Distribution Agreement

Signature:

____________________________ ____________________________

Chengxing Lu Date

Statistical Methods to Adjust for Misclassified Repeated Exposures in Modeling Disease-Exposure Associations

By

Chengxing Lu

Doctor of Philosophy

Biostatistics

Robert H. Lyles, Ph.D.
Advisor

Adolfo Correa-Villaseñor, Ph.D.
Committee Member

Eugene Huang, Ph.D.
Committee Member

John Williamson, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the Graduate School

Date

Statistical Methods to Adjust for Misclassified Repeated Exposures in Modeling Disease-Exposure Associations

By

Chengxing Lu
B.S., Zhongnan University of Economics and Law, 2001
M.S., Emory University, 2008

Advisor: Robert H. Lyles, Ph.D.

An abstract of

A dissertation submitted to the Faculty of the Graduate School of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Biostatistics

2008

Abstract

Statistical Methods to Adjust for Misclassified Repeated Exposures in Modeling Disease-Exposure Associations

By Chengxing Lu

In public health studies, it is common for exposure status to be misclassified. In this dissertation, statistical models to adjust for misclassification will be proposed to address four related questions of interest.

The first question focuses on exploring the association between a disease and the unobservable probability of true exposure given error-prone exposure replicates, in a case-control setting when the exposure is a binary variable. Assuming a beta distribution for the exposure probability, we obtain the estimated association by maximizing the marginal likelihood of the observed replicates and the disease status.

The second question is motivated by the same study, but our interest shifts to assessing the relationship between a disease and the true unknown binary exposure status. We generalize a regular latent class model and a latent class model with a random effect to incorporate a true disease model, for situations of both conditionally independent and dependent exposure replicates. A real data example from the Baltimore Washington Infant Study will be presented to demonstrate methods addressing the first two questions.

In general ANOVA settings, when the samples are misclassified in the study leading to an incorrect attribution of group membership, appropriate adjustments are necessary to obtain valid estimates and inferences. As a methodological transition into our next motivating example, we address this general misclassification problem as our third question, with focus on adapting both the classic regression calibration method and the likelihood method to correct for misclassification in this setting utilizing external or internal validation data.

Our fourth question aims to identify whether a subject's true mean and/or variability in exposure exceeding certain thresholds bears any association with a disease outcome. Misclassifications arise in the categorization of whether the continuous mean exposure or variance exceeds a relevant threshold, where the true mean or variance itself is unobserved. Methods to be discussed include derivations based on the matrix method, regression calibration, a full likelihood approach, and a two-stage empirical Bayes method incorporating categorizations based on both exposure means and variances. Simulation results and analysis of exposure and outcome data from the Mount Sinai Study of Women Office Workers will be presented.

Statistical Methods to Adjust for Misclassified Repeated Exposures in Modeling Disease-Exposure Associations

By

Chengxing Lu
B.S., Zhongnan University of Economics and Law, 2001
M.S., Emory University, 2008

Advisor: Robert H. Lyles, Ph.D.

A dissertation submitted to the Faculty of the Graduate School of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2008

ACKNOWLEDGEMENT

The pursuit of Ph.D has been an extraordinary journey to me. The whole process has not only taught me scientific techniques and skills of analytical thinking, but also approaches of conducting self-discovery, self-improvement, self-management and developing self-confidence. All those are valuable treasures to my life. Therefore, I would like to express my acknowledgement to the biostatistics department at Emory and everyone who helped to provide me this wonderful opportunity to learn, to experience and to grow.

Throughout the journey, I owe the deepest gratitude to my advisor, Dr. Robert H. Lyles for his constant guidance, support and encouragement. Without his enormous help, I would never be able to reach the point where I am. I always feel unbelievably fortunate to have such an advisor who showed me the ways of approaching science, who gave me the inspiration when I met difficulties, and who provided me the freedom to explore the unknown world.

Thanks are also due to Drs. Eugene Huang and John Williamson for their thoughtful comments and constructive suggestions to improve this dissertation at various stages of my research. Moreover, I would like to offer special thanks to Dr. Adolfo Correa-Villaseñor for his introduction to the epidemiologic research area, his insights into the motivating study and his supervision at the Centers for Disease Control and Prevention (CDC). To me he is a dedicated researcher and an intelligent leader. I would also like to thank Drs. Leila W. Jackson, Nancy Min, Michele Marcus and Chanley M. Small for their great help with my real-life data examples.

This is a great opportunity to express my appreciation to my colleagues, Dr. Csaba Siffel, Dr. Meekie Shin, Jim Kucik and Assia Miller, at the National Center for Birth Defects and Developmental Disabilities (NCBDDD), CDC. I cherish the experience of working for CDC and collaborating with every one of them.

Many friends have helped me stay cool in my ups and downs during the journey.

I am grateful for their friendship and deeply appreciate everything they have done for me. Special gratitude is also due to my boyfriend, Bo Yang, for his company in the past years and I look forward to sharing more great moments with him in the future.

Finally, I owe a lot to my parents for their unconditional love, patience and encouragement. Thanks to them for giving me life into this gorgeous world.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 OVERVIEW OF THE MOTIVATING EXAMPLES

In measurement error and misclassification adjustments, Carroll et al. (2006) suggest that we pay careful attention to the types and nature of the errors, and the source study based on which the errors arise and are modeled. The research background and source of misclassification for two motivating studies for this proposal are introduced in the following sections.

1.1.1 The Baltimore-Washington Infant Study (BWIS)

The core focus of modern epidemiologic studies is to identify associations between disease outcomes and potential exposures. In some situations challenges arise from the assessment of potential exposure, particularly in occupational or environmental epidemiologic studies. The common approaches of assessing potential exposures in case-control studies are either based on reported job histories linked to a job-exposure matrix, or based on industrial hygienists' expert opinions on potential exposures in the reported jobs. The latter approach is often considered to be less subject to differential recall than might occur when asking study participants themselves to recall past exposures incurred in the workplace (Bouyer & Hemon (1993), Stewart & Stewart (1994), Stewart et al. (1996)).

Assessments of potential on-the-job exposures by industrial hygienists can nevertheless be subject to error or exposure misclassification (Stewart (1999)). Because of this concern, some case-control studies of occupational exposures have attempted to quantify the error in industrial hygienists' exposure assessments by conducting various types of multiple assessments, including replicate assessments by the same industrial hygienist, by different industrial hygienists, or both. One of the analytical issues raised by such multiple assessments is how to take into account quantifiable uncertainty in the exposure assessment (i.e., misclassification) in the disease-exposure association study. Therefore, the analytical problem might be framed as how to explore the association between a disease and exposure given such replicates, in the absence of a gold standard for exposure, controlling for appropriate covariates.

This dissertation proposal is motivated in part by an occupational epidemiologic research study that poses such challenges. As a sub-study of the Baltimore-Washington infant study (BWIS) (Ferencz et al. (1993)), this initiative is aimed at identifying the association between parental occupational lead exposure and the total anomalous pulmonary venous return (TAPVR), a congenital cardiovascular malformation. Parental job histories were obtained by home interviews. Information including the type of employer, dates of employment, work areas (e.g., plant/section), and job activities were collected for each job held during the period 6 months before pregnancy to the end of pregnancy. These jobs were first screened by 2 epidemiologists, 1 toxicologist, and 1 industrial hygienist to determine jobs involving possible lead exposure. If any one of the screeners believed there was lead exposure in a job, the job was referred to the industrial hygienists for further evaluation. Jobs deemed to involve no lead exposure by all screeners were categorized as having no lead exposure in the analysis.

Three industrial hygienists, blinded to case-control status, were asked to assess the remaining jobs with potential lead exposure. One of the industrial hygienists (referred to as IH1) was selected a priori to be the primary industrial hygienist for the study based upon previous exposure assessment experience (Correa et al. (2006)). Details of the study are described elsewhere (Jackson (2003), Min (1995), Ferencz et al. (1993)). However, disagreements among the industrial hygienists were detected (Correa et al. (2006)), which implied misclassifications from some of the industrial hygienists, if not all of them. Then the primary problem falls into the general framework described above.

Statistical models motivated by this study are presented in chapter 2 and 3.

1.1.2 The Mount Sinai Study of Women Office Workers (MSSWOW)

Reproductive health is defined by the World Health Organization as a state of physical, mental, and social well-being, not merely the absence of disease or infirmity, in all matters relating to the reproductive system and to its functions and processes (Sadana (2002)). Reproductive health is a crucial issue of public health and a central feature of human development. It is a reflection of health in childhood, adolescence, adulthood and beyond reproductive years. Reproductive health status also influences the health of future generations. In the recent centuries, along with the process of industrialization, a greater challenge has been placed upon efforts to maintain and improve reproductive health.

The second motivating example for this dissertation proposal is provided by the reproductive health study MSSWOW. The prospective cohort study was conducted between 1991 and 1994 and designed to explore the effects of Video Display Terminal (VDT) use on rates of spontaneous abortion. 524 of 895 eligible females participated in the study and were interviewed to assess possible confounding factors such as recent medications, illicit drug use, caffeinated and alcoholic beverage consumption, smoking history, medical, gynecological, and reproductive history, partner characteristics and demographics. Other information such as hours of VDT use, exercise performed, stress level, whether sexual intercourse occurred, whether birth control was used, and when menstrual bleeding occurred were obtained from women's diaries. We are specifically interested in the repeated menstrual cycle length data, which can be derived from the women's diaries. However, when the subject-specific mean and variance describing each woman's cycle history are 'true' predictors of interest in models for reproductive health outcomes, the sample mean and variance researchers usually utilize as surrogates will introduce measurement errors to the study. A previous paper by Small et al. (2006) conducted careful analysis on this data and suggested that compared with 30- to 31-day cycles, women with shorter and longer cycles were more likely to experience spontaneously abortion. However, this paper did not address the misclassification issue as the sample means and variances were “plugged in” to the health outcome models in the analysis. However, in the concern of measurement error/misclassification, the errors may not be ignorable in some cases, especially for those women with small numbers of cycles observed in the study. Therefore, the classification, based on the surrogate of sample mean and variance, of women into “high” or “low” groups with regard to the subject-specific mean cycle length and variability can introduce misclassification bias, when the research question is to assess the association between this group membership status and some health outcome.

Statistical models motivated by this study are presented in chapter 5. Before we begin to describe the proposal research, however, some statistical background knowledge will be reviewed in the following sections.

1.2 BACKGROUND

1.2.1 General measurement error/misclassification modeling framework

Following Clayton (1992), three main statistical models are commonly considered in a traditional measurement error/misclassification problem: a "true disease model" (TDM), which models the relationship between the outcome and the true error-free exposure of interest; a "measurement error model" (MEM), which models the variable prone-to-error and its correspondent error-free variable; and an "exposure distribution model" (EDM), which is a model characterizing the distribution of the error-free exposure in the population under study (e.g., Lyles & Kupper (2000)). Note that among the three models, the TDM is usually the one of ultimate interest. The MEM is named and defined in terms of measurement error but could also be applied in the misclassification environment. An example of measurement error adjustment under a normal-normal paradigm will demonstrate how the three models can be specified.

Suppose we are interested in an epidemiologic study that measures exposure repeatedly for each subject. The essential goal of interest will be to determine the association between a continuous disease outcome and the true unobserved subject-specific mean exposure. Assume a one-way random effect ANOVA structure that defines the relation between the observed exposure and the true mean exposure, as follows:

$$X_{ij} = \mu + b_i + \epsilon_{ij} \qquad (i = 1, 2, \ldots, k; j = 1, 2, \ldots, n)$$

where X_{ij} is the j-th (out of n) observed exposure measurement for the i-th subject (out of k). b_i and ϵ_{ij} are assumed to follow normal distributions, i.e., $b_i \overset{iid}{\sim} N(0, \sigma_b^2)$ and $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_w^2)$, where the random variables b_i and ϵ_{ij} are mutually independent.

This model defines $\mu_i = \mu + b_i$ as the true unknown mean exposure for subject i, which is of primary interest.

To fit in the framework discussed above, assume a simple linear regression of a continuous outcome variable Y_i on μ_i is ultimately of interest. Further, assume the logical surrogate $\bar{X}_i = \frac{1}{n}\sum X_{ij}$ would replace μ_i in a "naive" analysis. Then

$$\begin{aligned} &\text{TDM:} \qquad Y_i = \alpha + \beta\mu_i + e_i \\ &\text{MEM:} \qquad \bar{X}_i = \mu_i + \bar{\epsilon}_i \\ &\text{EDM:} \qquad \mu_i \sim N(\mu, \sigma_b^2) \end{aligned}$$

A generalization to the above model (Lyles & Kupper (1997)) was altered to define mean exposure on a lognormal scale for several groups of workers. Specifically, they proposed a model with multiplicative-lognormal MEM structure after grouping the workers sharing the same job characteristics in the context of occupational epidemiology. Therefore, a log-transformed exposure measurement could be modeled as the following:

$$Y_{gij} = ln(X_{gij}) = \mu_{yg} + \delta_{gi} + \epsilon_{gij} \qquad (g = 1, \ldots, G; i = 1, \ldots, k_g; j = 1, \ldots, n_{gi})$$

where X_{gij} is the j-th (out of n_{gi}) exposure measurement on the i-th (out of k_g) worker in the g-th (out of G) group. Similar to the former example, δ_{gi} and ϵ_{gij} are assumed to be normally distributed and mutually independent: $\delta_{gi} \overset{iid}{\sim} N(0, \sigma_{bg}^2)$; $\epsilon_{gij} \overset{iid}{\sim} N(0, \sigma_{wg}^2)$

Suppose we are interested in the unobservable mean exposure $\mu_{xgi} = E(X_{gij}) = E(\exp(\mu_{yg} + \delta_{gi} + \epsilon_{gij})) = \exp(\mu_{yg} + \delta_{gi} + \sigma_{wg}^2/2)$ as the independent variable. Since μ_{xgi} is unobservable, we might take $\exp(\bar{Y}_{gi}) = \exp({n_{gi}}^{-1}\sum_{j=1}^{n_{gi}} Y_{gij})$ as the surrogate.

Therefore, the TDM-MEM-EDM settings might become:

$$\begin{array}{ll} \text{TDM:} & R_{gi} = \alpha + \beta\mu_{xgi} + \sum_{t=1}^{T} \gamma_t C_{git} + e_{gi} \\ \text{MEM:} & \exp(\bar{Y}_{gi}) = \mu_{xgi}(U_{gi}) \\ \text{EDM:} & \mu_{xgi} \sim \text{lognormal}[(\mu_{yg} + \sigma_{wg}^2/2), \sigma_{bg}^2] \end{array}$$

where R_{gi} is the health outcome variable, C_{git} represents the t-th out of T individual-specific covariates that may need to be controlled, and U_{gi} is the random disturbance relating the surrogate and the unobservable mean exposure. More details on the latter model examples are given by Lyles & Kupper (2000).

1.2.2 Beta-binomial regression model

In beta-binomial regression models, outcomes are usually grouped. Suppose there are k groups, and that within group i (i= 1,2,...,k) there are n_i subjects with observations denoted as X_{ij} (j=1,2,...,n_i). The binary responses from each member of a group are assumed to be mutually correlated. However, conditioned on the probability p_i, assumptions are made that the responses within the group are independently Bernoulli distributed, i.e., $X_{ij}|p_i \sim Bin(1, p_i)$. The p_i's are assumed to follow a beta distribution, i.e., $p_i \sim Beta(\alpha, \beta)$, where $\alpha > 0$ and $\beta > 0$. The probability density function of the beta distribution is

$$\frac{p_i^{\alpha-1}(1-p_i)^{\beta-1}}{B(\alpha,\beta)}, \qquad \alpha > 0, \beta > 0$$

where $B(\alpha, \beta)$ is the beta function, i.e.,

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)},$$

and $\Gamma(.)$ is the gamma function, defined by

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$$

The mean and variance of p_i are $\alpha/(\alpha+\beta)$ and $\alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)$. The marginal distribution for $X_i = \sum_{j=1}^{n_i} X_{ij}$ is a beta-binomial distribution BB(n_{ij}, α, β) with density

$$P(x_i; n_i, \alpha, \beta) = \frac{B(x_i+\alpha, n_i-x_i+\beta)\begin{pmatrix} n_i \\ x_i \end{pmatrix}}{B(\alpha,\beta)}, \quad x_i = 0, 1, \ldots, n_i$$

The beta-binomial distribution has proven useful in many settings involving clustered data, including toxicology studies (e.g., Williams (1975)). Under a different parameterization, let the parameter μ represent the mean of the beta distribution, i.e., $\mu = \alpha/(\alpha+\beta)$, and define the parameter $\theta = \mu/\alpha$. Since α and β are greater than zero, μ and θ have to be greater than zero as well. Then to incorporate covariates, the beta-binomial regresson model can be set up with a parameterization that dictates

$$\mu_i = \frac{exp(Z_i^T\gamma)}{1+exp(Z_i^T\gamma)}$$

where $\eta_i = Z_i^T\gamma$ is referred to as the linear predictor (e.g., Gange et al. (1996)). Then correspondingly,

$$\alpha_i = \frac{\mu_i}{\theta} = \frac{exp(Z_i^T\gamma)}{\theta(1+exp(Z_i^T\gamma))} \tag{1.1}$$

and

$$\beta_i = \frac{1}{\theta} - \alpha_i = \frac{1}{\theta} - \frac{\mu_i}{\theta} = \frac{1}{\theta(1+exp(Z_i^T\gamma))} \tag{1.2}$$

1.2.3 Latent class modeling

In epidemiologic or medical studies, researchers constantly struggle with the problem that the primary study variables are not always directly observable. For example, in diagnostic settings, certain diseases have no absolute gold standard test available or definable in real life, i.e., no current available test is believed to provide perfect sensitivity or specificity. Therefore, the true disease status is unknown to the researchers, and we will refer to this as a case of "unobservable outcome". Another example, to be discussed in detail later, is the case where the true exposure status in an epidemiological study is not directly observable or assessable. In such cases, multiple measurements or assessments may be conducted to bolster accuracy. Statistical analysis might show disagreement among the measurements, which suggests that misclassifications are present within individual measurements. Therefore, the questions include, whether there is one measurement/assessment that is more reliable than the others, and whether we can make inference on the unobservable exposure in a valid way. In contrast to the previous example, we refer to this as a case of "unobservable exposure". In both situations, biased estimates and potentially invalid inferences could result if the true outcome or exposure were not assessed properly.

Latent class analysis (LCA) was initially applied in the social sciences, such as psychology, education, and market research. Subsequently, researchers have applied LCA models in the medical world to deal with problems like those described above (Goodman (1974), Hui & Walter (1980), Walter (1984), Formann & Kohlmann (1996)). LCA models generally presume a parametric model associating multiple diagnoses or measurements with an the unobserved latent variable, and often assume conditional independence among the observed multiple diagnoses or measurements.

The conditional independence assumption presumes that within each latent class, replicate measurements are mutually independent. For example, within a class of unobserved medical status ("yes" or "no"), the presence/absence of symptoms based on one diagnosis is considered unrelated to the presence/absence decisions based on all others. Or in the "unobserved exposure" example, the conditional independence assumption implies that given the true unobserved exposure status is "exposed" or "unexposed", all the assessments are independent of each other.

In practice, the conditional independence assumption may or may not hold. In the case of diagnostic tests, a patient may exhibit clear characteristics such that the present/absent decision of multiple assessors are conditionally dependent. Or in the exposure assessment example, if one professional misclassifies the exposure status due to a worker's unrecorded hours of overtime, then very likely the other professionals could make the same mistake. Therefore, extensions of the traditional LCA with allowance for conditional dependency were proposed, e.g., by Qu et al. (1996). These authors introduced a common random effect for all the observed measurements when modeling the relationship between those measurements and the unobserved true variable, to take care of the correlations. Both LCA with and without the conditional independence assumption will be reviewed in the following sections.

Latent class model with conditional independence assumption

Let $\mathbf{W_i}$= $(w_{i1}, w_{i2}, \ldots, w_{in})'$ be the n observed error-prone binary variables for subject i, representing multiple diagnoses or measurements as indicated above. Suppose the true unobserved error-free binary exposure is X_i for subject i. Define

$$\pi_{jx} = P(w_{ij} = 1 | X_i = x) \quad (j = 1, 2, \ldots, n; x = 0, 1),$$

which is assumed to be identical across all k subjects for the same (jth) observation. Then the sensitivity and specificity for the jth observation are

$$\pi_{j1} = P(w_{ij} = 1 | X_i = 1) \tag{1.3}$$

and

$$1 - \pi_{j0} = 1 - P(w_{ij} = 1|X_i = 0) \tag{1.4}$$

respectively.

Therefore the probability of observing w_{ij} as the ith subject's jth measurement, given the true underlying exposure x, is

$$\begin{aligned} P(w_{ij}|X_i = x) &= P(w_{ij} = 1|X_i = x)^{w_{ij}}(1 - P(w_{ij} = 1|X_i = x))^{1-w_{ij}} \qquad (1.5) \\ &= \pi_{jx}^{w_{ij}}(1 - \pi_{jx})^{1-w_{ij}} \qquad (1.6) \end{aligned}$$

Under the conditional independence assumption, i.e., all the n observations are mutually independent within the ith subject, the probability of observing $\mathbf{W_i}$ given the true unobserved status of X_i is

$$P(\mathbf{W_i}|X_i = x) = \prod_{j=1}^{n} \pi_{jx}^{w_{ij}}(1 - \pi_{jx})^{1-w_{ij}}$$

Then by summing over the possible values that the unobserved X_i could possibly take, the probability of observing $\mathbf{W_i}$= $(w_{i1}, w_{i2}, \ldots, w_{in})'$ is

$$P(\mathbf{W_i}) = \sum_{x=0}^{1} [\tau_x \prod_{j=1}^{n} \pi_{jx}^{w_{ij}}(1 - \pi_{jx})^{1-w_{ij}}]$$

where $\tau_x = P(X = x)$.

Finally, the likelihood up to a proportionality constant for observing all the $\mathbf{W_i}$s is given by

$$\mathscr{L}(\mathbf{W_1}, \mathbf{W_2}, \ldots, \mathbf{W_n}; \pi_{jx}, \tau_x) = \prod_{i=1}^{k}\{\sum_{x=0}^{1} \tau_x[\prod_{j=1}^{n} \pi_{jx}^{w_{ij}}(1 - \pi_{jx})^{1-w_{ij}}]\} \tag{1.7}$$

The parameters can be estimated by maximizing the likelihood (1.7). Computational complications will be discussed in detail in later chapters.

Latent class model with random effect

To relax the conditional independence assumption, Qu et al. (1996) model the common characteristics among the multiple observed diagnoses or measurements in a given class of the latent variable by a latent random variable T, which is assumed to follow a standard normal distribution. Specifically,

$$P(w_{ij}=1|X_i=x,T=t)=F(a_{jx}+b_{jx}t) \tag{1.8}$$

where $t \sim N(0,1)$. Qu et al. select the probit function for F, but one could also use the logit or any other function appropriate for categorical outcomes.

Then the sensitivity and specificity for the jth measurement will be

$$P(w_{ij}=1|X_i=1)=\int_{-\infty}^{\infty}F(a_{j1}+b_{j1}t)f(t)dt \tag{1.9}$$

and

$$1-P(w_{ij}=1|X_i=0)=1-\int_{-\infty}^{\infty}F(a_{j0}+b_{j0}t)f(t)dt \tag{1.10}$$

where $f(t)$ is the probability density function of the standard normal distribution. If F is the probit function, explicit forms for sensitivity and specificity in (1.9) and (1.10) can be obtained as $\Phi(a_{j1}/\sqrt{1+b_{j1}^2})$ and $\Phi(-a_{j0}/\sqrt{1+b_{j0}^2})$, $j=1,2,\ldots,n$, where $\Phi(.)$ is the cumulative probability function for the standard normal distribution (Qu et al. (1996)).

Therefore the probability of observing $\mathbf{W_i}|X_i=x$ in the ith subject is

$$P(\mathbf{W_i}|X_i=x) \quad = \quad \int_{-\infty}^{\infty}\prod_{j=1}^{n}F(a_{jx}+b_{jx}t)^{w_{ij}}(1-F(a_{jx}+b_{jx}t))^{1-w_{ij}}f(t)dt \tag{1.11}$$

Then, up to a constant, the likelihood becomes

$$\mathscr{L}(\mathbf{W_1}, \mathbf{W_2}, \ldots, \mathbf{W_k}; a_{jx}, b_{jx}, \tau_x) = \prod_{i=1}^{k} (\sum_{x=0}^{1} \tau_x P(\mathbf{W_i}|X_i = x)) \tag{1.12}$$

where $P(\mathbf{W_i}|X_i = x)$ is defined in (1.11).

Integrations in the likelihood can be approximated by Gaussian-Hermite quadrature. Computational complications will again be discussed later.

In the research to date, most of the LCA applications are focused on dealing with the complications in the scenario of “unobserved outcome”. Relatively less research applies LCA with extension to the scenario of “unobserved exposure”.

Identifiability for latent class models

In this dissertation proposal, we will be building and discussing parametric models all throughout. A parametric model is one which specifies the main structure of the probability distribution but leaves some of the parameters to be estimated (Tsiatis (2006)). Bickel & Doksum (2001) defined identifiability in a parameterization in the following way:

Definition A parameterization is called identifiable if it is one-to-one. That is, let ξ_1 and ξ_2 be two parameter values with their corresponding distributions P_{ξ_1} and P_{ξ_2}, then $\xi_1 \neq \xi_2$ implies $\mathrm{P}_{\xi_1} \neq \mathrm{P}_{\xi_2}$.

Wieringen (2005) studied the regular latent class model with the conditional independence assumption, and pointed out that the model was generally not identifiable without any restriction. To demonstrate this, choose one set of parameters $\Psi = (\tau_1, \pi_{11}, \pi_{21}, \ldots, \pi_{k1}, \pi_{10}, \pi_{20}, \ldots, \pi_{k0})$ for the regular latent class model with conditional independence assumption. And then define

$$\Psi^* = (1 - \tau_1, \pi_{10}, \pi_{20}, \ldots, \pi_{k0}, \pi_{11}, \pi_{21}, \ldots, \pi_{k1})$$

We will find $\mathrm{P}(\Psi) = \mathrm{P}(\Psi^*)$ with the likelihood in (1.7), which violates the definition of identifiability above.

Wieringen (2005) proposed and proved the criteria for global identifiability for the regular latent class model, quoted as the following:

Theorem 1. *For global identifiability of model* (1.7) *it is sufficient to require*

$$0 < \tau_1 < 1 \textit{ and } 1 \geq \pi_{j1} > \pi_{j0} \geq 0 \qquad \textit{for j=1,...,n,} \quad \textit{AND}$$

$$-1 + \prod_{j=1}^{n} (l_j + 1) \geq 2n + 1$$

The l_j in the theorem is the maximum categorical level that the observed variable could take, counting from zero. In our situation, the observed assessments w_{ij} could take values 0 and 1. So all the l_js will be equal to 1 for every j, i.e., $l_1 = l_2 = \ldots = l_n = 1$.

Here, π_{j1} is the sensitivity and π_{j0} is one minus specificity by definition, which take expressions (1.3) and (1.4) in the latent class model under conditional independence and (1.9) and (1.10) in latent class models with the conditional dependence assumption. Therefore, following Theorem 1, the second sufficient condition in the theorem is equivalent to the restriction that the summation of sensitivity and specificity has to be greater than one.

The third sufficient condition in the theorem could be intuitively interpreted by the criteria discussed by McHugh (1956) and Goodman (1974) that to make one model identifiable, the number of parameters has to be less than or equal to the number of degrees of freedom of the model. For example, Wieringen (2005) assumes different sensitivity and specificity across observations within a subject. Therefore, there would be $2n$ π_{jx} parameters, plus one parameter τ_1, which equals the number in the right hand side of the third condition in Theorem 1, $2n + 1$. On the other hand, each observation takes values of 0 through l_j and there are n observations in total for each subject, so the total number of possible combinations of these n cells is $\prod_{j=1}^{n} (l_j + 1)$. However, the total number of the combinations we could see in the data has to be the total number of subjects k. Henceforth, the number of degrees of freedom is $\prod_{j=1}^{n}(l_j+1)-1$, which is the left hand side of the third condition in Theorem 1.

1.2.4 Misclassification corrections in case-control studies with validation sets

The common complication in exposure misclassification problems is that the categorical exposures are collected with error. People would naturally think of evaluating the sensitivities and specificities of the collected exposure surrogates, when dealing with such problems. Simple sensitivity and specificity estimates are usually obtained from validation sets, where the true exposures could be observed for a small group of subjects, from internal or external sources. However, when no validation sets are available, estimating sensitivity and specificity usually involves some complications.

In the early years of studies, researchers developed intuitively and computationally straightfoward methods to correct misclassification based on situations when estimated sensitivity and specificity were available. Barron (1977) proposed the "matrix method", by which the expectations of the unobserved cells (a, b, c and d from Table 1.1) are obtained by pre-multiplying the expectations of the observed cells (A, B, C and D from Table 1.1) by a matrix, including functions of the sensitivities and specificities estimated from a validation study, i.e.,

$$\begin{bmatrix} E(a) \\ E(b) \\ E(c) \\ E(d) \end{bmatrix} = \begin{bmatrix} s\hat{e}n & 1-s\hat{p}c & 0 & 0 \\ 1-s\hat{e}n & s\hat{p}c & 0 & 0 \\ 0 & 0 & s\hat{e}n & 1-s\hat{p}c \\ 0 & 0 & 1-s\hat{e}n & s\hat{p}c \end{bmatrix}^{-1} \times \begin{bmatrix} E(A) \\ E(B) \\ E(C) \\ E(D) \end{bmatrix} \tag{1.13}$$

Here, $s\hat{e}n$ and $s\hat{p}c$ are the estimated sensitivity and the specificity, respectively, which are defined by $sen = \mathrm{P}(W = 1|E = 1)$ and $spc = \mathrm{P}(W = 0|E = 1)$, if W is the binary surrogate and E is the true binary exposure.

Table 1.1: Data layout for the matrix method

	Observed		Unobserved	
D	W=1	W=0	E=1	E=0
1	A	B	a	b
0	C	D	c	d

An alternative representation of (1.13) would be in terms of probabilities, instead of expectations. Although they are equivalent, we find the probability representation easier to interpret in applications in section 5.2.

$$\begin{bmatrix} \mathrm{P}(E=1|D=1) \\ \mathrm{P}(E=0|D=1) \\ \mathrm{P}(E=1|D=0) \\ \mathrm{P}(E=0|D=0) \end{bmatrix} = \begin{bmatrix} \hat{sen} & 1-\hat{spc} & 0 & 0 \\ 1-\hat{sen} & \hat{spc} & 0 & 0 \\ 0 & 0 & \hat{sen} & 1-\hat{spc} \\ 0 & 0 & 1-\hat{sen} & \hat{spc} \end{bmatrix}^{-1} \times \begin{bmatrix} \mathrm{P}(W=1|D=1) \\ \mathrm{P}(W=0|D=1) \\ \mathrm{P}(W=1|D=0) \\ \mathrm{P}(W=0|D=0) \end{bmatrix} \tag{1.14}$$

After some reparameterizing and further developments stemming from Barron (1977), Marshall (1990) presented an alternative approach termed the “inverse matrix” method. The two methods were compared and extended by Morrissey & Spiegelman (1999) in situations of both differential and non-differential misclassification. Some clarifications regarding how these methods relate to maximum likelihood were later conducted by Lyles (2002).

1.2.5 Gaussian-Hermite quadrature

Gaussian-Hermite quadrature is a special kind of Gaussian quadrature, over the interval of ($-\infty$,∞). The fundamental theorem of Gaussian quadrature states that the optimal abscissas of the m point Gaussian quadrature formulas are precisely the roots of the orthogonal polynomial for the same interval and weighting function. Gaussian quadrature is optimal because it fits all polynomials up to degree $2m-1$ exactly. Particularly, the abscissas for the Gaussian-Hermite quadrature of order n are given by the roots x_i of the Hermite polynomials $H_n(x)$

$$H_n(x) = \frac{n!}{2\pi i} \oint e^{-t^2+2tx} t^{-n-1} dt \qquad (1.15)$$

and the weight is e^{-x^2} (Hildebrand (1956)). Therefore, if we want to integrate a function $f(x)$,

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} e^{-x^2}[e^{x^2} f(x)dx] \approx \sum_{k=1}^{n} w(x_k) e^{x_k^2} f(x_k)$$

where $w(x_k)$ is the weight e^{-x^2} and x_k is the kth root of (1.15) and it remains the same regardless of what specific function is being integrated.

1.2.6 Profile-likelihood-based confidence intervals

The classical Wald confidence intervals are mainly based on the asymptotic normality of the maximum likelihood estimate $\hat{\theta} \in \mathbb{R}^k$. However, the properties of $\hat{\theta}$ in small samples can be quite different from its asymptotic properties, and estimates of the asymptotic variances can be significantly biased (Evans & Kim (1996)). In these situations, profile-likelihood-based confidence intervals may be a more robust alternative.

Following Venzon & Moolgavkar (1988), let $\theta \in \mathbb{R}^k$ denote the parameter vector to be estimated, and $l(\theta)$ the log-likelihood for values of θ in the parameter space $\Theta \subseteq \mathbb{R}^k$. Suppose the MLE of θ is $\hat{\theta}$, i.e.,

$$l(\hat{\theta}) = \max_{\theta \in \Theta} l(\theta)$$

Denote the parameter of interest as θ_j. The profile likelihood approach considers the other k-1 parameters as nuisance parameters. Then consider a (k-1)-dimensional restricted parameter space $\Theta_j(\beta)$, where the parameter of interest is fixed at some value β. The profile likelihood for β is defined as

$$\tilde{l}_j(\beta) = \max_{\theta \in \Theta_j(\beta)} l(\theta)$$

where the above likelihood needs to be maximized with θ_j constrained to equal β, in accordance to the definition of profile likelihood. Therefore, an approximate 1-α profile-likelihood-based confidence interval for θ_j is given by

$$\beta : 2[l(\hat{\theta}) - \tilde{l}_j(\beta)] \leq q_1(1 - \alpha/2)$$

where $q_1(1 - \alpha/2)$ is the $(1 - \alpha/2)$ quantile of the chi-square distribution based on 1 degree of freedom.

Numerous researchers have studied the pros and cons of the likelihood-based confidence intervals and the Wald ones (Donaldson & Schnabel (1987); Meeker (1987); Ostrouchov & Meeker (1988); Evans & Kim (1996) etc.). A common consensus is that the likelihood-based confidence intervals are more advanced in terms of accuracy, however, computationally more challenging. Venzon & Moolgavkar (1988) and Cook & Weisberg (1990) introduced some algorithms that simplify the process of obtaining the profile-likelihood-based confidence intervals in certain situations.

1.3 OUTLINE

In this dissertation, we will focus on identifying associations between disease outcomes and potential exposures, when the observed exposure assessments are prone to misclassifications. Although the true binary exposure status is what the industrial hygienists seek to assess in one motivating study and is most relevant to individual workers, sometimes researchers may be interested in identifying the properties of the probability of exposure among individuals or within a subpopulation, and also the association between the disease outcome and the probability of exposure.

To address the research interest and also better understand the potential exposure distributions, the first chapter of this dissertation will discuss approaches to determine the disease-probability of exposure relationship based on certain assumptions. The second chapter will discuss approaches to examine disease-exposure associations with or without presuming the assumption of conditional independence among the industrial hygienists, when the true unobserved exposure status is of interest.

Prior to addressing a specific research question in epidemiology in the fourth chapter, we will discuss a more generic problem of misclassification in Chapter three: Two sample t-test and ANOVA have been widely applied and well developed statistical techniques. However, when the samples are misclassified to the groups that they do not belong to, appropriate adjustments are necessary to obtain valid inferences. We will focus on adapting the regression calibration method to correct the effect of misclassification in the two sample t-test or ANOVA setting when the group indicator is prone to misclassification. We discuss under the situations when external or internal validation data are available, the appropriate approaches of adjusting the misclassification and their respective advantages and disadvantages.

Back to epidemiology studies, a particular concern for the exposure measurements in the circumstance of evaluating a disease-exposure relationship is that the exposure itself is often highly variable in time, which makes it hard to measure the relevant subject-specific exposures directly (e.g., Brunekreef et al. (1987)). Therefore, the studies are usually designed to measure the exposures repeatedly over time at fixed sites to approximate the "true" subject-specific exposures. Under those study designs, the mean exposure during a certain time period is usually of interest because it is a reasonable measurement that smoothes the variability and also we may believe it is the chronic exposure, reflected by the mean, that makes people unhealthy more than any extreme one-time exposure.

Taking the mean continuous exposure as the independent variable of interest, the fourth chapter will discuss the potential misclassification problems within the threshold model structure. A threshold is defined by Wilson (1973) as a dose below which an outcome seen in excess of that is not produced. More specifically, a threshold is a cut-off point beyond which the exposures are believed to cause adverse impact on workers' health, while no significant adverse health impacts are believed to be brought on if the exposures are kept below the cut-off point (Hatch (1971), Wilson (1973), Haseman & Kupper (1979), Schwartz et al. (1995)). In our case, we would categorize the continuous mean exposure into two levels: above or below the threshold. The question of interest is, whether exposure to a certain toxic agent exceeding the threshold will cause an adverse health outcome.

Since the "true" mean of the exposure for a specific subject is unknown to the researchers, we would take a natural surrogate of that as the average of the exposures we observed. By this means, measurement errors will be introduced into the continuous mean exposures. As indicated by Flegal et al. (1991), misclassification in exposure categories stemming from a continuous variable arises because of the measurement errors in the continuous variable. In our case, misclassifications of whether mean exposure exceeds the threshold would occur when taking the surrogate for the unknown "true" mean exposure. Therefore, the analytical question in the fourth chapter goes to how to adjust the misclassifications and correct the exposure-outcome relationship based on a proper model.

In this dissertation, descriptions of the methodology, simulation results, and real data analysis results will be presented in each chapter. The last chapter will provide some brief summaries and future work plans.

Chapter 2

ASSESSING THE ASSOCIATION BETWEEN A HEALTH OUTCOME AND THE PROBABILITY OF EXPOSURE

2.1 INTRODUCTION

The common approaches of assessing the potential exposures in case-control studies are either based on reported job histories linked to a job-exposure matrix, or based on expert opinions on potential exposures in the reported jobs by industrial hygienists. In either approach, researchers often show interest in the categorical exposure status as well as the probability of exposure (Satten & Kupper (1993), Wijngaarden et al. (2003), Correa et al. (2006)). The probability of exposure is usually understood as the probability that a specific worker in a particular industry was exposed to some chemical or toxicant, but often defined as the percentage of workers in a particular industry or job category that were exposed.

In ecologic studies, people are sometimes interested in whether larger percentages of workers exposed coincide with larger proportions developing a disease. However, most of the studies base the exposure probability information directly on either the job-exposure matrix or the experts' assessments, which might be subjective. In a paper by Wijngaarden et al. (2003) for example, the probability of exposure information was obtained from either a job-exposure matrix or experts' assessments, based on availability. The continuous exposure probabilities were categorized into four levels to examine the association between the categorized probability of exposure and the outcome. In the interest of being more objective and adjusting possible misclassifications incurred from different experts' assessments, we would like to propose model-based probability estimates to assess the relationship between disease outcome and the probability of exposure.

When each job is assessed by different industrial hygienists repeatedly, the assessments on the same job are usually assumed to be correlated with each other. However, when conditioning on the same job, the assessments from different industrial hygeniests could be considered independent. If the predictor of interest in the health outcome model is the probability of exposure, which is a continous variable lying between 0 and 1, a beta distribution makes an appealling model for it. However, once the true probability of exposure is considered fixed, the binary observed exposure status would follow a Bernoulli distribution with the probability of success as the 'fixed' true probability of exposure. In this chapter, we work under these conditions and we assume the industrial hygienists are exchangeable so that their assessments on the same subject follow the same distribution. An underlying implication of the above statement is that on average, the estimated probability of exposure by the industrial hygienists equals the true probability of exposure. In other words, we assume the industrial hygienists have an unbiased probability of exposure in mind when they assess the individual binary exposure status. This is a reasonable assumption if we believe the industrial hygienists' assessments are not biased on the population level. As a hypothetical example, assume the industrial hygienists each assess 10,000 workers from the same working environment (assume the same exposure probability across those workers). Then, the proportion exposed from every industrial hygienist's assessment should be close to the true proportion exposed. Last, we make the non-differential measurement error assumption, which postulates that the surrogate is independent of the outcome, given the true unobserved exposure.

The other sections in the chapter will be organized as follows. A model for repeated exposure assessments based on beta-binomial assumptions and linked to a health outcome via logistic regression will be proposed in section 2.2. Then the BWIS example will be analyzed to demonstrate the proposed model in section 2.3. Interesting results with covariates in the model will also be presented. Simulation results will be shown in the next section, followed by some brief discussions.

2.2 MODEL

2.2.1 Model structure

Following the general structure introduced by Clayton (1992) (details were reviewed in 1.2.1), the modeling of the relation between a disease outcome and the exposure probability is divided into three parts: TDM, MEM and EDM.

$$\begin{array}{ll} \text{TDM:} & \text{logit}[\text{P}(Y_i = 1|P_i = p_i)] = \theta_0 + \theta_1 p_i \quad (i = 1, 2, \ldots, k) \\ \text{MEM:} & w_{ij} \overset{iid}{\sim} \text{Bernoulli}(p_i) \quad (j = 1, 2, \ldots, n_i) \\ \text{EDM:} & P_i \text{ follows a } Beta(\alpha, \beta) \text{ distribution} \end{array}$$

This model is analogous to the normal-normal model introduced in 1.2.1, in terms of data characteristics and corresponding model assumptions. For both models, the independent variable in the TDM is not observable directly, while its distribution is assumed to be given by the EDM. The repeatedly observed exposure surrogates or assessments from different experts (referred to as raters later), i.e., the X_{ij}s in the model in 1.2.1 and the w_{ij}s in the current model, are correlated. However, under the model assumptions, the observed exposure surrogates are no longer correlated, conditional on the independent variables in the TDM.

2.2.2 Marginal Likelihood

Following the notations in the first chapter, let $\mathbf{W_i} = (w_{i1}, w_{i2}, \ldots, w_{in})$ be the n observed exposure assessments from n raters for subject i. Suppose the true unobserved probability of exposure is P_i for subject i and Y_i is the binary disease outcome status. As introduced in 1.2.2, the mean of the beta-binomial distribution μ could be modeled as

$$\mu_i = \frac{\exp(Z_i^T \gamma)}{1 + \exp(Z_i^T \gamma)}$$

Under this new reparameterization of μ_i, the EDM, which is a beta distribution, takes a density in the form

$$f_{P_i}(p_i) = \frac{p_i^{\alpha-1}(1-p_i)^{\beta-1}}{B(\alpha,\beta)} \tag{2.1}$$

$$= \frac{p_i^{\frac{\exp(Z_i^T\gamma)}{\theta(1+\exp(Z_i^T\gamma))}-1}(1-p_i)^{\frac{1}{\theta(1+\exp(Z_i^T\gamma))}-1}}{B(\frac{\exp(Z_i^T\gamma)}{\theta(1+\exp(Z_i^T\gamma))}, \frac{1}{\theta(1+\exp(Z_i^T\gamma))})} \tag{2.2}$$

where α and β have been replaced by the expressions in (1.1) and (1.2).

We assume the observed exposure surrogates are independent and follow identical Bernoulli distributions given the true underlying exposure probability. Hence the probability of observing $\mathbf{W_i} = (w_{i1}, w_{i2}, \ldots, w_{in})$, given the exposure probability, is

$$\mathrm{p}(\mathbf{W_i}|P_i = p_i) = p_i^{\sum_{j=1}^n w_{ij}}(1-p_i)^{n-\sum_{j=1}^n w_{ij}} \tag{2.3}$$

When the TDM takes a logistic regression model as in 2.2.1, the likelihood of observing the disease outcome Y_i given the exposure probability P_i is

$$\mathrm{P}(Y_i|P_i = p_i) = (\frac{\exp(\theta_0+\theta_1 p_i)}{1+\exp(\theta_0+\theta_1 p_i)})^{Y_i}(\frac{1}{1+\exp(\theta_0+\theta_1 p_i)})^{1-Y_i} \tag{2.4}$$

Note that like other logistic models, covariates can be controlled for in the TDM.

When the true exposure probability is unknown, an intuitive approach is to take a surrogate of the exposure probability, based on the average of the industrial hygienists' assessments within a subject, i.e., $\frac{1}{n}\sum_{j=1}^n w_{ij}$ ($w_{ij} = 0$ or 1). A logistic model with the surrogate as the predictor and the disease status as the outcome is later refered to as the "naive" model. The estimate can be obtained by simply plugging $\frac{1}{n}\sum_{j=1}^n w_{ij}$ in for p_i in (2.4). However, taking the surrogate of the exposure probability as the explanatory variable would be a source of measurement error, especially when the number of repeated measurements is relatively small.

For a more appropriate analysis, note that the likelihood of observing both the disease outcome and the exposure surrogate for subject i will be

$$\begin{aligned} \mathrm{P}(Y_i, \mathbf{W_i}) &= \int_0^1 \mathrm{P}(Y_i|P_i = p_i, \mathbf{W_i})\mathrm{P}(\mathbf{W_i}|P_i = p_i) f_{P_i}(p_i) dp_i \\ &= \int_0^1 \mathrm{P}(Y_i|P_i = p_i)\mathrm{P}(\mathbf{W_i}|P_i = p_i) f_{P_i}(p_i) dp_i \qquad (2.5) \end{aligned}$$

Note that $f_{P_i}(p_i)$ is the distribution of the true unobserved exposure probability and $\mathrm{P}(\mathbf{W_i}|P_i = p)$ should be the probability of observing the exposure surrogates, given the exposure probability of the surrogate. As discussed in 2.2.1, we assume these two probability concepts to be equivalent when we believe the industrial hygienists' assessments to be unbiased.

The 2nd equation in (2.5) holds under the non-differential measurement error assumption mentioned in the last section, which is implied here by (2.3). This commonly applied assumption in measurement error/misclassification problems, supposes that the surrogate is independent of the outcome, given the true unobserved exposure. It usually takes the form as

$$\mathrm{P}(Y_i|X_i, \mathbf{W_i}) = \mathrm{P}(Y_i|X_i),$$

where X_i is the true exposure status. However, in the current modelling context, the analogous equation

$$\mathrm{P}(Y_i|P_i, \mathbf{W_i}) = \mathrm{P}(Y_i|P_i)$$

applies.

By taking the likelihood expressions in (2.2), (2.3) and (2.4) into (2.5), the joint marginal likelihood of observing the disease outcome status and the exposure surrogates is

$$
\begin{aligned}
& \mathscr{L}(Y_1, Y_2, \ldots, Y_k, \mathbf{W_1}, \mathbf{W_2}, \ldots, \mathbf{W_k}; \theta_0, \theta_1, \theta, \gamma) \\
= & \prod_{i=1}^{k} \mathrm{P}(Y_i, \mathbf{W_i}) \\
= & \prod_{i=1}^{k} \int_0^1 [(\frac{\exp(\theta_0 + \theta_1 p_i)}{1 + \exp(\theta_0 + \theta_1 p_i)})^{Y_i} (\frac{1}{1 + \exp(\theta_0 + \theta_1 p_i)})^{1-Y_i}] \\
& [p_i^{\sum_{j=i}^{n} w_{ij}} (1 - p_i)^{n - \sum_{j=i}^{n} w_{ij}}][\frac{p_i^{\frac{\exp(Z_i^T\gamma)}{\theta(1+\exp(Z_i^T\gamma))} - 1} (1 - p_i)^{\frac{1}{\theta(1+\exp(Z_i^T\gamma))} - 1}}{B(\frac{\exp(Z_i^T\gamma)}{\theta(1+\exp(Z_i^T\gamma))}, \frac{1}{\theta(1+\exp(Z_i^T\gamma))})}] dp_i \quad (2.6)
\end{aligned}
$$

The integration in the likelihood (2.6) can be numerically calculated, e.g., via the SAS QUAD function. The maximum likelihood estimates may be obtained by maximizing the likelihood in (2.6), using a Newton-Raphson approach. We carry out the optimization routine using SAS/IML (SAS Institute Inc (2001), SAS Institute Inc (2008)).

Further considering covariates in the model, let U_i be a covariate that needs to be controlled in the TDM. Hence, the likelihood of observing disease outcome Y_i given the exposure probability becomes

$$
\mathrm{P}(Y_i | P_i = p_i) = (\frac{\exp(\theta_0 + \theta_1 p_i + \theta_2 U_i)}{1 + \exp(\theta_0 + \theta_1 p_i + \theta_2 U_i)})^{Y_i} (\frac{1}{1 + \exp(\theta_0 + \theta_1 p_i + \theta_2 U_i))})^{1-Y_i} \quad (2.7)
$$

Note that U_i could as easily be a vector of covariates, and all as a subset of the U_i's might be included among the Z_i's accounted for in the EDM, depending on the research interests.

2.3 REAL-LIFE EXAMPLE

Returning to the BWIS example introduced in 1.1.1, recall that the outcome was whether or not a worker's child developed a congenital cardiovascular TAPVR. The exposure to lead was assessed by three industrial hygeniests. A total of 560 (53 cases and 507 controls) infants' paternal exposure assessments and corresponding disease status indicators were available. Among all the subjects, 391 (70%) were evaluated as not exposed by all three industrial hygeniests; 23 (4%) were assessed as exposed by one industrial hygienist but determined unexposed by the others; 75 (13%) were evaluated as exposed by two industrial hygienists but considered as unexposed by the other; and 71 (13%) cases were assessed as exposed by all.

To demonstrate incorporation of covariates, race of infants in two categories (black, white/other) was considered both in the TDM and EDM, indicating that different race could have different probabilities of developing disease and of exposure to lead.

Since low birth weight is usually highly correlated with birth defects, researchers also take low birth weight as a secondary research outcome in the BWIS (Min et al. (1996)). Min et al. (1996) selected research subjects among the eligible controls of the study, after excluding twins, infants with chromosomal abnormalities, syndromes, or organ abnormalities, and infants of race other than black or white. A total of 157 subjects each with three industrial hygeniests (1,2, and 3) assessing exposure status, were included in our analysis to demonstrate the proposed methods. Among all the subjects included in the study, 45 (29%) of them had low birth weight. In terms of the exposure assessment, both industrial hygienist 1 and 2 assessed 78 (50%) subjects as exposed and 79 (50%) as unexposed, but they did not agree on an individual by individual basis. The third industrial hygienist assessed 119 (76%) as exposed and 38 (24%) as unexposed.

Both models with and without covariates will be presented. For the outcome of TAPVR, we controlled for infants' race (white/other and black) as a covariate in the

TDM part. However, race of infants is not included in the BB part in the results to be presented, due to its insignificance based on model comparisons. Infants' race, maternal smoking (packs per day) and maternal weight gain were controlled for when modeling the outcome of low birth weight. The covariates are included in the TDM part, but since the they are all non-significant in the BB part, the results to be presented are the ones without incorporating the covariates in the BB part. Analysis results based on a naive method taking the probability of exposure of each subject as the observed proportion of exposures among the assessments within a subject (see section 2.2.2) were studied and comparied to the proposed method based on the beta-binomial model (BB).

Estimated odds ratios and 95% CIs for the effect of probability of exposure upon each disease outcome, with and without controlling for covariates, are presented in table 2.1.

Note that the naive method dramatically attenuated the effect of exposure probability upon the outcome of both TAPVR and low birth weight, no matter whether controlling for the racial effect or not. According to the results from the beta-binomial model without covariates, a higher probability of parental exposure to lead would coincide with a statistically significantly higher probability of both delivering a low birth weight infant and one with possible TAPVR. However, the naive method failed to capture this significance.

Controlling for the exposure probability, the two methods estimated the racial effect similarly with the beta-binomial model providing a narrower confidence interval, for both the TAPVR and the low birth weight outcome. Taking TAPVR as the outcome alone, both methods suggested that infants with race of white/other had a higher probability of developing TAPVR than those with race of black with the same extent of parental exposure to lead. Considering low birth weight as the outcome alone, the effects of maternal weight gain were also similar by the two methods, while

Table 2.1: Estimated Odds Ratios with 95% CIs

Factors	Results based on BB model		Results based on naive model	
	OR	95% CI	OR	95% CI
Results without controlling covariates				
TAPVR as the outcome				
Probability of Lead Exposure	2.03	1.07, 3.87	1.70	0.89, 3.16
Low birth weight as the outcome				
Probability of Lead Exposure	4.66	1.13, 19.40	2.34	0.74, 7.40
Results based on controlling covariates				
TAPVR as the outcome				
Probability of Lead Exposure	2.04	0.90, 4.59	1.05	0.58, 1.92
Race (white/other vs. black)	1.98	1.30, 3.02	1.98	1.11, 3.52
Low birth weight as the outcome				
Probability of Lead Exposure	4.18	0.41, 43.05	2.19	0.06, 76.89
Race (black vs. white)	3.60	0.63, 20.58	3.55	0.56, 22.66
Maternal smoking	1.72	0.72, 4.06	1.70	0.67, 4.36
Maternal weight gain	0.96	0.87, 1.06	0.96	0.87, 1.07

the beta-binomial model provided slightly narrower confidence intervals.

2.4 SIMULATION STUDY

Simulation experiments were conducted to compare the perfomance of the estimates based on the proposed method and the naive method, where the latter substitutes the proportion of ‘exposure’ ratings among all the assessments for a subject in place of the true probability of exposure. Since the number of replicates (here, the number of industrial hygienists) is critical for the analysis as well as for model convergence, we conducted the simulation experiments based on three sets of simulations where the number of replicates was taken to equal 3, 7 and 25. The results are shown in table 2.2.

As we can see, the proposed method performed uniformly better than the naive method, in terms of smaller bias in the estimates. The empirical standard deviations and the mean of the estimated standard errors were reasonably close to one another for the proposed method and the 95% coverages were reasonably close to 95 %. As the number of replicates increased from 3 to 25, the performance of the estimates was also enhanced correspondingly: the bias decreased, the empirical standard errors and the mean of the estimated standard errors got closer, and the coverage became closer to 95%. We can also see that when the number of replicates was 25, the performance of the naive method was quite close to the true parameter values, although not as close as the proposed method. This is reasonable because the more industrial hygienists there are, the more information we obtain via the observed proportion making ‘exposed’ assessments.

Table 2.2: Simulation results based on 500 simulated datasets, sample size = 300, and (θ_0, θ_1, α, β) = (-2.85, 3.50, 1.00, 2.00)

Para-meters	True values	Number of replicates	Results based on BB model			
			Mean Est.	Emp.SE	Mean Est. SE	95% coverage
θ_0	-2.85	3	-2.92	0.57	0.51	0.91
θ_1	3.50		3.61	1.25	1.12	0.92
		7	-2.88	0.44	0.40	0.92
			3.55	0.90	0.84	0.91
		25	-2.86	0.36	0.35	0.95
			3.51	0.75	0.72	0.94
Para-meters	True values	Number of replicates	Results based on naive model			
			Mean Est.	Emp.SE	Est. SE	95% coverage
θ_0	-2.85	3	-2.15	0.26	0.31	0.32
θ_1	3.50		1.68	0.46	0.51	0.21
		7	-2.43	0.30	0.32	0.73
			2.41	0.56	0.57	0.51
		25	-2.71	0.32	0.33	0.93
			3.11	0.65	0.63	0.91

Chapter 3

ASSESSING THE ASSOCIATION BETWEEN A HEALTH OUTCOME AND MISCLASSIFIED EXPOSURE

3.1 INTRODUCTION

The previous chapter proposed methods identifiying the association between a disease outcome and the probability of exposure. When possible, however, researchers would generally be more interested in determining the relationship between the disease outcome and the true unobserved exposure, instead of the probability. This is also a question more reasonable to address in the sense that a high probability of exposure does not necessarily imply that an individual was actually exposed. Therefore, this chapter will focus on considering the true underlying exposure status as the primary interest, in linking to the disease outcome. The intent is to make a natural step forward from the measurement error considerations of chapter 2, to consider misclassification issues related to the lack of a reliable true exposure indication.

As the study design and data have been introduced briefly in section 1.1.1, we re-emphasize the main characteristics of the data in the following:

- Binary outcome and binary exposure. The extension of the method to handle error-prone categorical exposure data with multiple levels would be straightforward, but for demonstration purposes we focus upon binary exposures and outcomes. Nevertheless, situations with continuous health outcomes will also be discussed briefly.

- Repeated exposure information available with possible within-subject correlation given the true exposure. More than one industrial hygenist assessed the exposure status. Given the true exposure status, the industrial hygienists' assessments might remain correlated due to their blindness to the actual working environment. For example, we might expect a residual connection, if one of the workers worked overtime unrecorded in an exposed environment. Then the industrial hygienists' assessments might all be biased due to their blindness to this fact.

- Exposure misclassified without a gold standard. The moderate agreement (kappa=0.5-0.6) in the industrial hygienists' assessments implies misclassifications of some of the exposure assessments, if not all of them (Correa et al. (2006)).

The following section 3.2 will propose a model in which the industrial hygienists' assessments are assumed to be independent, analogous to Liu & Liang (1991), and then extend the method to the situation where the industrial hygienists' assessments remain correlated given the true exposure status of a subject. A method for modelling continuous outcomes will also be proposed. Identifiability issues will be discussed in the model section, followed by some computational details. In section 3.3, we will get back to the BWIS example again, and apply the proposed models. Interesting results with covariates in the model will be presented. Finally, simulation results will be shown in section 3.4, followed by some brief discussions.

3.2 MODEL

3.2.1 Model structure

Following the general structure introduced by Clayton (1992) (details were reviewed in 1.2.1), the modelling of the association between the disease outcome and the true unobserved exposure status can be formulated into the three-model TDM, MEM and EDM paradigm:

TDM: $\text{logit}[\text{P}(Y_i = 1|X_i)] = \theta_0 + \theta_1 X_i$ for binary TDM

or $\text{E}[Y_i|X_i] = \theta_0 + \theta_1 X_i$ for linear TDM

MEM: $\text{P}(\mathbf{W_i}|X_i = x)$ relates observed surrogate ($\mathbf{W_i}$) to true exposure status (X_i)

EDM: $\text{P}(X_i = x) = \tau_x \quad (x = 0, 1)$

As pointed out in section 2.2.2 for modeling the exposure probability, additional covariates can be added to both TDM and MEM when the true exposure status is the independent variable. Details on the modeling strategies will be introduced in the following sections.

3.2.2 Marginal Likelihood

Binary TDM

Conditional independence Analogous to the notations in section 1.2.3, let $W_i = (w_{i1}, w_{i2}, \ldots, w_{in})$ be the observed exposure assessments from n industrial hygienists for subject i. Suppose the true unobserved error-free binary exposure status is X_i for subject i and Y_i is the disease outcome status. Define

$$\pi_{jx} = \text{P}(w_{ij} = 1|X_i = x) \quad (j = 1, 2, \ldots, n, x = 0 \text{ or } 1)$$

which is assumed to be identical across all k subjects for the same (jth) observation. Assumptions could be relaxed to allow heterogeneous π_{jx}'s for different subjects. However, to make the model identifiable, we would tend to need an unnecessarily large number of replicates to enable the heterogeneity assumption.

The sensitivity and specificity for the jth assessment are

$$\pi_{j1} = \mathrm{P}(w_{ij} = 1|X_i = 1)$$

and

$$1 - \pi_{j0} = 1 - \mathrm{P}(w_{ij} = 1|X_i = 0)$$

respectively.

Therefore the probability of observing w_{ij} for the ith subject's jth measurement, given the true underlying exposure x, is

$$\begin{aligned} \mathrm{P}(w_{ij}|X_i = x) &= \mathrm{P}(w_{ij} = 1|X_i = x)^{w_{ij}}(1 - \mathrm{P}(w_{ij} = 1|X_i = x))^{1-w_{ij}} \\ &= \pi_{jx}^{w_{ij}}(1 - \pi_{jx})^{1-w_{ij}} \end{aligned}$$

Under the conditional independence assumption, i.e., all the n observations are mutually independent within the ith subject given his or hers true exposure status, the probability of observing $\mathbf{W_i}$ given the true unobserved value of X_i is

$$\mathrm{P}(W_i|X_i = x) = \prod_{j=1}^{n} \pi_{jx}^{w_{ij}}(1 - \pi_{jx})^{1-w_{ij}} \tag{3.1}$$

Let the relationship between disease outcome and true exposure status follow a logistic model, i.e.,

$$\mathrm{logit}[\mathrm{P}(Y_i = 1|X_i)] = \log[\frac{\mathrm{P}(Y_i|X_i)}{1 - \mathrm{P}(Y_i|X_i)}] = \theta_0 + \theta_1 X_i$$

or equivalently

$$\mathrm{P}(Y_i = 1|X_i) = \frac{\exp(\theta_0 + \theta_1 X_i)}{1 + \exp(\theta_0 + \theta_1 X_i)},$$

where θ_0 and θ_1 are the primary parameter of interest.

Summing over the values that the unobserved X_i could possibly take, the joint likelihood of observing both the disease outcome and the observed exposure assessments for subject i will be

$$\begin{aligned}
\mathrm{P}(Y_i, \mathbf{W_i}) &= \sum_{x=0}^{1} \mathrm{P}(Y_i|X_i = x, \mathbf{W_i})\mathrm{P}(\mathbf{W_i}|X_i = x)P(X_i = x) \qquad (3.2) \\
&= \sum_{x=0}^{1} \mathrm{P}(Y_i|X_i = x)\mathrm{P}(\mathbf{W_i}|X_i = x)\mathrm{P}(X_i = x) \qquad (3.3)
\end{aligned}$$

Equation (3.3) holds under the non-differential measurement error assumption that the surrogate is independent of the outcome, given the true unobserved exposure, i.e.,

$$\mathrm{P}(Y_i|X_i, \mathbf{W_i}) = P(Y_i|X_i)$$

This is an intuitively reasonable assumption applied here in the sense that the industrial hygienists' assessments should be independent of the disease outcome, if the true exposure is known. In other words, the industrial hygienists know nothing better than the truth.

By expanding the assumed model in (3.3), the probability of observing Y_i and $\mathbf{W_i} = (w_{i1}, w_{i2}, \ldots, w_{in})$ is

$$\mathrm{P}(Y_i, \mathbf{W_i}) = \sum_{x=0}^{1} \tau_x \mathrm{P}(\mathbf{W_i}|X_i = x)[\frac{\exp(\theta_0 + \theta_1 x)}{1 + \exp(\theta_0 + \theta_1 x)}]^{Y_i}[\frac{1}{1 + \exp(\theta_0 + \theta_1 x)}]^{1-Y_i}$$

where $\mathrm{P}(\mathbf{W_i}|X_i = x)$ is derived in (3.1) and $\tau_x = P(X = x)$.

Finally the joint marginal likelihood of observing the disease and exposure over all subjects will be

$$\begin{aligned} & \mathscr{L}(Y_1, Y_2, \ldots, Y_n, W_1, W_2, \ldots, W_n; \theta_0, \theta_1, \pi_{jx}, \tau_x) \qquad (3.4) \\ = & \prod_{i=1}^{k} \{ \sum_{x=0}^{1} \tau_x \mathrm{P}(\mathbf{W_i}|X_i = x) [\frac{\exp(\theta_0 + \theta_1 x)}{1 + \exp(\theta_0 + \theta_1 x)}]^{Y_i} [\frac{1}{1 + \exp(\theta_0 + \theta_1 x)}]^{1-Y_i} \} \\ & \qquad (3.5) \end{aligned}$$

again defining $\mathrm{P}(\mathbf{W_i}|X_i = x)$ as in (3.1).

The parameters are estimated by numerically maximizing the likelihood (3.5). Computational complications will be discussed in detail in section 3.2.4. This analysis represents a fully likelihood-based analogous to the quasi-likelihood-based proposed by Liu & Liang (1991).

Conditional dependence As we discussed before, in some cases the assessments from different industrial hygienists may remain correlated, given the real exposure status. Following a model proposed by Qu et al. (1996) , the common characteristics among the industrial hygienists' evaluations given the true unobserved exposure status can be captured by a latent random variable T. As reviewed in section 1.2.3, Qu et al. (1996)'s model specifies

$$\mathrm{P}(w_{ij}|X_i = x, T = t) = F(a_{jx} + b_{jx} t) \qquad (3.6)$$

where $T \sim N(0, 1)$, and F could be the probit, logit or any other function appropriate for categorical outcomes.

General expressions for sensitivity and specificity were given in equations (1.9) and (1.10). In particular and as previously noted, if F is the probit function, explicit forms for sensitivity and specificity can be obtained as $\Phi(a_{j1}/\sqrt{1 + b_{j1}^2})$ and $\Phi(-a_{j0}/\sqrt{1 + b_{j0}^2})$, $j = 1, 2, \ldots, n$, where $\Phi(.)$ is the cumulative probability function for the standard normal distribution. However, if F takes a logit function form, one would have to evaluate the integrals in (1.9) and (1.10) numerically to obtain the sensitivity and specificity. In our real life example, as well as in simulation studies, we elect to use the probit link to stay consistent with Qu et al. (1996), and because we have found it equally stable in performance as the logit link in terms of convergence of the likelihood.

To review, Qu's model stipulates that the probability of observing $\mathbf{W_i}$, given the true unobserved exposure as x for the ith subject, is

$$\mathrm{P}(\mathbf{W_i}|X_i = x) \quad = \quad \int_{-\infty}^{\infty} \prod_{j=1}^{n} F(a_{jx} + b_{jx}t)^{w_{ij}} (1 - F(a_{jx} + b_{jx}t))^{1-w_{ij}} f(t) dt \quad (3.7)$$

Our extension is to link Qu's model for repeated error-prone binary exposure assessments to a model that relates a health outcome to true binary exposure status. Taking a logistic model to describe the relation between disease and true exposure, the likelihood of observing both the disease outcome and the observed exposure assessments for subject i will be

$$\mathrm{P}(Y_i, \mathbf{W_i}) = \sum_{x=0}^{1} \tau_x \mathrm{P}(\mathbf{W_i}|X_i = x)[\frac{\exp(\theta_0 + \theta_1 x)}{1 + \exp(\theta_0 + \theta_1 x)}]^{Y_i}[\frac{1}{1 + \exp(\theta_0 + \theta_1 x)}]^{1-Y_i}$$

with $\mathrm{P}(W_i|X_i = x)$ defined in (3.7).

Finally the joint marginal likelihood for observing the disease and exposure assessments for all subjects in the study will be

$$\begin{aligned} & \mathscr{L}(Y_1, Y_2, \ldots, Y_k, \mathbf{W_1}, \mathbf{W_2}, \ldots, \mathbf{W_k}; \theta_0, \theta_1, a_{jx}, b_{jx}, \tau_x) \\ = & \prod_{i=1}^{k} \{ \sum_{x=0}^{1} \tau_x \mathrm{P}(\mathbf{W_i}|X_i = x)[\frac{\exp(\theta_0 + \theta_1 x)}{1 + \exp(\theta_0 + \theta_1 x)}]^{Y_i}[\frac{1}{1 + \exp(\theta_0 + \theta_1 x)}]^{1-Y_i} \} \end{aligned} \quad (3.8)$$

Integrations in (3.7) can be approximated by Gaussian-Hermite quadrature, which was reviewed in detail in 1.2.5. Estimates can be obtained by numerically maximizing the log-likelihood corresponding to (3.8). Computational issues and complications will be discussed in the following sections.

Linear TDM

For continuous outcomes, a linear TDM should be attached to the latent class model instead of a logit one. For example, assume the continuous outcome Y_i for subject i follows a normal distribution with conditional mean $\theta_0 + \theta_1 X_i$ and standard deviation σ, i.e.,

$$Y_i = \theta_0 + \theta_1 X_i + \epsilon_i$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Therefore, the probability density function corresponding to of $Y_i|X_i$ would be

$$\mathrm{P}(Y_i|X_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-(Y_i - \theta_0 - \theta_1 X_i)^2/(2\sigma^2)) \tag{3.9}$$

Hence, replacing $[\frac{\exp(\theta_0+\theta_1 x)}{1+\exp(\theta_0+\theta_1 x)}]^{Y_i}[\frac{1}{1+\exp(\theta_0+\theta_1 x)}]^{1-Y_i}$ in likelihood functions (3.5) and (3.8) with (3.9), we obtain the likelihood for the model with linear TDM under both the conditional independence and conditional dependence assumptions. To obtain the full likelihood, the other components, such as $\mathrm{P}(\mathbf{W_i}|X_i = x)$ and $\mathrm{P}(X_i = x)$ remain the same as previously defined.

3.2.3 Identifiability issues

The basic identifiability ideas for latent class models were reviewed in section 1.2.3. In our model, we need to ensure that the sum of sensitivity and specificity is greater than 1, according to Theorem 1 from 1.2.3. Furthermore, to match the corresponding second criteria in Theorem 1, we need to be cautious as the number of parameters and the degrees of freedom change in our models. Since we have the outcome variable attached to the regular latent class model, we have two more cells to be taken into account either 0 or 1. Therefore, the degrees of freedom will be $2\prod_{j=1}^{n}(l_j+1)-1=2^{n+1}-1$ where l_j can only take the value 0 or 1 in our model. So the second condition in the theorem needs to be

$$2^{n+1}-1 \geq 2n+1 \tag{3.10}$$

in our model with the conditional independence assumption and

$$2^{n+1}-1 \geq 4n+1 \tag{3.11}$$

in the model with the conditional dependence assumption.

3.2.4 Estimation

We maximizing the likelihood functions with restrictions via trust region optimization (Conn et al. (1987)) with nonadaptive quadrature, where the quadrature points are centered at zero for each of the random effects and the current random-effects variance matrix is used as the scale matrix. Although many authors suggest the superiority of the adaptive quadrature (Liu & Pierce (1994), Pinheiro & Bates (1995), Lesaffre & Spiessens (2001)), empirical results suggest that the nonadaptive quadrature works better in our particular situation.

To enhance the model stability and speed up the convergence for the purpose of simulations, we assume the variances of the random effects are equal for all the assessments and regardless of the true exposure status, i.e., $b_{11}=b_{21}=\ldots=b_{j1}=b_{12}=b_{22}=\ldots=b_{j2}=b$; where $j=1,2,\ldots,n$. Note that this confers needed stability in the implementation of Qu et al. (1996)'s model when it is linked to a TDM such as in (3.8), but that differential misclassification can still be accomodated because the a_{jx} parameters are allowed to vary across true disease status and possibly across different raters or assessment methods.

The algorithm was implemented via the SAS procedure NLMIXED, which allows specification of a general log-likelihood computed via numerically integrating out normal random effects. An alternative program in SAS/IML was utilized as a second check (SAS Institute Inc (2001), SAS Institute Inc (2008)).

To test whether the repeated measurements are conditionally dependent, a likelihood ratio test is conducted taking a mixture chi-square distribution as the reference distribution for testing the variance components of the random effects, which are the b_{jx}'s in our case (Self & Liang (1987), Stram & Lee (1994), Stram & Lee (1995)). For an example, to test the hypothesis of q random effects vs. $q + 1$ random effects, the asymptotic null distribution of the -2 times log likelihood ratio is a mixture with equal weights 0.5 for χ^2_{q+1} and χ^2_q (Verbeke & Molenberghs (2000)).

3.3 REAL-LIFE EXAMPLE

To demonstrate the methodology proposed in this chapter, we also utilize the BWIS example introduced in 1.1.1 and 2.3. Similar to the example in chapter 2, the outcomes of interest are whether the infant developed a congenital cardiovascular TAPVR, and whether or not the birth weight was low. Both models, assuming conditional independence and conditional dependence, were fitted and compared.

In our case, the estimated random effect variance components b are quite close to zero with other estimates fairly similar for models with and without random effects, for both outcomes TAPVR and low birth weight (results not shown). Taking a 50:50 mixture of χ_0^2 and χ_1^2 as the reference distribution, a likelihood ratio test reveals no siginificance between the reduced model without random effect and the one with, for both outcomes. Therefore, we believe in this particular example the industrial hygienists' assessments are conditional independent. In the following tables reporting either the estimated sensitivity and specificity, or the estimated effects between the exposure, covariates and the adverse health outcomes, we focus on models with the conditional independence assumption.

Table 3.1 presents estimated sensitivities (SEN) and specificities (SPC) for individual IHs, assuming five possible situations: (1) the SEN and SPC for all three IHs are the same; (2) IH2 and IH3 share the same SEN and SPC and IH1 differs from them; (3) IH1 and IH3 share the same SEN and SPC and IH2 differs from them; (4) IH1 and IH2 share the same SEN and SPC and IH3 differs from them; (5) All three IHs have different SENs and SPCs.

Table 3.1 suggests that all the industrial hygenists have specificity close to 1, which implies that they are accurate in catching the truely unexposed. In another word, one was libely to be exposed when he was assessed as exposed by the IHs. In terms of catching the truely exposed, IH1 and IH2 are more accurate than IH3, who would approximately assess 50% exposed as unexposed but is quite accurate in assessing true unexposure. IH1 and IH2 are similar and both accurate in catching both truely exposed and unexposed. Table 3.1 also suggests that if discrepancies occured among the IH's assessments, we may consider IH2's opinion more seriously than the others, since IH2 had the highest estimated sensitivity and specificity among all. Based on Chi-square tests, model with the assumption that IH1 and IH2 are the same but differ from IH3 is in favor of the others. Therefore, this assumption is postulated throughout the rest of the modelling procedure.

Both models with adverse health outcome as TAPVR and low birth weight will be presented in 3.2. For the outcome of TAPVR, controlled for race (white/other or black). We controlled for race, maternal smoking (packs per day) and maternal weight gain for the outcome of low birth weight. Covariates are initally controlled in TDM as well as in sensitivity and specificity, i.e., assume different sensitivity and specificity for different race of infants etc. However, the estimates are similar to each other for different race of infants (results not shown), so the estimates reported in the tables are the ones controlling for covariates only in the TDM part. Naive models, assuming the subject to be exposed if two or more of the three industrial hygienists thought he was exposed, and unexposed if one or less than one hygienist assessed exposure, will also be fitted and compared. The odds ratio and the 95% CI for the effect of true unobserved binary exposure on the adverse outcome, with and without controlling for covariates, are presented in Table 3.2.

As seen in the table, the lead exposure effect is attenuated in the naive model for both outcomes, with or without controlling for covariates. Lead exposure plays no significant effect on both TAPVR and low birth weight, with or without controlling for covariates. As to the covariates, race of infants is not a significant risk factor for developing TAPVR. However, the proposed model suggests that infants of white/other tend to have a significantly higher odds of born low birth weight than infants of black, given the same lead exposure status, mother's smoking and maternal weight gain. The naive model implies significance in infants' race but suggests an opposite trend. Moreover, the proposed model suggested that the more packs a mother smokes a day, the higher the odds of her delievering an infant with low birth weight, given the same lead exposure status, infant race and maternal weight gain. Also suggested by the proposed model, that with the increase of maternal weight gain, the odds of developing a low birth weight infant decreases significantly given the same lead exposure status, infant race and maternal smoking. However, opposite trends suggested by the naive model for the association between delievering a low birth weight infant and maternal smoking, maternal weight gain seem contradicting to expectations.

Table 3.1: Estimated sensitivities (SEN) and specificities (SPC) for outcome TAPVR

	IH1 IH2 IH3 are all equal		IH1 differs from IH2 and IH3		IH2 differs from IH1 and IH3		IH3 differs from IH1 and IH2		IH1 IH2 IH3 are all different	
	Estimate	StdErr	Estimate	StdErr	Estimate	StdErr	Estimate	StdErr	Estimate	StdErr
SEN for IH1	0.76	0.01	0.95	0.02	0.71	0.03	0.98	0.005	0.96	0.03
SPC for IH1	1.00	–	1.00	–	0.98	0.002	0.97	0.005	0.96	0.01
SEN for IH2	0.76	0.01	0.69	0.01	1.00	–	0.98	0.005	0.99	0.01
SPC for IH2	1.00	–	0.99	0.004	0.99	0.02	0.97	0.005	0.99	0.01
SEN for IH3	0.76	0.01	0.69	0.01	0.71	0.03	0.50	0.04	0.49	0.04
SPC for IH3	1.00	–	0.99	0.004	0.98	0.002	1.00	–	1.00	–
-2loglik†	1585.6		1537.2		1531.5		1434.5		1429.1	

† the value of $-2 \times log(likelihood)$

Table 3.2: Estimated Odds Ratio with 95% CI

Factors	Results based on proposed model		Results based on naive model	
	OR	95% Wald CI	OR	95% Wald CI
Results for not controlling for covariates				
TAPVR as the outcome				
Lead Exposure	1.22	0.65, 2.29	0.72	0.39, 1.33
Low birth weight as the outcome				
Lead Exposure	2.05	0.93, 4.53	0.74	0.37, 1.50
Results based on controlling for covariates				
TAPVR as the outcome				
Lead Exposure	1.49	0.79, 2.82	0.73	0.39, 1.34
Race (white/other vs. black)	1.82	0.86, 3.84	0.51	0.24, 1.07
Low birth weight as the outcome				
Lead Exposure	2.05	0.86, 4.86	0.76	0.36, 1.63
Race (black vs. white)	3.63	1.65, 7.98	0.28	0.13, 0.61
Maternal smoking	1.57	1.08, 2.26	0.65	0.46, 0.93
Maternal weight gain	0.97	0.94, 0.99	1.03	1.01, 1.06

3.4 SIMULATION STUDY

In this section, we present simulation results for both the linear TDM when the outcome is continuous, and the binary TDM with a logit link when the outcome is binary, based on the joint likelihood including random effects as in model (3.5). For stable simulation results, we increased the number of replicates to 7 and added reasonable assumptions that (1) the variances of the random effects are equal for both truely exposed and unexposed cases; (2) the sensitivities and specificities are equal across different IHs (the first situation shown in Table 3.1), or equivalently, $a_{11} = a_{21} = \ldots = a_{j1} = a_1$; $a_{12} = a_{22} = \ldots = a_{j2} = a_2$; $b_{11} = b_{21} = \ldots = b_{j1} = b_{12} = b_{22} = \ldots = b_{j1} = b$; where j = 1,2,...,n

Simulation results for continuous outcomes with the conditional dependence assumption, based on the model described in 3.2.2, are presented in Table 3.3. Since the empirical standard deviations are larger than the mean of the estimated standard errors, we believe the estimated standard errors are not accurate. Therefore, profile-likelihood-based confidence intervals are calculated for θ_1 to provide more accuracy. The mean widths for the confidence intervals is 0.65 and the 95% coverage from profile-likelihood-based confidence intervals for θ_1 is presented in the table. Naive model results are also presented to show the extent of attenuation.

Simulation results for a binary outcome with logit link are shown in Table 3.4 with naive model results as comparison. Some of the 500 simulations produce unreasonably large estimated standard errors for $\hat{\theta}_1$. For example, 28 (5.6%) of the estimated standard errors were greater than 10. For demonstrative purposes, we refer to those simulated data with unreasonable estimated standard error for $\hat{\theta}_1$ as 'uninformative' data and those with reasonable estimated standard errors as 'informative' data. The reason that some estimated standard errors are large appears due to the flat profile likelihood over various possible θ_1 values, which suggests the likelihood function itself is not informative in these cases. A comparison of the profile likelihood functions

Table 3.3: Simulation results based on 500 simulated datasets, linear TDM, sample size = 600, number of replicates = 7, and (θ_0, θ_1, SEN, SPC, b, τ_1, σ) = (0.00, 1.00, 0.76, 0.86, 1.00, 0.30, 1.00)

Parameters	True values	Mean Est.	Emp.SD	Mean Est. SE
θ_0	0.00	0.004	0.11	0.06
θ_1†	1.00	1.05	0.23	0.16
θ_1^*‡	*1.00*	*0.73*	*0.10*	*0.10*
SEN	0.76	0.75	0.07	0.05
SPC	0.86	0.85	0.04	0.02
b	1.00	1.03	0.18	0.14
τ_1	0.30	0.29	0.06	0.04
σ	1.00	0.99	0.05	0.04

† The profile-likelihood-based CI coverage for θ_1 is 95.8%

‡ Estimates based on naive model

between the ‘uninformative’ data and ‘informative’ data is presented in Figure 3.1, where the true θ_1 is 1.00. Summary statistics for the parameters under estimation are also shown in Table 3.4 when the ‘uninformative’ simulated datasets were left out.

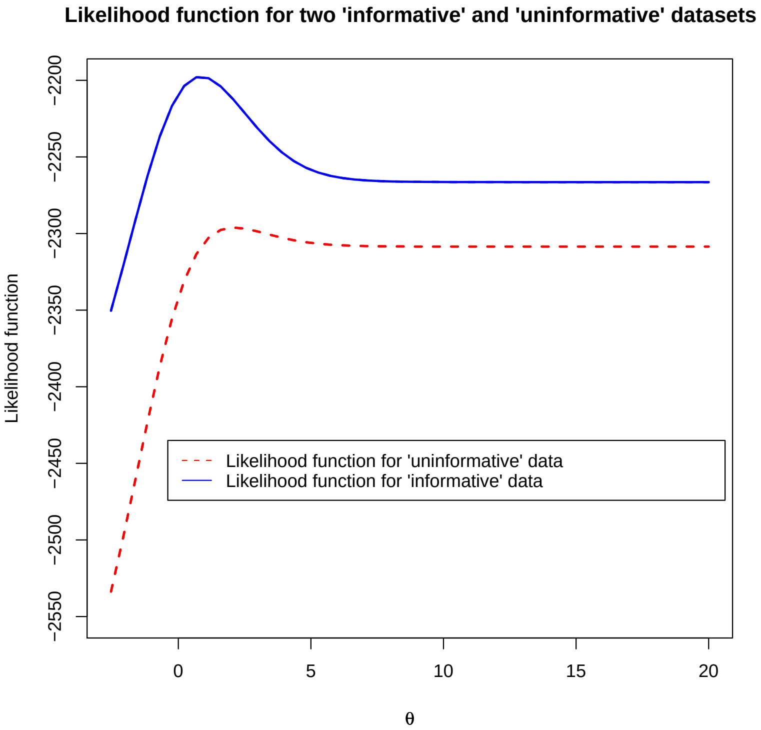


Figure 3.1: Likelihood function for an ‘informative’ and an ‘uninformative’ dataset

Table 3.4: Simulation results based on 500 simulated datasets,binary TDM, sample size = 600, number of replicates = 7, and (θ_0, θ_1, SEN, SPC, b, τ_1) = (0.00, 1.00, 0.76, 0.86, 1.00, 0.30)

Parameters	True values	Mean Est.	Emp.SD	Mean Est. SE
Results from all the 500 simulations				
θ_0	0.00	-0.05	0.75	1.67
θ_1	1.00	1.61	2.26	14.82
SEN	0.76	0.77	0.07	0.06
SPC	0.86	0.84	0.04	0.03
b	1.00	1.04	0.22	0.18
τ_1	0.30	0.29	0.08	0.06
Results from 472 simulations without the ‘uninformative’ cases				
θ_0	0.00	0.005	0.12	0.12
θ_1†	1.00	1.08	0.43	0.48
θ_1^*‡	*1.00*	*0.71*	*0.19*	*0.19*
SEN	0.76	0.77	0.06	0.06
SPC	0.86	0.85	0.04	0.03
b	1.00	1.01	0.19	0.18
τ_1	0.30	0.29	0.07	0.05

† The Wald CI coverage for θ_1 is 95.61%

‡ Estimates based on naive model from 500 simulations

3.5 DISCUSSION

In this section, we briefly comment on the connection between our work and prior work on exploring the exposure-disease association using the probability-of-exposure (POE) information by Satten & Kupper (1993). Both are focused on the examination of exposure-disease relationships and consider the POE as helpful information. We take a parametric approach, in which a subject's true unknown probability of exposure is a latent random variable defined via a beta-binomial model. We then extend the interest from POE-disease relationship to the exposure-disease association by extending latent class models to link with a disease outcome. We include the examination of POE and disease association as an initial aspect of the work with focus on exposure-disease association. In our case, the POE is not required information, or not directly related to the stage modeling of exposure-disease relationship. On the other hand, Satten & Kupper (1993) view the problem from a different aspect and incorporate the POE information directly into the exposure-disease modeling, assuming the POE can be obtained via measurement of a surrogate exposure variable. To meaningfully combine the two methods, one might consider estimating POE via posterior mean estimates from the beta-binomial model, which can be considered as future work.

Chapter 4

CORRECTING FOR MISCLASSIFICATION IN TESTING DIFFERENCES IN GROUP MEANS USING INTERNAL OR EXTERNAL VALIDATION DATA

4.1 INTRODUCTION

The two sample t-test to assess whether the means from two populations are significantly different is clearly among the most fundamental of all statistical methods. The corresponding approach for testing the differences in means from more than two populations is the method of analysis of variance (ANOVA) (D'Agostino et al. (2005)). For almost a century, these methods have been enriched in various ways and adapted to different practical problems (e.g., Welch (1938), Roy & Gnanadesikan (1959), Hasler et al. (2008), Miao & Chiou (2008), Dutilleul et al. (2008)). However, when observations are potentially misclassified in the study leading to an incorrect attribution of group memberships, appropriate adjustments would be required to the traditional two sample t-test or ANOVA setting for valid inferences. Despite a vast literature on misclassification problems, the discussion of the two group continuous outcome case and its extensions to the one-way ANOVA setting with the use of validation data has been limited.

When a "gold standard" measure of exposure is available but costly, a typical approach of adjusting for misclassification in exposure is to obtain a validation sample, in which the true sample population attributions are available. Based on the data sources, the validation data can be categorized into two main types, i.e., internal and external. In our setting, internal validation requires a random subsample of the main study, where the continuous response values and both the true and misclassified sample population attributions are available. External validation instead suggest an independent sample seperate from the main study, where typically the continuous response values would be unavailable (Carroll et al. (2006)). Although less costly, external validation data is generally less preferable to internal validation data because the misclassification structure and underlying assumptions in external validation data may not be "transportable" to the main study. More discussions on "transportability" can be found in Lyles et al. (2007) and Carroll et al. (2006). In addition to using internal or external validation data alone, recent researches also promotes designs combining both external and internal validation data. Detailed discussion relevant to the combined study designs can be found in (Greenland (1988), Thurston et al. (2003), Lyles et al. (2007).

In this paper, we will first focus on introducing a method to correct for two sample t-test results for misclassification of group membership situations where either internal or external validation data are available. For the internal validation data case, we examine and compare the performances of likelihood-based estimators and closed-form weighted average estimators. In both cases, the generalizations to correct the misclassifications of membership in multiple groups in ANOVA are also discussed. Simulation results will be provided to demonstrate the methods. These methods also serve as a methodological transition into chapter 5, in which a more complex but related misclassification problem is motivated by studies of menstrual cycle length.

4.2 MODEL

Notation and Assumptions

Testing mean differences across two groups

To compare the differences in means for two groups, let x denote the population group indicator so that $x = 1$ if the sample is from the first group, and $x = 0$ if the sample is from the other. Suppose x^* is the group indicator prone to misclassification. We let Y represent the continuous random variable of interest, and let $SE = \mathrm{P}(x^* = 1|x = 1)$ and $SP = \mathrm{P}(x^* = 0|x = 0)$ represent the sensitivity and the specificity of the misclassified group indicator, respectively. Define $\pi = \mathrm{P}(x = 1)$ and $\pi^* = \mathrm{P}(x^* = 1)$. Note here that π can be represented equivalently as $\pi = \frac{\pi^*+SP-1}{SE+SP-1}$, where π^* can be estimated from the main study by $\frac{n_1^*}{n_1^*+n_0^*}$ (n_j^* is the number of subjects with $x^* = j$ ($j = 0$ or 1)). As with the two sample t-test, we assume the conditional mean of the response given the true group indicator follows the model

$$\mathrm{E}[Y|x] = \delta_0 + \delta_1 x \quad (x = 0 \text{ or } 1)$$

In addition to testing whether the two sample means are equal, i.e., whether $\delta_1 = 0$, we also aim in this paper to provide valid and efficient point and confidence interval estimators for the differences in group mean δ_1 after adjusting for misclassification. The data observable in real study is (y, x^*). So ignoring misclassification the naive model will be $\mathrm{E}[Y|x^*] = \delta_0^* + \delta_1^* x^* \quad (x^* = 0 \text{ or } 1)$. We will discuss and compare the two models in later presentations.

If the misclassification structures are the same in the external validation data as the main study, i.e., they are “transportable”, we can use external validation data to adjust for misclassification in exposure. Table 4.1 displays the typical data layout from an external validation study.

Table 4.1: External Validation Study for two groups Data Layout

x	$x^* = 1$	$x^* = 0$	Total
1	n_{11}	n_{10}	$n_{1.}$
0	n_{01}	n_{00}	$n_{0.}$
Total	$n_{.1}$	$n_{.0}$	$n_{..}$

We impose the assumption of nondifferential misclassification (Thomas et al. (1993)) throughout the paper, which implies constant sensitivity and specificity for various outcomes y, i.e., $SE_y = SE$ and $SP_y = SP$ for all y, where $SE_y = \mathrm{P}(x^* = 1|x = 1, Y = y)$ and $SP_y = \mathrm{P}(x^* = 0|x = 0, Y = y)$.

Testing the differences in means across more than two groups

Suppose there are $k + 1$ groups in the one-way ANOVA setting. Let x_1, x_2, $\ldots$, x_k be the group indicators that if the sample is from population i ($i \neq k + 1$), then x_i=1 and $x_r = 0$ ($r \neq i$); if the sample is from population $k + 1$, then $x_1 = x_2 = \ldots = x_k = 0$. Again, x_1^*, x_2^*, $\ldots$, x_k^* will be the corresponding group indicators prone to misclassification. Also define $SE_{ij} = \mathrm{P}(x_i^* = 1|x_j = 1)$, $\pi_i = \mathrm{P}(x_i = 1)$ and $\pi_i^* = \mathrm{P}(x_i^* = 1)$, where ($i = 1, \ldots, k,$ and $j = 1, \ldots, k + 1$). Similar to the case of two groups, we assume the mean of the sample values and the group indicators follow such a model that

$$\mathrm{E}[Y|x_1, x_2, \ldots, x_k] = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k$$

ANOVA is testing wether all the groups means are equal, i.e., $\delta_0 = \delta_1 = \ldots = \delta_k = 0$. By providing valid point estimates and confidence intervals of $\delta_0, \delta_1, \ldots, \delta_k$, we will see if any mean from group $0, \ldots, k$ is significantly different from the mean in group $k + 1$, and also the estimated differences.

Take three groups as an example, Table 4.2 shows the notation for external validation data that we will mention as an example in later sections.

Table 4.2: External Validation Study for three groups Data Layout

x	$x^* = 1$	$x^* = 2$	$x^* = 3$	Total
1	n_{11}	n_{12}	n_{13}	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	$n_{2.}$
3	n_{31}	n_{32}	n_{33}	$n_{3.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{..}$

Regression Calibration Adaptation

In the regression calibration method, the regression coefficient δ_1 is adjusted for the misclassification by multiplying a correction coefficient (Carroll et al. (2006)). In our case, when there are only two groups, the mean of true group indicator given the misclassified group indicator can be written as

$$E(x|x^*) = \frac{(SE - \pi^*)(\pi^* + SP - 1)}{\pi^*(1 - \pi^*)(SE + SP - 1)}x^* + \frac{(1 - SE)(\pi^* + SP - 1)}{(1 - \pi^*)(SE + SP - 1)}$$

where x and x^* can be taken as 0 or 1.

The correction for δ_0 and δ_1 would be

$$\delta_0 = \delta_0^* - [\frac{(1 - SE)(\pi^* + SP - 1)}{(1 - \pi^*)(SE + SP - 1)}]\delta_1 \quad (4.1)$$

$$\delta_1 = \frac{\pi^*(1 - \pi^*)(SE + SP - 1)}{(SE - \pi^*)(\pi^* + SP - 1)}\delta_1^* \equiv \gamma\delta_1^* \quad (4.2)$$

where δ_0^* and δ_1^* are the intercept and slope of the regression from the naive model taking the misclassified group indicator as the predictor variable. In later discussions of this paper, we will refer to $\gamma = \frac{\pi^*(1-\pi^*)(SE+SP-1)}{(SE-\pi^*)(\pi^*+SP-1)}$ as the "correction coefficient". Details of the derivation are presented in the Appendix.

Both the correction for regression intercept and slope require the knowledge of π^*, SE and SP, or reasonable estimates. In the following sections, we will demonstrate the estimation of those parameters and also the coefficient corrections by using external and internal validation data.

When the research interest is to compare the differences in group mean for three groups, similar derivations can be conducted. The correction for regression intercept and slopes would be

$$\begin{aligned}
\delta_0 &= \delta_0^* - a_{01}\delta_1 - a_{02}\delta_2 && (4.3)\\
\delta_1 &= \frac{a_{22}\delta_1^* - a_{12}\delta_2^*}{a_{11}a_{22} - a_{12}a_{21}} && (4.4)\\
\delta_2 &= \frac{a_{21}\delta_1^* - a_{11}\delta_2^*}{a_{12}a_{21} - a_{11}a_{22}} && (4.5)
\end{aligned}$$

where

$$\begin{aligned}
a_{11} &= \frac{SE_{11}\mathrm{P}(x_1 = 1)}{\mathrm{P}(x_1^* = 1)} - \frac{(1 - SE_{11} - SE_{21})\mathrm{P}(x_1 = 1)}{\mathrm{P}(x_1^* = 0, x_2^* = 0)}\\
a_{21} &= \frac{SE_{21}\mathrm{P}(x_1 = 1)}{\mathrm{P}(x_2^* = 1)} - \frac{(1 - SE_{11} - SE_{21})\mathrm{P}(x_1 = 1)}{\mathrm{P}(x_1^* = 0, x_2^* = 0)}\\
a_{01} &= \frac{(1 - SE_{11} - SE_{21})\mathrm{P}(x_1 = 1)}{\mathrm{P}(x_1^* = 0, x_2^* = 0)}\\
a_{12} &= \frac{SE_{12}\mathrm{P}(x_2 = 1)}{\mathrm{P}(x_1^* = 1)} - \frac{(1 - SE_{12} - SE_{22})\mathrm{P}(x_2 = 1)}{\mathrm{P}(x_1^* = 0, x_2^* = 0)}\\
a_{22} &= \frac{SE_{22}\mathrm{P}(x_2 = 1)}{\mathrm{P}(x_2^* = 1)} - \frac{(1 - SE_{12} - SE_{22})\mathrm{P}(x_2 = 1)}{\mathrm{P}(x_1^* = 0, x_2^* = 0)}\\
a_{02} &= \frac{(1 - SE_{12} - SE_{22})\mathrm{P}(x_2 = 1)}{\mathrm{P}(x_1^* = 0, x_2^* = 0)}
\end{aligned}$$

and $\pi_i = \mathrm{P}(x_i = 1)\quad (i = 1 \text{ or } 2)$ can be calculated by

$$\begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix} = \begin{pmatrix} SE_{11} - SE_{13} & SE_{12} - SE_{13} \\ SE_{21} - SE_{23} & SE_{22} - SE_{23} \end{pmatrix}^{-1} \times \begin{pmatrix} \pi_1^* - SE_{13} \\ \pi_2^* - SE_{23} \end{pmatrix} \qquad (4.6)$$

Details on the derivation can be found in the Appendix.

Likewise, when there are more than three groups to compare the differences in group means, the correction for the regression coefficients would be

$$\begin{pmatrix} \delta_1 \\ \delta_2 \\ ... \\ \delta_k \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1k} \\ a_{21} & a_{22} & ... & a_{2k} \\ ... & ... & ... & ... \\ a_{k1} & a_{k2} & ... & a_{kk} \end{pmatrix}^{-1} \times \begin{pmatrix} \delta_1^* \\ \delta_2^* \\ ... \\ \delta_k^* \end{pmatrix}$$

where

$$\begin{aligned} a_{0i} &= \frac{(1 - SE_{1i} - SE_{2i} - \ldots - SE_{ki})\mathrm{P}(x_i = 1)}{\mathrm{P}(x_1^* = 0, x_2^* = 0, \ldots, x_k^* = 0)} \quad (i = 1, \ldots, k) \\ a_{ri} &= \frac{SE_{ri}\mathrm{P}(x_i = 1)}{\mathrm{P}(x_r^* = 1)} - \frac{(1 - SE_{1i} - SE_{2i} - \ldots - SE_{ki})\mathrm{P}(x_i = 1)}{\mathrm{P}(x_1^* = 0, x_2^* = 0, \ldots, x_k^* = 0)} \quad (0 < r \leq k) \end{aligned}$$

where $\pi_i = \mathrm{P}(x_i = 1) \quad (i = 1, \ldots, k)$ can be calculated by

$$\begin{pmatrix} \pi_1 \\ \pi_2 \\ \ldots \\ \pi_k \end{pmatrix} = \begin{pmatrix} SE_{11} - SE_{1,k+1} & SE_{12} - SE_{1,k+1} & \ldots & SE_{1,k} - SE_{1,k+1} \\ SE_{21} - SE_{2,k+1} & SE_{22} - SE_{2,k+1} & \ldots & SE_{2,k} - SE_{2,k+1} \\ \ldots & \ldots & \ldots & \ldots \\ SE_{k1} - SE_{k,k+1} & SE_{k2} - SE_{k,k+1} & \ldots & SE_{k,k} - SE_{k,k+1} \end{pmatrix}^{-1} \times \begin{pmatrix} \pi_1^* - SE_{1,k+1} \\ \pi_2^* - SE_{2,k+1} \\ \ldots \\ \pi_k^* - SE_{k,k+1} \end{pmatrix}$$

Hypothesis Testing

As stated earlier, the focus of this paper is examining the effect of misclassification on both hypothesis testing and estimation of the differences in group means. In this section, we discuss intuitively the validity of the hypothesis testing results based on naive models.

Taking the case of two group misclassification as an example under the non-differential measurement error assumption, the conditional likelihood of the continuous Y given the misclassified group indicator for one subject (the subject indicator i is suppressed) is

$$\begin{aligned} f(Y|x^*) &= f(Y, x=1|x^*) + f(Y, x=0|x^*) \\ &= f(Y|x=1, x^*)\mathrm{P}(x=1|x^*) + f(Y|x=0, x^*)\mathrm{P}(x=0|x^*) \\ &= f(Y|x=1)\mathrm{P}(x=1|x^*) + f(Y|x=0)\mathrm{P}(x=0|x^*) \\ &= w_1 \frac{1}{\sqrt{2\pi}\sigma_e} e^{-(Y-\delta_0-\delta_1)^2/(2\sigma_e^2)} + w_2 \frac{1}{\sqrt{2\pi}\sigma_e} e^{-(Y-\delta_0)^2/(2\sigma_e^2)} \end{aligned}$$

where the first line is written heuristically.

Therefore, the conditional distribution of $y|x^*$ is a mixture of two normal distributions with the weights w_1 and w_2 involving SE, SP and π^* . In general, the variance of $y|x^*$ depends on the misclassified group indicator x^*. So when $\delta_1 \neq 0$, the two sample t-test assumptions are not met for the naive model when taking x^* as the predictor variable. However, under the null hypothesis $H_0 : \delta_1 = 0$, the two normal distributions shrink to the same one, i.e., $f(Y|x^*) \sim N(\delta_0, \sigma^2)$ both for $x^* = 1$ and $x^* = 0$. Therefore, under the null, the hypothesis testing results based on the naive model remain valid. The simulation results in later sections will empirically demonstrate this point in more detail.

Misclassification Adjustment Using External Validation Data

When external validation data (Table 4.1) are available and “transportable”, we can estimate SE and SP solely based on the external validation data, and obtain the estimate of π^* separately based on the “main study” data. When there are only two groups to compare the means, as mentioned in the last section, the π^* can be estimated by $\hat{\pi}^* = n_1^*/(n_1^* + n_0^*)$ and the variance can be calculated by $\hat{Var}(\hat{\pi^*}) = \hat{\pi^*}(1-\hat{\pi^*})/(N_1 + N_0)$, where N_1 and N_1 are the number of observations from the two groups in the main study. In terms of the estimates and variances of SE and SP, taking the notations in Table 4.1, they can be obtained by

$$\begin{aligned}
\hat{SE} &= \frac{n_{11}}{n_{1.}} \\
\hat{SP} &= \frac{n_{00}}{n_{0.}} \\
\hat{Var}(\hat{SE}) &= \hat{SE}(1-\hat{SE})/(n_{1.}) \\
\hat{Var}(\hat{SP}) &= \hat{SP}(1-\hat{SP})/(n_{0.})
\end{aligned}$$

When π^*, SE and SP are estimated by the main study/external validation data, we can obtain the corrected coefficients for both δ_0 and δ_1 by plugging the point estimates into (4.1) and (4.2). The variance of the slope coefficient is also derived in the Appendix by the delta method.

For the cases of more than two groups, take three groups as an example, the estimated π_i^* and its variance can be obtained from the main study by $\pi_i^* = N_i/(N_1 + N_2 + N_3)$ and $\hat{Var}(\hat{\pi_i^*}) = \hat{\pi_i^*}(1-\hat{\pi_i^*})/(N_1 + N_2 + N_3)$, respectively. The estimates of SE_{ij} with their variances can be calculated from the external validation data taking the notations presented in Table 4.2

$$\begin{aligned}
\hat{SE}_{ij} &= \frac{n_{.i}}{n_{j.}} \\
\hat{Var}(\hat{SE}) &= \hat{SE}_{ij}(1-\hat{SE}_{ij})/(n_{j.})
\end{aligned}$$

Similar to the two-group case, plug in the above estimates of SE_{ij} into (4.4) and (4.5), we will obtain the corrected coefficients for all the regression parameters δ_0, δ_1 and δ_2.

Misclassification Adjustment Using Internal Validation Data

When internal validation data are available, we consider two approaches to correct the estimated coefficients for misclassification: the likelihood method and an alternative closed-form method.

Likelihood Method

Here, we seek to define and maximize the log-likelihood of the data observed in the main/internal validation study. For this purpose, we assume the continuous sample values from all groups follow normal distributions with the same variance σ_e^2. We also impose the non-differential measurement error assumption, i.e., $\mathrm{P}(Y|X, X^*) = \mathrm{P}(Y|X)$, implying that given the true group attribution information, the misclassified group indicator provides no additional information about the distribution of the continuous outcome Y. Under those assumptions, the likelihood for a single observation in the case of two groups with misclassification can be written as follows, with the subject indicator i suppressed

$$\begin{aligned}
& \mathscr{L}(Y|x, x^*; \delta_0, \delta_1, \theta) \\
= \;& \mathrm{P}(Y = y, X = x, X^* = x^*) \\
= \;& f(y|X = x)\mathrm{P}(X^* = x^*)\mathrm{P}(X = x) \\
= \;& R\{[\frac{1}{\sqrt{2\pi}\sigma_e} e^{-(y-\delta_0-\delta_1 x)^2/(2\sigma_e^2)}] SE^{xx^*}(1-SE)^{x(1-x^*)}(\frac{\pi^* + SP - 1}{SE + SP - 1})^x \\
& \quad (1-SP)^{(1-x)x^*} SP^{(1-x)(1-x^*)}(\frac{SE - \pi^*}{SE + SP - 1})^{1-x}\} + \\
& (1-R)\{SE^{x^*}(1-SE)^{1-x^*}(\frac{\pi^* + SP - 1}{SE + SP - 1})\frac{1}{\sqrt{2\pi}\sigma_e} e^{-(y-\delta_0-\delta_1)^2/(2\sigma_e^2)} \\
& \quad +(1-SP)^{x^*} SP^{1-x^*}(\frac{SE - \pi^*}{SE + SP - 1})\frac{1}{\sqrt{2\pi}\sigma_e} e^{-(y-\delta_0)^2/(2\sigma_e^2)}\} \qquad (4.7)
\end{aligned}$$

where the first equation is written heuristically. θ is the vector including all the nuisance parameters $(\pi^*, SE, SP, \sigma_e)$, and R is the validation selection indicator such that $R = 1$ if the sample is in the validation data, and $R = 0$ if not. The likelihood for all observations is the product of the likelihood from every single observation, i.e., $\mathbf{L} = \prod_{i=1}^{n} \mathscr{L}_i$, where $\mathscr{L}_i$ is defined in (4.7) with subscripts i added to the terms y, R, x^* and x.

Likewise, in the case of three groups with misclassification, the likelihood for a single observation can be written as

$$\begin{aligned}
&\mathscr{L}(Y|x_1, x_2, x_1^*, x_2^*; \delta_0, \delta_1, \delta_2, \theta) \\
= \quad & R\{[\frac{1}{\sqrt{2\pi}\sigma_e} e^{-(y-\delta_0-\delta_1 x_1-\delta_2 x_2)^2/(2\sigma_e^2)}] SE_{11}^{x_1 x_1^*} SE_{21}^{x_1 x_2^*} (1 - SE_{11} - SE_{21})^{x_1(1-x_1^*-x_2^*)} \pi_1^{x_1} \\
& \quad SE_{12}^{x_2 x_1^*} SE_{22}^{x_2 x_2^*} (1 - SE_{12} - SE_{22})^{x_2(1-x_1^*-x_2^*)} \pi_2^{x_2} SE_{13}^{(1-x_1-x_2)x_1^*} SE_{23}^{(1-x_1-x_2)x_2^*} \\
& \quad (1 - SE_{13} - SE_{23})^{(1-x_1-x_2)(1-x_1^*-x_2^*)} (1 - \pi_1 - \pi_2)^{1-x_1-x_2}\} \\
& +(1-R)\{SE_{11}^{x_1^*} SE_{21}^{x_2^*} (1 - SE_{11} - SE_{21})^{(1-x_1^*-x_2^*)} \frac{\pi_1}{\sqrt{2\pi}\sigma_e} e^{-(y-\delta_0-\delta_1)^2/(2\sigma_e^2)} \\
& \quad + SE_{12}^{x_1^*} SE_{22}^{x_2^*} (1 - SE_{12} - SE_{22})^{(1-x_1^*-x_2^*)} \frac{\pi_2}{\sqrt{2\pi}\sigma_e} e^{-(y-\delta_0-\delta_2)^2/(2\sigma_e^2)} \\
& \quad + SE_{13}^{x_1^*} SE_{23}^{x_2^*} (1 - SE_{13} - SE_{23})^{1-x_1^*-x_2^*} \frac{1 - \pi_1 - \pi_2}{\sqrt{2\pi}\sigma_e} e^{-(Y-\delta_0)^2/(2\sigma_e^2)}\} \qquad (4.8)
\end{aligned}$$

where $\pi_j = \mathrm{P}(x_j = 1) \quad (j = 1$ or $2)$ can be obtained by (4.6).

The joint log-likelihoods in (4.7) and (4.8) can be maximized numerically to obtain the MLE of the δ parameters and δ_1 and their standard errors by applying the Newton-Raphson optimization routine. The implementation is conducted by SAS procedure NLMIXED (SAS Institute Inc (2001)) and the program is available from the authors.

Closed-Form Method

Greenland (1988) proposed a weighted-average estimator δ_G as a closed-form log odds ratio estimator in the 2×2 table setting with exposure misclassification and internal validation data. This estimate is defined as

$$\hat{\delta_G} = \hat{w_G}\hat{\delta_I} + (1 - \hat{w_G})\hat{\delta_E}$$

where $\hat{w_G} = \frac{V\hat{a}r(\hat{\delta}_I)^{-1}}{V\hat{a}r(\hat{\delta}_I)^{-1} + V\hat{a}r(\hat{\delta}_E)^{-1}}$ and $\hat{\delta}_E$ is the corrected coefficient for δ_1 using the main study/internal validation data, treating the internal validation data as if they were external. $\hat{\delta}_I$ is the corresponding corrected coefficient for δ_1 using only the internal validation data by regressing the outcome on the true group indicator without misclassification. The variance of the estimate δ_G can be obtained by

$$V\hat{a}r(\hat{\delta_G}) = \frac{1}{V\hat{a}r(\hat{\delta}_I)^{-1} + V\hat{a}r(\hat{\delta}_E)^{-1}}$$

where $V\hat{a}r(\hat{\delta}_I)$ can be obtained by the regression procedure itself and $V\hat{a}r(\hat{\delta}_E)$ can be obtained by the delta method presented in the Appendix. Lyles et al. (2007) briefly discuss the rational for treating the weight $\hat{w}_G$ as fixed in this variance calculation.

The closed-form estimator $\hat{\delta}_G$ is obviously computational less demanding than the MLE with internal validation data. The simulation study in the next section will evaluate the performance of those two estimators in terms of bias and efficiency using the two-group main study/internal validation design. The performance of the estimator based solely on the external validation data will also be examined.

4.3 SIMULATION STUDY

Testing mean differences across two groups

We assume “transportability” in the sense that the sensitivity and specificity operating in the external validation study are the same as the main study, which are $SE = 0.8$, $SP = 0.7$ in our simulation. We also assume the exposure prevalence $\pi = 0.6$ and the variance of the continuous sample values $\sigma_e = 1$ for both groups. The sample size for the external validation data is 200, as oppose to 600 for the main study. For the simulation study with main/internal validation data, we assume the sampling rate = 25% and the other parameters remain the same. Table 4.3 displays the mean estimates, empirical standard deviation, mean estimated standard error, the 95% coverage for the estimates, the power of effect size 0.7 for sample size 600 and 1.0 for sample size 200 and the ‘pseudo-power’ by sticking into the true values of the nuisance parameter when calculating the power, for δ_0 and δ_1. Estimations based on external validation data, and main/internal validation data via both likelihood method and Greenland’s method are presented and compared. In presenting Greenland’s method, we provide only the estimates of the key parameter of interest (δ_1) and skip the details of the estimation of δ_0. The estimates based only on the misclassified group indicator will also be reported in the table as ‘naive’ estimates.

The simulation results show that the estimates based on a study with external validation data perform well, with reasonably small bias and similar empirical standard deviation and mean estimated standard error. The power of δ_1 is low from the external validation data with moderate sample size of 200 when the true δ_1 is away from the null. However, it is improved in the ‘pseudo-power’ calculation, mainly due to the fact that the variations from the nuisance parameters are eliminated. As expected, the performance of the estimates from external validation data is surpassed by the results from main/internal validation designs with the same sample size. The

Table 4.3: Simulation results for studies with external/internal validation data: 500 simulated data sets with main study sample size = 600 (or 200), external validation data sample size = 150 (or 50), internal validation sampling rate = 25%, sensitivity = 0.8, specificity = 0.7, variance $\sigma_e = 1$ for each group, true coefficient δ_1 = 1.000 or 0.000

Methods	Param.	True	Mean Est.	Emp.SD	Mean Est. SE	95% Cov.	Power†	Pseudo-Power
Main study sample size = 600, external validation data sample size = 150								
Naive Method	δ_0	0.000	0.297	0.074	0.070	2.40	–	–
Naive Method	δ_1	1.000	0.502	0.092	0.090	0.00	100.00	–
External Validation	δ_0	0.000	-0.033	0.200	–	–	–	–
External Validation	δ_1	1.000	1.047	0.274	0.273	96.4	99.8	98.6
Greenland's Method	δ_1	1.000	0.995	0.141	0.139	93.8	100.00	100.00
Likelihood Method	δ_0	0.000	0.002	0.094	0.090	94.0	–	–
Likelihood Method	δ_1	1.000	0.997	0.131	0.125	93.4	100.00	100.00
Naive Method	δ_0	0.000	-0.002	0.067	0.064	93.2	–	–
Naive Method	δ_1	0.000	0.002	0.084	0.083	94.2	–	–
External Validation	δ_0	0.000	-0.003	0.121	–	–	–	–
External Validation	δ_1	0.000	0.004	0.181	0.179	95.4	–	–
Greenland's Method	δ_1	0.000	0.0002	0.128	0.120	93.1	–	–
Likelihood Method	δ_0	0.000	0.002	0.087	0.086	94.6	–	–
Likelihood Method	δ_1	0.000	-0.004	0.130	0.127	93.6	–	–
Main study sample size = 200, external validation data sample size = 50								
Naive Method	δ_0	0.000	0.304	0.117	0.121	29.00	–	–
Naive Method	δ_1	1.000	0.499	0.151	0.157	8.80	88.4	–
External Validation	δ_0	0.000	-0.109	1.040	–	–	–	–
External Validation	δ_1	1.000	1.198	1.127	1.574	95.8	58.4	81.20
Greenland's Method	δ_1	1.000	0.997	0.245	0.242	93.8	98.6	99.2
Likelihood Method	δ_0	0.000	0.008	0.158	0.158	94.6	–	–
Likelihood Method	δ_1	1.000	0.988	0.216	0.217	94.6	98.4	99.2
Naive Method	δ_0	0.000	-0.002	0.106	0.112	94.8	–	–
Naive Method	δ_1	0.000	0.006	0.134	0.144	95.6	–	–
External Validation	δ_0	0.000	-0.002	0.716	–	–	–	–
External Validation	δ_1	0.000	0.009	0.792	2.291	98.4	–	–
Greenland's Method	δ_1	0.000	-0.004	0.238	0.217	94.2	–	–
Likelihood Method	δ_0	0.000	0.003	0.162	0.153	93.4	–	–
Likelihood Method	δ_1	0.000	-0.003	0.238	0.223	93.6	–	–

†The effect size for power is 0.7 for larger sample sizes and 1.0 for smaller sample sizes

estimates from the two methods using main/internal validation data have smaller bias and smaller variability. Consistent with the results from Lyles et al. (2007) for 2×2 table analysis, Greenland's estimator performs as well as if not better than the corresponding MLE in terms of bias and variability. The naive method, on the other hand, provides reliable estimates and hypothesis testing results only in the cases when the true parameters are zero.

When the research question goes to testing the differences in means across more than two groups, we conduct simulation in similar settings as there are two groups, with the same sample size (600 or 200) and true parameter values for δ_0, δ_1 and δ_2 and variance σ_e. However, we increse the validation sampling rate to 50% for more reliable performances in the estimates, and the true parameter values for the SE_{ij} ($i = 1, \ldots, 2$ and $j = 1, \ldots, 3$) are given as $SE_{11} = 0.8$, $SE_{21} = 0.1$, $SE_{12} = 0.2$, $SE_{22} = 0.7$, $SE_{13} = 0.1$, $SE_{23} = 0.2$. Bonferroni corrections are applied in calculating the 95% coverages and powers. For demonstration purpose, we only presented and compared the performances of the naive method and likelihood method in Table 4.4, although the simulation of other methods can be conducted in a similar manner as in the previous table.

Table 4.4 shows the likelihood method performs quite well under moderate sample size (n=200), with reasonably small bias and similar empirical standard deviation and mean estimated standard error. On the other hand, the naive method without any misclassification adjustment performs bad with large bias in estimating the mean differences across the groups, when the true parameters are away from zero. However, when the true parameters are zero, the naive method performs fairly good, in terms of estimates in the mean differences and hypothesis testing results.

Table 4.4: Simulation results for studies with internal validation data testing differences in mean across more than two groups: 1000 simulated data with main study sample size = 600 (or 200), internal validation sampling rate = 50%, $SE_{11} = 0.8$, $SE_{21} = 0.1$, $SE_{12} = 0.2$, $SE_{22} = 0.7$, $SE_{13} = 0.1$, $SE_{23} = 0.2$, variance $\sigma_e = 1$ for each group, the true coefficient $\delta_1 = 1.000$ (or 0.000), $\delta_2 = 0.500$ (or 0.000)

Methods	Param.	True	Mean Est.	Emp.SD	Mean Est. SE	95% Cov.	Power†	Test of equ.‡
Main study sample size = 600, internal validation data samples size = 300								
Naive Method	δ_0	0.000	0.177	0.081	0.081	57.1	–	
Naive Method	δ_1	1.000	0.620	0.105	0.108	12.8	100.00	0.0
Naive Method	δ_2	0.500	0.280	0.106	0.107	41.1	75.6	
Naive Method	$\delta_1 - \delta_2$	0.340	0.099					
Likelihood Method	δ_0	0.000	-0.005	0.088	0.086	98.0	–	
Likelihood Method	δ_1	1.000	1.004	0.123	0.126	98.6	99.5	0.0
Likelihood Method	δ_2	0.500	0.505	0.123	0.121	97.9	98.9	
Likelihood Method	$\delta_1 - \delta_2$	0.499	0.124					
Naive Method	δ_0	0.000	-0.002	0.076	0.077	98.5	–	
Naive Method	δ_1	0.000	0.0003	0.100	0.104	98.7	–	95.6
Naive Method	δ_2	0.000	0.0005	0.102	0.102	98.4	–	
Naive Method	$\delta_1 - \delta_2$	0.000	-0.0002	0.096				
Likelihood Method	δ_0	0.000	-0.003	0.085	0.085	98.6	–	
Likelihood Method	δ_1	0.000	0.001	0.123	0.126	98.2	–	94.4
Likelihood Method	δ_2	0.000	0.002	0.121	0.120	98.3	–	
Likelihood Method	$\delta_1 - \delta_2$	0.000	-0.0007	0.122				
Main study sample size = 200, internal validation data samples size = 100								
Naive Method	δ_0	0.000	0.172	0.136	0.139	88.1	–	
Naive Method	δ_1	1.000	0.625	0.181	0.188	66.4	83.0	14.1
Naive Method	δ_2	0.500	0.281	0.181	0.185	88.7	96.6	
Naive Method	$\delta_1 - \delta_2$	0.500	0.344	0.175				
Likelihood Method	δ_0	0.000	0.0001	0.158	0.148	96.6	–	
Likelihood Method	δ_1	1.000	1.002	0.229	0.217	96.8	98.4	1.0
Likelihood Method	δ_2	0.500	0.497	0.218	0.209	97.7	98.2	
Likelihood Method	$\delta_1 - \delta_2$	0.500	0.505	0.217				
Naive Method	δ_0	0.000	-0.004	0.131	0.134	98.6	–	
Naive Method	δ_1	0.000	0.004	0.174	0.180	98.5	–	95.4
Naive Method	δ_2	0.000	0.0002	0.175	0.177	98.4	–	
Naive Method	$\delta_1 - \delta_2$	0.005	0.168					
Likelihood Method	δ_0	0.000	-0.008	0.145	0.146	98.7	–	
Likelihood Method	δ_1	0.000	0.008	0.222	0.218	98.6	–	95.6
Likelihood Method	δ_2	0.000	0.008	0.204	0.207	98.3	–	
Likelihood Method	$\delta_1 - \delta_2$	0.0006	0.214					

†The effect size for power is 0.7 for sample sizes of 600 and 1.0 for sample sizes of 200

‡Percentage of failing to reject $H_0 : \delta_1 = \delta_2 = 0$

4.4 DISCUSSION

In this chapter, we discuss the misclassification adjustment approaches in the fundermental statistical ANOVA setting, when external or internal validation data are available. In terms of the study design, the methods are applicable in observational studies, such as cross-sectional studies and retrospective studies. The discussions in this chapter represent researches on correcting misclassification in a generic case when the outcome is continuous and the categorical exposures are misclassified. The availability of the validation data serves a critical part of the correction strategy. The following chapter will present another specific case when the binary exposure variable is misclassified and the correction strategy is mainly based on information from repeated measurements of the exposure data, motivated by a real-world example study of menstrual function.

APPENDIX

Derivations of the Regression Calibration Adaptation

Two-group case

Analogous to the regression calibration approach for measurement error correction, we first derive the mean of the unknown variable given its observed surrogate:

$$
\begin{aligned}
E(x|X^{*}=x^{*}) &= \mathrm{P}[(X=1)|x^{*}] \\
&= \begin{cases} \frac{SE\times \mathrm{P}(x=1)}{\mathrm{P}(x^{*}=1)}, \text{ if } x^{*}=1 \\ \frac{(1-SE)\times \mathrm{P}(x=1)}{\mathrm{P}(x^{*}=0)}, \text{ if } x^{*}=0 \end{cases} \\
&= \frac{SE\times \mathrm{P}(x=1)}{\mathrm{P}(x^{*}=1)}\times x^{*} \\
&\quad +\frac{(1-SE)\times \mathrm{P}(x=1)}{\mathrm{P}(x^{*}=0)}\times (1-x^{*}) \\
&= (\frac{SE\times \mathrm{P}(x=1)}{\mathrm{P}(x^{*}=1)}-\frac{(1-SE)\times \mathrm{P}(x=1)}{\mathrm{P}(x^{*}=0)})\times x^{*} \\
&\quad +\frac{(1-SE)\times \mathrm{P}(x=1)}{\mathrm{P}(x^{*}=0)} \\
&= \frac{SE\times \mathrm{P}(x=1)-\mathrm{P}(x=1)\times \mathrm{P}(x^{*}=1)}{\mathrm{P}(x^{*}=1)\times \mathrm{P}(x^{*}=0)}\times x^{*} \\
&\quad +\frac{(1-SE)\times \mathrm{P}(x=1)}{\mathrm{P}(x^{*}=0)}
\end{aligned}
$$

Since $\mathrm{P}(x=1)$ can be estimated by

$$
\mathrm{P}(x=1)=\frac{\mathrm{P}(x^{*}=1)+SP-1}{SE+SP-1}
$$

and also let P$(x^* = 1) = \pi^*$, the above representation can be written as

$$E(x|x^*) = \frac{(SE - \pi^*)(\pi^* + SP - 1)}{\pi^*(1 - \pi^*)(SE + SP - 1)}x^* + \frac{(1 - SE)(\pi^* + SP - 1)}{(1 - \pi^*)(SE + SP - 1)}$$

Then returning to the outcome model, we have

$$\begin{aligned} E(y_i|x^*) &= \delta_0 + \delta_1 P(x|x^*) \\ &= \delta_0 + \delta_1[\frac{(SE - \pi^*)(\pi^* + SP - 1)}{\pi^*(1 - \pi^*)(SE + SP - 1)}x^*] + \delta_1[\frac{(1 - SE)(\pi^* + SP - 1)}{(1 - \pi^*)(SE + SP - 1)}] \\ &= \delta_0^* + \delta_1^* x^* \end{aligned}$$

where δ_0^* and δ_1^* are the regression coefficients from the naive model.

Therefore the correction coefficients would be:

$$\begin{aligned} \delta_0 &= \delta_0^* - [\frac{(1 - SE)(\pi^* + SP - 1)}{(1 - \pi^*)(SE + SP - 1)}]\delta_1 \\ \delta_1 &= \frac{\pi^*(1 - \pi^*)(SE + SP - 1)}{(SE - \pi^*)(\pi^* + SP - 1)}\delta_1^* \end{aligned}$$

Three-group case

Similar to the derivation in two-group case, we first look at a conditional probability

$$\mathrm{P}[(x_1 = 1)|x_1^* = 1, x_2^* = 0] = \frac{\mathrm{P}(x_1 = 1, x_1^* = 1 \text{ and } x_2^* = 0)}{\mathrm{P}(x_1^* = 1 \text{ and } x_2^* = 0)}$$

Since $x_1^* = 1$ implies $x_2^* = 0$, the above conditional probability is equivalent to

$$\begin{aligned}
& \mathrm{P}[(x_1 = 1)|x_1^* = 1, x_2^* = 0] \\
= & \frac{\mathrm{P}(x_1 = 1 \text{ and } x_1^* = 1)}{\mathrm{P}(x_1^* = 1)} \\
= & \frac{SE_{11} \times \mathrm{P}(x_1 = 1)}{\mathrm{P}(x_1^* = 1)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
& \mathrm{P}[(x_1 = 1)|x_1^*, x_2^*] \\
= & \begin{cases} \frac{SE_{11} \times \mathrm{P}(x_1=1)}{\mathrm{P}(x_1^*=1)}, \text{ if } x_1^* = 1; \quad x_2^* = 0 \\ \frac{SE_{21} \times \mathrm{P}(x_1=1)}{\mathrm{P}(x_2^*=1)}, \text{ if } x_1^* = 0; \quad x_2^* = 1 \\ \frac{(1-SE_{11}-SE_{21}) \times \mathrm{P}(x_1=1)}{\mathrm{P}(x_1^*=0, x_2^*=0)}, \text{ if } x_1^* = 0; \quad x_2^* = 0 \end{cases} \\
= & \frac{SE_{11} \times \mathrm{P}(x_1 = 1)}{\mathrm{P}(x_1^* = 1)} \times x_1^* \\
& + \frac{SE_{21} \times \mathrm{P}(x_1 = 1)}{\mathrm{P}(x_2^* = 1)} \times x_2^* \\
& + \frac{(1 - SE_{11} - SE_{21}) \times \mathrm{P}(x_1 = 1)}{\mathrm{P}(x_1^* = 0, x_2^* = 0)} \times (1 - x_1^* - x_2^*) \\
= & (\frac{SE_{11} \times \mathrm{P}(x_1 = 1)}{\mathrm{P}(x_1^* = 1)} - \frac{(1 - SE_{11} - SE_{21}) \times \mathrm{P}(x_1 = 1)}{\mathrm{P}(x_1^* = 0, x_2^* = 0)}) \times x_1^* \\
& + (\frac{SE_{21} \times \mathrm{P}(x_1 = 1)}{\mathrm{P}(x_2^* = 1)} - \frac{(1 - SE_{11} - SE_{21}) \times \mathrm{P}(x_1 = 1)}{\mathrm{P}(x_1^* = 0, x_2^* = 0)}) \times x_2^* \\
& + \frac{(1 - SE_{11} - SE_{21}) \times \mathrm{P}(x_1 = 1)}{\mathrm{P}(x_1^* = 0, x_2^* = 0)} \\
= & a_{11}x_1^* + a_{21}x_2^* + a_{01}
\end{aligned}$$

where

$$\begin{aligned}
a_{11} &= \frac{SE_{11} \times \mathrm{P}(x_1 = 1)}{\mathrm{P}(x_1^* = 1)} - \frac{(1 - SE_{11} - SE_{21}) \times \mathrm{P}(x_1 = 1)}{\mathrm{P}(x_1^* = 0, x_2^* = 0)} \\
a_{21} &= \frac{SE_{21} \times \mathrm{P}(x_1 = 1)}{\mathrm{P}(x_2^* = 1)} - \frac{(1 - SE_{11} - SE_{21}) \times \mathrm{P}(x_1 = 1)}{\mathrm{P}(x_1^* = 0, x_2^* = 0)} \\
a_{01} &= \frac{(1 - SE_{11} - SE_{21}) \times \mathrm{P}(x_1 = 1)}{\mathrm{P}(x_1^* = 0, x_2^* = 0)}
\end{aligned}$$

Similarly,

$$\begin{aligned}
& \mathrm{P}[(x_2 = 1)|x_1^*, x_2^*] \\
= & \begin{cases} \frac{SE_{12} \times \mathrm{P}(x_2=1)}{\mathrm{P}(x_1^*=1)}, \text{ if } x_1^* = 1; x_2^* = 0 \\ \frac{SE_{22} \times \mathrm{P}(x_2=1)}{\mathrm{P}(x_2^*=1)}, \text{ if } x_1^* = 0; x_2^* = 1 \\ \frac{(1-SE_{12}-SE_{22}) \times \mathrm{P}(x_2=1)}{\mathrm{P}(x_1^*=0,x_2^*=0)}, \text{ if } x_1^* = 0; x_2^* = 0 \end{cases} \\
= & \frac{SE_{12} \times \mathrm{P}(x_2 = 1)}{\mathrm{P}(x_1^* = 1)} \times x_1^* \\
& + \frac{SE_{22} \times \mathrm{P}(x_2 = 1)}{\mathrm{P}(x_2^* = 1)} \times x_2^* \\
& + \frac{(1 - SE_{12} - SE_{22}) \times \mathrm{P}(x_2 = 1)}{\mathrm{P}(x_1^* = 0, x_2^* = 0)} \times (1 - x_1^* - x_2^*) \\
= & (\frac{SE_{12} \times \mathrm{P}(x_2 = 1)}{\mathrm{P}(x_1^* = 1)} - \frac{(1 - SE_{12} - SE_{22}) \times \mathrm{P}(x_2 = 1)}{\mathrm{P}(x_1^* = 0, x_2^* = 0)}) \times x_1^* \\
& + (\frac{SE_{22} \times \mathrm{P}(x_2 = 1)}{\mathrm{P}(x_2^* = 1)} - \frac{(1 - SE_{12} - SE_{22}) \times \mathrm{P}(x_2 = 1)}{\mathrm{P}(x_1^* = 0, x_2^* = 0)}) \times x_2^* \\
& + \frac{(1 - SE_{12} - SE_{22}) \times \mathrm{P}(x_2 = 1)}{\mathrm{P}(x_1^* = 0, x_2^* = 0)} \\
= & a_{12}x_1^* + a_{22}x_2^* + a_{02}
\end{aligned}$$

where

$$\begin{aligned}
a_{12} &= \frac{SE_{12} \times \mathrm{P}(x_2 = 1)}{\mathrm{P}(x_1^* = 1)} - \frac{(1 - SE_{12} - SE_{22}) \times \mathrm{P}(x_2 = 1)}{\mathrm{P}(x_1^* = 0, x_2^* = 0)} \\
a_{22} &= \frac{SE_{22} \times \mathrm{P}(x_2 = 1)}{\mathrm{P}(x_2^* = 1)} - \frac{(1 - SE_{12} - SE_{22}) \times \mathrm{P}(x_2 = 1)}{\mathrm{P}(x_1^* = 0, x_2^* = 0)} \\
a_{02} &= \frac{(1 - SE_{12} - SE_{22}) \times \mathrm{P}(x_2 = 1)}{\mathrm{P}(x_1^* = 0, x_2^* = 0)}
\end{aligned}$$

and $\pi_i = \mathrm{P}(x_i = 1)$ can be calculated by

$$\begin{pmatrix} \pi_1 \\ \\ \pi_2 \end{pmatrix} = \begin{pmatrix} SE_{11} - SE_{13} & SE_{12} - SE_{13} \\ \\ SE_{21} - SE_{23} & SE_{22} - SE_{23} \end{pmatrix}^{-1} \times \begin{pmatrix} \pi_1^* - SE_{13} \\ \\ \pi_2^* - SE_{23} \end{pmatrix}$$

Then returning to the outcome model, we have:

$$\begin{aligned} E(y|x_1^*, x_2^*) &= \delta_0 + \delta_1 P(x_1 = 1|x_1^*, x_2^*) + \delta_2 P(x_2 = 1|x_1^*, x_2^*) \\ &= \delta_0 + a_{01}\delta_1 + a_{02}\delta_2 + \delta_1[a_{11} \times x_1^* + a_{21} \times x_2^*] \\ & \quad +\delta_2[a_{12} \times x_1^* + a_{22} \times x_2^*] \\ &= \delta_0 + a_{01}\delta_1 + a_{02}\delta_2 + (a_{11}\delta_1 + a_{12}\delta_2)x_1^* + (a_{21}\delta_1 + a_{22}\delta_2)x_2^* \\ &= \delta_0^* + \delta_1^* x_1^* + \delta_2^* x_2^* \end{aligned}$$

Therefore,

$$\begin{aligned} \delta_0 &= \delta_0^* - a_{01}\delta_1 - a_{02}\delta_2 \\ \delta_1 &= \frac{a_{22}\delta_1^* - a_{12}\delta_2^*}{a_{11}a_{22} - a_{12}a_{21}} \\ \delta_2 &= \frac{a_{21}\delta_1^* - a_{11}\delta_2^*}{a_{12}a_{21} - a_{11}a_{22}} \end{aligned}$$

Delta Method-Based Variance Estimation for the Slope Coefficient

The delta method-based variance of the corrected δ_1 is

$$\hat{Var}(\hat{\delta}_1) = \hat{D}\hat{\Sigma}\hat{D}'$$

where $\hat{\Sigma}$ is the 4×4 diagonal matrix $Diag[\hat{Var}(\hat{\delta}_1^*), \hat{Var}(\hat{\pi}^*), \hat{Var}(\hat{SE}), \hat{Var}(\hat{SP})]$, $\hat{D} = (\hat{d}_1, \hat{d}_2, \hat{d}_3, \hat{d}_4)$, where

$$\begin{aligned}
\hat{d}_1 = \quad & \hat{\pi}^* \times (1-\hat{\pi}^*)(\frac{\hat{\pi}^* + \hat{SP} - 1}{\hat{SE} + \hat{SP} - 1}(\hat{SE} - \hat{\pi}^*))^{-1} \\
\hat{d}_2 = \quad & \hat{\delta}_1^*(\hat{SE} + \hat{SP} - 1)((1 - 2\hat{\pi}^*)(\hat{\pi}^* + \hat{SP} - 1)(\hat{SE} - \hat{\pi}^*) - \\
& \hat{\pi}^*(1 - \hat{\pi}^*)(\hat{SE} - 2\hat{\pi}^* - \hat{SP} + 1))(\hat{\pi}^* + \hat{SP} - 1)^{-2}(\hat{SE} - \hat{\pi}^*)^{-2} \\
\hat{d}_3 = \quad & \hat{\delta}_1^*\hat{\pi}^*(1 - \hat{\pi}^*)((\hat{\pi}^* + \hat{SP} - 1)(\hat{SE} - \hat{\pi}^*) - \\
& (\hat{SE} + \hat{SP} - 1)(\hat{\pi}^* + \hat{SP} - 1))(\hat{\pi}^* + \hat{SP} - 1)^{-2}(\hat{SE} - \hat{\pi}^*)^{-2} \\
\hat{d}_4 = \quad & \hat{\delta}_1^*\hat{\pi}^*(1 - \hat{\pi}^*)((\hat{\pi}^* + \hat{SP} - 1)(\hat{SE} - \hat{\pi}^*) - \\
& (\hat{SE} + \hat{SP} - 1)(\hat{SE} - \hat{\pi}^*))(\hat{\pi}^* + \hat{SP} - 1)^{-2}(\hat{SE} - \hat{\pi}^*)^{-2}
\end{aligned}$$

Chapter 5

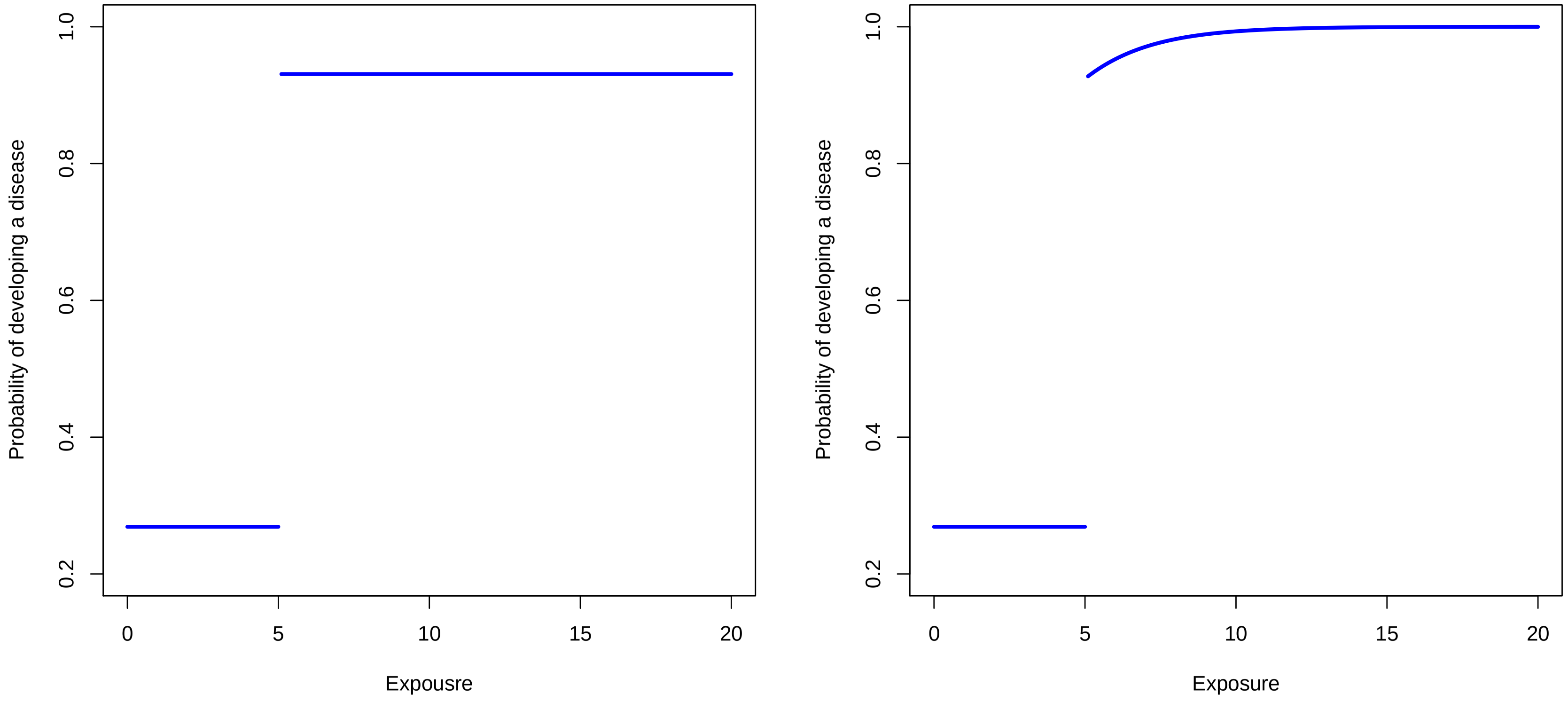
THRESHOLD MODELS FOR SUBJECT-SPECIFIC EXPOSURE MEANS AND VARIANCES

5.1 INTRODUCTION

In environmental epidemiology, the existence of exposure thresholds to some toxic reproductive and developmental agents is sometimes assumed by researchers (Hatch (1971), Wilson (1973), Haseman & Kupper (1979), Schwartz et al. (1995)). More specifically, Wilson (1973) pointed out that a threshold is a dose below which an outcome seen in excess of that is not produced, which can invalidate the non-threshold models. Therefore, instead of assuming a linear association between exposure and a health outcome, researchers sometimes assume a step-like relationship between them: the outcome takes on one distribution when the exposure is less than a certain dose, and takes on another when the exposure exceeds that. Usually the latter indicates a tighter exposure-disease association. The figures in 5.1 illustrate two types of threshold models in such a relationship, with the one on the left showing a constant exposure-outcome relationship below and above the threshold and the one on the right presenting a constant relationship below the threshold and a nonlinear relationship above the threshold.

For a long time, people have been developing appropriate methods to find an exact threshold dosage. The classical approach is to use the no observed adverse effect level (NOAEL) (Barnes & Dourson (1988)). A more general concept is termed the no observed effect level (NOEL), which is estimated by the largest experimental dose where the increase in response over the response in a control group is not statistically significant. For the advantages and disadvantages of NOEL in estimating the threshold, please see Gaylor (1983), Rodricks et al. (1987) and Chen & Kodell (1989). Crump (1984) later defined the benchmark dose (BD) as the lower confidence bound for the dose that causes a 1% to 10% increase in response over the baseline. Ulm (1989), Ulm (1990), Ulm (1991), Kodell et al. (1991), Silvapulle (1991) and Schwartz (1992) also proposed approaches for estimating the threshold dose as a parameter in single-agent dose-response models. A major advantage of estimating the threshold

Figure 5.1: A Illustration of Two Threshold Models: Figure on The Left Illustrates Constant Exposure-Outcome Relationship Below and Above Threshold; Figure on The Right Illustrates Constant Exposure-Outcome Relationship Below The Threshold and Nonlinear Relationship Above The Threshold



dose in this way is that it does not have to be a dose level conducted in the experiment, unlike NOEL or BD. In related work, researchers have also been interested in finding the threshold dose for subjects exposed to several agents in combination (Kavlock & Perreault (1993), Narotsky et al. (1995), Schwartz et al. (1995)).

Instead of making contributions to the topic of estimating the threshold dose itself, this dissertation aims to take a different view of the threshold problem and propose methods for solving misclassification problems in threshold settings. In this regard, suppose exposures to a toxic agent are measured repeatedly over time, and the research question of interest involves the relationship between an adverse health outcome and an individual's true mean exposure over time. This is a reasonable hypothesis in cases where chronic exposure is deemed to make people unhealthy, rather than an acute one-time exposure. However, since the true mean exposure for a specific subject is unknown to researchers in most studies, a natural surrogate for it is the arithmatic average of the exposures observed. Therefore, measurement error or misclassification would occur when taking this surrogate to define the explanatory variable instead of the true mean exposure itself. The analytical problem becomes how to adjust for the misclassification and correct the exposure-outcome relationship with a proper model.

As a nature extension of the generic exposure misclassification problem discussed in Chapter 4.2, this chapter presents models and implementations of a related misclassification problem when the repeated measurements serve as the key information for the correction strategy. Model assumptions and details of parameter estimation for both continuous health outcomes and categorical health outcomes will be presented in section 5.2. Special situations where all subjects have the same number of exposure measurements will be treated in detail and related to classical methodology for exposure misclassification, followed by generalizations to accommodate unbalanced data. Simulation results are shown in section 5.3, followed by a brief discussion.

5.2 MODEL

5.2.1 Homogeneous Within-Subject Variances

Binary Adverse Health Outcome

Matrix Method Suppose we are interested in the association between whether the individual mean exposure to a certain toxic agent exceeds a certain threshold, and a binary adverse health outcome. An example of such an outcome in the MSSWOW study (Section 1.1.2) is spontaneous abortion and the exposure is the mean and/or variability of menstrual cycle length. Then since the mean exposure of a subject is unknown, we impose a reasonable assumption on a single measurement of the exposure for a single subject, that the measurement fluctuates around the individual mean exposure μ_i via a random normal disturbance ϵ_{ij}. Also the mean exposure (μ_i) for subject i varies around the overall mean from the population μ via another random normal disturbance b_i, i.e.,

$$x_{ij} = \mu_i + \epsilon_{ij} \qquad i = 1, \ldots, k; \;\; j = 1, \ldots, n_i \tag{5.1}$$

and

$$\mu_i = \mu + b_i \quad i = 1, \ldots, k \tag{5.2}$$

where we assume $b_i \overset{iid}{\sim} N(0, \sigma_b^2)$, $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$, and b_i and ϵ_{ij} are mutually independent of each other. In this section 5.2.1, we assume homogeneous within-subject variance σ_ϵ; this assumption will be relaxed in 5.2.2.

To simplify the model, we combine (5.1) and (5.2) into a single familiar linear mixed model as follows:

$$x_{ij} = \mu + b_i + \epsilon_{ij} \qquad i = 1, \ldots, k; \;\; j = 1, \ldots, n_i \tag{5.3}$$

To model the relationship between the adverse health outcome and whether the mean exposure exceeds the threshold, we apply a simple logistic regression model, i.e.,

$$\mathrm{logit}[\mathrm{P}(Y_i = 1)] = \delta_0 + \delta_1 I(\mu_i > t) \tag{5.4}$$

where t is the exposure mean threshold level, and $I(.)$ is an indicator function, which takes value one if the criteria in the parentheses is satisfied and zero if not. Since the true individual mean exposure μ_i is not available directly, a natural surrogate of that, denoted as $\tilde{\mu_i}$, would be the average of the repeated exposure measurements taken on subject i, i.e., $\tilde{\mu_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$. Hence, a 'naive' approach would take $I(\tilde{\mu_i} > t)$ as the explanatory variable in (5.4) in stead of $I(\mu_i > t)$.

A special case in this setting is when we have balanced data so that every subject has the same number of exposure measurements, i.e., $n_1 = n_2 = \ldots = n_k = n$. In this case, we develop an approach as a special application of the matrix method proposed by Barron (1977) to correct for misclassification in binary exposure data.

Basic ideas of the matrix method were reviewed in 1.2.4, where in the current setting the events 'E=1' and 'W=1' correspond to '$\mu_i > t$' and '$\tilde{\mu_i} > t$'. Although in our case, no validation study is available, we could still estimate the sensitivity and specificity based on the normal distribution assumptions in (5.1) and (5.2), as suggested by Lyles & Xu (1999). More specifically,

$$\begin{aligned} sen &= \mathrm{P}(\tilde{\mu_i} > t|\mu_i > t) = \frac{\mathrm{P}(\tilde{\mu_i} > t \text{ and } \mu_i > t)}{\mathrm{P}(\mu_i > t)} \\ &= \frac{\mathrm{P}(\bar{\epsilon}_i > t - \mu - b_i \text{ and } b_i > t - \mu)}{\mathrm{P}(b_i > t - \mu)} \\ &= \frac{\int_{t-\mu}^{\infty} \Phi[\sqrt{n}(b_i - (t-\mu))/\sigma_\epsilon]\mathrm{exp}[-{b_i}^2/(2\sigma_b^2)]db_i}{\sqrt{2\pi}\sigma_b\Phi[(\mu - t)/\sigma_b]} \end{aligned} \tag{5.5}$$

where $\bar{\epsilon}_i = \frac{1}{n}\sum_{j=1}^{n}\epsilon_{ij}$. Equation (5.5) holds because $b_i \overset{iid}{\sim} N(0, \sigma_b^2)$ and $\bar{\epsilon}_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2/n)$.

Similarly,

$$sp = [\sqrt{2\pi}\sigma_b(1-p)]^{-1}\int_{-\infty}^{t-\mu}\Phi[\sqrt{n}((t-\mu)-b_i)/\sigma_\epsilon]\exp[-b_i^2/(2\sigma_b^2)]db_i \qquad (5.6)$$

where $p = \Phi[(\mu - t)/\sigma_b]$.

If we insert estimates of *sen* and *sp* in (5.5) and (5.6) into (1.14), and replace $\mathrm{P}(W=1|D=1)$, $\mathrm{P}(W=0|D=1)$, $\mathrm{P}(W=1|D=0)$, $\mathrm{P}(W=0|D=0)$ with the corresponding observed count proportions, (e.g., $\hat{P}(W=1|D=1)=\hat{P}(\tilde{\mu_i}>t|D=1)$), we will get the estimated probabilities of $\mathrm{P}(E=1|D=1)$, $\mathrm{P}(E=0|D=1)$, $\mathrm{P}(E=1|D=0)$, $\mathrm{P}(E=0|D=0)$ in (1.14). The true odds ratio for the unobserved cells takes representation

$$\frac{\mathrm{P}(\mu_i>t, y_i=1)\times\mathrm{P}(\mu_i<t, y_i=0)}{\mathrm{P}(\mu_i<t, y_i=1)\times\mathrm{P}(\mu_i>t, y_i=0)} = \frac{\mathrm{P}(E=1|D=1)\times\mathrm{P}(E=0|D=0)}{\mathrm{P}(E=0|D=1)\times\mathrm{P}(E=1|D=0)}$$

Thus an adaptation of the matrix method provides a viable approach to the mean threshold problem in the special case where exposure data are balanced, the outcome is binary, and there are no covariates.

Likelihood Method Given that the matrix method could only be adapted to estimate the odds ratio for balanced data, we seek to derive a maximum likelihood method to handle this misclassification problem with unbalanced data.

Given the random effect b_i under (5.4), the binary outcome Y_i takes value one with probability p_i, where

$$p_i = \mathrm{P}[Y_i=1|b_i] = \begin{cases} \frac{\exp(\delta_0+\delta_1)}{1+\exp(\delta_0+\delta_1)}, & b_i > t-\mu \\ \\ \frac{\exp(\delta_0)}{1+\exp(\delta_0)}, & b_i < t-\mu \end{cases}$$

Therefore, with the TDM specified in (5.4), and the MEM dictated in (5.3), and along with the non-differential measurement error assumption introduced in (3.2.2), we are able to write down the likelihood as

$$
\begin{aligned}
&\mathcal{L}(\theta; \mathbf{Y}, \mathbf{X}) \\
= \ & \prod_{i=1}^{k}\Big(\int_{-\infty}^{\infty}[f(y_i|(b_i, x_{ij})) \times (\prod_{j=1}^{n_i} f(x_{ij}|b_i)) \times f(b_i)]db_i\Big) \\
= \ & \prod_{i=1}^{k}\Big(\int_{-\infty}^{\infty}[f(y_i|b_i) \times (\prod_{j=1}^{n_i} f(x_{ij}|b_i)) \times f(b_i)]db_i\Big) \qquad (5.7) \\
= \ & \prod_{i=1}^{k}\Big(\int_{t-\mu}^{\infty}[(\prod_{j=1}^{n_i}\frac{exp(-\frac{(x_{ij}-\mu-b_i)^2}{2\sigma_\epsilon^2})}{\sqrt{2\pi}\sigma_\epsilon}) \times [\frac{exp(\delta_0+\delta_1)}{1+exp(\delta_0+\delta_1)}]^{y_i}[\frac{1}{1+exp(\delta_0+\delta_1)}]^{(1-y_i)} \\
& \times \frac{exp(-\frac{b_i^2}{2\sigma_b^2})}{\sqrt{2\pi}\sigma_b}]db_i \\
+ \ & \int_{-\infty}^{t-\mu}[(\prod_{j=1}^{n_i}\frac{exp(-\frac{(x_{ij}-\mu-b_i)^2}{2\sigma_\epsilon^2})}{\sqrt{2\pi}\sigma_\epsilon}) \times [\frac{exp(\delta_0)}{1+exp(\delta_0)}]^{y_i}[\frac{1}{1+exp(\delta_0)}]^{(1-y_i)} \times \frac{exp(-\frac{b_i^2}{2\sigma_b^2})}{\sqrt{2\pi}\sigma_b}]db_i\Big)
\end{aligned}
$$

(5.7) holds under the non-differential measurement error assumption, that the observed measurements give no more information about the outcome once the true mean exposure is known. This follows from the definition of the TDM in (5.4). We are able to estimate $\theta = (\mu, \sigma_b, \sigma_\epsilon, \delta_0, \delta_1)$ by maximizing the integrated likelihood above using a Newton-Raphson algorithm. The SAS NLPQN function in PROC IML can conduct this optimization. The SAS procedure NLMIXED also allows user specification of the observation-specific likelihoods and optimizes the full likelihood via various optimization algorithms (SAS Institute Inc (2001), SAS Institute Inc (2008)). We conducted the analysis in both PROC IML and PROC NLMIXED to assess numerical reliability.

For balanced data, the matrix method assuming MLEs are inserted into (5.5) and (5.6) is equivalent to the likelihood method but computationally less demanding. However, for unbalanced data, the matrix method is no longer available because the sensitivity and specificity calculated in (5.5) and (5.6) vary when n_i varies. Hence a constant expectation across different subjects is not available, which is required in the matrix method. On the other hand, the likelihood method is not subject to the restriction that the number of exposure measurements has to be equal across the subjects and it can also handle other covariates in the TDM and/or MEM. In other words, the likelihood method can be more generally applied than the matrix method. Another difference between the two methods is that the the matrix method can only obtain an estimated crude odds ratio, i.e., the crude association between the outcome and whether the individual mean exposure exceeds the threshold. Yet for other parameters, such as possible covariate effects in the TDM, only the likelihood method can provide appropriate estimates.

Continuous Adverse Health Outcome

Regression Calibration Approach Now we assume the adverse health outcome is continuous, while presuming the same mixed random effect model on exposures as in the case of the binary adverse health outcome (i.e, model (5.3)). Therefore, the TDM becomes

$$E(Y_i) = \delta_0 + \delta_1 I(\mu_i > t)$$

Taking the situation of balanced data as a special case again, an illustrative misclassification correction approach that derives from the idea of regression calibration (see Carroll et al. (2006)) is proposed here.

Regression calibration is a classic approach widely applied by resarchers to correct for measurement error in continuous exposures. However, in our case, the unknown latent variable is the categorical variable indicating whether the mean of exposure exceeds the threshold or not, i.e., $I(\mu_i > t)$. As we discussed in the previous section, the surrogate for the latent variable is whether the average of the exposure measurements taken for every subject exceeds the threshold or not, i.e., $I(\tilde{\mu}_i > t) = I(\bar{x}_i > t)$. Analogous to the regression calibration approach for misclassification correction derived in Chapter 4.2, we find

$$\delta_0 = \delta_0^* - \delta_1[\frac{(1-sen) \times \mathrm{P}(\mu_i > t)}{\mathrm{P}(\tilde{\mu}_i < t)}] \tag{5.8}$$

$$\delta_1 = \delta_1^* \frac{\mathrm{P}(\tilde{\mu}_i > t) \times \mathrm{P}(\tilde{\mu}_i < t)}{[sen - \mathrm{P}(\tilde{\mu}_i > t)]\mathrm{P}(\mu_i > t)} \tag{5.9}$$

where δ_0 and δ_1 are the intercept and slope for the desired model with the explanatory variable being the indicator for whether the true mean exposure exceeds the threshold or not, while δ_0^* and δ_1^* are the intercept and slope for the model with the surrogate as the explanatory variable.

Therefore, the procedure for conducting the analogue to regression calibration for this misclassification problem with balanced exposure data is as follows:

1. Estimate $(\mu, \sigma_b, \sigma_\epsilon)$ by fitting the mixed linear model (5.3) to exposure data x_{ij}

2. Calculate the estimated sensitivity, specificity and $\mathrm{P}(\tilde{\mu}_i > t)$, $\mathrm{P}(\mu_i > t)$ based on the estimated parameters $(\mu, \sigma_b, \sigma_\epsilon)$, with sen and sp defined as in (5.5) and (5.6)

3. Estimate the coefficients δ_0^* and δ_1^* by modelling the outcome y_i on the surrogate $\tilde{\mu}_i = \bar{x}_i$, the average of the x_{ij}s. That is, replace μ_i by $\tilde{\mu}_i$ in the TDM (5.2.1)

4. Obtain the estimate of δ_0 and δ_1 from (5.8) and (5.9), inserting the appropriate estimates in those expressions.

Likelihood Method As was mentioned in the case of the binary health outcome, the likelihood method is in particular demand when the data are unbalanced. In this case, unlike the binary health outcome, we assume a simple linear regression model to capture the relationship between the indicator variable for whether mean exposure μ_i exceeds the threshold t, and the adverse health outcome y_i:

$$y_i = \delta_0 + \delta_1 I(\mu_i > t) + e_i$$

where $e_i \overset{iid}{\sim} N(0, \sigma_e^2)$

Therefore, the mean and variance of the outcome, given the random effect b_i, are:

$$E[y_i|b_i] = \begin{cases} \delta_0 + \delta_1, & b_i > t - \mu \\ \delta_0, & b_i < t - \mu \end{cases}$$

$$Var[y_i|b_i] = \sigma_e^2$$

In that case, the likelihood becomes

$$\begin{aligned} \mathcal{L}(\mathbf{\Theta}; \mathbf{Y}, \mathbf{X}) &= \prod_{i=1}^{k} \Big(\int_{-\infty}^{\infty} [f(y_i|b_i) \times (\prod_{j=1}^{n_i} f(x_{ij}|b_i)) \times f(b_i) db_i] \Big) \\ &= \prod_{i=1}^{k} \Big(\int_{t-\mu}^{\infty} [\frac{exp(-\frac{(y_i - \delta_0 - \delta_1)^2}{2\sigma_e^2})}{\sqrt{2\pi}\sigma_e} \times (\prod_{j=1}^{n_i} \frac{exp(-\frac{(x_{ij} - \mu - b_i)^2}{2\sigma_\epsilon^2})}{\sqrt{2\pi}\sigma_\epsilon}) \times \frac{exp(-\frac{b_i^2}{2\sigma_b^2})}{\sqrt{2\pi}\sigma_b}] db_i \\ &+ \int_{-\infty}^{t-\mu} [\frac{exp(-\frac{(y_i - \delta_0)^2}{2\sigma_e^2})}{\sqrt{2\pi}\sigma_e} \times (\prod_{j=1}^{n_i} \frac{exp(-\frac{(x_{ij} - \mu - b_i)^2}{2\sigma_\epsilon^2})}{\sqrt{2\pi}\sigma_\epsilon}) \times \frac{exp(-\frac{b_i^2}{2\sigma_b^2})}{\sqrt{2\pi}\sigma_b}] db_i \Big) \end{aligned} \tag{5.10}$$

Again, we can estimate $\mathbf{\Theta} = (\mu, \sigma_b, \sigma_\epsilon, \sigma_e, \delta_0, \delta_1)$ by maximizing the likelihood above using Newton-Raphson iteration or similar optimization strategies.

Like the matrix method adaptation presented for binary outcome data, the regression calibration approach proposed here is only appropriate when the data are balanced. The corrections for both intercept and slope have an explicit form, which makes numerical iterations unnecessary. Therefore, the method is computationally superior. Also note that the normality assumption of the errors in the linear TDM is not necessary for the "regression calibration" approach. However, the likelihood method can provide the estimated standard errors for the parameters more conveniently than the matrix method adaptation, and more importantly, may be used for correcting misclassification for unbalanced data and for handling other covariates in the TDM or MEM.

5.2.2 Heterogeneous Within-Subject Variances

Likelihood Method

As our motivating example (1.1.2) requires modeling the variance of the menstrual cycle length, in this section we generalize the homogeneous within-subject variance assumption proposed in (5.3) in section 5.2.1. Instead of assuming a constant within-subject variance σ_ϵ for the error term ϵ, we postulate the within-subject variance v_i follows a lognormal distribution, as proposed by Lyles et al. (1999), i.e.,

$$X_{ij} = \mu + b_i + \epsilon_{ij} \qquad i = 1, \ldots, k; j = 1, \ldots, n_i$$

where

$$\epsilon_{ij}|v_i \overset{iid}{\sim} N(0, v_i), \qquad log(v_i) \overset{iid}{\sim} N(\alpha, \phi^2) \tag{5.11}$$

Therefore, the log-normally distributed v_i accounts for randomness in variability of women's cycle lengths (X_{ij}) in the same way that b_i accounts for randomness in the mean. The advantage of this extension is that it allows the true within-subject variances in cycle length (v_i), as well as the mean cycle lengths ($\mu_i = \mu + b_i$), to vary randomly across subjects. In the likelihood representation, v_i can be integrated out in the same way that b_i is in (5.8) and (5.10). For example, if we look at the exposure data only, the likelihood of observing the repeated exposures x_{ij} in all the subjects will need a double integration as in the following:

$$\mathscr{L}(\theta; \mathbf{X}) = \prod_{i=1}^{k} \{ \int_0^\infty \int_{-\infty}^\infty f(x_{ij}|b_i, v_i) f(b_i|v_i) f(v_i) db_i dv_i \}$$

If b_i and v_i are correlated, we presume b_i and $log(v_i)$ are bivariate normally distributed, as suggested by Lyles et al. (1999), i.e.,

$$\begin{bmatrix} b_i \\ log(v_i) \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ \alpha \end{pmatrix}, \begin{pmatrix} \sigma_b^2 & \sigma_{bv} \\ \sigma_{bv} & \phi^2 \end{pmatrix}\right] \tag{5.12}$$

where σ_{bv} is the covariance between b_i and $log(v_i)$. After algebra (Searle (1982)), we will find

$$(b_i|log(v_i) = lv) \sim N(\bar{\mu}, \bar{\Sigma})$$

where

$$\bar{\mu} = \sigma_{bv}(lv - \alpha)/\phi^2$$

$$\bar{\Sigma} = \sigma_b^2 - \sigma_{bv}^2/\phi^2$$

In order to incorporate both subject-specific mean and variance in a disease-exposure relationship, we consider generalized linear models of the following type:

$$g[E(Y_i)] = \delta_0 + \delta_1 I(\mu_i > \mu) + \delta_2 I(v_i > e^{\alpha}) + \delta_3 I(\mu_i > \mu) I(v_i > e^{\alpha})$$

where the threshold for mean cycle length is set at the overall mean μ, and the threshold for cycle length variance is set at its population median, e^{α}. Here, g is the link function accommodating different types of outcomes (e.g., continuous, binary, count etc.)

The structure of the threshold indicator could vary according to different research questions. For example, in the MSSWOW study, where the exposure pertains to women's menstrual cycles, researchers believe the risk of developing an adverse reproductive outcome is associated with abnormally long or short menstrual cycle length, and/or anomalously large variation in the cycle length (Small et al. (2006)). To be consistent with this hypothesis, the above generalized linear model can be adjusted to take the form:

$$g(E(y_i)) = \delta_0 + \delta_1 I(t_1 < \mu_i < t_2) + \delta_2 I(v_i < \omega) + \delta_3 I(t_1 < \mu_i < t_2) I(v_i < \omega) \quad (5.13)$$

where t_1, t_2 are the lower and upper bound of the normal lengths of menstrual cycles, and ω is the threshold of the normal variation in the lengths. All of those may be set as fixed known constants prior to the analysis, e.g., based on the prior results of Small et al. (2006).

Although Small et al. (2006) conducted an epidemiologically sophisticated study that involved careful statistical modelling, from a measurement error/misclassification perspective their approach is consistent with what might be termed a ‘naive’ method. That is, the models used were analogous to model 5.13, except with the sample mean ($\bar{x}_i$) and sample variance (s_i^2) of cycle lengths replacing the true quantities (μ_i and v_i).

The likelihood considering both disease outcome and the exposure measurements will be as follows:

$$\mathscr{L}(\mathbf{\Theta}; \mathbf{Y}, \mathbf{X}) = \prod_{i=1}^{k} \{ \int_0^\infty [\int_{-\infty}^\infty f(y_i | x_{ij}, b_i, v_i; \theta) f(x_{ij} | b_i, v_i; \theta) f(b_i | v_i; \theta) f(v_i; \theta) db_i] dv_i \}$$

where $\mathbf{\Theta}$ here includes all the parameters to be estimated, i.e., $(\delta_0, \delta_1, \delta_3, \sigma_e, \mu, \sigma_b, \alpha, \phi)$.

As with our previous work to adjust for misclassification, the advantage of this likelihood representation is that it allows direct estimates of and inference about the association between true cycle length means and variances and the outcome of interest.

Empirical Bayes (EB) Method

As an alternative to the naive method, where the subjects are grouped based on the raw sample quantities ($\bar{x}_i$ and s_i^2), the EB method will utilize the empirical Bayes predictors for μ_i and $log(v_i)$ as surrogates for the true classifications. These EB predictions are estimates of the corresponding posterior means, i.e.,

$$\tilde{\mu}_{i,B} = E(\mu_i|X_i) \quad and \quad \tilde{\eta}_{i,B} = E[ln(v_i)|X_i]$$

where $\eta_i = ln(v_i)$, and subscript B indicates the EB estimates. We obtain the EB predictions by utilizing the standard software, e.g., NLMIXED in SAS (SAS Institute Inc (2001), SAS Institute Inc (2008)).

Lyles et al. (1999) pointed out the tremendous amount of shrinkage by the EB estimates in comparison with the sample quantities. Therefore, sensitivity would generally be lower but specificity would be higher when taking EB predictions as the surrogates, rather than relying on the sample quantities. The details on comparing the two surrogates will be discussed in the simulation section.

5.3 SIMULATION STUDY

5.3.1 Homogeneous Within-Subject Variances

Binary Adverse Health Outcome

Matrix Method To demonstrate the performance of the matrix method adaptation, we simulated 500 datasets, each with a sample size of 300. The true parameter values are $(\delta_0, \delta_1, \mu, \sigma_b, \sigma_\epsilon)$ = (1, 1.5, 2.5, 1, 1.5) with the threshold t as 2.5. The number of repeated exposure measurements is 5 for each subject in the simulation.

Since we are interested in the association between the health outcome and whether the true mean exposure (μ_i) exceeds the threshold, and the matrix method can only provide direct estimates of the odds ratio, the simulation results in Table 5.1 will present estimates for δ_1 only based on 500 simulated data. Also in Table 5.1, five situations of different combinations of the overall mean and between- and within-subject variance components are listed to show how those parameters affect the sensitivity and specificity and the limiting value (δ_1^*) of the naive estimators. The threshold t remains constant at 2.5 throughout all situations. Figure 5.2 visually demonstrates how the sensitivity and specificity change as the mean and the between- and within-subject variance change. Each two-dimensional plot is generated when the other two parameters are fixed to the values of $(\mu, \sigma_b, \sigma_\epsilon)$= (2.5, 1, 1.5).

Table 5.1: Simulation results and hypothetical situations illustrating the matrix method for binary outcome: simulation results based on 500 simulated datasets, sample size = 300, t=2.5, number of replicates = 5, and $(\delta_0, \delta_1, \mu, \sigma_b, \sigma_\epsilon)$ = (0, 1)

μ	σ_b	σ_ϵ	SEN	SPC	δ_1^*	$\delta_1(SD)$
2.5	1	1.5	0.81	0.81	0.61	1.01 (0.41)
2.5	1	2.5	0.74	0.74	0.47	1.02 (0.53)
2.5	2	1.5	0.90	0.90	0.79	1.01 (0.32)
1.5	1	1.5	0.75	0.90	0.48	1.03 (0.75)
3.5	1	1.5	0.90	0.75	0.54	0.99 (0.57)

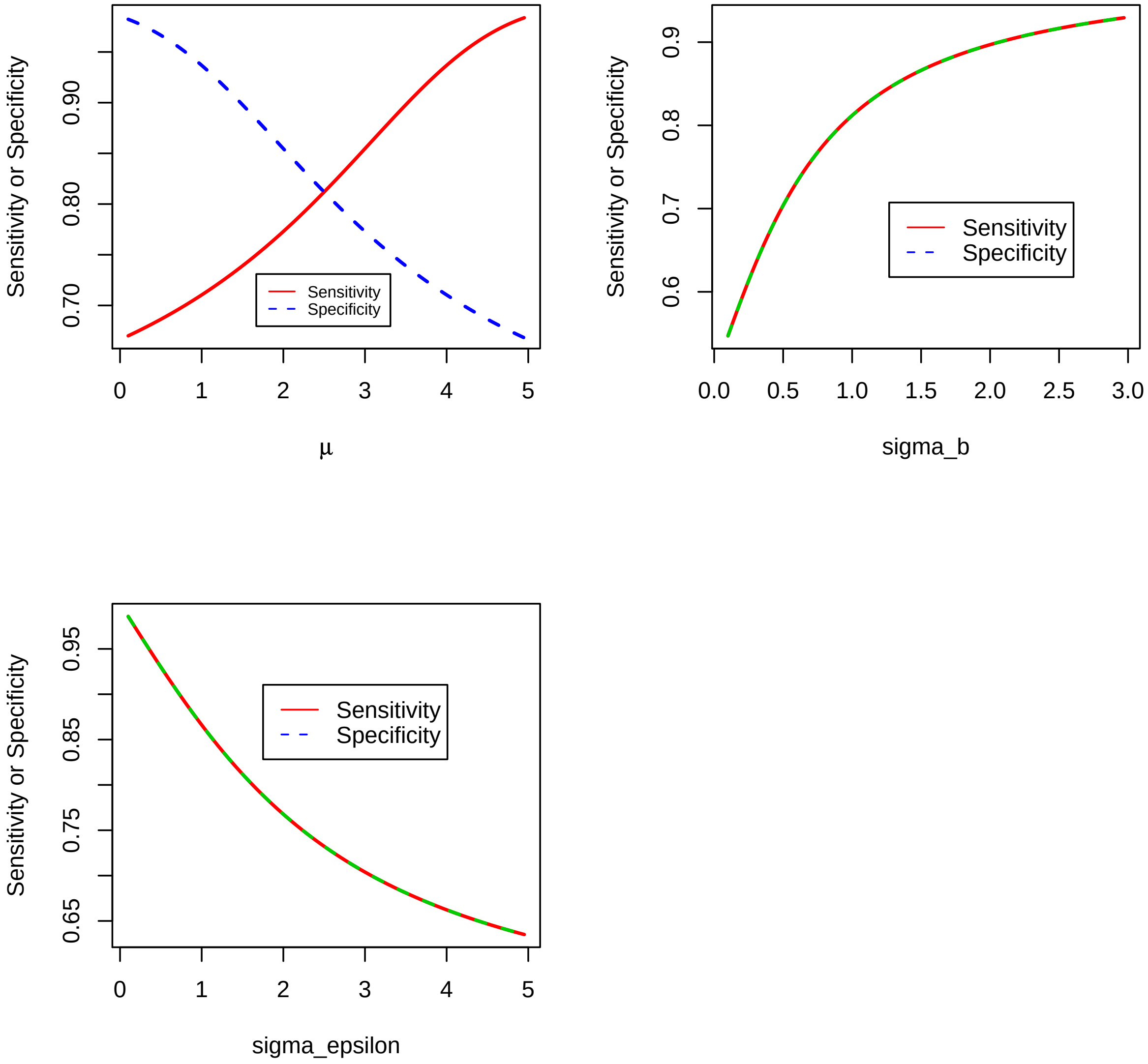


Figure 5.2: Sensitivity and Specificity versus μ, σ_b and σ_ϵ. When plotting sensitivity and specificity versus any one of the parameters μ, σ_b and σ_ϵ, the other parameters are fixed according to: $(\mu, \sigma_b, \sigma_\epsilon)$= (2.5, 1, 1.5). The threshold (t) = 2.5

From Table 5.1 we could see that the matrix method performs well, in the sense that the bias and the empirical standard deviation of the parameter δ_1 are reasonably small. The means of the estimated sensitivity and specificity were close to the ones calculated from the known parameters in the hypothetical situations. The listed hypothetical situations and also Figure 5.2 demonstrate that sensitivity and specificity are equal here given that the threshold equals the overall mean μ, which can also be easily detected from the representation of the sensitivity and specificity in (5.5) and (5.6). Sensitivity and specificity increase as the between-subject variance σ_b increases. The intuition behind this is that when the between-subject variance is large, the mean exposures of each subject are quite different from the threshold so that it would be easier to catch whether the mean exceeds the threshold or not. However, the sensitivity and specificity decrease when the within-subject variance increases. It is also reasonable because when the within-subject variance is relatively larger, the true mean exposure would be harder to approximate by the surrogate, the average of the measurements, which makes the sensitivity and specificity relatively lower. Again, the trends could also be perceived in formulas given in (5.5) and (5.6).

Likelihood Method For demonstrative and comparison purposes, we also assume balanced data in this section with the same simulation settings applied for the matrix method. The simulation results are shown in Table 5.2.

Table 5.2: Simulation illustrating the likelihood approach for binary outcome: results based on 500 simulated datasets, sample size = 300, t=2.5, number of replicates = 5, and $(\delta_0, \delta_1, \mu, \sigma_b, \sigma_\epsilon)$ = (0, 1, 2.5, 1, 1, 1.5)

Parameters	True values	Mean Est.	Emp.SD	Mean Est. SE	95% coverage
δ_0	0.00	0.000	0.23	0.20	0.92
δ_1	1.00	1.000	0.38	0.36	0.95
μ	2.50	2.505	0.07	0.07	0.94
σ_b	1.00	1.001	0.06	0.05	0.96
σ_ϵ	1.50	1.506	0.03	0.03	0.94

As we can see from Table 5.2, the bias associated with each MLE is quite small, indicating that the estimates are close to the true parameter values on average. The empirical standard deviations are not far off from the means of the estimated standard errors. The 95% coverages of the parameters δ_0, δ_1 and μ are close to 95%.

Continuous Adverse Health Outcome

Regression Calibration Approach For a continuous outcome, we utilize similar parameter settings as for the binary outcome but add one more variance component parameter, σ_e. The true parameter values in the simulation are $(\delta_0, \delta_1, \mu, \sigma_b, \sigma_e, \sigma_\epsilon) =$ (0, 1, 2.5, 1, 1, 1.5). As mentioned in the method comparison section, the regression calibration method corrects the estimates of the parameters δ_0 and δ_1, but provides no direct information on other parameters. Results are shown in Table 5.3. The hypothetical situations following the simulation results are a list of five possible numerical combinations of the overall mean and between- and within-subject variance components, demonstrating how those parameters affect the correction coefficient. In the hypothetical situations, we assume the true parameter δ_1 takes value 1, so the column for δ_1^* shows how the estimated parameters based on the naive method would attenuate the effect. Therefore, for demonstrative purpose, the estimates of δ_1^* from hypothetical situations are obtained directly from equation (5.9) assuming δ_1 is 1, without simulating data.

Figure 5.3 illustrates how the correction coefficient changes as the mean, between- and within-subject variance change. Again, the other parameters are set to constant as the values of $(\mu, \sigma_b, \sigma_\epsilon)$= (2.5, 1, 1.5), when one of the three parameters is plotted against the correction coefficient at one time.

The simulation results suggest the regression calibration approach performs reasonably well in the special case when the data are balanced, in that the mean estimated δ_0 and δ_1 are close to the true values and the empirical standard deviations

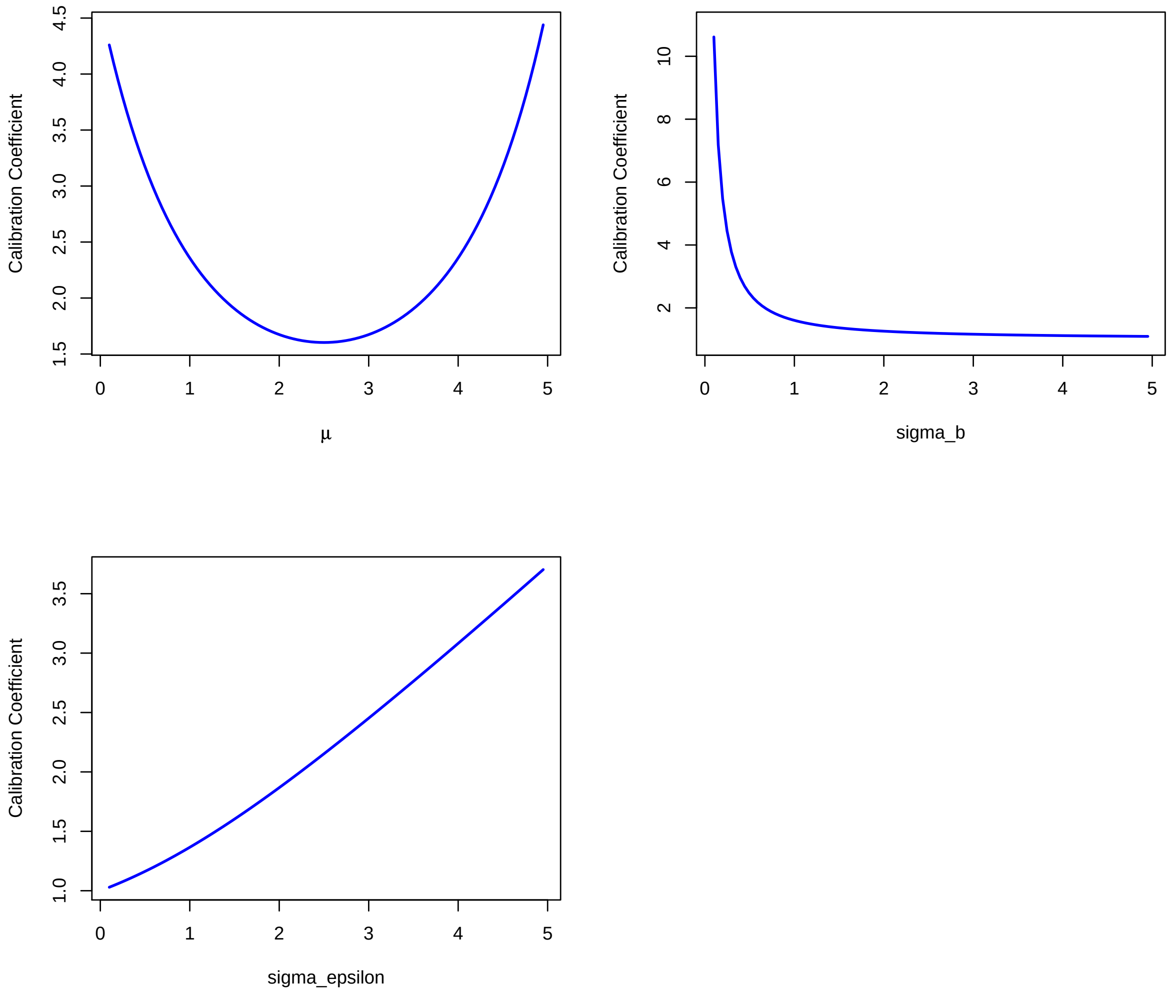


Figure 5.3: Regression calibration coefficient versus μ, σ_b and σ_ϵ. When plotting the regression calibration coefficient versus any one of the parameters μ, σ_b and σ_ϵ, the other parameters are fixed according to: $(\mu, \sigma_b, \sigma_\epsilon)$= (2.5, 1, 1.5)

are reasonably small. The hypothetical situations as well as the plots in Figure 5.3 illustrate that the correction coefficient gets to the lowest point when the overall mean of the exposures equals the threshold. The correction coefficient decreases as the between-subject variance σ_b increases, which is expected as Figure 5.2 suggests the sensitivity and specificity increase as σ_b increases. Also the correction coefficient increases while the within-subject variance σ_ϵ increases. The intuition is also that the larger the within-subject variance is, the less accurate the average of the repeated measurements could be, therefore the more correction would be needed.

Likelihood Method Simulation results based on the same simulation settings as demonstrated for the regression calibration method are displayed in Table 5.4.

The simulation results suggest that the performance of the likelihood method is quite good, in that the bias of the estimates is very small in each case. The means of the estimated standard errors are close to the empirical standard errors. And the 95% coverages are close to 95 for most of the parameters, including δ_1, which we are most interested in.

Table 5.3: Simulation results and hypothetical situations illustrating the regression calibration approach for continuous outcome: simulation results were based on 500 datasets, sample size = 300, t=2.5, number of replicates = 5, and $(\delta_0, \delta_1, \mu, \sigma_b, \sigma_e, \sigma_\epsilon)$ = (0, 1, 2.5, 1, 1, 1.5). The simulated results provide the mean estimates of the correction coefficients, δ_1^* and δ_1.

	μ	σ_b	σ_ϵ	cor. coeff.	δ_1^*	$\delta_1(SD)$
Simulation results	2.5	1	1.5	1.72	0.62	1.04 (1.23)
Hypothetical situations	2.5	1	1.5	1.55	0.65	1
	2.5	1	2.5	2.06	0.49	1
	2.5	2	1.5	1.23	0.81	1
	1.5	1	1.5	1.84	0.54	1
	3.5	1	1.5	1.84	0.54	1

Table 5.4: Simulation illustrating the likelihood approach for continuous outcome: results based on 500 simulated datasets, sample size = 300, t=2.5, number of replicates = 5, and $(\delta_0, \delta_1, \mu, \sigma_b, \sigma_e, \sigma_\epsilon)$ = (0, 1, 2.5, 1, 1, 1.5)

Parameters	True values	Mean Est.	Emp.SE	Mean Est. SE	95% coverage
δ_0	0.00	0.000	0.10	0.10	0.96
δ_1	1.00	1.001	0.16	0.16	0.93
μ	2.50	2.502	0.07	0.07	0.96
σ_b	1.00	1.002	0.12	0.12	0.94
σ_e	1.00	1.002	0.05	0.05	0.96
σ_ϵ	1.50	1.500	0.03	0.03	0.96

5.3.2 Heterogeneous Within-Subject Variances

Analogous to the real data example situation (details will be presented in the next section), and based on the model described in section 5.2.2, we conducted simulation where the outcome is binary, the variances of the exposure for each subject are considered to be different, and the mean and the variance of the exposures are assumed to be correlated. Simulation results are shown in Table 5.5

The simulation results show that the estimates of the parameters from the likelihood are reasonably close to the true parameter values, and the empirical standard deviations are close to the mean estimated standard errors. The 95% Wald confidence interval coverages for the model coefficients are near nominal, all of which suggest the favorable performance of the likelihood method in modeling both correlated mean and variance with the heterogeneous subject-specific variation assumption. Comparing to the likelihood estimates, those from both the EB method and the naive method tend to attenuate the effect to the null, which is consistent with the findings from much of the measurement error literature. We Observe from the simulation that the method of "plugging in" the EB predictors still introduces misclassification and does not guarantee smaller bias than the naive method, which makes sense given that sensitivity is generally decreased and specificity increased via this approach. Although

Table 5.5: Simulation illustrating the likelihood approach for Modeling Correlated Mean and Variance: results based on 500 simulated datasets, sample size = 800, threshold is [26.90, 30.65], number of replicates = 8, and $(\delta_0, \delta_1, \delta_2, \mu, \sigma_b, \alpha, \phi, \sigma_{bv})$ = (-0.8, -0.4, 0.3, 28, 2.8, 1.8, 1.0, 2.0)

	δ_0	δ_1	δ_2	μ	σ_b	α	ϕ	σ_{bv}
Results Based on Likelihood Method								
True Values	-0.8	-0.4	0.3	28	2.8	1.8	1.0	2.0
Mean Est.	-0.812	-0.371	0.302	27.982	2.742	1.803	0.970	1.902
Emp.SD	0.191	0.230	0.230	0.180	0.140	0.060	0.034	0.211
Mean Est. SE	0.190	0.215	0.221	0.093	0.053	0.038	0.031	0.151
95% coverage	95.4%	93.9%	95.8%	–	–	–	–	–
Results Based on EB Method								
True Values	-0.8	-0.4	0.3	–	–	–	–	–
Mean Est.	-0.801	-0.298	0.208	–	–	–	–	–
Emp.SD	0.147	0.167	0.161	–	–	–	–	–
Mean Est. SE	0.151	0.157	0.160	–	–	–	–	–
95% coverage	96.6%	89.0%	90.6%					
Results Based on Naive Method								
True Values	-0.8	-0.4	0.3	–	–	–	–	–
Mean Est.	-0.814	-0.286	0.140	–	–	–	–	–
Emp.SD	0.274	0.170	0.279	–	–	–	–	–
Mean Est. SE	0.263	0.156	0.270	–	–	–	–	–
95% coverage	95.6%	89.3%	84.4%	–	–	–	–	–

Wannemuehler & Lyles (2007) found for continuous exposures that the use of the surrogate of EB predictors performs similarly to the likelihood method, our simulation confirms the expected result that the estimates from dichotomizing the independent variable based on the EB predictors does not provide a consistent estimator.

5.4 REAL-LIFE EXAMPLE

To demonstrate the proposed method, we get back to the second motivating example of the MSSWOW study introduced in Section 1.1.2. The research objective here is to examine whether women with menstrual cycle lengths within a normal range, and/or with cycle length variation less than a certain threshold, are at lower risk of experiencing spontaneous abortion. Among a total number of 470 with menstrual cycle length information in the study (see Small et al. (2006) for details), 162 women are included in this analysis with pregnancy outcome of either live birth (118 women [73%]) or spontaneous abortion (44 women [27%]), including subclinical spontaneous abortion, blighted ovum and clinical spontaneous abortion. For those women with more than one pregnancy outcome, only the first outcome and cycles up to the first outcome are included in the analysis. We begin with a preliminary look into the exposure of menstrual cycle data with heterogeneous within-subject variance model structure when the mean and variance are correlated, as in (5.3), (5.11) and (5.12). Table 5.6 presents the estimated nuisance parameters μ, σ_b, α, ϕ and σbv based on the exposure data only.

Table 5.6: Estimated Nuisance Parameters μ, σ_b, α, ϕ and γ Based on Preliminary Analysis of the Exposure Data with Heterogeneous Within-Subject Variance Model Structure When the Mean and Variance Are Correlated

Parameters	Estimate	Standard Error
μ	28.78	0.24
σ_b	2.78	0.23
α	1.83	0.13
ϕ	1.03	0.12
σ_{bv}†	2.13	0.44

†The estimated correlation coefficient corr$(\mu_i, log(v_i))$ is 0.74

With the estimated nuisance parameters from the exposure model, we dichotomize the subjects' mean and variation of menstrual cycle lengths based on whether their mean cycle length in days is within the normal range of [26.90, 30.65] (corresponding to 25% and 75% percentile of the normal distribution with mean and variance from the estimated parameters) and the variance exceeds the threshold of 4.8 (mean of the lognormal distribution from the estimated parameters) or not. The threshold cut-offs for the mean are different from Small et al. (2006), due to taking advantage of the estimates from the exposure model with random effects. We apply the same thresholds across the three methods (likelihood method, EB method and naive method) for comparison purposes, although in the absence of a random effects exposure model, people often utilize the sample quantiles as the cut-off points.

Regarding the outcome model, the interaction between the mean and variance is perceived to be insignificant in our data example, and herefore is left out of the model. In addition to the dichotomized mean and variance of the cycle lengths as the independent variable, we control in the model for dichotomized maternal age by the cutoff of 35 years of age. In the analysis, we built models with cycle mean only, cycle variance only and both mean and variance in the model. Tables 5.7 - 5.9 present the estimates for those models from the likelihood method, empirical Bayes method and the naive method.

Tables 5.7 and 5.9 show that those women with cycle mean within the middle two quantitles [26.90, 30.65] bear lower probability of experiencing spontaneous abortion than those with cycle mean shorter or longer than the normal range, suggested consistently by the likelihood method, EB method and naive method. In terms of the magnitude of the estimated effects, the naive method and the EB method both appear to attenuate the effect, comparing to the likelihood method. On the other hand, Table 5.8 and 5.9 reveal that the dichotomized cycle variability appears to be an insignificant risk factor for spontaneous abortion as implied by all methods. The estimated variance effect from the models with only dichotomized cycle variance and maternal age from Table 5.8 is slightly further from the null by the naive method

Table 5.7: Estimated Odds Ratio for The Association Between Spontaneous Abortion and Whether the Menstrual Cycle Mean Exceed The Threshold Based on MSSWOW Data

	Results based on likelihood model		Results based on Emp. Bayes method		Results based on naive model	
Risk Factors	OR	95% Wald CI	OR	95%Wald CI	OR	95% Wald CI
Mean of the cyles ([29,33] vs. (<29 or >33))	0.23	0.05, 0.97	0.49	0.24, 0.99	0.44	0.19, 1.05
Maternal Age (<35 vs. ≥35)	0.58	0.22, 1.49	0.65	0.27, 1.56	0.59	0.24, 1.42

Table 5.8: Estimated Odds Ratio for The Association Between Spontaneous Abortion and Whether the Variance Exceed The Threshold Based on MSSWOW Data

	Results based on likelihood model		Results based on Emp. Bayes method		Results based on naive model	
Risk Factors	OR	95% Wald CI	OR	95%Wald CI	OR	95% Wald CI
Variance of the cycles (<23 vs. ≥23)	0.72	0.22, 2.37	0.91	0.44, 1.89	0.65	0.30, 1.37
Maternal Age (<35 vs. ≥35)	0.64	0.27, 1.52	0.64	0.27, 1.53	0.68	0.28, 1.61

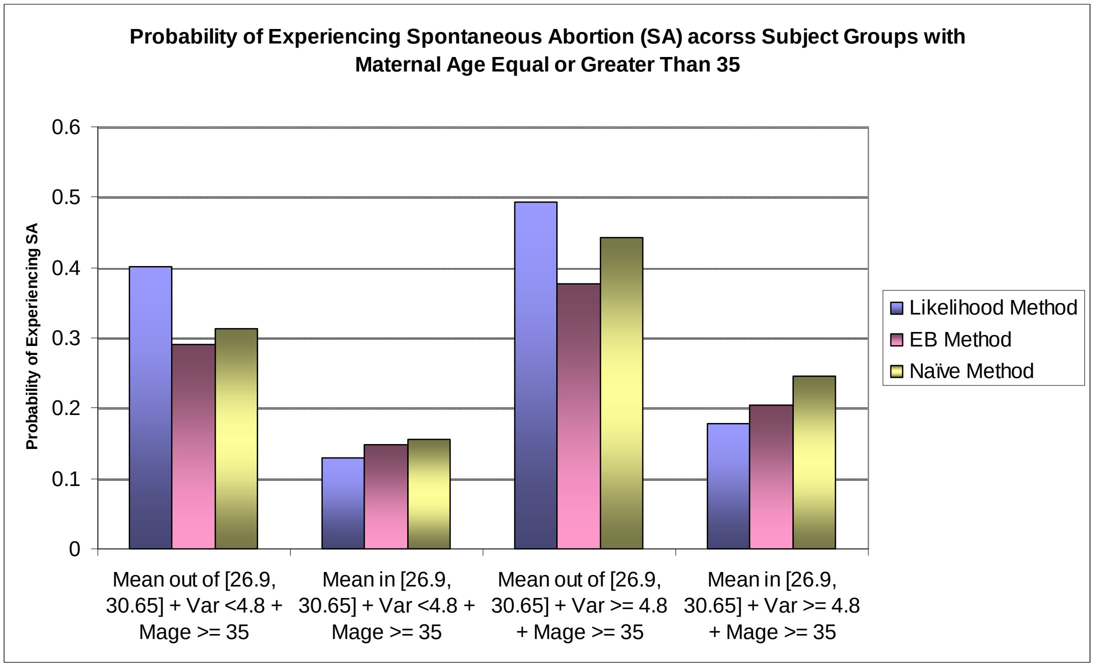
Table 5.9: Estimated Odds Ratio for The Association Between Spontaneous Abortion and Whether the Menstrual Cycle Mean and Variance Exceed Their Respective Thresholds Based on MSSWOW Data

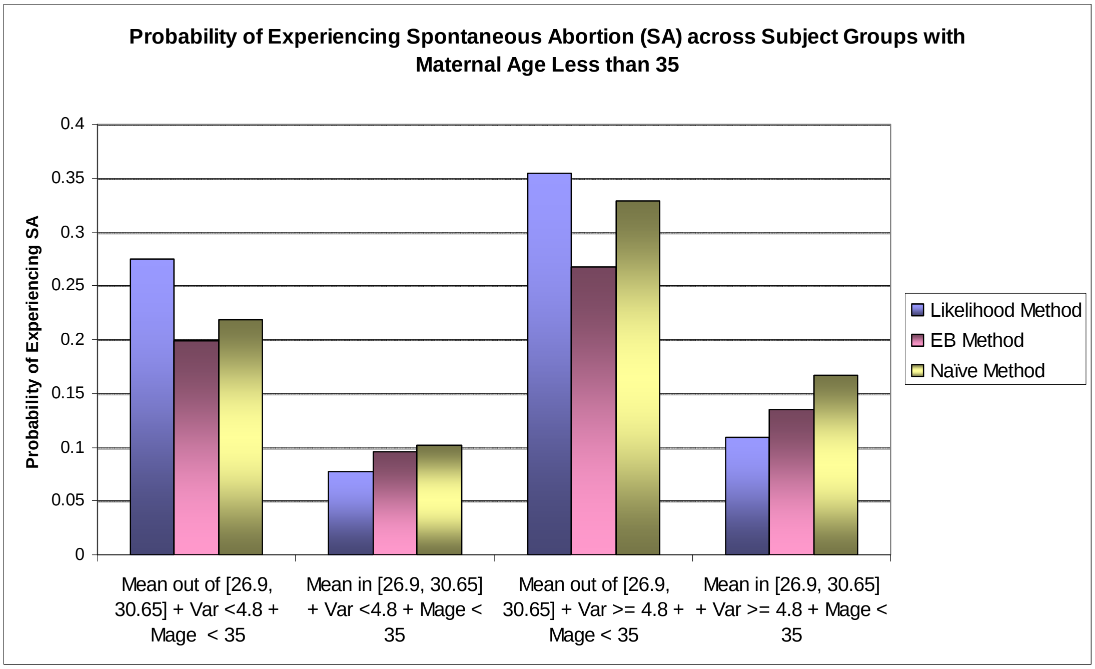
	Results based on likelihood model		Results based on Emp. Bayes method		Results based on naive model	
Risk Factors	OR	95% Wald CI	OR	95%Wald CI	OR	95% Wald CI
Mean of the cyles ([29,33] vs. (<29 or >33))	0.22	0.05, 0.97	0.42	0.20, 0.92	0.41	0.17, 0.99
Variance of the cycles (<23 vs. ≥23)	0.69	0.20, 2.39	0.68	0.31, 1.50	0.57	0.26, 1.24
Maternal Age (<35 vs. ≥35)	0.56	0.22, 1.47	0.61	0.25, 1.47	0.62	0.25, 1.49

as compared to the likelihood method, while the EB has the estimated effect closest to the null. However, with mean cycle length and maternal age controlled for in the model, the magnitute of the estimated variance effects are all similar across the three methods. The same occurs for maternal age, i.e., it is an insignificant covariate with very similar estimated effects from all the methods after adjusting for the cycle means and variations. The confidence intervals stemming from the EB method are narrower than those from the likelihood method, likely due to the fact that the variations of the EB predictions are not incorporated into the second stage model, where the predictions are "plugged in" as surrogates.

Figure 5.4, plotting the estimated probability of spontaneous abortion from likelihood method acorss different subject groups, shows the highest risk for those with cycle mean outside of the normal range and large cycle variation. The lowest risk is associated with cycle means within the normal range and smaller cycle variations. As the adjusted dichotomized maternal age effects are estimated to be similar across the three methods, the figure in the upper panel looks quite similar to that in the lower panel but on a different scale of probability.

Figure 5.4: Estimated Probability of Experiencing Spontaneous Abortion for Various Subject Groups with Young and Old Maternal Ages





Chapter 6

SUMMARY AND FUTURE WORK

6.1 Summary

This dissertation discusses statistical models in three scenarios, all of which focus on correcting for misclassification of exposures with repeated measurements. For all three scenarios, the true exposures are unknown to the researchers, and measurable only by means of an error-prone surrogate method, which explains why the misclassification of the exposures occurs. Although for each of the three study scenarios, the research question of interest is identifying the association between a disease outcome and exposure-related explanatory variables, the details on the exposure-related explanatory variables are different. One focuses on the probability of exposure, another on the exact binary exposure status, and the last on the whether the true mean and/or variance of exposures taken repeatedly over time exceeds a threshold. In the first two scenarios the unknown exposure is binary, and for the last one continuous exposures are converted to categories in a manner that is common in epidemiology.

The first two scenarios are motivated by the same data, from a sub-study of the Baltimore-Washington Infant Study (BWIS). The research question in the first scenario is whether those infants with higher parental probability of exposure to lead tend to have higher probability of developing TAPVR, a congenital cardiovascular malformation. By presuming a beta distribution on the true exposure probability, we obtain the estimated association by maximizing the marginal likelihood of the observed exposure replicates and the disease outcome status. The results suggest a significant association between developing TAPVR and the parental probability of exposure to lead without controlling for covariates, while after controlling for race as a covariate, the significance goes away. Simulation results suggest the naive model based on an obvious surrogate for the probability of exposure attenuates the extent of association.

The second research question is whether true parental binary exposure status to lead has any relation with the infant's probability of developing TAPVR. We apply the regular latent class model with a TDM in the situation that the replicated exposure assessments are independent. We also extend a previously proposed latent class model with random effects by attaching an additional TDM layer to the model, when we believe the replicated exposure assessments are correlated conditional on true exposure status. The estimated association obtained by maximizing the marginal likelihood suggested that those infants with parental lead exposure tend to have significantly higher probability of developing TAPVR without controlling for covariates. Again, this relationship became insignificant when covariates were controlled. Simulation results show noteworthy attenuation of the exposure-disease relationship by means of the naive model based on surrogates.

Unlike the first two scenarios, the third scenario is dealing with continuous exposures. Motivated by a reproductive health study, the research question is whether women's mean menstrual cycle length being abnormally long or short, and/or unusual cycle length variability, has any association with a reproductive outcome (e.g., a spontaneous abortion). By imposing a mixed random effects model on the exposure, the estimated associations between the disease outcome and the mean and variability of exposure are obtained by a maximum likelihood approach. The matrix and regression calibration methods are also derived to address special cases of the this scenario when all the subjects have the same numbers of repeated measurements. Simulation studies are conducted, and real-life examples based on the MSSWOW study are provided.

As a prelude to the third scenario discussed above, we also deliberate a generic exposure misclassification problem when the outcome is continuous, which is essentially a misclassification issue in a typical ANOVA setting. To adjust for the misclassification in this generic setting, we gain information from validation data to estimate the sensitivity and specificity. Details on estimation methods from different study designs with internal or external validation data are discussed. Two approaches, including regression calibration and a full likelihood method, are deliberated with simulation support.

6.2 Future Research

Potential future research topics will include:

- Generalizing the methods of Chapter 3 to situations involving both categorical explanatory variables and outcomes with more than two levels.

  Chapter 3 demonstrates methodologies to adjust for misclassification when both explanatory variables and outcomes and binary. The methods can be generalized to cases when predictors and outcomes are categorical variables with more than two levels. When the outcomes are categorical with more than two levels, one could substitute proportional logit model to the logistic regression model in Chapter 3. When the predictors are categorical with more than two levels, say high exposure, moderate exposure, low exposure etc., one could consider the predictor as ordinal in the proportional logit model.

- Exploring semiparametric models for the topic in Chapter 5, allowing for flexible model assumptions.

  Tsiatis and Ma (2006) proposed a local efficient estimate approach based on generalized estimating equations to allow for more flexible and robust models. The local efficient estimator is efficient when the underlying distribution assumption is correct, but the estimator remains consistent when the underlying distribution is misspecified. Possible future work includes applying the local efficient estimates to our topic in Chapter 5 to allow for flexible distributions (e.g., other than normal).

- Modeling the discrete time hazard of pregnancy as the outcome in the MSS-WOW data.

Small et al. (2006) considered the association between menstrual cycle characteristics and both the risk of spontaneous abortion and fecundity (time to pregnancy). They modeled the quantity of discrete time hazard, e.g., defined as the conditional probability that a woman became pregnant in a given menstrual cycle conditional on not being pregnant in the prior cycles. The reason for taking the conditional probability as the outcome was mainly due to the concern of the informative number of cycles. For example, if there is strong association between the mean cycle length and the fecundity, those women with shorter cycles contribute fewer cycles to the analysis because it takes less cycles for them to get pregnant. In contrast, by taking the advantage of the fact that the outcome of spontaneous abortion happened after the exposure window, utilizing it as the outcome avoids the complexity of the outcome model in order to keep the focus on demonstrating our misclassification adjustment methods. For further research, study of both outcomes would undoubtedly enrich our perspectives and insights into the misclassification adjustment method under various research questions.

- Analytically seeking the optimized threshold in association between menstrual cycle lengths and the reproductive outcome from the MSSWOW data.

  In the demonstrative example, we derived the threshold boundaries as estimated percentiles based on our random effects model for exposure. However, there may be optimized thresholds that best reveal the association between the cycle lengths and the reproductive outcome. Prior literature discusses parametric and semiparametric/nonparametric ways of detecting such thresholds. In future work, we will consider incorporating our misclassification framework into a reasonable threshold-exploring procedure to better reveal the underlying associations.

Bibliography

Barnes, D. & Dourson, M. (1988). Reference dose (rfd): Description and use in health risk assessment. *Regulatory Toxicology and Pharmacology* **8**, 471–486.

Barron, B. (1977). The effects of misclassification on the estimation of relative risk. *Biometrics* **33**, 414–418.

Bickel, P. & Doksum, K. (2001). *Mathematical Statistics: Basic Ideas and Selected Topics*. Vol. I. Prentice-Hall. New Jersey.

Bouyer, J. & Hemon, D. (1993). Studying the performance of a job exposure matrix. *Int J Epidemiol* **22**(Suppl.2), S65–S71.

Brunekreef, B., Noy, D. & Clausing, P. (1987). Variability of exposure measurements in environmental epidemiology. *American Journal of Epidemiology* **125**, 892–898.

Carroll, J., Ruppert, D., Stefanski, L. & Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models–A Modern Perspective*. 2nd edn. Taylor and Francis Group.

Chen, J. & Kodell, R. (1989). Quantitative risk assessment for teratological effects. *Journal of the American Statistical Association* **84**, 966–971.

Clayton, D. (1992). *Statistical Models for Longitudinal Studies of Health*. Oxford University Press. Oxford, U.K.. chapter Models for the longitudinal analysis of cohort and case-control studies with maccurately measured exposures.

Conn, A., Gould, N. & Toint, P. (1987). *Trust Region Method*. Society for Industrial Mathematics.

Cook, R. & Weisberg, S. (1990). Confidence curves in nonlinear regression. *Journal of the American Statistical Association* **85**, 544–551.

Correa, A., Yuan, M., Stewart, A., Lees, P., Breysse, P., Dosemeci, M. & Jackson, L. (2006). Inter-rater agreement of assessed prenatal maternal occupational exposures to lead. *Birth Defects Research Part A: Clinical and Molecular Teratology* **76**(11), 811–824.

Crump, K. (1984). A new method for determing allowable daily intakes. *Fundametal and Applied Toxicology* **4**, 854–871.

D'Agostino, R., Sullivan, L. & Beiser, A. (2005). *Introductory Applied Biostatistics*. Brooks Cole.

Donaldson, J. R. & Schnabel, R. (1987). Computational experience with confidence regions and confidence intervals for nonlinear least squares. *Technometrics* **29**, 67–82.

Dutilleul, P., Pelletier, B. & Alpargu, G. (2008). Modified f tests for assessing the mulitple correlation between one spatial process and several others. *Journal of statistical planning and inference* **138**, 1402–1415.

Evans, M. & Kim, H.M., O. T. (1996). An application of profile-likelihood based confidence interval to capture:recapture estimators. *Journal of Agricultural, Biological, and Environmental Statistics* **1**, 131–140.

Ferencz, C., Rubin, J., Loffredo, C. & Magee, C. (1993). *Epidemiology of congenital heart disease: the Baltimore-Washington infant study, 1981-1989*. Futura publishing.

Flegal, K., Keyl, P. & Nieto, F. (1991). Differential misclassification arising from nondifferential errors in exposure measurement. *American Journal of Epidemiology* **134**(10), 1233–1246.

Formann, A. & Kohlmann, T. (1996). Latent class analysis in medical research. *Statistical Methods in Medical Research* **5**, 179–211.

Gange, S., Muñoz, A., Sáez, M. & Alonso, J. (1996). Use of the beta-binomial distribution to model the effect of policy changes on appropriateness of hospital stays. *Applied Statistics* **45**(3), 371–382.

Gaylor, D. (1983). The use of safety factors for controlling risk. *Journal of Toxicology and Environmental Health* **11**, 329–336.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61**, 215–231.

Greenland, S. (1988). Variance estimation for epidemiologic effect estimates under misclassifiation. *Statistics in Medicine* **7**, 745–757.

Haseman, J. & Kupper, L. (1979). Analysis of dichotomous response data from certain toxicological experiments. *Biometrics* **35**, 281–292.

Hasler, M., Vonk, R. & Hothorn, L. (2008). Assessing non-inferiority of a new treatment in a three-arm trial in the presence of heteroscedasticity. *Statistics in Medicine* **27**, 490–503.

Hatch, T. (1971). Thresholds: Do they exist?. *Archives of Environmental Health* **22**, 687–689.

Hildebrand, F. (1956). *Introduction to Numerical Analysis*. McGraw-Hill. New York.

Hui, S. & Walter, S. (1980). Estimating the error rates of diagnostic tests. *Biometrics* **36**, 167–171.

Jackson, L. (2003). A case-control study of parental occupational lead exposure and total anomalous pulmonary venous return. PhD thesis. Johns Hopkins University.

Kavlock, R. & Perreault, S. (1993). *Toxicology of Chemical Mixtures: From Real Life Examples to Mechanisms of Toxicological Interactions*. San Diego: Academic Press. chapter Multiple Chemical Exposure and Risks of Adverse Reproductive Function and Outcome.

Kodell, R., Howe, R., Chen, J. & Gaylor, D. (1991). Mathematical modeling of reproductive and developmental toxic effects for quantitative risk assessment. *Risk nalysis* **11**, 583–590.

Lesaffre, E. & Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: an example. *Applied Statistics* **50**(3), 325–335.

Liu, Q. & Pierce, D. (1994). A note on gauss-hermite quadrature. *Biometrika* **81**, 624–629.

Liu, X. & Liang, K. (1991). Adjustment for nondifferential misclassification error in the generalized linear-model. *Statistics in Medicine* **10**(8), 1197–1211.

Lyles, H., Munõz, A., Xu, J., Taylor, J. & Chmiel, S. (1999). Adjusting for measurement error to assess health effects of variability in biomarkers. *Statistics in medicine* **18**, 1069–1086.

Lyles, R. (2002). A note on estimating crude odds ratios in case-control studies with differentially misclassified exposure. *Biometrics* **58**, 1034–1037.

Lyles, R. & Kupper, L. (1997). A detailed evaluation of adjustment methods for multiplicative measurement error in linear regression with applications in occupational epidemiology. *Biometrics* **53**, 1008–1025.

Lyles, R. & Kupper, L. (2000). *Handbook of Statistics*. Vol. 18. Elsevier Science. chapter Measurement Error Models for Environmental and Occupational Health Applications, pp. 501–517.

Lyles, R. & Xu, J. (1999). Classifying individuals based on predictors of random effects. *Statistics in Medicine* **18**, 35–52.

Lyles, R., Zhang, F. & Drews-Botsch, C. (2007). Combining internal and external validation data to correct for exposure misclassification a case study. *Epidemiology* **18**, 321–328.

Marshall, R. (1990). Validation study methods for estimating proportions and odds ratios with misclassified data. *Journal of Clinical Epidemiology* **43**, 941–947.

McHugh, R. (1956). Efficient estimation and local identifiability in latent class analysis. *Psychometrika* **21**, 331–347.

Meeker, W. (1987). Limited failure population life tests: Application to integrated circuit reliability. *Technometrics* **29**, 51–65.

Miao, W. & Chiou, P. (2008). Confidence intervals for the difference between two means. *Computational Statistics and Data Analysis* **52**, 2238–2248.

Min, Y. (1995). Low birth weight and parental occupational lead exposure: a case-control study. PhD thesis. Johns Hopkins University.

Min, Y., Correa, A. & Steward, P. (1996). Parental occupational lead exposure and low birth weight. *American Journal of Industial Medicine* **30**, 569–578.

Morrissey, M. & Spiegelman, D. (1999). Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons. *Biometrics* **55**, 338–344.

Narotsky, M., Weller, E., Chinchilli, V. & Kavlock, R. (1995). Nonadditive developmental toxicity in mixtures of trichloroethylene, di (-ethylhexyl) phthalate, and heptachlor in a 5 × 5 × 5 design. *Fundamental and Applied Toxicology.*

Ostrouchov, G. & Meeker, W. (1988). Accuracy of approximate confidence bounds computed from interval censored weibull and lognormal data. *Journal of Statistical Computation and Simulation* **29**, 43–76.

Pinheiro, J. & Bates, D. (1995). Approximations to the log-likelihood function in the non-linear mixed-effects model. *Journal of Computational Graphic Statistics* **4**, 12–35.

Qu, Y., Tan, M. & Kutner, M. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics* **52**, 797–810.

Rodricks, J., Frankos, V., Turnbull, D. & Tardiff, R. (1987). Risk assessment for effects other than cancer. *Food Protection Technology* pp. 61–74.

Roy, S. & Gnanadesikan, R. (1959). Some contributions to anova in one or more dimensions. *Annals of Mathematical statistics* **30**(2), 304–317.

Sadana, R. (2002). Definition and measurement of reproductive health. *Bulletin of the World Health Organization* **80**(5), 407–409.

SAS Institute Inc (2001). *SAS/IML Software: Changes and Enhancements, Release 8.2.* SAS Institute Inc. Cary, NC.

SAS Institute Inc (2008). *SAS/Share 9.2 User's Guide*. SAS Institute Inc. Cary, NC.

Satten, G. & Kupper, L. (1993). Inferences about exposure-disease association using probability-of-exposure information. *Journal of American Statistical Association* **88**(421), 200–208.

Schwartz, P. (1992). Threshold Models in Risk Assessment for a Combination of Agents. PhD thesis. Virginia Commonwealth University, Dept. of Biostatistics.

Schwartz, P., Gennings, C. & Chinchilli, V. (1995). Threshold models for combination data from reproductive and developmental experiments. *Journal of the Ametican Statistical Association* **90**(431), 862–870.

Searle, S. (1982). *Matrix Algebra Useful for Statistics*. Wiley. New York.

Self, S. & Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**(398), 605–610.

Silvapulle, M. (1991). On testing for threshold values. *Biometrics* **47**, 1628–1629.

Small, C., Manatunga, A.K., Klein, M., Feigelson, H., Dominguez, C., McChesney, R. & Marcus, M. (2006). Menstrual cycle characteristics - associations with fertility and spontaneous abortion. *Epidemiology* **17**(1), 52–60.

Stewart, P. (1999). Exposure assessment in community-based epidemiological studies. *Lancet* **353**, 1816–817.

Stewart, P., Stewart, W. & Heineman, E. (1996). A novel approach to data collection in a case-control study of cancer and occupational exposures. *International Journal of Epidemiology* **25**, 744–752.

Stewart, W. & Stewart, P. (1994). Occupational case-control studies: I. collecting information on work histories and work related exposures. *Am J Ind Med* **26**, 297–312.

Stram, D. & Lee, J. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50**, 1171–1177.

Stram, D. & Lee, J. (1995). Variance components testing in the longitudinal mixed effects model (vol 50, pg 1171, 1994). *Biometrics* **51**(3), 1196–1196.

Thomas, D., Stram, D. & Dwyer, J. (1993). Exposure measurement error: Influence on exposure-disease relationships and methods of correction. *Annual Reviews of Public Health* **14**, 69–93.

Thurston, S., Williams, P. & Hauser, R. (2003). A comparison of regression calibration approaches for designs with internal validation data. *Journal of Statistical Planning Inference* **131**, 175–190.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data.* New York: Springer.

Ulm, K. (1989). On the estimation of threshold values. *Biometrics* **45**, 1324–1326.

Ulm, K. (1990). Threshold models in occupational epidemiology. *Mathematical and Computer Modelling* **14**, 649–652.

Ulm, K. (1991). A statstical method for assessing a threshold in epidemiological studies. *Statistics in Medicine* **10**, 341–349.

Venzon, D. & Moolgavkar, S. (1988). A method for computing porfile-likelihood based confident intervals. *Applied Statistics* **37**, 87–94.

Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* Springer.

Walter, S. (1984). Measuring the reliability of clinical data: The case for using three observers. *Revue Epidemiologie et Sante Publique* **32**, 206–211.

Wannemuehler, K. & Lyles, R. (2007). Likelihood-based Measurement Error Adjustments in Occupational and Environmental Exposure Studies. PhD thesis. Biostatistics Dept., Emory University.

Welch, B. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika* **29**, 350–362.

Wieringen, W. (2005). On identifiability of certain latent class models. *Statistics and Probability Letters* **75**, 211–218.

Wijngaarden, E., Stewart, P., Olshan, A., Savitz, D. & Bunin, G. (2003). Parental occupational exposure to pesticides and childhood brain cancer. *American Journal of Epidemiology* **157**, 989–997.

Williams, D. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* **31**(4), 949–952.

Wilson, J. (1973). *Environmental and Birth Defects*. New York: Academic Press.