMEX is a self-contained simulation method, its computational cost is high. In addition, the error variance has to be known or estimated from replication or validation data a *priori*, and the choice of extrapolation methods is often not obvious.

1.4.2 Structural Approach

The structural approaches excel in their extensive applicability in more complex models and may improve the precision of parameter estimates (at the cost of computational complexity) (Schafer and Purdy, 1996). Two of the most common structural approaches are the maximum likelihood (ML) approach and the Bayesian method (Richardson and Gilks, 1993; Lyles et al., 1999; Spiegelman et al., 2000; Kulathinal et al., 2002; Ferrari et al., 2008; Hossain and Gustafson, 2009). A general concern about those structural approaches is the potential misspecification of the distribution of covariates. Since the true covariate is unknown due to measurement error, it is generally difficult to check whether the assumed model is a reasonable match for the unobserved values.

Imputation (IM) and multiple imputation (MI) methods have been well established for dealing with missing data, and have been gaining increasing attention for handling measurement error. Some of MI methods originate from the Bayesian approach, with missing values sampled from their posterior predictive distributions given the observed data. There are many ways to draw from the posterior distribution. In some situations, it can be relatively simple (e.g., continuous variables under

the assumption of multivariate normality). In more complex situations, iterative algorithms such as the Markov Chain Monte Carlo (MCMC) method or the Gibbs' sampler may be needed. More detailed information can be found in Schafer (1997).

In MI, multiple completed data sets are created by filling in values for the unobserved data. Each imputed data set is then analyzed as if it were a complete data set, and the inferences are drawn by combining the results from imputed data sets using combining rules suggested by Rubin (1987), Meng and Rubin (1992), and Little and Rubin (2002).

Thoresen and Laake (2000) compare the behavior of RC and ML in a simple, additive measurement error model, and indicate that the RC method has very good results regarding bias. Messer and Natarajan (2008) present a study comparing the ML, RC to MI methods via simulations in both realistic and extreme measurement error settings. Their results show that the behavior of each method depends on numerical approximation to the likelihood. Specifically, in the simulation scenarios where main study and validation sample sizes are large enough, ML works better than its competitor methods. Guolo and Brazzale (2008) evaluate the performance of the likelihood method in comparison with the RC and SIMEX methods under different structures of measurement error. They suggest that the choice of correction technique should be based on the measurement error structure. Cole et al. (2006) examine the MI and RC methods, and find that the MI correction works better in their simulation settings.

Most of the measurement error correction methods mentioned above are designed for situations where the variance is constant. However, in many situations, this assumption may not hold. In practice, heteroscedasticity can arise when the observations are obtained in nonhomogeneous conditions. That is, the data might

have been obtained from several studies where the populations are heterogeneous or from study designs where different measurement technologies are used (Walter, 1998; Strauss et al., 2001). Fuller (1987) points out that the measurement process might be subjective and different among all individuals, e.g., smoker/nonsmoker might be subject to a different contamination process. Another source of heteroscedastic data is biochemical assays in the clinical and biological sciences. Sadler and Smith (1985) and Gelman et al. (2004) show that assay data are commonly prone to heteroscedastic measurement error. Staudenmayer et al. (2008) find that measurement error may be heteroscedastic across individuals due to unequal sampling effort in a nutritional epidemiological study.

1.5 Overview

The primary objective of this dissertation is the development of statistical techniques to improve statistical inference in regression analysis by correcting for measurement error. Toward this end, several error-correction methods are designed, analyzed and empirically evaluated in this work.

Chapter 2 considers methods to deal with values below the detection limit. In many real world applications, raw data on the relationship between known and measured values of an analyte is collected and analyzed to determine the limit of quantification (LOQ) of an assay. Usually, researchers are given an observed value for the interested marker if this value is greater than the LOQ, and a missing value otherwise. From a statistical perspective, the implicit assumption is that there is no measurement error for values above the LOQ, and unacceptable measurement error for values under the LOQ. A more plausible assumption is that there is measurement error throughout the measure's support. To better represent characteristics

of measurement error, we describe a Bayesian measurement error model that yields prediction intervals for the true assay value throughout the range of analyte values, and allows for heteroscedasticity of the measurement errors. We illustrate our model on calibration data for fat-soluble vitamins, focusing particularly on Beta Cryptoxanthin. Our results show that prediction intervals for values above the LOQ are wide, and the width increases with the measured value. Prediction intervals below the LOQ provide more information than the statement that the value is less than the LOQ. Our findings solidify our argument that the current existing approach to transmit data from calibration assays is flawed, since it provides a distorted picture of the actual measurement error. Implications for subsequent analysis of assay measurements are discussed.

Chapter 3 considers the problem in regression analysis where covariate has heteroscedastic measurement error. A calibration sample that measures pairs of values of X and W is available; we consider calibration samples where Y is measured (internal calibration) and not measured (external calibration). One common approach for measurement error correction is Regression Calibration (RC), which substitutes the unknown values of X by predictions from the calibration curve of X on W. An alternative approach is to multiply impute the missing values of X given Y and W based on an imputation model, and then use multiple imputation (MI) combining rules for inferences. Recent work by Freedman et al. (2008) compares these two approaches, suggesting that RC is more efficient under plausible assumptions. However, their work assumes the measurement error of W has a constant variance, whereas in many situations, the variance varies as a function of X. We consider modifications of the RC method and the MI method that allow for heteroscedastic measurement error, and compare them by simulation. As will be shown in Section 3.5, the MI

method provides better inferences in this setting, in terms of small empirical bias, good precision, and confidence coverage.

Chapter 4 considers another aspect of measurement error correction for the situation that current information about the error gathered from calibration samples is insufficient to provide valid adjusted inferences. We consider the problem of estimating the regression of an outcome Y on covariates X and Z, where Y and Z are observed, X is unobserved, but a proxy variable W that measures X with error is observed. Data on the joint distribution of X and W (but not Y and Z) are recorded in a calibration experiment. In many situations, the data from the calibration experiment may not be directly available to public users, but summary statistics for the joint distribution of X and W can be provided. To exploit such properties, we describe a method that implement multiple imputations of the missing values of X in the regression sample, so that the regression coefficients of the regression Y on X and Z as well as associated standard errors can be estimated correctly using multiple imputation (MI) combining rules, under normal assumptions. The proposed method is demonstrated via a simulation study to have superior performance to existing methods, namely the naïve regression, classical calibration, and regression prediction.

Chapter 5 concludes this dissertation with a summary of contributions of this work, and a discussion of questions that remain open together with possible future research directions.

CHAPTER II

How Well Quantified is the Limit of Quantification?

2.1 Introduction

In epidemiological or other empirical studies that involve measurements of an analyte, the values supplied by a measuring instrument or assay are typically estimates and hence subject to measurement error. When the measurement error is a substantial fraction of the determined value, the measurement is considered to be too unreliable to report, and the information provided to the user is that the value is less than the detection limit. The limit of quantification (LOQ) and limit of detection (LOD) are two common types of the detection limit. The LOQ is the lowest concentration of the analyte in a sample that can be quantitatively determined with suitable precision and accuracy (Shah et al., 1992). A typical rule for determining the LOQ is to use the value at which the coefficient of variation (standard deviation/mean) of the measurement is greater than some threshold, such as 20%. The LOD is defined as the lowest concentration of the analyte that can be distinguished with reasonable confidence from background noise (Currie, 1968). The LOQ is different to (and usually higher than) the LOD, and measurements above the LOQ can be reported with a high degree of confidence (Armbruster et al., 1994).

Interest in this special issue concerns how to analyze data when some values are

below the LOQ. Strategies commonly applied to handle this type of data include deletion or replacement of values below the LOQ with the value of LOQ or some other imputed values (Hopke et al., 2001; Richardson and Ciampi, 2003; Schisterman et al., 2006). Our interest is the more basic question of the statistical properties of the LOQ method for communicating analyte values to users. The measured values are based on data that involves measurements for a calibration sample, where measured values and the underlying true values of the analyte are both recorded. We view the supplied values as summaries of the results of these calibration experiments. Intuitively, it seems that these summaries are to some degree distorted, in that the values above the LOQ are subject to measurement error that is effectively being ignored.

In this chapter, we describe a Bayesian model for the calibration data set, which provides a posterior distribution for the true values of future recordings from the measuring instrument. We show that this measurement error is substantial; indeed the prediction intervals for the values above the LOQ in our application are wider than the prediction intervals of values below the LOQ, arguably casting doubt on the practice of treating the error for cases above the LOQ as absent. On the other hand, the prediction intervals for values below the LOQ are considerably more informative about the true value than that provided by treating them simply as below the LOQ. Some implications of these distortions for subsequent analyses of the data are discussed.

Our approach draws on a considerable statistical literature devoted to calibration methods (Schwenke and Milliken, 1991; Giltinan and Davidian, 1994; Belanger et al., 1996; Higgins et al., 1998). In much of this work, the concentration estimates for the unknown samples are obtained by inverting a parametric regression of the fitted

concentration-response; and nonconstant variability is usually modeled as a function of the mean response. Beyond point estimation of unknown concentrations, the precision of calibration is also investigated by some authors. Giltinan and Davidian (1994) present a delta-method approximate variance to construct a confidence interval for the unknown concentration. Belanger et al. (1996) discuss large-sample confidence intervals, based on a model that allows for nonconstant variance. Dunsmore (1968) describes Bayesian approaches to the linear calibration problem. Racine-Poon (1988) implements a Bayesian approach to obtain the posterior distribution for unknown concentrations of interest in nonlinear calibration inference, but does not address the issue of nonconstant variance.

To our knowledge, little work has been done relating this prior work to the LOQ issue. Gelman et al. (2004) jointly estimate the calibration curve and the unknown concentrations using a Bayesian method and provide estimates for measurements below the detection limit, but they do not explicitly discuss the precision of the estimates.

The chapter is organized as follows. Section 2.2 describes the data used in this study, and provides definitions of the LOD and the LOQ. In Section 2.3, we describe a Bayesian measurement error model. Measurement error is constructed to be heteroscedastic. We consider both parametric (e.g., linear) regression and nonparametric regression (e.g., spline) for modeling the relationship between X and W. Section 2.4 contains results for estimates of the model parameters and predictive values for measurements below the LOQ. In Section 2.5, some implication is illustrated. In Section 2.6, conclusions are made and a discussion of our findings are presented. Some technical details of our methods are provided in Appendix 2.7.

2.2 Data Description

The data for this study is from calibration experiments for human serum fat soluble vitamins, measured by high performance liquid chromatography (HPLC). Three replicate calibration experiments were performed between May and November of 2005. Each experiment consists of eight samples with known concentrations of the analyte, from standard reference materials (SRMs) obtained from the National Institute of Standards and Technology (NIST). The eight samples have decreasing concentrations generated by performing a two-fold serial dilution. For each of the three replicate experiments, each of eight samples is analyzed 10 times by HPLC, yielding a total of 30 replicate measures for each dilution value. Let x_i denote the true concentration of the ith sample, and let w_{ij} be the HPLC measured value for replicate j of dilution i where $i = 1, ..., 8$ and $j = 1, ..., 10$. For simplicity we pool the replicates for each of the three experiments, which is justified here since the between-experiment error is comparable to the within-experiment variation of the replicates. This approach is supported by Analysis of Variance tests of the null hypothesis that the experiment means are equal – the P values are all greater than 0.8 for the eight vitamins that we investigate in this study. We mainly present data on one of the fat-soluble vitamins, Beta Cryptoxanthin, although we obtain similar results when we apply our methods to the other vitamins, namely Lutein, Retinol, Carotene, Lycopene, Gamma tocopherol, Delta tocopherol and Alpha tocopherol.

The LOD is here defined as $3 \times SD$, where SD is the standard deviation of the 30 replicate measurements at the lowest level of concentration of the analyte. The LOQ is defined by estimating the coefficient of variation (CV) of the replicates for each concentration, plotting these coefficients of variation against the concentration

values, and then fitting a nonlinear regression line to the data. The LOQ is the concentration corresponding to a predicted mean CV of 20%.

2.3 Models

We now consider a measurement w^* of the analyte from a subsequent data collection. Assuming a linear relationship between the true and measured values, the estimate of the true value of the analyte under the usual procedure is reported as:

$$\hat{x}(w^*) = \begin{cases} (w^* - b_0)/b_1 & if \ \ \hat{x}(z) \geq LOQ \\ < LOQ & if \ \ \hat{x}(z) < LOQ \end{cases} \tag{2.1}$$

where b_0 and b_1 are least squares estimates of the intercept and slope from the regression W on X fitted to the calibration data. In Eq.2.1, the estimated value is obtained from the calibration curve, providing that it is above the LOQ. The calibration curve assumes a linear relationship, which is empirically justified in this case, although nonlinear relationships can be fitted if necessary.

We now develop prediction intervals for the true value of x corresponding to w based on a Bayesian model fitted to the calibration data.

We assume

$$(w_{ij}|x_i, \theta) \sim_{ind} N(\mu(x_i; \beta), \sigma^2 x_i^{\alpha}) \tag{2.2}$$

where $\theta = (\beta, \log \sigma^2, \alpha)$, $N(a, b)$ denotes the normal distribution with mean a, variance b, and $\mu()$ denotes the mean function. We show results for a linear mean function

$$\beta = (\beta_0, \beta_1), \ \ \mu(x) = \beta_0 + \beta_1 x \tag{2.3}$$

which corresponds to the assumption applied in the usual method of Eq.2.1. To assess the linearity, we also fit the model (2.2), assuming a quadratic relationship between the measured and true values:

$$\beta = (\beta_0, \beta_1, \beta_2), \;\; \mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2 \tag{2.4}$$

as well as a more flexible curve using penalized splines:

$$\mu(x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{k=1}^{K} b_k (x - K_k)_+^p \tag{2.5}$$

where p is the degree of the spline, $\beta = (\beta_0, \beta_1, ..., \beta_p)$ are regression coefficients, $K_1 < K_2 < ...K_p$ are K fixed knots, and $\{b_1, ..., b_K\}$ are random effects, assumed normal with mean zero and variance τ^2. Here x_+ is equal to x if x is positive and zero otherwise.

The determination of a Bayesian predictive distribution for future values of X requires a prior distribution for these values. To study sensitivity of answers to the choice of this prior distribution, we present results for a dispersed lognormal prior distribution and a uniform prior distribution, as discussed below. For known θ, the posterior predictive distribution of x^* given w^* is given by Bayes' Theorem:

$$p(x^*|w^*, \theta) \propto p(x^*) p(w^*|x^*, \theta) \tag{2.6}$$

where $p(w^*|x^*, \theta)$ is given by Eq.2.2. We consider two approaches to deal with the fact that in practice θ is not known. One approach is to assume a prior distribution for θ and replace Eq.2.6 by

$$p(x^*|w^*, C) \propto \int p(w^*|x^*, \theta)p(\theta|C)d\theta \tag{2.7}$$

where $p(\theta|C)$ is the posterior distribution of θ given the calibration sample data C. We apply this approach with the non-informative prior distribution

$$p(\beta, \log\sigma, \alpha) = const., \quad -2 < \alpha < 2$$

for the parameters. The prior of α is assumed uniform in a finite range $(-2, 2)$ to assure a proper posterior distribution – in fact, preliminary analysis shows the value of α is clearly in the narrower range $(0, 1)$ for all eight vitamins analyzed. Draws from the posterior predictive distribution can be obtained by the Metropolis within Gibbs sampler, as discussed in Appendix 2.7.

We also provide prediction intervals where estimates $\hat{\theta}$ of the parameters θ are substituted in Eq.2.6, rather than integrating over their posterior distribution as in Eq.2.7. We call this approach empirical Bayes; it is inferior to full Bayes since it does not propagate uncertainty in θ, but it may be preferred by those seeking a more frequentist formulation of the problem. For known α, maximum likelihood (ML) estimates of the other parameters are easily computed by weighted least squares. The ML estimate of α can be computed using an iterated conditional modes algorithm (Besag, 1986). We present results for a simpler regression approach where α is estimated as the slope of a regression on the logarithm of X of the logarithm of the squared residuals of the regression of W on X. This approach is straightforward, and yields estimates of α close to the ML estimates when applied to our data.

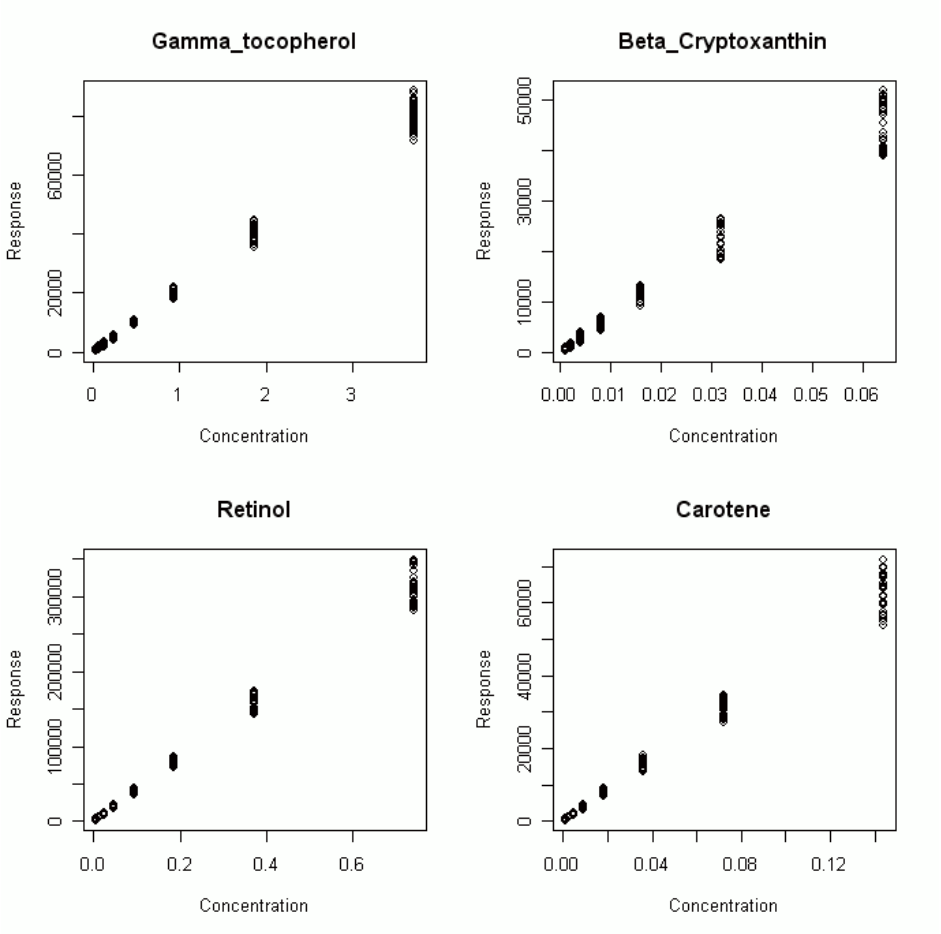

Figure 2.1: Calibration data for eight analytes

Table 2.1: Estimates of α for eight analytes by various methods

Analytes	Regression	ML	Bayes	
			Post. Mean	95%HPD.
Beta Cryptoxanthin	0.65	0.64	0.64	(0.58, 0.71)
Carotene	0.77	0.72	0.72	(0.67, 0.77)
Lutein	0.61	0.63	0.63	(0.58, 0.68)
Lycopene	0.70	0.72	0.71	(0.67, 0.76)
Retinol	0.71	0.68	0.67	(0.63, 0.72)
Alpha tocopherol	0.62	0.62	0.62	(0.57, 0.67)
Delta tocopherol	0.63	0.65	0.65	(0.53, 0.77)
Gamma tocopherol	0.56	0.56	0.56	(0.51, 0.61)

2.4 Results

An important feature of the model in Eq.2.2 is that the variance of w_{ij} is not assumed constant, but allowed to vary as a function of x_i, with the exponent α being treated as a parameter to be estimated. Figure 2.1 shows the concentration-response data for calibration samples of four analytes. These data exhibit heterogeneity of variance: variation in the response systematically increases with concentration level. It is thus important for the model to account for nonconstant variance, when fitting a concentration-response model to calibration sample (Belanger et al., 1996; Carroll and Ruppert, 1988). The assumption of constant variance ($\alpha = 0$) is not supported for the Beta Cryptoxanthin data and other analytes in our application, and we suspect the same holds in many other settings. The value of α has considerable influence on the width of the prediction interval for different measured values of W.

Table 2.1 presents estimates of α from ML, from our simple regression method, and the posterior mean and standard deviation of α from the Bayesian analysis, for the eight analytes for which we have calibration data. The ML and regression estimates are similar, and range from about 0.56 to 0.72. Uncertainty about α, reflected in the posterior variance from the Bayesian analysis, is surprisingly small.

Table 2.2: ML and Bayes parameter estimates, linear model, using Beta Cryptoxanthin 986 $C1$ calibration data

Parameters	ML	Bayes	
		Mean	95%HPD.
β_0	-0.003	-0.003	(-0.012, 0.006)
β_1	0.706	0.705	(0.689, 0.722)
σ^2	0.014	0.015	(0.012, 0.018)
α	0.642	0.643	(0.578, 0.707)

These results suggest that estimating α is clearly superior to an analysis based on the constant variance assumption $\alpha = 0$.

Parameter estimates for the linear model are shown in Table 2.2. The linear curve fits these data well, and accordingly results for the quadratic model and the spline model are similar and not presented. Summaries of the predicted distribution of the true values for the Beta Cryptoxanthin data are shown in Tables 2.3 and 2.4, for both the empirical Bayes and full Bayes methods. These prediction intervals indicate a range within which the true values are likely to occur, with 95% probability. To assess sensitivity to the choice of prior distribution of X^*, Table 2.3 presents results for the uniform prior distribution $p(x^*) = const.$, and Table 2.4 presents results for the dispersed lognormal prior distribution $\log x^* \sim N(0, 1000)$. In Tables 2.3 and 2.4, $X^* = \hat{x}(w^*)$ is the prediction for the standard calibration procedure, with predictions on the curve presented even if they are below the LOQ. Some conclusions from these results follow:

1. Results in Tables 2.3 and 2.4 are similar, indicating lack of sensitivity to the two choices of prior distribution of X^*.

2. The empirical Bayes and full Bayes procedures yield generally similar results, with the Bayesian prediction intervals being slightly wider since they propagate

error in the parameter estimates.

3. Prediction intervals are similar for the linear and quadratic models, suggesting that the linear model is reasonable. The predictive means of predictive distribution are generally quite close to the predictions using the classical procedure, although those predictive means are slightly higher.

4. The width of the prediction intervals increases with the value of X^*, reflecting the nonconstant variance – the estimate of the exponent of the variance is 0.64 with a standard error of 0.03.

5. There is substantial uncertainty in the predictions above the LOQ, which is 0.0062. For example, when $X^* = 0.0142$, a value more than the twice of the LOQ, the HPD prediction interval for the linear model with uniform prior is $(0.0101, 0.0203)$.

6. There is considerable information concerning the values of X below the LOQ. For example, when $X^* = 0.00286$, a value less than half the LOQ, the HPD prediction interval for the linear model with uniform prior is $(0.00168, 0.00523)$, which conveys considerably more information than the statement that the value is less than 0.0062 .

2.5 Analysis Implication

The Bayesian model presented before provides prediction intervals for the true values of the analyte X, which allow for the analysis of the data that can properly account for measurement error. Clearly, the marginal distribution of the estimated analyte values W is over-dispersed as an estimate of the distribution of X, and distorted by the heteroscedasticity of the prediction errors as a function of X.

Table 2.3: Prediction of the true values of X, from Bayes and empirical Bayes versions of the prediction model, with the assumption of linear relationship between W and X, using Beta Cryptoxanthin 986 $C1$ calibration data (LOQ = 0.00619, LOD = 0.00131). The prior distribution of X is uniform on the raw scale. CA: conventional approach, predicting the true values of X by inverting the fitted concentration-response calibration curve. HPD interval: the highest probability density interval.

CA	Empirical Bayes				Bayes			
X^*	Mean	SD	CV(%)	95%HPD	Mean	SD	CV(%)	95%HPD
283.80	288.76	33.45	11.58	(221.00, 357.00)	288.07	35.78	12.42	(218.00, 367.00)
142.03	146.49	22.85	15.60	(103.00, 201.00)	146.23	22.63	15.48	(101.00, 203.00)
56.96	60.17	12.61	20.96	(37.50, 86.50)	60.29	12.85	21.31	(36.55, 88.75)
42.79	45.65	10.62	23.26	(26.25, 68.75)	46.62	10.92	23.42	(27.20, 69.50)
28.62	31.55	8.67	27.48	(15.75, 49.75)	31.58	8.84	27.99	(16.75, 52.25)
10.19	12.79	5.09	39.81	(4.75, 25.75)	12.47	5.10	40.86	(4.75, 25.25)

Table 2.4: Prediction of the true values of X, from Bayes and empirical Bayes versions of the prediction model, with the assumption of linear relationship between W and X, using Beta Cryptoxanthin 986 $C1$ calibration data (LOQ = 0.00619, LOD = 0.00131). The prior distribution of X is lognormal with mean 0 and variance 1000. CA: conventional approach, predicting the true values of X by inverting the fitted concentration-response calibration curve. HPD interval: the highest probability density interval.

CA	Empirical Bayes				Bayes			
X^*	Mean	SD	CV(%)	95%HPD	Mean	SD	CV(%)	95%HPD
283.80	285.94	33.23	11.62	(221.00, 359.00)	286.05	33.72	11.79	(219.00, 361.00)
142.03	144.26	22.07	15.29	(105.00, 199.00)	144.98	22.13	15.26	(102.50, 197.50)
56.96	58.66	12.70	21.65	(36.50, 85.50)	59.20	12.85	21.71	(36.50, 86.55)
42.79	44.21	10.25	23.18	(26.50, 68.25)	44.85	10.29	22.94	(26.75, 67.25)
28.62	30.24	8.49	28.07	(15.25, 47.50)	30.69	8.65	28.19	(15.75, 49.25)
10.19	11.27	4.55	40.37	(4.25, 22.75)	11.68	4.69	40.15	(4.80, 23.62)

Consider inference for the linear regression of an outcome Y, with the analyte X treated as a covariate. In this case, the impact of measurement error with constant variance is well known (Fuller, 1987; Fox, 1997; Carroll et al., 2006), but the impact when the measurement error depends on X, as is the case here, has received little attention. For simplicity, we consider the simple regression of a dependent variable Y on the true analyte values, of the form

$$(Y_i|X_i) \sim_{ind} N(\gamma_0 + \gamma_1 X_i, \tau^2), \;\; i = 1, ..., n \tag{2.8}$$

where (γ_0, γ_1) are the parameters of interest, and n represents the number of subjects in the sample. We do not observe the actual values of X, but instead observe

$$(W_i|X_i) \sim_{ind} N(\beta_0 + \beta_1 X_i, \sigma^2 X_i^{\alpha}), \;\; i = 1, ..., n \tag{2.9}$$

as in Eq.2.2. Here, W_i and Y_i are conditionally independent given X_i, since W_i deviates from X_i by random measurement error. Finally, suppose the distribution of X is

$$X_i \sim_{ind} N(\mu_x, \sigma_x^2), \;\; i = 1, ..., n \tag{2.10}$$

Consistent estimation of γ_0 and γ_1 requires fitting the model defined by Eq.2.8–2.10 to the data. The standard approach is to estimate the values of X by $\hat{X}_i = (W_i - b_0)/b_1$, on the calibration curve Eq.2.1, and then regress Y on $\hat{X}$. When $\alpha = 0$, Eq.2.8–2.10 imply that the regression of Y on $\hat{X}$ is linear with mean

$$E(Y_i|X_i) = \gamma_0^* + \gamma_1^* X_i \tag{2.11}$$

where $\gamma_0^* = \gamma_0 + (1-\delta)\gamma_1\mu_x + \delta\gamma_1(b_0 - \beta_0)/\beta_1$, $\gamma_1^* = \delta b_1\gamma_1/\beta_1$ and $\delta = \beta_1^2\sigma_x^2(\beta_1^2\sigma_x^2 + \sigma^2)^{-1}$. Since (b_0, b_1) are unbiased for (β_0, β_1),

$$
\begin{aligned}
E(\gamma_0^*) &= \gamma_0 + (1-\delta)\gamma_1\mu_x \\
E(\gamma_1^*) &= \delta\gamma_1
\end{aligned}
\tag{2.12}
$$

Thus, the standard approach of regressing Y on $\hat{X}$ leads to an attenuation of the regression coefficient γ_1 by the factor δ, and a bias in the intercept of $(1-\delta)\gamma_1\mu_x$. When $\alpha \neq 0$ (the more realistic case for our data), the regression of Y on $\hat{X}$ is no longer even linear under Eq.2.8–2.10, because of the nonconstant variance in Eq.2.9. Approximating the intercept and slope from the covariance matrix of Y and D leads to Eq.2.12 with $\delta = \beta_1^2\sigma_x^2(\beta_1^2\sigma_x^2 + \sigma^2\kappa(\alpha))^{-1}$, where $\kappa(\alpha) = E(X^\alpha)$. Thus expressions for the bias of the intercept and the slope are as before, with this modified form for δ.

2.6 Conclusion and Discussion

The logic of the current approach to report analyte values is based on the idea that the coefficient of variation of the measurement error of the regression of W on X is the deciding factor – when the coefficient of variation is above the cutoff, uncertainty is ignored, and when the coefficient of variation is below the cutoff, uncertainty results in a value not being reported. There are two problems with this, from a statistical perspective. Firstly, the standard approach bases the prediction of X on the regression curve for the conditional distribution of W given X, but the correct predictive distribution for determining uncertainty about X is the distribution of X given W, not the distribution of W given X. The Bayesian paradigm followed here allows uncertainty to be focused on the right conditional distribution. This approach

has the difficulty of requiring a specification of the prior distribution of X. However, for the data set discussed here results are not very sensitive to two choices of this prior distribution. We note that the posterior means of X from our model in Table 2.3 are consistently lower than the standard predictions on the calibration curve, reflecting a bias in the standard method arising from the wrong choice of conditional distribution.

The second problem is that the standard deviation of the predictive distribution of X is more directly pertinent to distortion in statistical analysis than the coefficient of variation. In particular the former quantity directly determines the distortion of the marginal distribution of X when predictions X^* are used as proxies for the true values. Also measurement error models in statistics involve the standard deviation of the measurement error, not the coefficient of variation. From this perspective, the logic underlying the LOQ seems flawed, since the standard deviation of the predictive distribution increases with the value of X. Uncertainty is the rationale for not reporting values below the LOQ, but in absolute terms, the uncertainty in the reported values above the LOQ is actually greater than the uncertainty in the values below the LOQ.

Some implications of these findings for analysis are discussed in Section 2.5. In future work we plan to develop these ideas more fully, by providing techniques for correcting for measurement error as reflected in the posterior distribution of the true analyte values.

2.7 Appendix

Gibbs sampler is used to take draws of the parameter $\theta = (\beta, \sigma^2, \alpha)$ from its posterior predictive distribution. The loglikelihood function of θ is given observed

data (x_i, w_i) is given by

$$\begin{aligned} l &= P(\beta, \sigma^2, \alpha | X, W) \\ &= \frac{(\prod_{i=1}^n \phi_i)^{1/2}}{(2\pi\sigma^2)^{n/2}} exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n \phi_i (w_i - \beta_0 - \beta_1 x_i)^2\} \end{aligned}$$

where $\phi = 1/x_i^{\alpha}$. Assuming a prior distribution for the parameters is $p(\beta, \sigma^2, \alpha) \propto \sigma^{-1}$, by Baye's theorem, the joint posterior density is written as

$$\begin{aligned} &P(\beta, \sigma^2, \alpha | X, W) \\ &\qquad = \frac{p(\beta, \sigma^2, \alpha) \times l}{\int\int\int_\theta p(\beta, \sigma^2, \alpha) \times l d\alpha d\beta d\sigma} \\ &\qquad \propto \frac{(\prod_{i=1}^n \phi_i)^{1/2}}{(2\pi)^{n/2}\sigma^{n+1}} exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n \phi_i (w_i - \beta_0 - \beta_1 x_i)^2\} \end{aligned}$$

Then the posterior distribution of α given all the other parameters and the data can be

$$P(\alpha | \beta, \sigma^2, X, W) \propto (\prod_{i=1}^n \phi_i)^{1/2} exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n \phi_i (w_i - \beta_0 - \beta_1 x_i)^2\}$$

Since this distribution does not appear to have a closed-form solution, we generate α from it using Metropolis algorithm.

Given α, the model is reduced to weighted linear regression. Applying Bayesian theory, draws $\beta^{(d)}$ and $\sigma^{2(d)}$, can readily obtained from their posterior density as follows:

1. Draw a chi-square random variable, p , with $(n - p)$ degree of freedom, and define

$$\sigma^{2(d)} = \prod_{i=1}^n \phi_i (w_i - \beta_0 - \beta_1 x_i)^2 / q$$

2. Draw p standard normal deviates, $z = (z_1, ..., z_p)$, z_i $N(0, 1)$, $i = 1, ..., p$ and define

$$\beta^{(d)} = \hat{\beta} + A' z \sigma d$$

where A is the Cholesky factor of $(X\phi X)$, X is an $n \times p$ matrix with elements $x_i j$, $i = 1, ..., n$ and $j = 1, ..., p$, and W is an $n \times n$ matrix with diagonal elements ϕ_i and zero off-diagonal elements. $\hat{\beta}$ is weighted least square estimators:

$$\hat{\beta} = (X\phi X)^{-1} X\phi W$$

Thus, we can use Metropolis within Gibbs sampler to generate values form the joint posterior distribution in the order of α , σ^2 and β.

CHAPTER III

Regression Analysis on the Covariate with Heteroscedastic Measurement Error

3.1 Introduction

Measurement error is common in many empirical studies, arising from assay or instrumental error, biological variation, or errors in questionnaire-based self-report data. It is well known that in a regression analysis the estimated effect of a predictor variable may be attenuated when it is measured with error. In particular, Kipnis et al. (2003) find that measurement error in dietary intake assessment by using the Food Frequency Questionnaires leads to severe attenuation in the estimate of disease relative risk in a biomarker study. Cotton et al. (2005) show that measurement error can reduce the chance of accurate diagnosis of appropriate educational placement for children with reading difficulties. Wannemuehler et al. (2009) indicate that the impact of measurement error can be substantial in the assessment of the association between pollutant exposure and a health outcome, using surrogates for unobserved measurements of ambient concentrations.

Most of the research on measurement error in covariates assumes the variance of measurement error is constant. However, the variance of measurement error often increases with the true underlying value, as evidenced by the fact that the limit of quantification in assays is often defined in terms of the coefficient of variation

rather than the standard deviation. We consider here methods for correcting for heteroscedastic covariate measurement error. Our motivating example is provided by the BioCycle study, a study where one of the primary goals is to investigate the association between fat-soluble vitamins (e.g. β-carotene) and progesterone in human serum (Wactawski-Wende et al., 2009). The fat-soluble vitamins are measured with error using high performance liquid chromatography (HPLC). Guo et al. (2010) model calibration data on eight fat-soluble vitamin analytes with measurement variance $\sigma^2 X^{2\alpha}$, where X is the true value. They find that the constant variance assumption ($\alpha = 0$) is clearly violated, with estimates of α ranging from 0.5 to 0.8.

We consider data in the form displayed in Figure 3.1, where Y denotes a response variable, X denotes the covariate of interest, W denotes the error-prone measurement of X, and question marks denote unobserved values. The main study data consist of a sample of independent and identically distributed observations on (Y, W). The calibration data consist of a sample of independent and identically distributed observations on (X, W). Information about the measurement error is contained in calibration data with measured and true values of the covariate both recorded. We call the calibration data *internal* when they are a random sample from the main study, so they also contain observations of Y, as in Figure 3.1(a). We call the calibration data *external* when they are from another source, so information of Y is not available, as in Figure 3.1(b). We consider inference for the parameters of the regression of Y on X. The common case where there are additional error-free covariates is discussed in Section 3.7.

Comprehensive reviews of statistical methods for adjusting for measurement error include Fuller (1987) for linear models and Carroll et al. (2006) for nonlinear models. One commonly used and simple method is regression calibration (RC) (Carroll and

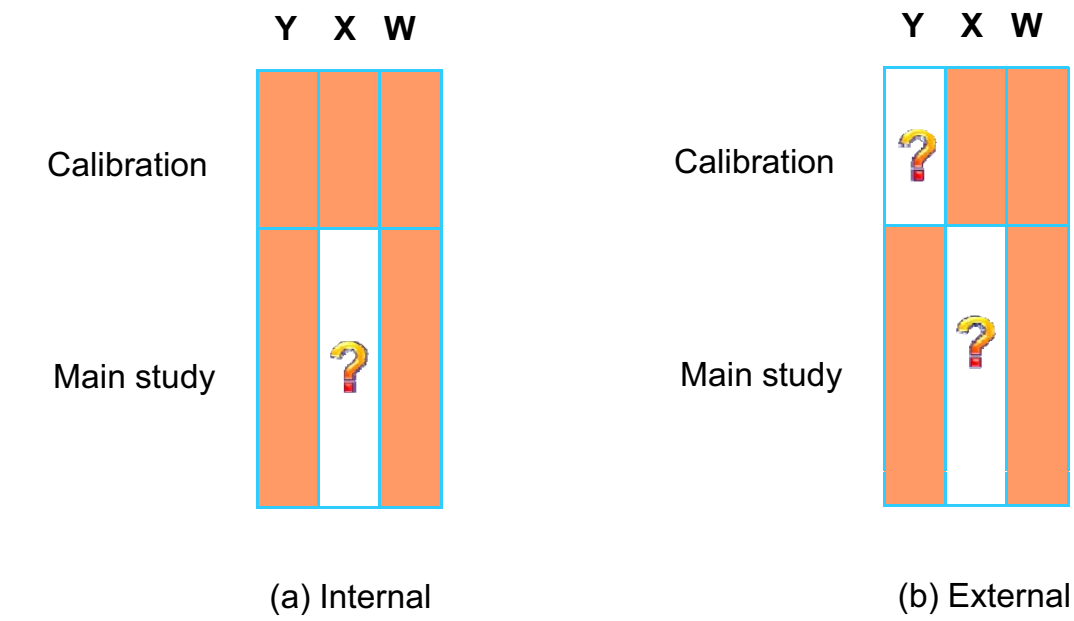

Figure 3.1: Validation calibration/main study design

Stefanski, 1990). The unobserved value X is imputed by the expected value of X given W, with coefficients estimated from the calibration data, and the regression of Y on X is then estimated using the filled-in data in the main study. A related method is moment reconstruction (MR), where imputed values are constructed to match the first two moments of Y and X (Freedman et al., 2004). This method is equivalent to RC in the linear regression case. For the case of internal calibration data, the estimate from RC can be combined with the estimate of the regression of Y and X computed directly from the calibration data, weighting the estimates from two sources according to their precisions. This method is known as efficient regression calibration (ERC); see Spiegelman et al. (2001).

An alternative approach is to use imputation or multiple imputation (MI) methods from the missing data literature to fill in the true values. Cole et al. (2006)

consider the MI method to remove the bias in the estimation of the hazard ratio for chronic kidney disease due to mismeasured covariates in a prospective cohort study. Schenker et al. (2009) describe MI to correct for measurement error of self-report data on health conditions in large-scale population surveys. Other Bayesian approaches for covariate measurement error are given in Clayton (1992); Richardson and Gilks (1993); Richardson and Green (1997); Hossian and Gustafson (2009); Yucel and Zaslavsky (2005).

Freedman et al. (2008) evaluate the performance of a number of these methods by simulation, for the case of internal calibration data. Several different scenarios are considered, including different choices of the measurement error variance and the strength of the response-covariate relationship. Data are simulated assuming non-differential measurement error (NDME), meaning that Y and W are independent given X. Their findings suggest that ERC is the preferred method. However, we note that unlike ERC, the version of MI used in this simulation does not exploit the NDME assumption. The MI methods described in this chapter are more efficient since they are based on an imputation model that makes the NDME assumption.

In addition, Freedman et al. (2008) assume that the variance of measurement error is constant, and do not assess methods when the measurement error is heteroscedastic. In that situation, existing error correction methods yield biased estimates, as our simulations demonstrate. We propose extensions of the methods compared in Freedman et al. (2008) to correct for heteroscedastic measurement error. We compare these methods through a simulation study, concluding that MI is the best of the methods compared in this setting.

The outline of the rest of this chapter is as follows. In Section 3.2, we specify models for the calibration data and main study data. In Section 3.3, we describe

various measurement error correction methods. We propose a simple extension of standard RC to deal with nonconstant variance. We also propose MI methods under both constant and nonconstant measurement variance, with and without the NDME assumption. MI methods are developed for both internal and external calibration designs. Our MI methods are Bayesian. The unobserved values of covariate X are replaced by draws generated using data augmentation (Tanner and Wong, 1987), and (in the nonconstant variance case) a Metropolis-Hastings (MH) algorithm (Hastings, 1970). A simplified approximate method that avoids the MH step is also developed. In Section 3.4, a simulation study is described, considering both constant and nonconstant measurement error variances, and both internal and external calibration study designs. Results from simulations are reported in Section 3.5. In Section 3.6, we illustrate the use of proposed methods on a real data example from the BioCycle study, where the effect of oxidative stress on female fecundity and fertility is investigated. In Section 3.7, we conclude with a discussion of the results and extensions of the proposed methods.

3.2 Models

Measurement error adjustments require an error model linking the true variable X to the surrogate measure W, which requires careful consideration in the context of the specific application (Heid et al., 2004; Guolo and Brazzale, 2008). The classical measurement error model assumes that

$$W = X + \xi \tag{3.1}$$

where ξ is random error with mean zero and constant variance (Dellaportas and Stephens, 1995; Hyslop and Imbens, 2001; Kuha and Temple, 2003). In our work we

consider a linear mean function and heteroscedastic measurement error, specifically:

$$p(W|X,\theta) \sim N(\beta_0 + \beta_1 X, \sigma^2 X^{2\alpha}) \tag{3.2}$$

where $\theta = (\beta_0, \beta_1, \alpha, \sigma^2)$, and the parameter α models heteroscedasticity. The measurement error variance is constant when $\alpha = 0$. In the main study, we assume a linear regression model of Y on X:

$$p(Y|X,\psi) \sim N(\gamma_0 + \gamma_X X, \tau^2) \tag{3.3}$$

where $\psi = (\gamma_0, \gamma_X, \tau)$, although more generally nonlinear relationships between Y and X can be modeled.

We assume that (Y, W) given X are bivariate normal with constant correlation ρ. Under the NDME assumption that Y and W are independent given X, $\rho = 0$ (Freedman et al., 2008). This assumption is common, and may be reasonable when measurement errors depend on bioassay techniques or laboratory experiments. NDME is less reasonable in retrospective case-control studies, where the disease status of subjects is known and the data about their exposure to risk factors are collected retrospectively, since recall error of past exposures is often thought to be more likely for cases than for controls (e.g., mothers of babies with a deformity may, on the average, have a different recall error about their early pregnancy drug intake than mothers of normal infants).

Further, we assume that the error model of W given X, and the regression model of Y given X hold with the same parameter values in both the main study sample and the calibration sample, when using external calibration data to assess measurement error. This assumption naturally holds for the internal calibration sample since it is

a subsample of the main study sample. Therefore, the final analysis can be applied to the completed data including both samples. The problem that models in the external calibration study may differ from those in the main study is discussed later in this chapter.

3.3 Measurement Error Correction Methods

3.3.1 Conventional Approach

The conventional approach (CA) fits an appropriate regression curve of W on X to the calibration data and estimates the true value of X using the value on the predicted calibration curve (Rodbard and Frazier, 1975; Finney, 1976; Higgins et al., 1998). For example, assuming a linear relationship between the true and measured values, the estimate of the true value is formed by inverting the calibration curve. That is, $\hat{X}_{CA} = (W - \hat{\beta}_0)/\hat{\beta}_1$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of the intercept and slope obtained from the regression of W on X using the calibration data. The estimate $\hat{X}_{CA}$ is then substituted for the unknown X in the main study data, and the regression model (3.3) is fitted to the data, yielding the CA estimate $\hat{\gamma}_{X,CA}$. The CA approach is biased for the regression coefficients but is nevertheless widely used in practice.

3.3.2 Regression Calibration

Regression calibration (RC) estimates the regression of X on W using the calibration data, and then substitutes the unknown values X in the main study with predictions $\hat{X}_{RC} = E(X|W)$ from this regression. The RC estimate $\hat{\gamma}_{X,RC}$ is then obtained by regressing Y on $\hat{X}_{RC}$. The RC method implicitly makes the NDME assumption.

The standard error of the RC estimate can be estimated using asymptotic cal-

culations (Spiegelman et al., 2001), or by bootstrapping the main and calibration samples. We create bootstrap samples from the calibration data and the main study data separately, and then combine them to compute RC estimates of the regression parameters. This procedure is repeated $B = 200$ times. The sample variance of the resulting B estimates is used to estimate the variance.

When the calibration data are internal, two estimates of the regression coefficient are available, the RC estimate $\hat{\gamma}_{X,RC}$, and the least squares estimate $\hat{\gamma}_{X,LSCalib}$ from fitting the linear regression model (3.3) to the calibration sample data on (Y, X). The ERC estimate is the inverse-variance weighted average of these two estimates,

$$\hat{\gamma}_{X,ERC} = w_{RC}\hat{\gamma}_{X,RC} + (1 - w_{RC})\hat{\gamma}_{X,LSCalib}$$

with weight

$$w_{RC} = \widehat{var}(\hat{\gamma}_{X,RC})^{-1}[\widehat{var}(\hat{\gamma}_{X,RC})^{-1} + \widehat{var}(\hat{\gamma}_{X,LSCalib})^{-1}]^{-1}$$

where $\widehat{var}(\hat{\gamma}_{X,RC})$ and $\widehat{var}(\hat{\gamma}_{X,LSCalib})$ are the estimated variances of $\hat{\gamma}_{X,RC}$ and $\hat{\gamma}_{X,LSCalib}$, respectively. The variance of $\widehat{\gamma}_{X,ERC}$ is computed approximately as $\widehat{var}(\hat{\gamma}_{X,ERC}) = [\widehat{var}(\hat{\gamma}_{X,RC})^{-1} + \widehat{var}(\hat{\gamma}_{X,calib})^{-1}]^{-1}$. ERC is more efficient than RC, particularly when the calibration data set is large.

We propose a modified version of RC, weighted RC (WRC), for situations where measurement error is heteroscedastic. This method estimates parameters of the regression model of X on W by weighted least squares. Specifically, we assume the regression of X on W can be approximated by the weighted regression model

$$p(X|W, \eta, \pi, \lambda) \sim N(\eta_0 + \eta_1 W, \pi^2 W^{2\lambda}) \tag{3.4}$$

The estimate $\hat{\gamma}_{X,WRC}$ is obtained by

- Estimating λ as the slope of a simple regression of logarithm of squared residuals of the regression of X on W on the logarithm of squared W using the calibration data.

- Estimating $\hat{\eta}_0$ and $\hat{\eta}_1$ by weighted least squares.

- Substituting unknown values X in the main study with the prediction, i.e., $\hat{X}_{WRC} = \hat{\eta}_0 + \hat{\eta}_1 W$.

- Estimating the coefficient $\hat{\gamma}_{X,WRC}$ of the regression of Y on $\widehat{X}_{WRC}$ from the main study data.

The associated standard error of the WRC estimate can be estimated using the bootstrapping approach mentioned above.

We can also modify the ERC estimate to account for heteroscedastic measurement error by replacing $\hat{\gamma}_{X,RC}$ with $\hat{\gamma}_{X,WRC}$. We call this the weighted ERC (WERC) estimate.

3.3.3 Multiple Imputation

We now develop MI methods based on a fully Bayesian model for the joint distribution of X, W and Y, which we write as $p(X, W, Y)$. These methods all create multiple imputations of the missing values – for the internal calibration design (Figure 3.1(a)), the missing values of X in the main sample, and for the external calibration design (Figure 3.1(b)), the missing values of X in the main sample and the missing values of Y in the calibration sample. In the work described in this chapter, we combine both the calibration sample and the main study sample in the final (post-imputation) analysis to provide inferences for the regression of Y on X. The MI estimates and associated standard errors are then obtained by applying standard

MI combining rules to the completed data including both samples (Rubin, 1987). Specifically, the MI estimate of γ_X is

$$\hat{\gamma}_{X,MI} = \frac{1}{m}\sum_{m=1}^{M} \hat{\gamma}_{X,MI}^{(m)}$$

and

$$Var(\hat{\gamma}_{X,MI}) = \bar{W} + B \times (M+1)/M$$

where B is the between-imputation variance, calculated as $B = \sum_{m=1}^{M}(\hat{\gamma}_{X,MI}^{(m)} - \hat{\gamma}_{X,MI})^2/(M-1)$; $\bar{W}$ is the average of the within-imputation variance, calculated as $\bar{W} = \sum_{m=1}^{M} var(\hat{\gamma}_{X,MI}^{(m)})/M$, and $var(\hat{\gamma}_{X,MI}^{(m)})$ is the standard variance estimator obtained from the m^{th} completed data. We apply this method to 16 completed data sets.

The form of the multiple imputations depends on the assumption made about α and ρ. When $\alpha = 0$, that is, the measurement error variance is constant, MI can be performed easily using standard Bayesian techniques for normal data described by Little and Rubin (2002). We assume uniform priors for the location parameters and log variances. The methods are labeled with a "0" to indicate the assumed value of α. For internal calibration data (Figure 3.1(a)), the imputation of X can be created assuming NDME ($\rho = 0$) – we label this method MIND0 – or not assuming NDME ($\rho \neq 0$) – we label this method MI0. The MI0 method is assessed in the simulation study of Freedman et al. (2008), but we also consider MIND0 to assess the gain of efficiency from basing imputations on the NDME assumption. For the external calibration design, the parameter ρ is not identified, so we only consider the MIND0 method that assumes NDME ($\rho = 0$).

For the cases where $\alpha \neq 0$ (i.e., the measurement error variance is not con-

stant), we consider two approaches for the case where α is unknown. One approach is to generate draws of this parameter from its posterior distribution, through a Metropolis-Hastings step. We describe here MI under the NDME assumption, where the joint distribution of W and Y given X can be factored as:

$$p(W, Y|X, \gamma, \tau, \beta, \sigma, \alpha) = p(W|X, \beta, \sigma, \alpha)p(Y|X, \gamma, \tau)$$

We add prior distributions for the marginal distribution of X and the parameters $(\gamma, \tau, \beta, \sigma, \alpha)$ to complete the fully Bayesian specification; Specifically, in the simulations we assume

$$p(X, \gamma, \tau, \beta, \sigma, \alpha) = p(X)p(\gamma, \tau, \beta, \sigma, \alpha)$$

where the prior distribution of X is a normal distribution with mean μ_x and variance σ_x^2, and the prior distribution of parameters $(\gamma, \tau, \beta, \sigma, \alpha)$ is a noninformative prior

$$p(\gamma, \log\tau, \beta, \log\sigma, \alpha) = const., \quad -2 < \alpha < 2$$

where the range $(-2, 2)$ for α includes values of that parameter thought likely to be of interest; the proper prior distribution for α is to ensure a proper posterior distribution (Guo et al., 2010). In this chapter, we consider a hierarchical normal structure for X, and assume a noninformative prior distribution for hyperparameters μ_x and σ_x^2, with $p(\mu_x, \log\sigma_x) = const.$, letting the posterior inferences be dominated by the observed data. Other choices of the prior distribution for X are discussed in the concluding section.

Draws can be conveniently computed using the data augmentation algorithm, which iteratively imputes missing values given observed data and draws of the parameters (*the imputation step*), and then draws parameters of the model from their posterior distribution given imputed values and observed data (*the posterior step*).

A Metropolis-Hastings step is required to generate draws of X. Details are given in Appendix 3.8. We label MI inferences from this algorithm MINDα.

We also develop a simpler version of MI that does not require the MH step, and can be viewed as a MI analog of the WRC method. The measurement error model is reformulated as the model (3.4), and the estimate $\hat{\lambda}$ of the parameter λ is substituted. For known λ, MI can be performed easily using standard Bayesian approaches for normal data. The estimate of λ can be computed using a simple regression approach, where λ is estimated as the slope of a regression of logarithm of squared residuals of the regression of X on W on the logarithm of squared W using the calibration data. We take into account the uncertainty of the estimation of λ by bootstrapping the calibration data to assure that MI is proper. We label this simplified MI method SMINDα.

3.4 Simulation Study

We assess the performance of the above methods by a simulation study. We consider both internal and external calibration data designs, vary the strength of the association of Y and X, the size of measurement error, and consider both homoscedastic and heteroscedastic measurement error. Simulation scenarios are generated by the following combinations of parameters:

Main study data: $\gamma_0 = 0$; $\gamma_x = 0.3$ or 0.6, $\tau^2 = 1$.

Measurement error data: $\beta_0 = 0$; $\beta_1 = 0.5$, 1 or 2; $\sigma^2 = 0.25$, 0.5 or 1; $\alpha = 0$ (homoscedastic measurement error) or 0.4 (heteroscedastic measurement error). The cases of $\sigma^2 = 0.5$ and 1 are investigated only for $\beta_1 = 0.5$, which results in five combinations.

To clarify the notation, "calib" and "main" will be attached to subscript to de-

note the calibration study and main study respectively. We simulate $n_{calib} = 100$ observations in the calibration sample and $n_{main} = 400$ observations in the main sample. For each scenario, 500 simulated data sets are generated.

The true X is first generated from a standard normal distribution with mean 0 and variance 1. Each main study data set is simulated by randomly generating the values for the response variable Y_i and the observed error-prone variable W_i for $i = 1, ..., n_{main}$, based on the models (3.2) and (3.3) respectively. For the external calibration data design, we randomly generate values of W_i from the measurement error model (3.2), for $i = 1, ..., n_{calib}$. For the internal calibration data design, we also generate responses Y_i, $i = 1, ..., n_{calib}$, using the model (3.3) with the same values of γ and τ^2 used to simulate the corresponding main study data.

For each of the 500 simulations across each of the simulation scenarios, we estimate the parameter of interest γ_X for each of the measurement error correction methods described above. All methods are compared with respect to bias, root mean squared error (RMSE) of the estimates and empirical non-coverage of confidence intervals. The empirical non-coverage is calculated as the proportion of simulated data sets for which the 95% confidence interval does not include the true value of γ_X. The proportions are multiplied by 1000 to avoid decimal points, and hence a nominal level of non-coverage is equal to 50.

3.5 Results

In Tables 3.1 - 3.3, we examine the performance of the naïve regression of Y on W (i.e., ignoring measurement error), and measurement error correction techniques CA, RC, and MI. We focus on the performance of various methods on inferences for the regression coefficient γ_X. We compare the results under the situations where the

Y - X association is week (small $\gamma_x = 0.3$) and strong (large $\gamma_x = 0.6$), where the magnitude of measurement error is small (minor error variance $\sigma^2 = 0.25$), moderate (fair error variance $\sigma^2 = 0.5$) and large (big error variance $\sigma^2 = 1$). Table 3.1 compares Naïve, CA, RC, ERC, LSCalib, MI0 and MIND0 for the case of internal calibration data with homoscedastic measurement error. Table 3.2 compares Naïve, CA, RC, MI0, MIND0, WRC, WERC, LSCalib, MINDα, and SMINDα for internal calibration data with heteroscedastic measurement error. Table 3.3 compares Naïve, CA, RC, MIND0, WRC, MINDα, and SMINDα for external calibration data with heteroscedastic measurement error; we do not evaluate the LSCalib, ERC and WERC methods since they are not applicable for this data structure.

Table 3.1 presents the results for the case of homoscedastic measurement error. As theory predicts, the estimates from the naïve analysis (using W in place of X) are attenuated towards 0, with the degree of attenuation varying with the magnitude of measurement error and the response-covariate association. In addition, the non-coverage rate of the naïve estimate is much higher than the nominal level of 50 in most of simulation scenarios. The CA method also performs poorly, with substantial bias and poor confidence interval coverage, particularly when the measurement error is large. The RC estimate is much less biased and has much better coverage than Naïve and CA, but has very large RMSE when the measurement error variance is large, suggesting that it is not very efficient. The estimates of ERC, LSCalib, MI0 and MIND0 methods have no bias or little bias. When measurement error is small or moderate (e.g. $\sigma^2 = 0.25$ or 0.5), the RMSE of the ERC estimate is generally smaller than that of the RC, LSCalib and MI0 estimates. This finding is consistent with the simulation results in Freedman et al. (2008). It is worth pointing out that the RMSE of the ERC estimate is better than that of the MI0 estimate but not for

Table 3.1: Empirical bias, RMSE and non-coverage of 95% confidence interval (nominal = 50) of estimates of γ_x with the internal calibration data based on 500 simulations, when the variance of measurement error is constant. All values are multiplied by 1000.

γ_x	β	σ^2	Estimation	Naïve	CA	RC	LSCalib	ERC	MI0	MIND0
0.3	2	0.25	Bias	159	19	1	0	1	2	1
			RMSE	160	51	52	98	46	50	45
			Non-coverage	1000	82	68	58	52	42	50
0.3	1	0.25	Bias	61	62	1	1	2	3	2
			RMSE	77	77	60	101	50	61	49
			Non-coverage	268	266	68	58	52	44	52
0.3	0.5	0.25	Bias	2	152	6	0	3	3	1
			RMSE	73	157	82	99	67	71	61
			Non-coverage	58	952	52	58	44	42	44
0.3	0.5	0.5	Bias	101	202	16	1	5	3	2
			RMSE	118	204	110	100	70	73	66
			Non-coverage	398	998	48	58	38	58	44
0.3	0.5	1	Bias	181	241	23	1	4	4	3
			RMSE	187	243	315	99	79	76	70
			Non-coverage	970	1000	40	58	46	52	40
0.6	2	0.25	Bias	318	38	1	0	0	3	1
			RMSE	319	63	54	98	45	51	45
			Non-coverage	1000	110	64	58	48	40	48
0.6	1	0.25	Bias	122	124	2	0	2	3	2
			RMSE	130	134	66	98	54	62	54
			Non-coverage	742	652	64	58	40	46	48
0.6	0.5	0.25	Bias	4	303	18	1	3	3	3
			RMSE	71	307	105	96	69	75	66
			Non-coverage	50	1000	52	58	42	48	42
0.6	0.5	0.5	Bias	201	401	40	0	4	4	3
			RMSE	209	404	151	98	77	79	69
			Non-coverage	894	1000	42	58	40	52	38
0.6	0.5	1	Bias	359	480	54	1	5	4	3
			RMSE	362	481	531	101	86	80	71
			Non-coverage	1000	1000	40	58	57	50	42

Naïve: naïve linear regression of Y on W; CA: conventional approach; RC: regression calibration; LSCalib: linear regression of Y on X using calibration data only; ERC: efficient regression calibration; MI0: multiple imputation without the NDME assumption; MIND0: multiple imputation with the NDME assumption.

MIND0 in small measurement error settings. As an example, when $\gamma_x = 0.3$, $\beta = 1$ and $\sigma^2 = 0.25$, the RMSE of the ERC, MI0 and MIND0 estimates are 50, 61, and 49, respectively. For large measurement error and strong response-covariate association, the performance of the ERC method becomes less impressive; MI0 and MIND0 both outperform ERC. Overall, the inferences from the MIND0 approach are superior to other methods, with small empirical bias, low RMSE and close to nominal levels of non-coverage.

Table 3.2 concerns heteroscedastic measurement error for the internal calibration data design, so the ERC, WERC and LSCalib methods are available for comparison. The MI methods taking into account heteroscedastic measurement error perform best among compared methods for all simulation scenarios considered, with respect to bias, RMSE and non-coverage of confidence interval. The simplified version SMINDα (i.e., estimating α using a simple regression) is comparable to MINDα (i.e., estimating α using a MH algorithm). Naïve and CA both yield seriously biased estimates of γ_x with high non-coverage. The RC estimate becomes badly biased with inflated RMSE, particularly when measurement error and the response-covariate effect increase. WRC has less empirical bias than RC, but some bias remains, and it becomes unstable with large RMSE when the magnitude of measurement error is large. The WERC estimate has smaller empirical bias and lower RMSE than WRC, especially for large measurement error. For example, when $\gamma_1 = 0.6$, $\beta_1 = 0.5$ and $\sigma^2 = 1$, the RMSE of the WERC estimate is 0.102, compared with 1.655 for the WRC estimate. The possible reason for the gain of WERC over WRC is that the inflated standard error of WRC reduces its effect on WERC, and the LSCalib estimate stabilizes the estimation. When there is little measurement error ($\sigma^2 = 0.25$), the WERC estimates are comparable to the MINDα estimates, with small bias and

Table 3.2: Empirical bias, RMSE and non-coverage of 95% confidence interval (nominal = 50) of estimates of γ_x with the internal calibration data based on 500 simulations, when the variance of measurement error is heteroscedastic. All values are multiplied by 1000.

γ_x	β	σ^2	Estimation	Naïve	CA	RC	LSCalib	MIND0	MI0	WRC	WERC	MINDα	SMINDα
0.3	2	0.25	Bias	175	47	3	5	2	3	1	0	0	0
			RMSE	177	69	61	96	50	59	61	53	48	49
			Non-coverage	1000	176	70	46	58	60	60	68	58	46
0.3	1	0.25	Bias	130	129	8	5	5	7	7	2	2	3
			RMSE	136	136	81	96	64	76	80	64	58	59
			Non-coverage	912	824	66	46	58	64	60	66	56	44
0.3	0.5	0.25	Bias	152	226	29	5	2	8	22	6	3	4
			RMSE	161	228	156	96	76	88	147	80	69	72
			Non-coverage	822	1000	48	46	54	60	46	48	46	50
0.3	0.5	0.5	Bias	216	258	116	5	8	7	20	10	4	5
			RMSE	220	259	1335	96	81	92	699	88	71	79
			Non-coverage	1000	1000	42	46	50	78	42	60	46	44
0.3	0.5	1	Bias	250	274	251	3	6	9	47	11	6	7
			RMSE	252	274	1951	106	83	94	882	99	72	81
			Non-coverage	1000	1000	40	60	58	68	40	58	52	44
0.6	2	0.25	Bias	349	94	5	5	3	6	3	0	0	1
			RMSE	350	109	68	96	54	63	67	57	52	53
			Non-coverage	1000	430	64	46	52	60	60	64	56	52
0.6	1	0.25	Bias	259	257	16	8	7	6	16	2	4	3
			RMSE	263	262	101	96	70	80	101	72	60	67
			Non-coverage	1000	988	72	46	46	70	56	54	46	48
0.6	0.5	0.25	Bias	302	450	57	5	6	13	45	8	3	9
			RMSE	307	451	230	96	78	90	221	87	70	78
			Non-coverage	998	1000	38	46	50	62	38	54	48	48
0.6	0.5	0.5	Bias	430	514	220	5	8	13	52	13	8	12
			RMSE	432	515	2196	96	82	94	1327	94	72	78
			Non-coverage	1000	1000	42	46	52	74	40	62	56	44
0.6	0.5	1	Bias	505	550	397	3	7	11	129	12	7	13
			RMSE	506	551	3457	106	87	96	1655	102	76	85
			Non-coverage	1000	1000	36	60	58	70	40	64	44	38

Naïve: naïve linear regression of Y on W; CA: conventional approach; RC: regression calibration; LSCalib: linear regression of Y on X using calibration data only; WRC: weighted regression calibration; WERC: weighted ERC; MI0: multiple imputation without the NDME assumption; MIND0: multiple imputation with the NDME assumption. MINDα: multiple imputation with the NDME and $\alpha \neq 0$ assumptions; SMINDα: simplified version of MINDα.

low RMSE. When measurement error is large and the response-covariate association is strong, the RMSE of the WERC estimate is larger than that of MINDα. The MI methods that fail to model the nonconstant measurement error variance (MI0 and MIND0) have larger RMSE and higher non-converge than the MI methods that allow for nonconstant variances.

Results of the external calibration design are presented in Table 3.3. We observe similar trends to those seen in Table 3.2. The Naïve and CA estimate are both subject to large bias, and the corresponding non-coverage is high. RC performs poorly with large bias and RMSE when the measurement error variance is large. The WRC method has reduced empirical bias, but is inefficient for large measurement error. The RMSE of the MIND0 estimate is generally greater than that of the MINDα estimate. Overall, the MINDα method and its simplified version SMINDα dominate all other correction methods.

3.6 Application

In this section, we perform an analysis of the dataset from the BioCycle study. This study was designed to assess the relationship between endogenous hormones and biomarkers of oxidative stress during the menstrual cycle. Two hundred and fifty nine regularly menstruating pre-menopausal women were followed for two menstrual cycles. A goal of the study was to investigate the association between carotenoids (β-carotene) and progesterone.

The calibration data were obtained from calibration experiments for human serum fat-soluble vitamins, measured by high performance liquid chromatography (HPLC). Three replicate calibration experiments were performed, with each experiment analyzing eight samples with known concentrations of the analyte carotene, from stan-

Table 3.3: Empirical bias, RMSE and non-coverage of 95% confidence interval (nominal = 50) of estimates of γ_x with the external calibration data based on 500 simulations, when the variance of measurement error is heteroscedastic. All values are multiplied by 1000.

γ_x	β	σ^2	Estimation	Naïve	CA	RC	MIND0	WRC	MINDα	SMINDα
0.3	2	0.25	Bias	172	46	3	5	2	0	0
			RMSE	174	66	57	61	57	52	58
			Non-coverage	1000	180	64	88	60	54	52
0.3	1	0.25	Bias	129	125	4	13	3	2	6
			RMSE	134	132	80	81	78	74	79
			Non-coverage	910	836	60	100	56	46	52
0.3	0.5	0.25	Bias	148	223	28	55	23	7	11
			RMSE	156	224	157	177	150	129	135
			Non-coverage	822	1000	40	128	52	56	46
0.3	0.5	0.5	Bias	212	255	72	86	58	8	14
			RMSE	216	256	389	246	378	162	171
			Non-coverage	1000	1000	44	196	42	48	42
0.3	0.5	1	Bias	250	274	251	68	109	10	12
			RMSE	252	274	1951	306	888	213	239
			Non-coverage	1000	1000	40	212	40	56	38
0.6	2	0.25	Bias	346	92	5	8	3	2	2
			RMSE	347	106	63	68	62	61	65
			Non-coverage	1000	422	66	106	64	58	54
0.6	1	0.25	Bias	255	253	5	27	4	6	7
			RMSE	259	258	98	107	97	92	100
			Non-coverage	1000	998	58	120	56	54	50
0.6	0.5	0.25	Bias	299	447	51	82	42	8	13
			RMSE	304	449	252	210	234	145	166
			Non-coverage	1000	1000	40	228	48	52	46
0.6	0.5	0.5	Bias	427	513	197	99	91	18	23
			RMSE	429	513	1211	241	1016	171	194
			Non-coverage	1000	1000	42	268	40	56	58
0.6	0.5	1	Bias	510	556	314	93	166	21	28
			RMSE	511	556	2526	306	1774	236	251
			Non-coverage	1000	1000	48	274	52	58	54

Naïve: naïve linear regression of Y on W; CA: conventional approach; RC: regression calibration; WRC: weighted regression calibration; MIND0: multiple imputation with the NDME assumption. MINDα: multiple imputation with the NDME and $\alpha \neq 0$ assumptions; SMINDα: simplified version of MINDα.

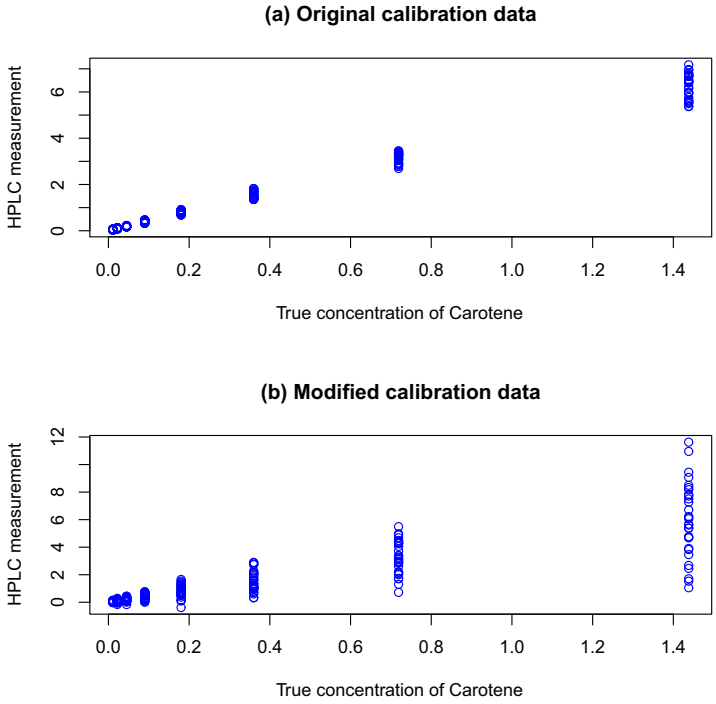

Figure 3.2: Calibration data of carotene from the BioCycle study

dard reference materials (SRMs) obtained from the National Institute of Standards and Technology (NIST). For each of three replicate experiments, each of eight samples were analyzed 10 times by HPLC, yielding a total of 30 replicate measures for each sample. In the calibration data, the true concentrations of carotene X are known, and HPLC measurements can be viewed as error-contaminated versions of X, denoted by W. This calibration data is external, since it only includes the information of W and X.

The main study consisted of 211 individual samples with complete information on both progesterone and carotene. Each individual sample had one measurement of outcome progesterone Y and one HPLC measurement of carotene W, but the true concentration of carotene X was unknown. We use the data collected at the visiting time $F1$.

Table 3.4: BioCycle data: estimated regression coefficients in a linear regression model with carotene as the covariate and progesterone as the dependent variable using data collected at visiting time F1. The calibration data of carotene, which is generated from standard reference materials (SRMs) 968 C1, is external. Standard error is shown in parentheses.

Parameters	Naïve	CA	RC	WRC	MINDα
intercept	0.5089(0.0238)	0.5089(0.0238)	0.5095(0.0244)	0.5099(0.0242)	0.5089(0.0239)
carotene	-0.0065(0.0024)	-0.0281(0.0104)	-0.0289(0.0107)	-0.0286(0.0106)	-0.0282(0.0107)

Figure 3.2(a) shows the original calibration data for carotene. It is clear that the variance of HPLC measurements increases as the true concentration of carotene increases, suggesting that the measurement error variance is not constant. To estimate the linear regression of progesterone Y on carotene X, we apply the naïve regression of Y on W, and four different measurement error correction methods (i.e., CA, RC, WRC and MI) to the data. Standard errors of the CA, RC and WRC methods are calculated using the bootstrapping method. Note that these methods are applied to correct only for measurement error from the assay, but not from other sources such as biological variation.

Table 3.4 presents the estimates and associated standard errors for regression coefficients. The naïve estimate indicates a weak association between progesterone and carotene – the change of progesterone is 0.0065 when carotene changes one unit. Adjusting for measurement error by the CA, RC, WRC and MI methods, we find a stronger association between carotene and progesterone. The error-corrected estimates are all four-fold greater than the uncorrected estimates. Specially, the CA method estimates a change of progesterone of 0.0281 for one unit change in carotene. The estimates obtained from the RC, WRC and MI methods are similar, and slightly larger than the CA estimate.

Differences between the correction methods are minor in this example. We be-

Table 3.5: Modified BioCycle data: estimated regression coefficients in a linear regression model with carotene as the covariate and progesterone as the dependent variable using data collected at visiting time F1. The calibration data is external. Standard error is shown in parentheses.

Parameters	Naïve	CA	RC	WRC	MINDα
intercept	0.5068(0.0238)	0.5066(0.0238)	0.5110(0.0242)	0.5085(0.0231)	0.5128(0.0253)
carotene	-0.0061(0.0024)	-0.0255(0.0099)	-0.0339(0.0123)	-0.0297(0.0113)	-0.0285(0.0105)

lieve this is because the magnitude of measurement error is not large relative to the size of the effect being estimated. To provide some empirical evidence, we analyze the calibration data of carotene and calculate the maximum likelihood estimator of the measurement error variance $\widehat{\sigma}^2$. We find that $\widehat{\sigma}^2 = 0.117$, which indicate the magnitude of measurement error is quite small. The simulation studies suggest that the performance of CA, RC and MI is more differentiated when the magnitude of measurement error is large.

To better illustrate our proposed approaches, we create a modified BioCycle data set, where carotene is measured with more error. Specifically, some random noise is added to HPLC measurements to increase the magnitude of measurement error. Figure 3.2(b) shows the modified calibration data, which maintain the same pattern as the original calibration data but are more dispersed, with $\widehat{\sigma}^2 = 2.496$. We analyze the modified data with our proposed methods, and results are summarized in Table 3.5.

As shown in Table 3.5, the naïve analysis attenuates the association between carotene and progesterone toward to the null, as expected. Noticeably, when applying our proposed error correction methods to the modified BioCycle data, there is appreciable difference in the estimates of regression coefficient of carotene. In particular, the estimates of the CA, RC, WRC and MINDα methods are -0.0255,

-0.0339, -0.0298 and -0.0280, respectively. We also observe that the RC estimate has a larger standard error than the WRC or MI estimate.

3.7 Conclusion and Discussion

The simulations in Freedman et al. (2008) show that ERC outperforms MI for the case of homoscedastic measurement error. The reason is that ERC exploits the NDME assumption, whereas the version of multiple imputation (MI0) considered by these authors does not. The simulations reported in Table 3.1 indicate that a version of MI that exploits the NDME assumption (MIND0) is similar or superior to ERC for the simulation conditions compared. This finding is to be expected, given the asymptotic efficiency of MI under a correctly specified model, as the number of imputations tends to infinity. Our simulation results also show the efficiency gains of MIND0 over MI0, demonstrating the utility of taking into account the NDME assumption when it is substantively reasonable. It may be seen that the RMSE of the MIND0 estimate is generally 10%-15% smaller than that of the MI0 estimate in our simulation settings.

The main focus of this chapter is on extending methods to the case of heteroscedastic measurement error, a situation where existing methods are biased. In particular, the RC method, which imputes a conditional mean of X given W, does not yield consistent estimates when the measurement error variance is not constant. Our modification WRC of RC, based on estimating the conditional means by weighted least squares, reduces but does not solve this problem. In contrast, the MI methods, which impute draws rather than means, can readily allow for nonconstant measurement variance by simply modifying the imputation model to reflect this feature. The RC method and its extensions rely on the assumption that measurement error

is non-differential, and in this chapter we focus on the performance of MI methods based on the NDME assumption. However, we note that the MI methods can also handle differential measurement error, provided that internal calibration data are available to identify the model parameters. In detail, the MINDα method designed under the NDME assumption can be easily extended to work in the case where the assumption may not hold, i.e., allowing for differential measurement error by modifying the imputation model in the following way — remodeling the measurement error as $p(W|X,Y)$, instead of $p(W|X)$.

Although the MINDα method performs well in the simulation studies, it is comparatively complex computationally, given its use of MCMC with a MH step. We provide an alternative. SMINDα is much simpler computationally since it avoids the MH step. It is based on an approximate model similar to WRC, so it lacks statistical rigor. However, SMINDα performs well and similarly to MINDα in the simulations, and it takes much less time to compute.

Our Bayesian MI methods require specification of prior distributions for the parameters, and also for $p(X)$. We chose a simple noninformative normal prior distribution for $p(X)$, and resulting estimates had good frequentist properties in the simulations. However, other choices may be worth considering. For example, mixtures of normal distributions for $p(X)$ have been proposed with a prespecified (Carroll et al., 1999) or unknown (Richardson et al., 1999) number of components for linear measurement error models. These methods could be adapted to our heteroscedastic setting. We assume the measurement error variance model is $\sigma^2 X^2$, indicating that the variance increases as the true value of X increases. This chosen nonconstant variance model is common (though others are possible). It is of interested, although more complex, to extend the proposed MI methods to allow for a more general form

of the nonconstant variance, like $\sigma^2 g^2(\beta, \alpha, X)$ with an unknown parameter vector (β, α).

We have restricted attention here to the case where a simple regression of Y on X is of interest. It is relatively straightforward to extend our MI methods to allow for other covariates Z, recorded without measurement error, since values of these variables can be conditioned in the MCMC analysis. Extensions to non-normal outcomes, as when Y is binary and follows a probit model, could also be developed without too much difficulty.

We assume that the same measurement error model holds for the calibration and main data sets. In some epidemiologic study designs, the calibration data are supplied by an external source, such as a pure standard sample, and the relationship of the true and measured variables might be different for the calibration and main study data, because the sample from each subject of the main study might have impurities that change the measurement error properties. In these situations where the assumption of the same models in both the calibration and main study samples is clearly violated, using external data to adjust for measurement error may introduce bias. Hence methods that allow a different measurement model are worthwhile in future research.

In applications, the choice of method should be considered based on the measurement error structure. In particular, the application to the BioCycle data suggests that the CA method may yield reasonable estimates, since the measurement error is small. The discrepancy between CA and the other various methods becomes more substantial as the measurement error increases, as demonstrated by the modified BioCycle data. The developed MI methods can be expected to have much appreciable impact when the response-covariate association is strong and the measurement

error is large.

3.8 Appendix

In this chapter, we apply the MI approach via data augmentation using a Markov Chain Monte Carlo (MCMC) algorithm. We consider measurement error problems in a missing data context where the true value of X is unoberved. The data augmentation method is a two-step iterative algorithm. The key idea is to iteratively generate draws for missing values given observed data and a set of parameters (I-step) and generate draws of the model parameters from their posterior distribution given complete data (P-step). These two steps are iterated long enough to the convergence, and then the results are reliable to be treated as multiply imputed data sets (Schafer, 1997).

Suppose we have data collected from the main study containing n_{main} observations (W_i, Y_i) and from the external calibration study containing n_{calib} observations (X_i, W_i). The observations are assumed to be independently and identically distributed. We let n denote the number of all observations, $n = n_{calib} + n_{main}$.

The posterior predictive distribution of X given W, Y and the parameters cannot be expressed in a closed form, however, by Bayes' Theorem, it can be factorized as

$$\begin{aligned} p(X|Y, W, \beta, \sigma, \alpha, \gamma, \tau, \mu_x, \sigma_x) &\propto p(W|X, Y, \beta, \sigma, \alpha)p(Y|X, \gamma, \tau)p(X|\mu_x, \sigma_x) \\ &\propto p(W|X, \beta, \sigma, \alpha)p(Y|X, \gamma, \tau)p(X|\mu_x, \sigma_x) \end{aligned}$$

This factorization represents three models: the measurement error model which links the true variable X and the observed variable W, the main study model which specifies the relationship between the response variable Y and the unobserved covariate X, and the prior distribution for X.

Assuming that the parameter vectors $\theta = (\beta, \sigma, \alpha)$, $\psi = (\gamma, \tau)$ and $\pi = (\mu_x, \sigma_x)$

are distinct and *priori* independent, the likelihood function can be factorized, and the joint posterior for the parameters given the complete data can be expressed as

$$p(\beta,\sigma,\alpha,\gamma,\tau,\mu_x,\sigma_x|Y,X,W) \propto p(Y,X,W|\beta,\sigma,\alpha,\gamma,\tau)p(\beta,\sigma,\alpha,\gamma,\tau,\mu_x,\sigma_x)$$
$$\propto p(W|X,\beta,\sigma,\alpha)p(\beta,\sigma,\alpha)p(Y|X,\gamma,\tau)p(\gamma,\tau)p(X|\mu_x,\sigma_x)p(\mu_x,\sigma_x)$$

Hence θ, ψ and π can be sampled separately.

For the external calibration/main design, we also need impute missing values for Y. The posterior distribution of Y is given straightforwardly as a normal distribution with mean $(\gamma_0 + \gamma_X X)$ and variance τ^2.

Having obtained the complete-data posteriors for the model parameters, and the condition predictive distribution for X and Y, our imputation procedure for the external calibration/main study composes of the following steps:

I-1 step: Generate imputed values of Y_i for i corresponding to the ith observation in the calibration study, $i = 1, ..., n_{calib}$, from the posterior density

$$p(Y_i|X_i,\psi) \sim N(\gamma_0 + \gamma_X X_i, \tau^2)$$

I-2 step: Generate imputed values of X_i for i corresponding to the ith observation in the main study, $i = 1, ..., n_{main}$, from the posterior density specified by

$$p(X_i|Y_i,W_i,\beta,\sigma,\alpha,\gamma,\tau) \propto \tau^{-2} exp(-\frac{1}{2\tau^2}(Y_i - \gamma_0 - \gamma_X X_i)^2)$$
$$\times \sigma^{-2} X_i^{-\alpha} exp(-\frac{1}{2\sigma^2 X_i^{2\alpha}}(W_i - \beta_0 - \beta_1 X_i)^2)$$
$$\times \sigma_x^{-2} exp(-\frac{1}{2\sigma_x^2}(X_i - \mu_x)^2)$$

P-1 step: Draw ψ from the posterior density $p(\psi|X,Y)$.

P-2 step: Draw θ from the posterior density $p(\theta|W,X)$.

P-3 step: Draw π from the posterior density $p(\pi|X)$.

For the internal calibration/main study design, we observe (Y, X, W) in the calibration study. Therefore, ML estimates for the regression parameters γ_0, γ_X and τ^2, as well as the measurement error model parameters β_0, β_1 and σ^2 can be computed using the calibration data by standard analysis (e.g., least squares methods), and used as initial values for the data augmentation algorithm. The sample mean and variance of X can also be calculated from the calibration data, and used as initial values for μ_x and σ_x^2. The initial value of α is estimated as the slope of regression of logarithm of squared residuals of the regression of X on W on the logarithm of squared W. For the external calibration/main study design, we still can obtain estimates of the measurement error parameters β_0, β_1, σ^2 and α, using the same approach for the internal calibration design. However, the initial values of γ_0, γ_X and τ^2 can not be computed straightforwardly because X and Y are not observed together in the whole data set. We apply the regression calibration method to obtain initial values. In detail, we first substitute unobserved values of X in the main study by the expectation $E(X|W)$, and then regress Y on the substituted values to obtain initial estimates for γ and τ^2.

The data augmentation method iterates between the P-step and the I-step for a large number of times until the algorithm converges. We use the method proposed by Gelman and Rubin (1992) for convergence diagnosis of the model parameters. Initial values of model parameters are chosen to be reasonably overdispersed by bootstrapping the original data set. For each of the bootstrap samples, estimates of model parameters are obtained using the method described above, and used as the initial points for the MCMC chains. After the algorithm converges, We discard data from the initial burn-in period before saving the values. The imputed complete data set is generated by using every d^{th} iteration after an adequate burn-in period,

to avoid possible autocorrelation between successive sets of imputed values, so that the imputed data sets can be treated as independent. For all the model parameters, we observe reasonable mixing and convergence after 2000 iterations of the MCMC chains. Hence we decide to discard the first 2000 iterations, and choose $d = 100$. We generate 16 imputed data sets, which can be analyzed by standard complete data inference.

The imputation procedure for the internal calibration/main study is similar to that of the external calibration/main study, except that the I-1 step can be omitted since Y is observed in this design.

The I-1 step can be performed easily by generating a random draw from a normal distribution. The I-2 step is not straightforward, since we don't have an analytical expression for the posterior density of X. We use a Metropolis-Hastings algorithm for this step to generate values of X_i. It consists of the following two steps:

- Generate a value $\boldsymbol{X_i}'$ from an appropriate candidate generating density $q_i(\boldsymbol{X_i}, \boldsymbol{X_i}')$, where $\boldsymbol{X_i}$ denote the current value
- Set

$$\boldsymbol{X_i}^{(j+1)} = \begin{cases} \boldsymbol{X_i}' & \text{with probability } \kappa = min[1, \frac{\pi(\boldsymbol{X_i}')q(\boldsymbol{X_i}', \boldsymbol{X_i}^{(j)})}{\pi(\boldsymbol{X_i}^{(j)})q(\boldsymbol{X_i}^{(j)}, \boldsymbol{X_i}')}], \\ \boldsymbol{X_i}^{(j)} & \text{otherwise.} \end{cases}$$

  where $\pi(X_i)$ is the target density from which we want to simulate X.

The choice of candidate generating density is arbitrary, but a correctly specified density can improve the efficiency of this algorithm. In our work, we generate a candidate $\boldsymbol{X_i}'$ from a Gaussian model centered on the current value $\boldsymbol{X_i}^{(j)}$ and variance σ_*^2, equal to a scaled sampling variance of the target density. That is,

$q(\boldsymbol{X_i}', \boldsymbol{X_i}^{(j)}) = N(\boldsymbol{X_i}^{(j)}, \sigma_*^2)$. The algorithm with the normal generating density is also called "a random walk Metropolis" algorithm.

The P-step requires generating the draws of the parameters from the complete-data posterior distribution. In P-1 step, by Bayes's rule, the posterior for $\psi = (\gamma, \tau)$ given the data (X, Y) can be factored as a multivariate normal distribution and a scaled inverted χ^2 -distribution, which make it easy to draw the values. In practice, we first draw τ^2 from $\tau^2 \sim inv - \chi^2(\upsilon, S^2)$; and then draw γ from multivariate normal distribution $N(\hat{\gamma}, \tau^2(X'X)^{-1})$, where $\hat{\gamma} = (X'X)^{-1}X'Y$ is the ordinary least squares estimate from the regression of all data, $S^2 = (Y - X\hat{\gamma})'(Y - X\hat{\gamma})$ is the corresponding sample variance, and $\upsilon = n - 2$ is the degree of freedom.

Drawing parameter θ from the posterior density $p(\theta|X, W)$ is a little complicated due to presence of the unknown parameter α. For known α, draws of β, σ^2 can be readily obtained from their posterior density using an approach similar to the P-1 step, with the least squares estimate replaced by the weighted least squares estimate. In detail, let $\omega_i = 1/X_i^{\alpha}$, we first generate values of α from its posterior distribution,

$$p(\alpha|\beta_0, \beta_1, \sigma^2, X, W) \propto (\prod_{i=1}^{n} \omega_i) exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \omega_i^2 (W_i - \beta_0 - \beta_1 X_i)^2\}, i = 1, ..., n$$

using a random walk Metropolis step. Given α, we draw σ^2 from $\sigma^2 \sim inv{-}\chi^2(\upsilon, S_w^2)$; and then draw β from multivariate normal distribution $N(\hat{\beta}, \sigma^2(X'\omega X)^{-1})$, where $\hat{\beta} = (X'\omega X)^{-1}X'\omega W$ is the weighted least squares estimate from the regression of all data, ω is a $n \times n$ weighted matrix with the diagonal element ω_i and the other elements equal to zero, $S_w^2 = (W - X\hat{\beta})'\omega(W - X\hat{\beta})$ is the corresponding weighted sample variance, and $\upsilon = n - 2$.

CHAPTER IV

Multiple Imputation for Covariate Measurement Error Correction based on Summary Statistics from External Calibration Data

4.1 Description of the Problem

Many studies in epidemiology involve biomarkers that are recorded with measurement error, and measurement error is known to distort inferences when these biomarkers are included as predictors in regression analysis. Specifically, it is well known that regression coefficients of variables subject to measurement error are attenuated, and when the variables subject to measurement error are covariates, treatment effects are potentially estimated with bias (Morgan and Elashoff, 1987; Richardson and Gilks, 1993; Zidek et al., 1996; Fung and Krewsk, 1999; Sarkar and Qu, 2007). Despite these facts, quantitative adjustments are rarely applied in epidemiological studies (Jurek et al., 2004).

Often information about measurement error is contained in a calibration experiment such as a bioassay, where samples with known values of the variable are analyzed by a measuring instrument, and the regression of the measured values on the true values is estimated, often in the form of a linear calibration curve (Higgins et al., 1998). The true values of future measurements are then simply estimated from the calibration curve, and these values are treated as the true values in the main anal-

ysis. Values that have high measurement error relative to the true values are often reported as below the limit of detection (LD). Browne et al. (2010) provide a review of methods for determining the LDs and related quantities. Simulations have shown that this approach, which we call classical calibration (CA), yields biased estimates when the measurement error is substantial (Guo et al., 2010). They note that this usual way of providing information from calibration experiments to users does not allow valid statistical inferences involving the true values of the biomarker, and hence better methods are needed.

We consider data from a main study and a calibration sample in the form of Figure 4.1. The main analysis concerns the regression of Y on X and Z, where, Y is the outcome of interest, X is the true value of the biomarker of interest, and Z denotes a set of other covariates, assumed to be measured without error. The data in the main study are a random sample on Y, Z and W, where W is the measured version of the biomarker, which we assume to be the true value X measured with error. Information relating W and X is gained from a calibration sample, which includes measurements on W and X. The question marks in Figure 4.1 denote unobserved values.

Figure 4.1 contrasts two calibration sample designs, which we call "internal calibration" and "external calibration". In internal calibration, Figure 4.1(a), values of X and W are available for a subsample of the main study participants, and hence values of Y and Z are also recorded for this subsample. External calibration is carried out independently of the main study, for example by an assay manufacturer, so values of Y and Z are not recorded for the calibration sample, yielding the more sparse missing data pattern of Figure 4.1(b).

External calibration is the harder problem, but the literature on measurement er-

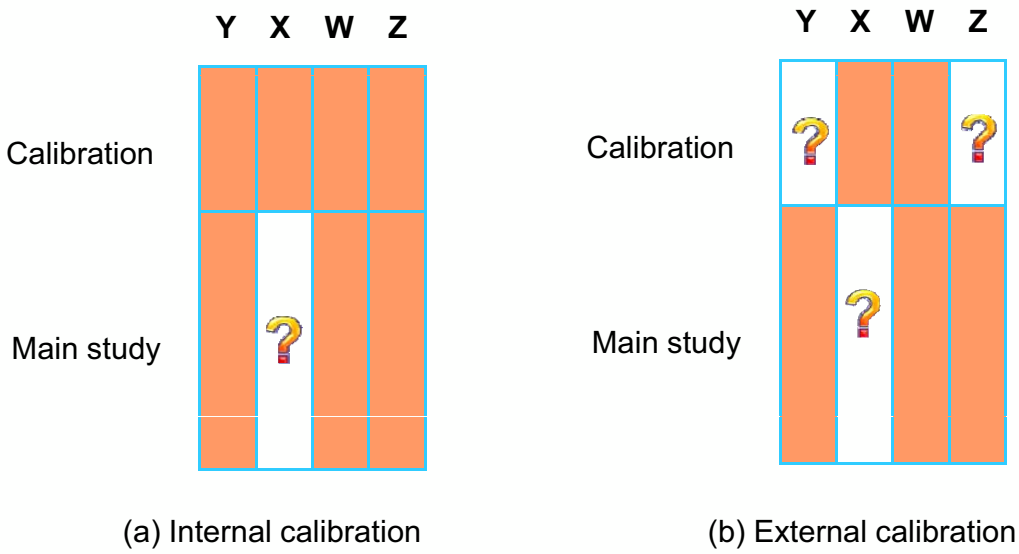

Figure 4.1: Internal and external calibration/main study design

ror adjustments has largely concerned the internal calibration design, with or without outcome Y. Regression calibration (RC) is one method for adjusting for measurement error using the internal calibration data. The method substitutes the estimated conditional expectation of the true biomarker given the observed surrogate and the other covariates into the primary regression model (Carroll and Stefanski, 1990). This method yields consistent estimates of the main regression parameters under the non-differential measurement error assumption that Y is independent of W given Z and X, which we denote NDME$(Y|Z, X)$. The standard errors of the estimates are calculated using either bootstrap or sandwich methods. Rosner et al. (1989, 1990) propose a related method, which they also call RC, and Thurston et al. (2003) demonstrate that this method yields identical estimated coefficients and asymptotic variances to those proposed by Carroll and Stefanski (1990), under fairly broad conditions. When values of Y are available in the calibration data, as shown in Figure

4.1(a), a refinement of RC is efficient regression calibration (ERC), which combines the RC estimates with direct estimates of the regression of Y on (X, Z) from the calibration sample (Spiegelman et al., 2001).

An alternative approach for dealing with measurement error problems with internal calibration data is multiple imputation (MI), a method developed to draw inferences from data sets with missing values (Little and Rubin, 2002). The true values of the biomarker X are imputed as draws from the conditional distribution of X given W ,Z and Y, estimated from the calibration subsample. This imputation step is repeated multiple times to create multiple completed data sets. Each completed data set is then analyzed using standard complete-data procedures, and estimates and standard errors from each complete-data analysis combined using MI combining rules given in Rubin (1987). Multiple imputation is increasingly available in widely-available statistical software (SAS PROC MI, IVEware, MICE), making the approach attractive practically. Cole et al. (2006) propose this approach with a survival outcome model when one covariate is measured with error. Messer and Natarajan (2008) present using MI for measurement error adjustment in logistic regression analysis. Raghunathan (2006) indicates self-reported values of health conditions collected in a large survey may be inaccurate, and proposes to multiply impute clinical (true) values of health conditions, based on the National Health and Nutrition Examination Survey which includes both self-reported values and clinical values from physical examinations. He and Zaslavsky (2009) show that cancer therapies for patients are likely underreported in registry systems, and propose to impute correct treatment status by using information from medical records collected in a calibration sample to improve analysis. Freedman et al. (2008) compare RC, ERC and multiple imputation under the multivariate normal model (MI-IC) for a single continuous

explanatory variable measured with error, for the internal calibration design. They show in simulations that RC is biased, and ERC can be more efficient than MI-IC because it exploits the NDME assumption. Guo and Little (2010) show that efficient versions of MI-IC are available that exploit the NDME assumption, and they extend MI-IC to handle the situation where the measurement error has nonconstant variance, a case that RC and ERC are not well equipped to handle.

Our focus in this chapter is on the external calibration design as shown in Figure 4.1(b). Since biomarkers are commonly calibrated by assay producers independently of the main study, this situation is much more common than that of the internal calibration, but methods for correctly adjusting for measurement error have not to our knowledge been developed. The classical CA method is biased as we already know from previous studies (in Chapter II), and there are two serious problems with RC, ERC and MI-IC in this external calibration setting. First, the calibration data may not be available to multiple researchers with various scientific purposes. Second, RC, ERC and MI-IC both require information in the calibration sample that is not available in the case of external calibration: for RC the values of Z, and for ERC and MI-IC the values of both Y and Z. If multiple imputation or regression calibration are based on the distribution of X given W, which can be estimated from the calibration data, they both yield biased inferences for the regression of Y on X and Z, as we demonstrate in simulations.

We propose a new MI method, multiple imputation for external calibration (MI-EC), which addresses these problems. First, it only requires summary statistics from the calibration sample. Second, it yields valid MI inferences for the regression of Y on X and Z, despite the fact that values of Y and Z are not measured in the calibration sample. Like MI-IC, it is based on a multivariate normal model, but it is not the

standard multivariate normal version of MI, as implemented in PROC MI in SAS; that method is actually not feasible for the missing-data pattern in Figure 4.1(b), since there is no information to estimate some imputation model parameters. Rather, estimates of model parameters for MI-EC can be obtained based on a multivariate normal model, and parameter restrictions under a NDME assumption that is very plausible in many settings. The algorithm for creating the MI-EC imputations, which is provided as an R package (details are given in Appendix 4.7), is remarkably simple, since it is a direct simulation method that does not require iterative Markov-chain Monte Carlo methods like the Gibbs sampler.

The remainder of this chapter is organized as follows. In Section 4.2, we describe the model that underlies MI-EC, and outline the algorithm for creating multiple imputations. A simulation study showing the superior performance of MI-EC over competing methods is reported in Section 4.3. Sensitivity analysis to examine the robustness of MI-EC is presented in Section 4.4. In Section 4.5, we analyze data from our motivating example. Discussion and future research is provided in Section 4.6.

4.2 Proposed Multiple Imputation Method

We write $U = (Y, Z)$, a vector of p variables, for the set of outcomes Y and covariates Z other than X, where Y has dimension q, Z has dimension r, and $p = q + r$. Since q and r may be greater than one, we allow for more than one dependent variable and/or covariate, so multivariate regression is included in our formulation. For simplicity we assume here that X and W are scalar, although our method can be extended to handle more than one variable X subject to measurement error.

We assume that in the main sample and the calibration sample, $(U, X|W)$ has

a joint $(p+1)$-variate normal distribution with a mean that is linear in W and a constant covariance matrix. Thus, the distribution of $(U, X|W)$ is assumed the same in the main study sample and the calibration sample, although we do not need to assume the same distribution of U in the two samples. This assumption is related to the "transportability across studies" assumption in Carroll et al. (2006). Further, we make the following non-differential measurement error assumption:

NDME(U): the distribution of U given W and X does not depend on W. (*)

This is assumption is stronger than the assumption NDME$(Y|Z, X)$ underlying RC for internal calibration, since a stronger assumption is needed, given the sparser information available from external calibration data. Nevertheless, it is reasonable if the measurement error is unrelated to values of $U = (Y, Z)$, as is plausible in many bioassays (Guolo and Brazzale, 2008). For MI, we need imputations of X from the conditional distribution of X given the observed variables in the main study sample, namely Y ,Z and W; let $\phi = (\lambda, \sigma_{x \cdot yzu})$ denote the vector of regression coefficients λ for the regression of X on (Y, Z, W), and corresponding residual variance for that regression. We describe two multiple imputation approaches, one of which is improper (Rubin, 1987) and based on the the maximum likelihood estimates $\widehat{\phi}$ of ϕ, and the other is proper and based on a draw $\phi^{(d)}$ from the Bayesian posterior distribution of ϕ, assuming a noninformative prior distribution of the parameters. In both cases, a set of D imputed data sets are created by filling in the missing values of X in the main study sample. For data set d, the improper method imputes the missing value x_i of X for the ith observation in the study sample by

$$\widehat{x}_i^{(d)} = E(x_i|y_i, z_i, w_i, \widehat{\phi}) + \varepsilon_i^{(d)} \tag{4.1}$$

where (y_i, z_i, w_i) are the values of (Y, Z, W) for observation i, $E(x_i|y_i, z_i, u_i, \widehat{\lambda})$ is the regression prediction with regression coefficients replaced by the ML estimate $\widehat{\lambda}$, and $\varepsilon_i^{(d)}$ is an independent normal deviate with mean zero and variance given by the ML estimate of the residual variance $\widehat{\sigma}_{x\cdot yzu}$ of the regression. The proper method imputes the missing value x_i by

$$\widehat{x}_i^{(d)} = E(x_i|y_i, z_i, w_i, \phi^{(d)}) + \varepsilon_i^{(d)} \tag{4.2}$$

where $\widehat{\phi}$ is replaced by a draw $\phi^{(d)}$ from the posterior distribution of ϕ. The proper method is better than the improper method because it takes into account uncertainty in estimating ϕ (Rubin, 1987), and it is actually not much harder computationally than the improper method. We now outline how $\widehat{\phi}$ and $\phi^{(d)}$ are computed, and why the algorithms work. Readers not interested in these statistical details can skip to the beginning of the next section.

The ML estimates $\widehat{\phi}$ are computed as follows:

Step (1) Let $\theta = (\theta_1, \theta_2, \sigma_{ux\cdot w})$, where θ_1 represents parameters of the normal distribution of X given W, θ_2 represents parameters of the normal distribution of U given W, and $\sigma_{ux\cdot w}$ represents the set of p partial covariances between U and X given W. Estimate θ_1 by $\widehat{\theta}_1$, the ML estimates based on the calibration sample on (X, W), and θ_2 by $\widehat{\theta}_2$, the ML estimates based on the main study sample on (U, W). These are the standard normal linear regression ML estimates for complete data, and the calculations involve standard least squares methods. Also note that $\widehat{\theta}_1$ can be computed from summary statistics on the calibration sample, namely the sample size, sample mean and sum of squares and cross products matrix of X and W.

Step (2) Estimate

$$\widehat{\sigma}_{ux\cdot w} = \widehat{\beta}_{uw\cdot w}\widehat{\sigma}_{xx\cdot w}/\widehat{\beta}_{xw\cdot w},$$

where $\widehat{\beta}_{uw\cdot w}$ is the $(p \times 1)$ vector of regression coefficients of U on W, estimated from the main sample, and $\widehat{\beta}_{xw\cdot w}$ and $\widehat{\sigma}_{xx\cdot w}$ is the regression coefficient of W and residual variance from regression of X on W, estimated from the calibration sample. This expression follows since, from properties of the multivariate normal distribution, $\beta_{uw\cdot w} - \sigma_{ux\cdot w}\beta_{xw\cdot w}/\sigma_{xx\cdot w}$ equals the set of regression coefficients of W in the regression of U on W and X, which are zero because of the NDME assumption (*).

Step (3) The ML estimates of the parameters of the distribution of $(U, X) = (Y, Z, X)$ given W are fully specified by the estimates in Steps (1) and (2). In fact, since the number of parameter restrictions from the NDME assumption, namely p, is the same as the number of parameters in $\sigma_{ux\cdot w}$ that are not estimable from the main and calibration samples – the model is technically "just identified". The parameter ϕ of the regression of X on Y, Z and W is a vector function of the parameters $(\theta_1, \theta_2, \sigma_{ux\cdot w})$, that is, $\phi = \phi(\theta_1, \theta_2, \sigma_{ux\cdot w})$. The ML estimate of ϕ is then $\widehat{\phi} = \phi(\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\sigma}_{ux\cdot w})$, obtained by substituting ML estimates of $(\theta_1, \theta_2, \sigma_{ux\cdot w})$ in this function. The details of this transformation are discussed in Little and Rubin (2002). Computation is straightforward using the SWEEP operator, which facilitates switching between parameters of different regressions derived from the multivariate normal distribution.

This completes the description of the ML algorithm, except for one minor caveat. The estimate of the residual variance of X given (Y, Z, W) could be negative, given the fact that estimates are being combined from two samples. If this happens, the residual variance should be set to zero. The resulting estimate of the covariance matrix lies in the parameter space. In particular, it is positive definite. This is always the case in our simulation studies, and in real applications this is likely to be the case unless the measurement error is very large.

As noted above, the imputations based on this procedure have the limitation that they do not reflect uncertainty in the ML estimates of ϕ. Fortunately, it is relatively easy to overcome this limitation, by replacing ML estimates $\widehat{\phi}$ of the parameters ϕ for the d^{th} imputed data set by a draw $\phi^{(d)}$ from the posterior distribution of ϕ. A noninformative Jeffrey's prior is assumed for the parameter (θ_1, θ_2). Then the ML estimates $(\widehat{\theta}_1, \widehat{\theta}_2)$ in Step (1) are replaced by draws $(\theta_1^{(d)}, \theta_2^{(d)})$ from their complete-data posterior distributions based on the calibration and main study samples, respectively. Draws from these posterior distributions are easily computed using chi-squared and normal deviates, as described in Little and Rubin (2002). Steps (2) and (3) are then as above, except that draws of $\sigma_{ux\cdot w}^{(d)}$, $\phi^{(d)}$ for $\sigma_{ux\cdot w}$ and ϕ are created using the draws $(\theta_1^{(d)}, \theta_2^{(d)})$ rather than $(\widehat{\theta}_1, \widehat{\theta}_2)$.

For biomarker data, since the calibration data are not a subset of the main study data, they are not included in the post-imputation analysis. Reiter (2008) shows that when the calibration data used to create the imputation models for imputing missing values of the covariate are not included for analysis, the usual multiple imputation variance estimator obtained from the MI combining rules outlined by Rubin (1987) is positively biased and confidence interval coverage exceeds 95%. He proposes a two-stage imputation procedure to generate imputations that enable unbiased estimation of variances. Here, we follow Reiter's procedure to impute unobserved true values of the covariate X. In detail, first, we draw the d^{th} values of model parameters $\phi^{(d)}$ by following Steps (1)-(3); second, for each $\phi^{(d)}$, $d = 1, ..., m$, we construct n imputed data sets by generating n sets of draws of X from the model (4.2). Finally, this procedure yields a collection of $m \times n$ imputed data sets, $D = D^{(d,i)} : d = 1, ..., m; i = 1, ..., n,$ which can be analyzed by standard complete data inference. The results from each imputed data set are then combined to obtain

valid inferences using the combining rules suggested by Reiter.

For $d = 1, ..., m$ and $i = 1, ..., n$, let $\hat{\gamma}^{(d,i)}$ and $var(\hat{\gamma}^{(d,i)})$ be the estimate of parameters of interest and the corresponding estimated variance computed with $D^{(d,i)}$ data set, respectively. The MI estimate of γ, $\hat{\gamma}_{MI}$, and associated variance T_{MI} are calculate as

$$\hat{\gamma}_{MI} = \sum_{d=1}^{m}\sum_{i=1}^{n}\hat{\gamma}^{(d,i)}/(mn) = \sum_{d=1}^{m}\overline{\gamma}_n^{(d)}/m$$

$$T_{MI} = U - W + (1 + 1/m)B - W/n$$

with

$$W = \sum_{d=1}^{m}\sum_{i=1}^{n}(\hat{\gamma}^{(d,i)} - \overline{\gamma}_n^{(d)})^2/(m(n-1))$$

$$B = \sum_{d=1}^{m}(\overline{\gamma}_n^{(d)} - \hat{\gamma}_{MI})^2/(m-1)$$

$$U = \sum_{d=1}^{m}\sum_{i=1}^{n}var(\hat{\gamma}^{(d,i)})/mn$$

The 95% confidence intervals for the MI estimate are calculated as $\hat{\gamma}_{MI} \pm t_{0.975,\upsilon}\sqrt{T_{MI}}$, with degree of freedom $\upsilon = [\frac{((1+1/m)B)^2}{(m-1)T_{MI}} + \frac{((1+1/n)W)^2}{(m(n-1))T_{MI}}]^{-1}$. When $T_{MI} < 0$, the variance estimator is recalculated as $(1 + 1/m)B$, and inferences are based on a t-distribution with $(m - 1)$ degrees of freedom.

In this study, we choose $m = 12$ and $n = 3$. When $n = 1$, the two-stage imputation is the normal multiple imputation method for handing missing data.

4.3 Simulation Study

Simulation studies are performed to investigate the performance of the proposed MI-EC method under an external/main study design with other existing methods,

including the naïve (ignoring measurement error), classical calibration (CA) and regression prediction (RP) methods.

4.3.1 Simulation Design and Setting

We consider a linear regression model for outcome Y on covariate X measured with error, and covariate Z measured without error,

$$f(Y|X, Z, \psi) \sim N(\gamma_0 + \gamma_X X + \gamma_Z Z, \tau^2) \tag{4.3}$$

where $\psi = (\gamma_0, \gamma_X, \gamma_Z, \tau^2)$ denotes the vector of regression coefficients and residual variance for the model. The parameters γ_X and γ_Z correspond to the covariate measured with and without error, respectively. We distinguish between different estimators of γ by a subscript. Our study aim is to estimate the parameters γ_X and γ_Z.

In the main study, the true value of X is not observed, instead we observe the surrogate measure W. The surrogate W is related to covariate X by a measurement error model. We consider a case of linear biased and non-differential measurement error process given as

$$f(W|Y, X, Z,) \sim N(\beta_0 + \beta_1 X, \sigma^2) \tag{4.4}$$

The performances of various methods are summarized using the absolute value of bias, the root mean square error (RMSE), and empirical non-coverage rate. The empirical non-coverage rate is calculated as the percentage of simulated data sets for which the 95% confidence intervals (CI) do not contain the true parameter values, γ_X, or γ_Z. The percentage is multiplied by 1000 to eliminate the decimal points, and hence a nominal value of non-coverage of 95% confidence interval is equal to 50.

To facilitate comparison of different simulated scenarios, the covariates X and Z are standardized to have mean 0 and variance 1. X and Z are constructed to be correlated with the correlation ρ. For simplicity, we fix $\beta_0 = 0$ and $\beta_1 = 1.1$, so that W is a linear biased surrogate for X. In the main study model, we also fix $\gamma_0 = 0$, $\gamma_Z = 0.4$, and $\tau^2 = 1$. We assess the performance of the estimator under different simulation scenarios by varying the remaining model parameters: γ_X, the regression coefficient of the covariate X, which is measured with error; σ^2, the variance of the measurement error model; and ρ, the correlation between X and Z. Specially, ρ is set to 0.3 for low correlation between X and Z, and 0.6 for high correlation. σ^2 is chosen to be 0.25, 0.5 and 0.75 to represent small, moderate and large measurement errors, respectively. γ_X is set to 0.4 and 1.2, which correspond to a small covariate effect and a large covariate effect respectively.

We generate 1000 simulations for each of the combinations of parameter values. For each of the simulations, a calibration sample and main study sample are generated. To generate the calibration study, n_{calib} observations of (X, Z) are first sampled from a bivariate normal distribution with zero mean, unit variance and correlation coefficient ρ, and then W is generated from the measurement error model as in Eq.4.4 conditioning on X, given the values of β_0, β_1, σ^2.

To generate the main study sample, we first sample n_{main} observations of (X, Z) from a bivariate normal distribution with zero mean, unit variance and correlation coefficient ρ. Next, W is generated from the model as in Eq.4.4 conditioning on X given values of β_0, β_1, σ^2, and Y is generated from the model as in Eq.4.3 conditioning on X and Z given γ_0, γ_X, γ_Z, τ^2, respectively.

Under each simulation setting, the sample size of the main study n_{main} is set to 400. The sample size of the calibration study n_{calib} is chosen as 100.

4.3.2 Existing Methods

- **Classical Calibration**

  The classical calibration approach (CA) is a widely used error-correction method in practice, especially when dealing with laboratory-related data. This method first fits a linear regression curve of W on X based on the calibration data, and then estimates the unknown value of X by $\hat{X}_{CA} = (W - \hat{\beta}_0)/\hat{\beta}_1$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of the intercept and slope obtained from the regression of W on X. The estimate $\hat{X}_{CA}$ is then substituted into the regression model (4.3) in place of the unknown X in the main study data, to yield the CA estimate of γ.

- **Regression Prediction**

  In this method, the unknown value of X in the main study data is substituted with $E[X|W]$. We first estimate the linear regression $E[X|W] = \alpha_0 + \alpha_1 W$ in the calibration study to get estimates $\hat{\alpha}_0$ and $\hat{\alpha}_1$, and then replace unknown values of X in the regression model (4.3) with the expectation of X given W, that is, $\hat{X}_{RP} = \hat{\alpha}_0 + \hat{\alpha}_1 W$, to estimate γ. Standard errors for the estimate of γ can be easily found by bootstrap methods.

  The RP method is known as the usual regression calibration method when there is no covaraite Z.

4.3.3 Results

The results of the simulation studies are shown in Table 4.1. We examine and compare the naïve regression of Y on W and Z, and various measurement error correction techniques, including aforementioned CA, RP, and MI-EC. We focus on the performance of various methods on inferences for the estimates of γ_X and γ_Z, with

respect to the bias, RMSE and empirical non-coverage of 95% confidence intervals.

We investigate the estimate for the inaccurately measured covariate X first. As expected, the naïve estimate γ_X obtained without adjustment for measurement error is attenuated toward the null value (zero) due to measurement error in X, and corresponding empirical non-coverage rate of 95% confidence intervals exceeds the nominal levels in all examined simulation scenarios. Similar to the naive method, the implementation of CA performs very poorly, with substantial bias and high non-coverage rate, particularly when the measurement error is large. RP has small bias when the correlation between X and Z is low. For high correlation, there is a large bias in RP, with the difference between RP and MI-EC increasing with large covariate effect and large measurement error. The poor results for the RP method may occur because imputing the missing X from the conditional distribution $X|W$ (ignoring Y and Z completely) has introduced extra noise. The non-coverage of confidence intervals for RP becomes substantial when the magnitude of measurement error and the correlation between X and Z increase. Under all simulation scenarios considered here, MI-EC has little or mild bias, and nominal level non-coverage of confidence intervals. The RMSE of MI-EC is comparable to or a little larger than that of RP in some situations. We believe that the loss of precision for MI-EC is because this method takes into account the correlation between X and Z. As demonstrated by the simulation study, under the same simulation settings, the modified version of MI-EC (ignoring the correlation between X and Z) is more efficient than RP. As the covariate effect and the $X - Z$ correlation become large, MI-EC has smaller RMSE than RP.

The estimate of the regression coefficient for Z, the covariate measured without error, is also shown in Table 4.1. The estimates of γ_Z obtained by the naïve method

Table 4.1: Empirical bias, RMSE, and non-coverage rate (*noncov.*) for the estimates of γ_X and γ_Z based on 1000 simulations. The calibration study sample size = 100 and the main study sample size = 400. The true value of γ_X is 0.4 or 1.2; the true value of γ_Z is 0.4. All values are multiplied by 1000.

Simulation parameters						X				Z			
γ_X	γ_Z	β	σ^2	ρ		Naïve	CA	RP	MI-EC	Naïve	CA	RP	MI-EC
0.4	0.4	1	0.25	0.3	Bias	105	75	7	1	23	23	23	0
					RMSE	113	90	60	61	57	57	57	54
					Noncov.	664	347	44	28	66	66	61	32
0.4	0.4	1	0.5	0.3	Bias	151	126	10	3	38	38	38	1
					RMSE	156	134	68	71	65	65	65	56
					Noncov.	961	784	48	42	111	111	117	41
0.4	0.4	1	0.75	0.3	Bias	185	163	13	5	49	49	49	3
					RMSE	189	169	75	80	72	72	72	59
					Noncov.	998	952	53	45	153	153	151	36
0.4	0.4	1	0.25	0.6	Bias	27	99	36	2	60	60	60	1
					RMSE	135	113	76	76	85	85	85	69
					Noncov.	710	416	83	36	159	159	165	50
0.4	0.4	1	0.5	0.6	Bias	181	158	56	9	95	95	95	6
					RMSE	186	166	92	95	112	112	112	79
					Noncov.	981	868	124	47	350	350	354	48
0.4	0.4	1	0.75	0.6	Bias	217	198	70	18	119	119	119	15
					RMSE	221	204	105	122	133	133	133	100
					Noncov.	1000	975	145	49	499	499	506	34
1.2	0.4	1	0.25	0.3	Bias	312	223	18	5	67	67	67	10
					RMSE	316	233	85	88	88	88	88	61
					Noncov.	1000	949	63	37	210	210	217	59
1.2	0.4	1	0.5	0.3	Bias	451	375	29	12	113	113	113	5
					RMSE	453	383	111	119	128	128	128	73
					Noncov.	1000	999	72	38	418	418	428	53
1.2	0.4	1	0.75	0.3	Bias	552	487	36	17	147	147	147	9
					RMSE	554	493	132	144	159	159	159	86
					Noncov.	1000	1000	77	34	621	621	621	53
1.2	0.4	1	0.25	0.6	Bias	377	294	104	9	177	177	177	6
					RMSE	381	303	137	111	188	188	188	85
					Noncov.	1000	988	248	37	766	766	760	47
1.2	0.4	1	0.5	0.6	Bias	539	472	166	25	284	284	284	20
					RMSE	541	478	197	166	291	291	291	124
					Noncov.	1000	1000	392	34	987	987	987	45
1.2	0.4	1	0.75	0.6	Bias	647	591	206	44	355	355	355	37
					RMSE	649	596	239	213	362	362	362	166
					Noncov.	1000	1000	466	37	999	999	1000	47

Naïve: naïve linear regression of Y on W; CA: classical calibration; RP: regression prediction; MI-EC: multiple imputation.

are considerately biased, with bias increasing with large measurement error, large covariate effect and high correlation between X and Z. The similar phenomenons are also observed for the CA and RP methods with large bias, poor precisian and high non-coverage rate. In contrast, our MI-EC method shows good performance in all scenarios used in our simulation study.

The results presented in Table 4.1 are based upon the two-stage imputation parameter setting $(m, n) = (12, 3)$. For a comprehensive evaluation, we also examine the performance of our MI-EC method under several other combinations of m and n settings, including $(20, 3)$ and $(12, 5)$. The results from the simulation study under those settings are close to those we present in Table 4.1, although the combination of $(20, 3)$ results in a slightly lower non-coverage rate.

We conclude that MI-EC is considerably superior to other existing methods for adjusting for covariate measurement error, mainly in eliminating measurement error bias and providing adequate coverage of confidence intervals. Its performance becomes more pronounced as the covariates X and Z are highly correlated, the covariate effect is large, or the measurement error is large.

4.4 Sensitivity Analysis to Multivariate Normal Assumption

As shown in previous results, the MI-EC method provides promising results to correct for covariate measurement error in regression analysis, for the situation where the calibration data is not directly available, and only summary statistics of the joint distribution of (X, W) can be collected and used. Recall that one key limitation of this method is that it is based upon certain model assumptions, including multivariate normality assumption and NDME assumption. The combination of these two assumptions allows us to impute X condition on all observed variables, and then

provide valid inference. The NDME assumption is common and reasonable, as presented and discussed in the measurement error literature (Spiegelman et al., 2001; Freedman et al., 2008; Guolo and Brazzale, 2008). Therefore, we focus on a sensitivity analysis to evaluate the robustness of the proposed method to the violation of the normality assumption in the following discussion.

We investigate the performance of our proposed method under two different covariate distribution misspecification: the case where the binary covariate Z is misspecified as normal, and the case where the log-normal covariate X is misspecified as normal.

4.4.1 Misspecification of Binary Covariate

In the simulation setting discussed in this subsection, we model Z as a binary covariate, which intentionally deviates our assumption that Z should be normal. We then apply our MI-EC method, which is originally designed for multivariate normal distribution, in this setting in order to examine and evaluate the robustness of the MI-EC method when its assumption does not hold.

For the evaluation completeness, we examine the performance of MI-EC under different simulation settings by varying the measurement error variance (σ^2), the correlation between X and Z (ρ), and the covariate effect of X (γ_X). The simulation parameter settings are similar to those presented in our previous simulation study as discussed in Section 4.3.

For simulation results present in Table 4.2, we first generate (X, Z^*) from a bivariate normal distribution with mean 0, variance 1, and correlation ρ. The binary variable Z is then generated being equal to 1 if $Z^* >= 0.8$, and equal to 0 otherwise; this setting results in an extreme (and rare) binary case where the probability $Pr.(Z = 1)$ equals to a low value 0.2. Besides the presented cut point 0.8, we have

Table 4.2: Sensitivity to multivariate normality assumption in the binary case. The table shows empirical bias, RMSE, and non-coverage rate (*noncov.*) for the estimates of the regression parameters (γ_X,γ_Z). All values are multiplied by 1000.

Simulation parameters						X				Z			
γ_X	γ_Z	β	σ^2	ρ		Naïve	CA	RP	MI-EC	Naïve	CA	RP	MI-EC
0.4	0.4	1	0.25	0.3	Bias	84	84	4	1	45	45	45	0
					RMSE	95	97	61	62	136	136	136	131
					Noncov.	439	452	49	31	64	64	65	30
0.4	0.4	1	0.5	0.3	Bias	139	138	5	3	74	74	74	2
					RMSE	145	146	70	73	149	149	149	137
					Noncov.	903	857	45	42	80	80	87	36
0.4	0.4	1	0.25	0.6	Bias	95	95	17	2	100	100	100	1
					RMSE	106	108	67	69	169	169	169	147
					Noncov.	487	504	59	43	109	109	111	38
0.4	0.4	1	0.5	0.6	Bias	153	153	27	5	161	161	161	7
					RMSE	159	160	77	81	210	210	211	160
					Noncov.	930	886	69	52	212	212	212	40
1.2	0.4	1	0.25	0.3	Bias	250	250	9	5	133	133	134	1
					RMSE	255	260	89	93	195	195	196	153
					Noncov.	996	976	60	39	145	145	148	71
1.2	0.4	1	0.5	0.3	Bias	414	413	13	12	220	220	220	7
					RMSE	417	420	120	125	267	267	267	182
					Noncov.	1000	1000	61	39	309	309	308	73
1.2	0.4	1	0.25	0.6	Bias	282	282	49	7	298	298	298	7
					RMSE	287	291	103	103	332	332	332	176
					Noncov.	998	986	100	37	499	499	509	52
1.2	0.4	1	0.5	0.6	Bias	457	456	77	18	480	480	480	25
					RMSE	460	462	139	141	504	504	504	234
					Noncov.	1000	1000	143	28	860	860	856	44

Naïve: naïve linear regression of Y on W; CA: classical calibration; RP: regression prediction; MI-EC: multiple imputation.

also examined several other cut points, such as 0.5, 0.6, and 0.7, and their results are analogous to those presented below. The surrogate W is generated from a simple unbiased measurement error model given as $W|X, Y, Z \sim N(X, \sigma^2)$. The outcome Y is generated to be related with covariates X and Z by a linear regression model, as given in model (4.3). To be consistent with previous setup, the sample size of the main study is chosen to be 400, and the sample size of the calibration study is chosen as 100. Under each simulation setting, the size of the simulated data sets is set to 1000.

Table 4.2 summarizes the results of the sensitivity analysis in the aforementioned binary case. We report the empirical bias, RMSE of the estimates for the regression parameters (γ_X, γ_Z), and the non-coverage rate of 95% confidence interval. The results presented in Table 4.2 clearly illustrate that in all the simulation settings the MI-EC method yields estimates with small empirical bias, and the non-coverage rate close to the 50 nominal level. From such promising performance, we can see that the MI-EC method is quite robust for the case of misspecified binary distribution.

4.4.2 Misspecification of Log-normal Covariate

The subsection follows a similar pattern of Section 4.4.1. We consider the simple case of unbiased measurement error process with $\beta_0 = 0$ and $\beta_1 = 1$. The ratios of measurement error variance, σ^2, to the variance of X is set to be equal to 0.25 and 0.5 to represent small and substantial measurement error, respectively. The outcome Y is generated from a linear regression model with regression parameters $\gamma_0 = 0$, $\gamma_0 = 0.4$, $\gamma_Z = 0.4$, and $\tau^2 = 1$. The true covariate X is generated from a log-normal distribution given as $X \sim LN(0, \phi^2)$, and Z is generated from a standard normal distribution with zero correlation between X and Z. We consider different degrees of skewness and heavy tails of the distribution of X by varying the parameter ϕ.

We set ϕ equal to 0.25, 0.5 and 1 to represent low, moderate and high skewness, respectively.

Table 4.3 presents results corresponding to the case of misspecification of log-normal covariate. As shown, unless in the scenario where the distribution of X is highly skewed and has very heavy tail (which mean that the actual distribution is extremely deviated from our normality assumption), the MI-EC method performs better than (or at least close to) the other methods in the aspects of reducing the measurement error bias and providing nominal coverage of confidence interval. This finding further consolidates our argument that MI-EC is a robust method.

4.5 Application to the Michigan Bone Health and Metabolism Study

The proposed method is illustrated in this section with a real-world data example from the Michigan Bone Health and Metabolism Study (MBHMS). One of the purposes of the MBHMS study is to assess the association between serum reproductive hormone concentrations and bone mineral density (BMD) loss in mid-life women. The primary explanatory variable we focus on in this application is regarding to sex hormone-binding globulin (SHBG) concentrations, which is the primary plasma transport protein for sex hormones. Due to a variety of reasons (e.g., assay imprecision), the measurements of the SHBG concentrations is found to contain substantial measurement error, which means that SHBG can be interpreted as X in the notation we defined before. In this analysis, main study data was collected in 81 white women, aged $44 - 64$ years, from the Michigan Bone Health Study cohort in 2008. In these data, the true SHBG concentration for each participant could not be directly observed, instead that an assay measure W was collected and used, which can be viewed as error-contaminated versions of true concentrations X. Besides the

Table 4.3: Sensitivity to multivariate normality assumption in the skew case. The table shows empirical bias, RMSE, and non-coverage rate (*noncov.*) for the estimates of the regression parameters (γ_X,γ_Z). All values are multiplied by 1000.

Simulation parameters			X				Z			
ϕ	*ratio*		Naïve	CA	RP	MI-EC	Naïve	CA	RP	MI-EC
0.25	0.25	Bias	82	82	0	0	0	0	0	0
		RMSE	188	190	218	214	49	49	50	49
		Noncov.	70	72	56	49	55	55	52	18
	0.5	Bias	135	135	3	2	0	0	0	0
		RMSE	204	207	239	238	49	49	50	49
		Noncov.	131	143	49	60	54	54	51	18
0.5	0.25	Bias	81	81	8	3	0	0	0	0
		RMSE	110	111	100	98	49	49	50	49
		Noncov.	187	200	51	48	52	52	58	19
	0.5	Bias	134	134	15	9	0	0	0	0
		RMSE	150	152	116	115	49	49	50	50
		Noncov.	509	527	43	58	52	52	57	22
1	0.25	Bias	86	86	35	25	1	1	1	1
		RMSE	92	93	81	78	52	52	52	53
		Noncov.	937	877	56	200	49	49	51	42
	0.5	Bias	141	140	69	45	1	1	1	1
		RMSE	145	146	131	126	54	54	54	59
		Noncov.	998	986	45	256	44	50	50	49

Naïve: naïve linear regression of Y on W; CA: classical calibration; RP: regression prediction; MI-EC: multiple imputation.

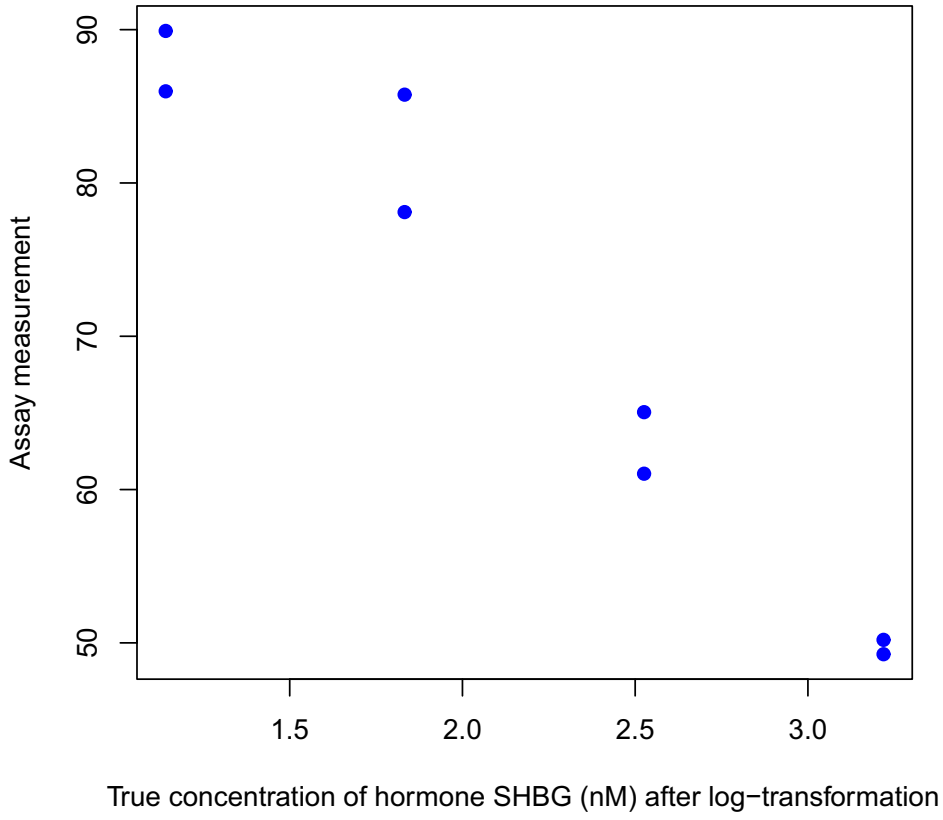

Figure 4.2: Calibration data of SHBG

measurement of W, several other covariates, which were measured at the same time when W was collected, may be related to the BMD too. In our study, we incorporate two of these variables in our model, including age and body mass index (BMI) variables, since they are potential risk factors of BMD as shown in other research work. We assume that age and BMI are measured without error. In our notation these variables are indicated by covariate Z. We are interested in examining the effect of SHBG concentrations on BMD after adjustment for age and BMI.

The calibration study of serum SHBG was constructed on the Bayer Diagnostic ACS-180 automated analyzer using the chemiluminescent technique (Bayer Corp.,

Table 4.4: Application to the MBHMS study. Parameter estimates in the linear regression of BMD on the logarithm of SHBG concentration, after adjustment for age and BMI.

	SHBG		
Methods	Est.	SE	P value
Naive	0.0052	0.0027	0.058
CA	-0.0995	0.0518	0.058
RP1	-0.1085	0.0610	0.086
RP2	-0.1054	0.0549	0.059
MI-EC	0.1062	0.0583	0.072

Norwood, MA) (Randolph et al., 2003). The SHBG assay is a competitive chemiluminescent immunoassay. This assay was developed *de novo* and used a commercially available rabbit anti-SHBG antibody, SHBG labeled with dimethylacridinium ester (DMAE), and a solid phase of goat anti-rabbit Immunoglobulin G (IgG) conjugated to paramagnetic particles (PMP). $40\mu L$ of serum was required for the assay in addition to sufficient dead volume for aspiration and repeat. The expected values were from 20 to $130nM$. The reporting range for the SHBG assay was 10 to $150nM$ (actual assay range: 1.95 to $250nM$). The assay was standardized against SHBG obtained from Wein Industries (Succasunna, NJ). The inter-assay and intra-assay coefficients of variation for SHBG were 17.9% and 9.3%, respectively. The calibration study recorded the true concentrations of SHBG for four samples, and corresponding assay measures. Figure 4.2 shows the scatter plot of the calibration data for SHBG. It is clear that the assay measures are with noise, suggesting that the measurement error exists.

In our analysis, we estimate the linear regression of BMD on the logarithm of SHBG concentration, age and BMI. We compare five different methods: the "naïve" analysis (where SHBG concentrations are represented by assay measures), and the

other four error correction methods including CA, RP1 (with adjusted standard error by use of the bootstrap method), RP2 (with "naïve" standard error without using the bootstrap method), and MI-EC. Table 4.4 presents the estimates, associated standard errors for the regression coefficient of SHBG, as well as corresponding P values. The naïve analysis provides an estimate of the influence of SHBG on BMD equal to 0.0052 (SE = 0.0027), controlling for age and BMI. This estimate indicates a positive association between SHBG on BMD. This association is marginally statistically significant. After adjustment for measurement error by various correction methods, we observe a negative association between SHBG and BMD. CA, RP1, RP2 and MI-EC result in estimates of the regression coefficient of SHBG equal to $-0.0995(0.0518)$, $-0.1085(0.0610)$, $-0.1054(0.0549)$, and $-0.1062(0.0583)$, respectively. These estimates indicates an even stronger association between SHBG and BMD than the naïve estimates. For example, the proposed MI-EC method estimate is approximately twenty times as great as the one provided by the naïve method. Recall that MI-EC is our proposed method and CA is commonly used in data analysis in epidemiology; we thus focus on the comparison between the CA method and MI-EC method in the following discussion. As shown in the table, the CA method provides an estimate approximately 7% smaller than that of MI-EC, and indicates a marginal significant effect of SHBG on BMD. In contrast, the MI-EC method estimate has standard error that is almost 1.3 times larger than that of the CA method, and consequently marginal statistical significance is no longer observed in MI-EC. Finally, in this MBHMS study, we do not observe significant difference between the performance of RP2 and MI-EC, though it may be worth pointing out that the RP1 estimate has smaller standard error than the RP2 estimate, which is not surprised since RP2 takes into account the uncertainty due to measurement error.

4.6 Conclusion and Discussion

In this chapter, we propose a novel multiple imputation method to correct for covariate measurement error in regression analysis, when the calibration data only provide information about X and W. As demonstrated in simulation studies, the MI-EC method performs well with small bias and accurate coverage.

Compared to other existing methods, our proposed method has some additional advantages. First, it is a simple and fast method, and doesn't need any iterative procedure. Second, MI-EC reduces the requirement for the internal calibration data, where information of (Y, X, W, Z) is available, or the external calibration data, where information of (X, W) is available. It requires only some simple summary statistics from the calibration sample to create the multiple imputations, hence its viability as a useful alternative to other simple methods such as RC has improved.

Although we make the multivariate normality assumption to allow conditional distributions characterized by linear regression relationships, the sensitivity analysis shows that our method is quite robust to the model misspecifications. We assume here that X and W are scalar; in the future we are interested to extend the proposed method to handle more than one covariate subject to measurement error.

4.7 Appendix

This section presents the R source Code to the MIEC algorithm introduced in Section 4.2.

```
#Calculate the summary statistics\\
calsummstat <-function(inputdata){

    x=inputdata[,1]
    w=inputdata[,2]
```

```
    n=nrow(inputdata)

    xbar=mean(x)
    wbar=mean(w)
    xx=sum(x^2)/n
    ww=sum(w^2)/n
    xy=sum(x*w)/n

    sxx=xx-xbar*xbar
    sww=ww-wbar*wbar
    sxy=xy-xbar*wbar

    b0=xbar-(sxy/sww)*wbar
    b1=sxy/sww
    sigmasq=sxx-sxy*sxy/sww

    param=c(b0,b1,sigmasq,sxx,sww,sxy,xbar,wbar)
    return(param)

}


#Draw parameters from their predictive distribution based on the
calibration data \\
generateRandomdDrawofMEMParam <- function(calibdata,n) {

    ndraw=1

    #Calculate the summary statistics of X and W
    iniparam=calsummstat(calibdata)

    betahat0=iniparam[1]
    betahat1=iniparam[2]
    rss=n*iniparam[3]
    sxx=iniparam[4]

    sigmasq=rss/rchisq(ndraw,(n-2))
```

```
    tmp0=rnorm(1)
    beta0=betahat0 + sqrt(sigmasq/n)*tmp0

    tmp1=rnorm(1)
    beta1=betahat1 + sqrt(sigmasq/(n*sxx))*tmp1

    param=c(beta0, beta1, sigmasq)
    return(param)
}

#Draw parameters from their predictive distribution based on the
main study data. \\
generateRandomMultivarRegreParam <- function(maindata,n,p) {

    ndraw=1

    w=maindata[,1]

    wmat=mat.or.vec(n,2)
    wmat[,1]=1
    wmat[,2]=w

    umat=maindata[,2:(p+1)]

    ww=solve(t(wmat)%*%wmat)
    coeffhat=ww%*%(t(wmat)%*%umat)
    residual=umat-wmat%*%coeffhat
    rss=t(residual)%*%residual
    a=t(chol(ww))
    invrss=solve(rss) #calculate inverse covariance matrix

    #Draw covariance of U given W from inverse
    #wishart distribution with df = (n-(k+p)+1)
    #where k is dimension of wmat and p is dimension of umat
    df=n-(p+2)+1
    covmatrix=riwish(df, rss)
    covmatrix=as.matrix(covmatrix)
```

```
    betavar= kronecker(covmatrix, ww)
    vec.coeffhat=as.vector(coeffhat)
    beta = mvrnorm(ndraw, vec.coeffhat, betavar)

    #regression coefficients of U on W
    multivarRegrecoeff=matrix(beta,2,p)

    #residual covariance matrix of U given W
    multivarResidcov=covmatrix

    param=rbind(multivarRegrecoeff,multivarResidcov)
    return(param)
}

#generate parameter of the distribution of surrogate W from the
posterior distribution \\
generateErrorvarParam <- function(calibdata,maindata,NCALIB,NSAMPLE)
{

    #combine the information about W from calibration data
    #and main study data
    w=c(calibdata[,2],maindata[,1])

    n=NCALIB+NSAMPLE

    ndraw = 1

    #Draw sigmaxsq from inverse chi-square distribution
    #with df = (n-2)
    muw = mean(w)
    rss = sum((w-muw)^2)
    sigmaxsq = rss/rchisq(ndraw,(n-1))

    #Draw mux from normal distribution
    tmp = rnorm(1)
    mu = muw + sqrt(sigmaxsq/n)*tmp
```

```
    param = c(mu, sigmaxsq) # two parameters: mean and variance

    return(param)
}

#create the initinal mean and covariance matrix \\
createMatrixonW <- function(memParam, dmParam,wParam, P){

    m = P+2+1
    matrixonW=matrix(NA, nrow = m, ncol = m)

    matrixonW[1,1] = -(1 + wParam[1]^2/wParam[2])
    matrixonW[1,2] = wParam[1]/wParam[2]
    matrixonW[2,2] = - 1/wParam[2]

    matrixonW[1:(m-1),3:(m-1)]=dmParam

    matrixonW[1:2,m] = memParam[1:2]
    matrixonW[m,m] = memParam[3]

    matrixonW[3:(m-1),m] = dmParam[2,] * memParam[3] / memParam[2]

    return(matrixonW)
}

#complete initinal mean and covariance matrix \\
completeGmatrix <- function(G){

    size = nrow(G)

    for (i in 1:size){
        for (j in 1:size){
            if ( is.na(G[i,j]) ) {
                if ( is.na(G[j,i]) == FALSE ) {
                    G[i,j] = G[j,i]
                } else {
                    print(sprintf("ERROR: both elements at [%d,%d]
                    and [%d,%d] are null", i, j, j, i));
```

```
                }
            }
        }
    }

    if (isSymmetric(G)) {
        return(G)
    } else {
        return(NULL)
    }
}

#check the symmetry of the matrix created by the sweep operator
isSymmetric <- function(G){

    size = nrow(G)

    for (i in 1:size){
        for (j in 1:size){
            if ( abs(G[i, j] - G[j, i]) > 1e-10) {
                print(sprintf("ERROR: elements not matched at [%d,%d]
                and [%d,%d]", i, j, j, i));
                return(FALSE);
            }
        }
    }
    return(TRUE);
}

#perform the sweep operator \\
sweep <- function(matrixonW){

    size = nrow(matrixonW)

    curH = completeGmatrix(matrixonW)
    newH= matrix(NA, nrow=size, ncol=size)

    for (k in 3:(size-1)){
```

```
        for (i in 1:size){
            for (j in 1:size){

                if (i==k && j==k)   {
                    newH[i,j] = -1/curH[k,k]
                } else if (i==k || j==k) {
                    newH[i,j] = curH[i,j]/curH[k,k]
                } else {
                    newH[i,j] = curH[i,j]
                      - curH[i,k]*curH[k,j]/curH[k,k]
                }
            }
        }

        curH = newH;
        newH= matrix(NA, nrow=size, ncol=size)
    }

    param=curH[,size]
    return(param)
}

#Create imputed values for unobserved covariate X
generateMissingvalue <- function(param, maindata){

    nsample=nrow(maindata)

    nparam=length(param)

    #test whether the estimate of the residual variance is negative,
    #and print a warning message
    if (param[nparam] < 0) {

        print(sprintf("%s", "Warning: The estimate of the
        residual variance of mismeaured covariate given
        the observed data is negative."))
        }
```

```
    mux = param[1]
    for (i in 2:(nparam-1)){

        mux = mux + param[i]*maindata[,i-1]

    }

    #Generate X from its posterior distribution with mean mux
    #and variance sigmasqxx_wzy
    imputedX = mux + sqrt(param[nparam])*rnorm(nsample, mean=0, sd=1)

    return(imputedX)
}


#title function \\
printTitle <-function(){

    print(sprintf("%s", "##########################################"))
    print(sprintf("%s", "## Loading required package: MIEC  ####"))
    print(sprintf("%s", "##########################################"))
}


#main function \\
MIEC <- function(maindata,calibdata,NCALIB,NSAMPLE,M,N,K,S) {

    printTitle()

    MIimputedXbasedoncalib=c()
    multipleImputedX=c()
    twostageMIimputedX=c()

    P=K+S

    count = 0
```

```
while (count < M) {

    #Step-1: draw parameters by regressing X on W based on
    #measurement error model
    memParam=generateRandomdDrawofMEMParam(calibdata,NCALIB)

    #Step-2: draw parameters by regressing (Y, Z) on W based on main
    #interested "disease" model
    dmParam=generateRandomMultivarRegreParam(maindata,NSAMPLE,P)

    #Step-3: generate draws of mean and variance of W
    wParam=generateErrorvarParam(calibdata,maindata,NCALIB,NSAMPLE)

    #Step-4: creating sweeping matrix on W by using parameters
    #obtained from step 1-3 and filling estimated covariance
    #parameter between U and X given W
    sweepMatrixonW=createMatrixonW(memParam, dmParam,wParam, P)

    #Step-5: calculate parameters of the imputation model for X
    #given (W,Y,Z) by the sweep operator
    impmodelParam = sweep(sweepMatrixonW)

    #Step-6: generate random draw for unknown X from its posterior
    #distribution, given W and Y
    varIndicator=length(impmodelParam)

    if (impmodelParam[varIndicator] >= 0){

        for (n in 1:N){

            secondStageDrawX = generateMissingvalue(impmodelParam,
            maindata)

            twostageMIimputedX=cbind(twostageMIimputedX,
            secondStageDrawX)

        }
        count=count+1
```

```
        }
    }

    #Output data with Y, Z, and multiply imputed X,
    #where maindata[,P+1] = U(Y,Z)
    twoStageMIimputeddata=cbind(maindata[,2:(P+1)], twostageMIimputedX)
    return(twoStageMIimputeddata)
}
```

CHAPTER V

Conclusion

In many epidemiological and clinical applications, accurate measurment is expensive or even impossible. Reasons include limitations of measurement instruments, biological variation, and recall bias. Consequently, observations are measured with error. Statistical models that fail to account for the measurement error yield distorted conclusions.

The work in this dissertation develops and evaluates Bayesian multiple imputation approaches for adjusting for measurement error based on information from a calibration sample. This concluding chapter summaries the main contributions of this work and describes promising future research directions.

5.1 Summary of Contributions

This work consider two important aspects of measurement error: the limit of detection and covariate measurement error.

5.1.1 Detection Limits

In Chapter II, the correction for measurement error below the limit of detection for the biomarker data is discussed. We develop a Bayesian measurement error model that yields prediction intervals for the true assay value throughout the range

of analyte values, and allows for heteroscedasticity of the measurement error. We illustrate the Bayesian model on calibration data for fat-soluble vitamins, focusing particularly on Beta Cryptoxanthin. The results confirm our hypothesis that prediction intervals for values above the LOQ are wide, and the width increases with the measured value; prediction intervals below the LOQ provide more information than the statement that the value is less than the LOQ. That is to say, our findings imply that the current approach to transmitting data from calibration assays is flawed, since it provides a distorted picture of the actual measurement error. Moreover, our proposed Bayesian MI method provides a general and fundamental framework, which can be extended to other generalized linear models without much difficulty.

5.1.2 Covariate Measurement Error

In Chapters III and IV, we develop new multiple imputation methods to deal with two aspects of covariate measurement error problems.

- **Heteroscedastic Measurement Error Correction**

  Chapter III presents our correction methods for heteroscedastic covariate measurement error in a linear regression analysis. This work develops an extended version of regression calibration, termed 'weighted regression calibration', and a novel multiple imputation approach computed using Bayesian Markov Chain Monte Carlo (MCMC) algorithms. The proposed methods are compared, in a simulation study, with the naive method that ignores measurement error, and several other existing error-correction methods, namely conventional approach, regression calibration, and efficient regression calibration.

  Simulation study and real-data analysis have shown that our first proposed method (weighted regression calibration) performs at least as well as existing

approaches in a wide variety of parameter settings, and significantly outperform others in the cases where the measurement error is substantial or the response-covariate association is strong. More importantly, the simulation studies present evidence that our second proposed approach (MI method) is superior to all other methods, with respect to empirical bias and precision of estimates of regression coefficients, and yields confidence intervals with close to nominal coverage.

- **Covariate Measurement Error Correction Based on Summary Statistics from External Calibration Data**

  In Chapter IV, we also consider the situation where interest concerns regression of outcomes on a variable subject to measurement error and other covariates, and information about measurement error is provided in the form of summary statistics from a bivariate calibration sample. We develop and evaluate a new correction method, named 'Multiple Imputation for External Calibration (MI-EC)', to tackle such situations, which yields valid inferences for the regression model parameters, using multiple imputation combining rules proposed by Reiter.

  The novelty of our approaches comes from two aspects. First, it is not built upon costly iterative sampling procedures like the MCMC techniques. Second, it only relies on summary statistics from the calibration data. The simulation evaluation has demonstrated that the MI-EC method works better than other existing methods to account for measurement error, in terms of correction for bias and achieving nominal confidence levels.

  The method is based on normal assumptions, and hence robustness to lack of normality is also assessed. More interestingly, the sensitivity study shows that

MI-EC still performs relatively well even when the multivariate normality assumptions upon which it is derived do not hold. We will analyze the underlying implication in the future.

5.2 Future Work

Although this dissertation presents a suite of strategies to improve statistical inference by taking into account measurement error in regression models, there is still much interesting work left in the research areas of this dissertation. Several promising future research directions are outlined below.

5.2.1 Extensions of Bayesian MI Method

For heteroscedastic measurement error settings in Chapter III, the attention has been restricted to the case where a linear regression of the outcome variable Y on the covariate X is of interest. Extensions to non-normal outcomes, such as when Y is binary and follows a probit model, could also be developed without too much difficulty. We will make further investigation in this field in the future.

Also, the Bayesian MI methods developed in Chapter III require specification of prior distributions for the unobserved true covariate, $p(X)$. A simple normal prior distribution for $p(X)$ is considered in this work. However, other choices may be worth considering too. As an example, the latent covariate distribution belongs to a flexible class of continuous distributions, like mixtures of the normal distributions. It may be possible to adapt it into our heteroscedastic measurement error setting, to achieve robustness of the inference of regression parameters against the model misspecification.

5.2.2 Extensions to Complex Covariate Data Structure

The work in Chapter IV has proposed a MI-EC method to deal with covariate measurement error for multivariate normal data, particularly for the case where the detailed calibration data is not directly available (e.g., due to information security reason), but only summary statistics of the joint distribution of (X, W) can be provided. Although the approach designed in this work is fairly robust to the misspeciation of model distributions, in the future, it is worth finding an exact solution for a regression model with mixtures of continuous and ordinal variables, where continuous variables are measured with error.

5.2.3 Further Extensions of MI Methods

A simplifying assumption made in this work is that there is only one covariate measured with error in the analysis model. It is worthwhile to take into account the situations where there may be more than one error-prone covariate in the subsequent research.

Also, in the future, we will study the problems with regard to the measurement error where a variable is measured multiple times and may be subject to measurement error. For example, in a longitudinal study, some risk factors are measured at each visiting time, and they are likely measured with error. Not surprisingly, it is a much more difficult task when incorporating repeated measure of covariate subject to measurement errors in the analysis. This requires a more complex model, which accounts for the distribution of the unobserved covariate, the correlation structure of the unobserved covariate and, of course, measurement error. We will systematically investigate this extension in the future.

5.3 Closing Remarks

While there are still many open questions (some of which were discussed in the previous section), this dissertation showed that the multiple imputation is an effective way for accounting for measurement error in regression analysis, and helped pave the way for future work on measurement error problems.

Bibliography

Armbruster, D. A., Tillman, M. D., and Hubbs, L. M. (1994). Limit of detection (lqd)/limit of quantitation (loq): comparison of the empirical and the statistical methods exemplified with gc-ms assays of abused drugs. *Clinical Chemistry* **40,** 1233–1238.

Belanger, B., Davidian, M., and Giltinan, D. (1996). The effect of variance function estimation in non-linear calibration inference in immunoassay. *Statistics in Medicine* **52,** 158–175.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)* **48,** 259–302.

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

Browne, R. W., Ocque, A. J., Ehrenstein, O., and Whitcomb, B. (2010). Practical experimental and statistical procedures for determination of detection limits: Application to high performance liquid chromatography analysis of fat soluble vitamins in human serum. *Epidemiology.* In press.

Carroll, R. J., Freedman, L. S., Kipnis, V., and Li, L. (1998). A new class of measuement-error models with applications to diery data. *The Canadian Journal of Statistics* **26,** 467–477.

Carroll, R. J., Kuchenhoff, H., Lombard, F., and Stefanski, L. A. (1996). Asymptotics for the simex estimator in structural measurement error models. *Journal of the American Statistical Association* **91,** 242–250.

Carroll, R. J., Maca, J. D., and Ruppert, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika* **86,** 541–554.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* Chapman and Hall, New York, NY.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall/CRC, Boca Raton, FL. Second Edition.

Carroll, R. J. and Spiegelman, C. H. (1986). The effect of ignoring small measurement errors in precision instrument calibraion. *Journal of Quality Technology* **18,** 170–173.

Carroll, R. J. and Stefanski, L. A. (1990). Approximate quasi-likelihood estimation in problems with surrogate predictors. *Journal of the American Statistical Association* **85,** 652–663.

Clayton, D. G. (1992). *Models for the analysis of cohort and case-control studies with inaccurately measured exposures In Statistical Models for Longitudinal Studies of Health*. Oxford University Press, Cary, NC. Dwyer, JH and Feinleib, M and Lippert, P and Hoffmeister, H (eds).

Cole, S. R., Chu, H., and Greenland, S. (2006). Multiple-imputation for measurement-error correction. *Statistics in Medicine* **35,** 1074–1081.

Cook, J. and Stefanski, L. A. (1994). A simulation extrapolation method for parametric measurement error models. *Journal of American Statistical Association* **89,** 1314–1328.

Cooper, G. S., Savitz, D. A., Millikan, R., and Kit, T. C. (2002). Organochlorine exposure and age at natural menopause. *Epidemiology* **13,** 729–733.

Cotton, S. M., Crewther, D. P., and Crewther, S. G. (2005). Measurement error: Implications for diagnosis and discrepancy models of developmental dyslexia. *Dyslexia* **11,** 186–202.

Currie, L. A. (1968). Limits for qualitative detection and quantitative determination–application to radiochemistry. *Analytical Chemistry* **40,** 586–593.

Dafni, U. G. and Tsiatis, A. (1998). Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics* **54,** 1445–1462.

Dahm, P. F., Gail, M. H., and Rosenberg, P. S. (1995). Determining the value of additional surrogate exposure data for improving the estimate of an odds ratio. *Statistics in Medicine* **14,** 2581–2598.

Dellaportas, P. and Stephens, D. A. (1995). Bayesian analysis of errors-in-variables regression models. *Biometrics* **51,** 1085–1095.

Dunsmore, I. R. (1968). A bayesian approach to calibration. *Journal of the Royal Statistical Society. Serie B (Methodological)* **30,** 396–405.

Ferrari, P., Carroll, R. J., Gustafson, P., and Riboli, E. (2008). A bayesian multilevel model for estimating the diet/disease relationship in a multicenter study with exposures measured with error: The epic study. *Statistics in Medicine* **57,** 6037–6054.

Finney, D. J. (1976). Radioligand assay. *Biometrics* **32,** 721–740.

Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods*. Sage, New York, NY.

Freedman, L. S., Fainberg, V., Kipnis, V., Midthune, D., and Carroll, R. J. (2004). A new method for dealing with measurement error in explanatory variables of regression models. *Biometrika* **60,** 172–181.

Freedman, L. S., Midthune, D., Carroll, R., and Kipnis, V. (2008). A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. *Statistics in Medicine* **27,** 5195–5216.

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

Fung, K. Y. and Krewsk, D. (1999). On measurement error adjustment methods in poisson regression. *Environmetrics* **10,** 213–224.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL. Second Edition.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* **7,** 457–472.

Giltinan, D. M. and Davidian, M. (1994). Assays for recombinant proteins: a problem in non-linear calibration. *Statistics in Medicine* **13,** 1165–1179.

Guo, Y., Harel, O., and Little, R. J. (2010). How well quantified is the limit of quantification? *Epidemiology* **21,** S1–S7.

Guo, Y. and Little, R. J. (2010). Regression analysis on the covariate with heteroscedastic measurement error. *Statistics in Medicine.* Accept, under the revision.

Guolo, A. and Brazzale, A. R. (2008). A simulation-based comparison of techniques to correct for measurement error in matched case-control studies. *Statistics in Medicine* **27,** 3755–3775.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika* **57,** 97–109.

He, Y. and Zaslavsky, A. M. (2009). Combining infomation from cancer registry and medical records data to improve analyses of adjuvant cancer therapies. *Biometrics* **65,** 946–952.

Hebert, J. R., Ebbeling, C. B., Matthews, C. E., Hurley, T. G., Ma, Y., Druker, S., and Clemow, L. (2002). Systematic errors in middle-aged women's estimates of energy intake: Comparing three self-report measures to total energy expenditure from doubly labeled water. *Annals in Epidemiology* **12,** 577–586.

Heid, I. M., Kuchenhoff, H., Miles, J., Kreienbrock, L., and E, W. H. (2004). Two dimensions of measurement error: Classical and berkson error in residential radon exposure assessment. *Journal of Exposure Analysis and Environmental Epidemiology* **14,** 365–377.

Higgins, K. M., Davidian, M., Chew, G., and Burge, H. (1998). The effect of serial dilution error on calibration inference in immunoassay. *Biometrics* **54,** 19–32.

Hopke, P. K., Liu, C., and Rubin, D. B. (2001). Multiple imputation for multivariate data with missing and below-threshold measurements: Time-series concentrations of pollutants in the arctic. *Biometrics* **57,** 22–33.

Hossain, S. and Gustafson, P. (2009). Bayesian adjustment for covariate measurement errors: A flexible parametric approach. *Statistics in Medicine* **28,** 1580–1600.

Hossian, S. and Gustafson, P. (2009). Bayesian adjustment for covariate measurement errors; a flexible parametric approach. *Statistics in Medicine* **28,** 1580–1600.

Hyslop, D. R. and Imbens, G. W. (2001). Bias from classical and other forms of measurement error. *Journal of Business and Economic Statistics* **19,** 475–481.

Jurek, A., Maldonado, G., Church, T., and Greenland, S. (2004). Exposure-measurement error is frequently ignored when interpreting epidemiologic study results. *European Journal of Epidemiology* **159,** S71–S76.

Kannel, W. B., Wentworth, J. D., Thomas, H. E., Stamler, J., Hulley, S. B., and Hulley, S. B. (1989). Overall and coronary heart disease mortality rates in relation to major risk factors in 325,348 men screened for mrfit. *Ametican Heart Journal* **112,** 825–836.

Kipnis, V., Subar, A. F., Midthune, D., and S, F. L. (2003). Structure of dietary measurement error: Results of the open biomarker study. *American Journal of Epidemiology* **158,** 14–21.

Ko, H. and Davidian, M. (2000). Correcting for measurement error in individual-level covariates in nonlinear mixed effects models. *Biometrics* **56,** 368–375.

Kuha, J. and Temple, J. (2003). Covariate measurement error in quadratic regression. *International Statistical Review* **71,** 131–150.

Kulathinal, S. B., Kuulasmaa, K., and Gasbarra, D. (2002). Estimation of an errors-in-variables regression model when the variances of the measurement errors vary between the observations. *Statistics in Medicine* **21,** 1089–1101.

Lin, X. and Carroll, R. J. (1999). Simex variance component tests in generalized linear mixed measurement error models. *Biometrics* **55,** 613–619.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley, Hoboken, NJ. Second Edition.

Lubin, J. H., Colt, J. S., Camann, D., Davis, S., Cerhan, J. R., Severson, R. K., Bernstein, L., and Hartge, P. (2004). Epidemiologic evaluation of measurement data in the presence of detection limits. *Environ Health Perspect* **112,** 1691–1696.

Lyles, R. H., Muno, A., Xu, J., Taylor, J. M. G., and Chmiel, J. S. (1999). Adjusting for measurement error to assess health effects of variability in biomarkers. *Statistics in Medicine* **18,** 1069–1086.

Meng, X. L. and Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* **79,** 103–111.

Messer, K. and Natarajan, L. (2008). Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment. *Statistics in Medicine* **57,** 6332–6350.

Morgan, T. and Elashoff, R. M. (1987). Effect of covariate measurement error in randomized clinical trials. *Statistical Science* **6,** 31–41.

Newman, M. C., Dixon, P. M., Looney, B. B., and Pinder, J. E. (2007). Estimating mean and variance for environmental samples with below delection limit observations. *Journal of the American Water Resources Association* **25,** 905–916.

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

Racine-Poon, A. (1988). A bayesian approach to nonlinear calibration problem. *Journal of the American Statistical Association* **83,** 650–656.

Racine-Poon, A., Weihs, C., and Smith, A. F. M. (1991). Estimation of relative potency with sequential dilution errors in radioimmunoassay. *Biometrics* **47,** 1235–1246.

Raghunathan, T. E. (2006). Combining information from multiple surveys for assessing health disparities. *Allgemeines Statistisches Archiv.* **90,** 515–526.

Randolph, J. F., Sowers, M., Gold, E. B., Mohr, B. A., Luborsky, J., Santoro, N., McConnell, D. S., Finkelstein, J. S., Korenman, S. G., Matthews, K. A., Sternfeld, B., and Lasley, B. L. (2003). Reproductive hormones in the early menopausal transition: Relationship to ethnicity, body size, and menopausal status. *Journal of Clinical Endocrinology & Metabolism* **88,** 1516–1522.

Reiter, J. P. (2008). Multiple imputation when records used for imputation are not used or disseminated for analysis. *Biometrika* **95,** 933–946.

Richardson, D. B. and Ciampi, A. (2003). Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *American Journal of Epidemiology* **157,** 355–363.

Richardson, S. and Gilks, W. R. (1993). Conditional independence models for epidemiological studies with covariate measurement error. *Statistics in Medicine* **12,** 1703–1722.

Richardson, S. and Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistics Society. Series B (Methodological)* **59,** 731–792.

Richardson, S., Leblond, L., Jaussent, I., and Green, P. J. (1999). Mixture models in measurement error problems, with reference to epidemiological studies. *Journal of the Royal Statistical Society. Series A* **165,** 549–566.

Rodbard, D. and Frazier, G. R. (1975). Statistical analysis of radioligand assay data. *Methods of Enzymology* **37,** 839–841.

Rosner, B., Spiegelman, D., and Willett, W. C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariate measured with error. *Statistics in Medicine* **132,** 734–745.

Rosner, B., Willett, W. C., and Spiegelman, D. (1989). Correction of logistic relative rsk estimates for non-random measurement error. *Statistics in Medicine* **8,** 1051–1069.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys (Wiley Series in Probability and Statistics).* Wiley, Hoboken, NJ.

Sadler, W. A. and Smith, M. H. (1985). Estimation of the response-error relationship in immunoassay. *Clinical Chemistry* **31,** 1802–1805.

Sarkar, S. and Qu, Y. (2007). Quantifying the treatment effect explained by markers in the presence of measurement error. *Statistics in Medicine* **26,** 1955–1963.

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

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman & Hall, London.

Schenker, N., Raghunathan, T. E., and Bondarenko, I. (2009). Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey. *Statistics in Medicine* **29,** 533–545.

Schisterman, E. F., Vexler, A., Whitcomb, B. W., and Liu, A. (2006). The limitations due to exposure detection limits for regression models. *American Journal of Epidemiology* **163,** 374–383.

Schwenke, J. R. and Milliken, G. (1991). On the calibration problem extended to nonlinear modelss. *Biometrics* **47,** 563–574.

Shah, V. P., Midha, K. K., Dighe, S., McGilveray, I. J., Skelly, J. P., Yacobi, A. Y., Layloff, T., Viswanathan, C. T., Cook, C. E., McDowall, R. D., Pittman, K. A., and Spector, S. (1992). Analytical methods validation: Bioavailability, bioequivalence and pharmacokinetic studies. *Pharmaceutical Research* **9,** 1233–1238.

Spiegelman, D., Carroll, R. J., and V, K. (2001). Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. *Statistics in Medicine* **20,** 139–160.

Spiegelman, D., McDermott, A., , and Rosner, B. (1997). Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *American Journal of Clinical Nutrition* **65,** 1179S–1186S.

Spiegelman, D., Rosner, B., and R, L. (2000). Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation designs. *Journal of the American Statistical Association* **95,** 51–61.

Spiegelman, D., Zhao, B., and Kim, J. (2005). Correlated errors in biased surrogates: Study designs and methods for measurement error correction. *Statistics in Medicine* **24,** 1657–1682.

Staudenmayer, J., Ruppert, D., and Buonaccorsi, J. P. (2008). Density estimation in the presence of heteroscedastic measurement error. *American Journal of Clinical Nutrition* **103,** 726–736.

Strauss, W. J., Carroll, R. J., Bortnick, S. M., Menkedick, J. R., and Schultz, B. D. (2001). Combining data sets to predict the effects of regulation of environmental lead exposure in housing stock. *Biometrics* **57,** 203–210.

Subar, A. F., Kipnis, V., Troiano, R. P., Midthune, D., Schoeller, D., Bingham, S., Sharbaugh, C., Trabulsi, J., Runowick, S., Ballard-Barbash, R., Sunshine, J., and Schatzkin, A. (2003). Using intake biomarkers to evalue the extent of dietary misreporting in a large sample of adults: The open study. *American Journal in Epidemiology* **158,** 1–13.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **52,** 528–540.

Thoresen, M. and Laake, P. (2000). A simulation study of measurement error correction methods in logistic regression. *Biometrics* **56,** 868–872.

Thurston, S. W., Spiegelman, D., and Ruppert, D. (2003). Equivalence of regression calibration methods in main study / external validation study designs. *Journal of Statistical Planning and Inference* **113,** 527–539.

Wactawski-Wende, J., Schisterman, E. F., Hovey, K. M., Howards, P., Browne, R. W., Hediger, M., Liu, A., and Trevisan, M. (2009). Biocycle study: design of the longitudinal study of the oxidative stress and hormone variation during the menstrual cycle. *Paediatric and Perinatal Epidemiology* **23,** 171–184.

Walter, S. D. (1998). Variation in baseline risk as an explanation of heterogeneity in meta-analysis. *Statistics in Medicine* **16,** 2883–2900.

Wannemuehler, K. A., Lyles, R. H., Waller, L. A., Hoekstra, R. M., Klein, M., and Tolbert, P. (2009). A conditional expectation approach for associating ambient air pollutant exposures with health outcomes. *Environmetrics* **20,** 877–894.

Willett, W. (1989). An overview of issues related to the correction of non-differential exposure measurement error in epidemiologic studies. *Statistics in Medicine* **8,** 1031 –1040.

Yucel, R. M. and Zaslavsky, A. M. (2005). Imputation of binary treatment variables with measurement error in administrative data. *Journal of the American Statistical Association* **100,** 1123–1132.

Zhao, L. P. and Lipsitz, S. (1992). Designs and analysis of two-stage studies. *American Journal of Epidemiology* **11,** 769–782.

Zidek, J. V., Wong, H., Le, N. D., and Burnett, R. (1996). Causality, measurement error and multicollinearity in epidemiology. *Environmentricse* **7,** 441–451.