CASE-CONTROL STUDIES OF GENETIC AND ENVIRONMENTAL FACTORS WITH ERROR IN MEASUREMENT OF ENVIRONMENTAL FACTORS

A Dissertation

by

IRYNA LOBACH

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2006

Major Subject: Statistics

CASE-CONTROL STUDIES OF GENETIC AND ENVIRONMENTAL FACTORS WITH ERROR IN MEASUREMENT OF ENVIRONMENTAL FACTORS

A Dissertation

by

IRYNA LOBACH

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Raymond J. Carroll
Committee Members,	Ruzong Fan
	Joanne Lupton
	Bani Mallick
	Naisyin Wang
Head of Department,	Simon J. Sheather

August 2006

Major Subject: Statistics

ABSTRACT

Case-Control Studies of Genetic and Environmental Factors with Error in Measurement of Environmental Factors. (August 2006)

Iryna Lobach, B.S., Belarusian State University

Chair of Advisory Committee: Dr. Raymond J. Carroll

It is widely believed that risks of many complex diseases are determined by genetic susceptibilities, including environmental exposures, and their interaction. Chatterjee and Carroll (2005) have recently developed an efficient retrospective maximum-likelihood method for analysis of case-control studies that exploits an assumption of gene-environment independence and leaves the distribution of the environmental covariates to be completely nonparametric. We generalize the semiparametric maximum-likelihood approach to situations when some of the environmental covariates are measured with error and allow genetic information to be missing on some subjects, e.g., unphased haplotypes. Profile likelihood techniques and an EM algorithm are developed, resulting in a relatively simple procedure for parameter estimation. We prove consistency and derive the resulting asymptotic covariance matrix of parameter estimates when variance of measurement error is known and when it is estimated using replications. The performance of the proposed method is illustrated using simulation studies emphasizing the case when genetic information is in the form of a haplotype and missing data arises from haplotype-phase ambiguity and missing genetic data. Inference is performed via a likelihood-ratio type procedure, one that we show has better small-sample performance than Wald-type inferences. An application of this method is illustrated using a case-control study of an association of calcium intake with early stages of colorectal tumor development.

To my loving family

ACKNOWLEDGEMENTS

It was an honor for me to have worked with Professor Raymond Carroll. A few words cannot do justice to his contribution. However, I would like to thank him for sharing his experience and knowledge, which made the Ph.D. process enjoyable in addition to beneficial. I am deeply indebted to him for his patience.

I would like to thank all the people in the Department of Statistics at the Texas A&M University, who made A&M a very welcoming place and created an exciting research atmosphere.

My special thanks to Professor Fred Dahm and Marilyn Randall for walking me through the complex bureaucratic processes of coming to and staying in the United States to study.

I would like to express my gratitude to all my friends for the great moments we've shared during our Ph.D. years. Additionally, many thanks to all my old friends from Belarus with whom bonds of friendship became even tighter despite the distance and time.

And most importantly, I am deeply grateful to my parents and the rest of my family for their whole-hearted love and care, and for being very close to me while living far away.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

TABLE Page

TABLE Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

CHAPTER I

INTRODUCTION

1.1 Motivation

Many health conditions, including cancer and psychiatric disorders, are believed to have a complex genetic bias, and genes and environmental factors are likely to interact in the presence and severity of these conditions. With the advent of modern genotyping technologies, epidemiologists have been increasingly interested in identifying genetically defined subgroups within a population with unusual resistence or suceptibility in environmental exposures both because such interactions may yield insight into mechanisms of action of exposures and because they can suggest disease prevention strategies. Case-control studies using unrelated individuals may be an effective approach to identifying genetic variants underlying complex traits.

1.2 Gene-Environment Interactions

The key objective of research in human genetics is to advance knowledge of how genetic and environmental factors combine to cause disease. As Clayton and McKeigue (2001) define, in statistical terms, gene-environment interaction is present when the effect of genotype on disease risk depends on the level of exposure to an environmental factor, or vice versa. This definition depends on how effects on risk are measured. The most usual measure of effect in epidemiology is the ratio of disease incidence between exposed and unexposed individuals, which, in case-control studies can be measured by an odds ratio.

The format and style follow that of *Biometrics*.

Case-control studies are often used to model gene-environment interactions. In recent years, a number of researchers have noted that in case-control studies of genetic epidemiology, the efficiency of standard analysis can be improved upon by exploiting certain natural model assumptions for the underlying genetic and the environmental covariates. In the context of haplotype-based analysis of case-control studies, Epstein and Satten (2003) and Satten and Epstein (2004) noted retrospective maximum likelihood methods can be more efficient than analogous prospective methods by taking full advantage of an assumption of Hardy-Weinberg-Equilibrium (HWE) for the underlying population. Chatterjee and Carroll (2005) exploited an assumption of gene-environment independence to yield more precise maximum-likelihood estimates of the odds-ratio parameters than those obtained from standard logistic regression analysis.

1.3 Prospective Analysis of Case-Control Studies

The common practice in biostatistics is to analyse retrospectively collected data as if it were collected prospectively, ignoring the fact that under this design subjects are sampled retrospectively conditional of their disease status. The validity of this approach relies on the classic results by Cornfield (1956) who showed the equivalence of prospective- and retrospective odds-rations. The efficiency of the approach was established in two other classic papers by Anderson (1970) and Prentice and Pyke (1979) who showed that standard prospective analysis of case-control data yields the proper maximum-likelihood estimates of the odds-ratio parameter under the retrospecive design as long as the distribution of the underlying covaiates are allowed to remain completely unrestricted (nonparametric).

This point has received recent attention. Roeder at al. (1996) extended the approach to the case of measurement error. Muller and Roeder (1997) took a nonparametric view of the relationship between a surrogate (W) and the latent variable (X) and developed Bayesian procedure that is computationally complex. At the cost of requiring a parametric form for

the distribution of $(W|X, D)$ the approach of Gustafson et al. (2002) is considerably simpler. Gustafson et al. (2002) echo the call of Roeder et al. (1996) indicating a surprisingly subtle problem, namely how best to deal with additional precisely measured covariates. Seaman and Richardson (2001) developed an approach to deal with situations with any number of categorical covariates. They illustrate the drawbacks of Bayesian methods for continuous variables, namely the difficulty of numerical integration over high-dimentional space and therefore in practice they are limited to a small number of covariates. Moreover, it is difficult to say how the methods could be generalized to allow for both continuous and discrete covariates. To do this would require a sutable flexible and uninformative prior density for an exposure space that combines continuous and discrete components.

1.4 Measurement Error in Epidemiologic Studies

Intakes of various foods and drugs (such as fat, fiber, fruits, vegetables, etc.) are prime examples of exposure variables of considerable interest in medical studies, while also being hard to measure. In the studies with large number of subjects resource limitations might only allow for a food-frequency questionnaire (FFQ), on which subjects report the frequency with which they consume specific foods. Measurement error arises from the fact that participants have imperfect recall when completing questionnaires. Moreover, individuals with high levels of disease-related variables might have tendency to blame suspected risk factors for their condition, or deny the possibility those factors caused their conditions, what can result in differential measurement error, see for example Gustafson (2004) and Weinberg et al. (1994).

A number of large epidemiologic studies of relationship between diet and cancer failed to find a consistent relationship between dietary components and cancers of the breast, colon, or rectum (Hunter et al. 1996; Fuchs et al. 1999; Freudenheim et al. 1988; Michels et al. 2002). This maybe explained by the lack of a true diet-cancer associations, or,

alternatively, by serious methodological limitations of such studies, especially due to the FFQ measurement error.

Since FFQs are subject to substantial error, both systematic and random, it can profoundly affect the interpretation of nutritional epidemiologic studies. Dietary mismeasurement often attenuates the estimates of disease relative risks and reduces statistical power to detect their significance. Hence the important relationship between diet and disease may be obscured.

In many analyses that arise in biostatistics and epidemiology involve discrete response variable, i.e. disease status. In such cases logistic regression is the most common inferential procedure. There is considerable literature on measurement error in binary regression models, though some of this focuses on probit regression rather than logistic regression for the sake of numerical tractability (Carroll et al. 1984). Closed form expression for the bias induced by measurement error do not exist for the logistic regression model. Another situation when the impact of misclassification is complex and hard to intuit is when a polychotomous (categorical with more than two levels) exposure variable is subject to misclassification. As pointed by Dosemeci at al. (1990), even the impact of nondifferential misclassification can be quite unpredictable.

1.5 Haplotype-Based studies

Haplotype-based studies are becoming increasingly popular, a number of researches have developed methods for logistic regression analysis of case-control studies in the presence of phase ambiguity. One well-established method for estimating haplotype frequencies is the EM algorithm (see for example Excoffier and Slatkin (1995), Fallin and Schork (2000)). This algorithm is particulary useful in the context of tightly linked markers where the number of observed haplotypes is much smaller than the number than the number of theoretically possible haplotype frequencies. Epstein and Sattern (2003) present an approach based

on retrospective likelihood for case-control data that integrates over the observed phase assignments. Zhao et al. (2003) propose a similar estimating equations approach, although under a rare disease assumption they calculate frequencies using only controls. Starm et al. (2003) make use of case and control sampling fractions. Incorporation of environmental factors, however, is complicated in these approaches, because the retrospective likelihood involves potentially high dimentional nuisance parameters that specify the distribution of the environmental factors in the underlying population.

The methodology developed by Chatterjee and Carroll overcomes the majority of described above difficulties and has several unique aspects. First, it is exact and does not require a rare disease assumption, what is very important for studying major genes. Second, the setting is very flexable and retains all the flexability of the traditional logistic regression analysis, such as continuous exposures, complex modeling of the regression effects of the risk factors. Third, it allows incorporation of the external information about the probability of disease in the population, hence improves efficiency. Finally, the methodology is developed in a semiparametric framework that allows the distribution of the environmental factors to be fully nonparametric.

In this dissertation we will extend the profile likelihood approach proposed by Chatterjee and Carroll to develop a relatively simple procedure for obtaining the efficient retrospective maximum likelihood estimator for case-control studies with missing genetic data and measurement error in environmental covariates.

CHAPTER II

METHODOLOGY

2.1 Model and Notations

Let D be the categorical indicator of disease status. To be general, we allow D to have $K+1$ levels with the possibility of $K \geq 1$ to accommodate different subtypes of a disease. Let $D = 0$ denote the disease-free (control) subjects and $D = k$, $k \geq 1$ denote the diseased (case) subjects of the k-th subtype. Let $H^{\text{dip}} = (H_1, H_2)$ denote the diplotype status, that is, the two haplotypes a subject carriers at the loci of interest on the pair of homologous chromosomes. Suppose there are M loci of interest within a genomic region. Let $H^{\text{dip}} = (H_1, H_2)$ denote the corresponding diplotype status for an individual, that is, the two haplotypes the individual carries in his/her pair of homologous chromosomes. Let $E = (X, Z)$ denote all of the environmental (non-genetic) covariates of interest with X denoting the factors susceptible to measurement errors. Given the environmental covariates X and Z and diplotype data H^{dip}, the risk of the disease in the underlying population is given by the polytomous logistic regression model

$$\text{pr}(D = d \geq 1 | H^{\text{dip}}, X, Z) = \frac{\exp\left\{\beta_{0d} + m(H^{\text{dip}}, X, Z, \beta)\right\}}{1 + \sum_{j=1}^{K} \exp\left\{\beta_{0j} + m(H^{\text{dip}}, X, Z, \beta)\right\}}. \quad (2.1)$$

Here $m(\cdot)$ is a known function parameterizing the joint risk of the disease from H^{dip}, X and Z in terms of the odds-ratio parameters β.

The model (2.1) cannot be used directly for analysis due to two reasons. First, the dipotype information H^{dip} is not measurable using standard genotyping technology. Typically, multi-locus genotype information, denoted by $\mathbf{G} = (G_1, G_2, \ldots, G_M)$, is available. Due to lack of haplotype-phase information, the same genotype data can be consistent with multiple configuration of haplotypes for a given subject. For example, if A/a and B/b de-

note the major/minor alleles in two bi-allelic loci, then subjects with genotypes (Aa) and (Bb) at the first and the second locus, respectively, are considered "phase ambiguous": their genotypes could arise from either the haplotype-pair (A-B,a-b) or the haplotype-pair (A-b, a-B). Let $\mathcal{H}^{\rm dip}$ denote the set of all possible diplotypes in the underlying population. Analogously, let $\mathcal{H}^{\rm dip}_{\mathbf{G}}$ denote the set of all possible diplotypes that are consistent with a particular genotype vector $\mathbf{G}$. We assume independence of $H^{\rm dip}$ and $E = (X, Z)$ in the underlying population. Moreover, we assume a parametric model of the form $\text{pr}(H^{\rm dip}) = Q(H^{\rm dip}, \theta)$. Note however that our method can be readily extended to a general parametric model for $H^{\rm dip}$ given (X, Z). For our numerical examples, we assume HWE so that the distribution of the diplotypes can be specified in terms of the frequency of the haplotypes. Our general framework, however, allows use of more flexible models than HWE (see e.g. Satten and Epstein, 2004; Lin and Zeng, 2006).

A second problem is that in our motivating example, covariate X is measured with error. Let W denote the error-prone version of X. We assume a parametric model of the form $f_{\rm mem}(w|X, H^{\rm dip}, Z, D; \xi)$ for the conditional distribution of W given the true exposure X, additional environmental factors Z and disease-status D. If measurement error can be assumed to be non-differential by disease status, then one can simplify the model as $f_{\rm mem}(w|X, H^{\rm dip}, Z, D; \xi) = f_{\rm mem}(w|X, H^{\rm dip}, Z; \xi)$. We assume that the joint distribution of the environmental factors in the underlying population can be specified according to a semiparametric model of the form $f_{X,Z}(x, z) = f_X(x|z; \eta) f_Z(z)$ where $f_Z(z)$ is left completely unspecified.

2.2 Semiparametric Inference Based on a Pseudo-likelihood

For $d \geq 1$, define n_d to be the number of subjects in the sample with disease at stage d, $\pi_d = n_d/\text{pr}(D = d)$, $\kappa_d = \beta_{0d} + \log(n_d/n_0) - \log(\pi_d/\pi_0)$, and $\widetilde{\kappa} = (\kappa_1, ..., \kappa_K)^{\rm T}$. Define $\kappa_0 = \beta_{00}$. Let $\widetilde{\beta}_0 = (\beta_{01}, ..., \beta_{0K})^{\rm T}$. Let $\Omega = (\widetilde{\beta}_0^{\rm T}, \beta^{\rm T}, \Theta^{\rm T}, \widetilde{\kappa}^{\rm T})^{\rm T}$, $\mathcal{B} = (\Omega^{\rm T}, \eta^{\rm T})^{\rm T}$ and

$\upsilon = (\eta^{\mathrm{T}}, \xi^{\mathrm{T}})^{\mathrm{T}}$. Make the definition

$$S(d, h^{dip}, x, z, \Omega) = \frac{\exp\left[I_{(d\geq 1)}(d)\left\{\kappa_d + m(h^{\mathrm{dip}}, x, z, \beta)\right\}\right]}{1 + \sum_{j=1}^{K} \exp\left\{\beta_{0j} + m(h^{\mathrm{dip}}, x, z, \beta)\right\}} Q(h^{\mathrm{dip}}, \Theta).$$

Consider a sampling scenario where each subject from the underlying population is selected into the case-control study using a Bernoulli sampling scheme, where the selection probability for a subject given his/her disease status $D = d$ is proportional to $\pi_d = n_d/\mathrm{pr}(D = d)$. Let $R = 1$ denote the indicator of whether a subject is selected in the case-control sample under the above Bernoulli sampling scheme. We propose parameter estimation using a pseudo-likelihood of the form

$$L^* = \prod_{i=1}^{N} \mathrm{pr}(D_i, W_i, \mathbf{G}_i | Z_i, R = 1)$$

Calculations given in the Appendix show that

$$\begin{aligned} L(d, g, w, z, \Omega, \eta, \xi) &\equiv \mathrm{pr}(D = d, W = w, \mathbf{G} = g | Z = z, R = 1) \\ &= \frac{\int \sum_{h^{\mathrm{dip}} \in \mathcal{H}_G^{\mathrm{dip}}} S(d, h^{\mathrm{dip}}, x, z, \Omega) f_{\mathrm{mem}}(w|d, h^{\mathrm{dip}}, x, z, \xi) f_X(x|z, \eta) dx}{\int \sum_{d_*=0}^{K+1} \sum_{h_*^{\mathrm{dip}} \in \mathcal{H}^{\mathrm{dip}}} S(d_*, h_*^{\mathrm{dip}}, x, z, \Omega) f_X(x|z, \eta) dx}. \end{aligned} \tag{2.2}$$

We observe that conditioning on Z in L^* allows it to be free of the non-parametric density function $f_Z(z)$, thus avoiding the need of estimating potentially high-dimensional nuisance parameters.

2.3 Population Stratification

Genetic association analysis of candidate genes can identify gene variants that are associated with disease by comparing allele frequencies of presumably biologically relevant genes in affected and control individuals. In a case-control study, when the allele frequencies of affected individuals are compared to those of controls, the control population should be carefully selected since there are natural variations in gene fequencies that occur between between ethnic groups. Hence this phenomenon should be taken into account and the presented methodology allows for it.

Moreover, gene-environment assumption can be relaxed by modeling genotype and environment conditionally on strata. Genetic susceptability and environmental exposures could be correlated on population level because of their dependence on other factors, such as ethnicity, smoking status, gender etc.

The proposed methodology can easily account for stratification. It would be necessary to spesify the distribution of genotype and possibly environment conditionally on strata S. We also allow the stratum covariate to be in the disease-risk model. The likelihood function then becomes

$$\begin{aligned} &\text{pr}(D = d, W = w, G|Z = z, S = s, R = 1) \\ &\quad = \frac{\int \sum_{h \in \mathcal{H}_G} S(d, h, x, z, s, \Omega) f_U(w|d, h, x, z, s, \xi) f_X(x|z, s, \eta) dx}{\int \sum_{d_*} \sum_{h_*} S(d_*, h_*, x, z, s, \Omega) f_X(x|z, s, \eta) dx} \\ &\quad = L(d, g, w, z, \Omega, \eta, \xi). \end{aligned} \tag{2.3}$$

The development of the methodology remains similar just with slight change in notation. An illustrative example of stratification is the Israeli Ovarian Cancer Study that can be found in Chatterjee and Carroll (2005).

CHAPTER III

ESTIMATION

3.1 Introduction

If a random sample has been drawn from a population involving unknown parameters, the latter may be estimated from the sample by the well-known technique of maximum likelihood that was first introduced and then extensively studies by R.A. Fisher. In this section we describe and investigate maximum likelihood estimating procedure based on the semiparametric likelihood function (2.3).

The widely used method to estimate parameters based on an incomplete data is the EM algorithm. The EM process is very attractive in part because of simplicity and generality of the theory, and in part because of the wide application. One of the earliest papers on EM algorithm is Hartley (1958), but the reference that formalized it and provided a proof of convergence is Dempster et al. (1977). The EM algorithm cosissts of two steps: E and M. The E-step requires us to estimate unobserved components given the observed and the current fitted parameters. The M-step is then quivalent to the complete-data maximization. In what follows the detailed description of E-steps is provided.

This chapter is organized as follows. In Section 3.2 we describe estimation procedure for the case when measurment error distribution is known. Section 3.3 provides the estimation procedure in the case when measurement error process is estimated using external replications. Section 3.4 describes steps of the EM algorithm.

3.2 Estimation with Known Measurement Error Distribution

In this section, we assume that the parameter ξ controlling the distribution of the measurement error is known. We show that maximization of L^*, although it is not the actual

retrospective-likelihood for case-control data, leads to consistent and asymptotically normal parameter estimates. Recall that $\mathcal{B} = (\Omega^{\mathrm{T}}, \eta^{\mathrm{T}})^{\mathrm{T}}$. Let $\Psi(d, g, w, z, \Omega, \eta, \xi)$ be the derivative of $\log\{L(d, g, w, z, \Omega, \eta, \xi)\}$ with respect to $\mathcal{B}$. Then define

$$\begin{aligned}
\mathcal{L}_n(\Omega, \eta, \xi) &= \sum_{i=1}^{n} \Psi(D_i, G_i, W_i, Z_i, \Omega, \eta, \xi); \\
\mathcal{I} &= -n^{-1} E\left[\partial\{\mathcal{L}_n(\Omega, \eta, \xi)\}/\partial\mathcal{B}^{\mathrm{T}}\right]; \\
\Lambda &= \sum_{d} \frac{n_d}{n} E\left\{\Psi(D, G, W, Z, \Omega, \eta, \xi)|D = d\right\} \\
&\qquad \times E\left\{\Psi(D, G, W, Z, \Omega, \eta, \xi)|D = d\right\}^{\mathrm{T}},
\end{aligned}$$

where all expectations are taken with respect to the case-control sampling design. We propose to estimate $\mathcal{B}$ as the solution to

$$0 = \mathcal{L}_n(\Omega, \eta, \xi) = \mathcal{L}_n(\mathcal{B}, \xi), \tag{3.1}$$

calling the solution $\widehat{\mathcal{B}} = (\widehat{\Omega}^{\mathrm{T}}, \widehat{\eta}^{\mathrm{T}})^{\mathrm{T}}$. Our main technical result, the proof of which is given in the Appendix, is the limiting properties of $\widehat{\mathcal{B}}$.

THEOREM **1.** The estimating function $\mathcal{L}_n(\Omega, \eta, \xi)$ is unbiased, i.e., has mean zero when evaluated at the true parameter values. In addition, under suitable regulatory conditions, there is a consistent sequence of solution to (3.1), with the property that

$$n^{1/2}(\widehat{\mathcal{B}} - \mathcal{B}) \Rightarrow \text{Normal}\{0, \mathcal{I}^{-1}(\mathcal{I} - \Lambda)\mathcal{I}^{-1}\}. \tag{3.2}$$

Remark 1: It is easy to obtain consistent estimates of both $\mathcal{I}$ and Λ. For example, to get an estimate $\widehat{\Lambda}$, in the definition of Λ, we can estimate $E\left\{\Psi(D, G, W, Z, \Omega, \eta, \xi)|D = d\right\}$ by $n_d^{-1}\sum_{i=1}^{n} I(D_i = d)\Psi(d, G_i, W_i, Z_i, \widehat{\mathcal{B}}, \xi)$. Similarly, $n^{-1}\partial\{\mathcal{L}_n(\widehat{\mathcal{B}}, \xi)\}/\partial\mathcal{B}^{\mathrm{T}}$ is a consistent estimate of $\mathcal{I}$. Alternatively, if $\widehat{\Sigma}$ is the sample covariance matrix of the terms $\Psi(D_i, G_i, W_i, Z_i, \widehat{\mathcal{B}}, \xi)$, then $\widehat{\Sigma} + \widehat{\Lambda}$ consistently estimates $\mathcal{I}$.

Remark 2: An EM-algorithm for computation, based along the lines of Spinka, et al. (2005) is given in the Appendix.

Remark 3: Similar to the settings of Chatterjee and Carroll (2005) and Spinka et al (2005), here, the intercept parameters ($\beta_{0d}, d \geq 1$) of the polytomous logistic regression model are theoretically identifiable from the pseudo-likelihood L^*, even though the sampling is retrospective. For rare diseases, however, $1 + \sum_{j=1}^{K} \exp\left\{\beta_{0j} + m(H^{\rm dip}, X, Z, \beta)\right\} \approx 1$ in and so L^* is expected to contain very little information about β_d. If information on $\Pr(D = d)$, is available externally, as could be the situation for population-based case-control studies, then π_d, $d \geq 1$, could be treated as fixed known parameters in the definition of κ_d allowing estimation of β_{0d} to be much more tractable. If $\Pr(D = d)$ is not known, one could employ the rare disease assumption under which β_{0d}'s disappear from the likelihood. Alternatively, one can estimate parameters (Ω, η, ξ) by maximizing the likelihood function for the values of π_d fixed on a grid and then performing a grid-search method to indentify the value of π_d that maximizes the profile likelihood $\mathcal{L}_n\{\Omega(\pi_d), \eta(\pi_d), \xi\}$.

3.3 Estimated Measurement Error Distribution

In practice, the parameter ξ controlling the measurement error distribution will be unknown, and typically additional data are necessary to estimate it. Here we consider the case of additive mean-zero measurement error with replications of W.

Our convention is that there are at most M replications of the W for any individual. Let W_i denote this ensemble of the M replicates, and let m_i be the number of replicates we actually observe. Let $f_{\rm mem}(w|d, h^{\rm dip}, x, z, m, \xi)$ be the joint density of the first m replicates for $m = 1, \cdots, M$; $\Psi(D, G, W, Z, \Omega, \eta, \xi, j)$, $\mathcal{I}_j$, and Λ_j be matrices defined in the Section 3.2 for the case with exactly $m = j$ replicates for each individual. Assume that m_i is independent of $(D_i, W_i, Z_i, G_i, X_i, H_i^{\rm dip})$ and that $\Pr(m_i = j) = p(j)$. Further, define $\mathcal{I} = \sum_{j=1}^{M} p(j)\mathcal{I}_j$. It is shown in appendix that the estimating function for $\mathcal{B} = (\Omega^{\rm T}, \eta^{\rm T}, \xi^{\rm T})^{\rm T}$

can be written in the form

$$0 = \sum_{i=1}^{n} \sum_{j=1}^{M} I_{(m_i=j)}(m_i) \Psi(D_i, G_i, W_i, Z_i, \Omega, \eta, \xi, j). \tag{3.3}$$

THEOREM **2.** The estimating function (3.3) is unbiased, i.e., has mean zero when evaluated at the true parameter values. In addition, under suitable regulatory conditions, there is a consistent sequence of solutions to (3.3), with the property that

$$n^{1/2}(\widehat{\mathcal{B}} - \mathcal{B}_0) \Rightarrow \text{Normal}[0, \mathcal{I}^{-1}\{\mathcal{I} - \sum_{j=1}^{M} p(j)\Lambda_j\}\mathcal{I}^{-1}]. \tag{3.4}$$

Remark 4: Consistent estimates of $\mathcal{I}$ and Λ_j can be obtained by applying formulas that are analagous to those outlined in the Remark 1.

3.4 EM Steps

In this section, we describe an EM algorithm for solving the score-equations associated with the pseudo-likelihood L^*. To facilitate the calculations, make the following definitions:

$$\begin{aligned}
T(d, h^{\text{dip}}, x, z, \Omega) &= \frac{\exp\left[I_{(d\geq 1)}(d)\left\{\kappa_d + m(h^{\text{dip}}, x, z, \beta)\right\}\right]}{1 + \sum_{j=1}^{K} \exp\left\{\beta_{0j} + m(h^{\text{dip}}, x, z, \beta)\right\}}; \\
\alpha(h^{\text{dip}}, d, z, w, \mathcal{B}, \xi) &= \int T(d, h^{\text{dip}}, x, z, \Omega) f_{\text{mem}}(w|d, h^{\text{dip}}, x, z, \xi) f_X(x|z, \eta) dx; \\
\gamma(h^{\text{dip}}, z, \mathcal{B}) &= \int \sum_{d_*} T(d_*, h^{\text{dip}}, x, z, \Omega) f_X(x|z, \eta) dx; \\
V_\beta(d, h^{\text{dip}}, x, z, \Omega) &= \frac{\partial m(h^{\text{dip}}, x, z, \beta)}{\partial \beta} \\
&\quad \times \left[\frac{1}{1 + \sum_{j=1}^{K} \exp\{\beta_{0j} + m(h^{\text{dip}}, x, z, \beta)\}} - I(d = 0)\right].
\end{aligned}$$

Note that neither $\alpha(\bullet)$ nor $\gamma(\bullet)$ depend on Θ.

We split up the EM calculations into a series of steps. All technical arguments are given in the Appendix C.

EM Algorithm for Θ Under Hardy - Weinberg Equilibrium, if θ_i is the frequency of haplotype h_i, pr$\{H^{\rm dip} = (h_i, h_j)|\Theta\} = \theta_i^2$ if $h_i = h_j$ and $= 2\theta_i\theta_j$ if $h_i \neq h_j$. Let $N_k(H^{\rm dip})$ be the number of copies of h_k in $H^{\rm dip}$, and note that as in Spinka, et al., $N_k(H^{\rm dip})/\theta_k = \partial \log\{{\rm pr}(H^{\rm dip})\}/\partial\theta_k$. Define

$$\begin{aligned}\mathcal{N}_k(\mathcal{B}) &= \sum_{i=1}^{n} E_{\mathcal{B}}\{N_k(H^{\rm dip})|G_i, D_i, W_i, Z_i, R_i = 1\};\\ \mathcal{V}_k(\mathcal{B}) &= 2\sum_{i=1}^{n} \frac{\sum_{h_s} Q\{(h_k, h_s), \Theta\}\gamma\{(h_k, h_s), Z_i, \mathcal{B}\}}{\sum_{h^{\rm dip}} Q(h^{\rm dip}, \Theta)\gamma(h^{\rm dip}, Z_i, \mathcal{B})}.\end{aligned}$$

Then if $\mathcal{B}^{(s)}$ is the current value of $\mathcal{B}$, we update θ_k to $\theta_k^{(s+1)}$ as

$$\theta_k^{(s+1)} = \mathcal{N}_k(\mathcal{B}^{(s)})\{\mathcal{V}_k(\mathcal{B}^{(s)})\}^{-1}. \tag{3.5}$$

Further, in each iteration we normalize $\theta_k^{(t+1)} = \theta_k^{(t+1)}/\sum_{k'=1}^{K_\Theta}\theta_{k'}^{(t+1)}$.

EM Algorithm For κ_d and β For $j = 1, .., K$, we update κ_j by solving the following equation for κ_j:

$$n_j = \sum_{i=1}^{n} \frac{\int \sum_{h_*^{\rm dip}} \sum_{d_*} I_{(d_*=j)}(d_*) S(d_*, h_*^{\rm dip}, x, Z_i, \Omega) f_X(x|Z_i, \eta) dx}{\int \sum_{h_*^{\rm dip}} \sum_{d_*} S(d_*, h_*^{\rm dip}, x, Z_i, \Omega) f_X(x|Z_i, \eta) dx}. \tag{3.6}$$

To update β, we solve

$$\begin{aligned}0 = \sum_{i=1}^{n} &\frac{\int \sum_{h^{\rm dip}\in\mathcal{H}_G^{\rm dip}} V_\beta(d, h^{\rm dip}, x, z_i, \Omega) S(d_i, h^{\rm dip}, x, z_i, \Omega)}{\int \sum_{h^{\rm dip}\in\mathcal{H}_G^{\rm dip}} S(d_i, h^{\rm dip}, x, z_i, \Omega) f_{\rm mem}(w|d_i, h^{\rm dip}, x, z_i, \xi) f_X(x|z_i, \eta) dx}\\ &\times f_{\rm mem}(w|d_i, h^{\rm dip}, x, z_i, \xi) f_X(x|z_i, \eta) dx\\ -\sum_{j=1}^{n} &\frac{\int \sum_{d_*} \sum_{h_*^{\rm dip}} V_\beta(d, h_*^{\rm dip}, x, z_i, \Omega) S(d_*, h_*^{\rm dip}, x, z_i, \Omega) f_X(x|z_i, \eta) dx dz}{\int \sum_{d_*} \sum_{h_*^{\rm dip}} S(d_*, h_*^{\rm dip}, x, z_i, \Omega) f_X(x|z_i, \eta) dx}.\end{aligned} \tag{3.7}$$

EM Algorithm for β_{0d} and η The updating schemes for β_{0d} and η are of the form (3.7) with $V_\beta(\bullet)$ replaced by $V_{\beta_{0d}}(d, h^{\rm dip}, x, z, \Omega) = -{\rm pr}(D = d \geq 1|h^{\rm dip}, x, z)$ and $V_\eta(x, z, \eta) = \partial \log\{f_X(x|z, \eta)\}/\partial\eta$ for β_{0d} and η, respectively.

3.5 Discussion

In our development, we have used a parametric model for the distribution of the environmental covariate measured with error, but we have not specified the form of this distribution. Our simulations and the example were based upon normal distributions, which seem reasonable in this context, but clearly more general models are possible, e.g., the semi-nonparametric family of Zhang and Davidian (2001). While such parametric assumptions can be wrong, often the resulting inferences are not badly affected, especially for logistic regression. For example, in the running Framingham data example in Carroll, et al. (2006), the underlying variable X (transformed systolic blood pressure) appears to be more accurately modeled by a t-distribution with 5 degrees of freedom, but the differences in inference compared to a normal distribution assumption are hardly noticeable.

For logistic regression when W is unbiased for X and with normally distributed measurement error, there are possible methods that can in principle avoid the use of distributional assumptions. The most widely used approach aimed at achieving this nonparametric feature is that of Stefanski and Carroll (1987), who use conditioning on sufficient statistics. Unfortunately, this approach will not work in our context, because in gene-environment interaction studies the sufficient statistic includes the underlying genetic variable, and hence cannot be allowed to be missing. Other methods that might be employed are SIMEX (Cook and Stefanski, 1995; Carroll, et al., 2006) and Monte-Carlo Corrected Scores (MCCS, Stefanski, et al., 2005, Carroll, et al., 2006). Neither method results in actual consistent estimation of the parameters, although the latter is generally close to unbiased. However, MCCS requires the use of complex variable calculations, which users may find to be a practical hinderance.

CHAPTER IV

INFERENCE

4.1 Introduction

This chapter concerns likelihood-ratio (LR) type inference for gene-environment interactions based on case-control studies when genetic information may be missing and some of the environmental variables are measured with error, thus causing biases in parameter estimates and possibly incorrect inferences, see Carroll, et al. (2006). Traditionally, case-control data are analyzed using prospective logistic regression method ignoring the fact that under this design subjects are sampled retrospectively conditional on their disease status. Hence the unique feature of the LR type inference procedure under investigation is that it should account for the fact that the data do not come from the parametric model the likelihood function is based upon, environmental factors are measured with error and genetic factor has missing values.

The Calcium Study we are investigating thus has the unique features described above, specifically the following.

- First, genetic information is missing. We wish to model the effect of CaSR haplotypes, but these are not observed, and instead we have unphased haplotype information in the form of the three SNPs. In haplotype-based studies, where the effect of a gene is studied in terms of 'haplotypes', the combination of alleles at multiple loci along individual chromosome, missing data arises due to intrinsic "phase ambiguity" of the locus-specific genotype data.

- Second, one of the environmental variables (calcium intake) is subject to substantial measurement error because of the use of a FFQ. It is well known that the FFQ as a

measure of long-term diet is subject both to biases and random errors, as illustrated in the OPEN study (Subar, et al., 2003).

In this setting it is undesirable to conduct inferences using the Wald-type procedure. Schafer and Purdy (1996) advocated likelihood analysis for regression models with errors in explanatory variables, for data problems in which the relevant distributions can be adequately modeled. They point out that the likelihood ratio tests and confidence intervals can be substantially better than tests and confidence intervals based on estimates and standard errors, since the sampling distribution of measurement-error corrected estimators are very often skewed, especially if the measurement errors are large.

4.2 Inference via Likelihood Ratio Techniques

The LR procedure for testing

$$H_0 : \mathcal{B} \in \mathcal{B}_0;$$
$$H_1 : \mathcal{B} \in \mathcal{B}_1, \tag{4.1}$$

is based on the following statistic

$$\lambda_n = \sup_{\mathcal{B}\in\mathcal{B}_0} \mathcal{L}_n(\mathcal{B}, \xi) / \sup_{\mathcal{B}\in\mathcal{B}_1} \mathcal{L}_n(\mathcal{B}, \xi). \tag{4.2}$$

Under the assuption of a correct model Wilks (1938) and Roy (1957) derived the limiting chi-square distribution of $-2\log(\lambda_n)$ using consistency and asymptotic normality of the maximum likelihood estimates. Kent (1982) examined the distribution of the LR statistic when the data do not come from the parametric model, but when the 'nearest' member of the parametric family still satisfies the null hypothesis. These arguments can be extended to find the limiting distribution of $-2\log(\lambda_n)$ based on a likelihood-type function, provided consistency and limiting distribution of the maximum likelihood has been established.

As it was proved in Theorems 1 and 2, the limiting covariance matrix of parameters estimates is in the form $\mathcal{I}^{-1}(\mathcal{I}-\Lambda)\mathcal{I}^{-1}$ and needs to be accounted for.

In what follows we discuss a likelihood ratio test procedure for testing simple and composite hypothesis based on the likelihood function (2.2). The critical technical part is that the data does not come from the parametric model the likelihood function is based on. Hence the asymototic distribution of the LR test statistic needs to be adjusted to take the sampling design into account.

4.2.1 Simple Hypothesis

First consider the null hypothesis of the form $\mathcal{B}=\mathcal{B}_0$. If the second derivative of $\mathcal{L}_n(\mathcal{B})$ is given as $\mathcal{L}_{\mathcal{B}\mathcal{B}}(\cdot)$, denote $\mathcal{S}^{-1}=\mathcal{I}^{-1}(\mathcal{I}-\Lambda)\mathcal{I}^{-1}$, then the estimate $\widehat{\mathcal{B}}$ satisfies

$$\mathcal{I}+o_p(1) \;=\; n^{-1}\mathcal{L}_{\mathcal{B}\mathcal{B}}(\mathcal{B}_0); \tag{4.3}$$

$$n^{1/2}(\widehat{\mathcal{B}}-\mathcal{B}_0) \;\Rightarrow\; \text{Normal}\left(0,\mathcal{S}^{-1}\right). \tag{4.4}$$

Our main technical result, the proof of which is given in the Appendix, is a limiting property of the test (4.1) based on a likelihood-type function $\mathcal{L}_n(\mathcal{B})$.

THEOREM **3.** Define $\mathcal{V}=\text{Normal}\,(0,\mathcal{S}^{-1})$. Using Cholesky decomposition the covariance matrix can be factored as $\mathcal{S}^{-1}=LL^{\mathrm{T}}$, where L is a lower-triangular matrix. Let $\lambda_i,\ i=1,\ldots,k$ be eigenvalues of the matrix $L\mathcal{I}L^{\mathrm{T}}$. Let $Z_1^2,Z_2^2,\ldots,Z_k^2$ denote independent χ_1^2 random variables. Then when H_0 is true, the likelihood-ratio type statistic based on the pseudo-likelihood $\mathcal{L}_n(\bullet)$ has the limiting distribution that is the same as

$$\mathcal{V}^{\mathrm{T}}\mathcal{I}\mathcal{V}\sim\sum_{i=1}^{k}\lambda_i Z_i^2. \tag{4.5}$$

Remark 5: To estimate λ_i's we propose to apply Cholesky decomposition to $\widehat{\mathcal{S}}=\widehat{L}\widehat{L}^{\mathrm{T}}$ and obtain $\widehat{\lambda}$'s as eigenvalues of $\widehat{L}\widehat{\mathcal{I}}\widehat{L}^{\mathrm{T}}$.

4.2.2 Composite Hypothesis

Let $\mathcal{B} = (\delta, \gamma)$, where δ is an r dimentional vector of interest and γ is $(k - r)$ dimentional nuisance vector. Let the null hypothesis be $\delta = \delta_0$ whatever γ may be. Define $\mathcal{B}_0 = \{(\delta, \gamma) : \mathcal{B} = \mathcal{B}_0, \gamma \in \Gamma\}$ and $\mathcal{B}_1 = \{(\delta, \gamma) : \mathcal{B} \neq \mathcal{B}_0, \gamma \in \Gamma\}$. Here we investigate the likelihood ratio test for (4.1) based on a likelihood $\mathcal{L}_n(\bullet)$.

Define S_{11} and S_{22} to be diagonal blocks of the covariance matrix $\mathcal{S}$ that correspond to parameters of interest and nuisance paramaters, respectively. Similary, the corresponding blocks of $\mathcal{I}$ are $\mathcal{I}_{11}$ and $\mathcal{I}_{22}$. Let $\mathcal{C} = \mathcal{S}_{11} - \mathcal{S}_{12}\mathcal{S}_{22}^{-1}\mathcal{S}_{21}$, $\mathcal{J} = \mathcal{I}_{11} - \mathcal{I}_{12}\mathcal{I}_{22}^{-1}\mathcal{I}_{21}$ and using Frobenius formula it can be easily seen that

$$n^{1/2}(\widehat{\delta} - \delta_0) \;\; \Rightarrow \;\; \text{Normal}\left(0, \mathcal{C}^{-1}\right).$$

The following theorem is an analog of the Theorem 1 for the case of composite hypothesis.
THEOREM **4.** Define $\mathcal{V}_1 = \text{Normal}\,(0, \mathcal{C}^{-1})$. Using Cholesky decomposition the covariance matrix can be factored as $\mathcal{C}^{-1} = LL^{\mathrm{T}}$, where L is a lower-triangular matrix. Let $\lambda_i,\ i = 1, \ldots, r$ be eigenvalues of the matrix $L\mathcal{J}L^{\mathrm{T}}$. Let $Z_1^2, Z_2^2, \ldots, Z_r^2$ denote independent χ_1^2 random variables. Then under H_0 likelihood-ratio type statistic based on the pseudo-likelihood $\mathcal{L}_n(\bullet)$ has the limiting distribution that is the same as

$$\mathcal{V}_1{}^{\mathrm{T}}\mathcal{J}\mathcal{V}_1 \sim \sum_{i=1}^{r} \lambda_i Z_i^2. \tag{4.6}$$

Remark 5: To estimate λ_i's we propose to apply procedure analogous to the one described in the Remark 1.

CHAPTER V

SIMULATION STUDY

5.1 The Binary Case

When all variables are binary, it is possible to compute the retrospective likelihood of case-control data. In this section, we compare of our pseudo-likelihood method with those based on the full retrospective likelihood, in the case that $H^{\rm dip}$ is binary and directly observable.

In this case, there are no covariates Z, the retrospective likelihood is given as follows. Define $H\{\beta_0 + m(h^{\rm dip}, x, \beta)\} = \text{pr}(D = 1|X, H^{\rm dip})$ and $\mathcal{H}(d, h^{\rm dip}, x, \beta_0, \beta) = [H\{\beta_0 + m(h^{\rm dip}, x, \beta)\}]^d[1 - H\{\beta_0 + m(h^{\rm dip}, x, \beta)\}]^{1-d}$. Then

$$\begin{aligned}&\text{pr}(W = w, H^{\rm dip} = h^{\rm dip}|D = d)\\ &\quad = \frac{\int \mathcal{H}(d, h^{\rm dip}, x, \beta_0, \beta)Q(h^{\rm dip}, \Theta)f_{\rm mem}(w|d, h^{\rm dip}, x, \xi)f_X(x|\eta)dx}{\int \sum_{h_*^{\rm dip}} \mathcal{H}(d, h_*^{\rm dip}, x, \beta_0, \beta)Q(h_*^{\rm dip}, \Theta)f_X(x|\eta)dx}.\end{aligned}$$

Because we have specified a distribution for X, all variables bare binary, and there is no Z, the parameters $(\beta_0, \beta, \theta, \eta)$ are sufficient to identify $\text{pr}(D = 1)$, i.e.,

$$\text{pr}(D = 1) = \int \sum_{h_*^{\rm dip}} H\{\beta_0 + m(h_*^{\rm dip}, x, \beta)\}Q(h_*^{\rm dip}, \Theta)f_X(x|\eta)dx. \tag{5.1}$$

Because of this, κ is identified from $(\beta_0, \beta, \theta, \eta)$ as well. Hence, simply using (2.3) as a likelihood function will be unstable. The obvious solution is to replace both β_0 and κ in (2.3) by the appropriate functions of $\text{pr}(D = 1)$ as given in (5.1) and the definition of κ, which is what we did.

We did a small simulation experiment in order to illustrate our approach in this simple case. We assumed that environmental variables (X, W), genetic variant (G), and disease status (D) are binary. Given the values of (G, X) we generated a binary disease outcome D from the following logistic model $\text{logit}\{\text{pr}(D|G, X)\} = \beta_0 + \beta_x X + \beta_g G + \beta_{xg} X * G$,

with parameters $(\beta_x, \beta_g, \beta_{xh}) = (1.099, 0.693, 0.693)$. The misclassification probabilities were pr$(W = 0|X = 1) = 0.20$ and pr$(W = 1|X = 0) = 0.10$.

We estimated parameters using the foregoing algorithm and investigated the effect of knowing the probability of disease. We found that our proposed method yielded estimates that were numerically identical to those based on the full retrospective likelihood: we believe but have not been able to show that this is true in general. Our method showed no noticeable bias in the parameter estimates, either in the risk parameters or in the genotype probabilities, whereas the naive analysis resulted in large biases (Tables 1 and 2).

Table 1: Biases and root mean squared errors for the ordinary logistic regression, retrospective and semiparametric (proposed) approaches, where disease status (D), genetic variant (G), and environmental covariate (X) are binary and probability of disease is known. Environmental variable is measured with error with misclassification probabilities pr$(W = 0|X = 1) = 0.20$ and pr$(W = 1|X = 0) = 0.10$. The results are based on a simulation study with 500 replications for 1000 cases and 1000 controls.

		Logistic		Retrospective		Semiparametric	
Parameter	True Value	Bias	RMSE	Bias	RMSE	Bias	RMSE
β_0	-5.000	4.294	4.295	-0.006	0.108	-0.006	0.108
β_g	0.693	0.239	0.323	-0.005	0.305	-0.004	0.305
β_x	1.099	-0.327	0.344	0.005	0.155	0.005	0.155
β_{xg}	0.693	-0.284	0.395	0.001	0.327	0.001	0.327
pr$(X = 1)$	0.100			0.002	0.021	0.002	0.022
pr$(G = 1)$	0.100			0.000	0.009	0.000	0.008

Table 2: Biases and root mean squared errors for the ordinary logistic regression, retrospective and semiparametric (proposed) approaches, where disease status (D), genetic variant (G), and environmental covariate (X) are binary and probability of disease is unknown. Environmental variable is measured with error with misclassification probabilities $\text{pr}(W = 0|X = 1) = 0.20$ and $\text{pr}(W = 1|X = 0) = 0.10$. The results are based on a simulation study with 500 replications for 1000 cases and 1000 controls.

		Logistic		Retrospective		Semiparametric	
Parameter	True Value	Bias	RMSE	Bias	RMSE	Bias	RMSE
β_0	-5.000	4.294	4.295	-1.016	2.042	-1.016	2.042
β_g	0.693	0.239	0.323	-0.009	0.306	-0.009	0.306
β_x	1.099	-0.327	0.344	0.004	0.155	0.004	0.155
β_{xg}	0.693	-0.284	0.395	0.013	0.333	0.013	0.333
$\text{pr}(X = 1)$	0.100			0.023	0.022	0.002	0.022
$\text{pr}(G = 1)$	0.100			0.000	0.009	0.000	0.009
$\text{pr}(D = 1)$	0.016			0.002	0.019	0.002	0.019

5.2 Continuous Simulations

In this simulation, we considered a continuous environmental variables and assumed that the genetic risk depends on the number of copies of a putative haplotype. We simulated the true environmental covariate (X) from Normal distribution with zero mean and variance 0.1. To simulate observed environmental variables we used additive model of the form $W = X + U$, where U is generated from the Normal distribution with zero mean and variance $\xi = 0.25$. Given the following haplotype frequencies $(h_1, h_2, h_3, h_4, h_5, h_6) = (0.25, 0.15, 0.25, 0.1, 0.1, 0.15)$ we generated diplotypes for each subject under the assumption of Hardy-Weinberg Equilibrium. Given the diplotype information H^{dip} and environmental covariate X we generated binary disease status according to the following model

$$\text{logit}\{\text{pr}(D = d|H^{\text{dip}}, X)\} = \beta_0 + \beta_x X + \beta_g N_3(H^{\text{dip}}) + \beta_{xg} X N_3(H^{\text{dip}})$$

where $N_3(H^{\text{dip}})$ is the number of copies of h_3 in H^{dip}. In this setting we are interested in estimating the relative risk parameters and the frequency of haplotype h_3. To estimate probability of disease we used grid-search method by maximizing likelihood function for

values of probability of disease fixed on a grid and then performing grid-search method to identify the value of probability of disease that maximized the likelihood. Moreover, we assessed the effect of missing data by assuming that 50% of subjects were not genotyped and for those who were genotyped linkage phase is unknown.

We found that for our method there is no noticeable bias in parameter estimates, whereas the naive approach that ignores existence of the measurement error results in substantial bias, as illustrated in the Table 3. It is somewhat remarkable that even with 50% of the genotypes are missing, our method still remains largely unbiased.

Table 3: Biases and root mean squared errors for the naive approach that ignores existence of measurement error and the proposed method. The results are based on a simulation study with 500 replications for 1000 cases and 1000 controls, where disease status (D) is binary, environmental variables (X, W) are continuous, genetic variant is in the form of diplotype. Environmental variable is measured with error and error variance is assumed to be 0.25. Furthermore, the simulation is used to assess the effect of missing genetic data.

			Naive Approach		Proposed Method	
	Parameter	True Value	Bias	RMSE	Bias	RMSE
Complete Data	β_0	-5.000	1.207	1.459	0.230	0.086
	β_g	0.693	0.080	0.011	-0.001	0.007
	β_x	1.099	-0.797	0.645	0.001	0.137
	β_{xg}	0.693	-0.478	0.235	0.006	0.088
	pr$(G = 1)$	0.250	0.005	0.000	0.000	0.000
	pr$(D = 1)$	0.046	-0.032	0.001	0.008	0.000
	η_1	0.000			0.003	0.001
	η_2	0.100			-0.001	0.000
50% of genetic information is missing	β_0	-5.000	1.206	1.460	0.228	0.084
	β_g	0.693	0.082	0.015	-0.002	0.007
	β_x	1.099	-0.794	0.647	0.013	0.161
	β_{xg}	0.693	-0.477	0.243	0.011	0.102
	pr$(G = 1)$	0.250	0.004	0.000	0.000	0.000
	pr$(D = 1)$	0.046	-0.032	0.001	0.008	0.000
	η_1	0.000			0.003	0.001
	η_2	0.100			-0.002	0.000

5.3 Inference in the Binary Case

We estimated parameters using the foregoing algorithm and performed inferences based on the Wald, Proposed Likelihood Ratio and Naive Likelihood Ratio procedures for small $(n = 400)$ and moderate $(n = 2000)$ sample sizes. The results are presented in the Table 4 and using histograms in the Appendix G. We found that the proposed method closely achieves the nominal coverage, while Wald test resulted in rather elevated error rates and the effect is more apparent for small sample size. The distribution of parameter estimates is skewed in the presence of measurement error and small sample size. Our simulations showed that in this setting it is approximately correct to use the standard asymtotics for the Likelihood Ratio procedure.

Table 4: Coverage probabilities of the 95% Wald and LR confidence intervals for interaction parameters. The results are based on simulation studies with 1000 relications of 200 cases and 200 controls $(n = 400)$; and 1000 relications of 1000 cases and 1000 controls $(n = 2000)$. Disease status (D), genetic (H) and environmental (X) factors are binary with pr$(D = 1) = 0.0163,$ pr$(G = 1) = 0.1,$ pr$(X = 1) = 0.5$.

	n = 400	n = 2000
True value of β_{xg}	0.693	0.693
Mean of $\widehat{\beta}_{xg}$ over all simulated datasets	0.848	0.692
Median of $\widehat{\beta}_{xg}$ over all simulated datasets	0.695	0.669
Variance of $\widehat{\beta}_{xg}$ over all simulated datasets	0.707	0.105
5% trimmed mean estimate of variance of $\widehat{\beta}_{xg}$	1.005	0.091
Coverage of the Wald CI	0.937	0.931
Coverage of the Likelihood Ratio CI	0.954	0.949

5.4 Inference in the Continuous Case

Here we performed inferences in the Continuous Case assuming probability of disease is known. Results are presented in the Table 5. The sampling distribution of the parameter estimates is slightly skewed, as it is illustrated using histograms in the Appendix H, and skewness is more pronounced for small sample sizes. Hence it is undesirable to use Wald-type confidence intervals, since they are based on asymptotic normality. The Likelihood Ratio Type Test inferences performed substantially better and our simulations showed that in this setting it is approximately correct to use the standard asymtotics for the Likelihood Ratio procedure.

Table 5: Coverage probabilities of the 95% Wald and adjusted retrospective LR confidence intervals for interaction parameters with different amounts of measurement error. The results are based on simulation study with 1000 cases and 1000 controls $(n = 2000)$, where disease status (D) is binary, environmental variables (X, W) are continuous and the genetic variant h_3 is in the form of diplotype. The environmental variable is measured with error and the error variance is set to be ξ.

Measurement Error Variance ξ	0.25
True value of β_{xg}	0.693
Mean of $\widehat{\beta}_{xg}$	0.678
Median of $\widehat{\beta}_{xg}$	0.700
Robust variance estimate of $\widehat{\beta}_{xg}$	0.102
5% trimmed mean estimate of variance of $\widehat{\beta}_{xg}$	0.048
Coverage of the Wald test	0.697
Coverage of the Likelihood Ratio test	0.941

CHAPTER VI

CALCIUM DATA ANALYSIS

6.1 Introduction

Here we analyse a case-control study of colorectal adenoma (Peters et al., 2004) designed to investigate the interactions of dietary calcium intake and genetic variants in the calcium-sensing receptor (CASR) region. In this study, a total of 772 cases and 778 controls were sampled from the screening arm of the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial. Information on dietary food intake on the participants were available from a baseline food frequency questionnaire (FFQ). Genotype data were available on three non-synonymous single neucleotide polymorphisms (SNP) in the CASR region. One of the major goals of the study was to investigate the interaction of dietary calcium and the CASR gene based on "haplotypes", that is the combinations of alleles at three different CASR loci along individual chromosomes. Two technical problems arose. First, as typical, we only had locus-specific genotype data which provides information on two alleles a subject carries on the pair of homologous chromosome, at each locus separately. Such genotype data lacks the phase information that is which combinations of allele arise together on the individual chromosomes giving rise to an interesting missing data problem. Second, it is well known that FFQ as an instrument for measuring dietary intake is prone to both bias and random error. We will use data from an external study (Potischman, et al., 2002) to form estimates of the bias and variance of the measurement error. The availability of such external data gave rise to the opportunity for studying calcium-CASR interaction after correcting for measurement error due to use of FFQ.

6.2 Modeling

Here we analyze the colorectal adenoma study data described in the introduction. The genetic data observed were three SNPs in the calcium receptor gene CaSR, the environmental variable X measured with error was log(1+calcium intake), which was measured by W, the result of a food frequency questionnaire. The variables Z measured without error were age, sex and race. The possible haplotypes in the data were ACG, ACT, AGG, GCG, AGT, GGG, and GCT. Since haplotypes AGT, GGG, GCT are rare, we pooled them with the next common haplotype AGT. A few subjects do not have measurements of calcium intake and we eliminated them from the analysis.

Given calcium intake (X) and diplotype information ($H^{\rm dip}$) we considered the following risk model

$$\begin{aligned}\text{logit}\{\text{pr}(D=1|H^{\rm dip},X)\} &= \beta_0+\beta_x*X+\beta_{h2}*N_2(H^{\rm dip})+\beta_{h4}*N_4(H^{\rm dip})\\ &+ \beta_{h5}*N_5(H^{\rm dip})+\beta_{xh2}*X*N_2(H^{\rm dip})\\ &+ \beta_{xh4}*X*N_4(H^{\rm dip})+\beta_{xh5}*X*N_5(H^{\rm dip}),\end{aligned}$$

where $N_2(H^{\rm dip})$ is the number of haplotypes ACT observed in a diplotype, $N_4(H^{\rm dip})$ is the number of haplotypes GCG observed in a diplotype and $N_5(H^{\rm dip})$ is number of haplotypes AGG, AGT, GGG, or GCT observed in a diplotype.

Unfortunately, there is no direct information in the study to assess the measurement error properties of calcium intake W. We used a combination of outside data and sensitivity analysis instead. The outside data come from the WISH Study (Potischman, et al., 2002). There were $\approx$ 400 women in this study, which used the same FFQ as in the colorectal adenoma study and also include the results of 6 24-hour recall measurements, which we

denote by T_{ij} for the i^{th} individual and j^{th} replicate. The model for these data are that

$$\begin{aligned} W_i &= \alpha_0 + \alpha_1 X_i + U_i; \\ T_{ij} &= X_i + V_{ij}, \end{aligned}$$

where $U_i = \text{Normal}(0, \sigma_u^2)$ and $V_{ij} = \text{Normal}(0, \sigma_v^2)$. Using variance components analysis, we estimated $(\alpha_0, \alpha_1, \sigma_u^2)$, and took these as fixed and known in the colorectal adenoma study, although we also varied σ_u^2. The distribution of X was taken to be Gaussian with mean linear in Z and variance ξ. We used the method of Fuller (1987, Chapter 2,5) and found estimates $\widehat{\alpha}_0 = 0.22$, $\widehat{\alpha}_1 = 0.75$, $\widehat{\sigma}_u^2 = \widehat{\xi} = 0.65$. To asses sensitivity to the measurement error model specification we considered several scenarios by imposing measurement error structure estimated using WISH data and varying it.

6.3 Estimation

Four sets of parameter estimates presented in the Table 6 correspond to different values of the measurement error variance.

The probability of disease was set to be on the interval $(0.001, 0.5)$, but the likelihood function was flat either as a function of the probability of disease, or, equivalently, as a function of the intercept parameter β_0. However, estimates of the risk parameters are unchanged for different values of probability of disease.

Results presented in the Table 7 illustrate the importance of assessing the measurement error process, as its incorrect specification results in substantial biases.

Based on estimates of the main effects, we first observed that subjects with the diplotype (h_1, h_1), a unit increase in calcium intake is associated with decreased risk of the colorectal adenoma development with odds ratio $\exp(-0.1507) = 0.8601$, assuming that $\xi = 0.65$. Inspection of the interaction parameter estimates suggests that among carriers of h_4 and h_5 haplotypes, increased calcium intake is associated with an even greater decrease

in risk of colorectal tumor development, especially for larger error variance.

Table 6: Estimates of risk parameters for the colorectal adenoma study assuming different variances (ξ) of the measurement error.

Parameter	Naive	$\xi = 0.10$	$\xi = 0.60$	$\xi = 0.65$	$\xi = 0.70$
β_{h_2}	-0.2087	-0.1866	-0.1606	-0.1770	-0.1365
β_{h_4}	-0.1663	-0.1908	-0.3710	-0.4289	-0.5377
β_{h_5}	-0.2770	-0.3670	-0.6609	-0.7584	-0.9379
β_x	-0.0852	-0.0683	-0.1402	-0.1507	-0.1850
β_{xh2}	0.0398	0.0394	0.1296	0.1044	0.2224
β_{xh4}	-0.1886	-0.1749	-0.5192	-0.5817	-0.8124
β_{xh5}	-0.2804	-0.2361	-0.7136	-0.8885	-1.1234

Table 7: Standard errors of risk parameter estimates for the colorectal adenoma study assuming different variances (ξ) of the measurement error.

Parameter	Naive	$\xi = 0.10$	$\xi = 0.60$	$\xi = 0.65$	$\xi = 0.70$
β_{h_2}	0.1132	0.1058	0.1175	0.1188	0.1201
β_{h_4}	0.1451	0.1304	0.1511	0.1532	0.1554
β_{h_5}	0.1815	0.1348	0.1686	0.1719	0.1752
β_x	0.0683	0.0679	0.1580	0.1672	0.1764
β_{xh2}	0.0851	0.0838	0.1890	0.1997	0.2105
β_{xh4}	0.0907	0.0895	0.1924	0.2027	0.2130
β_{xh5}	0.1203	0.1004	0.2028	0.2132	0.2236

6.4 Inference

We performed inference based on the Naive Likelihood Ratio, Proposed Likelihood Ratio, and Wald testing procedures. We found that Wald confidence intervals are generally wider than Likelihood Ratio confidence intervals, and hence Wald test announced some of the parameters to be not significantly different from 0, while the Likelihood Ratio procedure proved they are significant. For majority of cases λ was very close to 1 and there was no noticable difference between Naive and Proposed Likelihood Ratio intervals.

We found that at 0.05 significance level the data does not have enough evidence to indicate that β_{xh2} is significantly different from 0, as illustrated on the Figure 1. Hence we considered reduced model by setting β_{xh2} to be 0. Analysis of the reduced model showed that β_{xh5} is significantly different from 0 for all measurement error model specifications we considered. Wald test announced β_{xh4} as significant for measurement error variance 0.5 and greater, while the Likelihood Ratio test proved it is significantly different from 0 for measurement error variance of 0.4 and larger, what can be seen on the Figure 2.

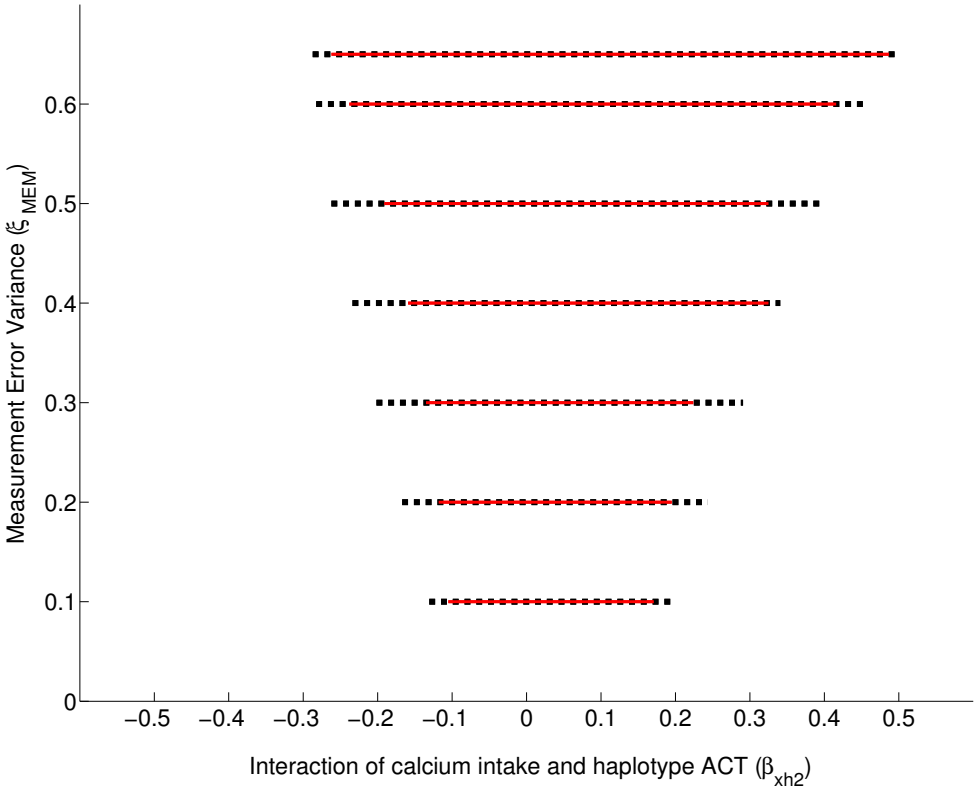

Figure 1: Wald (dashed black line) and likelihood ratio (red line) confidence intervals for β_{xh2} in the full model for different values of measurement error variance ξ_{MEM}.

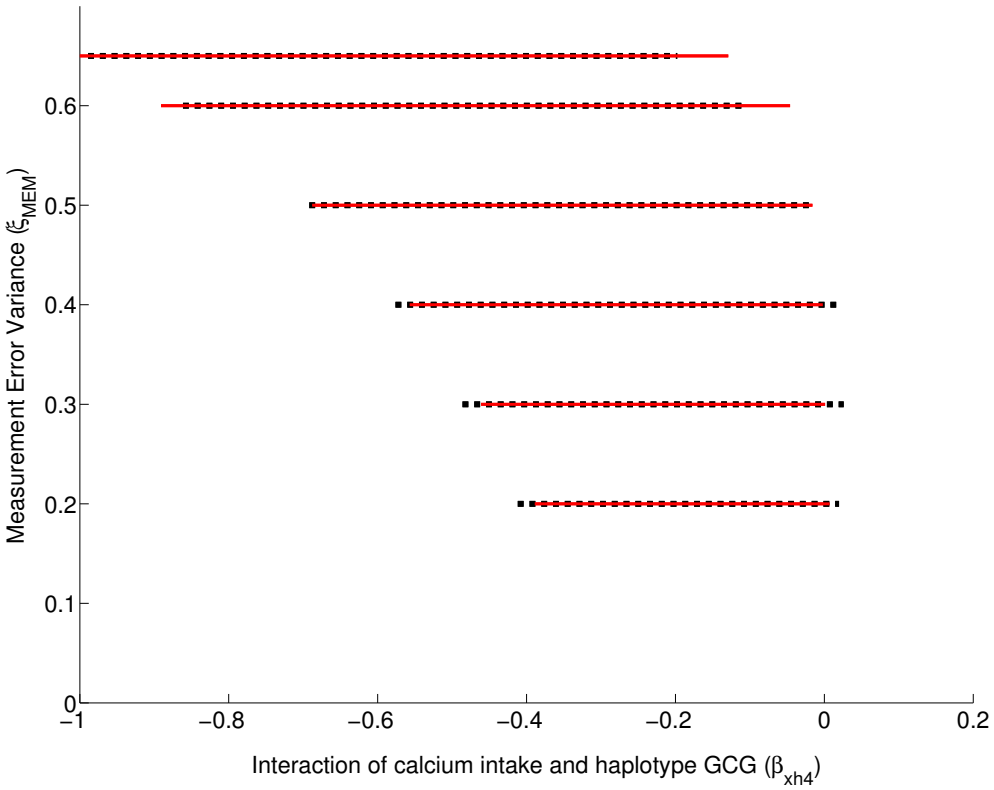

Figure 2: Wald (dashed black line) and likelihood ratio (red line) confidence intervals for β_{xh4} in the reduced model with $\beta_{xh2} = 0$ for different values of measurement error variance ξ_{MEM}.

CHAPTER VII

SUMMARY AND FUTURE RESEARCH

7.1 Summary

We have considered the problem of relating risk of a complex disease to genetic susceptibilities, environmental exposures, and their interaction when the environmental covariates are measured with error and some of the genetic information is missing. Utilizing a polychotomous logistic regression model, profile likelihood and a model for the distribution of underlying gene information, we constructed a relatively simple yet efficient semiparametric algorithm for parameter estimation. We have shown that the resulting estimates are consistent and derived their asymptotic variance when the distribution of measurement error is known, and when it is estimated from replications.

Our simulation results illustrate that for large studies there is no noticeable bias in our parameter estimates, whereas the naive approach that ignores the existence of the measurement error results in substantial bias.

We developed a LR-type procedure investigating significance of intraction parameters, as well as main effects. In our setting it is undesirable to use Wald-type procedure since it proved to behave aberrantly in the binomial logit model and it can suffer in presence of measurement error. The LR-type procedure we developed proved to be a successful alternative. Particularly, in small-sample setting Wald test resulted in rather elevated error rates, while LR-type procedure closely achieved the nominal coverage.

The methodology was applied to the analysis of the Calcium Study, the main goal of which was to investigate interaction between dietary calcium intake and CaSR haplotypes.

7.2 Future Research

This work has several interesting extentions. First, to accomodate different types of disease we allowed disease status to take $K + 1$ unordered levels. In many situations, i.e. cancers, the disease stage is an ordered categorical variable. Hence, it could prove useful to model ordered disease status. Second, it is often the case that genotypes are misclassified. Therefore, it would be beneficial to model genotyping errors. Third, sometimes genotypes are missing informatively, that is the probability of missingness depends on what is being measured. It would be interesting to investigate robustness of our methodology to the missing at random assumption and possibly develop a robust version.

REFERENCES

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C.M. (2006). *Measurement Error in Nonlinear Models*, Second Edition, London: Chapman & Hall CRC Press.

Carroll, R. J., Spiegelman, C. H., Lan, K. K. G., Bailey, K. T., and Abbott, R.D. (1984). On errors-in-variables for binary regression models. *Biometrika*, **71**, 19-25.

Chatterjee, N. and Carroll, R. J. (2005). Semiparametric maximum likelihood estimation in case-control studies of gene-environmental interactions. *Biometrika*, **92**, 399-418.

Clayton, D. and McKeigue, P. M. (2001). Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet*, **358**, 1356-60.

Cook, J. and Stefanski, L. A. (1995). A simulation extrapolation method for parametric measurement error models. *Journal of the American Statistical Association*, **89**, 1314–1328.

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

Dosemeci, M., Wacholder, S., and Lubin, J. H. (1990). Does nondifferential misclassification of exposure always bias a true effect toward the null value? *American Journal of Epidemiology*, **132**, 746-748.

Epstein, M. and Sattern, G. (2003). Interence of haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics*,**73**, 1316-1329.

Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in diploid population. *Molecular Biology Evolution*, **12**, 921-927.

Fallin, D. and Schork, N. (2000). Accuracy of haplotype frequency estimation for bial-

lelic loci, via the expectation - maximization algorithm for unphased diplotype data. *American Journal of Human Genetics*, **67**, 947-959.

Freudenheim, J. L. and Marshall, J. R. (1988). The problem of profound mismeasurement and the power of epidemiologic studies of diet and cancer. *Nutrition and Cancer*, **11**, 243-250.

Fuchs, C. S., Giovannucci, E. L. and Colditz, G.A. (1999). Dietary fiber and the risk of colorectal cancer and adenoma in women. *New England Journal of Medicine*, **340**, 169–176.

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

Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology*. Chapman and Hall.

Gustafson, P., Le, N.D., and Vallee, M. (2002). A bayesian approach to case-control studies with errors in covariates. *Biostatistics*, **3**, 229-243.

Hartley, H. (1958). Maximum likelihood estimation from incomplete data. *Biometrics*, **14**, 174-194.

Hunter, D.J., Spiegelman, D., Adami, H.O., Beeson, L., van den Brandt, P.A., et al. (1996). Cohort studies of fat intake and the risk of breast cancer - a pooled analysis. *New England Journal of Medicine*, **334**, 356-361.

Kent, J. T. (1982). Robust properties of likelihood ratio test. *Biometrika*, **69**, 19-27.

Michels, K.B., Giovannucci, F., Joshipura, K.J., Rosner, B.A., Stampfer, M.J., Fuchs, C.S., Golditz, G.A., Speizer, F.E., and Willett, W.C. (2002). Prospective study of fruit and vegetable consumption and incidence of colon and rectal cancers. *Journal of the National Cancer Institute*, **92**, 1740-1752.

Muller, P. and Roeder, K. (1997). A bayesian semiparametric model for case-control studies with errors in variables. *Biometrika*, **84**, 523-537.

Peters, U., Chatterjee, N., Yeager, M., Chanock, S.J., Schoen, R.E., McGlynn, K.A., Church, T.R., Weissfeld, J.L., Schatzkin, A. and Hayes, R.B. (2004). Association of genetic variants in the calcium-sensing receptor with risk of colorectal adenoma. *Cancer Epidemiol Biomarkers Prev*, **13**(12): 2181-2186.

Potischman, N., Coates, R. J., Swanson, C. A., Carroll, R. J., Daling, J. R., Brogan, D. R., Gammon, M. D., Midthune, D., Curtin, J. and Brinton, L. A. (2002). Increased risk of early stage breast cancer related to consumption of sweet foods among women less than age 45. *Cancer Causes and Control*, **13**, 937-46.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, **66**, 403-412.

Roeder, K., Carroll, R. J. and Lindsay B. G. (1996). A semiparametric mixture approach to case-control studies with errors in covariates *Journal of the American Statistical Association*, **91**, 722-732.

Roy, K. P. (1957). A note on asymptotic distribution of likelihood ratio. *Calcutta Statistical Association Bulletin*, **1**, 60-62.

Satten, G. A. and Epstein M.P. (2004). Comparison of prospective and retrospective methods for haplotype inference in case-control studies. *Genetic Epidemiology*, **27**, 192-2001.

Seaman, S. R. and Richardson, S. (2001). Bayesian analysis of case-control studies with categorical covariates *Biometrika*, **88**, 1073-1088.

Spinka, C., Carroll, R. J. and Chatterjee, N. (2005). Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-

phase ambiguity. *Genetic Epidemiology*, **29**, 108-127.

Starm, D., Haiman, C., Hirschhorn, J., Altshuler, D., Kolonel, L., Henderson, B. and Pike, M. (2003). Choosing haplotype-tagging SNPs based on uphased genotype data using as preliminary sample of unrelated subjects with an example from the multiethnic cohort study. *Human Heredity*, **55**, 27-36.

Stefanski, L.A., Novick, S.J. and Devanarayan, V. (2005). Estimating a nonlinear function of a normal mean, *Biometrika*, **92**, 732-736.

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

Subar, A.F., Kipnis, V., Troiano, R.P., Midthune, D., Bingham, S., et al. (2003). Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: The Observing Protein and Energy Nutrition (OPEN) study. *American Journal of Epidemiology*, **158**, 1-13.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, **54**, 426-485.

Wilks, S. S.(1938). The large-sample distribution of the likelihood ratio for testing composite hypothesis. *Annals of Mathematical Statistics*, **7**, 73-77.

Zhang, D. and Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, **57**, 795-802.

Zhao, L., Li, S., and Khalid, N. (2003). A method for the assesment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *American Journal of Human Genetics*, **72**, 1231-1250.

APPENDIX A

PROOF OF (2.2)

The proof of (2.3) is straightforward. Note that

$$
\begin{aligned}
&\text{pr}(D = d, H^{\text{dip}} = h^{\text{dip}}, X = x, W = w | R = 1, Z = z) \\
&\quad \propto \text{pr}(D = d, H^{\text{dip}} = h^{\text{dip}}, X = x, W = w, R = 1 | Z = z) \\
&\quad \propto \frac{n_d}{\pi_d}[1 + \sum_{j=1}^{m} \exp\{\beta_{0j} + m(h^{\text{dip}}, x, z, \beta)\}]^{-1} \\
&\qquad \exp[I_{(d \geq 1)}(d)\{\beta_{0d} + m(h^{\text{dip}}, x, z, \beta)\}] \\
&\qquad\quad \times Q(h^{\text{dip}}, \Theta) f_{\text{mem}}(w | d, h^{\text{dip}}, x, z, \xi) f_X(x | z, \eta) \\
&\quad \propto \frac{n_0}{\pi_0} S(d, h^{\text{dip}}, x, z, \Omega) f_{\text{mem}}(w | d, h^{\text{dip}}, x, z, \xi) f_X(x | z, \eta) \\
&\quad = \frac{S(d, h^{\text{dip}}, x, z, \Omega) f_{\text{mem}}(w | d, h^{\text{dip}}, x, z, \xi) f_X(x | z, \eta)}{\sum_{d_*} \sum_{h_*^{\text{dip}}} \int S(d_*, h_*^{\text{dip}}, x, z, \Omega) f_{\text{mem}}(w | d_*, h_*^{\text{dip}}, x, z, \xi) f_X(x | z, \eta) dw dx} \\
&\quad = \frac{S(d, h^{\text{dip}}, x, z, \Omega) f_{\text{mem}}(w | d, h^{\text{dip}}, x, z, \xi) f_X(x | z, \eta)}{\sum_{d_*} \sum_{h_*^{\text{dip}}} \int S(d_*, h_*^{\text{dip}}, x, z, \Omega) f_X(x | z, \eta) dx}.
\end{aligned}
$$

Equation (2.3) now follows by appropriate summation over $h^{\text{dip}} \in \mathcal{H}_G^{\text{dip}}$ and integration over x.

APPENDIX B

PROOF OF THEOREM 1

The proof consists of two steps. The first shows that the estimating equation has mean zero when evaluated at the true parameters. We then show that the estimating function evaluated at the true parameters has a covariance matrix of the form $\mathcal{I} - \Lambda$.

We first consider the derivative with respect to Ω. Denote the first partial derivative of $S(d, h^{\rm dip}, x, z, \Omega)$ with respect to Ω by $S_\Omega(d, h^{\rm dip}, x, z, \Omega)$. The semiparametric profile likelihood score for $\mathcal{B}(\Omega^{\rm T}, \eta^{\rm T})^{\rm T}$ is the derivative of the logarithm of (2.3) with respect to Ω and is given as

$$n^{-1} \sum_{i=1}^{n} \left\{ C_1(D_i, Z_i, W_i, G) - C_2(Z_i) \right\},$$

where $C_1(\bullet) = \{A_1^{\rm T}(\bullet), B_1^{\rm T}(\bullet)\}$, $C_2(\bullet) = \{A_2^{\rm T}(\bullet), B_2^{\rm T}(\bullet)\}$,

$$A_1(d, z, w, {\rm g}) = \frac{\sum_{h^{\rm dip} \in \mathcal{H}_G^{\rm dip}} \int S_\Omega(d, h^{\rm dip}, x, z, \Omega) f_{\rm mem}(w|d, h^{\rm dip}, x, z, \xi) f_X(x|z, \eta) dx}{\sum_{h^{\rm dip} \in \mathcal{H}_G^{\rm dip}} \int S(d, h^{\rm dip}, x, z, \Omega) f_{\rm mem}(w|d, h^{\rm dip}, x, z, \xi) f_X(x|z, \eta) dx},$$

$$A_2(z) = \frac{\int \sum_{d_*} \sum_{h_*^{\rm dip}} S_\Omega(d_*, h_*^{\rm dip}, x, z, \Omega) f_X(x|z, \eta) dx}{\int \sum_{d_*} \sum_{h_*^{\rm dip}} S(d_*, h_*^{\rm dip}, x, z, \Omega) f_X(x|z, \eta) dx},$$

and where $B_1(\bullet)$ and $B_2(\bullet)$ are defined by replacing $S_\Omega(\bullet)$ by $S(\bullet) s_X(x|z, \eta)$, where $s_X(x|z, \eta) = \partial \log\{f_X(x|z, \eta)\}/\partial \eta$.

It is useful to note that the density of Z and (W, G, Z) given $D = d$ can be written as

$$[Z|D = d] = f_Z(z) \frac{n_0}{\pi_0 n_d} \int \sum_{h_*^{\rm dip}} S(d, h_*^{\rm dip}, x, z, \Omega) f_X(x|z, \eta) dx; \tag{B.1}$$

$$[W, G, Z|D = d] = f_Z(z) \frac{n_0}{\pi_0 n_d} \int \sum_{h^{\rm dip} \in \mathcal{H}_G^{\rm dip}} S(d, h^{\rm dip}, x, z, \Omega) f_{\rm mem}(w|d, h^{\rm dip}, x, z, \xi) \times f_X(x|z, \eta) dx. \tag{B.2}$$

Then it follows from (B.1) that

$$\begin{aligned}\mathrm{E}\left\{A_2(Z)\right\} &= \sum_{d_*} \frac{n_{d_*}}{n} \mathrm{E}\left\{A_2(Z)|D=d_*\right\} \\ &= \int \sum_{d_*} \frac{n_0}{n\pi_0} \sum_{h_*^{\mathrm{dip}}} S(d_*, h_*^{\mathrm{dip}}, x, z, \Omega) f_X(x|z,\eta) f_Z(z) A_2(z) dx\, dz \\ &= \int \sum_{d_*} \frac{n_0}{n\pi_0} \sum_{h_*^{\mathrm{dip}}} S_\Omega(d, h_*^{\mathrm{dip}}, x, z, \Omega) f_X(x|z,\eta) f_Z(z) dx\, dz.\end{aligned}$$

It is also straightforward using (B.2) to show that

$$\begin{aligned}\mathrm{E}\left\{A_1(D,Z,W,G)\right\} &= \sum_{d_*} \frac{n_{d_*}}{n} \mathrm{E}\left\{A_1(D,Z,W,G|D=d_*)\right\} \\ &= \frac{n_0}{n\pi_0} \int \sum_{d_*} \sum_{h_*^{\mathrm{dip}}} S_\Omega(d_*, h_*^{\mathrm{dip}}, x, z, \Omega) f_X(x|z,\eta) f_Z(z) dx\, dz,\end{aligned}$$

thus showing that the top part of (2.3) has mean zero. That the bottom part also has mean zero is shown similarly.

Much the same argument holds for the estimating function for ξ. Define $\mathcal{C}(w|d, h^{\mathrm{dip}}, x, z, \xi)$ to be the derivative of $\log\{f_{\mathrm{mem}}(w|d, h^{\mathrm{dip}}, x, z, \xi)\}$ with respect to ξ, and define

$$\begin{aligned}&A_\xi(d, w, z, \mathrm{g}) \\ &\quad = \frac{\sum_{h^{\mathrm{dip}} \in \mathcal{H}_G^{\mathrm{dip}}} \int S(d, h^{\mathrm{dip}}, x, z, \Omega) \mathcal{C}(w|d, h^{\mathrm{dip}}, x, z, \xi)}{\sum_{h^{\mathrm{dip}} \in \mathcal{H}_G^{\mathrm{dip}}} \int S(d, h^{\mathrm{dip}}, x, z, \Omega) f_{\mathrm{mem}}(w|d, h^{\mathrm{dip}}, x, z, \xi) f_X(x|z,\eta) dx} \\ &\qquad \times f_{\mathrm{mem}}(w|d, h^{\mathrm{dip}}, x, z, \xi) f_X(x|z,\eta) dx.\end{aligned}$$

Then it is easy to show that

$$\begin{aligned}&\mathrm{E}\left\{A_\xi(D,W,Z,G)\right\} = \sum_{d_*} \frac{n_{d_*}}{n} \mathrm{E}\left\{A_\xi(D,Z,W,G|D=d_*)\right\} \\ &\quad = \frac{n_0}{n\pi_0} \int \sum_{d_*} \sum_{h_*^{\mathrm{dip}}} S(d_*, h_*^{\mathrm{dip}}, x, z, \Omega) f_X(x|z,\eta) f_Z(z) \\ &\qquad \times \Big\{ \int f_{\mathrm{mem}}(w|d_*, h_*^{\mathrm{dip}}, x, z, \xi) \mathcal{C}(w|d_*, h_*^{\mathrm{dip}}, x, z, \xi) dw \Big\} dx\, dz = 0,\end{aligned}$$

the interior integral being equal to zero by standard likelihood results.

As described above, the estimating equation is given as (2.3). Define

$$C_3(d) = \{A_3^{\mathrm{T}}(d), B_3^{\mathrm{T}}(d)\}^{\mathrm{T}} = E[\{C_1(D, Z, W, G) - C_2(Z)\}|D = d].$$

Then, when evaluated at the true parameters, the estimating function takes the form

$$n^{-1/2} \sum_{i=1}^{n} \{C_1(D_i, Z_i, W_i, G) - C_2(Z_i) - C_3(D_i)\},$$

which is a sum of independent, mean zero random variables. It follows directly that, when evaluated at the true parameters, the estimating function has covariance matrix

$$\Sigma_* = n^{-1} \sum_{i=1}^{n} E\left[\{C_1(D_i, Z_i, W_i, G) - C_2(Z_i)\}\{C_1(D_i, Z_i, W_i, G) - C_2(Z_i)\}^{\mathrm{T}}\right] - \Lambda. \tag{B.3}$$

Make the definitions

$$\begin{aligned}
\mathcal{Q}_1(d, g, w, z, \mathcal{B}, \xi) &= \int \sum_{h^{\mathrm{dip}} \in \mathcal{H}_G^{\mathrm{dip}}} \left\{S_\Omega^{\mathrm{T}}(d, h^{\mathrm{dip}}, x, z, \Omega), S(d, h^{\mathrm{dip}}, x, z, \Omega) s_X^{\mathrm{T}}(x|z, \eta)\right\}^{\mathrm{T}} \\
&\qquad \times f_{\mathrm{mem}}(w|d, h^{\mathrm{dip}}, x, z, \xi) f_X(x|z, \eta) dx; \\
\mathcal{Q}_2(d, g, w, z, \mathcal{B}, \xi) &= \int \sum_{h^{\mathrm{dip}} \in \mathcal{H}_G^{\mathrm{dip}}} S(d, h^{\mathrm{dip}}, x, z, \Omega) f_{\mathrm{mem}}(w|d, h^{\mathrm{dip}}, x, z, \xi) f_X(x|z, \eta) dx; \\
\mathcal{Q}_3(z, \mathcal{B}, \xi) &= \int \sum_{d_*} \sum_{h_*^{\mathrm{dip}}} \left\{S_\Omega^{\mathrm{T}}(d_*, h_*^{\mathrm{dip}}, x, z, \Omega), S(d_*, h_*^{\mathrm{dip}}, x, z, \Omega) s_X^{\mathrm{T}}(x|z, \eta)\right\}^{\mathrm{T}} \\
&\qquad \times f_X(x|z, \eta) dx; \\
\mathcal{Q}_4(z, \mathcal{B}, \xi) &= \int \sum_{d_*} \sum_{h_*^{\mathrm{dip}}} S(d_*, h_*^{\mathrm{dip}}, x, z, \Omega) f_X(x|z, \eta) dx.
\end{aligned}$$

Then it is easy to show that (B.3) can be rewritten as

$$\begin{aligned}
\Sigma_* &= \mathcal{A}_1 - \mathcal{A}_2 - \Lambda; \\
\mathcal{A}_1 &= \frac{n_0}{n\pi_0} \int \sum_{d_*} \sum_{g_*} \frac{\mathcal{Q}_1(d_*, g_*, w, z, \mathcal{B}, \xi) \mathcal{Q}_1^{\mathrm{T}}(d_*, g_*, w, z, \mathcal{B}, \xi)}{\mathcal{Q}_2(d_*, g_*, w, z, \mathcal{B}, \xi)} dw f_Z(z) dz; \\
\mathcal{A}_2 &= \frac{n_0}{n\pi_0} \int \frac{\mathcal{Q}_3(z, \mathcal{B}, \xi) \mathcal{Q}_3^{\mathrm{T}}(z, \mathcal{B}, \xi)}{\mathcal{Q}_4(z, \mathcal{B}, \xi)} f_Z(z) dz.
\end{aligned}$$

We claim that $\mathcal{I} = \mathcal{A}_1 - \mathcal{A}_2$. By a direct calculation, $\mathcal{I} = \mathcal{I}_1 - \mathcal{I}_2$, where using (B.1),

$$\begin{aligned}\mathcal{I}_2 &= -\sum_{d_*} \frac{n_{d_*}}{n} E\left[\frac{\partial}{\partial \mathcal{B}^{\mathrm{T}}}\left\{\frac{\mathcal{Q}_3(Z,\mathcal{B},\xi)}{\mathcal{Q}_4(Z,\mathcal{B},\xi)}\middle| D = d\right\}\right] \\ &= -\frac{n_0}{n\pi_0}\frac{\partial^2}{\partial \mathcal{B}\partial \mathcal{B}^{\mathrm{T}}}\int \sum_{d_*}\sum_{h_*^{\mathrm{dip}}} S(d_*, h_*^{\mathrm{dip}}, x, z, \Omega) f_X(x|z,\eta) f_Z(z) dx dz + \mathcal{A}_2.\end{aligned}$$

In addition, using (B.2), we find that

$$\begin{aligned}\mathcal{I}_1 &= -\sum_{d_*} \frac{n_{d_*}}{n} E\left[\frac{\partial}{\partial \mathcal{B}^{\mathrm{T}}}\left\{\frac{\mathcal{Q}_1(d_*,G,W,Z,\mathcal{B},\xi)}{\mathcal{Q}_2(d_*,G,W,Z,\mathcal{B},\xi)}\middle| D = d\right\}\right] \\ &= -\frac{n_0}{n\pi_0}\frac{\partial^2}{\partial \mathcal{B}\partial \mathcal{B}^{\mathrm{T}}}\int \sum_{d_*}\sum_{h_*^{\mathrm{dip}}} S(d_*, h_*^{\mathrm{dip}}, x, z, \Omega) f_X(x|z,\eta) f_Z(z) dx dz + \mathcal{A}_1,\end{aligned}$$

completing the proof.

APPENDIX C

EM CALCULATIONS

In what follows, we will need the following identities:

$$\text{pr}(H^{\text{dip}} = h^{\text{dip}}|Z = z, R = 1) = \frac{Q(h^{\text{dip}}, \Theta)\gamma(h^{\text{dip}}, z, \mathcal{B})}{\sum_{h_*^{\text{dip}}} Q(h_*^{\text{dip}}, \Theta)\gamma(h_*^{\text{dip}}, z, \mathcal{B})}; \tag{C.1}$$

$$\begin{aligned}\text{pr}(H^{\text{dip}} = h^{\text{dip}}|G, D = d, W = w, Z = z, R = 1) \\ = \frac{Q(h^{\text{dip}}, \Theta)\alpha(h^{\text{dip}}, d, z, w, \mathcal{B}, \xi)}{\sum_{h^{\text{dip}} \in \mathcal{H}_G^{\text{dip}}} Q(h^{\text{dip}}, \Theta)\alpha(h^{\text{dip}}, d, z, w, \mathcal{B}, \xi)};\end{aligned} \tag{C.2}$$

$$\begin{aligned}\text{pr}(X = x, H^{\text{dip}} = h^{\text{dip}}|D, G, W, Z, R = 1) \\ = \frac{S(D, h^{\text{dip}}, x, Z, \Omega) f_{\text{mem}}(W|D, h^{\text{dip}}, x, Z, \xi) f_X(x|Z, \eta)}{\int \sum_{d_*} \sum_{h^{\text{dip}} \in \mathcal{H}_G^{\text{dip}}} S(d_*, h^{\text{dip}}, x, z, \Omega) f_{\text{mem}}(w|d_*, h^{\text{dip}}, x, z, \xi) f_X(x|z, \eta) dx};\end{aligned} \tag{C.3}$$

$$\begin{aligned}\text{pr}(D = d, H^{\text{dip}} = h^{\text{dip}}, X = x|Z, R = 1) \\ = \frac{S(d, h^{\text{dip}}, x, Z, \Omega) f_X(x|Z, \eta)}{\int \sum_{d_*} \sum_{h_*^{\text{dip}}} S(d_*, h_*^{\text{dip}}, x, Z, \Omega) f_X(x|Z, \eta) dx}.\end{aligned} \tag{C.4}$$

Argument for (3.5)

As in Spinka, et al., the estimating equation for θ_k is

$$\begin{aligned}0 \;=\; & \sum_{i=1}^{n} \text{E}_{(\Omega,\eta)} \left[\frac{\partial \log\{Q(H^{\text{dip}}, \theta)\}}{\partial \theta_k} | G, D_i, W_i, Z_i, R_i = 1 \right] \\ & - \sum_{i=1}^{n} \text{E}_{\mathcal{B}} \left[\frac{\partial \log\{Q(H^{\text{dip}}, \theta)\}}{\partial \theta_k} | Z_i, R_i = 1 \right] + \lambda.\end{aligned}$$

Note that

$$\begin{aligned}\frac{\partial \log[\text{pr}\{H^{\text{dip}} = (h_i, h_j)|\theta\}]}{\partial \theta_k} \;&=\; 2/\theta_k, \text{ if } h_i = h_j = h_k; \\ &=\; 1/\theta_k, \text{ if } h_i = h_k \text{ and } h_j \neq h_k, \text{ or } h_j = h_k \text{ and } h_i \neq h_k; \\ &=\; 0, \text{ if } h_i \neq h_k \text{ and } h_j \neq h_k.\end{aligned}$$

and

$$\mathrm{E}\left[\frac{\partial \log Q\{H^{\mathrm{dip}}|\theta\}}{\partial\theta_k}\right] = (2/\theta_k)\mathrm{pr}\{H^{\mathrm{dip}} = (h_k, h_k)\} + (1/\theta_k)\sum_{h\neq h_k}\mathrm{pr}\{H^{\mathrm{dip}} = (h_k, h)\} + (1/\theta_k)\sum_{h\neq h_k}\mathrm{pr}\{H^{\mathrm{dip}} = (h, h_k)\} = 2\theta_k + 2(1-\theta_k).$$

Since $\sum_{k=1}^{K_\Theta}\{2\theta_k + 2(1-\theta_k)\} = 2K_\Theta$, therefore $\lambda = 0$. Using (C.1) and (C.2), we arrive at (3.5).

Argument for (3.6)

It is readily seen that the estimating function for κ_j is

$$0 = \sum_{i=1}^{n}\mathrm{E}_{(\Omega,v)}\left[\frac{\partial \log\{T(D, H^{\mathrm{dip}}, X, Z, \Omega)\}}{\partial\kappa_j}|G_i, D_i, W_i, Z_i, R_i = 1\right] - \sum_{i=1}^{n}\mathrm{E}_{\mathcal{B}}\left[\frac{\partial \log\{T(D, H^{\mathrm{dip}}, X, Z, \Omega)\}}{\partial\kappa_j}|Z_i, R_i = 1\right].$$

Since $\partial \log\{T(D, H^{\mathrm{dip}}, X, Z, \Omega)\}/\partial\kappa_j = I_{(D=j)}(D)$, using (C.4), estimation can be performed by iteratively solving (3.6).

Argument for (3.7)

The estimating function for β is

$$0 = \sum_{i=1}^{n}\mathrm{E}_{(\Omega,v)}\left[\frac{\partial \log\{T(D, H^{\mathrm{dip}}, X, Z, \Omega)\}}{\partial\beta}|G_i, D_i, W_i, Z_i, R_i = 1\right] - \sum_{i=1}^{n}\mathrm{E}_{\mathcal{B}}\left[\frac{\partial \log\{T(D, H^{\mathrm{dip}}, X, Z, \Omega)\}}{\partial\beta}|Z_i, R_i = 1\right].$$

Using (C.3) and (C.4), we arrive at (3.7). The arguments for updating the β_{0d} and η are similar.

APPENDIX D

PROOF OF THEOREM 2

The estimating function for $\mathcal{B}$ can be written in the form

$$0 = \sum_{i=1}^{n} \sum_{j=1}^{M} I_{(m_i=j)}(m_i) \mathcal{C}(D_i, Z_i, W_i, G_i, m_i, \mathcal{B}),$$

where

$$\mathcal{C}(D_i, Z_i, W_i, G_i, m_i, \mathcal{B}) = \begin{bmatrix} A_1(D_i, Z_i, W_i, G_i, m_i, \mathcal{B}) - A_2(Z_i, \mathcal{B}) - A_3(D_i, \mathcal{B}) \\ A_4(D_i, Z_i, W_i, G_i, m_i, \mathcal{B}), \end{bmatrix}$$

$A_2(\bullet)$ and $A_3(\bullet)$ are independent of m and are given in the Appendix B,

$$A_1(d, z, w, g, m, \mathcal{B}) = \mathcal{Q}_1(d, g, w, z, \mathcal{B}, \xi)\{\mathcal{Q}_2(d, g, w, z, \mathcal{B}, \xi)\}^{-1},$$

and

$$\begin{aligned} &A_4(d, z, w, g, m, \mathcal{B}) \\ &\quad = \int \sum_{h^{\text{dip}} \in \mathcal{H}_G^{\text{dip}}} \frac{\partial}{\partial \xi^{\text{T}}} \log\{f_{\text{mem}}(w|d, h^{\text{dip}}, x, z, m, \xi)\}\{\mathcal{Q}_2(d, g, w, z, \mathcal{B}, \xi)\}^{-1} \\ &\qquad \times S(d, h^{\text{dip}}, x, z, \Omega) f_{\text{mem}}(w|d, h^{\text{dip}}, x, z, m, \xi) f_X(x|z, \eta) dx, \end{aligned}$$

where $\mathcal{Q}_1(\bullet)$ and $\mathcal{Q}_2(\bullet)$ are defined in the Appendix B. The expectation of the right hand side of (3.3) is

$$\sum_{j=1}^{m} p(j) \, \text{E} \left\{ \sum_{i=1}^{n} \mathcal{C}(D_i, Z_i, W_i, m_i = j, \mathcal{B}) \right\} = 0,$$

since we have shown that the expectation is zero if the same number of replicates are used. Similarly, -(Hessian) of the right hand side of (3.3) is

$$-\sum_{i=1}^{n} \sum_{j=1}^{M} I_{(m_i=j)}(m_i) \frac{\partial}{\partial \mathcal{B}^{\text{T}}} \mathcal{C}(D_i, Z_i, W_i, G_i, m_i, \mathcal{B}),$$

and this has expectation $\sum_{j=1}^{M} p(j)\mathcal{I}_j = \mathcal{I}$. Finally, the covariance matrix of the right hand side of (3.3) is

$$\begin{aligned}
&\mathrm{E}\left\{\sum_{i=1}^{n}\sum_{j=1}^{M} I_{(m_i=j)}(m_i)\mathcal{C}(D_i, Z_i, W_i, G_i, m_i, \mathcal{B})\mathcal{C}^{\mathrm{T}}(D_i, Z_i, W_i, G_i, m_i, \mathcal{B})\right\} \\
&\quad = \sum_{j=1}^{M} p(j)\Sigma_j = \sum_{j=1}^{M} p(j)(\mathcal{I}_j - \Lambda_j) = \mathcal{I} - \sum_{j=1}^{m} p(j)\Lambda_j.
\end{aligned}$$

This then shows (3.4).

APPENDIX E

PROOF OF THEOREM 3

The prof consists of two steps. First we will show that the limiting distribution of the likelihood ratio test statistic is of the form (4.5). We then will show that it is distributed as a weighted sum of χ^2_1 random variables.

Using usual likelihood ratio argument one can easily see that

$$2\left\{\mathcal{L}(\mathcal{B}_0) - \mathcal{L}(\widehat{\mathcal{B}})\right\} = \left(\mathcal{B}_0 - \widehat{\mathcal{B}}\right)^{\mathrm{T}} \mathcal{L}_{\mathcal{B}\mathcal{B}}(\theta_*) \left(\mathcal{B}_0 - \widehat{\mathcal{B}}\right).$$

where $\mathcal{B}_*$ is between $\mathcal{B}_0$ and $\widehat{\mathcal{B}}$.

Now use (4.3) and (4.4), so that

$$\begin{aligned} 2\left\{\mathcal{L}(\mathcal{B}_0) - \mathcal{L}(\widehat{\mathcal{B}})\right\} &= \left\{n^{1/2}(\widehat{\mathcal{B}} - \mathcal{B}_0)\right\}^{\mathrm{T}} \mathcal{I} \left\{n^{1/2}(\widehat{\mathcal{B}} - \mathcal{B}_0)\right\} + o_p(1) \\ &= \mathcal{V}^{\mathrm{T}}\mathcal{I}\mathcal{V} + o_p(1). \end{aligned}$$

Since the covariance matrix $\mathcal{S}^{-1}$ is symmetric and positive definite, using Cholesky decomposition it can be factored as $\mathcal{S}^{-1} = LL^{\mathrm{T}}$ where L is a lower-triangular matrix. Define P to be an orthogonal matrix of eigenvectors of $L\mathcal{I}L^{\mathrm{T}}$ and Λ is a diagonal matrix of eigenvalues of $L\mathcal{I}L^{\mathrm{T}}$. Since $L\mathcal{I}L^{\mathrm{T}}$ is square and symmetric, Singular Value Decomposition can be applied to it in the following manner $P^{\mathrm{T}}L\mathcal{I}L^{\mathrm{T}}P = \Lambda$. Let $\mathcal{V}_1 = L^{-1}\mathcal{V}$ and $\mathcal{V}_2 = P\mathcal{V}_1$. Note that the distribution of $\mathcal{V}^{\mathrm{T}}\mathcal{I}\mathcal{V}$ is the same as the distribution of $\mathcal{V}_2^{\mathrm{T}}\Lambda\mathcal{V}_2$. It can be easily seen that $\mathcal{V}_2$ has limiting Normal$(0, E)$ distribution, where E is an identity matrix. The fact that quadratic form $\mathcal{V}_2^{\mathrm{T}}E\mathcal{V}_2$ is distributed as $\sum_{i=1}^{k} \lambda_i Z_i^2$ completes the prof.

APPENDIX F

PROOF OF THEOREM 4

Following ideas the of Roy (1957) it is readily seen that the likelihood ratio takes the following form

$$\begin{aligned} \mathcal{L}_n(\widehat{\mathcal{B}}) - \mathcal{L}_n(\delta_0, \widehat{\gamma}) &= (\widehat{\mathcal{B}} - \mathcal{B})^{\mathrm{T}} \mathcal{L}_{\mathcal{B}\mathcal{B}}(\mathcal{B}_*)(\widehat{\mathcal{B}} - \mathcal{B}) \\ &- (\widehat{\gamma} - \gamma)^{\mathrm{T}} \mathcal{L}_{\mathcal{B}\mathcal{B}}(\delta_0, \gamma_*)(\widehat{\gamma} - \gamma), \end{aligned} \tag{E.1}$$

where $\mathcal{B}_*$ is a point between $\mathcal{B}$ and $\widehat{\mathcal{B}}$, likewise γ_* is a point between γ and $\widehat{\gamma}$.

Using arguments of Roy (1957) and Wald (1943), it can be seen that $(E.1)$ for large samples is equivalent to

$$\{n^{-1/2}(\widehat{\delta} - \delta_0)\}^{\mathrm{T}} \mathcal{J} \{n^{-1/2}(\widehat{\delta} - \delta_0)\}.$$

Applying arguments used while proving the Theorem 1 we arrive to (4.6).

APPENDIX G

HISTOGRAMS OF INTERACTION PARAMETER ESTIMATES IN THE BINARY CASE

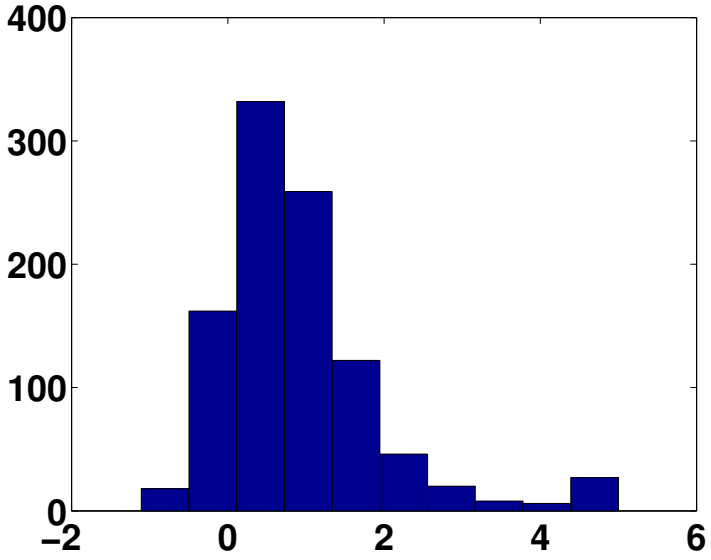

Figure 3: Histogram of $\widehat{\beta}_{xg}$ over 1000 simulations. Disease status (D), genetic variant (G), and environmental covariate (X) are binary and probability of disease is unknown. Environmental variable is measured with error with misclassification probabilities pr$(W = 0|X = 1) = 0.20$ and pr$(W = 1|X = 0) = 0.10$. The results are based on a simulation study of 200 cases and 200 controls.

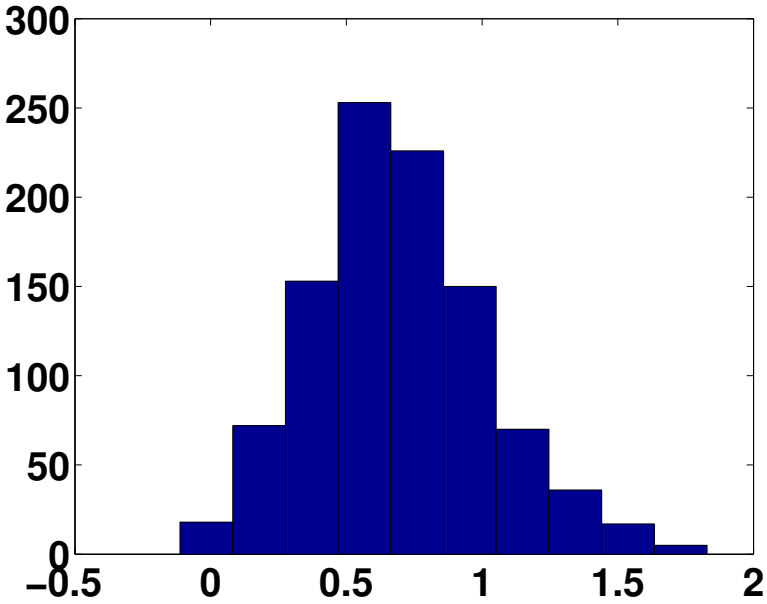

Figure 4: Histogram of $\widehat{\beta}_{xg}$ over 1000 simulations. Disease status (D), genetic variant (G), and environmental covariate (X) are binary and probability of disease is unknown. Environmental variable is measured with error with misclassification probabilities pr$(W = 0|X = 1) = 0.20$ and pr$(W = 1|X = 0) = 0.10$. The results are based on a simulation study of 1000 cases and 1000 controls.

APPENDIX H

HISTOGRAMS OF INTERACTION PARAMETER ESTIMATES IN THE CONTINUOUS CASE

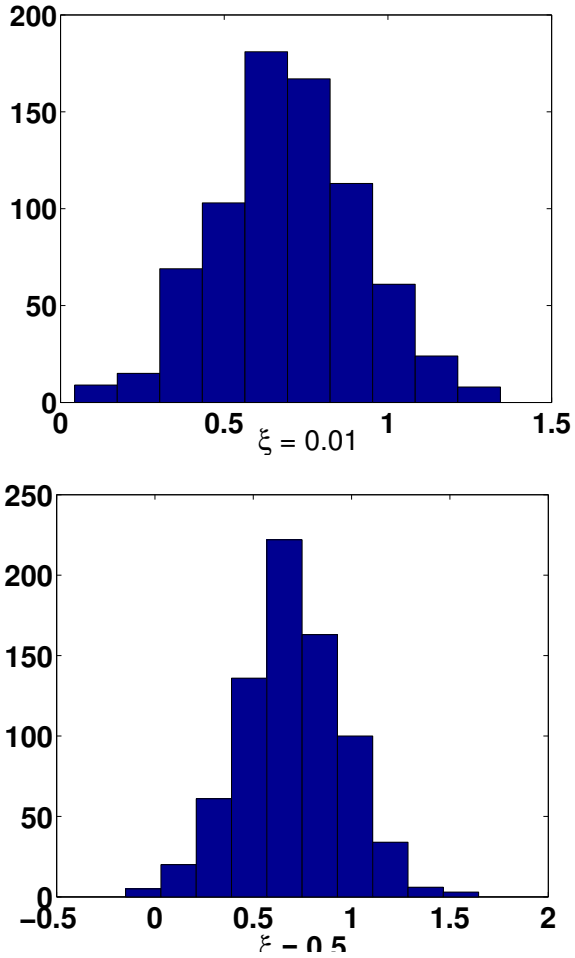

Figure 5: Histogram of $\widehat{\beta}_{xg}$ for different amounts of measurement error: $\xi = 0.01$ and $\xi = 0.05$. Disease status (D), genetic variant (G), and environmental covariate (X) are binary and probability of disease is unknown. Environmental variable is measured with error with misclassification probabilities pr$(W = 0|X = 1) = 0.20$ and pr$(W = 1|X = 0) = 0.10$. The results are based on 1000 replications of 1000 cases and 1000 controls.

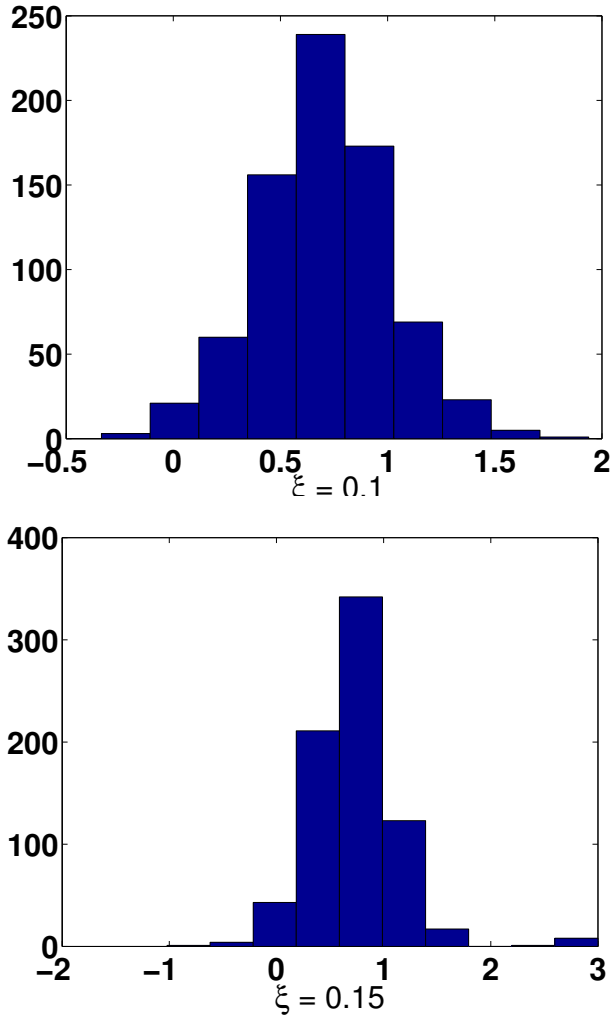

Figure 6: Histogram of $\widehat{\beta}_{xg}$ for different amounts of measurement error: $\xi = 0.10$ and $\xi = 0.15$. Disease status (D), genetic variant (G), and environmental covariate (X) are binary and probability of disease is unknown. Environmental variable is measured with error with misclassification probabilities pr$(W = 0|X = 1) = 0.20$ and pr$(W = 1|X = 0) = 0.10$. The results are based on 100 replications of 1000 cases and 1000 controls.

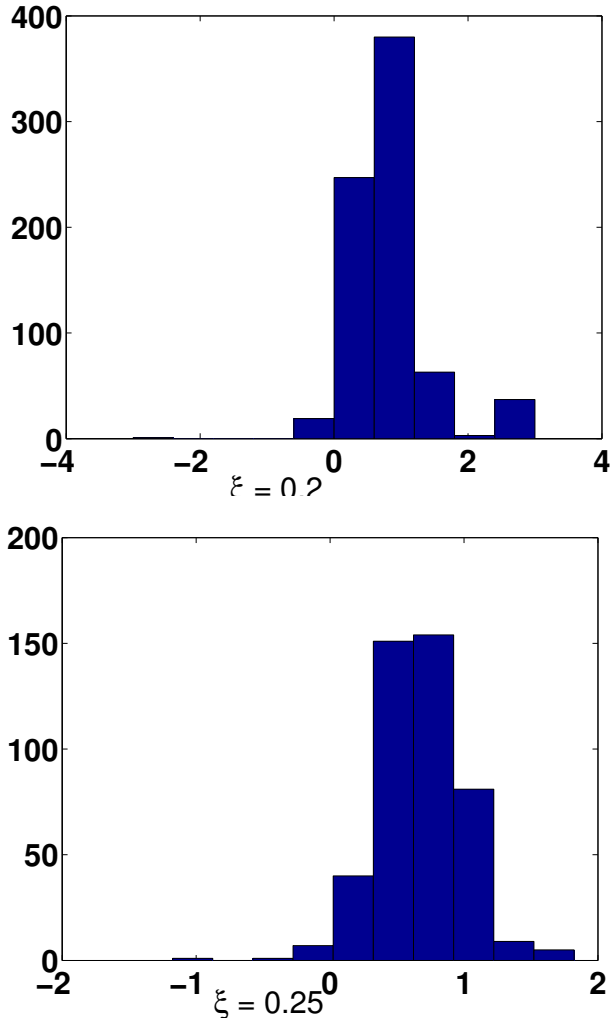

Figure 7: Histogram of $\widehat{\beta}_{xg}$ for different amounts of measurement error: $\xi = 0.20$ and $\xi = 0.25$. Disease status (D), genetic variant (G), and environmental covariate (X) are binary and probability of disease is unknown. Environmental variable is measured with error with misclassification probabilities pr$(W = 0|X = 1) = 0.20$ and pr$(W = 1|X = 0) = 0.10$. The results are based on 1000 replications of 1000 cases and 1000 controls.

VITA

Iryna Lobach was born in Minsk, Belarus. She is the first daughter of Volha V. Lobach and Viktar I. Lobach. Iryna received Bachelor of Science degree from the Belarusian State University, Applied Mathematics and Computer Science Department in Minsk, Belarus in May 2001. That same year she was admitted to the Ph.D. program in the Department of Statistics, at the Texas A&M University. Iryna received her Ph.D. degree in August 2006.

Iryna's permanent address is

Timoshenko 24-1-272

Minsk, Belarus 220134