Semiparametric Regression Methods with Covariate Measurement Error

Nels G. Johnson

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

Inyoung Kim, Chair
Pang Du
Scotland Leman
George Terrell

November 02, 2012
Blacksburg, Virginia

Keywords: Bayesian methods, error-in-covariates, generalized linear models, matched case-control studies, mixed models, semiparametric regression

Semiparametric Regression Methods with Covariate Measurement Error

Nels G. Johnson

Abstract

In public health, biomedical, epidemiological, and other applications, data collected are often measured with error. When mismeasured data is used in a regression analysis, not accounting for the measurement error can lead to incorrect inference about the relationships between the covariates and the response. We investigate measurement error in the covariates of two types of regression models. For each we propose a fully Bayesian approach that treats the variable measured with error as a latent variable to be integrated over, and a semi-Bayesian approach which uses a first order Laplace approximation to marginalize the variable measured with error out of the likelihood.

The first model is the matched case-control study for analyzing clustered binary outcomes. We develop low-rank thin plate splines for the case where a variable measured with error has an unknown, nonlinear relationship with the response. In addition to the semi- and fully Bayesian approaches, we propose another using expectation-maximization to detect both parametric and nonparametric relationships between the covariates and the binary outcome. We assess the performance of each method via simulation terms of mean squared error and mean bias. We illustrate each method on a perturbed example of 1–4 matched case-control study.

The second regression model is the generalized linear model (GLM) with unknown link function. Usually, the link function is chosen by the user based on the distribution of the response variable, often to be the canonical link. However, when covariates are measured with error, incorrect inference as a result of the error can be compounded by incorrect choice of link function. We assess performance via simulation of the semi- and

fully Bayesian methods in terms of mean squared error. We illustrate each method on the Framingham Heart Study dataset.

The simulation results for both regression models support that the fully Bayesian approach is at least as good as the semi-Bayesian approach for adjusting for measurement error, particularly when the distribution of the variable of measure with error and the distribution of the measurement error are misspecified.

Dedication

To Maurice.

Acknowledgments

This has been a tough place to get to and I certainly didn't get here on my own. There are enough people that really made an impact over the years that it would be hard to even list just the most important ones without making errors of omission. So I'll keep it brief and nondescript: thank you to my advisor, my committee, other mentors, family and friends. It means everything.

Contents

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

In this chapter we briefly introduce the classical measurement error model, the two semiparametric regression topics we consider in the presence of covariate measurement error, and then provide some remarks about this document. In Section 1.1 we introduce the classical measurement error model, it's notation, and assumptions. In Section 1.2 we will introduce our work on low-rank thin-plate splines for matched case-control studies. In Section 1.3 we will introduce our work on unknown link function estimation in generalized linear models (GLM). In Section 1.4 we make a statement of purpose and provide an outline for the rest of this document.

1.1 Classical Measurement Error Model

Regression models with covariate measurement error are models for describing the relationship between some response variable Y and some predictor variables (X,Z), where the variables X are measured with error and Z without error. Ignoring that X is measured with error can result in incorrect inference being made about all predictors in the model, including Z. The goal of measurement error methods is to correct for the mismeasurement and make inference as if no measurement error were observed.

For the entirety of this work we consider the classical measurement error model. That is, instead of observing the continuous covariates, X, we observe W, where $W = X + U$, $U \sim N(0, \Sigma_u)$. The joint likelihood used for analysis in such model is $L(Y, W|X, \Sigma_u, -) \propto L(Y|X, -)N(W|X, \Sigma_u)$. We address two main approaches for handling the classical measurement error model:

- Fully Bayesian approach: Make an assumption *a priori* about the distribution of X, say, $X \sim N(\mu_x, \Sigma_x)$, then treat X as a latent variable to be integrated over.

- Semi-Bayesian approach: Marginalize X out of $L(Y, W|X, \Sigma_x, -)$ using a first order Laplace approximation.

Both of these approaches rely upon $k = 1, 2, ..., K_{ij}$ repeated measurements w_{ijk} of x_{ij}, where i indexes observation and j indexes the variable, $K_{ij} > 1$ for some i of every j.

1.2 Matched Case-Control Studies Using Splines

The $1 - M$ matched case-control study is a study design for clustered binary outcomes, often used in public health and epidemiology. The outcomes are clustered intro strata S. In each stratum there is one "case" ($Y = 1$) and M "controls" ($Y = 0$). The model we consider takes the form:

$$P(Y = 1|x, Z, \beta_z, S) = H^{-1}[m(x) + Z\beta_z + q(S)],$$

where $H(\cdot)$ is the probit link function, $m(\cdot)$ is an unknown function of the single covariate measured with error x, the covariates measured without error, Z, have linear relationship β_z, and $q(S) \sim N(0, \sigma^2_\beta)$ are the random block effects for each stratum.

We use low-rank thin-plate splines of order p to nonparametrically estimate $m(\cdot)$:

$$m(x) \approx \beta_0 + \left[x, ..., x^{p-1}\right] \beta_x + \left[|x - \xi_1|^{2p-1}, ..., |x - \xi_\kappa|^{2p-1}\right] \beta_L,$$

where ξ are the knots.

We investigate, via simulation, four methods for simultaneously accounting for error in measurements of x and properly estimating $m(\cdot)$ using the low-rank thin-plate splines. The first and second methods are the fully Bayesian and semi-Bayesian approaches described in Section 1.1. The third and fourth are expectation-maximization (EM) approaches, both of which also marginalize x out of the likelihood. We also apply the methods on a 1–4 matched case-crossover study investigating juvenile aseptic meningitis in children in South Korea.

1.3 Unknown Link Function

Generalized linear models are a flexible class of regression models for handling data where Y follows a distribution of the exponential family, e.g., Normal, Bernoulli, Poisson. A model for fitting GLMs could be written as follows:

$$\begin{aligned} Y &\sim \text{Exponential Family}(g(\mu), \phi), \\ E(Y) &= \mu = g^{-1}(\eta), \\ \eta &= \beta_0 + X\beta_x + Z\beta_z, \end{aligned}$$

where $g(\cdot)$ is the link function and ϕ is the dispersion parameter.

Traditionally $g(\cdot)$ is an invertible function chosen by the user to map η onto the domain of μ. However, the choice of $g(\cdot)$ can affect model fit, so to relax the assumption of known link, we estimate $\mu = g^{-1}(\eta)$ nonparametrically by choosing a baseline link $g_0(\cdot)$, setting $\beta_0 = 0$, and using a specially chosen transformation $T(\cdot)$ of a mixture of R beta distribution functions (i.e., regularized incomplete beta functions):

$$\mu = g^{-1}(\eta) \approx T^{-1}\left(\sum_{r=1}^{R} \omega_r \text{IB}\left\{T\left[g_0^{-1}(\eta)\right], a_r, b_r\right\}\right),$$

where ω are the mixture weights and IB is the distribution function for a beta distribution with parameters (a_r, b_r).

Similarly to our work on matched case-control studies, we investigate the fully Bayesian and semi-Bayesian approach for accounting for measurement error via simulation. We also apply the methods on the Framingham Heart Study data for coronary heart disease, which has binary response.

1.4 Remarks

The focus of this dissertation is to: (1) describe some methods for handling covariate measurement error in semiparametric regression models, and (2) assess the performance of these methods on variety of correctly and incorrectly specified models.

The rest of this document is organized as follows: In Chapter 2 we look at the prospective analysis of matched case-control studies using low-rank thin-plate splines and in

Chapter 3 we look at unknown link function estimation in generalized linear models. Finally, in Chapter 4 we reiterate common themes and discuss possible extensions to the work describe herein.

CHAPTER 2

SEMIPARAMETRIC APPROACHES FOR MATCHED CASE-CONTROL STUDIES WITH ERROR-IN-COVARIATES

Abstract

The matched case-control study is a popular design in public health, biomedical, and epidemiological research for human, animal, and other subjects for clustered binary outcomes. Often covariates in such studies are measured with error. Not accounting for this error can lead to incorrect inference for all covariates in the model. The methods for assessing and characterizing error-in-covariates in matched case-control studies are quite limited. In this article we propose several approaches for handling error-in-covariates that detect both parametric and nonparametric relationships between the covariates and the binary outcome. We propose a fully Bayesian approach, a semi-Bayesian approach, and an approach using expectation-maximization for addressing error-in-covariates that is additive and Gaussian, where the variable measured with error has an unknown, nonlinear relationship with the response. The Bayesian approaches use a latent variable probit model. All methods are developed using the nonparametric method of low-rank thin-plate splines. We assess the performance of each method in terms of mean squared error and mean bias in both simulations and a perturbed example of 1–4 matched case-control study.

2.1 Introduction

In case-control studies the response variable Y is dichotomous, e.g. presence or absence of some disease or injury. Subjects where $Y = 1$ are called cases and subjects where $Y = 0$ are called controls. Often there are potential confounding variables that are not of interest. Subjects with similar responses on these variables are considered part of the same stratum S. Matching subjects based on their stratum can reduce the effect of the confounding. A case-control study where 1 case is matched with M controls within the same stratum is called a 1–M matched case-control study (Agresti, 2002; Hosmer and Lemeshow, 2000). A special case of the matched case-control study is a matched case-crossover study where the stratum is the subject. Matched case-control studies are popular in public health, biomedical, and epidemiological applications, e.g., vaccine studies (Whitney et al., 2006), organ transplant studies (Peleg et al., 2007), and studies on traffic safety (Tester et al., 2004).

Often covariates in such studies are measured with error. Not accounting for this error can lead to incorrect inference for all covariates in the model. A prospective semi-parametric model for matched case-control studies with covariates measured with error is,

$$P(Y = 1|\tilde{X}, Z, S) = H^{-1}[m^*(\tilde{X}, Z) + q(S)],$$

where $H(\cdot)$ is the link function, $q(S)$ is the stratum random block effect, $m^*(\cdot,\cdot)$ is some function of Z, the covariates measure without error, and $\tilde{X}$, the covariates measured with error. Often H is chosen to be the logit link function and then a retrospective model is analyzed using conditional logistic regression in order to avoid estimating $q(S)$. An alternative approach, and the approach we take in this paper, is to estimate the prospective model directly and model $q(S)$ as a random effect. For both approaches, the methods for assessing and characterizing error-in-covariates are quite limited. To our knowledge the only methods available were proposed by McShane et al. (2001) and Guolo and Brazzale (2008) for matched case-control studies analyzed using conditional logistic regression.

There are many methods for error-in-covariates for regression with normal or binary responses. Guolo (2008) use structural approaches where the unknown true covariate

is treated as a random variable. These methods require knowledge of the true exposure rate and the measurement error distribution, including parameters. The methods of Buzas and Stefanski (1996) and Stefanski and Carroll (1987) use functional approaches where the unobserved true covariate is unknown, but considered to be fixed and consequently no assumption is made regarding the distribution of the unobserved true covariate. Berry et al. (2002) propose a fully Bayesian approach for covariate measurement error in semiparametric regression models for normal responses Y. We adapt their fully Bayesian approach for use with matched case-control studies, which treats $\tilde{X}$ as a latent variable to be integrated over. We also develop a semi-Bayesian approach which uses a first order Laplace approximation (Tierney and Kadane, 1986) to marginalize $\tilde{X}$ out of the likelihood. The last method we develop is an approximate expectation-maximization (EM) method (Dempster et al., 1977) uses a first order Laplace approximation to perform the E-step. For a thorough review of non-Bayesian and Bayesian error-in-covariate methods, see Carroll et al. (2006).

For estimating $m^*(\cdot,\cdot)$, we assume $m^*(\tilde{X},Z) = m(x)+Z\beta_z$, where $m(\cdot)$ is a smooth function that can be approximated by the user's favorite spline method, and where only one variable x is measured with error. Our focus will then be on a semiparametric mixed model approach for estimating $m(x)+Z\beta_z$ that addresses covariate measurement error in x in 1–M matched case-control studies. Existing methods for characterizing error-in-covariates in models with clustered binary outcomes cannot estimate nonparametric relationships between the clustered binary outcome and covariates measured with error. Hence, the method we propose are unified approaches in their ability to handle error-in-covariates and detect both parametric and nonparametric relationships between clustered binary outcome and error-in-covariates. The Bayesian approaches are developed using a latent variable probit model (Albert and Chib, 1993), for computation convenience, the EM approach with the probit link, and all are developed with low-rank thin-plate splines. We show through both simulations and a perturbed example of a 1–4 matched case-control study that the fully Bayesian approach is not worse than the semi-Bayesian approach, and that it tends to perform better under model misspecification. We also show the EM approach has computational issues that make it difficult to

implement effectively.

This chapter is organized as follows: In Section 2.2, we describe a semiparametric mixed model with error-in-covariates and estimate it using low-rank thin-plate splines. In Section 2.3, we develop the semi- and fully Bayesian approaches based on the latent variable probit model, as well as an expectation-maximization approach. In Section 2.4, we conduct a simulation study to compare our methods. In Section 2.5, we apply each approach to a 1-4 matched case-case control study for juvenile aseptic meningitis. Section 2.6 contains concluding remarks and possible future work.

2.2 Semiparametric Mixed Model with Error-in-Covariates

We approximate $m(x)$ using low-rank thin-plate splines. That is, $m(x) = X^*\beta_x^* + L_p^*(x)\beta_L^*$, where $X^* = (1, x, \ldots, x^{p-1})$, $\beta_x^* = (\beta_0, \beta_1, \ldots, \beta_{p-1})^T$, $L_p^*(x) = (|x-\xi_1|^{2p-1}, |x-\xi_2|^{2p-1}, \ldots, |x-\xi_\kappa|^{2p-1})$, for some $p = 1, 2, \ldots$, with knots $(\xi_1, \xi_2, \ldots, \xi_\kappa)$ chosen *a priori*. The penalty for β_L^* could be handled using a Bayesian framework by setting the prior for β_L^* to be $N(0, \sigma_\beta^2 \Omega^{-1})$, where the the (r,c)th element of the penalty matrix Ω is $|\xi_r - \xi_c|^{2p-1}$. However, Ω is not positive definite, so we use singular value decomposition to find $(\Omega^{-1/2})^T(\Omega^{-1/2}) = \Omega^{-1}$. We can then transform $\Omega^{1/2}\beta_L^* = \beta_L$ and $L_p^*(x)\Omega^{-1/2} = L_p(x)$. The basis $L_p(x)$ is now orthogonal, so the prior distribution (i.e. penalty) for β_L is $N(0, \sigma_\beta^2 I)$.

For error-in-covariates in matched case-control studies, we assume that we observe $w_{ijk} = x_{ij} + u_{ijk}$, however x_{ij} is unobserved, the measurement error $u_{ijk} \sim N(0, \sigma_u^2)$, $i = 1, 2, \ldots, N$ (the number of strata), $j = 1, 2, \ldots, M+1$ (the number of subjects in each strata), and $k = 1, 2, \ldots, K_{ij}$ is the number of replicated measurements for subject j in strata i. In order to properly estimate σ_u^2, K_{ij} must be greater than or equal to 2 for at least one ij.

To ease computations, we adapt the latent variable probit model of Albert and Chib (1993). We consider a latent variable l such that,

$$l_{ij} \sim N[m(x_{ij}, Z_{ij}) + q(S_i), 1] \times [\delta_{(l_{ij} \geq 0)}\delta_{(y_{ij}=1)} + \delta_{(l_{ij}<0)}\delta_{(y_{ij}=0)}],$$

where δ is an indicator variable. We can summarize our model as follows,

$$
\begin{aligned}
y_{ij}|x_{ij},Z_{ij},S_i &\sim \text{Bernoulli}(\pi_{ij}),\\
\pi_{ij} &= \Phi(l_{ij}),\\
l_{ij} &\sim \begin{cases} N^+(\eta_{ij},1), & y_{ij}=1\\ N^-(\eta_{ij},1), & y_{ij}=0 \end{cases},\\
\eta_{ij} &= m(x_{ij})+Z_{ij}\beta_z+q(S_i),\\
&\approx X^*_{ij}\beta^*_x+L_p(x_{ij})\beta_L+Z_{ij}\beta_z+q(S_i),\\
q(S_i) &\sim N(0,\sigma^2_q),\\
\beta_L &\sim N(0,\sigma^2_\beta),\\
w_{ijk} &= x_{ij}+u_{ijk},\\
u_{ijk} &\sim N(0,\sigma^2_u),
\end{aligned}
$$

where $N^+(\cdot,1)$ and $N^-(\cdot,1)$ are truncated normal distributions, to the left and to the right of zero, respectively.

2.3 Methods

We develop a fully Bayesian (FB) approach and a semi-Bayesian (SB) approach using first order Laplace approximation in Sections 2.3.1 and 2.3.2, respectively. We also propose an expectation-maximization (EM) approach in Section 2.3.3.

2.3.1 Fully Bayesian Approach

To improve computations, we let the intercept β_0 be absorbed into $q(S)$. We work with X_{ij}, which is X^*_{ij} without the column of ones, and β_x, which is β^*_x without β_0. Then, a natural parameterization of the likelihood for modeling additive Gaussian measurement error is as follows,ling additive Gaussian measurement error is as follows,

$$
L(W,Y|l,x,Z,\beta,\sigma^2_u)\propto\prod_{i=1}^{N}\prod_{j=1}^{M+1}\{N[l_{ij};\eta_{ij},1]\times[\delta_{(l_{ij}\geq 0)}\delta_{(y_{ij}=1)}+\delta_{(l_{ij}<0)}\delta_{(y_{ij}=0)}]
$$

$$\times N(W_{ij}; x_{ij}, \sigma_u^2 I_{K_{ij}\times K_{ij}})\}$$

$$\eta_{ij} = Z_{ij}\beta_z + X_{ij}\beta_x + L_p(x_{ij})\beta_L + q(S_i).$$

As mentioned previously, the prior on β_L should be chosen to be $\pi(\beta_L|\sigma_\beta^2) \sim N(0, \sigma_\beta^2)$, where σ_β^2 is a hyperparameter. In practice, the prior distributions placed on x and on $q(S)$ should be chosen to reflect the data collected. For instance, Carroll et al. (1999) discuss methods for modeling x using a mixture of normals. For this article, we choose $\pi(x|\mu_x, \sigma_x^2) \sim N(\mu_x, \sigma_x^2)$ and $\pi[q(S)|\beta_0, \sigma_q^2] \sim N(\beta_0, \sigma_q^2)$, where μ_x, β_0, σ_x^2, and σ_q^2 are hyperparameters. The prior distributions for the other parameters are: $\pi(\sigma_u^2) \sim IG(\sigma_u^2; A_u, B_u)$, and $\pi[\beta_z, \beta_x] \sim N(\beta_z, \beta_x; g_\beta, t_\beta^2)$. Finally, we take the prior distributions for the hyper parameters as follows: $\pi(\mu_x) \sim N(\mu_x; g_\mu, t_\mu^2)$, $\pi(\sigma_x^2) \sim IG(\sigma_x^2; A_x, B_x)$, $\pi(\sigma_\beta^2) \sim IG(\sigma_\beta^2; A_{\sigma_\beta^2}, B_{\sigma_\beta^2})$, $\pi(\beta_0) \sim N(\beta_0; g_0, t_0^2)$, and $\pi(\sigma_q^2) \sim IG(\sigma_q^2; A_q, B_q)$. Both the likelihood and prior structure are adapted from Berry et al. (2002), which defaults to normal priors on mean-like parameters and inverse-gamma priors on variance parameters. In practice, careful choice of an informative prior structure can further improve inference. For example, it may be more appropriate to restrict the support of the prior on σ_u to be less than the value of σ_w since σ_w should be greater than σ_u when errors are additive and Gaussian.ling x using a mixture of normals. For this article, we choose $\pi(x|\mu_x, \sigma_x^2) \sim N(\mu_x, \sigma_x^2)$ and $\pi[q(S)|\beta_0, \sigma_q^2] \sim N(\beta_0, \sigma_q^2)$, where μ_x, β_0, σ_x^2, and σ_q^2 are hyperparameters. The prior distributions for the other parameters are: $\pi(\sigma_u^2) \sim IG(\sigma_u^2; A_u, B_u)$, and $\pi[\beta_z, \beta_x] \sim N(\beta_z, \beta_x; g_\beta, t_\beta^2)$. Finally, we take the prior distributions for the hyper parameters as follows: $\pi(\mu_x) \sim N(\mu_x; g_\mu, t_\mu^2)$, $\pi(\sigma_x^2) \sim IG(\sigma_x^2; A_x, B_x)$, $\pi(\sigma_\beta^2) \sim IG(\sigma_\beta^2; A_{\sigma_\beta^2}, B_{\sigma_\beta^2})$, $\pi(\beta_0) \sim N(\beta_0; g_0, t_0^2)$, and $\pi(\sigma_q^2) \sim IG(\sigma_q^2; A_q, B_q)$. Both the likelihood and prior structure are adapted from Berry et al. (2002), which defaults to normal priors on mean-like parameters and inverse-gamma priors on variance parameters. In practice, careful choice of an informative prior structure can further improve inference. For example, it may be more appropriate to restrict the support of the prior on σ_u to be less than the value of σ_w since σ_w should be greater than σ_u when errors are additive and Gaussian.

We use Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) and Gibbs (Geman and Geman, 1984) algorithms to sample the joint posterior of these parameters

using Markov chain Monte Carlo (MCMC). The joint posterior distribution of x uses a Metropolis-Hastings step, while all other parameters can be sampled using Gibbs steps. The conditional posterior distributions for each parameter are as follows:

- Full conditional for x_{ij} is:

$$[x_{ij}|-] \propto L(Y_{ij}, W_{ij}|l_{ij}, x_{ij}, Z_{ij}, \beta, q(S), \sigma_u^2) \times N(x_{ij}; \mu_x, \sigma_x^2),$$

- Full conditional for σ_u^2 is:

$$[\sigma_u^2|-] \sim IG\Bigg[\sigma_u^2; (1/2)\sum_{i=1}^{N}\sum_{j=1}^{M+1} K_{ij} + A_u, \\ \sum_{i=1}^{N}\sum_{j=1}^{M+1}(W_{ij} - x_{ij})^T(W_{ij} - x_{ij})/2 + B_u\Bigg],$$

- Full conditional for μ_x is:

$$[\mu_x|-] \sim N\Bigg\{\mu_x; \Big[t_\mu^2 \sum_{i=1}^{N}\sum_{j=1}^{M+1} x_{ij} + g_\mu \sigma_x^2\Big]/[N(M+1)t_\mu^2 + \sigma_x^2], \\ t_\mu^2 \sigma_x^2/[N(M+1)t_\mu^2 + \sigma_x^2]\Bigg\},$$

- Full conditional for σ_x^2 is:

$$[\sigma_x^2|-] \sim IG[\sigma_x^2; N(M+1)/2 + A_x, (x - \mu_x)^T(x - \mu_x)/2 + B_x],$$

- Full conditional for (β_x, β_z) is:

$$[\beta_x, \beta_z|-] \sim MN\{\beta_x, \beta_z; \\ [(X,Z)^T(X,Z) + I/t_\beta^2]^{-1}(X,Z)^T[l - L_p(x)\beta_L - Jq(S)], \\ [(X,Z)^T(X,Z) + I/t_\beta^2]^{-1}\},$$

- Full conditional for β_L is:

$$[\beta_L|-] \sim MN\{\beta_L; \\ [L_p(x)^T L_p(x) + I/\sigma_\beta^2]^{-1} L_p(x)^T[l - X\beta_x - Z\beta_z - Jq(S)], \\ [L_p(x)^T L_p(x) + I/\sigma_\beta^2]^{-1}\},$$

- Full conditional for σ^2_β is:

$$[\sigma^2_\beta|-] \sim IG(\sigma^2_\beta; \kappa/2 + A_\beta, \beta_L^T \beta_L/2 + B_\beta),$$

- Full conditional for $q(S)$ is:

$$\begin{aligned}[q(S)|-] \sim & MN[q(S); \\ & (J^T J + I/\sigma^2_q)^{-1}\{\beta_0/\sigma^2_q + J^T[l - X\beta_x - Z\beta_z - L_p(x)\beta_L]\}, \\ & (J^T J + I/\sigma^2_q)^{-1}],\end{aligned}$$

- Full conditional for σ^2_q is:

$$[\sigma^2_q|-] \sim IG\{\sigma^2_q; N/2 + A_q, [q(S) - \beta_0]^T [q(S) - \beta_0]/2 + B_q\},$$

- Full conditional for β_0 is:

$$[\beta_0|-] \sim N\left\{\beta_0; \left[t_0^2 \sum_{i=1}^{N} q(S_i) + g_0 \sigma^2_q\right] / (N t_0^2 + \sigma^2_q), t_q^2 \sigma^2_q / (N t_q^2 + \sigma^2_q)\right\},$$

where J is a $N(M+1) \times N$ matrix such that,

$$J = \begin{bmatrix} 1_{(M+1)\times 1} & 0 & \dots & 0 \\ 0 & 1_{(M+1)\times 1} & 0 & \vdots \\ 0 & 0 & \ddots & 0 \\ 0 & \dots & 0 & 1_{(M+1)\times 1} \end{bmatrix}.$$

When choosing a proposal distribution for x_{ij}, we take the advice of Berry et al. (2002) and use $x_{ij}^{(t)} \sim N(x_{ij}^{(t-1)}, 2^2 \sigma_u^{2^{(t-1)}} / K_{ij})$, where $2^2 \sigma_u^2 / K_{ij}$ is chosen as the proposal variance because it covers about 95% of the sampling distribution for $\bar{w}_{ij\cdot} = K_{ij}^{-1} \sum_{k=1}^{K_{ij}} w_{ijk}$.

2.3.2 Semi-Bayesian Approach

Updating each x_{ij} via Metropolis-Hastings can be quite time consuming computationally. An alternative that would reduce computations dramatically would be to use the unconditional likelihood $\int L(Y, W|l, x, Z, S, \beta, \sigma^2_u) \mathrm{d}x = L(Y, W|l, Z, S, \beta, \sigma^2_u)$. This integration is intractable due to the spline portion of the linear predictor. If we use a

first order Laplace approximation (Tierney and Kadane, 1986) to solve the integration, we get $\int L(Y,W|l,x,Z,S,\beta,\sigma_u^2)\,\mathrm{d}x \approx L(Y|l,x=\bar{w},Z,S,\beta)$ (See Appendix 2.A.1 for derivation). This produces a simple way of handling error-in-covariates by substituting $\bar{w}_{ij\cdot}$ in for x_{ij} and then proceeding as if there were no measurement error. That is, we use the following likelihood in our Bayesian analysis,

$$\begin{aligned} L(Y|l,\bar{w},Z,\beta) \propto & \prod_{i=1}^{N}\prod_{j=1}^{M+1}\{N[l_{ij};\eta_{w,ij},1] \\ & \times [\delta_{(l_{ij}\geq 0)}\delta_{(y_{ij}=1)} + \delta_{(l_{ij}<0)}\delta_{(y_{ij}=0)}], \\ \eta_{w,ij} = & Z_{ij}\beta_z + \bar{W}_{ij\cdot}\beta_x + L_p(\bar{w}_{ij\cdot})\beta_L + q(S_i). \end{aligned}$$

where $\bar{W}_{ij\cdot} = (\bar{w}_{ij\cdot}, \bar{w}^2_{ij\cdot}, \ldots, \bar{w}^{p-1}_{ij\cdot})$.

The prior structure we adopt for the rest of this model, i.e. β, $q(S)$, and their hyperparameters, is the same as in Section 2.3.1. We then obtain the same conditional posteriors them as well.

2.3.3 Expectation-Maximization Approach

We also develop an EM approach (Dempster et al., 1977) for the original parameterization of the model for traditional probit analysis:

$$\begin{aligned} Y_{ij}|x_{ij},Z_{ij},S_i,\beta \sim & \text{Bernoulli}(\pi_{ij}), \\ \pi_{ij} = & \Phi(\eta_{ij}), \\ \eta_{ij} = & X^*\beta_x^* + L_p(x_{ij})\beta_L + Z_{ij}\beta_z + q(S_i), \\ q(S_i) \sim & N(0,\sigma_q^2), \\ \beta_L \sim & N(0,\sigma_\beta^2), \\ w_{ijk} = & x_{ij} + u_{ijk}, \\ u_{ijk} \sim & N(0,\sigma_u^2). \end{aligned}$$

The EM algorithm is as follows:

- Step 0: (Initial step):

- Set σ_β^2, σ_q^2, μ_x, σ_x^2, and σ_u^2 to values that will remain fixed.
- Initialize β_z, β_x^*, β_L, and $q(S)$ to values that will be updated.
- We define $L_{\text{old}} = -\sum_{i=1}^{N}\sum_{j=1}^{M+1}[Y_{ij}\log(\pi_{ij})+(1-Y_{ij})\log(1-\pi_{ij})] - \frac{\beta_L^T\beta_L}{2\sigma_\beta^2} - \frac{q(S)^Tq(S)}{2\sigma_q^2}$.

- Step 1: (E-step) Find expected log-likelihood with respect to x and update as follows;
  - Set $x_{ij} = \tilde{x}_{ij}$, for each ij; we describe how to calculate $\tilde{x}_{ij}$ in the end of this section.
  - Calculate $Q = \log\{L[Y|x=\tilde{x}, Z, \beta, q(S)]\}$.

- Step 2: (M-step) Find parameters that maximize Q and update them as follows;
  - Update $[\beta_z, \beta_x^*]^T = [\beta_z, \beta_x^*]^T - H_\beta^{-1} g_\beta$, where, $g_\beta = [Z, X^*]^T Y^*$ and $H_\beta = s^*[Z, X^*]^T[Z, X^*]$.
  - Update $\beta_L = \beta_L - H_L^{-1} g_L$, where, $g_L = L_p(x)^T Y^*$ and $H_L = s^* L_p(x)^T L_p(x)$.
  - Update $q(S) = q(S) - H_q^{-1} g_q$, where, $g_q = J^T Y^*$ and $H_q = s^* J^T J$,

  where,

$$Y^* = \left[Y\frac{\phi(\eta)}{\Phi(\eta)} - (1-Y)\frac{\phi(\eta)}{1-\Phi(\eta)} \right],$$
$$s^* = \sum_{i=1}^{N}\sum_{j=1}^{1+M} \phi(\eta_{ij}) \left[Y_{ij}\frac{\phi(\eta_{ij})}{\Phi(\eta_{ij})} + (1-Y_{ij})\frac{1-\phi(\eta_{ij})}{1-\Phi(\eta_{ij})} \right],$$

  and define $L_{\text{new}} = -\sum_{i=1}^{N}\sum_{j=1}^{M+1}[Y_{ij}\log(\pi_{ij})+(1-Y_{ij})\log(1-\pi_{ij})] - \frac{\beta_L^T\beta_L}{2\sigma_\beta^2} - \frac{q(S)^Tq(S)}{2\sigma_q^2}$.

- Step 3: Repeat Step 1 and step 2 until convergence, $L_{\text{new}} - L_{\text{old}} < \text{prec}$, for some small value prec.

We consider three methods, E1, E2, and E3, for finding $\tilde{x}_{ij}$ in the E-step:

E1: Set $\tilde{x}_{ij} = \bar{w}_{ij\cdot}$.

E2: Set $\tilde{x}_{ij} = \frac{K_{ij}\bar{w}_{ij\cdot}\sigma_x^2 + \mu_x\sigma_u^2}{K_{ij}\sigma_x^2 + \sigma_u^2}$.

E3: Generate n samples $x_{ij}^{(t)}$ from:

$$(x_{ij}|-) \propto [Y_{ij}\log(\pi_{ij}) + (1-Y_{ij})\log(1-\pi_{ij})] \times N(W_{ij}; x_{ij}, \sigma_u^2) \times N(x_{ij}; \mu_x, \sigma_x^2),$$

using Metropolis-Hastings, for each ij. Set $\tilde{x}_{ij} = n^{-1}\sum_{t=1}^{n} x_{ij}^{(t)}$.

The first two methods, E1 and E2, ignore the $L(Y|x)$ part of the likelihood $L(Y,W|x)$, and are thus not using all the information available about x. E1 is based on the first order Laplace approximation in Section 2.3.2. While E2 is based the first order Laplace approximation of $N(W|x,\sigma_u^2) \times N(X;\mu_x,\sigma_x^2)$ (see Appendix 2.A.2 for derivation). E1 is essentially plugging in the MLE of x, while E2 is plugging in the Bayes estimator of x. The third method, E3, uses a Bayesian approach to perform the integration, updating x_{ij} to be its posterior mean. It does use all the available information about x in the joint likelihood $L(Y,W|x)$, however, it is computationally very expensive. Updating $N(M+1)$ x_{ij}'s in the E-step, the solutions $\tilde{x}_{ij}$ move around too much unless a very large number of posterior samples n is used. Additionally, the algorithm can be easily confused and accidentally indicate convergence much too early. For this reason, we will not investigate the performance of third method further in this chapter. Therefore, we propose two EM algorithms: one developed using an E1 and M step (EMM1) and the other using an E2 and M step (EMM2).

2.4 Simulation Study

We considered four methods for adjusting for error in covariates, (1) the fully Bayesian approach of Section 2.3.1, (2) the semi-Bayesian approach of Section 2.3.2, (3) the EM approach of Section 2.3.3 using E1 for the E-step, and (4) using E2 for the E-step. To assess the adequacy of each approach for correcting for covariate measurement error, we conducted a simulation study to address performance in terms of minimizing both the mean squared error and the mean bias. In Section 2.4.1 we address model performance when the assumptions concerning the covariate measurement error are met. In Section 2.4.2 we address the robustness of each method when there is model misspecification error in the distribution of x and u. In Section 2.4.3 we describe the results,

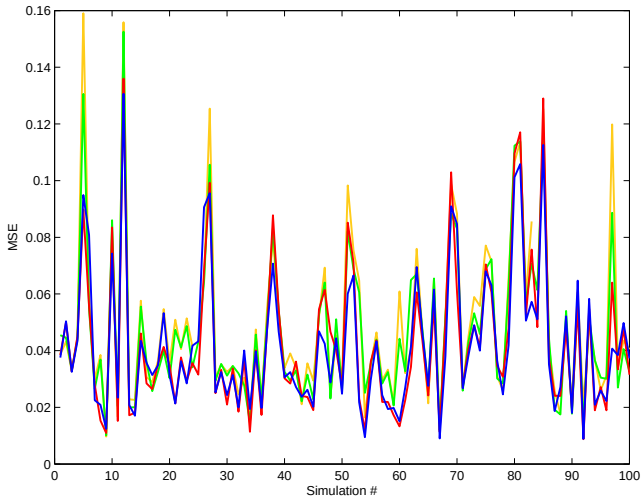

Figure 2.1: The mean squared error for computed for each simulated data set for, where $m(x) = x^2/6$, $N = 25$, and $\sigma_u = 0.5$. The yellow line is for the EM algorithm using method 1 for the E-step, the green line is for the EM algorithm using method 2 for the E-step, the red line is for the semi-Bayesian method, and the blue line is for the fully Bayesian method. All methods perform similarly for this simulation, so it is difficult to tell their performance apart in this figure.

which will mainly focus on comparisons of the semi-Bayesian and fully Bayesian results due to large inconsistencies in the performance of the EM methods and issue regarding invertibility of the Hessian matrix during the M-step, preventing convergence.

For all simulations we set $K_{ij} = 2$ for all ij, $M = 4$. We look at only a single covariate z measured without error, with $\beta_z = -0.5$. We simulate $z \sim N(0,1)$ and $q(S) \sim N(0, 0.1^2)$. We also look at two functions $m(x) = x^2/6$ and $m(x) = \sin(\pi x/2)$. To generate the clustered binary outcomes, $1+M$ binary outcomes were generated from $P(Y = 1|x,z,S,\beta_z) = \Phi[m(x) + z\beta_z + q(S)]$ until $\sum_{j=1}^{1+M} Y_j = 1$. This was repeated for each of the N strata. This process is repeated to produce 100 datasets for each simulation setup.

For each Bayesian approach, we use the same prior structure as noted in Section 2.3.1, where $\{A_u, B_u, A_x, B_x, A_{\sigma_\beta^2}, B_{\sigma_\beta^2}, A_q, B_q\} = 0.1$, $\{g_\mu, g_\beta, g_0\} = 0$, and $\{t_\mu^2, t_\beta^2, t_0^2\} = 5^2$. For the EM approach using E2, we set $\mu_x = N^{-1}\sum_{i=1}^{N}(M+1)^{-1}\sum_{j=1}^{M+1}\bar{w}_{ij\cdot}$, $\sigma_x^2 = [N(M+1)]^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M+1}(\bar{w}_{ij\cdot} - \mu_x)^2$, and $\sigma_u^2 = [N(M+1)]^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M+1}K_{ij}^{-1}\sum_{k=1}^{K_{ij}}(W_{ijk} - \bar{w}_{ij\cdot})^2$. Also for the EM methods we set σ_β^2 and σ_q^2 to the posterior mean of each found using the fully Bayesian approach.

For estimation, we use low-rank thin-plate splines with $\kappa = 10$ knots, chosen at

evenly spaced percentiles of $\bar{w}$, and with order $p = 2$, for all methods. The mean squared error $[N(M+1)]^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M+1}\left(\hat{\eta}_{ij}^{(\cdot)} - \hat{\eta}_{ij}^{(T)}\right)^2$ and mean bias $[N(M+1)]^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M+1}\left(\hat{\eta}_{ij}^{(\cdot)} - \hat{\eta}_{ij}^{(T)}\right)$ is computed for each simulation dataset, where $\hat{\eta}^{(\cdot)}$ is the estimated linear predictor using one of the proposed methods, and $\hat{\eta}^{(T)}$ is the estimated linear predictor for the fully Bayesian approach with perfect measurements for x.

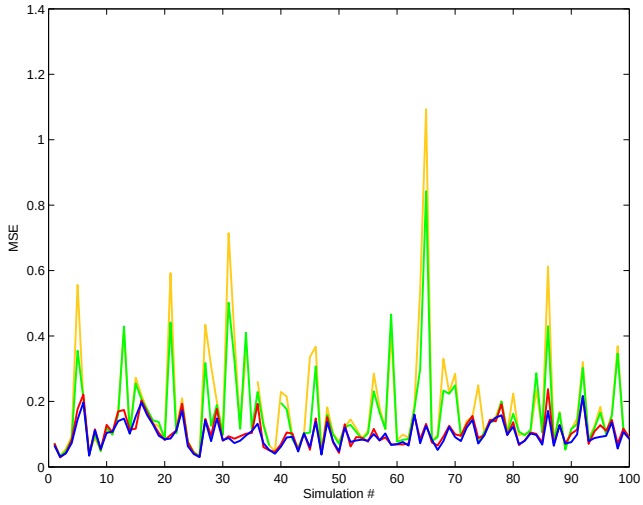

Figure 2.2: The mean squared error for computed for each simulated data set for, where $m(x) = \sin(\pi x/2)$, $N = 25$, and $\sigma_u = 0.5$. The yellow line is for the EM algorithm using method 1 for the E-step, the green line is for the EM algorithm using method 2 for the E-step, the red line is for the semi-Bayesian method, and the blue line is for the fully Bayesian method. All methods perform similarly for this simulation, so it is difficult to tell their performance apart in this figure.

2.4.1 Correctly Specified Model

In this first set we look at $x \sim N(0,1)$ and u such that $\sigma_u = \{0.1, 0.3, 0.5\}$, corresponding to small, large, and very large amounts of measurement error when $\sigma_x = 1$ according to Parker et al. (2010). We also consider small and large sample situations with the number of strata $N = \{25, 100\}$.

2.4.2 Model Mis-specification

In this simulation set we look at three cases of model misspecification. One case where the distribution of x is misspecified, two where the distribution of u is misspeci-

fied, and then the last case where both distributions are misspecified:

- $2^{3/2} \times (x+4) \sim \chi^2_4$ and $u \sim N(0, \sigma_u = 0.5)$
- $x \sim N(0,1)$ and $u \sim \text{Laplace}\left[0, \text{scale} = 2^{-3/2}\right]$
- $2^{3/2} \times (x+4) \sim \chi^2_4$ and $u \sim \text{Laplace}\left[0, \text{scale} = 2^{-3/2}\right]$

The misspecified distributions are chosen such that $\sigma_u/\sigma_x = 0.5$ for all cases. For this set, we consider a moderate sample size with the number of strata $N = 50$.

σ_u	N		SB	FB
0.1	25	MMB	**0.0509**	-0.2130
		MMSE	0.3220	**0.2957**
	100	MMB	0.0498	**0.0148**
		MMSE	**0.1427**	0.1503
0.3	25	MMB	**-0.1358**	-0.6797
		MMSE	**1.9079**	1.9413
	100	MMB	0.1393	**-0.0481**
		MMSE	0.7767	**0.7505**
0.5	25	MMB	**-0.1692**	-1.2819
		MMSE	4.2503	**4.2035**
	100	MMB	0.2307	**-0.1485**
		MMSE	1.7862	**1.7210**

Table 2.1: The MMB$\times 10^2$ and the MMSE$\times 10^2$ of the linear predictors $\hat{\eta}^{(\cdot)}$ for the case where $m(x) = x^2/6$ for comparing the semi-Bayesian method (SB) and the fully Bayesian method (FB).

2.4.3 Simulation Results

Firstly, our approach for selecting values for σ^2_k and σ^2_q for both EM approaches led to inconstancies in performance levels for each simulation setup. For example, the MSE of the correctly specified model where $m(x) = x^2/6$, $N = 25$, and $\sigma_u = 0.5$ is plotted over simulation number in Figure 2.1. We see both EM approaches performing very well in comparison the the semi- and fully Bayesian methods. However if we look at the corresponding simulation where $m(x) = \sin(\pi x/2)$, shown in Figure 2.2, the performance of both EM methods is substantially worse than the semi- and fully Bayesian methods.

Due to the inconsistency in performance, and computational difficulties, we do not investigate the EM approaches further. A more detailed explanation for this choice can be found in the Appendix 2.C.

σ_u	N		SB	FB
0.1	25	MMB	**0.0716**	-0.0954
		MMSE	0.6372	**0.6363**
	100	MMB	**0.0929**	0.0990
		MMSE	**0.4688**	0.4791
0.3	25	MMB	1.3661	**0.1965**
		MMSE	4.2483	**4.1058**
	100	MMB	1.0568	**0.3580**
		MMSE	3.2328	**3.0816**
0.5	25	MMB	2.3124	**0.0997**
		MMSE	10.4693	**9.6985**
	100	MMB	2.3269	**0.7694**
		MMSE	7.1887	**6.8378**

Table 2.2: The MMB$\times 10^2$ and the MMSE$\times 10^2$ of the linear predictors $\hat{\eta}^{(\cdot)}$ for the case where $m(x) = \sin(\pi x/2)$ for comparing the semi-Bayesian method (SB) and the fully Bayesian method (FB).

Tables 2.1 and 2.2 present the results for when the distribution of x and u are correctly specified, for the quadratic $m(x) = x^2/6$ and sinusoidal $m(x) = \sin(\pi x/2)$ cases, respectively. We observe that for the quadratic cases, neither method consistently reduces the bias or MSE over the other. However, for the sinusoidal cases the fully Bayesian approach provides greater reduction in MSE and bias than the semi-Bayesian approach in almost all cases. These results suggest that the fully Bayesian is at least as good as the semi-Bayesian approach.

Distribution of X	Distribution of U		SB	FB
Normal	Laplace	MMB	**0.0853**	-0.5722
		MMSE	2.7317	**2.6548**
χ^2	Normal	MMB	**-0.1753**	-0.7628
		MMSE	2.9160	**2.7046**
χ^2	Laplace	MMB	**-0.0929**	-0.8170
		MMSE	2.6917	**2.5126**

Table 2.3: The MMB$\times 10^2$ and the MMSE$\times 10^2$ of the linear predictors $\hat{\eta}^{(\cdot)}$ for assessing robustness of the semi-Bayesian method (SB) and the fully Bayesian method (FB) to model misspecification of the distribution of x and u when $m(x) = x^2/6$.

Distribution of X	Distribution of U		SB	FB
Normal	Laplace	MMB	2.3627	**0.3828**
		MMSE	8.4782	**7.9205**
χ^2	Normal	MMB	2.0390	**-0.2838**
		MMSE	8.1365	**7.7085**
χ^2	Laplace	MMB	1.8992	**-0.1211**
		MMSE	7.8498	**7.3134**

Table 2.4: The MMB$\times 10^2$ and the MMSE$\times 10^2$ of the linear predictors $\hat{\eta}^{(\cdot)}$ for assessing robustness of the semi-Bayesian method (SB) and the fully Bayesian method (FB) to model misspecification of the distribution of x and u when $m(x) = \sin(\pi x/2)$.

Tables 2.3 and 2.4 present the results for when the distribution of x and u are incorrectly specified, for the quadratic $m(x) = x^2/6$ and sinusoidal $m(x) = \sin(\pi x/2)$ cases, respectively. We observe that for the quadratic cases, the fully Bayesian approach provides better reduction in MSE than the fully Bayesian approach for all misspecification types. The opposite is observed for bias, where semi-Bayesian approach provides better reduction than the fully Bayesian for all misspecification types. However, this is not true for the sinusoidal cases where the fully Bayesian approach reduces both the bias and MSE more than the semi-Bayesian approach for all misspecification types. These results suggest the semi-Bayesian approach might be at least as good at reducing bias, however the fully Bayesian approach is at least as good at reducing MSE, both when the distribution of x or u is misspecified.

There be some concern that the results of performance for the semi- and fully Bayesian methods may depend too heavily on the priors chosen for σ^2_β and σ^2_q. A small simulation study addressing this issue can be found in Appendix 2.B. Based on the results of that small simulation, we do not feel the need to change our conclusions concerning the semi- and fully Bayesian methods.

2.5 Application: Juvenile Aseptic Meningitis Data

Aseptic meningitis is a viral infection that causes inflammation of the membrane that covers the brain and spinal chord. It is rarely fatal, but can take about two weeks to recover from fully. The study design was for a 1-4 case-crossover study with 64 strata

	SB	FB
MB	**0.8112**	-1.0794
MSE	4.6578	**4.4914**

Table 2.5: MB$\times 10^2$ and MSE$\times 10^2$ of the posterior-mean fitted values for the juvenile aseptic meningitis data.

(i.e. subjects). Water turbidity (the amount of suspended matter in the water) is believed to affect the risk of aseptic meningitis. For this study water turbidity was measured by a nephelometer. A nephelometer shoots a beam of light at water then measures the scattered light. It then uses a formula to determine the turbidity, measured in Nephelometric Turbidity Units (NTU). The device is susceptible to miscalibration, and can be thrown off by air bubbles that may make water appear cloudy that does not actually contain any suspended particles. This study design was not setup for multiple measurements of NTU, so for illustrative purposes we center and scale NTU, then add noise to every measurement to get two observations measured with error, where the amount of error is $u \sim N(0, \sigma_u^2 = 0.5^2)$. We use the same model specification as used in our simulations. For Z we use the centered and scaled body temperature of the subjects in degrees Celsius.

Figure 2.3 shows a plot of the posterior mean fitted value for $m(\cdot)$ using the data measured without error, the semi-Bayesian and fully Bayesian methods. Introducing the measurement error has resulted in some deviation from the fit without error. From Table 2.5 we can see a reduction of the MSE from 0.0466 to 0.0449 by using the fully Bayesian method instead of the semi-Bayesian. This corresponds to only a 3.7% improvement. We can also see that the semi-Bayesian method reduced the absolute bias by about 25% compared the fully Bayesian method, -0.011 to 0.008. The posterior mean for σ_u^2 from the fully Bayesian model was 0.2502, which is very close to the true value of 0.25. The 90% equal tail credible interval for σ_u^2 was $[0.2201, 0.2833]$. From Figure 2.4 we can see that the distribution of NTU is not normally distributed. In violation of this information, we placed a normal prior on the distibution of NTU, making a model misspecification error. From our simulations we would expect fully Bayesian method to be better at reducing the MSE in such a situation, but might not be better at reducing the

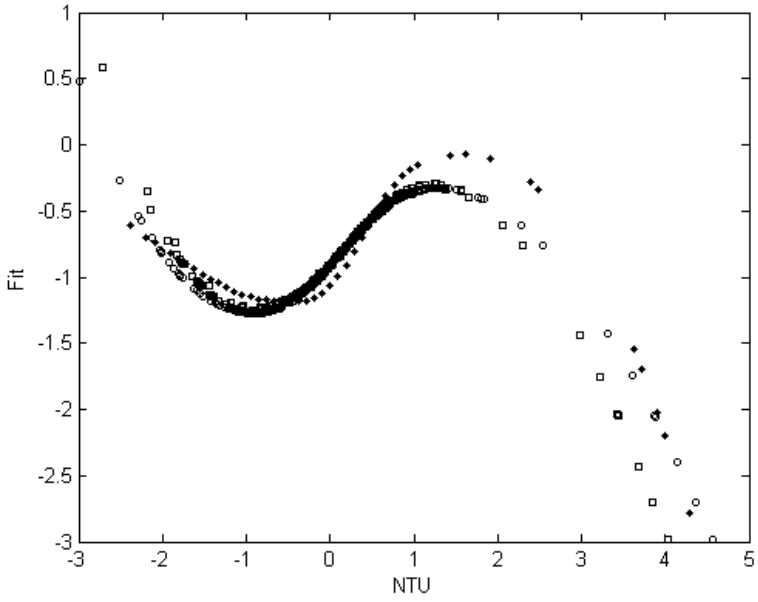

Figure 2.3: The posterior-mean fits of $m(x)$ for the aseptic meningitis data, where x is Nephelometric Turbidity Units (NTU) and z is body temperature. Both x and z have been centered and scaled. The black dots are the posterior-means of the centered fitted values for NTU when there is no measurement error, evaluated at NTU measured without error x. Similarly, the black circles are for the semi-Bayesian method evaluated at $\bar{w}$ for NTU, and the black squares are for the fully Bayesian method evaluated at the posterior-mean $\hat{x}$ of NTU.

bias, which is what we observed for this example.

2.6 Discussion

We have proposed a fully Bayesian and a semi-Bayesian approach for handling a semiparametric mixed model with error-in-covariates for matched case-control studies. These approaches are developed using low-rank thin-plate splines and a latent variable probit model. The strength of these methods is that they can handle both error-in-covariates and explain nonlinear relationships between matched binary outcomes and covariates measured with error. Additionally, we have shown these methods exhibit some robustness to model misspecification of x and u. Based on our knowledge, there is no existing methodology that has been shown to do these in matched case-control studies. We also proposed two EM approaches. However, these critically rely on the

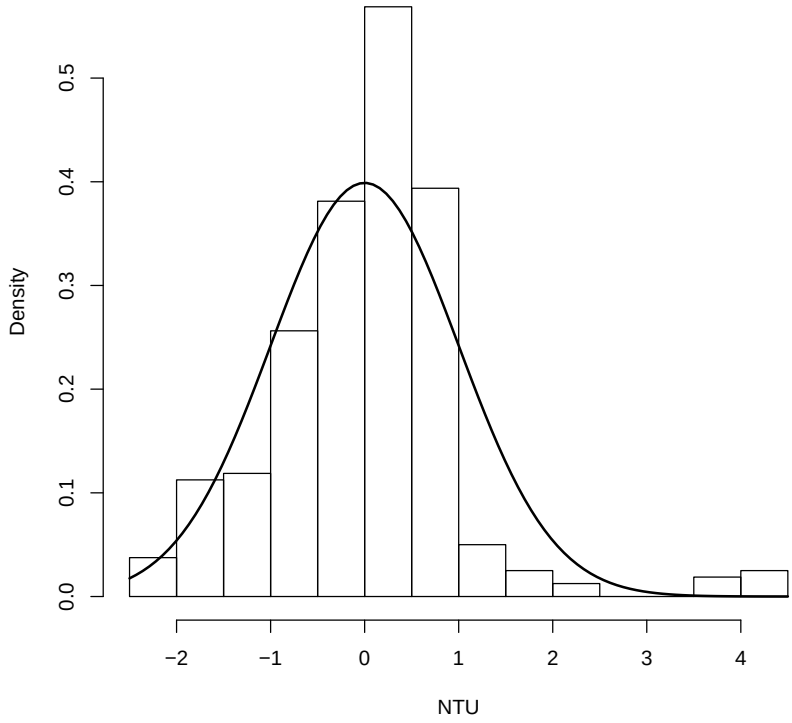

Figure 2.4: Histogram of the centered and scaled values of x, Nephelometric Turbidity Units (NTU), from the aseptic meningitis data. The black line is a fit from a normal distribution. The measurement for NTU are somewhat symmetric and are unimodal, but normality does not hold.

choice of σ_k^2 and σ_q^2, and thus were not investigated with the same detail as the semi- and fully Bayesian approaches. We conjecture that instead of choosing σ_k^2 and σ_q^2, using a regularized estimator of each could reduce these issues. Based on some preliminary simulations, omitted here, we did find mild improvement. However, not enough for us to recommend and investigate further.

The fully-Bayesian approach treats x as a latent variable and then integrates it out. The semi-Bayesian approach uses a first order Laplace approximation to the likelihood, marginalizing out x. We have shown that the fully Bayesian approach is at least as good as the semi-Bayesian approach, and in some cases superior, in terms of reducing MSE. However, this is true of reducing mean bias, where the best approach appears to depend on the true function $m(x)$.

We note that our approach was developed for the univariate x. Our approach can be generalized for several covariates x measured with error into an additive model,

$$m^*(\tilde{X},Z)=\sum_{r_1=1}^{R_1} m_{r_1}(x_{r_1})+\sum_{r_2=1}^{R_2} m_{r_2}(z_{r_2}),$$

where there are R_1 covariates measured with error and R_2 covariates measured without error. Generalization to a nonadditive model will be an interesting and challenging problem because of the unknown interaction structure among unknown covariates. We illustrated our technique using low-rank thin-plate splines, however, it is straight forward to change the spline basis to any other where the smoothness penalty can be thought of as a $N(0,\sigma^2_\beta)$ prior on β_L the spline coefficients (Ruppert et al., 2003).

We assumed that the measurement error u was additive and normally distributed. This assumption creates a computational convenience, as we can choose an inverse-gamma conjugate prior for σ^2_u so that it can be sampled in a Gibbs step. If we change the distributional assumption on u, this convenience will be lost. More complicated measurement error distributions that may depend on x or Y are worthwhile future research problems. Another choice of computational convenience was latent variable probit approach. We were then able to place conjugate normal priors on β_x, β_z, β_L and $q(S)$ and sample them using Gibbs steps. Changing the link function would also remove this convenience.

Finally, we assumed the distribution of x was normal. In practice this might not be the case, and was not the case for our data analysis example. We showed the fully and semi-Bayesian methods were somewhat robust to violations of this assumption, as well as assumptions about the distribution of u. However, flexible methods for properly modeling the distribution of x and u should improve performance of Bayesian error-in-covariates models.

Acknowledgments

We would like to thank Pang Du, Leanna House, Scotland Leman, George Terrell, and Matt Williams for their advice and assistance. We would also like to thank Ho Kim for

supplying the aseptic meningitis data.

Bibliography

Agresti, A. (2002), *Categorical Data Analysis*, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., 2nd ed.

Albert, J. and Chib, S. (1993), "Bayesian-Analysis of Binary and Polytochtomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Berry, S. M., Carroll, R. J., and Ruppert, D. (2002), "Bayesian smoothing and regression splines for measurement error problems," *Journal of the American Statistical Association*, 97, 160–169.

Buzas, J. S. and Stefanski, L. A. (1996), "A note on corrected-score estimation," *Statistics & Probability Letters*, 28, 1–8.

Carroll, R., Roeder, K., and Wasserman, L. (1999), "Flexible parametric measurement error models," *Biometrics*, 55, 44–54.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006), *Measurement Error in Nonlinear Models: A Modern Perspective*, Monographs on Statistics and Applied Probability, Boca Raton, FL: Chapman and Hall/CRC, 2nd ed.

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

Geman, S. and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Guolo, A. (2008), "A Flexible Approach to Measurement Error Correction in Case-Control Studies," *Biometrics*, 64, 1207–1214.

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

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109.

Hosmer, Jr., D. W. and Lemeshow, S. (2000), *Applied Logistic Regression*, Wiley Series in Probability and Statistics, Hoboken, NJ: John Wiley & Sons, Inc., 2nd ed.

McShane, L., Midthune, D., Dorgan, J., Freedman, L., and Carroll, R. (2001), "Covariate measurement error adjustment for matched case-control studies," *Biometrics*, 57, 62–73.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equation of State Calculations by Fast Computing Machines," *The Journal of Chemical Physics*, 21, 1087–1092.

Parker, P. A., Vining, G. G., Wilson, S. R., Szarka, III, J. L., and Johnson, N. G. (2010), "The Prediction Properties of Classical and Inverse Regression for the Simple Linear Calibration Problem," *Journal of Quality Technology*, 42, 332–347.

Peleg, A. Y., Husain, S., Qureshi, Z. A., Silveira, F. P., Sarumi, M., Shutt, K. A., Kwak, E. J., and Paterson, D. L. (2007), "Risk factors, clinical characteristics, and outcome of nocardia infection in organ transplant recipients: A matched case-control study," *Clinical Infectious Diseases*, 44, 1307–1314.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge series on statistical and probabilistic mathematics, New York: Cambridge University Press.

Stefanski, L. A. and Carroll, R. J. (1987), "Conditional Scores and Optimal Scores for Generalized Linear Measurement-Error Models," *Biometrika*, 74, 703–716.

Tester, J., Rutherford, G., Wald, Z., and Rutherford, M. (2004), "A matched case-control study evaluating the effectiveness of speed humps in reducing child pedestrian injuries," *American Journal of Public Health*, 94, 646–650.

Tierney, L. and Kadane, J. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86.

Whitney, C. G., Pilishvili, T., Farley, M. M., Schaffner, W., Craig, A. S., Lynfield, R., Nyquist, A.-C., Gershman, K. A., Vazquez, M., Bennett, N. M., Reingold, A., Thomas, A., Glode, M. P., Zell, E. R., Jorgensen, J. H., Beall, B., and Schuchat, A. (2006), "Effectiveness of seven-valent pneumococcal conjugate vaccine against invasive pneumococcal disease: a matched case-control study," *Lancet*, 368, 1495–1502.

2.A Laplace Approximation

2.A.1 First order Laplace Approximation for Semi-Bayesian Approach

The goal of this section is to show using first order Laplace approximation that:

$$\int L(Y,W|l,x,Z,S,\beta,\sigma_u^2)\,\mathrm{d}x \approx L(Y|l,x=\bar{w},Z,S,\beta).$$

Note that:

$$\begin{aligned}
L(Y_{ij},W_{ij}|l_{ij},x_{ij},Z_{ij},S,\beta,\sigma_u^2) &= L(Y_{ij}|l_{ij},x_{ij},Z_{ij},S_i,\beta)\times N(W_{ij};x_{ij},\sigma_u^2)\\
&= L(Y_{ij}|l_{ij},x_{ij},Z_{ij},S_i,\beta)(2\pi\sigma_u^2)^{K_{ij}/2}\\
&\quad\times\exp[-(W_{ij}-x_{ij})^T(W_{ij}-x_{ij})/(2\sigma_u^2)].
\end{aligned}$$

We will write $A(x_{ij}) = L(Y_{ij}|l_{ij},x_{ij},Z_{ij},\beta)(2\pi\sigma_u^2)^{-K_{ij}/2}$ and $h(x_{ij}) = (W_{ij}-x_{ij})^T(W_{ij}-x_{ij})/(2\sigma_u^2)$. It is easy to show $h(x_{ij})$ has unique maximum $\bar{w}_{ij\cdot}$, since $h(\cdot)$ is a quadratic form, and that the second derivative $h''(x_{ij})=1/\sigma_u^2$, both for all ij. Tierney and Kadane (1986) show that:

$$\int A(x_{ij})\exp[-h(x_{ij})]\,\mathrm{d}x_{ij} \approx A(\tilde{x})\exp[-h(\tilde{x})]\sqrt{\frac{2\pi}{K_{ij}h''(\tilde{x})}},$$

where $\tilde{x}$ is the value that maximizes $h(\cdot)$. We then get:

$$\begin{aligned}
\int A(x_{ij})\exp[-h(x_{ij})]\,\mathrm{d}x_{ij} &\approx L(Y_{ij}|l_{ij},\bar{w}_{ij},Z_{ij},S_i,\beta)\left[K_{ij}(2\pi\sigma_u^2)^{K_{ij}-1}\right]^{-1/2}\\
&\quad\times\exp[-(W-\bar{w})^T(W-\bar{w})/(2\sigma_u^2)]\\
&\propto L(Y_{ij}|l_{ij},\bar{w}_{ij},Z_{ij},S_i,\beta).
\end{aligned}$$

It is clear then that:

$$\begin{aligned}
\int L(Y,W|l,x,Z,S,\beta,\sigma_u^2)\,\mathrm{d}x &= \prod_{i=1}^{N}\prod_{j=1}^{M+1}\int L(Y_{ij},W_{ij}|l_{ij},x_{ij},Z_{ij},S_i,\beta,\sigma_u^2)\,\mathrm{d}x_{ij}\\
&\approx \prod_{i=1}^{N}\prod_{j=1}^{M+1}\Big\{L(Y_{ij}|l_{ij},\bar{w}_{ij},Z_{ij},S_i,\beta)\left[K_{ij}(2\pi\sigma_u^2)^{K_{ij}-1}\right]^{-1/2}\\
&\quad\times\exp[-(W_{ij}-\bar{w}_{ij})^T(W_{ij}-\bar{w}_{ij})/(2\sigma_u^2)]\Big\}\\
&\propto \prod_{i=1}^{N}\prod_{j=1}^{M+1}L(Y_{ij}|l_{ij},\bar{w}_{ij},Z_{ij},\beta)\\
&= L(Y|l,\bar{w},Z,\beta).
\end{aligned}$$

2.A.2 First Order Laplace for E2 Approach to E-step

Consider now where the goal is to use first order Laplace approximation to find:

$$E\{\log[L(Y|x,Z,S,\beta)]\} = \int \log[L(Y|x,Z,S,\beta)]N(W;x,\sigma_u^2)N(x;\mu_x,\sigma_x^2)\,\mathrm{d}x,$$
$$\approx \log[L(Y|\tilde{x},Z,S,\beta)],$$
$$\tilde{x}_{ij} = \frac{K_{ij}\bar{w}_{ij.}\sigma_x^2 + \mu_x\sigma_u^2}{K_{ij}\sigma_x^2 + \sigma_u^2}.$$

We can rewrite the integration as follows:

$$\begin{aligned}
&\int \log[L(Y_{ij}|x_{ij},Z_{ij},S_i,\beta)]N(W_{ij};x_{ij},\sigma_u^2)N(x_{ij};\mu_x,\sigma_x^2)\,\mathrm{d}x_{ij} \\
&\quad = \int \log[L(Y_{ij}|x_{ij},Z_{ij},S_i,\beta)](2\pi\sigma_u^2)^{-K_{ij}/2} \\
&\quad \times \exp\left[-0.5\sum_{k=1}^{K_{ij}}(w_{ijk}-x_{ij})^2/\sigma_u^2\right] \\
&\quad \times (2\pi\sigma_x^2)^{-1/2}\exp\left[-0.5(x_{ij}-\mu_x)^2/\sigma_x^2\right]\,\mathrm{d}x_{ij} \\
&\quad \approx \log[L(Y|\tilde{x},Z,S,\beta)], \\
&\quad = \int A(x_{ij})\exp\left[-h(x_{ij})\right]\,\mathrm{d}x_{ij},
\end{aligned}$$

where:

$$A(x_{ij}) = \log[L(Y_{ij}|x_{ij},Z_{ij},S_i,\beta)](2\pi\sigma_u^2)^{-K_{ij}/2}(2\pi\sigma_x^2)^{-1/2}, \quad \text{and}$$
$$h(x_{ij}) = \frac{\sum_{k=1}^{K_{ij}}(w_{ijk}-x_{ij})^2 + (x_{ij}-\mu_x)^2}{2\sigma_x^2\sigma_x^2}.$$

It should be clear since $h(x_{ij})$ is the sum of two quadratic functions of x_{ij}, that the unique maximum of $h(\cdot)$ is the Bayes estimator $\tilde{x}_{ij} = \frac{K_{ij}\bar{w}_{ij.}\sigma_x^2+\mu_x\sigma_u^2}{K_{ij}\sigma_x^2+\sigma_u^2}$. Also the second derivative $h''(x_{ij}) = \frac{1}{\sigma_x^2\sigma_u^2}$. It follows then that:

$$\begin{aligned}
\int A(x_{ij})\exp\left[-h(x_{ij})\right]\,\mathrm{d}x_{ij} &\approx A(\tilde{x}_{ij})\exp\left[-h(\tilde{x}_{ij})\right]\sqrt{\frac{2\pi\sigma_u^2\sigma_x^2}{K_{ij}+1}}, \\
&\propto \log[L(Y_{ij}|\tilde{x}_{ij},Z_{ij},S_i,\beta)].
\end{aligned}$$

It follows from a similar argument as in 2.A.1 that:

$$E\{\log[L(Y|x,Z,S,\beta)]\} \approx \log[L(Y|\tilde{x},Z,S,\beta)].$$

2.B Impact of Prior Selection

Due to sensitivity of the choice of prior for σ_q^2 and σ_β^2 on the posterior of $q(S)$ and of β_L, there may be some concern that the preferred approach for handling measurement error, the semi-Bayesian approach or fully Bayesian approach, may be sensitive to this as well. Namely, the choice of A_β, B_β, A_q and B_q, when the inverse-gamma prior is chosen, as in Section 2.3.1. To investigate the possible effect, we conduct a small simulation study using two sets 10 simulated 1-4 matched case control studies. The first set has $N = 25$ and the second has $N = 100$. The data were simulated as in Section 2.4.1 with $\sigma_u = 0.5$, $\sigma_q^2 = 0.05$ however, instead of choosing $m(x)$ to be a quadratic, sine, or some other known function, we chose $m(x) = -0.5x + L_p(x)\beta_L$, where $p = 2$, using knots $(\xi_1, ..., \xi_5)$ chosen to be the quantiles of $\bar{w}$ and the spline coefficients $\beta_L = (-0.1403, -0.1742, 0.0569, 0.4927, -0.0868)$ were generated from $N(0, \sigma_\beta^2 = 0.1)$. In Figure 2.5 we see that this choice of $m(x)$ looks nearly quadratic with slight bumps near zero. However, it does have very large (negative) tails, which will not bit fit well by a probit link. The four priors we investigated were chosen as follows:

P1: Weak prior: $A_\beta = B_\beta = A_q = B_q = 0.1$. Infinite prior mean and variance. Mass near 0.09.

P2: Strong prior, poorly chosen, mass too large: $A_\beta = A_q = 3$ and $B_\beta = B_q = 2$. Prior mean and variance of 1.

P3: Strong prior, well chosen: $A_\beta = A_q = 3$, $B_\beta = 2\sigma_\beta^2$, and $B_q = 2\sigma_q^2$. Prior means of σ_β^2 and σ_q^2, prior variances are the means squared.

P4: Strong prior, poorly chosen, mass too small: $A_\beta = A_q = 3$, $B_\beta = 2\sigma_\beta^2/10$, and $B_q = 2\sigma_q^2/10$. Prior means of $\sigma_\beta^2/10$ and $\sigma_q^2/10$, prior variances are the means squared.

We compute the MMSE and MMB as described in Section 2.4 and the results are displayed in Table 2.6. We can see that for priors all priors except P4, the fully Bayesian approach is preferred for reducing MMB and MMSE, in both small and large sample cases. For prior P4, the semi-Bayesian approach appears to be preferred for both MMSE

and MMB. In the small sample $N = 25$ case, the choice of prior did have some impact on model estimation. However, for $N = 100$, the choice of prior on σ_q^2 and σ_β^2 did not impact performance of any method. Figures 2.7 and 2.6 compare the posterior mean estimate for $m(x)$ under each prior for the semi- and fully Bayesian methods, respectively, when $N = 100$. There is very little discernible difference.

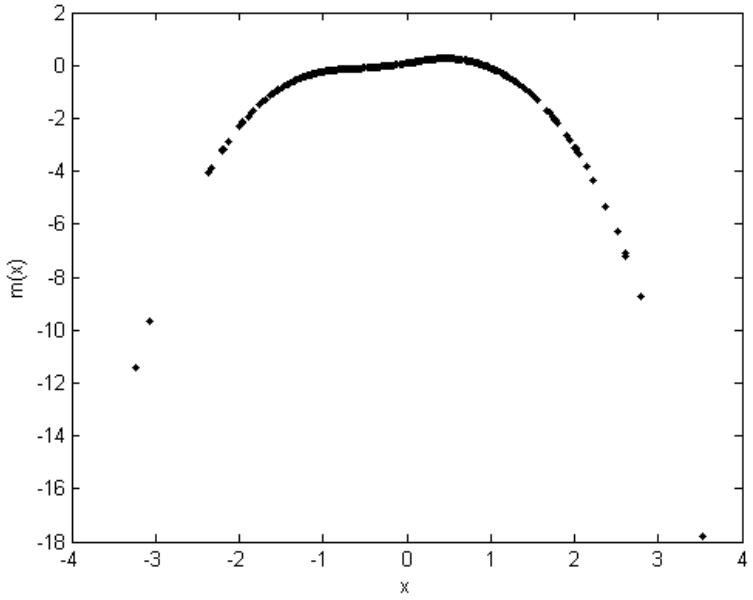

Figure 2.5: Scatterplot of $m(x) = -0.5x + L_p(x)\beta_L$, where $p = 2$ and $\beta_L = (-0.1403, -0.1742, 0.0569, 0.4927, -0.0868)$, used for evaluating the robustness of the semi- and fully Bayesian approaches to the choice of prior on σ_q^2 and σ_β^2.

2.C EM Performance Issue

Our EM approach is essentially a maximum likelihood approach, since the E-steps E1 and E2 do not dependent on each iteration of the M-step. As a result, issues dominating its performance are related to the M-step, and picking initial, and the model itself. The performance of all approaches, semi- and fully Bayesian, as well as EM approaches are dependent on the variance parameters σ_β^2 and σ_q^2. For all EM simulations, a single value was chosen based on the fully Bayesian approach, and kept constant within each simulation setting. It's possible that in the simulations where $m(x) = \sin(\pi x/2)$

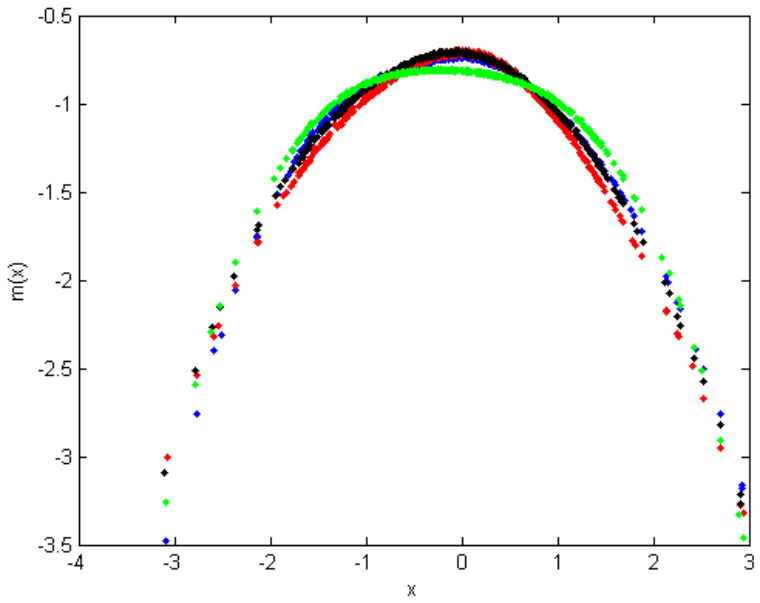

Figure 2.6: Scatterplot of the posterior mean of $m(x) = \beta_0 + x\beta_1 + L_p(x)\beta_L$, evaluated at the posterior mean $x_{ij} = \hat{x}_{ij}$ for the fully Bayesian method, where prior P1 is in blue, prior P2 is in red, prior P3 is in black, and prior P4 is in green.

this was too restrictive. Currently better ways of selecting or updating σ_β^2 and σ_q^2 for both EM and Bayesian methods are desirable. Initial values for parameters updated in the M-step were randomly generated on a small range. It's possible then that we have converged on local modes due to poor initial values. However, the next best option for selection of initial values is, say, the posterior mean of the semi-Bayesian approach. In robust simulations, there were heavy convergence issues due to nearly-singular Hessian matrices. This is a result of the outlying x-values when $x \sim \chi^2$.

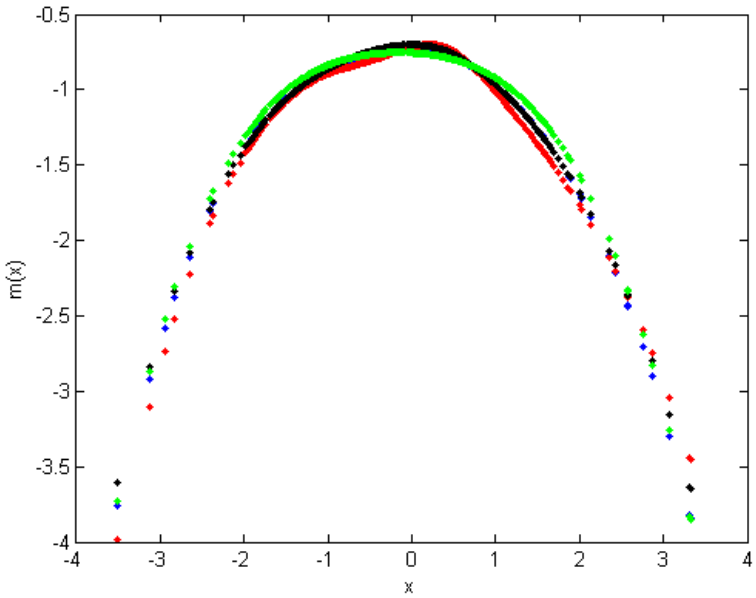

Figure 2.7: Scatterplot of the posterior mean of $m(x) = \beta_0 + x\beta_1 + L_p(x)\beta_L$, evaluated at $x_{ij} = \bar{w}_{ij}$ for the semi-Bayesian method, where prior P1 is in blue, prior P2 is in red, prior P3 is in black, and prior P4 is in green.

N	Prior		SB	FB
25	P1	MMB	0.0337	**0.0182**
		MMSE	0.0917	**0.0868**
	P2	MMB	0.0316	**0.0104**
		MMSE	0.1133	**0.0990**
	P3	MMB	0.0182	**0.0097**
		MMSE	0.0739	**0.0711**
	P4	MMB	0.0193	**0.0139**
		MMSE	**0.0435**	0.0467
100	P1	MMB	0.0157	**0.0127**
		MMSE	0.0572	**0.0555**
	P2	MMB	0.0203	**0.0104**
		MMSE	0.0622	**0.0589**
	P3	MMB	0.0180	**0.0148**
		MMSE	0.0574	**0.0561**
	P4	MMB	**0.0140**	0.0190
		MMSE	**0.0522**	0.0540

Table 2.6: The MMB and the MMSE of the linear predictors $\hat{\eta}^{(\cdot)}$ for assessing robustness of the semi-Bayesian method (SB) and the fully Bayesian method (FB) to the choice of prior on σ_β^2 and σ_q^2, where P1 is a weakly informative prior, P2 is a strongly informative prior centering mass above the correct value, P3 is a strongly informative prior centering mass at the correct value, and P4 is a strongly informative prior centering mass below the correct value.

CHAPTER 3

GENERALIZED LINEAR MODELS WITH COVARIATE MEASUREMENT ERROR AND UNKNOWN LINK FUNCTION

Abstract

Generalized linear model (GLM) with error in covariates are often useful in epidemiological research using human, animal, and other subjects. The link function is chosen by the user depending on the type of outcomes. It is often chosen to be the canonical link function such as logit or log for binary and count outcomes, respectively. However, when covariates are measured with error, incorrect inference can be made, compounded by incorrect choice of link function. In this article we propose two flexible approaches for handling error-in-covariates and estimating an unknown link simultaneously. The first approach is uses a fully Bayesian hierarchal framework, treating the unobserved covariate as a latent variable to be integrated over. The second is a semi-Bayesian approach using a Laplace approximation to marginalize the variables measure with error out of the likelihood. Our simulation results show support that the fully Bayesian approach is often a better choice than the semi-Bayesian approach for adjusting for measurement error. The usefulness of our approaches is demonstrated through its applications with binary and count outcomes.

3.1 Introduction

Generalized Linear Models (GLM) (McCullagh and Nelder 1989) are a class of linear regression models for data where the response y is distributed according to a member of the exponential family of distributions. This is a rich class of regression models capable of handling a wide variety of response variables. We could write the model as follows:

$$\begin{aligned} y_i &\sim \text{Exponential Family}[g(\mu_i), \phi], \\ E(y_i) = \mu_i &= g^{-1}(\eta_i), \\ \eta_i &= \beta_0 + x_i^* \beta, \end{aligned}$$

where $g(\cdot)$ is the link function, ϕ the dispersion parameter, X^* is an $N \times p$ matrix of p covariates, β is the $p \times 1$ vector of their coefficients, and β_0 is the intercept.

In error-in-covariables problems, we say that X^* can be subset into X, the p_1 covariates measured with error, and Z, the p_2 covariates measured without error. Similarly we subset β into β_x and β_z. One approach to handling measurement error is to assume the "classical" measurement error model. That is, instead of observing X we observe $W = X + U$, where U is the measurement error. When X is are continuous covariates, the classical measurement error model is often assumed, $U \sim N(0, \Sigma_u)$, which is what we address in this paper. If Σ_u is unknown, replicated measurements w_{ijk}, for $j = 1, 2, ..., p_1$, $k = 1, 2, ..., K_{ij}$, of x_{ij} are necessary to identify Σ_u, where $K_{ij} > 1$ for some i for every j (depending on the structure of Σ_u). In a Bayesian framework, a very informative prior on Σ_u can be used to create identifiability if $K_{ij} = 1$ for all i for some j. Also in the Bayesian framework, X is treated as a latent variable that is integrated over. The joint likelihood of (Y, W) is:

$$L(Y, W|X, Z, \beta, \phi, g, \Sigma_u) \propto \prod_{i=1}^{N} f(y_i; x_i, z_i, \beta, \phi, g) N(W_i; x_i, \Sigma_u),$$

where f is the density or mass function for the distribution of Y.

The function $g^{-1}(\cdot)$ is called the inverse-link function, but sometimes it is also called the mean function since it maps the linear predictor η_i onto the domain of μ_i, Ω. Using a link function $g(\cdot)$ is advantageous as Ω is not the same for all members of the exponential

family and it allows for unrestricted estimation of η_i, while maintaining μ_i in the proper domain given the distribution of y_i. Traditionally the link function $g(\cdot)$ is chosen by the user, often to be the canonical link function, e.g. the logit function for $y_i \sim$ Bernoulli and the natural log function for $y_i \sim$ Poisson. The canonical link function is not guaranteed to map μ_i onto $\Re$, e.g., the inverse function for $y_i \sim$ Gamma. In these cases another function that meets this requirement is often chosen instead, e.g., natural log function for $y_i \sim$ Gamma. The choice of link function can be a nontrivial one and it affects how well the data are fit by the model, which is exacerbated by covariate measurement error. In order to improve the flexibility of our model to fit the data we treat the link function as unknown and estimate it nonparametrically.

There have been many Bayesian and frequentist papers addressing covariate measurement error for a variety of GLM and related models. For a thorough review of many Bayesian and frequentist methods for generalized linear models and others, see the text Carroll et al. (2006). Some other interesting methodology has been developed by Müller and Roeder (1997) for case-control studies using mixtures of Dirichlet processes (Escobar and West 1995). A more recent variation of their methodology can be found in Sinha et al. (2010), who also incorporate smoothing spline methodology (Ruppert et al. 2003). However, the methodology in both these papers use what's called the instrumental variable model (Buzas and Stefanski 1996; Carroll et al. 2004), rather than the classical model. Mallick and Gelfand (1996) develop a method for handling unknown link function when there is classical measurement error in the response Y and in the predictors X. They do so by extending the methods of Mallick and Gelfand (1994) for handling unknown link functions in GLMs. However, they take a calibration approach and rely on strong prior assumptions to identify the measurement error variance since they do not use replicated measurements of X. For our approach, we also extend the method of Mallick and Gelfand (1994) for unknown link function estimation to handle covariate measurement error. We propose two methods based on a study design with replicated measurements of X. The first, uses a Bayesian hierarchical framework to simultaneously model the unknown link function and the covariate measurement error, treating X as a latent variable and integrating over it. The second uses a semi-Bayesian approach which

marginalizes X out of the likelihood. These approaches do not rely on *a priori* knowledge of the amount of measurement error. We assess the performance of each approach under both correctly and incorrectly specified models.

The remainder of this chapter is organized as follows: In Section 3.2 we describe a generalized linear model with unknown link function and error in covariates. In Section 3.3 we develop the Bayesian hierarchical models used in the semi- and fully Bayesian approaches. In Section 3.4 we compare the performance of these two approaches both when the model is correctly and incorrectly specified. In Section 3.5 we illustrate each method on two real datasets. Finally, in Section 3.6 we have our concluding remarks.

3.2 Generalized Linear Model with Error in Covariates and Unknown Link Function

Our GLM with error in covariables could be written as follows:

$$\begin{aligned} y_i &\sim \text{Exponential Family}[g(\mu_i), \phi], \\ E(y_i) = \mu_i &= g^{-1}(\eta_i), \\ \eta_i &= \beta_0 + x_i\beta_x + z_i\beta_z, \\ w_{ijk} &= x_{ij} + u_{ijk}, \\ U_k &\sim N(0, \Sigma_u). \end{aligned}$$

To estimate the unknown link function $g(\cdot)$, we use the method of Mallick and Gelfand (1994). First we need a strictly increasing differentiable transformation T from Ω into $(0,1)$. Then $J(\eta) = T\left[g^{-1}(\eta)\right]$ is a distribution function. Hence, modeling $g(\cdot)$ is the equivalent to modeling a distribution function. A finite mixture of beta distribution functions (i.e., regularized incomplete beta functions) provides a rich class of models for approximating distribution functions. In order to do this we select a baseline link function $g_0(\cdot)$ and set $J_0(\eta) = T\left[g_0^{-1}(\eta)\right]$, so now:

$$J(\eta) \approx \sum_{r=1}^{R} \omega_r \text{IB}\left[J_0(\eta); a_r, b_r\right],$$

where ω_r is the weight for mixand r, where $\sum_{r=1}^{R} \omega_r = 1$, $\text{IB}(\cdot; a_r, b_r)$ is the distribution function for beta distribution with parameters a_r and b_r, and R is the number of mixands, usually chosen to be 3 or 4. To find $g^{-1}(\eta)$, we compute $T^{-1}\left[J(\eta)\right]$. It should be noted that g is not identifiable for linear predictors with unknown intercept β_0, so β_0 is set to zero.

Combining the measurement error portion of the model with the unknown link function portion of the model, the joint likelihood of (Y, W) becomes:

$$
\begin{aligned}
L(Y, W|X, Z, \beta, \phi, \omega, \Sigma_u) &\propto \prod_{i=1}^{N} f(y_i; x_i, z_i, \beta, \phi, \omega) N(W_i; x_i, \Sigma_u), \\
g^{-1}(\eta) &= T^{-1}\left[J(\eta)\right], \\
J(\eta) &\approx \sum_{r=1}^{R} \omega_r \text{IB}\left[J_0(\eta); a_r, b_r\right], \\
\sum_{r=1}^{R} \omega_r &= 1, \\
\eta &= X\beta_x + Z\beta_z,
\end{aligned}
$$

where f is the density or mass function for the distribution of Y.

3.3 Bayesian Hierarchical Model

For Bayesian inference we describe our selection of priors in Section 3.3.1. In Section 3.3.2 we describe how to perform inference on the posterior using Markov chain Monte Carlo (MCMC). We also describe a semi-Bayesian approach in Section 3.3.3.

3.3.1 Prior Selection

Our approach for selecting priors is to use common default choices where relevant. That is, normal priors for mean-like parameters, inverse-Wishart priors for covariance matrices, inverse-gamma priors for variance-link parameters, and Dirichlet priors weighting parameters. For the generalized linear model portion of the parameters, we select $(\beta_x) \sim N(g_x, t_x^2)$, $(\beta_z) \sim N(g_z, t_z^2)$, and $(\phi) \sim \text{IG}(A_\phi, B_\phi)$.

For the measurement error portion of the model we select this common hierarchical structure: $(X|\mu_x,\Sigma_x) \sim N(\mu_x,\Sigma_x)$, $(\mu_x) \sim N(g_\mu, t_\mu^2)$, $(\Sigma_x) \sim \text{IW}(S_x, d_x)$, and $(\Sigma_u) \sim \text{IW}(S_u, d_u)$. When $p_1 = 1$, then $\Sigma_x = \sigma_x^2$ and $\Sigma_u = \sigma_u^2$, and we select $(\sigma_x^2) \sim \text{IG}(A_x, B_x)$ and $(\sigma_u^2) \sim \text{IG}(A_u, B_u)$. For $p_1 > 1$ it still may be reasonable to use the inverse-gamma distribution as the prior and make the assumption that cov(U) or cov(X) are diagonal matrices. We will make this assumption for the rest of the paper. In practice, it is important to select the prior for X such that it is close to the actual distribution of X.

And for the link function estimation, we select $(\omega) \sim \text{Dir}(c)$, where c is the $R \times 1$ vector of concentration parameters. A prior is not placed on R since 3 or 4 mixands is simple to implement and flexible enough to cover most situations.

3.3.2 Posterior Sampling via MCMC

To sample the joint posterior we use Metropolis-Hastings (Metropolis et al. 1953 and Hastings 1970), Gibbs (Geman and Geman 1984), and Metropolis-within-Gibbs algorithms (Mallick and Gelfand 1994). These methods require the knowledge of the conditional distribution for each parameter in the model. The full conditionals for this model are as follows:

$$
\begin{aligned}
(\beta_x, \beta_z|-) &\propto \prod_{i=1}^{N} f(y_i; x_i, z_i, \beta, \phi, \omega) \times N\left(\beta_x; g_x, t_x^2\right) N\left(\beta_z; g_z, t_z^2\right), \\
(\phi|-) &\propto \prod_{i=1}^{N} f(y_i; x_i, z_i, \beta, \phi, \omega) \times \text{IG}(\phi; A_\phi, B_\phi), \\
(\omega|-) &\propto \prod_{i=1}^{N} f(y_i; x_i, z_i, \beta, \phi, \omega) \times \text{Dir}(\omega; c), \\
(x_{ij}|-) &\propto f(y_i; x_i, z_i, \beta, \phi, \omega) N\left(w_{ij}; x_{ij}, \sigma_{u_j}^2 I\right) N\left(x_{ij}; \mu_{xj}, \sigma_{xj}^2\right), \\
(\mu_{xj}|-) &\sim N\left(\mu_{xj}; \frac{N\bar{x}_{\cdot j} t_\mu^2 + g_\mu \sigma_{xj}^2}{N t_\mu^2 + \sigma_{xj}^2}, \frac{\sigma_{xj}^2 t_\mu^2}{N t_\mu^2 + \sigma_{xj}^2}\right), \\
(\sigma_{xj}^2|-) &\sim \text{IG}\left[\sigma_{xj}^2; N/2 + A_x, \sum_{i=1}^{N} (x_{ij} - \mu_{xj})^2/2 + B_x\right], \\
(\sigma_{uj}^2|-) &\sim \text{IG}\left[\sigma_{xj}^2; \sum_{i=1}^{N} K_{ij}/2 + A_u, \sum_{i=1}^{N}\sum_{k=1}^{K_{ij}} (w_{ijk} - x_{ij})^2/2 + B_u\right].
\end{aligned}
$$

The parameters μ_x, σ_x^2, and σ_u^2 can be sampled in Gibbs steps since their full conditionals are closed form. The rest of the parameters, β, ϕ, ω, and X, must be updated in Metropolis-Hastings steps. Mallick and Gelfand (1994) suggest using Metropolis-within-Gibbs instead of Metropolis-Hastings to improve mixing since ω mixes very slowly. The Metropolis-within-Gibbs algorithm we use is as follows:

- Step 1: Initialize $\left[\beta_x^{(1)}, \phi^{(1)}, w^{(1)}, x_{ij}^{(1)}, \sigma_{xj}^{2(1)}, \sigma_{uj}^{2(1)}\right]$.
- Step 2: At the t_1th iteration
  - Step 2.1: Gibbs steps for $\mu_x^{(t_1)}$, $\sigma_x^{2\,(t_1)}$, and $\sigma_u^{2\,(t_1)}$;
    - Step 2.1.1: Draw samples from $\left(\mu_x^{(t_1)}|-\right)$, $\left(\sigma_x^{2\,(t_1)}|-\right)$, and $\left(\sigma_u^{2\,(t_1)}|-\right)$.
  - Step 2.2: Initialize $\beta_*^{(1)} = \beta^{(t_1-1)}$, $\phi_*^{(1)} = \phi^{(t_1-1)}$, $\omega_*^{(1)} = \omega^{(t_1-1)}$, and $X_*^{(1)} = X^{(t_1-1)}$.
  - Step 2.3: Metropolis-Hastings steps for $\beta_*^{(t_2)}$, $\phi_*^{(t_2)}$, $\omega_*^{(t_2)}$, and $X_*^{(t_2)}$;
    - Step 2.3.1: At the t_2 iteration, Draw samples from $\left(\beta_*^{(t_2)}|-\right)$, $\left(\phi_*^{(t_2)}|-\right)$, $\left(\omega_*^{(t_2)}|-\right)$, and $\left(X_*^{(t_2)}|-\right)$.
    - Step 2.3.2: Repeat Step 2.3.1 T_2 times
  - Step 2.4: Set $\beta^{(t_1)} = \beta_*^{(T_2)}$, $\phi^{(t_1)} = \phi_*^{(T_2)}$, $\omega^{(t_1)} = \omega_*^{(T_2)}$, and $X^{(t_1)} = X_*^{(T_2)}$.
- Step 3: Repeat Step 2, T_1 times

Essentially, we run the MCMC sampler for $t_1 = 1, 2, ..., T_1$ iterations and $t_2 = 1, 2, ..., T_2$ subiterations for each iteration t_1. At each of the t_1 iterations we sample the parameters that can be sampled using Gibbs steps. Also at each iteration t_1, instead of a single Metropolis-Hastings update for each parameter needing Metropolis-Hastings, we make T_2 updates, and keep the sample from the last update T_2 as the sample for iteration t_1.

3.3.3 Semi-Bayesian Approach

In a fully Bayesian approach, there are $N \times p_1$ latent variables x_{ij} that need to be updated via MCMC, which can be time consuming computationally for large N or p_1. One possible solution is instead of using the full likelihood, use the marginal likelihood:

$$\int L(Y, W|X, Z, \beta, \phi, \omega, \sigma_u^2) \mathrm{d}X = L(Y, W|Z, \beta, \phi, \omega, \sigma_u^2).$$

However, this integration is intractable, so we use a first order Laplace approximation (Tierney and Kadane 1986). The marginal likelihood then becomes:

$$L(Y,W|Z,\beta,\phi,\omega,\sigma_u^2) \propto L(Y|X=\overline{W},Z,\beta,\phi,\omega).$$

That is we get the simple alternative of using $\bar{w}_{ij\cdot} = K_{ij}^{-1}\sum_{k=1}^{K_{ij}} w_{ijk}$ in place of x_{ij} and proceeding as if there were no measurement error. Details in Appendix 3.A. All other aspects of prior selection and posterior sampling of β, ϕ, and ω can be adopted as described in Sections 3.3.1 and 3.3.2.

3.4 Simulation Study

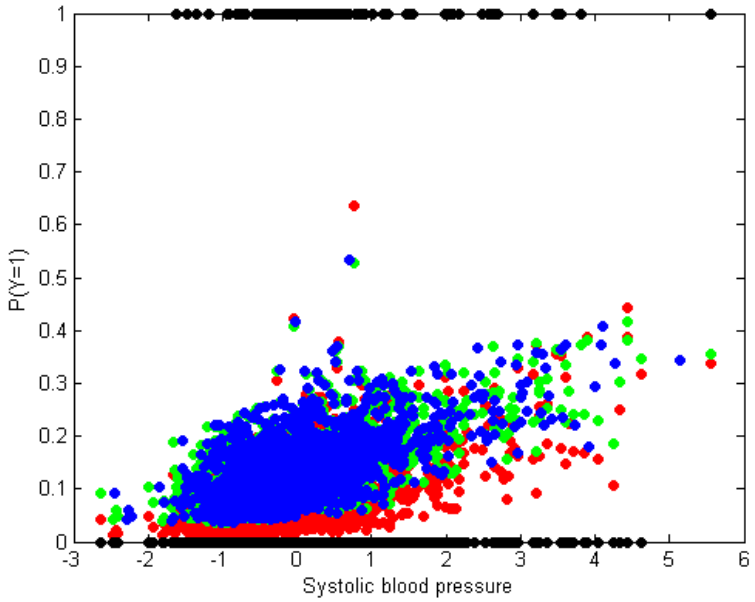

Figure 3.1: Posterior mean fitted values of the Framingham Data. The fitted values using the logit link are in red, evaluated at the average systolic blood pressure measurement $\overline{\text{SBP}} = \bar{w}$, the unknown link using the semi-Bayesian approach is in green evaluated at $\bar{w}$, and the fully Bayesian approach is in blue evaluated at the posterior mean of $x = \text{SBC}$, $\hat{x}$. The black circles are for the indicator of CHD y evaluated at $\bar{w}$.

For our simulations we look at two popular members of the exponential family, the Bernoulli distribution and the Poisson distribution. We address three key comparisons

for these models: (1) whether models with unknown link functions are more heavily impacted by covariate measurement error than models with known link function; (2) which of the proposed fully Bayesian and semi-Bayesian methods for adjusting for covariate measurement error performs better when our model has been properly specified; and (3) which of the two proposed methods performs better under model misspecification of the distribution of X and of U.

All simulations share the following settings:

- We set: $\phi = 1$, $p_1 = p_2 = 1$, $\beta_x = 0.5$, $\beta_z = -0.5$, $N = 100$, $\{A_u, B_u, A_x, B_x\} = 0.1$, $\{g_\mu, g_x, g_z\} = 0$,$a_r = \lambda r$, $b_r = \lambda(R - r + 1)$, $\lambda = 1$, and c is a vector of ones.

- For the Bernoulli cases we set: $\{t_\mu^2, t_x^2, t_z^2\} = 2.5^2$, $R = 3$, $g_0(\cdot)$ as the logit link, and $T(\cdot)$ as the identity function.

- For the Poisson cases we set: $\{t_\mu^2, t_x^2, t_z^2\} = 10^2$, $R = 4$, and $g_0(\cdot)$ and $T(\cdot)$ such that $J_0(\eta) = [1 + \exp(-\eta)]^{-1}$ and $T^{-1}[J(\eta)] = \frac{J(\eta)}{1-J(\eta)}$.

For performing Metropolis-Hastings steps, we select the following proposal distributions:

$$\begin{aligned}
\beta^{(t)} &\sim N(\beta^{(t-1)}, S), \\
\omega^{(t)} &\sim \text{Dir}\left[c_0 \omega^{(t-1)}\right], \\
x_i^{(t)} &\sim N\left(x_i^{(t-1)}, 2^2 \sigma_u^{2\,(t)} / K_i\right),
\end{aligned}$$

where S and c_0 are chosen to produce acceptance rates between 0.2 and 0.4. The proposal for x_i was chosen because $2\sqrt{\sigma_u^2 / K_i}$ covers about 95% of the sampling distribution for $\bar{w}_{i\cdot}$.

3.4.1 Impact of Link Function

For the first set of simulations, we see the impact of not accounting for error-in-covariables when estimating an unknown link function. We set $K_{ij} = 1$, for all ij, and simulate $x \sim N(0,1)$ and $z \sim N(0,1)$. We look at two levels of measurement error, $\sigma_u =$

Measurement error?	σ_u	logit	MG
Yes	0.3	178.7221	**177.3832**
	0.7	172.9418	**172.0213**
No	0.3	68.7430	**63.4383**
	0.7	63.4489	**58.3382**

Table 3.1: The $100^2 \times$ MMSE for comparing how the unknown link (MG) versus the logit link is impacted by measurement error, $x_{ij} \equiv w_{ij}$, with $\sigma_u = \{0.3, 0.7\}$, when $y \sim$ Bernoulli. MG is a link defined by our simulation setup with $g^{-1}(\eta) = T^{-1}\left[J(\eta)\right]$ and $J(\eta) \approx \sum_{r=1}^{R} \omega_r \text{IB}\left[J_0(\eta); a_r, b_r\right]$ where $R = 4$, $a_r = \lambda r$, $b_r = \lambda(R - r + 1)$, $\lambda = 1$, $J_0(\eta) = [1 + \exp(-\eta)]^{-1}$, $T^{-1}[J(\eta)] = \eta/(1-\eta)$. The $100^2 \times$ MMSE of the unknown link versus logit link is also shown for these cases with the measurement error removed, $x_{ij} \equiv x_{ij}$.

$\{0.3, 0.7\}$. For the Bernoulli case we simulate Y using the logit link function. And for the Poisson case, we simulate Y using the log link function.

We estimate the model using true link function, and using the unknown link function method, both using the variable measured without error x and the variable measured with error w in place of x. Though in both cases Y is simulated from a linear predictor with $\beta_0 = 0$, we must estimate β_0 when we use the true link function. Doing so makes the estimate of μ comparable across the two methods.

To assess performance for each simulation in this section we compute $\hat{\mu}^{(T,K)}$, $\hat{\mu}^{(T,U)}$, $\hat{\mu}^{(W,K)}$, and $\hat{\mu}^{(W,U)}$, where:

- $\hat{\mu}^{(T,K)}$ is the posterior mean for the methods using true x measured without error with known link function
- $\hat{\mu}^{(T,U)}$ is the posterior mean for the methods using for true x measured without error with unknown link function
- $\hat{\mu}^{(W,K)}$ is the posterior mean for the methods using the data measured with error w with known link function
- $\hat{\mu}^{(W,U)}$ is the posterior mean for the methods using the data measured with error w with unknown link function.

We also compute $\mu = [1 + \exp(-X\beta_x - Z\beta_z)]^{-1}$. Using these we compute the mean mean

squared error (MMSE) of 100 simulated datasets:

$$\text{MMSE} = (100N)^{-1} \sum_{s=1}^{100} \sum_{i=1}^{N} \left(\hat{\mu}_{is}^{(\cdot,\cdot)} - \mu_{is} \right)^2,$$

where $(\cdot,\cdot)$ is either (T,K), (T,U), (W,K), or (W,U).

The goal is to compare the MMSE of the case with unknown link function with measurement error to the MMSE of the case with known link with measurement error and see if one method is more heavily impacted by the measurement error than the other in terms of MMSE.

3.4.2 Comparison of Approaches

To compare performance of the fully Bayesian with semi-Bayesian method, we set $K_{ij} = 2$, for all ij, and simulate $x \sim N(0,1)$ and $z \sim N(0,1)$. We look at two levels of measurement error, $\sigma_u = \{0.3, 0.7\}$. For the Bernoulli case, we consider the class of link functions with inverse defined as $H^{-1}(\eta) = [1+\exp(-\eta)]^{-a}$. We simulate from the cases where $a = \{0.2, 1, 4\}$. For the Poisson distribution we simulate from the log link, and from a link defined by our simulation setup with $R = 4$, $a_r = \lambda r$, $b_r = \lambda(R-r+1)$, $\lambda = 1$, $J_0(\eta) = [1+\exp(-\eta)]^{-1}$, $T^{-1}[J(\eta)] = \eta/(1-\eta)$, and $\omega = \{0.15, 0.35, 0.35, 0.15\}$ and another case where $\omega = \{0.35, 0.15, 0.15, 0.35\}$. The latter two link functions are referred to as MG1 and MG2, respectively.

To assess performance in this simulation we compute the posterior mean for the mean $\hat{\mu}^{(T)}$, $\hat{\mu}^{(SB)}$, and $\hat{\mu}^{(FB)}$, for using the true x measured without error, semi-Bayesian method, and fully Bayesian method, respectively. We make our comparisons based on the MMSE of 100 simulated datasets:

$$\text{MMSE} = (100N)^{-1} \sum_{s=1}^{100} \sum_{i=1}^{N} \left(\hat{\mu}_{is}^{(\cdot)} - \hat{\mu}_{is}^{(T)} \right)^2,$$

where $(\cdot)$ is either (SB) or (FB). The goal is to see which method has better performance in terms of MMSE.

3.4.3 Robustness to Model Misspecification

To assess the robustness of our proposed method to model misspecification errors, we consider three cases where the ratio of the standard deviations $\sigma_u/\sigma_x = 0.5$.

- Case 1: $2^{3/2} \times (X+4) \sim \chi^2_4$, $U \sim N(0, \sigma_u = 0.5)$; it considers misspecification of the distribution on X.
- Case 2: $X \sim N(0,1)$ and $U \sim \text{Laplace}\left(0, \text{scale} = 2^{-3/2}\right)$; it is for a misspecification in the distribution of U.
- Case 3: $2^{3/2} \times (X+4) \sim \chi^2_4$, $U \sim \text{Laplace}\left(0, \text{scale} = 2^{-3/2}\right)$; it considers a misspecification of the distribution on both X and U.

We set $K_{ij} = 2$ for all ij and look at the same link function setup as in Section 3.4.2.

To assess performance in this simulation we compute the posterior mean for the mean $\hat{\mu}^{(T)}$, $\hat{\mu}^{(SB)}$, and $\hat{\mu}^{(FB)}$, for using the true x measured without error, semi-Bayesian method, and fully Bayesian method, respectively. Again, we make our comparisons based on the MMSE of 100 simulated datasets. However, for Cases 2 and 3, we only compute $\hat{\mu}_{is}$ when $x_{is} \in [-1.5, 2]$, so:

$$\text{MMSE} = \sum_{s=1}^{100} (100 N_s^*)^{-1} \sum_{i=1}^{N_s^*} \left(\hat{\mu}_{is}^{(\cdot)} - \hat{\mu}_{is}^{(T)}\right)^2.$$

Note that instead of using N, we use N_s^*. In Case 1, $N_s^* = N$ for all s. However in Case 2 and 3, N_s^* is the number of $x_s \in [-1.5, 2]$. The interval $[-1.5, 2]$ covers 95% of the interior of the distribution of x. The goal is to see which method has better performance in terms of MMSE on the interior of the distribution of x.

3.4.4 Results

Tables 3.1 and 3.2 present the results for impact of the link function of Section 3.4.1, for the Bernoulli and Poisson cases respectively. We observe that for the Bernoulli case estimating the unknown link provided an improvement, about an 8% reduction in MMSE over the known link, while for the Poisson case, the MMSE is similar in both cases. This

Measurement error?	σ_u	log	MG
Yes	0.3	10.1260	**10.0663**
	0.7	25.5381	**25.3898**
No	0.3	5.5845	**5.5481**
	0.7	**6.3291**	6.3668

Table 3.2: The $100 \times$ MMSE for comparing how the unknown link (MG) versus the log link is impacted by measurement error, $x_{ij} \equiv w_{ij}$, with $\sigma_u = \{0.3, 0.7\}$, when $y \sim$ Poisson. MG is a link defined by our simulation setup with $g^{-1}(\eta) = T^{-1}\left[J(\eta)\right]$ and $J(\eta) \approx \sum_{r=1}^{R} \omega_r \text{IB}\left[J_0(\eta); a_r, b_r\right]$ where $R = 4$, $a_r = \lambda r$, $b_r = \lambda(R - r + 1)$, $\lambda = 1$, $J_0(\eta) = [1 + \exp(-\eta)]^{-1}$, $T^{-1}[J(\eta)] = \eta/(1-\eta)$. The $100 \times$ MMSE of the unknown link versus log link is also shown for these cases with the measurement error removed, $x_{ij} \equiv x_{ij}$.

result suggest only a minor loss of efficiency when compared to the fully known link approach, if there is a loss of efficiency at all. However, for both the Bernoulli and Poisson cases with measurement error, the MMSE is similar for the known and unknown link methods. This means when the link function is known, the unknown link method is still comparable to the known link function method, even when measurement error is not accounted for.

a	σ_u	FB	SB
1	0.3	7.0865	**6.9963**
	0.7	**32.9700**	33.7233
0.2	0.3	**2.2008**	2.2392
	0.7	**12.7531**	12.9832
4	0.3	3.4942	**3.4889**
	0.7	17.3425	**17.1887**

Table 3.3: The $100^2 \times$ MMSE for comparing the fully Bayesian (FB) to semi-Bayesian (SB) approach for handling measurement error when $y \sim$ Bernoulli, the true link function is of the form $g^{-1}(\eta) = [1 + \exp(-\eta)]^{-a}$, with $a = \{0.2, 1, 4\}$, and the measurement error standard deviation is $\sigma_u = \{0.3, 0.7\}$.

Tables 3.3 and 3.4 contain the results for the comparison of the fully Bayesian and semi-Bayesian methods, for the Bernoulli and Poisson cases, respectively. There is no substantial improvement by using the fully Bayesian method over the semi-Bayesian method for the Bernoulli case. However for the Poisson case, the fully Bayesian approach was always superior to the semi-Bayesian approach. While in some cases this improvement was negligible, such as for the log link with $\sigma_u = 0.3$, we observed up to a

True link	σ_u	FB	SB
log	0.3	**2.8729**	2.8851
	0.7	**11.8855**	13.1001
MG1	0.3	**5.4887**	5.6256
	0.7	**23.0256**	27.5496
MG2	0.3	**1.5878**	1.5947
	0.7	**8.4817**	8.6440

Table 3.4: The $100^2 \times$MMSE for comparing the fully Bayesian (FB) to semi-Bayesian (SB) approach for handling measurement error when $y \sim$ Poisson, with true link function as log, MG1, or MG2, where MG1 is a link defined by our simulation setup with $g^{-1}(\eta) = T^{-1}\left[J(\eta)\right]$ and $J(\eta) \approx \sum_{r=1}^{R} \omega_r \text{IB}\left[J_0(\eta); a_r, b_r\right]$ where $R = 4$, $a_r = \lambda r$, $b_r = \lambda(R - r + 1)$, $\lambda = 1$, $J_0(\eta) = [1 + \exp(-\eta)]^{-1}$, $T^{-1}[J(\eta)] = \eta/(1-\eta)$, and $\omega = \{0.15, 0.35, 0.35, 0.15\}$ and MG2 is the same as MG1 except for $\omega = \{0.35, 0.15, 0.15, 0.35\}$. The measurement error standard deviations are $\sigma_u = \{0.3, 0.7\}$.

15% reduction in MMSE with the log and MG1 links when $\sigma_u = 0.7$.

a	Distribution of X	Distribution of U	FB	SB
1	Normal	Laplace	19.5711	**19.4281**
	χ^2	Normal	**18.0662**	18.2447
	χ^2	Laplace	**17.1173**	17.4892
0.2	Normal	Laplace	5.2722	**5.1426**
	χ^2	Normal	**4.7857**	4.8597
	χ^2	Laplace	**5.6301**	5.8233
4	Normal	Laplace	10.4845	**10.1958**
	χ^2	Normal	**8.7519**	8.9303
	χ^2	Laplace	**10.0208**	10.1009

Table 3.5: The $100^2 \times$MMSE for comparing the fully Bayesian (FB) to semi-Bayesian (SB) approach for handling measurement error when $y \sim$ Bernoulli, the true link function is of the form $g^{-1}(\eta) = [1 + \exp(-\eta)]^{-a}$, with $a = \{0.2, 1, 4\}$, and the distributions of x and/or u has been misspecified. When the distribution of x is a transformed χ^2, the MMSE is only calculated for $-1.5 < x < 2$ where a domain that covers 95% of the interior of the distribution.

Tables 3.5 and 3.6 show the results for the robustness of the fully Bayesian and semi-Bayesian methods, for the Bernoulli and Poisson cases, respectively. For the Bernoulli case there is no substantial improvement in the fully Bayesian approach over the semi-Bayesian approach. However, there was slight preference for the fully Bayesian method when $x \sim \chi^2$, and slight preference for the semi-Bayesian method when $x \sim$ Normal. For the Poisson cases, the only case favoring the semi-Bayesian approach is with the MG2

Link	Distribution of X	Distribution of U	FB	SB
log	Normal	Laplace	**7.0399**	7.2264
	χ^2	Normal	**4.5578**	4.9735
	χ^2	Laplace	**5.1663**	5.6432
MG1	Normal	Laplace	**12.9994**	13.9717
	χ^2	Normal	**8.9164**	10.6451
	χ^2	Laplace	**8.2902**	9.6387
MG2	Normal	Laplace	3.9417	**3.8532**
	χ^2	Normal	**3.1956**	3.4664
	χ^2	Laplace	**3.1239**	3.4890

Table 3.6: The $100^2 \times$MMSE for comparing the fully Bayesian (FB) to semi-Bayesian (SB) approach for handling measurement error when $y \sim$ Poisson, with true link function as log, MG1, or MG2, where MG1 is a link defined by our simulation setup with $g^{-1}(\eta) = T^{-1}\left[J(\eta)\right]$ and $J(\eta) \approx \sum_{r=1}^{R} \omega_r \text{IB}\left[J_0(\eta); a_r, b_r\right]$ where $R = 4$, $a_r = \lambda r$, $b_r = \lambda(R - r + 1)$, $\lambda = 1$, $J_0(\eta) = [1 + \exp(-\eta)]^{-1}$, $T^{-1}[J(\eta)] = \eta/(1-\eta)$, and $\omega = \{0.15, 0.35, 0.35, 0.15\}$ and MG2 is the same as MG1 except for $\omega = \{0.35, 0.15, 0.15, 0.35\}$. The distributions of x and/or u has been misspecified. When the distribution of x is a transformed χ^2, the MMSE is only calculated for $-1.5 < x < 2$ where a domain that covers 95% of the interior of the distribution.

link when $x \sim$ Normal and $u \sim$ Laplace. This improvement is only about 2%. In all other cases, the fully Bayesian method is the superior choice, at times showing improvement as large as 16%, and most frequently around 7–8%.

3.5 Application: the Framingham Heart Study

The Framingham Heart Study is a longitudinal study of cardiovascular disease (CVD). The first cohort of the study began in 1948: 5,209 men between 30 and 62 years of age from Framingham, MA who had not developed overt symptoms of CVD or suffered a heart attack or stroke. Each participant returns to the doctor for a full check-up and evaluation every two years after the initial visit. To investigate the performance of each of these methods on a real data set, we use the version of the Framingham Heart Study data from the text Carroll et al. (2006). It is available at `http://www.stat.tamu.edu/~carroll/eiv.SecondEdition/data.php`. Our goal in this section is to use measurements taken at the second check-up for each participant to predict the occurrence of coronary heart disease (CHD) in the first eight years after the second check-up. A

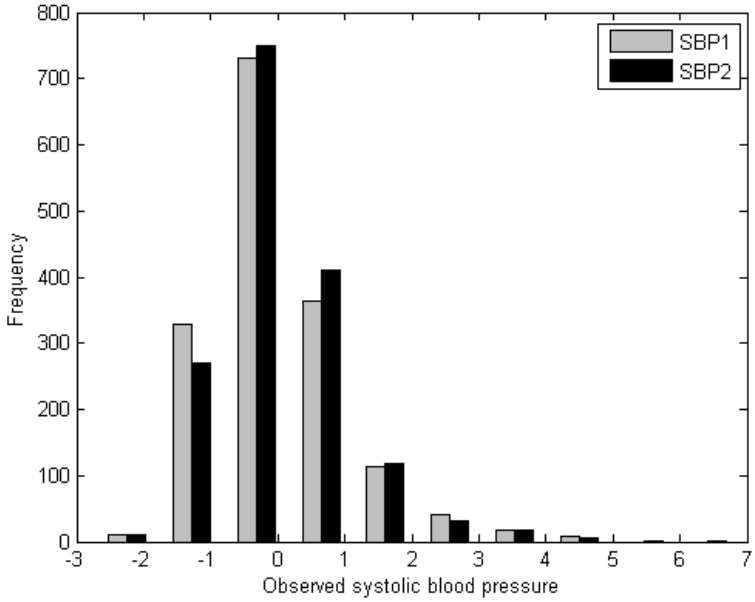

Figure 3.2: A histogram of the centered and scaled measurements of systolic blood pressure, SBP1 and SBP2 in grey and black, respectively. The distribution is skewed to the right.

description of the variables used is as follows:

- $N = 1615$
- Predictors measured with error, x:
  - SBP: systolic blood pressure at second check-up, $m = 2$.
- Predictors measured without error, Z:
  - AGE: age at second check-up.
  - CHOLEST: cholesterol at second check-up.
- Response variable, y:
  - FIRSTCHD: indicator for subject showed signs of CHD within an eight year follow-up period of exam 2.

The response y, "showed signs of CHD", is binary, so logistic regression for Bernoulli response data would be the default choice of GLM with known link function. All predic-

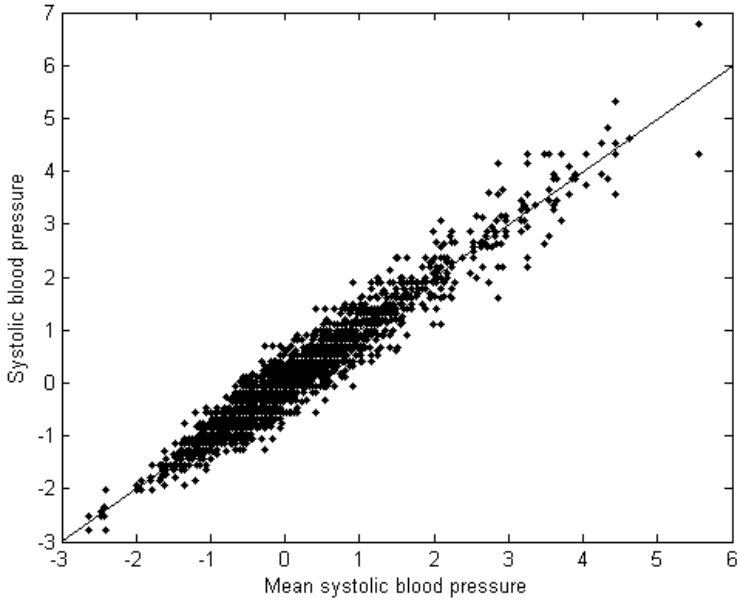

Figure 3.3: A scatter plot of the centered and scaled measurements of systolic blood pressure plotted against the average of the two measurements for each subject. The black line is a one-to-one line the average measurement falls on.

tors are centered and scaled to improve performance of the MCMC sampler, as well as allow us to use the prior specification outlined in Section 3.3, which was chosen with predictors with zero mean and variance one in mind. We estimate the model using the semi-Bayesian approach with both the logit link and unknown link, as well as the fully Bayesian unknown link approach. We follow the choices described in Section 3.4 for the Bernoulli case when selecting hyper parameters for priors, proposal schemes for MCMC, and number of mixands, and their parameters, used in the unknown link method.

We estimate the model twice, once using a random subset of size $N = 100$, to make the results comparable to our simulation that used $N = 100$. The other uses the full data $N = 1615$. Plots of the posterior mean of the fitted values versus an estimate of systolic blood pressure can be seen in Figures 3.4 and 3.1, for the $N = 100$ and $N = 1615$, respectively. We can see that in both cases, estimating the link function provided a different solution to the fitted values. This can also been seen in Table 3.7 where we would expect the 90% credible intervals to be quite similar for each parameter if the logit

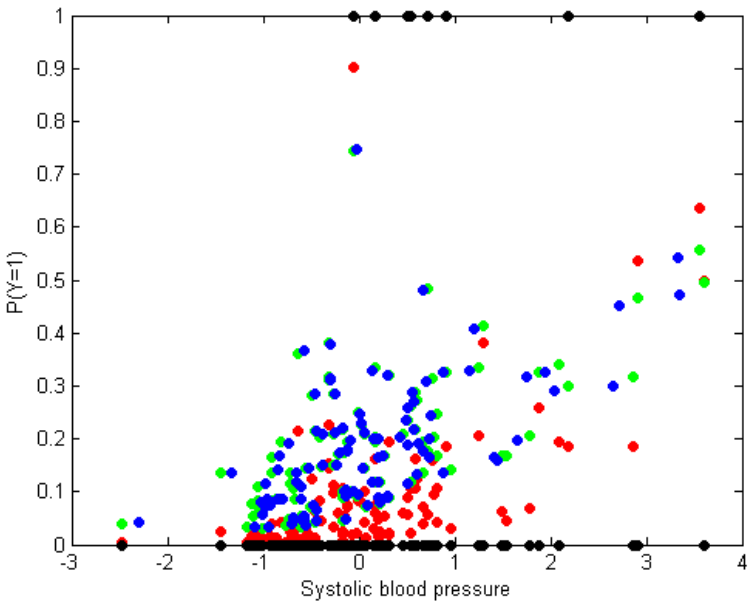

Figure 3.4: Posterior mean fitted values based on a subset of size $N = 100$ of the Framingham Data. The fitted values using the logit link are in red, evaluated at the average systolic blood pressure measurement $\overline{\text{SBP}} = \bar{w}$, the unknown link using the semi-Bayesian approach is in green evaluated at $\bar{w}$, and the fully Bayesian approach is in blue evaluated at the posterior mean of $x = \text{SBC}$, $\hat{x}$. The black circles are for the indicator of CHD y evaluated at $\bar{w}$.

link were the true link function. Instead, the credible intervals for age and cholesterol using the logit link don't even overlap with the credible intervals using the unknown link methods. This is a good indicator that unknown link method was the better modeling choice.

Figure 3.3 contains a scatter plot of the two measurements of SBP for each subjected versus the average SBP measurement for each subject. While we cannot infer whether the assumption that the measurement errors for SBP have zero mean based on $m = 2$, there does appear to be some mild heteroskedasticity issues with the measurement error variance, which is not something we addressed in this paper. Figure 3.2 contains side by side histograms of the two centered and scaled measurements of the systolic blood pressure. The measurements are skewed to the right. If the zero mean of the the measurement errors is an appropriate assumption, this must mean that the distribution of SBP itself is skewed to the right. Based on our simulation results in Section 3.4.4

N	Method	Predictor	LB	UB
100	logit	SBP	0.2041	1.3896
		AGE	-0.4952	0.9580
		CHOLEST	0.4506	1.7603
	SB	SBP	-0.0967	0.6127
		AGE	-0.2304	0.5834
		CHOLEST	0.1265	0.6971
	FB	SBP	-0.0663	0.6248
		AGE	-0.2067	0.5630
		CHOLEST	0.1209	0.7222
1615	logit	SBP	0.1335	0.4131
		AGE	0.3015	0.6312
		CHOLEST	0.2717	0.5620
	SB	SBP	0.0428	0.2103
		AGE	0.1095	0.2617
		CHOLEST	0.1015	0.2563
	FB	SBP	0.0523	0.2043
		AGE	0.1037	0.2679
		CHOLEST	0.1105	0.2557

Table 3.7: 90% credible intervals, [LB, UB], for three models using known logit link, and unknown link using the semi-Bayesian (SB) and fully Bayesian (FB) approaches, respectively; 90% equal tail credible intervals were computed for (β_x, β_z) for each predictor, systolic blood pressure (SBP), age, and cholesterol (CHOLEST). This was done both on a small subset of the data ($N = 100$) and the full data ($N = 1615$).

we concluded that there was minimal difference between the performance of the semi-Bayesian and fully Bayesian when the distribution of x was skewed, however the fully Bayesian approach was favored. For this reason, we might prefer the fully Bayesian approach for this analysis. While it's not clear how the increase in sample size from $N = 100$ to $N = 1615$ would alter the simulation results, if at all, what can be said is that computationally the semi-Bayesian was significantly faster (20 versus 240 minutes when we ran it). For very large datasets such as the Framingham Heart Study, possibly with multiple variables measured with error, this time savings could be preferential over slightly improved model fits.

3.6 Discussion

For generalized linear models we have proposed two flexible approaches for handling error-in-covariates in the presence of an unknown link function. The first is a fully Bayesian approach that treats x as a latent variable to be integrated over, while the second is a semi-Bayesian approach that could be motivated several ways, such as via a first order Laplace approximation, marginalizing X out of the likelihood. There is some existing methodology to tackle this problem (Mallick and Gelfand 1996), however it uses an alternative parameterization of the model, as well as relying on very informative priors in place of replicated measurement of x in order to induce identifiability of the model. By using replicated measurements, our methodology no longer hinges upon informative prior information. In addition, we investigate the performance of our methods both when the model has been correctly and incorrectly specified.

Our simulations show that: (1) The unknown link function method of Mallick and Gelfand (1994) is more robust to unaccounted for covariate measurement error than using the canonical link. And (2) the fully Bayesian approach is either as good as the semi-Bayesian approach, as in the Bernoulli cases, or better, as in the Poisson cases. However, the semi-Bayesian approach is much easier and faster to implement computationally, which could make it a better choice in some circumstances. Also for the robust simulations, we only considered the domain of x that covered 95% of the interior of the distribution. The reason for this choice is that for the Poisson case, the fully Bayesian approach did not perform as well on the tails of x. Though this choices was also made for the Bernoulli case to maintain consistancy, we did not observe this loss in performance in the tails. When considering the number of replicated measurements, we looked mainly at the case where $K_{ij} = 2$ for all ij. We suspect that if $K_{ij} = 1$ for some ij, we would begin to see even greater favoring of the fully Bayesian approach. If $K_{ij} = 1$ for all ij, then the performance of one over the other will depend on the user's ability to correctly specify the prior for σ_u.

The approaches presented here are for the classical form of measurement error $W = X + U$, where X is never observed. It would be an interesting and challenging prob-

lem to extending our approach for non-additive measurement error, as well as for high dimensional measurement error. Also, while we would expect the extension of this methodology to the instrumental variable approach for handling measurement error to be straightforward, investigating its properties and performance would be very useful given the popularity of instrumental variable methodology.

Acknowledgments

We would like to thank Pang Du, Leanna House, Scotland Leman, George Terrell, and Matt Williams for their advice and assistance.

Bibliography

Buzas, J. and Stefanski, L. (1996), "Instrumental variable estimation in generalized linear measurement error models," *Journal of the American Statistical Association*, 91, 999–1006.

Carroll, R., Ruppert, D., Crainiceanu, C., Tosteson, T., and Karagas, M. (2004), "Nonlinear and nonparametric regression and instrumental variables," *Journal of the American Statistical Association*, 99, 736–750.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006), *Measurement Error in Nonlinear Models: A Modern Perspective*, Monographs on Statistics and Applied Probability, Boca Raton, FL: Chapman and Hall/CRC, 2nd ed.

Escobar, M. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *J. Amer. Statist. Assoc.*, 90, 577–588.

Geman, S. and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109.

Mallick, B. and Gelfand, A. (1996), "Semiparametric errors-in-variables models - A Bayesian approach," *Journal of Statistical Planning and Inference*, 52, 307–321.

Mallick, B. K. and Gelfand, A. E. (1994), "Generalized Linear-Models With Unknown Link Functions," *Biometrika*, 81, 237–245.

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

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equation of State Calculations by Fast Computing Machines," *The Journal of Chemical Physics*, 21, 1087–1092.

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

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge series on statistical and probabilistic mathematics, New York: Cambridge University Press.

Sinha, S., Mallick, B. K., Kipnis, V., and Carroll, R. J. (2010), "Semiparametric Bayesian Analysis of Nutritional Epidemiology Data in the Presence of Measurement Error," *Biometrics*, 66, 444–454.

Tierney, L. and Kadane, J. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86.

3.A Marginal likelihood using Laplace approximation

The goal of this section is to show using first order Laplace approximation that:

$$\int L(y,w|x,z,\beta,\sigma_u^2)\,\mathrm{d}x \approx L(y|x=\bar{w},z,\beta).$$

Note that:

$$\begin{aligned}
L(y_i,W_{ij}|x_i,z_i,\beta,\sigma_u^2) &= L(y_i|,x_i,z_i,\beta)\times N(W_{ij};x_{ij},\sigma_u^2)\\
&= L(y_i|,x_i,z_i,\beta)(2\pi\sigma_u^2)^{K_{ij}/2}\\
&\quad\times\exp[-(W_{ij}-x_{ij})^T(W_{ij}-x_{ij})/(2\sigma_u^2)].
\end{aligned}$$

We will write:

$$\begin{aligned}
A(x_{ij}) &= L(y_i|x_i,z_i,\beta)(2\pi\sigma_u^2)^{-K_{ij}/2}, \qquad \text{and}\\
h(x_{ij}) &= (W_{ij}-x_{ij})^T(W_{ij}-x_{ij})/(2\sigma_u^2).
\end{aligned}$$

It is easy to show that for all ij, $h(x_{ij})$ has unique maximum $\bar{w}_{ij\cdot}$, since $h(\cdot)$ is a quadratic form, and that the second derivative $h''(x_{ij})=1/\sigma_u^2$. Tierney and Kadane (1986) show that we can approximate $\int A(x_{ij})\exp[-h(x_{ij})]\,\mathrm{d}x_{ij}$ by $A(\tilde{x})\exp[-h(\tilde{x})]\sqrt{\frac{2\pi}{K_{ij}h''(\tilde{x})}}$, where $\tilde{x}$ is the value that maximizes $h(\cdot)$. We then get:

$$\begin{aligned}
\int A(x_{ij})\exp[-h(x_{ij})]\,\mathrm{d}x_{ij} &\approx L(y_i|x_{i,-j},x_{ij}=\bar{w}_{ij},z_i,\beta)\left[K_{ij}(2\pi\sigma_u^2)^{K_{ij}-1}\right]^{-1/2}\\
&\quad\times\exp[-(W_{ij}-\bar{w}_{ij\cdot})^T(W-\bar{w}_{ij\cdot})/(2\sigma_u^2)]\\
&\propto L(y_i|x_{i,-j},x_{ij}=\bar{w}_{ij\cdot},z_i,\beta).
\end{aligned}$$

It is clear then that:

$$\begin{aligned}
\int L(y,W|x,z,\beta,\sigma_u^2)\,\mathrm{d}x &= \prod_{i=1}^{N}\prod_{j=1}^{p_1}\int L(y_i,W_{ij}|x_{ij},z_i,\beta,\sigma_u^2)\,\mathrm{d}x_{ij}\\
&\approx \prod_{i=1}^{N}\prod_{j=1}^{p_1}\left\{L(y_i|x_i=\bar{w}_i,z_i,\beta)\left[K_{ij}(2\pi\sigma_u^2)^{K_{ij}-1}\right]^{-1/2}\right.\\
&\quad\left.\times\exp[-(W_{ij}-\bar{w}_{ij\cdot})^T(W_{ij}-\bar{w}_{ij\cdot})/(2\sigma_u^2)]\right\}\\
&\propto \prod_{i=1}^{N}L(y_i|\bar{w}_i,z_i,\beta)\\
&= L(y|\bar{w},z,\beta).
\end{aligned}$$

CHAPTER 4

DISCUSSION

In this dissertation, we proposed a semi- and a fully Bayesian approach for handling covariate measurement error in both semiparametric regression for matched case-control studies and generalized linear models with unknown link function, described in Chapters 2 and 3, respectively. Using a simulation study, we compared the performance of each method under correctly and incorrectly specified models in terms of mean squared error. Finally, we also illustrated each method on a real dataset. Additionally, for semiparametric regression methods for matched case-control studies, we proposed two EM approaches, as well as consider performance in terms of mean bias in the simulation study.

Our conclusions concerning the performance of the semi-Bayesian versus the fully Bayesian approach in terms of MSE was consistent across both applications: the fully Bayesian approach is at least as good as the semi-Bayesian approach in reducing MSE, and performs particularly well when there is model misspecification the distribution of X or U. For semiparametric regression in matched case-control studies we found that the performance of the EM approaches depended too heavily on the choice of σ_q^2 and σ_k^2 to be considered in the comparison of methods. We conjectured that instead of choosing them, using a regularized estimation technique for each could alleviate this issue. However based on some preliminary simulations, omitted from this dissertation, we found this not to be the case. When comparing the semi- and fully Bayesian approaches in terms of mean bias, we found that the choice of superior method depended on the unknown function $m(x)$.

4.1 Future Work

In Chapter 2 we used low-rank thin-plate splined to estimate an unknown function of a variable measured with error $m(x)$. Little work has been done to investigate performance of such spline methods in generalized additive models. Additionally, low-rank thin-plate splines have a natural extension for use as a high dimensional smoother and little work has been done evaluating their performance when dealing with a large number of variables measured with error.

In Chapter 3 we used a special transformation of a mixture of beta distribution functions to approximate the mean response in generalized linear models. Many of the possible extensions of this work could additionally be considered extensions of the work described in Chapter 2. We showed some robustness to misspecification of X and U when using the semi- and fully Bayesian approaches. In general, investigation into matters of robustness is not done enough when investigating performance of measurement error methods. It is our opinion that more time should probably be spent on these matters than simply proof of concept for models correctly specified. In practice, correctly specifying the distribution can be difficult. This is why work on flexible nonparametric approaches for relaxing some distributional assumptions about X and U could provide even better performance in the future, rather than relying on the robustness of the methods proposed, that often is not investigated. Additionally, we could consider other measurement error models such as the instrumental variable model, which is becoming more popular in public health applications. For example,increasingly work using Dirichlet process mixtures is used in the instrumental variable model for covariate measurement error.