A Bayesian Approach to Case-control Studies with Errors in the Covariates

by

Marc Vallée

B.Sc, Université de Montréal, 1996

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE STUDIES

(Department of Statistics)

we accept this thesis as conforming
to the required standard

The University of British Columbia

December 1998

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date Dec. 9th, 1998

Abstract

It is not uncommon to be faced with imprecise exposure measurements when dealing with case-control data. In cancer case-control studies, for instance, smoking histories may be unreliable. The usual methods of analysis involve logistic regression with different correction factors. The approach we adopt involves Bayesian fitting of a retrospective discriminant analysis model. The parameters of interest are the regression coefficients in the prospective log-odds ratio for disease. Under a standard non-informative prior, the posterior means of these parameters are infinite. Posterior medians, however, perform reasonably relative to other estimators that adjust for covariate imprecision. For models with only continuous exposures, the Bayesian inference can be implemented with exact posterior simulation.

The presence of binary covariates requires some elements of a covariance matrix to be fixed. We develop a general approach for sampling such a constrained covariance matrix. The Bayesian inference in this context now demands the use of a Gibbs sampling algorithm.

Contents

Abstract ii

Contents iii

List of Figures v

Acknowledgements vii

Dedication viii

1 Introduction 1

2 Literature Review 5

2.1 Basic Terminology 5

2.2 Correction Methods 7

2.2.1 The Armstrong, Whittemore and Howe Method 7

2.2.2 The Rosner, Willet and Spiegelman Methods 9

2.2.3 Other Methods 10

3 The PMED Approach 14

3.1 Methodology 14

3.2 Known Measurement Error Variance 18

3.3 Unknown Measurement Error Variance 21

3.4 Simulation Study 23

3.5 Discussion 30

4 Multivariate Exposure **34**

4.1 Methodology 35

4.2 Binary Covariates 38

4.2.1 The Gibbs Sampler 40

4.2.2 Application 42

4.3 Covariance Matrix With Partially Fixed Diagonal 44

4.3.1 A Simple Case 49

4.3.2 More than Two Binary Covariates 53

4.4 Discussion 56

5 An Application to Real Data **57**

5.1 Bladder Cancer Case-Control Data 58

5.1.1 A Closer Look at the Data 59

5.2 The Analysis 61

6 Conclusion **71**

Bibliography **74**

List of Figures

3.1 Comparison of the PMED and the AWH estimators in the known τ case. The first two panels give $(0.1, 0.3, 0.5, 0.7, 0.9)$ quantiles of the estimators as a function of $\sqrt{\tau}$. The median is displayed with the solid line. The long-dashed line indicate true value of β. The AWH estimator is only considered when τ is sufficiently small that $P(\hat{\lambda} < 0)$ does not exceed 0.05. The third panel gives empirical coverage probabilities of 80% credible intervals, with error bars corresponding to plus and minus two standard errors. 32

3.2 Comparison of the PMED, AWH, RWS1 and RWS2 estimators in the unknown τ case. The format is as per Figure 3.1 33

5.1 Histograms for the distribution of the age and smoking history variables 66

5.2 Diagnostic plots to verify the normality of the logarithmic transformation of the age and smoking history variables 67

5.3 Plots of the coefficients from the logistic regression of the disease status on increasing stratificaions of both the age and logarithmic transformation of the smoking history variables 68

5.4 Plot of the coefficients from the logistic regression of the disease status on increasing stratificaions of the logarithmic transformation of the smoking history variable 69

5.5 Histograms of the posterior distributions of the parameter estimates for β_1, β_2 and β_3 obtained with the assumption there was no measurement error, 10% and finally 25% measurement error in the smoking history variable. 70

Acknowledgements

Sincères remerciements à Paul Gustafson pour son support, sa patience, son ingéniosité et sa grande disponibilité. Paul a su me redonner confiance dans les moments difficiles.

Merci aussi à Nhu Le pour ses précieux conseils et son expérience. Merci à tous les membres du département pour avoir fait de ces deux années une expérience extraordinaire.

Et finalement, merci à ma famille pour ses indispensables encouragements et son support inconditionel.

Marc Vallée

The University of British Columbia

December 1998

A ma famille...

Chapter 1

Introduction

The development of case-control methodology is of great importance to the field of epidemiology. A common problem which seems to arise in these types of studies is the mismeasurement of some of the covariates. Our objective is to develop, using Bayesian methods, an approach that will allow us to reach reasonable conclusions from case-control studies where some of the variables are measured imprecisely.

Three major components are present here: Case-control studies, Errors in covariates and Bayesian methods.

Case-contol studies are usually conducted to investigate the relationship between the presence of disease and of specific risk factors often referred to as the exposure. The central idea, as stated in Breslow (1996): "is the comparison of a group having the outcome of interest to a control group with regard to one or more characteristics." The work of Cornfield (1951) was a great contribution to the development of case-control methodology. He demonstrated that the exposure odds ratio for cases versus controls equals the disease odds ratio for exposed versus unexposed. Also, provided the disease is rare, the exposure odds ratio (known as the relative risk) approximates the disease rate ratio. Basic case-control data can be regarded as two independent samples of covariable vectors, $\{x_{0i}\}_{i=1}^{n_0}$ from $X|D = 0$ (the controls), and $\{x_{1i}\}_{i=1}^{n_1}$ from $X|D = 1$ (the cases), where D denotes the disease status (0 for healthy and 1 for diseased) and X is the covariable vector. The typical analysis of such data is done by fitting a prospective logistic regression model for $D|X$ to the retrospectively sampled data. This procedure is described and justified by Prentice and Pyke (1979).

It is not uncommon to have binary risk factors, in which case errors can occur by misclassification. The approach we will introduce later on, and most of the methods we will review, are designed to deal with continuous exposure variables. These are usually covariates that are more complex to measure as opposed to a simple classification; in which case a surrogate exposure X^* is observed rather than the true exposure X.

Examples of commonly, mismeasured covariates include information about diet or nutrient intake, past history of cigarette or alcohol use, and radiation exposure. The obvious solution is better measuring instruments and techniques, but this is not always feasible. It is widely recognized that in epidemiology or in studies of relationships between a response variable and a set of covariables, these covariates are often measured with error, which can seriously affect the statistical analysis. If ignored, these measurement errors may lead to biased results; the usual consequence is the attenuation of the strength of the relationship. General discussion on classical statistical procedures for dealing with errors in covariates can be found in Fuller (1987) and Carroll (1989). Errors in covariates appear to be difficult to deal with classically; it seems a number of assumptions and approximations have to be introduced in order to proceed with the analysis. A natural alternative is the use of Bayesian methods.

The key feature of the Bayesian approach is that it uses probability theory to describe uncertainty about both parameters and observables. Bayesians think of model parameters as random variables, therefore having a certain probability distribution which can be interpreted as belief about the possible values this parameter can take. This differs from the frequentists who consider model parameters to be fixed. In the Bayesian approach, the information that is available to the experimenter before data are observed and his belief contribute to the specification of a *pior distribution.* A sample is then taken and the data observed so that the prior distribution can be updated with this sample information. This updated prior distribution is called the *posterior distribution.* Inference is then made on the basis of this posterior distribution which is proportional to the product of the likelihood and the prior distribution. The recent developments in algorithms for performing Bayesian computations make this approach particularly appealing. We are referring here to Markov chain Monte Carlo (MCMC) methods; reviews of these methods can be found in both Smith and Roberts (1993) and Besag, Green, Higdon, and

Mergensen (1995).

This makes the Bayesian approach to errors in covariates in case-control studies very sensible and reasonably simple. With the use of a measurement error model we will produce a distribution over the plausible values of the parameters of interest. Although quite natural, this approach has not received much attention in the literature. Richardson and Gilks (1993) and Muller and Roeder (1997) explore Bayesian approaches to errors in covariates in case-control studies. Most of the papers related to this topic suggest classical methods, among those are Armstrong, Whittemore, and Howe (1989), Rosner, Willet, and Spiegelman (1989) and Carroll, Gail, and Lubin (1993). In Mallick, and Gelfand (1995) a Bayesian approach to errors in covariates is presented, but it is not specifically applicable to case-control studies.

In the following chapter we will review the main findings from many of the previously mentioned papers amongst others. In Chapter 3 we will introduce a univariate version of our method along with a simulation study. Chapter 4 will exhibit the generalization to multivariate applications. Finally, an analysis of bladder cancer case-control data will be presented in Chapter 5.

Chapter 2

Literature Review

In this Chapter we will review the methods that have been suggested for correcting for measurement errors. As it was previously mentioned, most of the correction methods make use of the classical theory but we did find results that were obtained within the Bayesian framework.

2.1 Basic Terminology

First, we believe a brief summary of the basic terminology could be quite helpful. Errors can be "random" or "systematic", the key feature of random errors is that the law of large numbers applies; if we were to repeat the measurement many times the mean of these replicates would provide an unbiased estimate of the true quantity we are trying to measure. As opposed to random errors, unbiasedness does not hold for systematic errors; the mean of many repeated measures would not necessarily converge toward the true value.

Errors can also be either "differential" or "nondifferential". This depends on whether the errors are related to the disease outcome D. An example of differential errors would be if the mismeasurement or misclassification of the exposure was dependent on the disease status. Such errors can be the cause of serious bias, but fortunately good study design can help guard against these. A measurement error is nondifferential if it arises in the same way for cases and controls; the errors distribution is independent of the disease outcome. A more formal definition of nondifferential errors can be stated as $X^*|X, D \equiv X^*|X$, or equivalently as $D|X, X^* \equiv D|X$, which means that when the true exposure is known the measured one does not add any additional information. Nondifferential random errors can also be the cause of biased results, and most of the correction methods we will present deal with this type of errors.

Measurement error has traditionally been modeled in two different ways: The "classical error" model and the "Berkson error" model. In the classical error formulation, the conditional distribution of the surrogate X^* given the true value X is specified, while in the Berkson formulation, it is the conditional distribution of X given X^* which is specified.

In order to get an assessment of the measurement error distribution a "validation study" is often used. In a validation study a "gold standard" measurement of X is obtained and compared to the surrogate measure X^* which will be used in the main study. In this manner a direct estimate of the error distribution is provided, but since gold standard measurements are generally quite expensive, they can only be performed on a small portion of the whole sample.

This short terminology review should make the subsequent sections easier to follow. Now we will present techniques suggested by different authors, some on which our own approach was based, and others that served as comparison for our work.

2.2 Correction Methods

In this section we introduce different methods to analyse case-control data with imprecise exposure measurement. Two of these methods are presented in greater details because of the similarity of their initial setting to ours.

2.2.1 The Armstrong, Whittemore and Howe Method

In their 1989 paper, Armstrong, Whittemore and Howe [hereafter AWH] suggest a method of correcting a standard logistic regression analysis to account for measurement errors. In order to adjust for the effect of the measurement error on the logistic regression coefficients obtained from case-control data, a multivariate discriminant analysis model is assumed for the joint distribution of the true covariate values and errors among cases and controls. Their approach is applicable to multiple strata designs, but for simplicity we will present here the one-stratum, one-dimensional case.

They consider x_D as a p-dimensional (p=1, for the present case) row vector of unknown true covariates for case (D=1) and control (D=0). Following the dicriminant analysis model:

$$X_D \sim N(\mu + D\Delta, \sigma),$$

where $N(\cdot, \cdot)$ stands for the normal distribution with mean $\mu{+}D\Delta$ and variance σ. This notation for the normal distribution will be used throughout the thesis. The cases and controls have a common variance σ, μ is the common part of the mean for both groups and Δ represents by how much the case mean differs from the control one. The information that is available is from l_D measurements of the flawed X_D^* (we will consider $l_D = 1$ here) with:

$$X_D^* = X_D + e_D,$$

where

$$e_D \sim N(\gamma + D\delta, \tau).$$

The e_Ds are the error components and are independent of each other and of the X_Ds. In their model the parameter γ stands for the mean error common to cases and controls and δ represents the systematic difference in error between cases and controls, therefore allowing for differential errors. The parameter τ is the error variance. These assumptions imply the following:

$$X_0^* \sim N(\mu + \gamma, \sigma + \tau), \qquad X_1^* \sim N(\mu + \gamma + \Delta + \delta, \sigma + \tau).$$

Then using a well known relationship between discriminant analysis and logistic regression, they obtain a model of the form:

$$logit\{P(D = 1|X = x)\} = \alpha + \beta^T x.$$

Fitting this model with x_0^* and x_1^*, realizations of X_0^* and X_1^* respectively, they get a 'naive' estimate $\hat{\beta}$ of β. They propose an estimator of the prospective log-odds ratio $\beta_{AWH} = \beta/\lambda$, where λ is the correction factor which adjusts for measurement error. An estimate of β_{AWH} can be obtained by the 'naive' $\hat{\beta}$ and the appropriate estimate of λ. In the unidimensional case this correction factor is $\lambda = \sigma/(\tau + \sigma)$; the estimators are $\hat{\lambda} = 1 - (\tau/S_p^2)$ if τ is known, and $\hat{\lambda} = 1 - (S_v^2/S_p^2)$ if τ is unknown. Here $S_p^2 = (SS_0 + SS_1)/(n_0 + n_1 - 2)$ is the pooled estimator of the variance $\tau + \sigma$ and $SS_D = \sum_i (x_{Di}^* - \bar{x}_D^*)^2$, D=0,1 and n_D is the sample size. While $S_v^2 = n_2^{-1} SS_2$ estimates τ from a validation sample which would be needed when τ is unknown and $SS_2 = \sum_i (x_{2i}^* - x_{2i})^2$.

In Chapter 3 a simulation study that gives a comparison of the AWH approach with the one proposed in this work will be presented.

2.2.2 The Rosner, Willet and Spiegelman Methods

Another approach is proposed by Rosner, Willet, and Spiegelman (1989) [hereafter RWS]. In their paper two methods are provided to correct relative risk estimates obtained from logistic regression models for measurement error in continuous covariates. Both methods require a separate validation study to estimate the regression coefficient λ relating the surrogate measure to the true exposure.

They first assume that the model relating a single-dimensional true exposure X and the probability of disease D is of the logistic form:

$$logit\{P(D = 1|X = x)\} = \alpha + \beta x.$$

Then a linear relationship is assumed to exist between true exposure X and observed exposure X^* of the form:

$$X = \alpha' + \lambda X^* + \varepsilon, \text{ where } \varepsilon \sim N(0, \sigma^2).$$

Finally, they assume nondifferential errors and that the conditional distribution of X given X^* and the marginal distribution of X^* are the same for the main and validation study populations.

Their first method is a linear approximation which yields an estimate of β_{RWS1} of the same form as β_{AWH} $(\hat{\beta}/\hat{\lambda})$. Here $\hat{\lambda}$ is obtained by regressing X on X^* using the above linear model with ordinary least squares and the validation study data.

Their second method is a likelihood approximation where a second-order Taylor series expansion is used to approximate the logistic function, enabling closed-form likelihood estimation of β_{RWS2}.

Again, performance of these methods will be examined in the simulation study presented in Chapter 3. A multivariate versions of their approaches is presented in Rosner, Spiegelman, and Willet (1990).

2.2.3 Other Methods

Following the likelihood approach Carroll, Gail, and Lubin (1993) derive pseudolikelihoods on which to base estimators of the parameters of a prospective logistic model for case-control data that are, of course, measured with error. They also examine computationally simpler methods where the conditional expectation of the true covariate X knowing the surrogate X^* is substituted for X in the logistic model.

Prentice and Pyke (1979) established the equivalence of the asymptotic distributions of estimators based on the prospective and retrospective likelihoods under case-control sampling, and Breslow and Day (1980) suggested the use of estimators based on the conditional likelihood. Forbes and Santner (1995) continue in the same vein looking at the effect of measurement error on the conditional maximum likelihood estimator (CMLE). They then go on to suggest three alternative estimators correcting for the measurement error: The first based on a correction for the asymptotic bias of the CMLE, the second is a functional estimator, and the third is a "transformed" estimator obtained by computing the CMLE using transformed covariates.

In Buonaccorsi (1990), he makes use of the normal discriminant analysis model setting and the correction for measurement error is made possible by a double sampling scheme in which the surrogates are collected on all units and true values are obtained on a random subset of units, allowing the consideration of a large set of measurement error models.

Yet another approach is presented in Roeder, Carroll, and Lindsay (1996), in which a semiparametric mixture method is introduced. By using a mixture model, the relationship between the surrogate X^* and the true covariate X can be modeled. The likelihood depends on the marginal distribution X and the measurement error density; this measurement error density is parametrically based on a validation sample, and the marginal of X is modeled using a nonparametric mixture distribution.

Work in the classical setting has also been done on sample size calculations for case-contol studies with errors in the covariates, both McKeown-Eyssen and Thomas (1985), and McKeown-Eyssen and Tibshirani (1994) discuss that matter.

Although many problem in epidemiology can be naturally formulated in the Bayesian framework, the approach was not pursued due to computational complexities. But the introduction of Markov chain Monte Carlo sampling methods has really opened the way, and the literature on this subject is now more common.

Richardson and Gilks (1993) take a Bayesian perspective on measurement error problems in epidemiology. The authors construct what they call a conditional independence model which is equivalent to considering nondifferential errors. They also introduce a graphical representation to this type of model. Then they indicate how Bayesian estimation can be carried out in these settings using a Gibbs sampler.

Also using the Bayesian approach is Muller and Roeder (1994), paper in which they present a semiparametric model for case-control studies with errors in variables. The approach proposed in this paper is based on a nonparametric model for the exposure and a parametric disease model. The model they present is complex in structure, but is said to be simple to implement.

We have reviewed here the papers we think are most relevant to the problem we will be tackling ourselves in the subsequent Chapters. The main role the material found in these articles will be playing is guiding us in building our own initial model, thus making comparisons possible. The methodology we will be presenting will combine parts of the different approaches found in the literature that have not been tried together.

Chapter 3

The PMED Approach

To develop our own correction method we investigate the use of a retrospective discriminant analysis model for the unobserved real exposure, which leads to a Bayesian variance component model. In this chapter we present the basic methodology of our approach and illustrate it in the simple context of a univariate exposure. The method is applied to both scenarios where the measurement error variance is known and unknown. A simulation study is then carried out, and results are compared to the ones obtained with established methods.

3.1 Methodology

As for the methods we reviewed earlier we have D and X that respectively represent an individual's disease status and covariable vector, with D coded as 0 (healthy) or 1 (diseased). Case-control data are obtained retrospectively and, as it was mentioned in the introduction, can be regarded as two independent covariable vectors, $\{x_{0i}\}_{i=1}^{n_0}$ from $X|D=0$ (the controls), and $\{x_{1i}\}_{i=1}^{n_1}$ from $X|D=1$ (the cases). We want to perform a fully Bayesian analysis of the retrospective data which demands a likelihood function based on the sampled distribution $X|D$, as opposed to the typical fitting of a prospective logistic regression model for $D|X$. Since our main objective is to learn about the prospective relationship $D|X$, we have to examine the form of the prospective model implied by a specified retrospective model.

Assuming for now that all the covariables are continuous, a simple model to consider is one suggested by AWH, the normal discriminant analysis model:

$$X|D=0 \sim N(\mu_0,\Sigma_0), \qquad X|D=1 \sim N(\mu_1,\Sigma_1). \tag{3.1}$$

The logistic regression model can be rewritten as:

$$\begin{aligned} logit\{P(D=1|X=x)\} &= log\left\{\frac{P(D=1|X=x)}{1-P(D=1|X=x)}\right\} \\ &= log\left\{\frac{P(D=1|X=x)}{P(D=0|X=x)}\right\} \\ &= log\left\{\frac{P(X=x|D=1)P(D=1)}{P(X=x|D=0)P(D=0)}\right\} \\ &= log\left\{\frac{P(X=x|D=1)}{P(X=x|D=0)}\right\} + log\left\{\frac{P(D=1)}{P(D=0)}\right\} \\ &\equiv log\left\{\frac{f(X|D=1)}{f(X|D=0)}\right\} + \text{constant}. \end{aligned}$$

Plugging in the distributions suggested by the normal discriminant analysis model (3.1), we obtain:

$$log\left\{\frac{f(X|D=1)}{f(X|D=0)}\right\} = log\left\{\frac{c_1 \exp\left(-\frac{1}{2}(x-\mu_1)'\Sigma_1^{-1}(x-\mu_1)\right)}{c_0 \exp\left(-\frac{1}{2}(x-\mu_0)'\Sigma_0^{-1}(x-\mu_0)\right)}\right\}$$

$$\begin{aligned} &\propto -\frac{1}{2}\{(x-\mu_1)'\Sigma_1^{-1}(x-\mu_1)+(x-\mu_0)'\Sigma_0^{-1}(x-\mu_0)\} \\ &\propto (\Sigma_1^{-1}\mu_1-\Sigma_0^{-1}\mu_0)x+x^T((\Sigma_0^{-1}-\Sigma_1^{-1})/2)x. \end{aligned}$$

Hence, the retrospective model implies a prospective model of the form

$$logit\{P(D=1|X=x)\} \;=\; \alpha+\beta^T x+x^T Cx, \tag{3.2}$$

where $\beta = \Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0$, and $C = (\Sigma_0^{-1} - \Sigma_1^{-1})/2$, Given the fact that β and C depend only on the parameters of (3.1), they can be estimated from case-control data. It is not the case for the intercept α. Because it also depends on the marginal distribution of D, α is not estimable from case-control data alone. Information about the disease prevalence would be needed. The relationship between (3.1) and (3.2) is well known; it is exploited by Armstrong, Whittemore and Howe (1989) and Muller, Parmigiani, Schildkraut and Tardella (1997) amongst others.

For the remainder of this chapter ideas will be presented in the univariate exposure setting, i.e. with X being a scalar. We will also assume a common variance $\nu = \Sigma_0 = \Sigma_1$ for both cases and controls. Model (3.1) therefore becomes:

$$X|D=0\sim N(\mu_0,\nu), \qquad X|D=1\sim N(\mu_1,\nu), \tag{3.3}$$

which implies the prospective relationship

$$logit\{P(D=1|X=x)\} \;=\; \alpha+\beta x, \tag{3.4}$$

with $\beta = (\mu_1 - \mu_0)/\nu$ being the parameter of interest. Typically, the interpretation for β is that $\exp\{\beta(x'-x)\}$ is the disease odds ratio for exposure

x' compared to exposure x. Additionally, for rare diseases the relative risk of disease is approximately equal to this odds ratio.

An important point that has to be considered is how much stronger an assumption is (3.3) compared to (3.4). If we consider f_i to be the density of $X|D = i$, then we get exactly (3.4) when

$$\frac{f_1(x)}{f_0(x)} \propto \exp(\beta x). \tag{3.5}$$

(3.5) can obviously be satisfied by densities f_0 and f_1 which are not normal, but it does restrict their tail behavior. If we assume without loss of generality that $\beta > 0$, then (3.5) implies that the right tail of f_0 and the left tail of f_1 fall off faster than $\exp(-\beta|x|)$. If these 'thinner than exponential' tails are in fact normal, then the other tails (the left tail of f_0 and the right tail of f_1) will also be normal. This suggests that (3.3) is likely to be appropriate in at least some situations where (3.4) holds. In practice transformations are often used to approximate normality. If such a transformation was used on the sample of control exposures for example, then in order to proceed with the retrospective model suggested here, the same transformation would have to yield approximate normality with a similar variance for the other exposure sample.

We now have a model for the relationship between the exposure and the disease outcome. Next we have to consider the measurement error which will arise in that a surrogate exposure X^* will be observed rather than the true exposure X. So we also have to model this error, we suggest the following simple classical (non-Berkson) model:

$$X^*|X \sim N(X, \tau) \tag{3.6}$$

under which X^* is an unbiased but noisy measurement of X. We assume the measurement error is nondifferential; that is, it arises in the same way for cases and controls. Using the formal definition of nondifferential error given in chapter 2, (3.3) and (3.6) can be collapsed to obtain the next model, from which the data arise:

$$X^*|D=0 \sim N(\mu_0, \tau+\nu), \qquad X^*|D=1 \sim N(\mu_1, \tau+\nu). \tag{3.7}$$

So a direct consequence of the measurement error is some extra variability in the data compared to (3.3). This extra variability τ (the measurement error variance) could be known from an external validation of the measurement process, or it could be estimated with the use of a validation sample as it was illustrated in chapter 2. The following two sections will illustrate our approach for either situation.

3.2 Known Measurement Error Variance

In the present section we will assume τ to be known. Since we are conducting a Bayesian analysis, we need a prior distribution in addition to the likelihood based on (3.7). This will yield a posterior distribution on the unknown parameters (μ_0, μ_1, ν) and enable us to estimate β. The prior distribution on which our inference will be based is a standard noninformative prior for variance component models,

$$\pi(\mu_0, \mu_1, \nu) \propto (\tau + \nu)^{-1}. \tag{3.8}$$

Formally, this prior is known as the reference prior (Berger and Bernardo, 1992) when ν is the parameter of interest and (μ_0, μ_1) are nuisance parameters. The reference prior for β as the parameter of interest was also considered, but the fact that it did not have a simple form made it unappealing.

Given the prior distribution in (3.8) and the realization $x^* = \{x_{10}^*, \cdots, x_{n_00}^*, x_{11}^*, \cdots, x_{n_11}^*\}$ from the normal discriminant analysis model in (3.7), the resulting posterior is of the following form:

$$\begin{aligned}
\pi(\mu_0, \mu_1, \nu | x^*) \quad &\propto \quad \overbrace{(\nu + \tau)^{-1}}^{\text{prior}} \\
&\quad \times \prod_{i=1}^{n_0} (\nu + \tau)^{-1/2} \exp\left(-\frac{(x_{i0}^* - \mu_0)^2}{2(\nu + \tau)}\right) \\
&\quad \times \prod_{i=1}^{n_1} (\nu + \tau)^{-1/2} \exp\left(-\frac{(x_{i1}^* - \mu_1)^2}{2(\nu + \tau)}\right) \\
&\propto \quad \left(\frac{1}{\nu + \tau}\right)^{\frac{n_0+n_1}{2}+1} \\
&\quad \times \exp\left(-\frac{n_0(\mu_0 - \bar{x}_0^*)^2}{2(\nu + \tau)}\right) \exp\left(-\frac{\sum_{i=1}^{n_0}(x_{i0}^* - \bar{x}_0^*)^2}{2(\nu + \tau)}\right) \\
&\quad \times \exp\left(-\frac{n_1(\mu_1 - \bar{x}_1^*)^2}{2(\nu + \tau)}\right) \exp\left(-\frac{\sum_{i=1}^{n_1}(x_{i1}^* - \bar{x}_1^*)^2}{2(\nu + \tau)}\right)
\end{aligned} \tag{3.9}$$

Our Bayesian inference can be implemented by simulating independent draws from this posterior distribution. We can then integrate out μ_0 and μ_1 from (3.9) to get

$$\pi(\nu | x^*) \propto \left(\frac{1}{\nu + \tau}\right)^{\frac{n_0+n_1}{2}} \exp\left(-\frac{\sum_{i=1}^{n_0}(x_{i0}^* - \bar{x}_0^*)^2 + \sum_{i=1}^{n_1}(x_{i1}^* - \bar{x}_1^*)^2}{2(\nu + \tau)}\right). \tag{3.10}$$

In order to make sampling easier, it is convenient to reparametrize from (μ_0, μ_1, τ) to (μ_0, μ_1, γ), where $\gamma = \tau + \nu$. The posterior distribution of $\gamma|x^*$ is then easily obtained from (3.10) and can be expressed as

$$\gamma|x^* \equiv G|G > \tau, \tag{3.11}$$

where G has an inverse gamma distribution with shape parameter $(n_0 + n_1 - 2)/2$ and scale parameter $(SS_0 + SS_1)/2$, with $SS_i = \sum_j (x^*_{ij} - \bar{x}^*_i)^2$. Sampling from $G|G > \tau$ is easily implemented by repeatedly sampling G until $G > \tau$. Then ν is taken to be the difference between the sampled G and the known τ. Now that we have sampled γ (or ν) it is easy to see from (3.9) that we have

$$\begin{aligned} \mu_0|\gamma, x^* &\sim N(\bar{x}^*_0, n_0^{-1}\gamma), \\ \mu_1|\gamma, x^* &\sim N(\bar{x}^*_1, n_1^{-1}\gamma), \end{aligned} \tag{3.12}$$

where μ_0 and μ_1 are independent given γ and x^*. A draw from the joint posterior $\mu_0, \mu_1, \gamma|x^*$ can be obtained by sampling from $\gamma|x^*$ and then from $\mu_0, \mu_1|\gamma, x^*$. Thus we have a simple algorithm for exact posterior simulation in a variance components model. This computational approach is pursued in greater generality by Wolfinger and Kass (1996).

Commonly, Bayesian parameter point estimation is done by using the posterior distribution's mean. It can be seen by the following that in the present case $E(\beta|x^*)$ is infinite;

$$\begin{aligned} E\{(\mu_1 - \mu_0)\nu^{-1}|x^*\} &= E\{\nu^{-1}E(\mu_1 - \mu_0|\nu, x^*)\} \\ &= E\{\nu^{-1}(\bar{x}^*_1 - \bar{x}^*_0)|x^*\} \Leftarrow \text{ by } (3.12) \\ &= (\bar{x}^*_1 - \bar{x}^*_0)E\{\nu^{-1}|x^*\} \end{aligned}$$

but

$$E\{\nu^{-1}|x^*\} = \frac{\int_\tau^\infty (\nu-\tau)^{-1}\nu^{-(\alpha+1)}\exp(-\omega/\nu)}{\int_\tau^\infty \nu^{-(\alpha+1)}\exp(-\omega/\nu)},$$

where α and ω are both positive and are respectively the shape and scale parameter of an inverse gamma distribution, the numerator's integral blows up near τ, therefore $E\{\nu^{-1}|x^*\} = +\infty$. Intuitively this occurs because, according to both the likelihood based on (3.7) and the prior (3.8), zero is a plausible value for ν which is the denominator of β. Since the posterior mean of β does not exist, we use the posterior median of our parameter of interest as a point estimator. From now on we will refer to this posterior median as PMED.

This concludes the known measurement error variance case. We will come back to it in a subsequent section to reveal the procedure and the results of a simulation study.

3.3 Unknown Measurement Error Variance

We will now examine the situation where τ is unknown, in which case it will have to be estimated along with ν. The estimation of this error variance will be done with the use of a validation sample. In addition to the surrogate exposures for cases and controls, we assume an independent data set is available. This extra data set will consist of measurements of both the surrogate exposure X^* and the true exposure X for each subject. We have assumed nondifferential measurement error, thus the subjects selection procedure for this validation sample does not matter; we can view the true exposure X as being sampled from an arbitrary distribution, so long as the measurement error model $X^*|X \sim N(X, \tau)$ is the same for both the main and validation samples. The observed validation sample is denoted $\{(x^*_{2i}, x_{2i})\}_{i=1}^{n_2}$, with $SS_2 = \sum_i (x^*_{2i} - x_{2i})^2$ defined for subsequent use.

To conduct our analysis we can proceed in a very similar manner as in the known τ case. We begin by doing a reparametrization similar to the one done in the previous case. The variance components (ν, τ) are reparametrized to (γ, τ), where $\gamma = \tau + \nu$. We again use the reference prior for a variance components model but this time when (γ, τ) are the parameters of interest

$$\pi(\mu_0, \mu_1, \gamma, \tau) \propto \gamma^{-1}\tau^{-1} \qquad (0 \leq \tau < \gamma < \infty) \tag{3.13}$$

Once more we have selected in (3.13) a noninformative prior. Combined with the likelihood given by the normal discriminant analysis model in (3.7), we can write the resulting posterior distribution from which we will be sampling:

$$\begin{aligned} \pi(\mu_0, \mu_1, \gamma, \tau | x^*) \;\; \propto \;\; & \overbrace{(\gamma^{-1}\tau^{-1}) I\{\gamma > \tau\}}^{\text{prior}} \\ & \times \left(\frac{1}{\gamma}\right)^{\frac{n_0}{2}} \exp\left(-\frac{n_0(\mu_0 - \bar{x}^*_0)^2}{2\gamma}\right) \exp\left(-\frac{\sum_{i=1}^{n_0}(x^*_{i0} - \bar{x}^*_0)^2}{2\gamma}\right) \\ & \times \left(\frac{1}{\gamma}\right)^{\frac{n_1}{2}} \exp\left(-\frac{n_1(\mu_1 - \bar{x}^*_1)^2}{2\gamma}\right) \exp\left(-\frac{\sum_{i=1}^{n_1}(x^*_{i1} - \bar{x}^*_1)^2}{2\gamma}\right) \\ & \times \left(\frac{1}{\tau}\right)^{\frac{n_2}{2}} \exp\left(-\frac{\sum_{i=1}^{n_2}(x^*_{i2} - x_{2i})^2}{2\tau}\right). \end{aligned} \tag{3.14}$$

Once again we can integrate out μ_0 and μ_1, this time obtaining:

$$\begin{aligned} \pi(\gamma, \tau | x^*) \;\; \propto \;\; & \left(\frac{1}{\gamma}\right)^{\frac{n_0-1}{2} + \frac{n_1-1}{2} + 1} \exp\left(-\frac{\sum_{i=1}^{n_0}(x^*_{i0} - \bar{x}^*_0)^2 + \sum_{i=1}^{n_1}(x^*_{i1} - \bar{x}^*_1)^2}{2\gamma}\right) \\ & \left(\frac{1}{\tau}\right)^{\frac{n_2-1}{2} + 1} \exp\left(-\frac{\sum_{i=1}^{n_2}(x^*_{i2} - x_{2i})^2}{2\tau}\right) I\{\gamma > \tau\}. \end{aligned} \tag{3.15}$$

Hence from (3.15) the marginal posterior distribution of the variance components is

$$(\gamma, \tau)|x^* \equiv (G, T)|G > T, \tag{3.16}$$

where G and T are independent, with

$$G \sim IG\left(\frac{n_0 + n_1 - 2}{2}, \frac{SS_0 + SS_1}{2}\right), \quad T \sim IG\left(\frac{n_2 - 1}{2}, \frac{SS_2}{2}\right), \tag{3.17}$$

and the posterior conditional distribution for μ_0 and μ_1 once γ and τ have been sampled is the same as in the known τ case

$$\begin{aligned} \mu_0|\gamma, \tau, x^* &\sim N(\bar{x}_0^*, n_0^{-1}\gamma), \\ \mu_1|\gamma, \tau, x^* &\sim N(\bar{x}_1^*, n_1^{-1}\gamma). \end{aligned} \tag{3.18}$$

Thus again exact posterior sampling can be implemented, by simulating (G, T) pairs from the distributions in (3.17) until $G > T$, and then sampling μ_0 and μ_1 from (3.18) in order to get a set of β's. The posterior mean for the parameter of interest is infinite in this case too, so again our PMED is used as a point estimator.

3.4 Simulation Study

Situations with known and unknown measurement error variance have now been addressed. In both cases our approach led to exact posterior sampling, which should make implementation simple and sampling procedures reasonably efficient In order to assess the performance of our approach we conduct the analysis of simulated data sets and compare our results to established methods previously illustrated in chapter 2.

Simulations are carried out based on samples size $n_0 = n_1 = 50$ from the retrospective model (3.3) with parameter values (μ_0, μ_1, ν)=$(0, 1, 1)$. Under these parameters, the prospective logistic regression coefficient is $\beta = 1$. The simulations are done with different values of $\sqrt{\tau}$ ranging from 0.0 to 2.0 by jumps of 0.1. In one case τ is assumed to be known. In the other it has to be estimated from a simulated validation sample of size $n_2 = 50$. In either situation, the PMED estimate of β is computed as the median of 500 independent and identically distributed draws from the posterior distribution.

Here are the multiple steps of the simulation process for the τ known case:

1. Generate 500 samples of size $n_0 = 50$ from a normal distribution with mean 0 and variance $\nu = 1$.

2. Generate 500 samples of size $n_1 = 50$ from a normal distribution with mean 0 and variance $\nu = 1$.

3. Create a vector of $\sqrt{\tau}$'s, (0.0, 0.1, 0.2,...1.8, 1.9, 2.0).

4. Use the 500 samples generated in step 1. to create 500 new samples that will be from a normal distribution with mean $\mu_0 = 0$ and variance τ (the controls).

5. Use the 500 samples generated in step 2. to create 500 new samples that will be from a normal distribution with mean $\mu_1 = 1$ and variance τ (the

cases).

6. Sample in turn ν, μ_0 and μ_1.

   - Sample γ from (3.11) to get ν.
   - Use ν to sample both μ_0 and μ_1 from (3.12).
   - Compute $\beta = \nu^{-1}(\mu_1 - \mu_0)$.

7. Repeat step 6. 500 times, which gives us 500 independent and identically distributed draws from the posterior distribution.

8. Compute the PMED estimate of β, the median of the 500 draws from the posterior.

9. Repeat steps $4., 5., 6.$ and 7. with every value $\sqrt{\tau}$ was given.

The reason for steps 1. and 2. is that initializing all the future samples with these (which come from standard normal distributions), makes comparison from one measurement error level to another easier and more precise.

The process is very similar for the τ unknown case, except one step is added between steps 2. and 3. and between steps 5. and 6. to create the validation sample, and step 6. is slightly different. Here are these two added steps and the correction for step 6.:

1. Generate 500 samples of size $n_2 = 50$ from a normal distribution with mean 0 and variance $\nu = 1$.

2. Use the 500 samples generated in the previous step to create 500 new samples that will be from a normal distribution with mean $\mu_2 = 0$ and variance τ (the validation sample).

3. Sample in turn ν, μ_0 and μ_1.

   - Sample γ and τ from (3.16, 3.17) to get ν.
   - Use ν to sample both μ_0 and μ_1 from (3.18).
   - Compute $\beta = \nu^{-1}(\mu_1 - \mu_0)$.

Quite obviously we did not generate both surrogate and true exposures in our validation sample, it actually consists in $\{(x^*_{2i} - x_{2i})\}_{i=1}^{n2}$, the differences between the surrogate and true exposures. These differences should follow a normal distribution with mean 0 and variance τ.

The simulation routines were coded in C, and although probably not using the most efficient algorithms, they compared very favorably to S-plus with respect to running time and memory resources needed.

The PMED estimator of β is compared to other estimators suggested in the literature. For the known τ case, the comparison is done with a method by AWH, illustrated in chapter 2, in which a correction factor λ is used to adjust the β estimate obtained from a 'naive' logistic regression. For the τ unknown case, the AWH method is still applicable and two others by RWS, also discussed in chapter 2, are performed on our simulated data for comparison. The first by RWS, which we refer to as RSW1, takes the form of the AWH estimator, except that the correction factor λ is estimated by the fitted slope of a regression of

X on X^* using the validation sample. The second estimator, referred to as RWS2, is based on an approximate likelihood function. The RWS method needs to fulfill more assumptions then the PMED approach in order to be applicable. It requires that the variance of X be the same for the validation sample as for the case and control samples. If this requirement is not met, the fitted slope will no longer estimate λ correctly. This assumption is not needed for the PMED method to work, it is applicable to any distribution of X in the validation sample. This seems more sensible as it is parameters of the conditional distribution for $X^*|X$ which must be estimated.

The AWH estimator has a definite disadvantage in that it performs poorly when faced with substantial measurement error. The estimate $\hat{\lambda}$ of the correction factor can sometimes take on a negative value, in fact, $P(\hat{\lambda} < 0)$ can be quite high when τ is sufficiently large. We can obtain $\hat{\lambda}$'s distribution and derive the following:

$$P\left(\hat{\lambda} < 0\right) = F_{n_0+n_1-2}\left(\frac{n_0+n_1-2}{1+\nu/\tau}\right), \tag{3.19}$$

where F_k is the chi-square distribution function with k degrees of freedom. The derivation of this probability is done using information on the AWH approach presented in Chapter 2. We illustrated there that in the known τ scenario λ can be estimated by $\hat{\lambda} = 1 - (\tau/S_p^2)$, where

$$\left(\frac{n_0+n_1-2}{\nu+\tau}\right) S_p^2 \sim \chi^2_{n_0+n_1-2}.$$

From there we can write:

$$P\left(\hat{\lambda} < 0\right) = P\left(1 - (\tau/S_p^2) < 0\right)$$

$$\begin{aligned} &= P\left(\tau/S_p^2 > 1\right) \\ &= P\left(S_p^2 < \tau\right) \\ &= P\left(\left(\frac{n_0+n_1-2}{\nu+\tau}\right)S_p^2 < \left(\frac{n_0+n_1-2}{\nu+\tau}\right)\tau\right) \\ &= P\left(\chi^2_{n_0+n_1-2} < \left(\frac{n_0+n_1-2}{1+\nu/\tau}\right)\right) \end{aligned}$$

Similarly, in the unknown τ case,

$$P\left(\hat{\lambda} < 0\right) \;=\; F_{n_0+n_1-2,n_2}\left(\frac{1}{1+\nu/\tau}\right), \qquad (3.20)$$

where $F_{k,l}$ is the F distribution function with k and l degrees of freedom. This time the probability was derived given that in the unknown τ scenario λ is estimated by $\hat{\lambda} = 1 - S_v^2/S_p^2$, and

$$\frac{S_p^2/(\nu+\tau)}{S_v^2/\tau} = \left(\frac{1}{1+\nu/\tau}\right)\frac{S_p^2}{S_v^2} \sim F_{n_0+n_1-2,n_2}.$$

Again we can write:

$$\begin{aligned} P\left(\hat{\lambda} < 0\right) &= P\left(1 - S_v^2/S_p^2 < 0\right) \\ &= P\left(S_v^2/S_p^2 > 1\right) \\ &= P\left(S_p^2/S_v^2 < 1\right) \\ &= P\left(\left(\frac{1}{1+\nu/\tau}\right)\frac{S_p^2}{S_v^2} < \left(\frac{1}{1+\nu/\tau}\right)\right) \\ &= P\left(F_{n_0+n_1-2,n_2} < \left(\frac{1}{1+\nu/\tau}\right)\right) \end{aligned}$$

Because of this the the AWH estimator is implemented in our simulation study only when the measurement error τ is small enough to ensure that either (3.19) or (3.20) does not exceed 0.05.

Results of the simulations for τ known and τ unknown are illustrated respectively in Figure 3.1 and Figure 3.2, which can be found at the end of the present chapter. In the first case, for each value of τ 500 independent data sets are simulated, for each data set PMED and AWH estimates are computed. Using the sampling methods presented above in the different steps of the simulation we obtain samples of these estimates. The empirical $(0.1, 0.3, 0.5, 0.7, 0.9)$ quantiles of each estimator are displayed as functions of $\sqrt{\tau}$ in the first two panels of Figure 3.1. We can see from these two panels that for small measurement error both methods yield sampling distributions of the estimators that are very similar, but as τ gets larger the PMED estimator is more tightly centered about the true value $\beta = 1$. Even when (3.19) becomes non-negligeable and the AWH estimate is considered inappropriate, the PMED estimator keeps performing quite reasonably.

We also proceed to construct Bayesian credible intervals for β. If we take the $\alpha/2$ and $1 - \alpha/2$ quantiles of the posterior distribution of β, the interval between these constitutes an equal-tailed $(1 - \alpha)$ credible interval for β. These posterior quantiles can be estimated by empirical quantiles from the simulated posterior sample. Based on the 500 data sets, empirical coverage probabilities of 80% credible intervals are displayed in the third panel of Figure 3.1. The associated 'error bars' (plus and minus two standard errors) indicate that the credible intervals can be reasonably interpreted as 80% frequentist confidence intervals.

For the unknown τ case, Figure 3.2 compares the PMED, AWH, RWS1 and RWS2 estimators. The format is the same as in Figure 3.1. In the first four panels the sampling distributions of the estimators are presented by five empirical quantiles. The comparison between the PMED and AWH estimators is similar to that in the known τ case, but the AWH estimate becomes inappropriate because (3.20) becomes non-negligeable even faster than in the previous case. The RWS estimators behave quite similarly to the PMED estimator, and in fact have slightly better performance for large measurement error variance τ. However, as noted previously, the PMED estimator works under less restrictive assumptions than the RWS estimators. In the last panel are displayed the empirical coverage probabilities of 80% credible intervals. Again these seem consistent with the confidence interval interpretation.

3.5 Discussion

We have now introduced methodology in a general setting up to a certain point after which we have illustrated our approach in the simple context of a univariate exposure for two particular cases. The simulations seem to indicate the equivalent or superior performance of the posterior median estimator compared to the others that were studied here in the analysis of case-control data with imprecise exposure measurements. We have discussed the fact that the assumption of a normal retrospective model compared to the assumption of a prospective logistic regression model is somewhat stronger, though in an applied situation it should be reasonably easy to verify the retrospective assumptions. Also, transformations on exposure variable to approximate normality may often satisfy the retrospective model.

Given the promising results of this simple case, the next step will be to make this Bayesian approach more practical by incorporating the realities of complex data sets. Thus in the next chapter we will generalize our methodology to deal with multiple covariates, some of which could be binary.

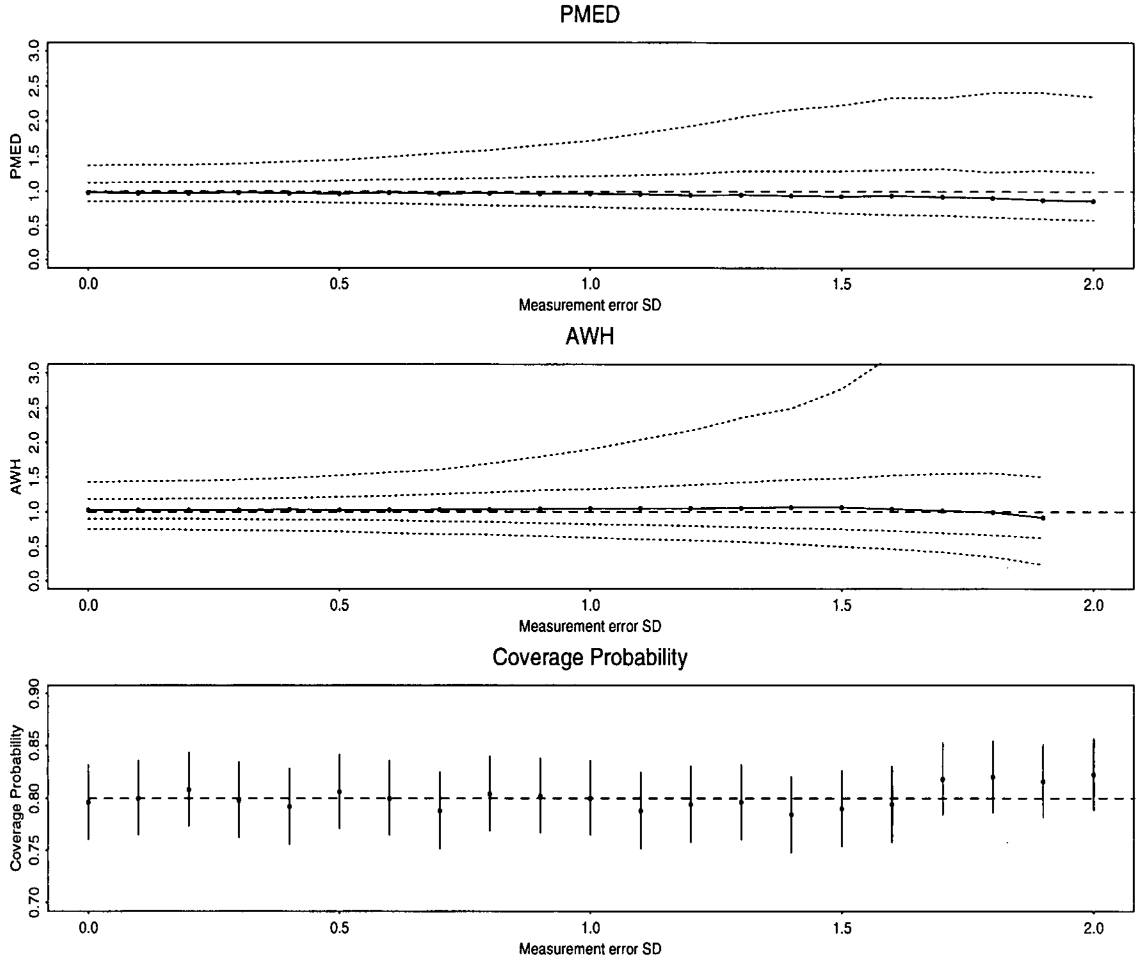


Figure 3.1: Comparison of the PMED and the AWH estimators in the known τ case. The first two panels give $(0.1, 0.3, 0.5, 0.7, 0.9)$ quantiles of the estimators as a function of $\sqrt{\tau}$. The median is displayed with the solid line. The long-dashed line indicate true value of β. The AWH estimator is only considered when τ is sufficiently small that $P(\hat{\lambda} < 0)$ does not exceed 0.05. The third panel gives empirical coverage probabilities of 80% credible intervals, with error bars corresponding to plus and minus two standard errors.

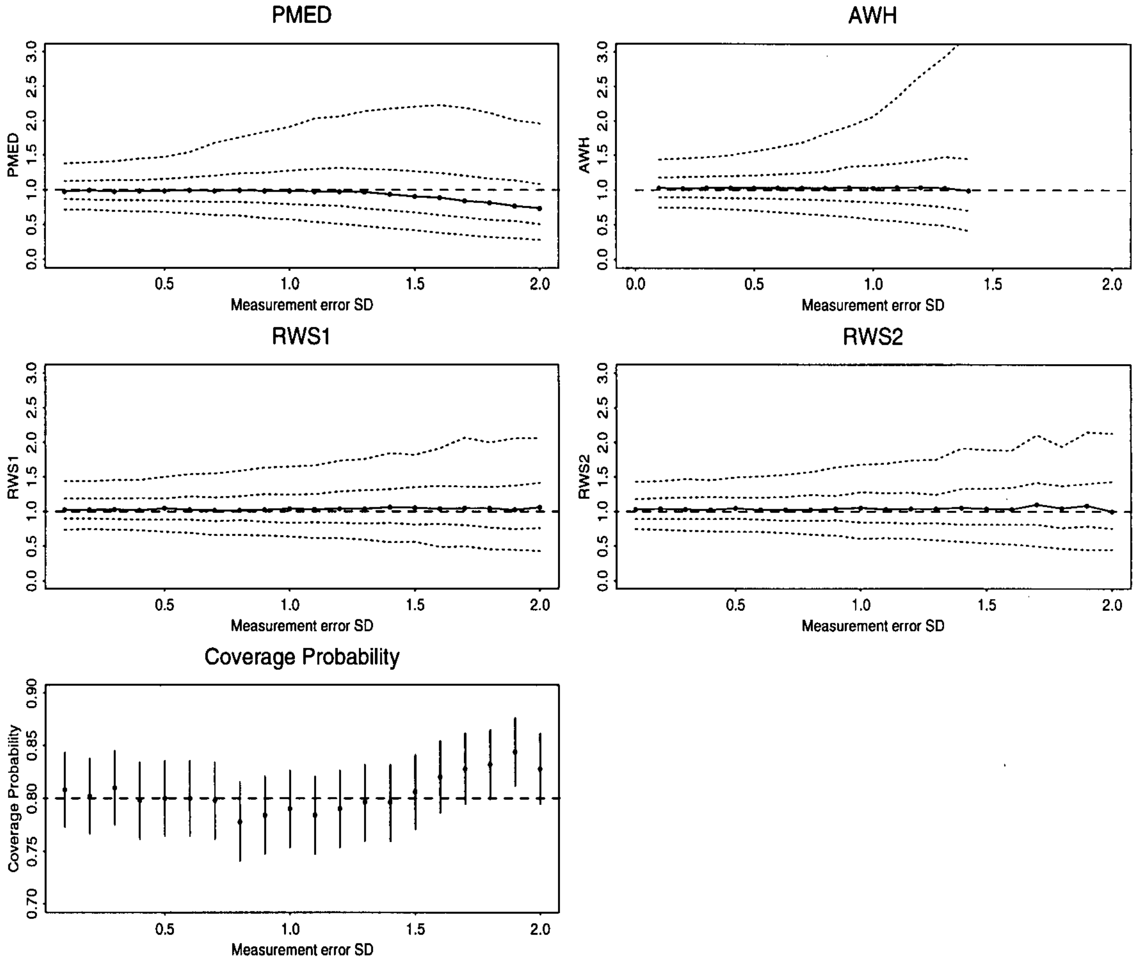


Figure 3.2: Comparison of the PMED, AWH, RWS1 and RWS2 estimators in the unknown τ case. The format is as per Figure 3.1

Chapter 4

Multivariate Exposure

The previous chapter established the validity of our work in a simple setting. The next logical step is to consider the more realistic situation in which we would be faced with multiple covariates. Developing this multivariate methodology will undoubtedly mean an increase in the level of complexity of the algebraic manipulations, the distribution identification and the sampling procedures from these distributions.

The present chapter will illustrate the development of this multivariate approach, first with the general multivariate exposure methodology, then with the introduction in the model of binary covariates, and finally with the development of a sampling method for covariance matrices with partially fixed diagonal elements.

4.1 Methodology

The initial setting will not be much different from the univariate exposure one. We are still dealing with case-control data for which the cases and the controls can be regarded as independent samples. The difference is that we are now faced with covariable matrices; $\{x_{0ij}\}_{i=1,j=1}^{n_0,d}$ from $X|D=0$ (the controls), and $\{x_{1ij}\}_{i=1,j=1}^{n_1,d}$ from $X|D=1$ (the cases), where d is the dimension of the exposure covariate. Our inferential objective remains the same, that is, to learn about the prospective relationship $D|X$. But again, to conduct a Bayesian analysis, we will need a likelihood based on the sampled distribution $X|D$. Therefore, assuming all the covariables are continuous, we consider once more the normal discriminant analysis model:

$$X|D=0 \sim N(\mu_0,\Sigma_0), \qquad X|D=1 \sim N(\mu_1,\Sigma_1). \tag{4.1}$$

which still 'implies' the prospective model of the form

$$logit\{P(D=1|X=x)\} \;=\; \alpha+\beta^T x + x^T C x, \tag{4.2}$$

where $\beta=\Sigma_1^{-1}\mu_1-\Sigma_0^{-1}\mu_0$.

Up until now, this is exactly as in chapter 3 with the exception that we do not restrict X to be unidimentional. Let's assume we have a common covariance matrix for the cases and the controls, $\Sigma=\Sigma_0=\Sigma_1$, (4.1) becomes

$$X|D=0 \sim N(\mu_0,\Sigma), \qquad X|D=1 \sim N(\mu_1,\Sigma). \tag{4.3}$$

Leading us once more to a standard logistic regression for the prospective relationship

$$logit\{P(D=1|X=x)\} \;=\; \alpha+\beta x, \tag{4.4}$$

where this time the parameter of interest $\beta = (\mu_1 - \mu_0)\Sigma^{-1}$.

The measurement error arises in a similar fashion. It is still assumed to be nondifferential and follows the same model

$$X^*|X \sim N(X, \tau) \tag{4.5}$$

where τ is now a $d \times d$ covariance matrix. Consequently, the observed data arise from

$$X^*|D = 0 \sim N(\mu_0, \tau + \Sigma), \qquad X^*|D = 1 \sim N(\mu_1, \tau + \Sigma). \tag{4.6}$$

We assume the error covariance matrix τ to be known from an external validation of the measurement process. As in the univariate case we base our inference on the noninformative prior

$$\pi(\mu_0, \mu_1, \Sigma) \propto |(\tau + \Sigma)|^{-(d+1)/2}. \tag{4.7}$$

This is a standard noninformative prior for variance component models. Our prior in (4.7) and our likelihood obtained from (4.6) yield the following posterior distribution:

$$\begin{aligned}
\pi(\mu_0, \mu_1, \Sigma|x^*) \quad \propto \quad & |(\Sigma + \tau)|^{-(d+1)/2} \\
& \times \prod_{i=1}^{n_0} |(\Sigma + \tau)|^{-1/2} \exp\left(-\frac{1}{2}(x^*_{i0} - \mu_0)'(\Sigma + \tau)^{-1}(x^*_{i0} - \mu_0)\right) \\
& \times \prod_{i=1}^{n_1} |(\Sigma + \tau)|^{-1/2} \exp\left(-\frac{1}{2}(x^*_{i1} - \mu_1)'(\Sigma + \tau)^{-1}(x^*_{i1} - \mu_1)\right) \\
\propto \quad & |(\Sigma + \tau)|^{-(n_0+n_1+d+1)/2} \\
& \times \exp\left(-\frac{1}{2}\sum_{i=1}^{n_0}(x^*_{i0} - \mu_0)'(\Sigma + \tau)^{-1}(x^*_{i0} - \mu_0)\right) \\
& \times \exp\left(-\frac{1}{2}\sum_{i=1}^{n_1}(x^*_{i1} - \mu_1)'(\Sigma + \tau)^{-1}(x^*_{i1} - \mu_1)\right)
\end{aligned}$$

$$
\begin{aligned}
&\propto \; |(\Sigma+\tau)|^{-(n_0+n_1+d+1)/2} \\
&\quad \times \exp\{-(1/2)\ \text{trace}((\Sigma+\tau)^{-1}S_0)\} \\
&\quad \times \exp\{-(1/2)\ \text{trace}((\Sigma+\tau)^{-1}S_1)\} \qquad (4.8)
\end{aligned}
$$

where for $c = 0$ or 1 (controls or cases)

$$
\begin{aligned}
S_c \;=\; & \sum_{i=1}^{n_c} \begin{pmatrix} (\mu_c^{(1)} - \bar{x}_c^{*(1)})^2 & \cdots & (\mu_c^{(1)} - \bar{x}_c^{*(1)})(\mu_c^{(d)} - \bar{x}_c^{*(d)}) \\ \vdots & \ddots & \vdots \\ (\mu_c^{(1)} - \bar{x}_c^{*(1)})(\mu_c^{(d)} - \bar{x}_c^{*(d)}) & \cdots & (\mu_c^{(d)} - \bar{x}_c^{*(d)})^2 \end{pmatrix} \\
& + \sum_{i=1}^{n_c} \begin{pmatrix} (x_{ic}^{*(1)} - \bar{x}_c^{*(1)})^2 & \cdots & (x_{ic}^{*(1)} - \bar{x}_c^{*(1)})(x_{ic}^{*(d)} - \bar{x}_c^{*(d)}) \\ \vdots & \ddots & \vdots \\ (x_{ic}^{*(1)} - \bar{x}_c^{*(1)})(x_{ic}^{*(d)} - \bar{x}_c^{*(d)}) & \cdots & (x_{ic}^{*(d)} - \bar{x}_c^{*(d)})^2 \end{pmatrix},
\end{aligned}
$$

where $\bar{x}_c^{*(j)}$ is a scalar representing the mean of the realizations of the jth covariate. We can integrate out μ_0 and μ_1 from (4.8) which gives us the marginal posterior distribution of the variance components

$$
\begin{aligned}
\pi((\Sigma+\tau)|x^*) \;\propto\; & |(\Sigma+\tau)|^{-(n_0+n_1+d+1)/2} \\
& \times \exp\{-(1/2)\ \text{trace}((\Sigma+\tau)^{-1}S)\} \qquad (4.9)
\end{aligned}
$$

where

$$
\begin{aligned}
S \;=\; & \sum_{c=0}^{1}\sum_{i=1}^{n_c}(x_{ic}^* - \bar{x}_c^*)(x_{ic}^* - \bar{x}_c^*)' \qquad (4.10) \\
=\; & \sum_{c=0}^{1}\sum_{i=1}^{n_c} \begin{pmatrix} (x_{ic}^{*(1)} - \bar{x}_c^{*(1)})^2 & \cdots & (x_{ic}^{*(1)} - \bar{x}_c^{*(1)})(x_{ic}^{*(d)} - \bar{x}_c^{*(d)}) \\ \vdots & \ddots & \vdots \\ (x_{ic}^{*(1)} - \bar{x}_c^{*(1)})(x_{ic}^{*(d)} - \bar{x}_c^{*(d)}) & \cdots & (x_{ic}^{*(d)} - \bar{x}_c^{*(d)})^2 \end{pmatrix}.
\end{aligned}
$$

So we have

$$(\Sigma + \tau)|X^* \equiv G|[(G - \tau) \text{ positive definite}], \tag{4.11}$$

where G has an inverse Wishart distribution with $n_0 + n_1$ degrees of freedom and scale matrix S. Sampling from $G|[(G-\tau)$ positive definite] can be implemented by repeatedly sampling G and subtracting the known error covariance matrix from it until we get one that is positive definite.

The posterior conditional distribution for μ_0 and μ_1 once Σ has been sampled is easily obtained from (4.8), both follow multivariate normals of dimension d

$$\begin{aligned} \mu_0|\Sigma, \tau, x^* &\sim N(\bar{x}_0^*, n_0^{-1}(\Sigma + \tau)), \\ \mu_1|\Sigma, \tau, x^* &\sim N(\bar{x}_1^*, n_1^{-1}(\Sigma + \tau)). \end{aligned} \tag{4.12}$$

Using (4.11) to sample Σ and (4.12) to get μ_0 and μ_1 gives a reasonably simple algorithm for exact posterior simulation in a variance components model. From this algorithm we can get a sample of our d-dimensional parameter of interest $\beta = (\mu_1 - \mu_0)\Sigma^{-1}$. This multivariate exposure setting will also call on our PMED estimator, in this case the sample of β's will be split in d subsamples for each exposure variable and a median will be computed for each of these subsamples.

4.2 Binary Covariates

Our approach is now applicable to multidimensional exposure variable, but an important assumption that we have made is that all variables are continuous.

However, in applied situations it is common to be faced with binary exposure variables, for which individuals are classified to be either exposed or unexposed. Even more common is a combination of both continuous and binary exposures. For this reason, it should be a priority for us to integrate this type of variable in our model.

Since our methodology has so far been developed for continuous variables, we will try to find a continuous representation of these binary covariates. A method that could be used to obtain such a representation is described in Albert and Chib (1993), where they regress a unidimensional binary response variable on a set of covariates.

They suppose N independent binary random unidimensional variables $Y_1, \cdots, Y_N$ are observed, where Y_i has a Bernouilli distribution with probability of success p_i. These p_i are related to a set of covariates that may be continuous or discrete. They define a binary regression model as

$$p_i = H(x_i^T \beta), i = 1, \cdots, N,$$

where β is a $k \times 1$ vector of unknown parameters, $x_i^T = (x_{i1}, \cdots, x_{ik})$ is a vector of known covariates, and $H(\ )$ is a known function linking the probabilities p_i with the linear structure of β. Taking H to be the standard normal cdf $\Phi(\ )$ yields what is called the probit model. In order to get this continuous representation they introduce N latent variables $Z_1, \cdots, Z_N$, where these Z_i are independent $N(x_i^T \beta, 1)$, and define $Y_i = 1$ if $Z_i > 0$ and $Y_i = 0$ otherwise. It can easily be shown that the Y_i are independent Bernoulli random variables with $p_i = P(Y_i = 1) = \Phi(x_i^T \beta)$.

The joint posterior density of β and $Z = (Z_1, \cdots, Z_N)$ given the data $y = (y_1, \cdots, y_N)$ is given by

$$\begin{aligned}\pi(\beta, Z|y) &= C\pi(\beta)\prod_{i=1}^{N}\{I(Z_i > 0)I(y_i = 1) + I(Z_i \leq 0)I(y_i = 0)\} \\ &\quad \times\phi(Z_i; x_i^T\beta, 1),\end{aligned}$$

where $\pi(\beta)$ is a prior on β, $\phi(\ ; \mu, \sigma^2)$ is the $N(\mu, \sigma^2)$ pdf, $I(X \in A)$ is an indicator function that is equal to 1 if the random variable X is contained in the set A, and here and henceforth C is a proportionality constant. It is very difficult to sample directly from this distribution. But computation of the marginal posterior distribution of β using the Gibbs sampling algorithm requires only the posterior distribution of β conditional on Z and the posterior distribution of Z conditional on β, and fortunately these full conditional distribution are of standard forms. The posterior density of β conditioned on Z is given by

$$\pi(\beta|y, Z) = C\pi(\beta)\prod_{i=1}^{N}\phi(Z_i; x_i^T\beta, 1), \tag{4.13}$$

and the posterior density of Z conditioned on β results in the random variables $Z_1, \cdots, Z_N$ being independent with

$$Z_i = \begin{cases} Z_i^*|Z_i^* > 0 & \text{, if } y_i = 1 \\ Z_i^*|Z_i^* < 0 & \text{, if } y_i = 0 \end{cases} \quad \text{where } Z_i^* \sim N(x_i^T\beta, 1). \tag{4.14}$$

4.2.1 The Gibbs Sampler

We mentioned above the Gibbs sampling algorithm. Here is a brief review of this useful technique. Suppose one is interested in simulating from the posterior distribution of θ partitioned into the vector components $\theta = (\theta_1, \cdots, \theta_p)$.

Although it may be difficult to sample from the joint posterior, suppose that it is easy to sample from the conditional distributions $\pi(\theta_k|\{\theta_j, j \neq k\})$. To implement the Gibbs sampler, one starts with initial guesses of the θ_i, say $\theta_1^{(0)}, \cdots, \theta_p^{(0)}$ and then simulates in turn

$$\begin{aligned} &\theta_1^{(1)} \quad \text{from} \quad \pi(\theta_1|\{\theta_j^{(0)}, j \neq 1\}) \\ &\theta_2^{(1)} \quad \text{from} \quad \pi(\theta_2|\theta_1^{(1)}, \{\theta_j^{(0)}, j > 2\}) \\ &\vdots \\ &\theta_p^{(1)} \quad \text{from} \quad \pi(\theta_p|\{\theta_j^{(1)}, j < p\}). \end{aligned} \tag{4.15}$$

The cycle in (4.15) is iterated t times, generating the sample $\theta^{(t)} = (\theta_1^{(t)}, \cdots, \theta_p^{(t)})$. As t approaches infinity, it can be shown that the joint distribution of $\theta^{(t)}$ approaches the joint distribution of θ, in practice this convergence is usually quite fast. So for sufficiently large t, say t^*, $\theta^{(t^*)}$ can be regarded as one simulated value from the posterior distribution of θ. Repeating this process m times yields the sample $\{(\theta_{1j}^{(t^*)}, \theta_{2j}^{(t^*)}, \cdots, \theta_{pj}^{(t^*)}), j = 1, \cdots, m\}$, which can be used for statistical inference. In practice, instead of restarting the algorithm once the convergence is obtained we will just keep for our posterior sample the $\theta_{ij}^{(t)}$ for $t = t^*, \cdots, t^* + m$. This concludes the review of the Gibbs sampler.

Returning to the method suggested by Albert and Chib, now given a previous value of β, one cycle of the Gibbs algorithm would produce Z and β from the distributions (4.14) and (4.13).

Advances developed from this approach are presented in Chib and Greenberg (1998), in which they construct a multivariate probit model; the binary response variable is allowed to be multidimentional. This generalization to multidimensional response makes this technique applicable to our own analysis.

4.2.2 Application

We are faced with b-dimensional binary data $Y_{ijk} = 0$ or 1, $i = 1, \cdots, n_j$, where j is either 0 or 1 for controls or cases and $k = 1, \cdots, b$, where b is the dimension of the binary covariate we want to include in our model. We must introduce the latent variables $Z_{1jk}, \cdots, Z_{n_j jk}$, where the Z_{ij} are independent multinormal $N(\eta_j, \Sigma_{22})$. We are using η and not $x_i^T \beta$ as the mean because our application is not in a regression context. The reason for the Σ_{22} notation of the covariance matrix will be explained later on. It is necessary for identifiability reasons to fix as 1's the diagonal elements of this covariance matrix. This way the variance of Z_{ij} is 1 as in the univariate case from the Albert and Chib method presented earlier, but the covariance parameters are free. In their multivariate development Chib and Greenberg simply refer to this as the necessity of Σ_{22} to be in correlation form.

Then we define $Y_{ijk} = 1$ if $Z_{ijk} > 0$ and $Y_{ijk} = 0$ otherwise. The Y_{ij} are independent vectors of 0's and 1's with $p_{jk} = P(Y_{ijk} = 1) = \Phi(\eta_{jk})$, where $\Phi(\ )$ is the standard normal cdf, which gives us the model known as the probit link model. Since the y_i are observed we can estimate the p_{jk} and using the probit link obtain estimates for the η_{jk} as starting values for our Gibbs sampler. As a starting value for Σ_{22} we can use 1's for the variance (on the diagonal) and 0's for the covariance elements (off the diagonal).

So the next step is to go ahead and implement our Gibbs sampler. First by sampling Z from its posterior distribution conditioned on η, Σ_{22} and Y, then by sampling Σ_{22} conditioned on η and Z, finally by sampling η conditioned on Σ_{22} and Z. Repeating this enough times will at some point be equivalent to sampling from the joint posterior.

In order to proceed with this sampling scheme we have to determine these conditional densities. The random variables $Z_{1j}, \cdots, Z_{n_j j}$ will be independent with b-dimensional "truncated multinormal" distributions

$$Z_{ij}|y, \eta, \Sigma_{22} \sim N(\eta_j, \Sigma_{22}) \qquad (4.16)$$

$$\text{with } Z_{ijk} > 0 \text{ if } y_{ijk} = 1,$$

$$\text{and } Z_{ijk} < 0 \text{ if } y_{ijk} = 0.$$

This way the initial binary covariates are represented by continuous variables following a multinormal distribution. It is therefore possible to include binary exposures in our model, and their connection to the normal distribution will make covariance estimation feasible. The η_j, which are vectors of size b, will also follow a multinormal distribution just like μ_0 and μ_1 in (4.12). In fact, when we will juxtapose the Z's to the observed continuous X^*'s to deal with both the continuous and binary variables simultaneously, the η_{jk} will simply be the last b components of the μ_j vectors. So we can sample η_j from the following

$$\eta_j|Z_j, \Sigma_{22} \sim N(\bar{Z}_j, \Sigma_{22} n_j^{-1}) \qquad (4.17)$$

Finally we need to determine the conditional distribution of Σ_{22}. We know from (4.11) that its distribution will somehow be related to an inverse Wishart, but we don't know yet how the correlation form condition (1's on the diagonal) will influence the actual distribution from which our sample has to be drawn.

The next section will concentrate on determining this distribution and will present an approach for sampling such a covariance matrix.

4.3 Covariance Matrix With Partially Fixed Diagonal

The addition of these binary covariates in a continuous form will not change model (4.6) from which the data arise. The Z mentioned in the previous section will just have to be juxtaposed to X^*, forming a new X^* consisting of c continuous exposure variables and b binary exposure variables in a continuous form, where $c + b = d$ the total dimension of the exposure. For later use, the binary covariates will always be placed at the end of the exposure vector; so we will have first the c continuous X^* and then the b binary ones. This will yield a covariance matrix Σ splitted in the following way

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \tag{4.18}$$

where Σ_{11} is $c \times c$, Σ_{12} is $c \times b$, Σ_{21} is of course the transpose of Σ_{12} and Σ_{22} is $b \times b$.

We also make the important assumption that there is no misclassification for the binary covariates, therefore the errors in variables will only be occurring for continuous variables. Also, with no influence on our previous work, some of the continuous exposures could be measured without error. This disposition of the continuous variables followed by the binary ones combined with the no-misclassifications assumption means that the lower diagonal of the covariance matrix Σ (the diagonal of Σ_{22}) will consist in a series of b 1's, as will the lower diagonal of $(\Sigma + \tau)$, since the lower diagonal of τ will consist of a series of b 0's.

The major change comes from the new prior we will have to use. In (4.7) we suggested Jeffrey's noninformative prior, but now part of the covariance matrix is known. Thus we try the prior

$$\pi(\mu_0, \mu_1, \Sigma) \propto |(\Sigma + \tau)|^{-(d+1)/2} I(\text{diag}(\Sigma_{22} + \tau_{22}) = 1), \tag{4.19}$$

where $I(\ )$ is an indicator function equal to 1 if the diagonal elements of Σ_{22} are 1's and equal to 0 otherwise. Do note that τ_{22} is a matrix of 0's since we have assumed no misclassification for the binary covariates. The addition of this new prior will operate a slight change in equation (4.9), the marginal posterior of the variance components now becomes

$$\begin{aligned} \pi((\Sigma + \tau)|X^*) \quad \propto \quad & |(\Sigma + \tau)|^{-(n_0+n_1+d+1)/2} \\ & \times \exp\{-(1/2) \text{ trace}((\Sigma + \tau)^{-1} S)\} \\ & \times I(\text{diag}(\Sigma_{22}) = 1), \end{aligned} \tag{4.20}$$

where S is as defined in (4.10), thereby transforming (4.11) to

$$(\Sigma + \tau)|X^* \equiv G|[(G - \tau) \text{ positive definite and diag}(G_{22}) = 1)], \tag{4.21}$$

Where G still has an inverse Wishart distribution with $n_0 + n_1$ degrees of freedom and scale matrix S.

Sampling $(\Sigma + \tau)$ from (4.21) is not a trivial task anymore. We need to introduce a matrix decomposition presented in Le and Zidek (1992). Take a covariance matrix Σ which follows an inverse Wishart distribution with m degrees of freedom and scale matrix Ψ, which can be partitioned in the following way

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where Σ_{11} is $u \times u$ and Σ_{22} is $g \times g$. Through the Bartlett decomposition Σ can be written as

$$\Sigma = \begin{pmatrix} \Sigma_{1|2} + \psi \Sigma_{22} \psi^T & \psi \Sigma_{22} \\ \Sigma_{22} \psi^T & \Sigma_{22} \end{pmatrix}, \tag{4.22}$$

where $\Sigma_{1|2} \equiv \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, a $u \times u$ matrix and $\psi = \Sigma_{12}\Sigma_{22}^{-1}$, a $u \times g$ matrix. The inverse Wishart distribution of Σ can be equivalently presented in terms of the new parameters $(\Sigma_{22}, \Sigma_{1|2}, \psi)$ as

$$\begin{aligned} &\Sigma_{22} | \Psi, m \sim IW(\Psi_{22}, m - u) \\ &\Sigma_{1|2} | \Psi, m \sim IW(\Psi_{1|2}, m) \\ &\psi | \Sigma_{1|2}, \Psi \sim N(\Psi_{12}\Psi_{22}^{-1}, \Sigma_{1|2} \otimes \Psi_{22}^{-1}), \end{aligned} \tag{4.23}$$

where $IW(A, B)$ stands for inverse Wishart with scale matrix A and degrees of freedom B, Σ_{22} is independent of $(\Sigma_{1|2}, \psi)$, and $\otimes$ symbolizes the Kroneker product.

This decomposition can obviously be applied to our own covariance matrix $\Sigma + \tau$ which has to be sampled from (4.21), and the partitioning will be the one suggested earlier in (4.18)

$$\Sigma + \tau = \begin{pmatrix} \Sigma_{11} + \tau_{11} & \Sigma_{12} + \tau_{12} \\ \Sigma_{21} + \tau_{21} & \Sigma_{22} + \tau_{22} \end{pmatrix} = \begin{pmatrix} (\Sigma + \tau)_{11} & (\Sigma + \tau)_{12} \\ (\Sigma + \tau)_{21} & (\Sigma + \tau)_{22} \end{pmatrix},$$

where $\Sigma_{22} + \tau_{22} = \Sigma_{22}$ since we assume there is no misclassification for the binary covariates. Therefore u will be equal to our number of continuous variables c, g equal to the number of binary variables b, m equal to the degrees of freedom $n_0 + n_1$ of our covariance matrix's inverse Wishart distribution and Ψ equal to the scale matrix S of this same inverse Wishart distribution. Based on (4.22) $\Sigma + \tau$ can be reparametrized, leading to

$$\Sigma + \tau = \begin{pmatrix} (\Sigma + \tau)_{1|2} + \psi\Sigma_{22}\psi^T & \psi\Sigma_{22} \\ \Sigma_{22}\psi^T & \Sigma_{22} \end{pmatrix}.$$

However, $\Sigma + \tau$ does not follow exactly an inverse Wishart distribution, since it has the extra conditions that it must be positive definite once τ is subtracted from it and the elements of the lower diagonal (diag(Σ_{22})) have to be all 1's. Work has to be done to see how these conditions influence the distributions suggested in (4.23).

First, the positive definite condition can be verified once the sampling is done, so this condition will not affect the distributions we will be sampling from. The diagonal elements set to be 1 will however make the sampling procedure more complex. For later use let us introduce the following notation

:

$$IW^*(Scale, df) \equiv G|\text{diag}(G) = 1, \text{ where } G \sim IW(Scale, df).$$

Given the independence between Σ_{22} and $((\Sigma+\tau)_{1|2}, \psi)$, sampling $(\Sigma+\tau)_{1|2}$ and ψ will always be trivial once the non-diagonal elements of Σ_{22} are identified. The case where we have only one binary variable is quite simple. Σ_{22} would be a scalar set to be 1, independent of $(\Sigma+\tau)_{1|2}$ and ψ which could be sampled from the distributions given in (4.23) without being affected by $\Sigma_{22} = 1$. Once these are sampled, $\Sigma + \tau$ can be computed from (4.22), then τ subtracted so that the positive definite condition can be verified. If the condition is not met we need to resample these parameters.

The process becomes more complex when we are faced with more then one binary covariate, in which case $b > 1$ and Σ_{22} is not a simple scalar equal to 1, it has off-diagonal elements. In such a case an iterative process will have to be used. The initial step will be to perform the decomposition as it was done in the "1 binary covariate" case. This time the decomposition gives us a matrix Σ_{22} of dimension $b > 1$. From (4.23) we know that Σ_{22} follows an inverse Wishart with $n_0 + n_1 - c$ degrees of freedom and scale matrix S_{22}, and its diagonal elements are all 1's. To keep the notation as simple as we can we will now take $\Sigma^* = \Sigma_{22}$ and $S^* = S_{22}$. Given Σ^*'s distribution, the next step is to perform the same decomposition on Σ^*, in which case the split has to be done in the following manner

$$\Sigma^* = \begin{pmatrix} \Sigma^*_{11} & \Sigma^*_{12} \\ \Sigma^*_{21} & \Sigma^*_{22} \end{pmatrix},$$

so that $\Sigma^*_{11} = 1$ and Σ^*_{22} is a $(b-1) \times (b-1)$ matrix still with 1's on its diagonal. From (4.22) Σ_{22} (now Σ^*) can be written differently as

$$\Sigma^* = \begin{pmatrix} \Sigma^*_{1|2} + \phi\Sigma^*_{22}\phi^T & \phi\Sigma^*_{22} \\ \Sigma^*_{22}\phi^T & \Sigma^*_{22} \end{pmatrix},$$

where $\Sigma^*_{1|2} \equiv \Sigma^*_{11} - \Sigma^*_{12}\Sigma^{*-1}_{22}\Sigma^*_{21}$, but $\Sigma^*_{1|2} + \phi\Sigma^*_{22}\phi^T = 1$, and $\phi = \Sigma^*_{12}\Sigma^{*-1}_{22}$, a vector of dimension $b-1$. We have to keep using the decomposition until the bottom right corner, Σ^*_{22} for the above decomposition, is a scalar set to be 1. So if $b = 2$, two iteration are needed; the decomposition has to be performed twice in order to get $\Sigma^*_{22} = 1$, once to split the continuous and binary variables and once more to obtain the scalar form of the bottom right corner.

We saw in (4.23) that the decomposition specifies the distributions of the different parameters obtained from the decomposition. These distributions will be slightly different from the second iteration and onward.

4.3.1 A Simple Case

We will first deal with a fairly simple situation, that is the case where the number of binary covariates b is 2. In such a case the initial decomposition would lead us to

$$\Sigma + \tau = \begin{pmatrix} (\Sigma + \tau)_{1|2} + \psi\Sigma_{22}\psi^T & \psi\Sigma_{22} \\ \Sigma_{22}\psi^T & \Sigma_{22} \end{pmatrix}, \tag{4.24}$$

where $(\Sigma + \tau)_{1|2} + \psi\Sigma_{22}\psi^T$ is a $c \times c$ matrix (remember that c is the number of continuous variables), and Σ_{22} is a 2×2 matrix with 1's on its diagonal.

From (4.23) we know that

$$
\begin{aligned}
&\Sigma_{22}|S, n_0, n_1 \sim IW^*(S_{22}, n_0 + n_1 - c) \\
&(\Sigma + \tau)_{1|2}|S, n_0, n_1 \sim IW(S_{1|2}, n_0 + n_1) \qquad (4.25) \\
&\psi|(\Sigma + \tau)_{1|2}, S \sim N(S_{12}S_{22}^{-1}, (\Sigma + \tau)_{1|2} \otimes S_{22}^{-1}),
\end{aligned}
$$

where Σ_{22} is independent of $((\Sigma + \tau)_{1|2}, \psi)$.

For simplicity we will again use $\Sigma^* = \Sigma_{22}$ and $S^* = S_{22}$. We now perform the decomposition on Σ^*, this time resulting in

$$
\Sigma^* = \begin{pmatrix} \Sigma^*_{1|2} + \phi\Sigma^*_{22}\phi^T & \phi\Sigma^*_{22} \\ \Sigma^*_{22}\phi^T & \Sigma^*_{22} \end{pmatrix}, \qquad (4.26)
$$

where both Σ^*_{22} and $\Sigma^*_{1|2} + \phi\Sigma^*_{22}\phi^T$ are equal to 1. Once again we use (4.23) to obtain

$$
\begin{aligned}
&\Sigma^*_{22} = 1 \\
&\Sigma^*_{1|2}|S^*, n_0, n_1, c \sim IW(S^*_{1|2}, n_0 + n_1 - c) \qquad (4.27) \\
&\phi|\Sigma^*_{1|2}, S^* \sim N(S^*_{12}S^{*-1}_{22}, \Sigma^*_{1|2} \otimes S^{*-1}_{22}),
\end{aligned}
$$

the last two distributions are subject to the condition, $\Sigma^*_{1|2} + \phi\Sigma^*_{22}\phi^T = 1$, which can be reduced to $\Sigma^*_{1|2} + \phi^2 = 1$, making the condition independent of Σ^*_{22} which will not be the case for situations with more than 2 binary covariates. $\Sigma^*_{1|2}$'s distribution is also simplified, given the fact that it is a scalar the inverse Wishart distribution now becomes an inverse gamma, and its shape and scale parameters respectively are $(n_0 + n_1 - c)/2$ and $S^*_{1|2}/2$.

In order to find the distribution for $\Sigma^*_{1|2}$ and ϕ under this condition we use the following reparametrization: $Z_1 = \Sigma^*_{1|2} + \phi^2$ and $Z_2 = \phi$. The Jacobian for this transformation is $J = 1$.

We already have the joint distribution for $\Sigma^*_{1|2}$ and ϕ

$$f_{\Sigma^*_{1|2},\phi}(\Sigma^*_{1|2}, \phi) = f(\phi|\Sigma^*_{1|2})f(\Sigma^*_{1|2}),$$

from which we can derive the joint distribution of our new variables Z_1 and Z_2

$$f_{Z_1,Z_2}(z_1, z_2) = f_{\Sigma^*_{1|2},\phi}(z_1 - z_2^2, z_2)J.$$

We now want to obtain the distribution of Z_2 given $Z_1 = 1$, we can not get this exactly but do work out a function proportional to this conditional distribution:

$$\begin{aligned} f_{Z_2|Z_1=1}(z_2|z_1 = 1) &= f_{Z_1,Z_2}(z_1, z_2)/f_{Z_1}(z_1 = 1) \qquad (4.28) \\ &\propto f_{Z_1=1,Z_2}(1, z_2) \\ &\propto (1 - z_2^2)^{-(m+3)/2} \\ &\quad \times \exp\left(-\frac{1}{2}\frac{\left(z_2 - \frac{S^*_{12}}{m}\right)^2 + \frac{S^*_{11}}{m} + \frac{S^*_{22}}{m-1} - \left(\frac{S^*_{12}}{m}\right)^2}{(1 - z_2^2)/m}\right), \end{aligned}$$

where $m = n_0 + n_1 - c$. This is not a standard distribution, but still we want to sample from it. In order to do so we use the rejection sampling technique as presented in Ross(1997), to sample Z from a certain density proportional to $f(\,)$:

1. Generate Y from a density proportional to $g(\,)$, where $g(\,)$ is a function that bounds $f(\,)$.

2. Generate U from a standard uniform distribution.

3. If $U \leq f(Y)/g(Y)$, take $Z = Y$, else go back to 1.

Therefore we need to find a bound for (4.28). We proceed in doing so by splitting (4.28) into two parts

$$A = (1 - z_2^2)^{-(m+3)/2} \exp\left(-\frac{1}{2}\frac{S_{11}^*/m + S_{22}^*/m - 1 - (S_{12}^*/m)^2}{(1 - z_2^2)/m}\right),$$
$$B = \exp\left(-\frac{1}{2}\frac{(z_2 - S_{12}^*/m)^2}{(1 - z_2^2)/m}\right).$$

To obtain an upper bound for A we note that

$$\text{argmax}\left(x^{-a}\exp(b/x)\right) = b/a, \tag{4.29}$$

the resulting bound is a constant obtained by plugging in A

$$(1 - z_2^2) = \frac{m}{m+3}(S_{11}^*/m + S_{22}^*/m - 1 - (S_{12}^*/m)^2).$$

The bound for B is

$$\exp\left(-\frac{1}{2}m(z_2 - S_{12}^*/m)^2\right),$$

which is proportional to a normal density with mean S_{12}^*/m and variance $1/m$, making it easy to sample from the bounding function.

Once Z_2 is sampled, this means we have an estimate for ϕ, therefore we can complete Σ^* or Σ_{22} in (4.26). Then use the distributions given in (4.25) to sample the necessary parameters to complete $\Sigma + \tau$ in (4.24). Following carefully all of the previous steps we are able to obtain an estimate of the covariance matrix with parts of it being fixed, then all that needs to be done is to subtract τ and verify the positive definite condition.

4.3.2 More than Two Binary Covariates

We must now be able to deal with a greater number of binary covariates, this will increase the number of iterations and will again slightly alter the distributions. Let's assume we still have c continuous variables and b binary ones, but this time b is greater than 2.

We proceed as in the previous simple case and operate the initial decomposition and obtain a matrix just like in (4.24) except Σ_{22} is now $b \times b$, $b > 2$. The distributions in (4.25) still stand, therefore we can go ahead with the second decomposition as in the simple case and obtain Σ^* just like in (4.26) with the exception that even with this second iteration Σ^*_{22} is not yet a scalar equal to 1, it is now a $(b-1) \times (b-1)$ matrix. There is now a slight transformation to the distributions we had obtained in (4.27)

$$\begin{aligned} &\Sigma^*_{22}|S^*, m \sim IW^*(S^*_{22}, m-1) \\ &\Sigma^*_{1|2}|S^*, m \sim IW(S^*_{1|2}, m) \\ &\phi|\Sigma^*_{1|2}, S^* \sim N(S^*_{12}S^{*-1}_{22}, \Sigma^*_{1|2} \otimes S^{*-1}_{22}), \end{aligned} \tag{4.30}$$

where $m = n_0 + n_1 - c$, Σ^*_{22} is independent of $(\Sigma^*_{1|2}, \phi)$ and these are this still subject to the condition $\Sigma^*_{1|2} + \phi\Sigma^*_{22}\phi^T = 1$. Since this time Σ^*_{22} is not yet a scalar equal to 1, the condition $\Sigma^*_{1|2} + \phi\Sigma^*_{22}\phi^T = 1$ can not be simplified and therefore depends on Σ^*_{22} which will once more influence our sampling.

We now want to sample from a density proportional to

$$f(\Sigma^*_{22})f(\Sigma^*_{1|2})f(\phi|\Sigma^*_{1|2})I(\text{diag }(\Sigma^*_{22}) = 1)I(\Sigma^*_{1|2} + \phi\Sigma^*_{22}\phi^T = 1).$$

First we will consider the distribution of $(\Sigma^*_{22}, \Sigma^*_{1|2}, \phi|\text{diag}(\Sigma^*_{22}) = 1)$ to be

$g(\Sigma^*_{22})f(\Sigma^*_{1|2})f(\phi|\Sigma^*_{1|2})$, where $g(\Sigma^*_{22})$ is not the same as $f(\Sigma^*_{22})$. Thus we want to sample from the density proportional to

$$g(\Sigma^*_{22})f(\Sigma^*_{1|2})f(\phi|\Sigma^*_{1|2})I(\Sigma^*_{1|2}+\phi\Sigma^*_{22}\phi^T=1), \tag{4.31}$$

where both $f(\Sigma^*_{1|2})$ and $f(\phi|\Sigma^*_{1|2})$ are known. We now reparametrize by introducing $Z_1=\Sigma^*_{1|2}+\phi\Sigma^*_{22}\phi^T$ and $Z_2=\phi$, once again the Jacobian for this transformation is $J=1$. From this (4.31) becomes

$$g(\Sigma^*_{22})f(z_1-z_2\Sigma^*_{22}z_2^T)f(z_2|z_1-z_2\Sigma^*_{22}z_2^T)I(z_1=1),$$

therefore we want to sample from a density proportional to

$$g(\Sigma^*_{22})f(1-z_2\Sigma^*_{22}z_2^T)f(z_2|1-z_2\Sigma^*_{22}z_2^T). \tag{4.32}$$

By putting in (4.32) the respective distribution functions and doing a little algebra we get that (4.32) is proportional to

$$\begin{aligned} g(\Sigma^*_{22})\,&\overbrace{(1-z_2\Sigma^*_{22}z_2^T)^{-(m+3)/2}\exp\left(-\frac{S^*_{1|2}}{2(1-z_2\Sigma^*_{22}z_2^T)}\right)}^{A}\\ &\times\underbrace{\exp\left(-\frac{(z_2-S^*_{12}S^{*-1}_{22})S^*_{22}(z_2-S^*_{12}S^{*-1}_{22})^T}{2(1-z_2\Sigma^*_{22}z_2^T)}\right)}_{B}. \end{aligned} \tag{4.33}$$

To sample from this distribution we can apply the rejection sampling technique, therefore we need to bound (4.33). To find an upper bound for A we use (4.29). The resulting bound is a constant obtained by plugging in A

$$1-z_2\Sigma^*_{22}z_2^T=S^*_{12}/(m+3).$$

B can easily be bounded by

$$\exp\left(-\frac{1}{2}(z_2-S^*_{12}S^{*-1}_{22})S^*_{22}(z_2-S^*_{12}S^{*-1}_{22})^T\right),$$

which is to a near constant a multinormal distribution of dimension $b-(i-1)$ with mean vector $S^*_{12}S^{*-1}_{22}$ and covariance matrix S^{*-1}_{22}, where i is the number of iterations performed to this point. This will again make it easy to sample from the bounding distribution.

The whole bound is

$$g(\Sigma^*_{22})\left(\frac{S^*_{1|2}}{m+3}\right)^{(m+3)/2}\exp\left(-\frac{m+3}{2}\right)$$
$$\times\exp\left(-\frac{1}{2}(z_2-S^*_{12}S^{*-1}_{22})S^*_{22}(z_2-S^*_{12}S^{*-1}_{22})^T\right),$$

thus in both the bound and the distribution we want to sample from in (4.33) we have $g(\Sigma^*_{22})$. These will cancel out in the third stage of the rejection sampling, avoiding the task of determining them. However we see that in (4.33) we need Σ^*_{22}, therefore we need to iterate the process and apply the decomposition on Σ^*_{22} until we have reached its bottom right corner which is 1. Once this is done we can move up one dimension at a time and use the preceding technique to sample the off-diagonal elements of the covariance matrix.

As an example, suppose the number of continuous variables c is 1, and that the number of binary covariates b is 3. We apply the decomposition once to split the continuous part of the covariance matrix from the binary one. We apply the decomposition once more resulting in a 2×2 Σ^*_{22} matrix, then apply the decomposition one last time to obtain $(\Sigma^*_{22})_{22}=1$. Once this is done we can work our way back up. Knowing $(\Sigma^*_{22})_{22}=1$ we can sample from (4.33) with $m=n_0+n_1-c-2$ using the rejection sampler. This will give us the off-diagonal elements of Σ^*_{22} which we can use again in (4.33) with this time

$m = n_0 + n_1 - c - 1$ giving us the off-diagonal elements of Σ_{22}, and now that we have Σ_{22} we can use it to sample from the last two distributions of (4.25).

It is now possible for us to run our Gibbs sampling algorithm of Section 4.2.2 since all the conditional distribution needed have been determined.

4.4 Discussion

We have seen that in the generalization of our approach to a multivariate setting it is feasible to incorporate binary variables to our model. The use of a variation of the probit regression model enables us to get a continuous representation of these binary covariates. This continuous representation is obtained by generating values from truncated normal distributions which have their variance set to 1. Then an iterative process which makes it possible to estimate a covariance matrix with a partially fixed diagonal allows us to compute estimates of our parameter of interest which is $\beta = (\mu_1 - \mu_0)\Sigma^{-1}$ the main parameter of a standard logistic regression.

Chapter 5

An Application to Real Data

The work presented in the previous chapters has resulted in an approach that should be generally applicable to actual data. One of the generalizations allowed for multivariate exposures and another made the addition of binary covariates to the analysis possible. We will now verify via the analysis of real case-control data if our methodology is suitable to applied situations and how results are influenced by our correction for error.

The present chapter will first describe the data we will be working on and expose the results obtained in a previously conducted analysis. We will also verify if the basic assumptions of our method are met and if so, will proceed with the analysis using our own approach.

5.1 Bladder Cancer Case-Control Data

The data we will be conducting our analysis on is a fraction of a dataset which was obtained for a large-scale study. The objective of this study was to identify occupational cancer risk factors. Information on occupation, smoking and alcohol consumption histories was collected by means of a self-administered questionnaire from male cancer patients aged 20 years and over ascertained from the British Columbia population-based cancer registry from 1983 to 1990.

To estimate smoking relationships for types of cancer known to be strongly associated with cigarette smoking, patients suffering from cancer types with no such associations were used as controls. So the control group consisted of all cancer types with the exception of those known to be associated with smoking. The analysis was performed by matching cases and controls on exact age and diagnosis year. The patients smoking histories were measured in pack-years (number of years of smoking 20 cigarettes a day). Odds ratios for different stratifications of these pack-years were estimated using conditional logistic regression. Based on this analysis, a statistically significant relationship was found between bladder cancer and cigarette smoking; the odds ratio would get bigger as the pack-years increased. This analysis was conducted assuming no measurement error was present. The details of this study are presented in Band et al. (to appear in the Journal of Occupational and Environmental Medicine).

In this large-scale study many types of cancer were considered, but the fraction we focus on deals with bladder cancer and contains information on the disease status (case or control), the age, the smoking history and the diagnosis year.

Here we apply the Bayesian approach to re-examine the relationship between cigarette smoking and the risk of developing bladder cancer, taking into consideration measurement error of the smoking history information.

5.1.1 A Closer Look at the Data

In light of these results we are interested in seeing if our method would lead to similar conclusions. A first step in doing so is to take a closer look at the data and see if the assumptions required by our approach are met.

Our dataset consists of 1038 cases and 7006 controls. We have, for each individual, their bladder cancer status (healthy:0, diseased:1), their age (in years), their smoking history (in pack-years) and their year of diagnosis (from 1983 to 1990). We initially get histograms for the age and the smoking history variables. The one for age has a longer and thicker left tail, but a transformation could probably approximate normality. The histogram for the smoking history, however, reveals a problem of greater consequence. There is an important concentration of 'never-smokers' or of individuals whose smoking history is 0 pack-years. In fact there are 1782 'never-smokers', of which 1658 are controls and 124 are cases, together they represent close to 20% of the whole group. Therefore, no transformations could lead us to an approximately normal distribution for the smoking history.

Since the data in the present format fail to meet the normality assumption, we need to consider a slightly different objective. We now consider 'smokers' only, or the individuals that did not have 0 pack-years for their smoking history. We examine the dose-response, that is, how the risk of bladder cancer is influenced by the amount of smoking a person has done throughout her life. Since we ignore the 'non-smokers' we are now left with 914 cases and 5348 controls. Once more we proceed to construct histograms for the age and the smoking history variables, these are shown in Figure 5.1. Based on these histograms, it seems age is quite close to normality, however smoking history will obviously require a transformation in order to approximate normality. The two bottom graphs of Figure 5.2 indicate that although it is not perfect, a logarithmic transformation of the smoking history could be considered to be approximately normal. To see if the age and the logarithmic transformation of the smoking history can both be used as continuous variables, we proceed with a logistic regression of the disease status on different stratifications of these two variables. The parameters obtained for these regressions are plotted against the increasing stratifications in Figure 5.3. We notice that for the regression on age the coefficients seem to follow a curvilinear trend, perhaps a logarithmic transformation of the age would correct for that. The coefficients for the logistic regression on the logarithmic transformation of the smoking history however increase in a relatively linear fashion, indicating that the logarithm of the pack-years can be used as a continuous variable. Given the non-linearity of the coefficients for the age, we try the logarithmic transformation of the age. We 'logistically' regress on stratifications of the logarithm of the age, a plot of the obtained parameters shown in Figure 5.4 increases in a slightly more linear fashion than for the untransformed age, indicating that it is reasonable to use the logarithm of the age as a continuous variable. We also need to verify the normality assumption for this transformation of the age. The upper panels of Figure 5.2 indicate that normality of the logarithmic transformation of the age is reasonably acceptable.

5.2 The Analysis

Based on the verification of assumptions done in the previous section we can now proceed to do an analysis of this bladder cancer case-control data using the methodology presented in the earlier chapters. The variables we use are the following: The logarithmic transformation of both the age and the smoking history, that we respectively name X_1 and X_2, and a binary variable indicating if the cancer patient was diagnosed before 1987 that we name Y_3. That last variable mainly serves as a check to see if diagnostic methods have changed between these two four years periods from 1983 to 1986 and 1987 to 1990. Using an indicator for each year from 1983 to 1990 would have given us eight binary variables, which our approach can handle, however a normal representation of these variables would not have been appropriate given that a 1 for one of them means 0 for all the others, setting the sum to 1. The normal representation which sets the sign of the normal given the binary response can deal with correlation but not so much as to fix the sum. The analysis is thus performed following the model:

$$logit\{P(D = 1|X = x)\} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 Y_3 \tag{5.1}$$

where $X = (X_1, X_2, Y_3)$ and α is not estimable from case-control data alone.

Remember that the original analysis was performed assuming no measurement error was present. However, it is important to note that the information on the smoking history of these patients was obtained from a self-administered questionnaire. For this reason smoking history is believed to be reported with some uncertainty. Some people convinced that smoking is undoubtedly the cause of their cancer may tend to overestimate it. Others who put the blame on different factors like pollution or work environment could tend to undermined the effect of their smoking habits on their illness and thereby underestimate it. We are interested in examining how different error levels in this variable would influence the results. In order to see how the parameter estimates would be affected, we conduct the analysis three times, each time assuming a different level of error on the smoking history variable. In the first case we assume no error is present, in the second case we assume a 10% error and finally we assume a 25% error. Note that the error when assumed known was incorporated in our method by assuming that the true data covariance matrix Σ was inflated with the addition of the errors' covariance matrix τ, leading to the observed data covariance matrix $\Sigma + \tau$. We are considering the smoking history data to be arising from the following:

$$\log(X_2^*) = \log(Z) + \log(X_2), \tag{5.2}$$

where X_2^* is the observed pack-years, X_2 is the true pack-years and $\log(Z)$ is the error component, which is equivalent to:

$$X_2^* = Z \times X_2. \tag{5.3}$$

In (5.3) the error component Z is now a multiplicative factor of the true pack-years, so obtaining the desired error levels can be done in the following way:

No error: $\tau_{22} = 0$

10% error: $1.1 \cong \exp(.1)$ and $0.9 \cong \exp(-.1) \Rightarrow \tau_{22} = (.1)^2$

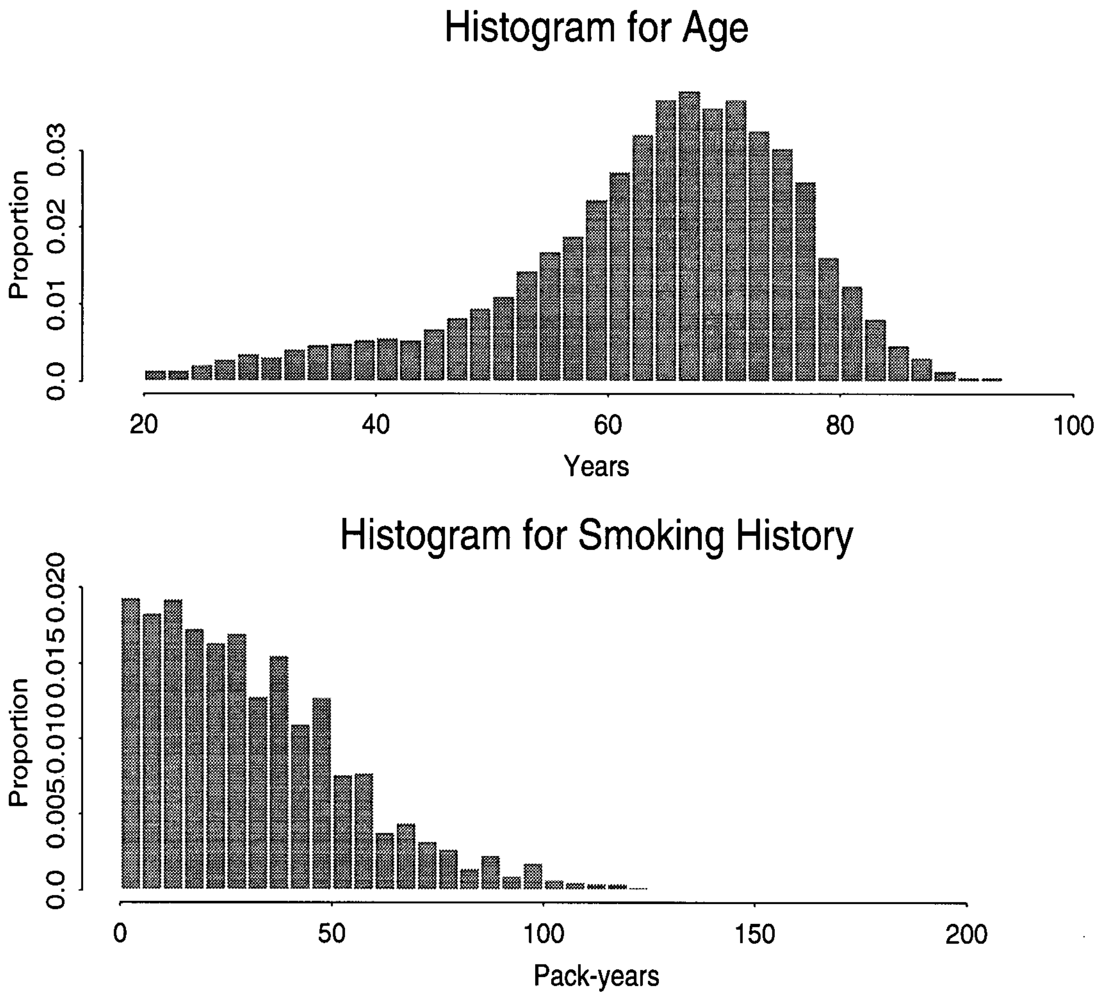
25% error: $1.25 \cong \exp(.25)$ and $0.75 \cong \exp(-.25) \Rightarrow \tau_{22} = (.25)^2$

Note that only the $(2,2)$ element of τ is affected by the error since we assume error only for the smoking history and that this error is not correlated to the other variables. Our correction therefore consists in subtracting from the estimated observed data covariance matrix $\Sigma + \tau$ the covariance matrix τ we have assumed for the error, all this within runs of our Gibbs sampling algorithm because of the presence of a binary variable.

Results of the analysis for each of these error level assumptions are summarized in Figure 5.5, in which can be found histograms of the posterior distribution of the parameters β_1, β_2 and β_3. We can see from the histograms that the parameters β_1 and β_3, respectively related to the logarithm of the age and the diagnosis period, do not seem to be influenced by the different error levels we have assumed for the pack-years. The range and the centrality measures of the posteriors for both β_1 and β_3 remain relatively constant while the error level is changing. Now what happens to the parameter β_2? It is the one we expect should be mainly affected by the assumed error on the reported pack-years and in fact it is. Based on the histograms of Figure 5.5 and some summary statistics for each of these posteriors we can detect an increase in β_2 as the error is increasing. This is exactly what was expected since the estimated parameters are obtained from the difference between the estimated case mean and control mean multiplied by the inverse of the estimated covariance matrix ($\beta = \Sigma^{-1}(\mu_1 - \mu_0)$). When this covariance matrix is estimated from the observed data, it will be overestimated if there is error in the data. Therefore when correcting for the error the estimate of the covariance matrix will be reduced, which means division by a smaller value leading to an increased estimate of the parameter β.

As it was mentioned earlier, the interpretation of the parameter β is that it is the coefficient in the prospective log-odds ratio and that for rare disease the relative risk of disease can be approximated by the odds ratio. Thus, the basic conclusion that can be drawn from the obtained results is that the risk of getting bladder cancer given the smoking history is underestimated if there is in fact error in the observed data. Meaning that an analysis performed assuming no error to be present gives conservative estimates of the relative risk of disease.

It should be noted that in the present analysis it was impossible, as it is for many exposures, to get a validation sample in which we would have had both the true smoking history and the observed one. We could only guess on what the error was. Thus, our approach here was mainly used as a validation, but the potential of the methodology should not be overlooked.



Histogram for Age

Histogram for Smoking History

Figure 5.1: Histograms for the distribution of the age and smoking history variables

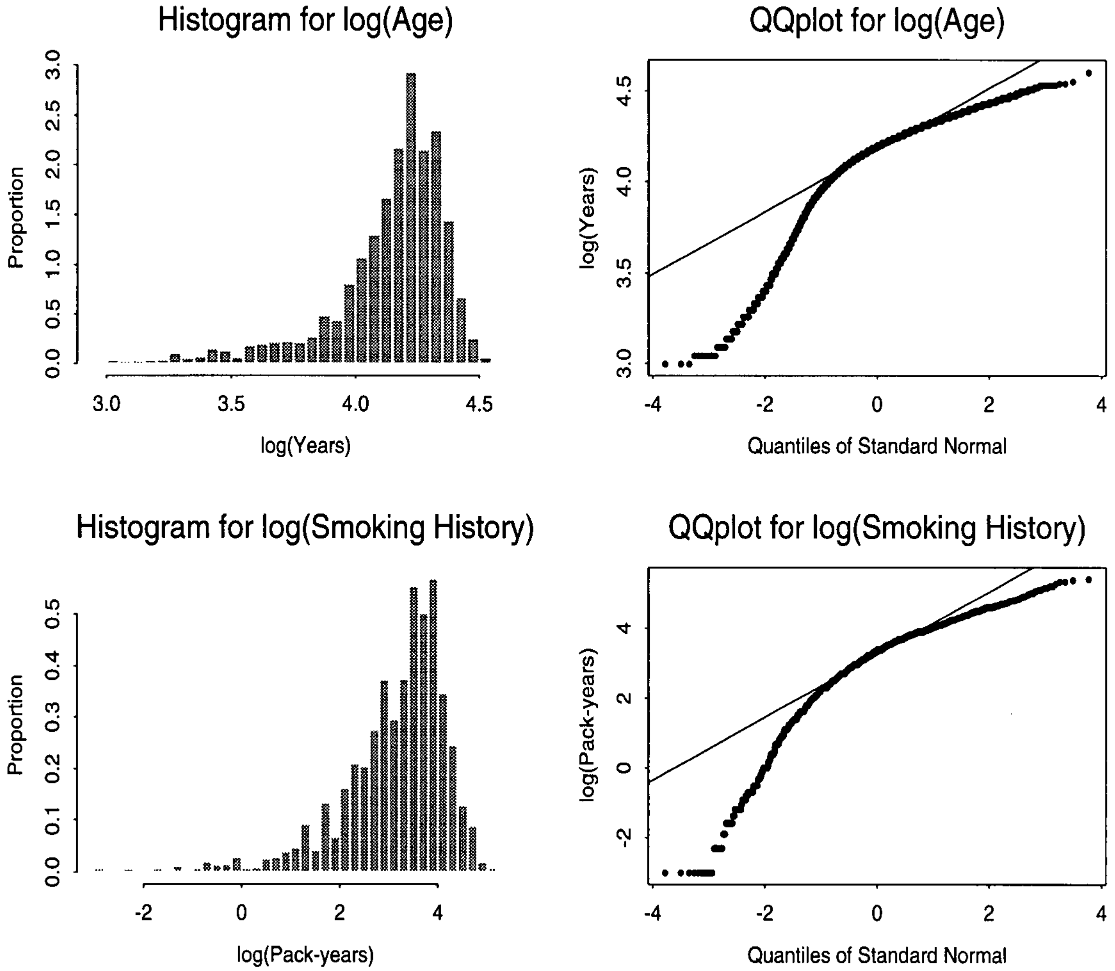


Figure 5.2: Diagnostic plots to verify the normality of the logarithmic transformation of the age and smoking history variables

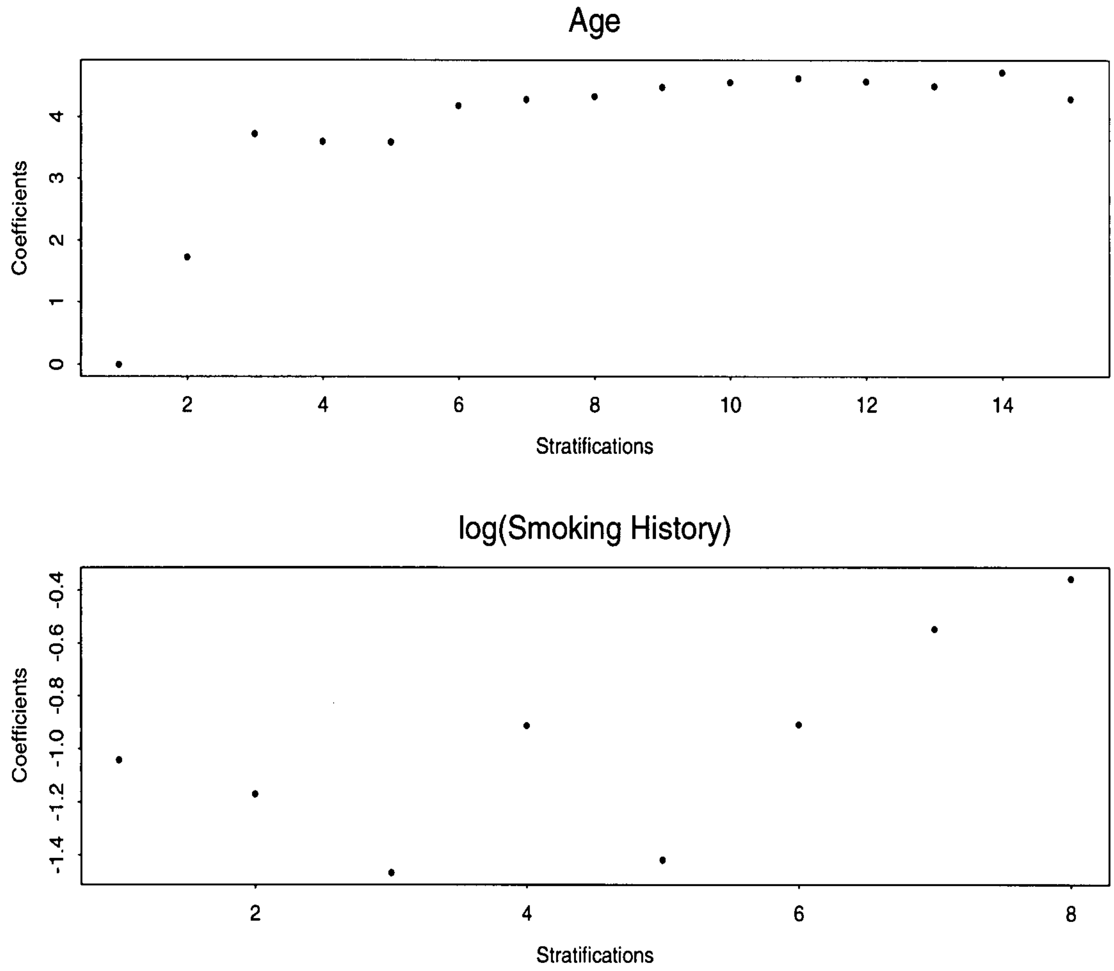


Figure 5.3: Plots of the coefficients from the logistic regression of the disease status on increasing stratificaions of both the age and logarithmic transformation of the smoking history variables

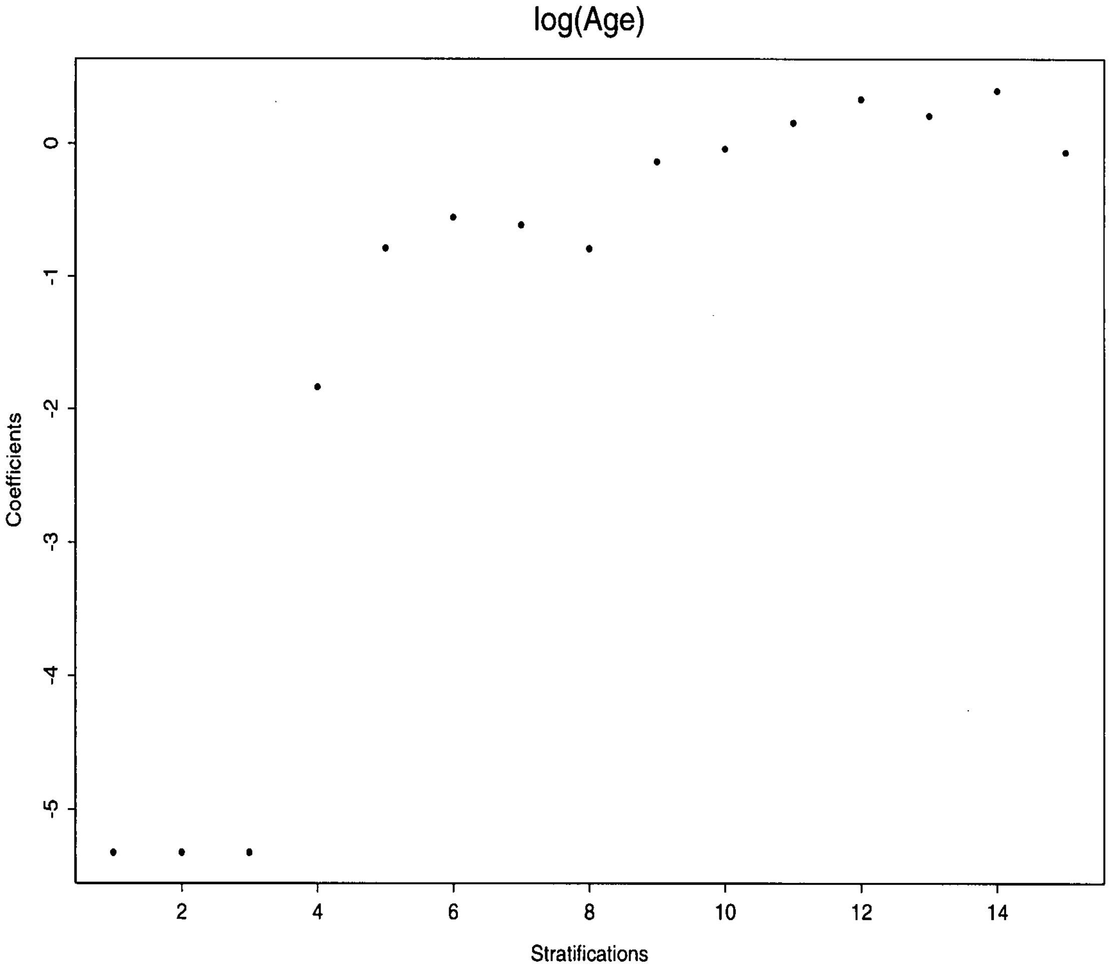


Figure 5.4: Plot of the coefficients from the logistic regression of the disease status on increasing stratificaions of the logarithmic transformation of the smoking history variable

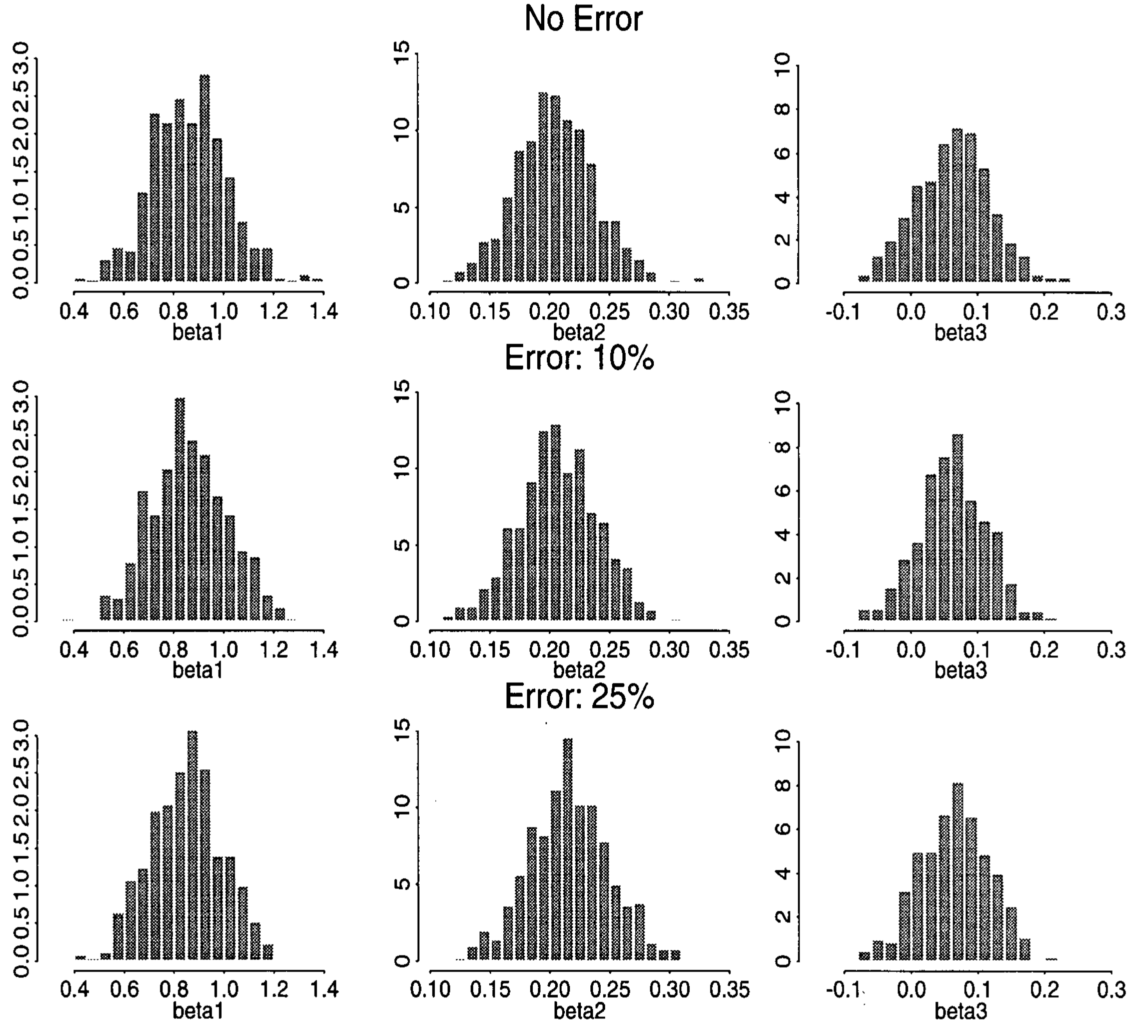


Figure 5.5: Histograms of the posterior distributions of the parameter estimates for β_1, β_2 and β_3 obtained with the assumption there was no measurement error, 10% and finally 25% measurement error in the smoking history variable.

Chapter 6

Conclusion

At the risk of repeating ourselves, imprecise exposure measurements are common in case-control studies. Since perfecting measuring instruments is not always a feasible option, accepting the errors and developing methodology that accounts for them in the analysis has generated a fair amount of literature. Chapter 2 reviewed and summarized some of the work that had been done on the subject. Prospective logistic regression being the typical approach for analyzing case-control data, most of the methods that offered a correction for these errors were based on this methodology which fits a prospective model to the retrospectively sampled data. The alternative we adopted called on the Bayesian approach which revealed to be a fairly natural way of incorporating uncertainty about the unobserved true exposure.

The simulations and comparisons to known procedures carried out in Chapter 3 found our method to perform reasonably. Although the assumptions of the normal discriminant analysis model may be somewhat stronger than those of the prospective logistic regression model they should be met fairly easily in an applied situation. In light of these encouraging results, we pursued our work to make this approach more adaptable to the realities of complex datasets.

In the generalization to multivariate exposure all the basic elements from the univariate setting carried over, allowing the simplicity of the procedure to be kept intact. Simple yet efficient, for models with only continuous exposures the Bayesian inference could be performed with exact posterior sampling. The addition to the model of binary covariates presented a challenging problem. Fortunately, the use of a variation of the probit regression model enabled us to get a continuous representation of these binary covariates. This required some elements of a covariance matrix to be fixed, which led to the development of a general algorithm for sampling such a constrained covariance matrix. The Bayesian inference in this context required the use of a Gibbs sampling algorithm.

Analyzing real case-control data showed the method could be applied fairly easily, and produced results that were in line with what was theoretically expected.

Finally, although we could not explore these in the scope of this thesis, the following suggestions could be considered for further development of the presented methodology. One would be to generalize the approach to make it applicable to matched case-control studies. The other, to explore other applications of the algorithm for sampling covariance matrices with partially fixed diagonal elements, for example multivariate or multinomial probit models.

Bibliography

[1] Albert, M.H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669-679.

[2] Armstrong, B.G., Whittemore, A.S. and Howe, G.R. (1989). Analysis of case-control data with covariate measurement error: application to diet and colon cancer. *Statistics in Medicine* **8**, 1151-1163.

[3] Berger. J.O. and Bernardo, J.M. (1992). On the development of reference priors. *Bayesian Statistics 4* 35-49.

[4] Breslow, N.E. and Day, N.E. (1980). *Statistical Methods in Cancer Research.* Lyon: International Agency for Reseach on Cancer.

[5] Breslow, N.E. (1996) Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association* **91**, 14-28.

[6] Buonaccorsi, J.P. (1990). Double sampling for exact values in the normal discriminant model with application to binary regression. *Commun. Statist. -Theory Meth.* **19**, 4569-4586

[7] Carroll, R.J., Gail, M.H. and Lubin, J.H. (1993). Case-control studies with errors in covariates. *Journal of the American Statistical Association* **88**, 185-199

[8] Carroll, R.J., Roeder, K. and Wasserman, L. (1996). Flexible parametric measurement error models. Technical Report 648, Department of Statistics, Carnegie Mellon University.

[9] Casella, G. and George, E.I. (1992). Explaining the Gibbs sampler. *The American Statistician* **46**, 167-174.

[10] Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit model. *Biometrika* **85**, 347-361.

[11] Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix. *Journal of the National Cancer Institute* **11**, 1269-1275.

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

[13] Forbes, A.B. and Santner, T.J. (1995) Estimators of odds ratio regression parameters in matched case-control studies with covariate measurement error. *Journal of the American Statistical Association* **90**, 1075-1084

[14] Le, N.D. and Zidek, J.V. (1992). Interpolation with uncertain spatial covariances: A Bayesian alternative to kriging. *Journal of Multivariate Analysis* **43**, 351-374.

[15] Mallick, B.K. and Gelfand, A.E. (1996). Semiparametric errors-in-variables models: A Bayesian approach. *Journal of Statistical Planning and Inference* **52**, 307-321.

[16] McKeown-Eyssen, G.E. and Thomas, D.C. (1985). Sample size determination in case-control studies: The influence of the distribution of exposure. *Journal of Chronic Dis.* **38**, 559-568

[17] McKeown-Eyssen, G.E. and Tibshirani, R. (1994). Implications of measurement error in exposure for the sample sizes of case-control studies. *American Journal of Epidemiology* **139**, 415-421.

[18] Muller, P., Parmigiani, G., Schildkraut, J. and Tardella, L. (1996). A Bayesian hierarchical approach for combining case-control and prospective studies. Discussion Paper 96-29, Institute of Statistics and Decision Sciences, Duke University.

[19] Muller, P. and Roeder, K. (1994). A Bayesian semiparametric model for case-control studies with errors in variables. Discussion Paper 94-28, Institute of Statistics and Decision Sciences, Duke University.

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

[21] Richardson, S. and Gilks, W.R. (1993). A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *American Journal of Epidemiology* **138**, 430-442.

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

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

[24] Rosner, B., Spiegelman, D. and Willett, W.C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple covariates measured with error. *American Journal of Epidemiology* **132**, 734-745.

[25] Rosner, B., Willett, W.C. and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine* **8**, 1051-1069.

[26] Ross, S. (1997). *Introduction to Probability Models.* 6th ed. San Diego: Academic Press.

[27] Thomas, D., Stram, D. and Dwyer, J. (1993). Exposure measurement error: Influence on exposure-disease relationships and methods of correction. *Annu. Rev. Publ. Health* **14**, 69-93

[28] Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics* **22**, 1701-1728.

[29] Wolfinger, R.D. and Kass, R.E. (1996). Bayesian analysis of variance component models via rejection sampling. Technical Report 642, Department of Statistics, Carnegie Mellon University.