AN
INTRODUCTION TO
BAYESIAN
ANALYSIS

MTH535A

A Comparative study of Bayesian approach to case-control studies with errors in covariables

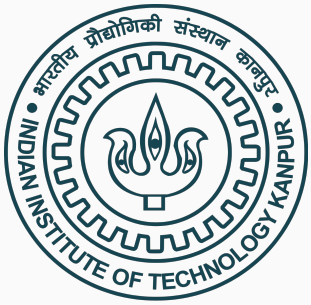
Soumya Paul, Sankhadeep Mitra, Sampriti Dutta

Under the guidance of

Prof. Arnab Hazra

Department of Mathematics and Statistics,

Indian Institute of Technology Kanpur



Contents

Abstract

We would like to develop a Bayesian methodology for the analysis of case-control data with covariate uncertainty. Pretending that the distribution of imprecisely measured covariates is discrete on a heuristically chosen support set provides a method that is relatively easy to implement and applicable to a variety of study designs. Further development of the method highlights the interplay between retrospective and prospective analyses. We illustrate this method with simulated data.

0.1 Introduction

A *case-control* study is a study of a medical condition/disease by comparing patients with the condition/disease (*"cases"*) to otherwise similar patients (*"control"*) who do not have the condition/disease. It is commonly used to identify possible contributing factors. Case-control studies are an important and useful method for studying health outcomes, and many methods have been developed to analyze case-control data of exposed subjects between cases and controls.

Our analysis seeks to examine the effects of covariates on disease status using samples collected as a function of disease status. In many cases, some covariates cannot be measured precisely. It is also well known that an analysis that treats inaccurate measurements as accurate can lead to distorted results. In this paper, we presented a Bayesian methodology to account for covariate measurement errors in case-control analysis.

As of late, Bayesian strategies have been effectively applied to a wide run of measurment error issues *[Carroll et al.(1999), Dellaportas and Stephens(1995), Mallick and Gelfand(1996), Gilks et al.(1995), Richardson and Gilks(1993)]*. The victory of these strategies stems from a direct conceptual and computational approach to averaging over unobserved quantities such as the genuine covariate values, within the posterior distribution.

If the precise covariates are denoted by X, and disease status by D (where $D = 0$ and $D = 1$ corresponds to disease free and diseased respectively), then a case–control study involves sampling "controls" from $X \mid D = 0$ and "cases" from $X \mid D = 1$. Therefore, an exact analysis requires a likelihood based on the conditional distribution of X given D, the so-called ***retrospective model***.

However, the standard approach to analysis is to pretend that the data were collected prospectively and use likelihood based on the ***prospective model*** $D|X$ that is usually assumed to follow a logistic regression model. A justification for this is provided by *[PRENTICE and PYKE(1979)]*. In the case of measurement error, the imprecise measurement W stands in for X.

Here we used an approach that derives the posterior distribution directly from the retrospective model. In fact, as an approximation of the retrospective posterior distribution, we get the prospective posterior distribution. This sheds new light on the interplay of retrospective and prospective analyses.

This method can be applied to different measurement error scenarios. One scenario is ***validation design*** where both W and X are measured for some subjects (full data) and only W for the rest (reduced data). If the exact form of the distribution of $W \mid X, D$ contains unknown parameters, the full data will contain information about these parameters. Therefore, both full and reduced data contain information about the relationship between X and D.

Another scenario arises when the distribution of $W \mid X, D$ is completely known from an external study and the case-control study consists only of reduced data. This situation is called ***external validation***. For our project, we will limit ourselves to external validation.

The approach of *[Gustafson P. Le N.(2000)]* contains the normal retrospective model for $X \mid D$ and the normal measurement error model for $W \mid X, D$. The assumption of normal distribution is limited in scope. The approach described here is much more widely applicable.

Section 0.2 describes the model and derives some variants of the posterior and prior distributions of

interest. *Section 0.5* test the method on simulated data. *Sections 0.3* and *0.4* details the Markov Chain Monte Carlo (MCMC) scheme used.

0.2 Methodology

Suppose for a particular individual,

$$D = \begin{cases} 1 & \text{if the individual is diseased} \\ 0 & \text{if the individual is disease-free} \end{cases} \quad \text{and } \boldsymbol{X} = (X_1, X_2, ..., X_p) \text{ : a vector of } p \text{ covariates.}$$

Generally, the scale at which measurement error is additive may not be same as the scale at which the covariate is related with the disease model for imprecisely measured covariates ($\boldsymbol{W}$). Therefore, we assume that there exists a suitable transformation $s(.)$ on $\boldsymbol{X}$ such that the measurement error has an additive effect on the components of $\boldsymbol{X}$.

Our analysis requires the likelihood based on the retrospective model. The standard approach to analysis, is to pretend the data are sampled *prospectively* and use a likelihood based on $D|\boldsymbol{X}$, which is typically assumed to follow a *logistic regression model*. Thus we specify the disease's *prospective model* as

$$\log\left(\frac{\mathbb{P}(D=1|\boldsymbol{X}=\boldsymbol{x})}{\mathbb{P}(D=0|\boldsymbol{X}=\boldsymbol{x})}\right) = \alpha^* + \beta^T s(\boldsymbol{x}) \tag{0.2.1}$$

, where $s(\boldsymbol{x}) = (s_1(x_1), s_2(x_2), ..., s_p(x_p))^T$, with $s_i : \mathbb{R} \to \mathbb{R}$ being a known function.

We are interested in knowing the value of the parameter set $\{\alpha^*, \beta_1, \beta_2, ..., \beta_p\}$.

The data are collected retrospectively, i.e., the *controls* comprise a sample of size n_1 drawn from the distribution of $\boldsymbol{X} \mid D = 0$, while the *cases* comprise a sample of size n_2 drawn from the distribution of $\boldsymbol{X} \mid D = 1$. For future use we define the total sample size as $n = n_1 + n_2$, and the sampling fractions as $r_i = \dfrac{n_i}{n};\ i = 1, 2$

In terms of log-odds there is a fundamental link between *retrospective* and *prospective* model:

$$\begin{aligned} \log\left(\frac{f(\boldsymbol{x} \mid D=1)}{f(\boldsymbol{x} \mid D=0)}\right) &= \log\left(\frac{\mathbb{P}(D=1|\boldsymbol{X}=\boldsymbol{x})\, f(\boldsymbol{x})\, \mathbb{P}(D=0)}{\mathbb{P}(D=0|\boldsymbol{X}=\boldsymbol{x})\, f(\boldsymbol{x})\, \mathbb{P}(D=1)}\right) \\ &= \log\left(\frac{\mathbb{P}(D=1|\boldsymbol{X}=\boldsymbol{x})}{\mathbb{P}(D=0|\boldsymbol{X}=\boldsymbol{x})}\right) - \log\left(\frac{\mathbb{P}(D=1)}{\mathbb{P}(D=0)}\right) \\ &= \alpha^* + \beta^T s(\boldsymbol{x}) - \log\left(\frac{\mathbb{P}(D=1)}{\mathbb{P}(D=0)}\right) \\ &= \alpha + \beta^T s(\boldsymbol{x}) \text{ where } \alpha = \alpha^* - \log\left(\frac{\mathbb{P}(D=1)}{\mathbb{P}(D=0)}\right) \end{aligned} \tag{0.2.2}$$

Using the retrospective data we cannot estimate α^* without knowing the value of $\mathbb{P}(D=1)$.

The equation (0.2.2) can be rewritten as equation which is parametrized according to (α, β, h)

$$f_{\boldsymbol{X}|D}\left(\boldsymbol{x} \mid d\right) = \exp\left(d\left(\alpha + \beta^T s\left(\boldsymbol{x}\right)\right)\right) h\left(\boldsymbol{x}\right); \text{ where } d = 0, 1 \text{ and } h\left(x\right) : \text{density of } X \mid D = 0$$

If $d = 1$, then we get the density of $\boldsymbol{X} \mid D$, i.e.,

$$\begin{aligned} f_{\boldsymbol{X}|D}\left(x \mid d = 1\right) &= \exp\left(\left(\alpha + \beta^T s\left(\boldsymbol{x}\right)\right)\right) h\left(\boldsymbol{x}\right) \qquad (0.2.3) \\ &\text{and} \\ \int_{-\infty}^{\infty} f_{\boldsymbol{X}|D}\left(\boldsymbol{x} \mid d = 1\right) d\boldsymbol{x} &= 1 \\ \implies \int_{-\infty}^{\infty} \exp\left(\alpha + \beta^T s\left(\boldsymbol{x}\right)\right) h\left(\boldsymbol{x}\right) d\boldsymbol{x} &= 1 \\ \implies E_h\left(\exp\left(\alpha + \beta^T s\left(\boldsymbol{x}\right)\right)\right) &= 1 \qquad (0.2.4) \end{aligned}$$

We have equation (0.2.3) with constraint (0.2.4) and three unknown parameters (α, β, h). So, α can be written as a function of β and h i.e. $\alpha = \alpha\left(\beta, h\right)$.

Let's talk about measurement error.Suppose we have only one covariate $(p = 1)$ for which the presence of measurement error is externally validated i.e. we have observed W where the actual value X is unobserved for all subjects and the measurement error density, $f(w|x, d)$ is completely specified. Instead of restricting ourselves to a specific form of the density, it is useful for what follows to define $\tau^2 = Var(W|X, D)$ as the variance of imprecise measurement given actual measurement. Here, we presume that D has no effect on the conditional variance. More broadly, measurement error is frequently considered to be *nondifferential*, in which case the conditional distribution of $W|X, D$ is independent of D.

Now, the joint distribution of (W, X, D) can be presented as

$$f\left(w, x, d\right) = f\left(w \mid x, d\right) f\left(x \mid d\right) f\left(d\right); \text{ where } f\left(x \mid d\right) \text{ depends on } h \text{ and } \beta$$

For a given prior $f\left(\beta, h\right)$ the posterior distribution of (X, β, h) is

$$\begin{aligned} f(x, \beta, h|w, d) &\propto f(w, x, d|\beta, h) f(\beta, h) \\ &= \left(\prod_{i=1}^{n} f\left(w_i \mid x_i, d_i\right) f\left(x_i \mid d_i, \beta, h\right) f\left(d_i \mid \beta, h\right)\right) f\left(\beta, h\right) \end{aligned}$$

Usually β is the parameter of interest and density h is the nuisance parameter. As per *[BUONACCORSI(1990)]*, If we assume h belongs to a simple parametric family, there is a risk of misspecifying the model. On the other hand, the flexible model of h proposed by *[Muller(1997), Muller et al.(1999)]* failed to completely eliminate erroneous specifications, leading to complex model-tuning procedures. To find a compromise between simplicity and flexibility, here we consider approximating h by a discrete distribution.

Instead of modeling h itself as a discrete distribution, it makes sense to reparameterize from (β, h) to

(β, g), where the density of X is given by

$$g(x) = r_1 h(x) + r_2 \exp\{\alpha(\beta, h) + \beta s(x)\} h(x) \tag{0.2.5}$$

Suppose λ parameterizes g and for simplicity, assume β and λ are a priori independent. Then (**??**) reduces to

$$f(x, \beta, \lambda | w, d) \quad \propto \quad \left(\prod_{i=1}^{n} f(w_i \mid x_i, d_i) f(x_i \mid d_i, \beta, \lambda) f(d_i \mid \beta, \lambda) \right) f(\beta) f(\lambda) \tag{0.2.6}$$

To approximate g with a discrete distribution, we need to specify a grid of points to act as supports. Let's say the grid-points are

$$z[1] < z[2] < ..... < z[m]$$

, and define $\theta = \{\theta_1, \theta_2, ..., \theta_n\}$ to match the sample drawn from distribution of X to the grid via

$$x_i = z[\theta_i] \ ; \theta_i \in \{1, 2, ..., m\}$$

Further, let λ_j be the probability that g assigns to $z[j]$, and say that λ has a uniform, or $Dirichlet(1, ..., 1)$, prior on the m-dimensional probability simplex. Under this formulation, (0.2.6) becomes

$$f(\theta, \beta, \lambda | w, d) \quad \propto \quad \left(\prod_{i=1}^{n} f(w_i \mid z[\theta_i], d_i) f(z[\theta_i] \mid d_i, \beta, \lambda) f(d_i \mid \beta, \lambda) \right) f(\beta) \tag{0.2.7}$$

Strategy for choosing grid points $z[1], z[2], ..., z[m]$ depend on the nature of the data at hand .

For the measurement error model if we assume additive relationship between precise and imprecise measurements as

$$W = X + u \ ; E(u) = 0, Var(u) = \tau^2 \tag{0.2.8}$$

Then

$$E(W|X) = X, \text{ and } Var(W|X) = \tau^2$$

When $\tau = 0$, the distribution of $W|X, D$ is degenerated at the point x. We simply take the grid-points to be the observed x values.

When $\tau > 0$,in the case of external validation there are competing desiderata in choosing the m grid-points. On the one hand, we want m to be small in hopes of getting reasonable accuracy in estimating g. But on the other hand, to properly update the x_i values, we need a fine enough grid to correctly capture the shape of $f(w_i|x_i, d_i)$ as a function of x_i. According to [Gustafson(2002)] we will choose an equally spaced grid, with

$$z[1] = \min_i \{w_i\} - 2.5\tau, z[m] = \max_i \{w_i\} + 2.5\tau \text{ with a spacing of } \tau/4.$$

0.3 MCMC Scheme

Our main interest is to draw samples from $f(\theta, \beta, \lambda|w, d)$ as presented in (0.2.7) using MCMC scheme.

For the measurement error model

$$W \mid X \sim N\left(X, \tau^2\right)$$

$X \mid D, \beta, \lambda$ is approximated through $z[\theta]$

$$logit\left\{\mathbb{P}\left(D = 1 \mid X\right)\right\} = -3 + 0.5\exp\left(X\right)$$

So, true value of β is 0.5. The prior for β is

$$\beta \sim N\left(0, 100^2\right)$$

Here it is not possible to specify a conditionally conjugate prior. Also we assumed the candidate distribution to be normal which is symmetric. So, we have used Metropolis sampling technique to replace draws from the exact full conditional distribution with the draw from a candidate distribution followed by an accept or reject state.

Algorithm 0.1 Metropolis Algorithm

Initialize $\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_p^{(0)}\right)$
for $s \leftarrow 1$ to S **do**
  for $j \leftarrow 1$ to p **do**
    sample $\theta_j^* \sim q_j\left(\theta_j \mid \theta_j^{(s-1)}\right)$ for symmetric q_j
    set $\boldsymbol{\theta}^* = \left(\theta_1^{(s)}, ..., \theta_{j-1}^{(s)}, \theta_j^*, \theta_{j+1}^{(s-1)}, ..., \theta_p^{(s-1)}\right)$
    set $R = \dfrac{f\left(\boldsymbol{Y} \mid \theta^*\right)\pi\left(\theta^*\right)}{f\left(\boldsymbol{Y} \mid \theta^{(s-1)}\right)\pi\left(\theta^{(s-1)}\right)}$
    sample $U \sim \text{uniform}(0, 1)$
    if $U < R$ **then**
      set $\theta_j^{(s)} = \theta_{j-1}^*$
    else
      set $\theta_j^{(s)} = \theta_j^{(s-1)}$

0.4 Computation

Let us assume that the precise measurement $X \sim N\left(\mu, \tau^2\right)$. By equation (0.2.8)

$$W \mid X \sim N\left(X, \tau^2\right)$$

The model for $D \mid X$ is

$$logit\left\{\mathbb{P}\left(D = 1 \mid X\right)\right\} = \alpha + \beta\exp\left(X\right) \tag{0.4.1}$$

We are interested in the posterior distribution of $X, \beta \mid W, D$.

$$\begin{aligned} f(x,\beta|w,d) &\propto f\left(w \mid x,d,\beta\right) f\left(x,d,\beta\right) \\ &= f\left(w \mid x,d\right) f\left(x \mid d,\beta\right) f\left(d,\beta\right) \\ &= f\left(w \mid x,d\right) f\left(x \mid d\right) f\left(d \mid \beta\right) f\left(\beta\right) \\ &= \left(\prod_{i=1}^{n} f\left(w_i \mid x_i,d_i\right) f\left(x_i \mid d_i\right) f\left(d_i \mid \beta\right)\right) f\left(\beta\right) \end{aligned} \tag{0.4.2}$$

where

$$f\left(w_i \mid x_i,d_i\right) = \frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{\left(w_i - x_i\right)^2}{2\tau^2}\right)$$

The expressions of $f\left(x_i \mid d_i\right)$ and $f\left(d_i \mid \beta\right)$ are derived as follows.

The m grid-points $z\left[j\right]$'s are created as mentioned in Section 0.2. To each grid-point we associate probability λ_j, where $\left(\lambda_1, ..., \lambda_m\right) \sim Dirichlet\left(1, 1, ..., 1\right)$. We approximate the unobserved precise covariate X by the grid-points. So, $f\left(x \mid d\right)$ is a PMF with the m grid-points as the support.

$$f\left(x_j \mid d_j\right) = \frac{\lambda_j f\left(w_j \mid x_j\right)}{\sum_{j=1}^{m} \lambda_j f\left(w_j \mid x_j\right)},\ j = 1\left(1\right)m \tag{0.4.3}$$

Note that here we have m x_j's and w_j's, whereas equation (0.4.2) has n many x_i's and w_i's. To overcome this issue we simulate w_j from $N\left(z\left[j\right], \tau^2\right),\ j = 1, 2, ..., m$ and calculate $f\left(x_j \mid d_j\right)$. Then we generate a sample of size n with replacement from the distribution given by (0.4.3).

The posterior of X is given by

$$f\left(x \mid w,d\right) \propto \prod_{i=1}^{n} f\left(w_i \mid x_i,d_i\right) f\left(x_i \mid d_i\right) \tag{0.4.4}$$

To draw samples from (0.4.4) we implemented MCMC using Metropolis sampling algorithm.

From equation (0.4.1)

$$\begin{aligned} logit\left\{\mathbb{P}\left(d_i = 1 \mid \bar{x}_i\right)\right\} &= \alpha + \beta \exp\left(\bar{x}_i\right) \\ \implies \frac{\mathbb{P}\left(d_i = 1\right)}{\mathbb{P}\left(d_i = 0\right)} &= \exp\left(\alpha + \beta \exp\left(\bar{x}_i\right)\right) \\ \implies \frac{p_i}{1 - p_i} &= \exp\left(\alpha + \beta \exp\left(\bar{x}_i\right)\right) \text{ where } p_i = \mathbb{P}\left(d_i = 1\right) \\ \implies p_i &= \frac{\exp\left(\alpha + \beta \exp\left(\bar{x}_i\right)\right)}{1 + \exp\left(\alpha + \beta \exp\left(\bar{x}_i\right)\right)} \end{aligned} \tag{0.4.5}$$

where $\bar{x}_i$ is the post burn-in sample mean of X_i.

So,

$$d_i \mid \beta \sim Ber\left(p_i\right)$$

The posterior of β is given by

$$f(\beta \mid w, d) \propto \left(\prod_{i=1}^{n} f(d_i \mid \beta)\right) f(\beta) \tag{0.4.6}$$

, where $f(\beta)$ is PDF of $N(0, 100^2)$

To draw samples from (0.4.6), again we implemented Metropolis sampling algorithm.

Algorithm 0.2 Computation

Set $\tau \in \{0, 0.25, 0.5, 0.75\}$, $\mu = 2$, $\alpha = -3$, $\beta = 0.5$ and $n \in \{30, 60, 120, 160\}$
Generate $\boldsymbol{X} = (X_1, X_2, .., X_n)$ such that $X_i \sim N(\mu, \tau^2)$ (n even)
Generate $\boldsymbol{D} = (D_1, D_2, .., D_n)$ with $n/2$ 0's and $n/2$ 1's such that
if $\dfrac{\exp(\alpha + \beta \exp(X))}{1 + \exp(\alpha + \beta \exp(X))} < \dfrac{1}{2}$ **then**
    $D = 0$
else
    $D = 1$
Generate $\boldsymbol{W} = (W_1, W_2, .., W_n)$ such that
for $i \leftarrow 1$ to n **do**
    $W_i \sim N(X_i, \tau^2)$
Generate m gridpoints $Z[1] < Z[2] < ... < Z[m]$ such that
for $j \leftarrow 1$ to m **do**
    set $Z[j] = \min_i \{w_i\} - 2.5\tau + (j-1)\dfrac{\tau}{4}$
Draw $(\lambda_1, \lambda_2, ..., \lambda_m) \sim Dirichlet(1, 1, ..., 1)$.
Generate $\boldsymbol{W} = (W_1, W_2, .., W_m)$ and $f(\boldsymbol{w}) = (f(W_1), f(W_2), .., f(W_m))$such that
for $j \leftarrow 1$ to m **do**
    Generate $W_j \sim N(Z[j], \tau^2)$
    set $f(W_j) = \dfrac{1}{\tau\sqrt{2\pi}} \exp\left(-\dfrac{(W_j - Z[j])^2}{2\tau^2}\right)$
Generate $\boldsymbol{p} = (p_1, p_2, ..., p_m)$such that
for $j \leftarrow 1$ to m **do**
    set $p_j = \dfrac{\lambda_j f(w_j)}{\sum_{k=1}^{m} \lambda_j f(w_j)}$
function CLOSEPROB(a)
    return the probability of closest point of a
function LOGPOSTERIORX(x)
    return log-Posterior of x
set $S = 10000$
Apply *Metropolis Sampling* with initial value $\boldsymbol{X}^{(0)} = \boldsymbol{W}$, $p = n$ and candidate: $N\left(\boldsymbol{X}^{(s-1)}, \tau^2\right)$
Consider samples $\boldsymbol{X}^* = (X_1^*, X_2^*, ..., X_n^*)$ after burn-in period of 2000 samples
Obtain $\widehat{E(X)} = \dfrac{1}{8000n} \sum_{i=1}^{8000} \sum_{j=1}^{n} X_{ij}^*$, $\widehat{Var(X)} = \dfrac{1}{8000n} \sum_{i=1}^{8000} \sum_{j=1}^{n} \left(X_{ij}^* - \bar{X}_j^*\right)^2$, Mean distance $= \|\boldsymbol{X} - \bar{\boldsymbol{X}}^*\|$

where $\bar{\boldsymbol{X}}^* = \left(\bar{X}_1^*, \bar{X}_2^*, ..., \bar{X}_n^*\right)$

function LOGPOSTERIORBETA(β)

    return log-Posterior of β

Apply *Metropolis Sampling* with initial value $\beta^{(0)} = 50$, $p = 1$ and candidate: $N\left(\beta^{(s-1)}, 100^2\right)$

Consider samples β^* after burn-in period of 2000 samples

Obtain $\widehat{E(\beta)} = \dfrac{1}{8000}\sum_{i=1}^{8000}\beta_i^*$, $\widehat{MSE(\beta)} = \dfrac{1}{8000}\sum_{i=1}^{8000}(\beta_i^* - \beta)^2$, $\widehat{Var(\beta)} = \frac{1}{8000}\sum_{i=1}^{n}\left(\beta_i^* - \widehat{E(\beta)}\right)^2$

Compute 95% credible interval $\left(\widehat{E(\beta)} - z_{0.025}\sqrt{\dfrac{\widehat{Var(\beta)}}{8000}}, \widehat{E(\beta)} + z_{0.025}\sqrt{\dfrac{\widehat{Var(\beta)}}{8000}}\right)$

Repeat (Step 2) to (Step 35) 10 times and obtain Coverage probability=Proportion of times credible interval contains the true value of β.

0.5 Results

0.5.1 Scenario A

n_i	$\widehat{E(\beta)}$	$\widehat{MSE(\beta)}$
15	0.5649	0.0128
30	0.5594	0.0073
60	0.5273	0.0034
120	0.5116	0.0019

Table 1: Simulation results for scenario A. This scenario involves no measurement error and various sample sizes (with $n_1 = n_2$). The results are based on 10 replicated data sets. The reported MSE is the MSE for $E(\beta \mid data)$.

0.5.2 Scenario B

Simulation results for scenario B for various values of τ are considered. This scenario involves measurement error and fixed sample size (with $n_1 = n_2 = 80$). The results are based on 10 replicated data sets.

τ	$\widehat{E(\beta)}$	$\widehat{MSE(\beta)}$	Coverage Probability
0.25	0.5256	0.0022	0.1
0.5	0.4763	0.0052	0.3
0.75	0.4490	0.0097	0.3

Table 2: Simulation results for scenario B for various values of τ are considered, with $n_1 = n_2 = 80$. Means and MSEs of $\hat{\beta} = E(\beta \mid data)$ are reported, along with the coverage probability (CP) of the 95% equal-tailed credible interval for β. The true value is $\beta = 0.5$ throughout.

τ	$\widehat{E(X)}$	$\sqrt{\widehat{Var(X)}}$	Mean distance
0.25	1.9094	0.2543	2.3072
0.5	1.9842	0.4835	8.0644
0.75	1.9301	0.7559	6.4974

Table 3: Simulation results for scenario B for various values of τ are considered, with $n_1 = n_2 = 80$. Means and standard deviations of $\hat{X} = E(X \mid data)$ are reported, along with the Mean distance between the true. X and predicted X.

0.6 Conclusion

- In Scenario A involving no measurement error, we observe that the estimated value of effect β of the covariate is close to the true value. The estimated Mean squared error of β decreases as the sample size increases and the posterior of β yields samples with better accuracy i.e., the effect of covariate can be prominently observed as the sample size increases.

- Scenario B involves measurement error for a fixed sample size (reasonably large) the estimated Mean squared error of β increases as the variance of error increases and the expected value of β is not as close to the true value as in Scenario A. The method results in a coverage probability of the covariate effect around 30% when applied on only 10 replicated datasets. It is expected that this probability will increase as we increase the number of datasets.

- The mean distance between the true and estimated covariate values is quite small and they approximately resemble the distribution of the true covariate. Moreover, the estimated covariates from the posterior can be used in cases where we need the covariates but do not actually have it.

0.7 Contribution

- ***Soumya Paul:*** Provided theoretical inputs,Writing code (25%), Preparing report(50%), Presentation (35%)
- ***Sankhadeep Mitra:*** Finding the paper (50%), Writing code (50%), Presentation(35%),Preparing report (25%)
- ***Sampriti Dutta:*** Finding the paper (50%), Writing code (25%),Preparing report(25%), Presentation (35%)

Bibliography

[BUONACCORSI(1990)] J. P. BUONACCORSI. 1990. Double sampling for exact values in the normal discriminant model with application to binary regression. *Communications in Statistics* (1990), 4569â4586.

[Carroll et al.(1999)] R. J. Carroll, K. Roeder, and L. Wasserman. 1999. Flexible parametric measurement error models. *Biometrics* 55 (Mar 1999), 44–54. Issue 1.

[Dellaportas and Stephens(1995)] Petros Dellaportas and David A. Stephens. 1995. BAYESIAN ANALYSIS OF ERRORS-IN-VARIABLES REGRESSION MODELS. *Biometrics* 51 (1995), 1085–1095.

[Gilks et al.(1995)] W.R. Gilks, S. Richardson, and David Spiegelhalter (Eds.). 1995. Markov Chain Monte Carlo in Practice. (dec 1995). https://doi.org/10.1201/b14835

[Gustafson(2002)] P. Gustafson. 2002. A Bayesian approach to case-control studies with errors in covariables. *Biostatistics* 3, 2 (jun 2002), 229–243. https://doi.org/10.1093/biostatistics/3.2.229

[Gustafson P. Le N.(2000)] Valleqe M. Gustafson P. Le N. 2000. Parametric Bayesian analysis of case-control data with imprecise exposure measurements. *Statistics and Probability Letters* (2000), 357–363.

[Mallick and Gelfand(1996)] Bani K. Mallick and Alan E. Gelfand. 1996. Semiparametric errors-in-variables models A Bayesian approach. *Journal of Statistical Planning and Inference* 52, 3 (1996), 307–321. https://doi.org/10.1016/0378-3758(95)00139-5

[Muller(1997)] P Muller. 1997. A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika* 84, 3 (sep 1997), 523–537. https://doi.org/10.1093/biomet/84.3.523

[Muller et al.(1999)] Peter Muller, Giovanni Parmigiani, Joellen Schildkraut, and Luca Tardella. 1999. A Bayesian Hierarchical Approach for Combining Case-Control and Prospective Studies. *Biometrics* 55, 3 (sep 1999), 858–866. https://doi.org/10.1111/j.0006-341x.1999.00858.x

[PRENTICE and PYKE(1979)] R. L. PRENTICE and R. PYKE. 1979. Logistic disease incidence models and case-control studies. *Biometrika* 66, 3 (1979), 403–411. https://doi.org/10.2307/2335158

[Richardson and Gilks(1993)] Sylvia Richardson and Walter R. Gilks. 1993. A Bayesian Approach to Measurement Error Problems in Epidemiology Using Conditional Independence Models. *American Journal of Epidemiology* 138, 6 (sep 1993), 430–442. https://doi.org/10.1093/oxfordjournals.aje.a116875