Dealing with Measurement Error in Covariates with Special Reference to Logistic Regression Model: A Flexible Parametric Approach

by

Md. Shahadut Hossain

M.Sc., University of British Columbia, 2003

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy in THE FACULTY OF GRADUATE STUDIES (Statistics)

The University of British Columbia

November 2007

© Md. Shahadut Hossain, 2007
Abstract

In many fields of statistical application the fundamental task is to quantify the association between some explanatory variables or covariates and a response or outcome variable through a suitable regression model. The accuracy of such quantification depends on how precisely we measure the relevant covariates. In many instances, we can not measure some of the covariates accurately, rather we can measure noisy versions of them. In statistical terminology this is known as measurement errors or errors in variables. Regression analyses based on noisy covariate measurements lead to biased and inaccurate inference about the true underlying response–covariate associations.

In this thesis we investigate some aspects of measurement error modelling in the case of binary logistic regression models. We suggest a flexible parametric approach for adjusting the measurement error bias while estimating the response–covariate relationship through logistic regression model. We investigate the performance of the proposed flexible parametric approach in comparison with the other flexible parametric and nonparametric approaches through extensive simulation studies. We also compare the proposed method with the other competitive methods with respect to a real–life data set. Though emphasis is put on the logistic regression model the proposed method is applicable to the other members of the generalized linear models, and other types of non–linear regression models too. Finally, we develop a new computational technique to approximate the large sample bias that my arise due to exposure model misspecification in the estimation of the regression parameters in a measurement error scenario.
# Table of Contents

Abstract .................................................................................................................. ii

Table of Contents .................................................................................................. iii

List of Tables ........................................................................................................... viii

List of Figures ......................................................................................................... ix

Acknowledgement ................................................................................................... xii

Dedication .............................................................................................................. xiv

1 Measurement Errors in Covariates ............ 1

1.1 Introduction ...................................................................................................... 1

1.2 A Typical Measurement Error Scenario ....................................................... 2

1.3 Review of Literature on Adjusting the Impact of Measurement Errors .......... 4

1.4 Comparative Discussion of the Methods Dealing with Measurement Error Problems 5

1.5 Conditional Independence Models for Measurement Error Problem .......... 9
1.5.1 Formulation of the Posterior Distribution under Conditional Independence Modeling Assumption in the Presence of Measurement Errors . . . 11

1.6 Specification of the Exposure Model ............................. 13

1.7 Discussion ......................................................... 15

2 Impact of the Misspecified Exposure Model 16

2.1 Introduction ....................................................... 16

2.2 Simulation Design ............................................... 17

2.2.1 Simulation set-up ............................................. 17

2.2.2 Prior Distributions for the Model Parameters ................. 19

2.3 Results ............................................................. 20

2.4 Discussion .......................................................... 26

3 Flexible Exposure Models to Adjust the Effect of Measurement Errors in Logistic Regression 28

3.1 Introduction ....................................................... 28

3.2 Skew-normal Distribution and Its Variants ...................... 30

3.3 Performance of Skew-normal Exposure Model: Simulation Studies ........... 31

3.4 Measurement Error Models for Logistic Regression .............. 32

3.4.1 Joint and Conditional Posterior Distributions of the Model Parameters . 34

3.4.2 Simulation Set-up ............................................. 35
3.4.3 Monitoring the Convergence of MCMC Algorithm for the Model Parameters 36
3.4.4 Convergence of the Shape Parameter $\omega$ and the Scale Parameter $\lambda^2$ . . . 36
3.4.5 Convergence of the Regression Parameters $\beta$ ................................. 40
3.5 Results and Discussion ................................................................. 46

4 More Flexible Exposure Models: Flexible Extensions of Skew–normal Distribution 53
4.1 Introduction ................................................................. 53
4.2 Flexible Generalized Skew–normal Distribution ................................. 54
4.2.1 Joint and Conditional Posteriors and the MCMC Algorithm under FGSN Exposure Model ................................................................. 55
4.3 Flexible Generalized Skew–t Distribution ........................................ 56
4.3.1 Definition and the Basic Characteristics of FGST Distribution .......... 62
4.3.2 Parameter Estimation with FGST Distribution ................................ 63
4.4 Results ........................................................................... 67
4.5 Discussion .................................................................... 75

5 Performance of the Proposed Flexible Parametric Approach Compared to other Flexible Parametric and Popular Non-parametric Approaches 76
5.1 Introduction ................................................................. 76
5.2 FGST Exposure Model versus Normal Mixture Exposure Model ........ 77
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2.1</td>
<td>Simulation Study Designs</td>
<td>78</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Implementation, and Investigation of the Performance of the MCMC Algorithms</td>
<td>78</td>
</tr>
<tr>
<td>5.2.3</td>
<td>FGST Exposure Model versus Normal-Mixture Exposure Model: Results of the Simulation Studies</td>
<td>85</td>
</tr>
<tr>
<td>5.2.4</td>
<td>Application to the Study of Cholesterol and Coronary Heart Disease</td>
<td>88</td>
</tr>
<tr>
<td>5.3</td>
<td>Flexible Parametric Approach versus Parametric Correction (PC) and Non-Parametric Correction (NPC) Approaches</td>
<td>89</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Flexible Parametric Approach versus PC and NPC Approaches: Simulation Studies</td>
<td>93</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Posterior Conditionals and Estimation of the Model Parameters</td>
<td>94</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Results</td>
<td>96</td>
</tr>
<tr>
<td>6</td>
<td>A New Computational Method to Evaluate Large Sample Bias in the Estimate of Regression Coefficients due to Exposure Model Misspecification in the Presence of Measurement Error</td>
<td>102</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>102</td>
</tr>
<tr>
<td>6.2</td>
<td>Mathematical Details of the New Computational Technique</td>
<td>104</td>
</tr>
<tr>
<td>6.3</td>
<td>Application of the New Computational Technique: Simulation Studies</td>
<td>111</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Results</td>
<td>111</td>
</tr>
<tr>
<td>6.4</td>
<td>Discussion</td>
<td>113</td>
</tr>
</tbody>
</table>
7 Discussion and Scope of Future Research 115

7.1 Discussion ................................................................. 115

7.2 Future Research Work .................................................. 117

Bibliography 119
List of Tables

4.1 MSE of the estimates of $\beta_1$ ........................................... 74

5.1 Sensitivity to exposure model misspecification: comparison of FGST and Gaussian-mixture exposure models in the bimod case ........................................... 86

5.2 Sensitivity to exposure model misspecification: comparison of FGST and Gaussian-mixture exposure models in skew case ........................................... 87

5.3 Sensitivity to exposure model misspecification: comparison of FGST model with PC and NPC methods in single-covariate logistic regression model ........................................... 98

5.4 Sensitivity to exposure model misspecification: comparison of FGST model with PC and NPC methods in double-covariate logistic regression model ........................................... 100

6.1 Sensitivity to exposure model misspecification: comparison of repeated sample results with those obtained by the suggested large sample method ........................................... 113

viii
List of Figures

2.1 Tail behavior of log-normal distribution. The solid curve represents the pdf of $X \sim LN(0, \lambda^2 = 1)$, the dashed curve represents the pdf of $X \sim LN(0, \lambda^2 = 0.5)$, and the dotted curve represents the pdf of $X \sim LN(0, \lambda^2 = 0.25)$ 18

2.2 Empirical kernel densities for the estimates of regression parameters under simulation setting 1 21

2.3 Empirical kernel densities for the estimates of regression parameters under simulation setting 2 22

2.4 Empirical kernel densities for the estimates of regression parameters under simulation setting 3 23

2.5 Empirical kernel densities for the estimates of regression parameters under simulation setting 4 24

2.6 Empirical kernel densities for the estimates of regression parameters under simulation setting 5 25

3.1 Convergence plots for $\omega$ and $\lambda^2$ under simulation set-up 1 38

3.2 Convergence plots for $\omega$ and $\lambda^2$ under simulation set-up 2 38

3.3 Convergence plots for $\omega$ and $\lambda^2$ under simulation set-up 3 39

3.4 Convergence plots for $\omega$ and $\lambda^2$ under simulation set-up 4 39
3.5 Convergence plots for $\omega$ and $\lambda^2$ under simulation set-up 5 ................. 40

3.6 Convergence plots of $\beta$ under simulation set-up 1 .......................... 41

3.7 Convergence plots of $\beta$ under simulation set-up 2 .......................... 42

3.8 Convergence plots of $\beta$ under simulation set-up 3 .......................... 43

3.9 Convergence plots of $\beta$ under simulation set-up 4 .......................... 44

3.10 Convergence plots of $\beta$ under simulation set-up 5 .......................... 45

3.11 Kernel densities of the estimates of regression parameters under simulation design 1 47

3.12 Kernel densities of the estimates of regression parameters under simulation design 2 48

3.13 Kernel densities of the estimates of regression parameters under simulation design 3 49

3.14 Kernel densities of the estimates of regression parameters under simulation design 4 50

3.15 Kernel densities of the estimates of regression parameters under simulation design 5 51

4.1 Kernel densities of the estimates of regression parameters under simulation design 1 57

4.2 Kernel densities of the estimates of regression parameters under simulation design 2 58

4.3 Kernel densities of the estimates of regression parameters under simulation design 3 59

4.4 Kernel densities of the estimates of regression parameters under simulation design 4 60

4.5 Kernel densities of the estimates of regression parameters under simulation design 5 61

4.6 Kernel density plots of the estimates of regression parameters under simulation
  design 1 ................................................................. 69

4.7 Kernel density plots of the estimates of regression parameters under simulation
  design 2 ................................................................. 70
4.8 Kernel density plots of the estimates of regression parameters under simulation design 3 ................................................................. 71

4.9 Kernel density plots of the estimates of regression parameters under simulation design 4 ................................................................. 72

4.10 Kernel density plots of the estimates of regression parameters under simulation design 5 ................................................................. 73

5.1 MCMC plots for degree of freedom v in the bimod case ......................... 81

5.2 MCMC plots for degree of freedom v in the skew case .......................... 82

5.3 Histogram and empirical density estimate (superimposed) of the true covariate $X$ in the skew case ......................................................... 84
Acknowledgements

First of all, I thank my advisor, Paul Gustafson, for his support, both academic and financial, and continuous guidance that was the key to the completion of this thesis. Paul is a great mentor, and I feel privileged and honored to have had the chance to work with him over many years.

It has been a great opportunity for me over the past few years to have had the chance to come across with many outstanding faculty at Statistics Department at UBC. Their great mentorship, enthusiasm and approachability create very congenial environment for learning and developing new ideas. Therefore, I would like to extend my thanks to all the faculty members at the Department of Statistics, UBC. I am indebted to the office staff, Christine Graham, Rhoda Morgan, Elaine Salameh, and Peggy Ng for all their help with administrative matters. Finally, I express my sincere thanks to Lang Wu and Nhu Le for their being in my thesis committee, and for their helpful comments on my dissertation.

On a personal note, I would like to express my gratitude to my family. First, to my parents for providing me with all the opportunities up to early years of my university education despite all the limitations they had in those days.

Finally, my wife, Hasina Akther, deserves very special thanks for her all-out support, mental courage throughout my entire study period at UBC. My little daughter, Purbita Hossain, always brings endless joy and happy moments in our lives. I look forward to spending more joyous moments with my family in the time to come.
To my Parents, my wife Hasina, and my daughter Purbita.
Chapter 1

Measurement Errors in Covariates

1.1 Introduction

In numerous statistical applications quantification of the association between the explanatory variables (or covariates) and response is the main objective of the research. For example, in biostatistics and epidemiology the responses are usually the disease outcomes and the explanatory variables are human exposures of different kinds. Statistical models, e.g., different types of regression models, are applied to quantify the association between such disease outcomes and the suspected human exposures. The accuracy of such quantification depends on the quality of measurements of the covariates. In many fields of statistical applications imperfectly measured covariates are very common phenomena. Common examples are nutritional, environmental and occupational epidemiology, where some covariates are difficult to measure directly, and only surrogate measures of them are available. In statistical terminology this is known as measurement errors or errors in covariates when the explanatory variable or the exposure in question is continuous in nature. The similar phenomenon in the case of categorical covariates (e.g. ethnicity, gender etc.) is termed as misclassification. To encompass the imperfection in the measurements of both types of covariates the term "mismeasurement" is used in Gustafson (2004).

In the presence of mismeasurement in explanatory variables, modeling association between response and explanatory variables based on the surrogate measurements produces biased estimates of the true underlying association. The degree and direction of such bias depends on various aspects, e.g., the degree of association between the true covariates, the type of measurement error present with respect to the response, etc. Gustafson (2004) investigates various
aspects of biasedness in the estimates of regression coefficients in the presence of measurement errors in covariates. In multivariate models the presence of measurement error in covariates poses a more serious problem. In this case, the association between the response and a precisely measured covariate becomes biased if the precisely measured covariate is correlated with the error-prone covariate. In short, mismeasurement in covariates distorts the regression relationship and thus obscures the true underlying picture of the response–covariate association. Furthermore, the presence of such mismeasurement not only biases the estimates of regression coefficients of the outcome model used to quantify the association between the response and the covariates, but also causes the confidence limits of the coefficients to be artificially narrow. That is, the presence of mismeasurement in the explanatory variables leads to biased and falsely precise inference about underlying disease–exposure relationship.

In this dissertation, we investigate particular aspects of mismeasurement in explanatory variables. We particularly investigate the impact of the misspecification of a particular model (known as exposure model) while adjusting the bias induced by the mismeasurement in the explanatory variables. We also investigate the applicability of some flexible models for the true but unobserved exposure. Throughout this dissertation, we focus on the mismeasurement in continuous explanatory variables only. That is, in this dissertation we deal with errors in covariates or simply measurement errors. We mainly focus on the logistic regression model for a binary outcome, though the methodology we investigate is unified and can be extended or adapted to the other regression models too, such as Poisson regression model, Cox regression model, and many other non-linear regression models.

1.2 A Typical Measurement Error Scenario

Since in a measurement error scenario some of the covariates can not be measured precisely, the types of data formats required to carry out inference need some degree of elaboration. For the typical scenario, we consider a regression situation relating a response, $Y$, to the true predictors $X$ and $Z$. But instead of observing $(Y, X, Z)$ we observe $(Y, X^*, Z)$, where $X^*$ is a noisy version of the true covariate $X$. We assume that $Z$ is measured precisely. Here, the interest is to relate $Y$ to $X$ and $Z$, that is, the parameters of the joint distribution of $(Y, X, Z)$ are of interest. However, it is not possible to estimate the parameters of the joint distribution of $(Y, X, Z)$ on the basis of the information contained in the observed sample on $(Y, X^*, Z)$. Some features of the joint distribution of $(Y, X, X^*, Z)$ must be known or be estimated to be able to estimate the parameters of interest consistently. More specifically, the joint distribution of $(Y, X, X^*, Z)$
can be decomposed as

\[ f(y, x, x^*, z|\theta) = f(y|x, z, \theta_R)f(x^*|x, z, y, \theta_M)f(x|z, \theta_E), \]

(1.1)

where \( \theta = (\theta_R, \theta_M, \theta_E) \) is the vector of the model parameters. In measurement error terminology, the first component in the right hand side of (1.1) is called the outcome model, the second one is known as the measurement model, and the third component is known as the exposure model, respectively. To gain complete identifiability of the parameters of joint distribution \( f(y, x, x^*, z|\theta) \) one must need complete knowledge of the conditional distribution \( f(x^*|x, z, y, \theta_M) \). That is, to be able to estimate the parameter vector \( \theta \) consistently we need to know the distribution of the measurement error completely, given the true covariate \( X \), with \( Z \) and \( Y \) possibly influencing the process. The details about the parameter identifiability in the context of covariate measurement errors are provided in Gustafson (2004). If the parameter of the measurement model is not known completely then different sets of values of the parameter vector \( \theta \) will lead to same marginal distribution \( f(y, x^*, z|\theta) \) of \( (Y, X^*, Z) \). That is, unique estimate of the model parameters \( \theta \) does not exist and thus leaving us with the problem of parameter non-identifiability. Therefore, unknown parameters of the conditional distribution of \( (X^*|X, Z, Y) \) must be known or be estimated from the data for the consistent estimate of the model parameter \( \theta \). For the complete knowledge about the conditional distribution of \( (X^*|X, Z, Y) \) additional data are needed or \( \theta_M \) must be known in advance. So, in measurement error scenario, in order to have parameter identifiability additional information in any of the following three formats are needed

(i) observed sample on \( (Y, X^*, Z) \) with \( \theta_M \) known from previous studies

(ii) having repeated measurements on \( X^* \) for some of the study subjects. That is, having measurements in the format \( (y_i, x^*_{i1}, x^*_{i2}, \ldots, x^*_{im_i}, z_i) \), where, \( x^*_{i1}, x^*_{i2}, \ldots, x^*_{im_i} \) are the replicates on the surrogate \( X^* \) for the \( i \)-th study subjects. Note that, the replicated surrogates are true replicate in the sense that they are conditionally independent given the true covariate \( X \).

(iii) \( X \) is measured itself for a randomly selected subset of the study subjects, in addition to \( (Y, X^*, Z) \). The subset on which \( X \) is measured is referred to as validation sample. So, for the validation sample measurements are observed on \( (Y, X, X^*, Z) \).
1.3 Review of Literature on Adjusting the Impact of Measurement Errors

In view of the deleterious impact of treating the mismeasured (surrogate) covariates as if they were true covariates, there has been growing interest in the development of statistical methods to correct it. Such type of correction is necessary to assure the scientific validity of the estimate of true underlying regression relationship.

In the literature, methods dealing with measurement error problems differ from various perspectives. The main difference lies in the strength of modeling assumptions that are made. From this perspective, the existing literature on measurement error problem can be categorized as structural (or fully parametric) and functional approaches. The difference between structural and functional approaches is mainly due to the specification of the distribution of true but unobserved exposure. As an aside, it is already mentioned that for all measurement error analyses usually three submodels are considered—the outcome model relating the outcome to the true exposure, the measurement model relating the mismeasured exposure to the true exposure, and the model for the true exposure. The details of the component models in measurement error scenario are provided in Section 1.5.

Since, in the presence of covariate measurement errors the true covariate is not observed, it is very hard to check empirically whether a postulated model is reasonable for the true but unobserved exposure. Hence, some researchers advocate for trying to avoid any parametric specification of the exposure distribution. Therefore, a substantial volume of research for adjusting the impact of covariate measurement error is based on functional approaches. In the context of functional approaches, the regression calibration (RC) is quite straightforward and widely used. Carroll et al. (1995) provides a detailed and comprehensive discussion of regression calibration. Rosner et al. (1989), and Carroll and Stefanski (1990) studied the extension of regression calibration to the case of logistic regression and a general class of regression models. Simulation—extrapolation (SIMEX) is another popular functional approach for bias correction in regression parameters. SIMEX was developed by Cook and Stefanski (1995) and summarized in detail in Carroll et al. (1995). The conditional score approach (Stefanski and Carroll, 1987) and the corrected score approach (Nakamura, 1990) are the two other attractive alternatives in the functional approach paradigm. The idea behind the corrected score approach is to remove the bias by correcting the biased score function that results due to measurement errors. On the other hand, the idea behind the conditional score approach is to condition away the nuisance random parameters (e.g., true but unobserved exposure in the case of measurement error prob-
lems) based on sufficient statistics. Recently, Huang and Wang (2001, 2000) develop parametric and non-parametric correction approaches for logistic and Cox's regression models. Other functional methods that are worth mentioning are modified GLIM procedure (Armstrong, 1985), instrumental variable approach (Fuller, 1987), and semi-parametric likelihood methods (Hu, Tsiatis and Davidian, 1998; Song, Davidian and Tsiatis, 2002b; Roeder et al., 1996).

Structural approaches are those approaches which assume full parametric specifications for all the three submodels in measurement error problem. The maximum likelihood (ML) approach and the Bayesian approach based on conditional independence models (Richardson and Gilks, 1993) are the examples of structural approaches. In the context of measurement error problem, the advantage of likelihood methods relative to functional approaches have been studied in Schafer and Purday (1996), and Kuchenhoff and Carroll (1997). The Bayesian approach is different from both the functional and structural approaches mentioned so far. Section 1.4 provides a comparative discussion about the methods used in the context of measurement error problem with special emphasis on the differences between Bayesian and non-Bayesian approaches.

1.4 Comparative Discussion of the Methods Dealing with Measurement Error Problems

It has been already mentioned that the methods dealing with the problem of measurement errors in covariates can be categorized from two different perspectives—(i) structural versus functional approaches from the viewpoint of model formulation, and (ii) Bayesian versus non-Bayesian approaches from the viewpoint of the method of inference and computation. Functional approaches are appealing because they provide reasonable inference regardless of the distribution of true exposure. On the other hand, structural approaches can yield large gain in efficiency and allow construction of likelihood based confidence intervals (credible intervals in case of Bayesian approach) that have coverage probabilities closer to the nominal level than large sample normal theory intervals used with the functional methods. The biggest advantage of structural approaches is that they are likelihood based approaches and hence can be applied to more general problems, including those with discrete covariates subject to misclassification. However, they require distributional assumption about true but unobserved exposure around which much of the criticism of the structural approaches revolves. Therefore, some flexible parametric models need to be explored in order to achieve robustness of the parameter estimates against possible
misspecification of the exposure model. If such robustness can be achieved through flexible parametric modeling for the true but unobserved exposure, the structural model can be much more appealing because of its general applicability, computational simplicity and potential gain in efficiency.

The second aspect of comparison among the methods dealing with measurement error problem is based on whether prior specifications are made for the unknown model parameters. Bayesian approaches assume prior distributions for the unknown model parameters but frequentist approaches (both structural and functional) do not assume such specifications. The main advantage of Bayesian approaches is that they can incorporate the proper parameter uncertainty into account in estimation and inferential processes. Even if not much prior information is available and consequently diffuse priors are used, Bayesian methods can still take the amount of uncertainty supported by the data into account by averaging over all plausible values of the parameters of interest in the light of sample data. In frequentist approaches there is less opportunity to take the parameter uncertainty into account. Bayesian approaches also provide a unified framework allowing propagation of uncertainty from one level to another level in the hierarchy that arise in measurement error modeling. By taking parameter uncertainty into account and by propagating the uncertainty from one level to another level of the hierarchy the Bayesian approach can provide appropriate amount of uncertainty assessment about the model parameters and thus can provide better calibrated inference about them. Furthermore, like classical approach, Bayesian approach also provides the flexibility to accommodate non-parametric or flexible parametric specification of the exposure distribution (Gustafson, 2004). In classical/frequentist literature the nearest cousin to Bayesian approach is the maximum likelihood (ML) method. Besides providing proper uncertainty assessment the Bayesian approach also provides computational advantages over ML approach in measurement error scenario. Gustafson (2004) provides a detailed comparative discussion of the ML and the Bayesian approaches in the context of measurement errors.

Among the functional approaches, though RC and SIMEX are simple they have many limitations. For example, RC approach can be ineffective in reducing the measurement error induced bias while estimating the disease-exposure relationship through non-linear regression models when (a) the effect of the covariate $X$ on the outcome $Y$ is large, i.e. large odds ratios in logistic regression; or (b) the measurement error variance is large. SIMEX incrementally adds measurement errors to the surrogate $X^*$ using computer-simulated random errors and computes the corresponding regression parameter estimate (simulation step). The extrapolation step models the relation between the parameter estimate ($\hat{\theta}$) and the magnitude of the measurement errors
(manifested by measurement error variance $\tau^2$). SIMEX estimate is the extrapolation of this relation to the case of zero measurement error. Since, for SIMEX a functional form relating the parameter estimates $\hat{\theta}$ to the corresponding measurement variances $\tau^2$ is required, it is critical to find a proper functional form to relate $\hat{\theta}$ to $\tau^2$. In addition, there is always a danger in extrapolating a function beyond the range where the function values are observed. For elaborate discussion about the limitations of RC and SIMEX readers are referred to Gustafson (2004).

Both the conditional score (CS) and the corrected score (CTS) approaches suffer from serious root finding failure as well as the problem of multiple roots, especially when the measurement error is substantial. For multiple covariate regression, the problem of root finding failure is even more serious. Furthermore, the corrected score approach is not applicable for logistic regression, and the conditional score approach is not consistent beyond normally distributed measurement errors. Conditional score equations are very complicated to work out for highly non-linear models. Considering the limitations of the existing methods, especially the dearth of widely applicable methods for logistic regression, Huang and Wang (2001) propose parametric and non-parametric correction approaches for logistic regression. The proposed methods can provide estimates which are consistent, asymptotically normal and locally efficient. Though the parametric and the non-parametric correction approaches are less problematic in terms of consistency and efficiency of the resulting estimates they still have some limitations. For example,

(a) they are not globally efficient; rather they are efficient in the class of estimators based on correction amenable estimating equations

(b) construction of parametric and non-parametric estimating equations is mathematically intensive

(c) both the methods are not based on likelihood function. In the absence of a likelihood function the only choice of testing hypothesis about regression parameter $\beta$ of the outcome model is the Wald test, which behaves aberrantly in logistic regression when $|\hat{\beta}|$ is large (Hauck and Donner, 1977). So, extra effort, e.g. bootstrap, is required to conduct test and construct confidence intervals for $\beta$

(d) non-parametric correction approach is applicable only when replicate measurements are available for the mismeasured covariates on the entire study sample. But this is not always the case in practice. Sometimes replicate measurements are available only on a sub-sample of the cases. In some other cases, even no replicate measurements are available, rather a small validation sample is available
simulation results in Huang and Wang (2001) reveal that the finite sample performances of parametric and non-parametric correction approaches are not always satisfactory in logistic regression model. In fact, simulation studies demonstrate that in almost all the scenarios considered in Huang and Wang (2001) both the methods appear to be less efficient than the RC method. The loss of efficiency increases with the magnitude of regression parameters of the imprecisely measured covariate. Furthermore, in many cases both the methods appear to be less effective than RC approach in bias correction. Their simulation results also reveal that the bias correction ability of their proposed methods decreases as the correlation coefficient between the true covariates increases from 0 to 0.5. In their study Huang and Wang (2001) consider a maximum correlation of 0.5 between the two covariates. But in practice the correlation among the covariates may be much stronger than this, and in those cases the bias correction performance of their proposed methods may deteriorate substantially.

The issue of taking the correlation among the true exposures into account is crucial because existence of such correlation increases the bias in the estimates of the regression coefficients in the presence of measurement errors. By being devoid of likelihood function the CS approach, the parametric and non-parametric correction approaches (all are based on unbiased estimating function rather than likelihood function) cannot take the correlation among the covariates into account. It is very intuitive to argue that since the existence of correlation among covariates results in stronger biases, the extent to which these methods can correct the extra bias induced by the correlation among the covariates is a cause of concern. A Bayesian structural approach based on a conditional independence model has the potential to overcome most of the limitations of the existing functional approaches. Bayesian structural approach uses likelihood function and hence can incorporate correlation among the true covariates even if it considers non-parametric approach to model the exposure distribution. When a parametric exposure model is assumed and is correct, it provides estimates of the regression parameters that have all the nice large sample properties of maximum likelihood estimators in addition to providing better uncertainty assessment, computational ease and exact, rather than approximate, inference. When the exposure distribution is misspecified, both the Bayesian structural approach and the ML approach may provide inconsistent estimates. Therefore, we are in need of a flexible parametric model for the true but unobserved exposure.

Furthermore, it is always a question that how well the existing functional methods perform in practice, especially, in finite sample cases and in the presence of correlations among the covariates. There is also the question of efficiency of the estimates. Therefore, studying how well
the competitive methods can achieve a trade-off between consistency and efficiency is worthwhile. Particularly, the relative performance of structural Bayesian approach with misspecified exposure model and the functional approaches in different finite sample scenarios is worth investigating. If structural Bayesian approach with misspecified exposure model is found to perform poorly then another option is to investigate the relative performance of Bayesian approach with flexible exposure distribution and those of the popular functional approaches (e.g., parametric and non-parametric correction approaches).

We emphasize using structural models for exposure distribution simply because if the estimates of the regression parameters are insensitive to the misspecification of the exposure distribution then by being structural we can gain in efficiency of the estimates. Another appealing property of the structural Bayesian approach is its flexibility to incorporate additional complexity of epidemiological designs, e.g., incorporating subject-specific random effects and incorporating the covariate measurements obtained through several instruments (Richardson and Gilks, 1993) and the case of differential measurement error. By differential measurement error we mean that the measurement process depends not only on the true but unobserved exposure $X$, but also on the outcome $Y$ and the other precisely measured covariate $Z$. This means that the distribution of the mismeasured exposure $X^*$ given the associated true exposure $X$ is dependent on the outcome $Y$ and the precisely measured covariates $Z$. None of the functional approaches offer an obvious flexibility to incorporate such complicated design set-up in the presence of measurement errors. Even though ML method can incorporate such complicated data structure, more and more approximation is required to implement it and consequently the finite sample performance of ML estimate becomes questionable. For example, in the case of random effects model the approximate ML method for non-Gaussian response may produce a variance-covariance matrix of the estimates which is not positive definite (Breslow and Clayton, 1993) even in the case of no measurement errors.

1.5 Conditional Independence Models for Measurement Error Problem

Since the Bayesian approach to measurement error problem is based on conditional independence model, we elaborate the concept of such model in this section. Models formulated based on the conditional independence relationship are known as conditional independence models (Richardson and Gilks, 1993). A typical conditional independence model in measurement error
scenario involves three submodels (Clayton, 1992). Let $Y$ be the response variable, and let $X$ be the covariate which is not observed but a surrogate $X^*$ is observed instead. Again, let $Z$ be the vector of precisely measured covariates. If the measurement error mechanism is viewed in terms of the distribution of the surrogate $X^*$ given true $X$, the joint density of all the relevant variables given in (1.1) can be factored as

$$f(y, x, x^*, z; \theta_R, \theta_M, \theta_{E_1}, \theta_{E_2}) = f(y \mid x, z, \theta_R) f(x^* \mid x, y, z, \theta_M) \times f(x \mid z, \theta_{E_1}) f(z \mid \theta_{E_2})$$

(1.2)

where $\theta = (\theta_R, \theta_M, \theta_{E_1}, \theta_{E_2})$ is the vector of the model parameters. The first term on the right hand side of (1.2), the conditional density of $(y \mid x, z)$, known as the response model (outcome model), describes the relationship between the response $Y$ and the true explanatory variables $(X, Z)$. The second term $f(x^* \mid x, y, z)$ is known as the measurement model, which describes how the surrogate $X^*$ arises from the true explanatory variable $X$, with $Y$ and $Z$ possibly influencing the process. Often it is assumed that $f(x^* \mid x, y, z, \theta_M) = f(x^* \mid x, \theta_M)$, i.e. given the true explanatory variable $X$, the distribution of the surrogate $X^*$ does not depend on the response $Y$ and the vector of precisely measured explanatory variables $Z$. In measurement error terminology, this kind of measurement error is known as non-differential measurement error. The third and fourth term together give the joint distribution of $(X, Z)$ and is known as exposure model with parameter vector $\theta_E = (\theta_{E_1}, \theta_{E_2})$. In practice, for the purpose of inference of the parameters in the outcome model the fourth term does not play any role. The elaboration on this point is provided in Section 1.5.1. So, by the term “exposure model” we usually mean the third term, which is the conditional distribution of $X$ given $Z$.

If the vector $(Y, X, X^*, Z)$ were observed for each study subject then the likelihood function could be constructed from (1.2) by taking the products over all the study subjects. But in reality only the vector $(Y, X^*, Z)$ is observed, so that

$$f(y, x^*, z; \theta_R, \theta_M, \theta_{E_1}, \theta_{E_2}) = \int f(y, x, x^*, z; \theta_R, \theta_M, \theta_{E_1}, \theta_{E_2}) dx$$

(1.3)

is required to form the likelihood. In some problems, e.g., for binary response variable, the above integral does not have a closed form. Bayesian MCMC approach provides an easy route to evaluate the integral in (1.3). To evaluate such integral the Bayesian MCMC method does not require (1.3) explicitly. Instead it works with (1.2) and evaluates the integral required in (1.3) implicitly.
1.5.1 Formulation of the Posterior Distribution under Conditional Independence Modeling Assumption in the Presence of Measurement Errors

Assume that there are \( n \) study subjects whose observations are independent of each other. Using the similar notation used in the previous section the joint distribution of all the relevant variables is already defined in equation (1.2). Throughout, this dissertation we only assume non-differential measurement error. By non-differential measurement error we mean that the distribution of measurement errors depends only on the true covariate \( X \), not on the outcome \( Y \) and the precisely measured covariate \( Z \). However, as already mentioned, the Bayesian structural approach can handle the differential measurement error too. Under the assumption of non-differential measurement error, the measurement model defined in equation (1.2) can be written as \( f(x^*|x,y,z,\theta_M) = f(x^*|x,\theta_M) \). As it is mentioned in section 1.2, for parameter identifiability additional information about parameter \( \theta_M \) of the measurement model can be obtained in three formats. The first scenario is one where the parameter \( \theta_M \) is known. This might occur if the measurement error mechanism is thoroughly investigated in previous external studies. The problems associated with using information from the external studies are discussed in Gustafson (2004). Under the assumption mentioned so far in this section, the joint distribution of all the relevant quantities is given as

\[
f(y,x,x^*,z,\theta) = \prod_{i=1}^{n} f(y_i|x_i,z_i,\theta_R)f(x_i^*|x_i)f(x_i,z_i|\theta_E) \times \pi(\theta_R,\theta_{E_1},\theta_{E_2}),
\]

(1.4)

where \( \pi(\theta_R,\theta_{E_1},\theta_{E_2}) \) is the prior distribution of the model parameters. Usually, prior independence of the model parameters is assumed. According to Bayes theorem, the density of unobserved quantities \( U \) given the observed quantities \( O \) is proportional to the joint density of \( U \) and \( O \). In the context of structural measurement error modelling we have \( U = (x,\theta_R,\theta_{E_1},\theta_{E_2}) \) and \( O = (x^*,y,z) \). Therefore, the posterior density of \( U \) given \( O \) is proportional to

\[
f(x,\theta|x^*,y,z) \propto \left[ \prod_{i=1}^{n} f(y_i|x_i,z_i,\theta_R)f(x_i^*|x_i)f(x_i,z_i|\theta_{E_1})f(z_i|\theta_{E_2}) \right] \pi(\theta_R,\theta_{E_1},\theta_{E_2}) \]

(1.5)

For actual normalized posterior density of the unobserved quantities \( U \), given the observed quantities \( O \), we must compute the integral of (1.5) over \( U \), given the fixed values of \( O \). Computing integral of (1.5) is not possible in closed form in the presence of measurement error in the cases other than linear regression model. However, to implement Bayes–MCMC approach one does not need to evaluate the normalized posterior density. For Bayes–MCMC approach (1.5) is enough to carry out inference about the model parameters \( \theta \). This is the biggest computational advantage of Bayes–MCMC inference in problems involving mismeasured, missing, or censored data over the ML and other approaches.
Now, in the case of prior independence of the model parameters, \( \theta_{E_2} \) in (1.5) is independent of \( (\theta_R, \theta_{E_1}) \). Therefore, \( \theta_{E_2} \) will also be independent of \( (\theta_R, \theta_{E_1}) \) a posteriori as (1.5) can be factored into a term not containing \( \theta_{E_2} \) and a term not containing \( (\theta_R, \theta_{E_1}) \). That is,

\[
f(x, \theta_R, \theta_{E_1}, \theta_{E_2} \mid x^*, y, z) = f(x, \theta_R, \theta_{E_1} \mid x^*, y, z)f(\theta_{E_2} \mid z),
\]

where

\[
f(x, \theta_R, \theta_{E_1} \mid x^*, y, z) \propto \left[ \prod_{i=1}^{n} f(y_i \mid x_i, z_i, \theta_R) f(x_i^* \mid x_i) f(x_i \mid z_i, \theta_{E_1}) \right] \times \pi(\theta_R, \theta_{E_1}), \tag{1.6}
\]

and

\[
f(\theta_{E_2} \mid z) \propto \left[ \prod_{i=1}^{n} f(z_i \mid \theta_{E_1}) \right] \times \pi(\theta_{E_2}). \tag{1.7}
\]

As long as the main interest is on the parameter \( \theta_R \) of the outcome model the marginal posterior \( f(\theta_{E_2} \mid z) \) does not contain any information about \( \theta_R \). Therefore, one can simply work with (1.6) and not make any modelling assumption about the precisely measured covariate \( Z \). This implies that we can treat \( f(x \mid z, \theta_{E_1}) \), instead of \( f(x \mid z, \theta_{E_1}, \theta_{E_2}) \), as exposure model. So, from now on, by exposure model we will mean \( f(x \mid z, \theta_{E_1}) \), the conditional distribution of unobserved covariate \( X \) given the other covariate \( Z \), which is measured precisely.

In the second situation of measurement error scenario, we assume that the measurement model is known up to a parameter \( \theta_M \), i.e. the measurement model \( f(x^* \mid x, \theta_M) \) has an unknown parameter \( \theta_M \). We also assume that for a randomly selected subset of the study subjects the true covariate \( X \) is itself measured, in addition to \((Y, X^*, Z)\). Commonly, this is referred to as validation sample. The subjects with \((Y, X, X^*, Z)\) measurements are called the complete cases or the validation sub-study, while the subjects with \((Y, X^*, Z)\) comprise the reduced cases. Validation designs are fairly common as for some exposure it is possible to make gold-standard measurements on small number of study subjects. Such gold-standard measurements are usually too expensive to be used for the entire study sample. Greenland (1988) and Spiegelman (1994) discuss issues related to cost-information tradeoff and the size of the gold-standard sample in relation to the main study sample. If, in the presence validation sample, we denote the complete cases by \( x_c \) and the reduced cases by \( x_r \), so that the entire study sample on the exposure \( X \) becomes \( x = (x_c, x_r) \), the posterior density for the unobserved quantities given the observed quantities then takes the form

\[
f(x_r, \theta_M, \theta_R, \theta_{E_1} \mid x^*, y, x_c, z) \propto \left[ \prod_{i=1}^{n} f(y_i \mid x_i, z_i, \theta_R) \right] \times \left[ \prod_{i=1}^{n} f(x_i^* \mid x_i, \theta_M) \right] \\
\times \left[ \prod_{i=1}^{n} f(x_i \mid z_i, \theta_{E_1}) \right] \times \pi(\theta_R, \theta_M, \theta_{E_1}). \tag{1.8}
\]
Though the right hand side of (1.8) does not appear to distinguish between the observed and unobserved component of \( x \), the MCMC sampling from (1.8) can still provide a principled way to make simultaneous inferences about \( \theta_M, \theta_R, \) and \( \theta_E_1 \).

The third and final scenario in measurement error problem arises when repeated measurements of \( X^* \) are available for at least some study subjects. Suppose, \( m_i \) replicate measurements are made for the \( i \)-th study subject. If we retain the non-differential measurement error model assumption parameterized by \( \theta_M \), and assume that the replicate measurement of \( X^* \) are conditionally independent given the true value of \( X \), then the posterior density of the unobserved quantities given the observed quantities can be derived as

\[
f(x, \theta_M, \theta_R, \theta_E_1 | x^*, y, z) \propto \prod_{i=1}^{n} f(y_i | x_i, z_i, \theta_R) \times \prod_{i=1}^{n} \prod_{j=1}^{m_i} f(x_{ij}^* | x_i, \theta_M) \times \prod_{i=1}^{n} f(x_i | z_i, \theta_E_1) \times \pi(\theta_R, \theta_M, \theta_E_1). \tag{1.9}
\]

Here \( x_{ij}^* \) is the \( j \)-th of the \( m_i \) replicated \( X^* \) measurements for the \( i \)-th study subject. It is worth noting that assessment of whether the replicated \( X^* \) measurements are really conditionally independent given \( X \) is a difficult task because \( X \) is not observed. Typically this cannot be verified empirically, and thus is subject to the consideration of scientific plausibility in the subject-area context at hand.

### 1.6 Specification of the Exposure Model

The main criticism of structural modeling approaches for exposure distribution lies in the fear for the misspecification of it and the subsequent impact of such misspecification on the parameters in the outcome model. Hence many authors strongly advocate for the functional approaches or some other flexible modeling approaches (e.g., using mixture models) for the true exposure distribution. In Bayesian paradigm Muller and Roeder (1997), Carroll Roeder and Wasserman (1999), and Richardson, Leblond, Jaussent and Green (2002) use mixture models for the exposure distribution.

Though in the literature much attention is paid on the specification of the exposure model, in measurement error problems the main objective lies in the parameters of the outcome model, which is one step apart from the exposure model. In the structural modeling hierarchy for adjusting for measurement error, the measurement model lies in between the outcome and the
exposure models. That is, the measurement model acts as a buffer between the outcome and the exposure models. So, it is intuitive to argue that the misspecification of the exposure models may have little impact on the parameters in the outcome model (Gustafson, 2004). Fuller (1987) and Gustafson (2004) show, in the case of linear outcome model, that regardless of the real shape of the exposure model the postulated normal exposure model suffices to give the same limiting estimates as would have been obtained by applying the response model alone if the true exposure were observed. They also show that in the case of linear response model the asymptotic variance of the estimates of the regression parameters depends only on the variance of the exposure model, not on the shape of the exposure distribution. Thus, in the case of linear outcome model, the estimates of the regression parameters remain asymptotically unaffected even though the exposure model is misspecified.

Unfortunately, unlike linear models, the effect of the misspecified exposure model can not be investigated in closed form for the responses arising from the other members of generalized linear models (GLM). But in the Statistics literature, it is believed that the phenomena arising in normal linear models tend to be approximately manifested in the members of GLM. If this is the case, then one can assume a simpler model (e.g. normal model) for the continuous exposure distribution and carry on with the analysis to adjust for the measurement errors. This way, there may not be any asymptotic bias in the estimates of the regression parameters of the outcome model. However, some recent work provides evidence against this idea in the case of logistic regression model. Richardson and Leblond (1997) show that in logistic regression model the use of normal exposure model produces biased estimates of the regression parameters of the outcome model when the exposure model is in fact not normal. A limitation in their study is that they consider the model for the exposure $X$ alone in the estimation procedure while they simulate data sets which involve positive correlation between the true exposure $X$ and the precisely measured covariate $Z$. So, they do not acknowledge a second aspect of misspecification by ignoring the correlation between $X$ and $Z$. Therefore, it is worthwhile to investigate the performance of the misspecified normal exposure model in correcting the measurement error induced bias by acknowledging dependence between the true unobserved covariate $X$ and the other precisely measured covariate $Z$. In Chapter 2, we investigate the mechanism of bias correction in the estimates of the regression coefficients in the outcome model when exposure model is misspecified as normal distribution. In the process of our investigation we consider different scenarios that may arise in practice.
1.7 Discussion

We devote this Chapter discussing the issues of covariate measurement error, their impacts on the estimates of the response–covariate relationship and the existing methodologies (both functional and structural) to minimize such impacts. We also draw a comparative picture between functional and structural approaches by mentioning the strengths of the structural approaches in the measurement error scenario. Provided that the robustness of the estimates of the outcome model parameters to the potential misspecification of the exposure model can be ensured through flexible parametric modelling for the unobserved exposure, structural approaches can offer more flexibility, efficiency gain and computational simplicity as compared to functional approaches. Finally, we carry on discussing the comparative advantages of Bayesian structural approach and the frequentist structural approach (e.g. ML approach). As already mentioned, the biggest advantage of Bayesian structural approach to covariate measurement error scenarios is its flexibility to incorporate complicated design set-up and easy adaptation of it in scenarios like longitudinal studies, non-linear models, etc. Extension of functional approaches to longitudinal studies and adaptation of them to highly non-linear models are still far off. Though structural frequentist approach based on ML estimation procedure can accommodate the complicated design set-up, their implementation remains a formidable challenge in the measurement error scenarios. Furthermore, in ML approach there is less scope for propagating uncertainty arising from estimating one parameter to another. Bayesian structural approaches can overcome all these problems. They are as flexible as the frequentist structural approaches in terms of flexibility to incorporate complicated data structure, including longitudinal studies and non-linear regression models. The extra advantage of Bayesian structural approaches is computational due to MCMC algorithms. Unlike ML approach, Bayesian approach can offer exact (exact in the sense that the Monte-Carlo error can be minimized by through large Monte-Carlo sample), not approximate, inferences for the parameters of interest. The main criticism of Bayes approach remains in the choice of prior distributions of the model parameters. Such criticism can be mitigated by choosing diffuse proper priors. With such priors the Bayes point estimates will be similar to those could be obtained by ML approach, but will still possess better uncertainty assessment, in addition to overwhelming computational advantage of Bayes approaches in measurement error scenarios. Keeping all these arguments in favor of Bayes structural approaches in mind, throughout this dissertation we stick to this approach as a method of our choice for adjusting measurement error induced bias in the estimate of underlying disease–exposure relationship.
Chapter 2

Impact of the Misspecified Exposure Model

2.1 Introduction

It is already noted in the earlier chapter that regardless of the real shape of the true exposure distribution, the normal exposure model suffices in eliminating the measurement error induced bias in the case of linear regression model. As it is believed that phenomena arising in normal linear models tend to be approximately manifested in generalized linear models as well, the insensitivity to exposure model misspecification arising with linear outcome models might be manifested in logistic outcome models too. Some recent work, however, speaks against this. Richardson and Leblond (1997) investigate the sensitivity of the estimates of the regression parameters of the logistic outcome model to exposure model misspecification. They consider a two-covariate logistic regression model for the binary outcome $Y$ given the imprecisely measured exposure variable (covariate) $X$ and a precisely measured covariate $Z$. For the unknown exposure they assume normal distribution while the true $X$ values are generated from distributions other than normal. They apply the Bayesian structural approach which is implemented through Bayes–MCMC method. Their results demonstrate that the estimates of the regression parameters are biased. That is, the estimated disease–exposure relationship are not robust to the exposure model misspecification in the case of logistic outcome model. Regarding the study of Richardson and Leblond (1997), we note that their results is based on only 10 simulated data sets. That is, the reported estimated regression parameters is the average of the estimates obtained from 10 simulated data sets. Results based on the averaging over this small number of data sets may be influenced by high sampling fluctuation, rather than bias. Therefore, it is worth investigating the sensitivity to exposure model misspecification with a large number
of data sets to minimize the effect of the sampling fluctuation on the estimated regression coefficients. Hence, we perform systemic simulation studies to investigate the bias correction mechanism in the estimated regression coefficients when the exposure model is misspecified. The simulation design, and the different simulation set-ups that are considered are outlined in the next two consecutive sections.

2.2 Simulation Design

For our simulation studies we consider a two-variable logistic regression model. One variable, $X$, is measured with error and the other one, $Z$, is measured accurately. Suppose, the outcome $Y_i$ follows a Bernoulli distribution with success probability $p_i$ so that we have,

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_i + \beta_2 Z_i. \quad (2.1)$$

Instead of $X$ a surrogate $X^*$, which is subject to measurement errors, is observed. Throughout our simulation studies we consider the simple case of unbiased and non-differential measurement error process, i.e. conditional on $X$ and $Y, Z$ the surrogate $X^*$ follows a normal distribution with mean $X$ and variance $\tau^2$:

$$X_i^*|X_i, Y_i, Z_i \sim N(X_i, \tau^2).$$

That is, in all our simulation settings, we consider the normal measurement error distribution. Furthermore, we assume that only the exposure model is misspecified. That is, in the process of parameter estimation through Bayesian structural approach we do assume that the distribution of measurement errors is normal as they are simulated from normal distribution.

2.2.1 Simulation set-up

For computational simplicity we assume that measurement process is completely known, i.e., the value of $\tau^2$ is known. All through our simulation studies, the response $Y$ is generated assuming a logistic regression model with parameters fixed at $\beta_0 = -1, \beta_1 = 1$ and $\beta_2 = 0.5$. The true exposure $X$, conditional on $Z$, is generated from different log-normal distributions, i.e., $X_i|Z_i \sim LN(\alpha_0 + \alpha_1 Z_i, \lambda^2)$, and $Z$ is generated from standard normal distribution. Different degrees of skewness and heavy-tailedness for the distribution of $X$ are considered, along with varying degrees of correlation $\rho_{XZ}$ between $X$ and $Z$. For the purpose of parameter estimation using
the structural Bayesian model the log-normal exposure distribution is misspecified as normal distribution. Following are the different simulation set-up for the true exposure distribution that are considered in this study.

1. \((a_0, a_1, \lambda^2) = (0, 0.3, 0.25)\). For this set-up \(\rho_{xz} = 0.47\);
2. \((a_0, a_1, \lambda^2) = (0, 0.75, 0.12)\) so that \(\rho_{xz} = 0.76\);
3. \((a_0, a_1, \lambda^2) = (0, 0.3, 1)\) so that \(\rho_{xz} = 0.39\);
4. \((a_0, a_1, \lambda^2) = (0, 0, 1)\), where \(\rho_{xz} = 0\);
5. \((a_0, a_1, \lambda^2) = (0, 0.6, 0.1)\). In this set-up, \(\rho_{xz} = 0.785\).

For all the simulation set-up the ratios of measurement error variance, \(\tau^2\), to the expected variance of true exposure, \(E[\text{var}(x|z)]\), are taken to be 0.20 and 0.40 to represent small and substantial amount of measurement errors, respectively. Note that larger values of \(\lambda^2\) gives log-normal distributions with high skewness and thicker tails. Figure 2.1 demonstrates how the tail behavior of log-normal distribution changes depending on the values of the scale parameter \(\lambda^2\). The density plots in Figure 2.1 demonstrate that larger \(\lambda^2\) values give rise to log-normal

![Figure 2.1: Tail behavior of log-normal distribution. The solid curve represents the pdf of \(X \sim LN(0, \lambda^2 = 1)\), the dashed curve represents the pdf of \(X \sim LN(0, \lambda^2 = 0.5)\), and the dotted curve represents the pdf of \(X \sim LN(0, \lambda^2 = 0.25)\).](image)
distributions with thicker tails.

Simulation set-up 1, represents a case where the log–normal distribution does not have any heavy–tailedness. In such case, the misspecified normal exposure model is expected to pick up the shape of the true exposure distribution quite well. Set–up 2 and 3 represent two contrasting situations with the former yielding a situation with log–normal distribution which has no thicker tail but strong correlation between the two covariates, and the latter representing a scenario with an exposure distribution with thicker tail but a weak correlation (0.39) between $X$ and $Z$. Similarly, set–up 4 and 5 are considered to represent two opposite situations with 4 representing an exposure distribution with very thick tail and zero correlation between $X$ and $Z$, while 5 represents the situation with a well–behaved log–normal exposure distribution but a very high correlation between $X$ and $Z$. These contrasting situations are considered to investigate whether strong correlation between the covariates or the wild observations (represented by the thicker tails) in the exposure distribution is the barrier to correct the bias in the regression parameters by using misspecified normal exposure distribution. In all, the simulation set–ups are considered to mimic different possible practical situations.

### 2.2.2 Prior Distributions for the Model Parameters

For the full Bayesian treatment, prior distributions for the unknown parameters $\alpha = (\alpha_0, \alpha_1), \beta = (\beta_0, \beta_1, \beta_2), \tau^2$ and $\lambda^2$ need to be specified. It is very common to assume the prior independence among the unknown parameters so that

$$f(\alpha, \beta, \tau^2, \lambda^2) = f(\alpha)f(\beta)f(\tau^2)f(\lambda^2).$$

For $\alpha$ and $\beta$, the common choice is the locally uniform priors, i.e., $f(\alpha) \propto 1$ and $f(\beta) \propto 1$. Locally uniform priors represent prior ignorance about the parameters of interest, i.e., these priors do not favor any particular value over the other values of the parameters of interest. This is appealing in the sense that in absence of any prior knowledge these priors allow the data completely to tell about the parameters of interest. For the variance components, $\tau^2$ and $\lambda^2$, diffuse proper priors are considered. In this study, unit information inverted gamma prior is considered for both $\tau^2$ and $\lambda^2$. That is, $\lambda^2 \sim IG(\frac{1}{2}, \frac{1}{2})$ and $\tau^2 \sim IG(\frac{1}{2}, \frac{1}{2})$, where $IG$ stands for inverse–gamma distribution. The centre of such prior distribution is roughly at $\frac{1}{1/2} = 1$ so that prior guesses for both $\lambda^2$ and $\tau^2$ are 1. Taking unit information inverse–gamma prior for the variance components is tantamount to considering an effective sample size of 1 for the priors for them. In other words, we are assuming that \textit{apriori} we have information worth a
single data point about the variance components. So, we are letting the data to take the main role in estimating the values of $\lambda^2$ and $\tau^2$, which is very natural when we have little prior information. The prior distributions considered above are all close to being non-informative, and hence we expect the results to be robust with respect to the prior distributions. The reason for choosing non-informative type of priors is that in absence of any plausible prior information it is reasonable to allow the data to take the prime role in drawing the inferences about the model parameters.

2.3 Results

In each case of the simulation set-up, 50 independent data sets, each of size $n = 500$, are generated. For each data set, to have posterior estimate of each of the regression parameters a posterior sample of size 1500 is generated through an appropriate MCMC algorithm. For the convergence and mixing of MCMC sampler we visually investigate sequences of the posterior draws for the regression parameters for few of the generated data sets under each simulation setting. These visual inspections confirm that in each of the cases considered the MCMC sampler converges well before 500 posterior draws. Also, each of the MCMC chains investigated exhibits good mixing. So, out of 1500 MCMC draws first 500 iterations are deemed to be sufficient as burn-in. The remaining 1000 draws are used to estimate the regression parameters of interest. For the parameter estimate, the posterior mean is used. To investigate the degree of bias in the estimated regression parameters we construct the kernel density plots of them. The kernel density estimators here actually estimates the sampling distribution of the estimators of the corresponding regression parameters of the logistic outcome model. That is, in the case of each regression parameter the corresponding kernel density plot shows the sampling distribution of the 50 estimates obtained based on 50 different samples. To have a comparative picture, we plot the kernel density of the estimates obtained under misspecified normal exposure model along with the kernel densities of the naive estimates and of the estimates based on the true values of the mismeasured exposure. The naive estimates are the estimates obtained by ignoring the measurement errors and treating the surrogate $X^*$ as if they were true $X$. The kernel density plots of the parameter estimates are displayed in Figures 2.2-2.6.

In each case, figures in panel (a) correspond to the case of low measurement errors and the figures in panel (b) correspond to the case of high measurement errors, e.g., figures in panel 2.2(a) and 2.2(b) in each plot represent the empirical kernel density estimates of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ for the cases of low and substantial measurement errors, respectively. Kernel density plots
Figure 2.2: Empirical kernel densities for the estimates of regression parameters under simulation setting 1. In each case the solid curve gives the empirical kernel density of the naive estimates, the dashed curve gives the kernel density of the estimates based on true exposure values, and the dotted curve gives the kernel density of measurement error adjusted estimates.
Figure 2.3: Empirical kernel densities for the estimates of regression parameters under simulation setting 2. In each case the solid curve gives the empirical kernel density of the naive estimates, the dashed curve gives the kernel density of the estimates based on true exposure values, and the dotted curve gives the kernel density of measurement error adjusted estimates.
Figure 2.4: Empirical kernel densities for the estimates of regression parameters under simulation setting 3. In each case the solid curve gives the empirical kernel density of the naive estimates, the dashed curve gives the kernel density of the estimates based on true exposure values, and the dotted curve gives the kernel density of measurement error adjusted estimates.
Figure 2.5: Empirical kernel densities for the estimates of regression parameters under simulation setting 4. In each case the solid curve gives the empirical kernel density of the naive estimates, the dashed curve gives the kernel density of the estimates based on true exposure values, and the dotted curve gives the kernel density of measurement error adjusted estimates.
Figure 2.6: Empirical kernel densities for the estimates of regression parameters under simulation setting 5. In each case the solid curve gives the empirical kernel density of the naive estimates, the dashed curve gives the kernel density of the estimates based on true exposure values, and the dotted curve gives the kernel density of measurement error adjusted estimates.
of the estimated parameters demonstrate that in simulation set-up 1, Bayesian adjustment based on misspecified normal exposure model seems to be adequate in both cases of low and substantial measurement errors. In both cases, the sampling distributions of the measurement error adjusted estimates and the estimates based on true exposure values are found to be centered around the true parameter values, while the kernel density for the naive estimates centers around a value far away from the corresponding true parameter values. However, the center of the measurement error adjusted estimates in the case of high measurement errors is a little further away from the center of the estimates based on the true exposure values as compared to the case of low measurement errors. For the contrasting situations represented by simulation set-up 2 and 3 the use of misspecified exposure model seems to be inadequate in correcting the bias caused by measurement errors (Figures 2.3(a), 2.3(b) and 2.4(a), 2.4(b)). From the empirical kernel density plots it is also clear that bias correction performance of the misspecified normal exposure model is even worse in set-up 3, which represents a scenario of having a true exposure distribution with thicker tails but low correlation coefficient between the true covariates.

Finally, from the contrasting situations 4 and 5 similar trend of bias correction performance is observed as is observed in set-up 2 and 3. However, in set-up 5 (representing a scenario with a well-behaved log-normal exposure distribution but very high correlation between true covariates) the performance of misspecified normal exposure model is far better as compared to its contrasting scenario produced by the simulation set-up 4 (Figures 2.5(a), 2.5(b) and 2.6(a), 2.6(b)).

In conclusion, simulation results reveal that the influence of the misspecification of the unknown exposure distribution on regression parameter estimates is a cause of concern. Simulation results also reveal that both the presence of thick tails in the true exposure distributions and the presence of correlation between the true covariate measurements aggravate the biases in the regression parameter estimates and consequently the misspecified exposure model performs poorly in correcting these biases. In the former scenario the misspecified exposure model is found to perform worse.

2.4 Discussion

In this chapter we review some aspects of the bias correction in regression parameter estimates in logistic outcome model when some covariates are measured with errors. Particularly, we
investigate how well the postulated normal distribution for the unobserved continuous exposure variable performs in correcting the bias when the true exposure distribution is, in fact, not normal. Initially, we think that insensitiveness of the regression parameter estimates to the misspecification of the exposure model in the case of linear regression will also be retained for the logistic regression. But our simulation studies rule out our initial conjecture. We observe that the postulated normal distribution for the unobserved exposure can not correct the bias adequately in some cases. The bias correction performance of the structural approaches using a misspecified normal exposure model deteriorates when the true exposure distribution has wild observations (or thicker tails) and when the covariates in the model are correlated. Simulation studies also reveal that the presence of skewness and/or thicker tailedness in the true exposure distribution is a more serious problem than the presence of correlation among the exposures with respect to bias correction performance of the misspecified normal exposure model.
Chapter 3

Flexible Exposure Models to Adjust the Effect of Measurement Errors in Logistic Regression

3.1 Introduction

Though inference about the regression parameters in the linear regression model is not sensitive to the choice of exposure distribution in the presence of measurement errors in the explanatory variables (or exposures), the phenomenon no longer holds for the logistic regression model. Simulation studies in the previous chapter and in Richardson and Leblond (1997) reveal this fact. It can be noted that while adjusting for the effect of measurement errors on the estimates of regression parameters of the outcome models, it is typical to assume normality of the true but unobserved continuous exposure. But if the distribution of the true exposure is not normal the parametric methods based on misspecified normal exposure model fail to reduce the bias in the estimates of the regression parameters of interest adequately in the case of logistic regression. The performance is worse for heavy-tailed exposure distributions. In all, inferences are sensitive to the assumption about the distribution of true but unobserved exposure. So, it is worthwhile to search for flexible modelling approaches for the unobserved exposure that can reduce the sensitivity to assumption about the true exposure distribution while retaining the efficiency of the parametric inference. In structural approach paradigm, Richardson, Jaussent and Green (2002) attempt flexible parametric models based on mixture of normal distributions for the unobserved exposure. Huang, Stefanski and Davidian (2006) use second-order semi-nonparametric density as a flexible parametric exposure model. The second-order semi-nonparametric density is a special case of semi-nonparametric (SNP) density introduced by Gallant and Nychka (1987). The SNP density can capture a broad range of non-normal behav-
ior controlled by user–chosen tuning parameter. This density includes normal distribution as a special case. For elaborate discussion we refer readers to Davidian and Giltinan (1995), and Zhang and Davidian (2001). Huang, Stefanski and Davidian (2006) argue that if the assumed distribution for the unobserved exposure is sufficiently flexible that the moments, on which the estimation of the outcome model parameters depends, of the true exposure model are estimated consistently then the full robustness of the estimated outcome model parameters is possible. Once such robustness can be established, parametric structural methods become much more appealing in many ways than the functional approaches.

Though semiparametric and non-parametric approaches have the virtue of being more robust to the modeling assumption about the true but unobserved exposure they can lead to considerable loss of efficiency, in general. Furthermore, as it is already mentioned earlier in Chapter 1, in the case of measurement error problems in covariates it is not obvious how to incorporate complex epidemiological designs in the available non–parametric or semi–parametric approaches. Hence, in this thesis we take the parametric approach with flexible parametric models for the true but unobserved exposure. To this end one natural choice found in literature is the mixture of normal distributions with known or unknown number of components. With unknown number of components the parameter estimation becomes very challenging as it requires reversible–jump MCMC algorithm to implement. Besides this, simulation studies in Richardson et al. (2002) reveal that mixture of normal distribution does not perform well when the true exposure distribution is skewed. For substantial measurement errors, use of normal mixture virtually provides no improvement over the normal exposure model when the true exposure distribution is skewed and/or heavy–tailed. Use of SNP model may be another alternative in the context of structural parametric approaches. However, Ma, Genton and Davidian (2004) mention that the use of SNP distribution as a flexible parametric model for unobserved quantities is known to produce artifactual modes when the true distribution is skewed. Huang, Stefanski and Davidian (2006) investigate the performance of SNP distribution as a flexible exposure model in the case of a single covariate logistic outcome model. In their simulation studies they generate true the $X$ from a normal distribution and a mixture of normal distributions. Therefore, more investigation is needed to see how well the SNP distribution perform in correcting the measurement error bias in the case of very skewed exposure distribution, and in the cases where the true exposure distributions have thicker tails. In this thesis we take an alternative flexible modelling approach for the unobserved exposure. We consider using the skew-normal distribution and its more flexible variants as flexible models for the unobserved exposure.
3.2 Skew-normal Distribution and Its Variants

Azzalini (1985) introduces the skew–normal distribution. It represents a class of distributions which can reflect varying degrees of skewness, and includes the normal distribution as special case. It governs the skewness by an additional shape parameter. It is a natural choice in all practical situations where there is some skewness present. Mathematically, skew-normal (SN) class is nicely tractable and have a good number of properties in common with the standard normal distribution. The skew-normal distribution is defined as follows.

**Definition**: A random variable $X$ is said to have a skew–normal distribution with parameters $\mu, \sigma, \omega$, i.e. $X \sim SN(\mu, \sigma, \omega)$, if its density is of the form

$$f(x; \mu, \sigma, \omega) = \frac{2}{\sigma} \phi \left( \frac{x - \mu}{\sigma} \right) \Phi \left[ \omega \left( \frac{x - \mu}{\sigma} \right) \right]$$

$$-\infty < x, \mu, \omega < \infty, 0 < \sigma < \infty,$$  

where $\phi(.)$ and $\Phi(.)$ denote the standard normal density and distribution functions, respectively.

In (3.1), $\mu$ and $\sigma$ are the location and scale parameters, respectively, and $\omega$ is the shape parameter, which governs the skewness of the distribution. The distribution in (3.1) reduces to normal distribution when $\omega = 0$, while $\omega < 0$ and $\omega > 0$ correspond to the negatively skewed and the positively skewed densities, respectively. Again, as $\omega \to \infty$, SN density tends to $\frac{2}{\sigma} \phi \left( \frac{x - \mu}{\sigma} \right) I_{(\mu, \infty)}(x)$, which is the folded half-normal pdf. Similarly, as $\omega \to -\infty$, SN density tends to negative folded half-normal density. For the purpose of inference, methods of moments (MM) or maximum likelihood (ML) method can be used. However, both of these classical methods pose challenging problems in inferential procedure for SN class of distribution and its generalized variants (Lesio, 2004; Pewsey, 2000). In this dissertation, the generalized variants of SN distribution are introduced in Chapter 4.

Furthermore, the parametric approaches to adjust for the covariate measurement errors give rise to hierarchical models (a hierarchy of outcome model, measurement error model and exposure model), which further complicates the inferential process through ML method in presence of measurement errors in non-linear regression model. The difficulty arises because the true exposure is missing, and hence defining a likelihood function of the observed data requires integrating out the unobserved exposure. In measurement error problems with SN or its more generalized variants as exposure model such integrations do not have closed form. So, numerical integration methods are required. But numerical integration methods are usually far from
accurate in presence of measurement errors because we need to define grids for something which is not observed. Apart from computational difficulties with ML approach there are also technical difficulties including problem of convergence of the maximization algorithm used, and the possibility of negative estimate of the variance component of the true exposure distribution (Bolfarine et al., 2006). To elaborate on this point let $X$ and $Z$ be the two covariates on which the binary outcome $Y$ is regressed. Further, suppose that instead of the true covariate $X$ a noisy version, $X^*$, of $X$ is observed, and the other covariate $Z$ is measured precisely. Under the conditional independence assumption described in Section 1.5 of Chapter 1, the measurement model is given by $X^* \mid X \sim N(X, \tau^2)$ and the SN exposure model is given as $X \mid Z \sim SN(\alpha_0 + \alpha_1 Z, \lambda^2, \omega)$. Then the marginal distribution of $X^*$ given $Z$ is also a skew–normal distribution (Bolfarine et al., 2006) given by $X^* \mid Z \sim SN(\alpha_0 + \alpha_1 Z, \sigma^2_{x*}, \omega_{x*})$, where $\sigma^2_{x*} = \lambda^2 + \tau^2$ and $\omega_{x*} = \frac{\omega \lambda}{\sqrt{\lambda^2 + \tau^2(1 + \omega^2)}}$. The maximum likelihood estimate of the scale parameter, $\lambda$, of the distribution of the true exposure $X$ is $\hat{\lambda}^2 = \hat{\sigma}^2_{x*} - \hat{\tau}^2$, which may not be necessarily positive. On the other hand, unlike the ML approach, the Bayesian approach provides estimates of all the model parameters whose values are within the plausible range of values of the corresponding true parameters. Moreover, the use of MCMC algorithms makes a Bayesian approach to SN class and its generalized variants quite appealing. Bayesian MCMC approach provides a relatively simple numerical tool that allows us to bypass the analytical difficulties inherent in the ML approach. The Bayesian method accomplishes complicated integration behind the scene by just drawing Monte–Carlo samples from the posterior distributions of all the unobservables. Given all the nice features of Bayesian approach including computational advantage, especially in the case of measurement error problems, we resort to Bayesian approach to inference while adjusting for the bias in the estimate of the regression parameters of the logistic regression model.

3.3 Performance of Skew–normal Exposure Model: Simulation Studies

Our endeavor is to search for a flexible exposure model which is relatively robust to departure from the true underlying exposure distribution while retaining the efficiency of parametric inference. For this purpose, we attempt to investigate the performance of SN distribution as an exposure model through simulation studies. We consider exactly the same simulation settings as are considered in Section 2.2 of Chapter 2. For each of the simulation designs, we generate 50 data sets and estimate the regression coefficients of the given logistic regression model by
using SN distribution as exposure model. We compare the results with those obtained by using misspecified normal exposure model as well as with the naive estimates and estimates based on true values of the exposure. In Section 3.4, we provide a discussion about the prior specifications as well as a mathematical sketch of deriving the conditional posterior distributions for the model parameters. Simulation results including discussion are presented in Section 3.5.

3.4 Measurement Error Models for Logistic Regression

We consider a two-covariate logistic regression model given as
\[
\logit(p_i) = \beta_0 + \beta_1 X + \beta_2 Z,
\]
where the continuous covariate \(X\) is not observed, rather a noisy version, \(X^*\), of \(X\) is observed. The other covariate \(Z\) is measured without error. To be able to carry out inference about the regression parameters we need the joint distribution of the triplet \((Y, X, X^*)\) conditional on \(Z\), i.e., we need \(f(y, x, x^* | z) = f(y | x, z)f(x^* | x, y, z)f(x | z)\). The logistic model defined in (3.2) provides the conditional distribution \(f(y | x, z)\), which is known as outcome model. To define the joint distribution of \((Y, X, X^* | Z)\) we also need two more conditionals, \(f(x^* | x, y, z)\) and \(f(x | z)\), which are known as measurement model and the exposure model, respectively. As in the earlier chapter, we assume normal measurement error model with known variance for simulation studies. Since our aim is to investigate the relative robustness of the SN distribution as exposure model we assume that unobserved exposure \(X\), conditional on \(Z\), follows SN distribution though we generate the true exposure \(X\), conditional on \(Z\), from the log-normal distribution. That is, for simulation studies we generate data from the following hierarchical measurement error model.

\[
\logit(p_i) = \beta_0 + \beta_1 X + \beta_2 Z
\]
\[
X^*_i | X_i \sim N(X_i, \tau^2)
\]
\[
X_i | Z_i \sim LN(\alpha_0 + \alpha_1 Z_i, \lambda^2).
\]

We take \(\tau^2\) to be known. In each case, two values of \(\tau^2\) are considered so as to make the ratios of \(\tau^2\) to \(E[Var(x | z)]\) 20% and 40% to reflect low and substantial amount of measurement errors in the data, respectively. Having the three sub–models defined in (3.3), (3.4) and (3.5), and the priors for the unknown parameters put together we get the joint posterior distribution.
of the model parameters as
\[
\pi(x, \alpha, \beta, \lambda^2, \omega \mid y, x^*, z) = \frac{f(y \mid x, z, \beta) f(x^* \mid x, y, z) f(x \mid z, \alpha, \lambda^2, \omega) \times}{\pi(\alpha, \beta, \lambda^2, \omega)},
\]
where \(\pi(\alpha, \beta, \lambda^2, \omega)\) is the joint prior distribution of the model parameters. As an aside, we take \(\alpha = (\alpha_0, \alpha_1)\) and \(\beta = (\beta_0, \beta_1, \beta_2)\). Without any prior knowledge about the dependence of the unknown parameters it is a common practice to assume prior independence among them so that \(\pi(\alpha, \beta, \lambda^2, \omega) = \pi(\alpha) \pi(\beta) \pi(\lambda^2) \pi(\omega)\). In regression problem, the common choice for the regression parameters are the locally uniform priors, i.e., \(\pi(\alpha) \propto 1\) and \(\pi(\beta) \propto 1\). For variance parameters (scale parameters), \(\lambda^2\), diffuse proper priors are usually used to ensure the propriety of the joint posterior distribution defined in (3.6). In this study, we choose unit information inverted Gamma prior for \(\lambda^2\), i.e., \(\pi(\lambda^2)\) is taken to be \(IG(1, 2)\). Details about \(IG(q, D)\) prior have been provided in Section 2.2.2 of Chapter 2.

For our estimation purpose, the log-normal exposure distribution is misspecified as SN distribution, i.e., we assume \(X \mid Z \sim SN(\alpha_0 + \alpha_1 Z, \omega)\), where \(\omega\) is the shape parameter. Now, for the full Bayesian specification we need a prior distribution for the shape parameter \(\omega\). In choosing the prior for \(\omega\) our guiding principles are to center the prior at zero and allow a large variance. A zero value for \(\omega\) is equivalent to getting back to normal exposure model. Allowing large variability for the prior for \(\omega\) is tantamount to being close to noninformative. In such case, the posterior inference of \(\omega\) will largely depend on the information contained in the data about it. According to our guiding principle we choose a normal prior with a mean zero and a variance equal to 100. So, for our estimation and inference purposes the full Bayesian measurement error model for the logistic regression takes the form
\[
\text{logit}(p_i) = \beta_0 + \beta_1 X + \beta_2 Z
\]
\[
X^*_i \mid X_i \sim N(X, \tau^2)
\]
\[
X_i \mid Z_i \sim SN(\alpha_0 + \alpha_1 Z_i, \lambda^2, \omega)
\]
\[
\theta \sim \pi(\theta),
\]
where \(\theta = (\beta_0, \beta_1, \beta_2, \alpha_0, \alpha_1, \lambda^2, \omega)\) and \(\pi(\theta)\) is the prior for \(\theta\). Under the assumption of prior independence among the components of \(\theta\) we can write \(\pi(\theta) = \pi(\beta) \pi(\alpha) \pi(\lambda^2) \pi(\omega)\).
3.4.1 Joint and Conditional Posterior Distributions of the Model Parameters

In Bayesian approach, inferences are based on the posterior distributions, especially on the conditional posterior distributions of the respective model parameters. Under model (3.7), the joint posterior distribution of the model parameters, given the data, is obtained as

\[
\pi(x, \theta | y, x^*, x) = \left[ \prod_{i=1}^{n} f(y_i | x_i, z_i, \beta) f(x_i^* | x_i) f(x_i | z_i, \alpha, \lambda^2, \omega) \right] \pi(\alpha, \beta, \lambda^2, \omega)
\]

\[
\times \left[ \prod_{i=1}^{n} \frac{\exp \left[ y_i (\beta_0 + \beta_1 x_i + \beta_2 z_i) \right]}{1 + \exp (\beta_0 + \beta_1 x_i + \beta_2 z_i)} \right] \left[ \prod_{i=1}^{n} e^{-\frac{1}{2} \lambda^2 (x_i^* - x_i)^2} \right] \left[ \prod_{i=1}^{n} \left( \frac{1}{\lambda^2} \right)^{\frac{3}{2}} e^{-\frac{1}{2\lambda^2} (x_i - a_0 - a_1 z_i)^2} \right] \Phi \left[ \frac{\omega (x_i - a_0 - a_1 z_i)}{\lambda} \right] \left( \frac{1}{\lambda^2} \right)^{\frac{3}{2}} e^{-\frac{1}{2\lambda^2} w_0^2} \left( \frac{1}{\lambda^2} \right)^{\frac{3}{2}} e^{-\frac{1}{2\lambda^2} \omega^2}. \tag{3.8}
\]

From (3.8), the conditional posterior distributions of each of the model parameters, given the data, can be derived. The conditional posterior distributions are

(i) for \( \beta, \pi(\beta | y, x, x^*, z, \theta - \beta) \propto \left[ \prod_{i=1}^{n} \frac{\exp [y_i (\beta_0 + \beta_1 x_i + \beta_2 z_i)]}{1 + \exp (\beta_0 + \beta_1 x_i + \beta_2 z_i)} \right] \exp \left[ \frac{\lambda^2}{2} \left( x - \frac{\alpha^2 + w_0^2}{\lambda^2} \right)^2 \right] \Phi \left[ \frac{\omega (x - w_0)}{\lambda} \right], \]

where \( \theta - \beta \) is the vector of all model parameters other than \( \beta \).

(ii) for \( X, \pi(x | y, x, x^*, z, \theta) \propto \left[ \prod_{i=1}^{n} \frac{\exp [y_i (x_0 + x_1 x_i + x_2 z_i)]}{1 + \exp (x_0 + x_1 x_i + x_2 z_i)} \right] \exp \left[ \frac{\lambda^2}{2} \left( x - \frac{\alpha^2 + w_0^2}{\lambda^2} \right)^2 \right] \Phi \left[ \frac{\omega (x - x_0)}{\lambda} \right], \]

where \( w = (1, z), \alpha = (a_0, a_1) \) and \( 1 \) is a column vector of 1s of length \( n \), the sample size.

(iii) for \( \alpha, \pi(\alpha | y, x, x^*, z, \theta - \alpha) \propto \exp \left( -\frac{1}{2\lambda^2} \left[ x - \frac{\alpha^2 + w_0^2}{\lambda^2} \right] \right) \left[ \prod_{i=1}^{n} \Phi \left[ \frac{\omega (x - \omega_0)}{\lambda} \right] \right]. \]

(iv) for \( \lambda^2, \pi(\lambda^2 | y, x, x^*, z, \theta - \lambda^2) \propto \left( \frac{1}{\lambda^2} \right)^{\frac{n+1}{2}} e^{-\frac{0.5 + 0.5 (x_0 - w_0)^2}{\lambda^2}} \left[ \prod_{i=1}^{n} \Phi \left( \frac{\omega (x - \omega_0)}{\lambda} \right) \right]. \]

(v) for \( \omega, \pi(\omega | y, x, x^*, z, \theta - \omega) \propto e^{-\frac{1}{2\lambda^2} \omega^2} \left[ \prod_{i=1}^{n} \Phi \left( \frac{\omega (x_i - \omega_0)}{\lambda} \right) \right]. \]

Once we have the posterior conditionals of the model parameters, we can use suitable MCMC algorithms for posterior simulations from the respective posterior conditionals. For details of different types of MCMC algorithms researchers are referred to Gilks et al. (1998), and Robert and Casella (1998). The type of MCMC algorithm needed to implement the computation depends on the form of the conditional posteriors of the model parameters. For our current model,
the conditional posterior distribution of $\beta$ does not follow any known distribution and hence random-walk Metropolis–Hastings (RWMH) algorithm is used for drawing posterior simulations for $\beta$. The posterior conditionals for $X$, $\alpha$ and $\lambda^2$ are composed of two parts with one part (the first part) having the form of known densities, and the second part involving the standard normal cumulative distribution function or the product of standard normal cumulative distribution functions. For instance, for the vector of the values of the true exposure $X$, we have $\pi(x | y, x^*, z, \theta) \propto N \left( \frac{\lambda^2 x^* + \tau^2 w \alpha}{\tau^2 + \lambda^2}, \frac{\tau^2 \lambda^2}{\tau^2 + \lambda^2} \right) \left[ \Phi \left( \frac{w(x - w\alpha)}{\lambda} \right) \right]$, where the first part is the normal density. Similarly, the first part of the posterior conditional for $\alpha$ is also normal with mean $(w'w)^{-1}w'x$ and covariance matrix $\lambda^2(w'w)^{-1}$. Note that the first part of conditional posterior for $\lambda^2$ is inverted Gamma distribution with scale parameter $n+1$ and shape parameter $0.5 + 0.5(x - w\alpha)'(x - w\alpha)$. Again, the first parts of the posterior conditionals for $X$, $\alpha$ and $\lambda^2$ do not depend on the respective parameters. Hence, we can use independence sampler MH algorithm (with first part of the respective posterior conditionals being the candidate generating densities for the respective parameters) to carry out posterior computations. For independence sampler MH algorithm, the proposal density for $X$ will be $q_x = N \left( \frac{\lambda^2 x^* + \tau^2 w \alpha}{\tau^2 + \lambda^2}, \frac{\tau^2 \lambda^2}{\tau^2 + \lambda^2} \right)$, which is independent of $X$, the condition needs to be satisfied to be able to use independent sampler MH algorithm. The candidate generating densities for $\alpha$ and $\lambda^2$ would be $q_{\alpha} = N \left( (w'w)^{-1}w x, (w'w)^{-1}\lambda^2 \right]$ and $q_{\lambda^2} = \text{inv-gamma} \left[ \frac{n+1}{2}, 0.5 + 0.5(x - w\alpha)'(x - w\alpha) \right]$, respectively. Unlike RWMH algorithm, the independence sampler MH algorithm does not require any tuning during the implementation. Implementation of the independence sampler MH algorithm is as simple as that of the accept–reject algorithm. Actually, this algorithm is a generalization of the accept–reject algorithm (Robert and Casella, 1998). So, from the perspective of implementation independent sampler MH algorithm is much easier than the RWMH algorithm. However, with independence sampler MH algorithm we have to be careful about the acceptance rate and the convergence of the MCMC sequence. To be comparable with the RWMH algorithm the acceptance rate for the independence sampler MH algorithm should not be less than 20%, the widely suggested lower limit of acceptance rate for the RWMH algorithm. Finally, for posterior computation of $\omega$, RWMH algorithm is used.

### 3.4.2 Simulation Set-up

In order to investigate the performance of the misspecified SN exposure model we conduct simulation studies under several different predetermined simulation settings. The simulation set-up considered here are exactly same as those are considered in Section 2.2 of Chapter 2. We consider five different simulation designs to generate true exposure from log–normal distribution
of different degrees of skewness. In each of the simulation designs we simulate 50 data sets of size $n = 500$. Five different set-ups are considered in order to investigate how the SN model for exposure distribution works in adjusting the bias in the estimates of the regression parameters of the logistic outcome model for different degrees of skewness of the unobserved true exposure distribution and different degrees of correlation between the true exposure and the precisely measured covariate.

3.4.3 Monitoring the Convergence of MCMC Algorithm for the Model Parameters

There are two fundamental issues with inferences using Bayes-MCMC methods: (i) the slow movement of the chain toward the target distribution exhibited by low acceptance rate, and (ii) slow mixing of the MCMC chain. Both problems result in slow convergence or lack of convergence of the MCMC chain. So, in practice it is always desirable to have a chain which mixes and moves well at the same time to ensure adequate coverage of the entire area of target posterior distribution.

The general and widely used approach to monitor the convergence of MCMC chain is to plot the MCMC chain against the time sequence. Gelman (1996) suggests examining multiple chains generated with widely dispersed starting points in the target distribution. Gelman (1996) argues that starting points that are far apart can make lack of convergence apparent and ensure that all major regions of the target distribution are represented in the simulations. The monitoring of convergence of MCMC chains for different parameters under different simulation designs is presented in the Sections 3.4.4 and 3.4.5.

3.4.4 Convergence of the Shape Parameter $\omega$ and the Scale Parameter $\lambda^2$

For SN distribution estimation of the shape parameter $\omega$ is problematic. In the case of ML method, estimation of the shape parameter of skew-normal distribution poses some problems, especially the divergence of the maximization algorithm. Pewsey (2000) discusses the problems of inference about the shape parameter of skew-normal distribution. The Bayesian approach does not suffer from the problems those arise in ML estimation method because of the calibration of the likelihood by the prior information. However, since the estimation of shape parameter is problematic it is important to investigate the convergence and mixing of the MCMC chain for
the shape parameter $\omega$.

For $\omega$ we need to use RWMH algorithm to generate MCMC chain from the posterior distribution of $\omega$. With RWMH algorithm we need to choose appropriate tuning parameter (jump size for the proposal distribution) to have an MCMC chain for the parameter of interest which mixes and moves well in the region of the posterior distribution. There is no hard and fast rule to determine the appropriate jump size. Usually, a trial and error method is adopted. A considerable volume of research work has been carried out in this regard and suggestions have been made to monitor the mixing and convergence of the MCMC chains. A nice reference in this regard is Gilks et al. (1996). Research findings suggest that for better mixing and convergence a desirable acceptance rate is around 20%-50% for RWMH algorithm. So, we need to adjust the jump size so as to get an MCMC chain which mixes well and has an acceptance rate of around 20%-50%.

For posterior simulation of the scale parameter $\lambda^2$ we use independence sampler MH algorithm because the first part of the conditional posterior distribution of $\lambda^2$ is an inverted Gamma distribution. The nice thing about the posterior simulation of $\lambda^2$ is that we can use an algorithm whose implementation is in no way more difficult than that of the Gibbs sampler.

In order to monitor the convergence of the MCMC chain for the shape parameters $\omega$ and the scale parameter $\lambda^2$ we plot two chains started with two different and widely spread starting values on the same plotting window for a randomly selected data set under each of the simulation settings we consider in this study. These convergence plots are presented in Figures 3.1-3.5.

In each plot, the upper panel represents the convergence plot for the MCMC chains of the shape parameter $\omega$ and the lower panel represents the same of the scale parameter $\lambda^2$. For both $\omega$ and $\lambda^2$ good mixing and convergence are demonstrated by the MCMC plots. For all the simulation set-up, the two independent chains starting with two different widely dispersed initial values become indistinguishable for both the parameters well before 5000 iterations, the burn-in length considered for the MCMC chains. Indistinguishable MCMC chains starting with widely spread initial values for a parameter lends more credence about the convergence of the MCMC chains to the true parameter value. Again, all the chains presented in different figures exhibit good mixing, which is an indication of the coverage of the entire area of the target posterior distribution. The acceptance rates for the MCMC simulations for $\omega$ are found to vary from 22% to 59% for the 50 data sets in each of the simulation designs considered in this study. The acceptance rates for $\lambda^2$ are found to be around 90% for all the data sets in each case of the simulation designs. Such high acceptance rates and good convergence of the MCMC chains for
Figure 3.1: Convergence plots for $\omega$ and $\lambda^2$ under simulation set-up 1

Figure 3.2: Convergence plots for $\omega$ and $\lambda^2$ under simulation set-up 2
Figure 3.3: Convergence plots for $\omega$ and $\lambda^2$ under simulation set-up 3

Figure 3.4: Convergence plots for $\omega$ and $\lambda^2$ under simulation set-up 4
Figure 3.5: Convergence plots for $\omega$ and $\lambda^2$ under simulation set-up 5

$\lambda^2$ indicate that independence sampler MH algorithm works nicely for $\lambda^2$ for all the simulated data sets.

3.4.5 Convergence of the Regression Parameters $\beta$

Estimating the regression parameters of the logistic outcome model is the primary goal of inference, because these parameters represent the strength of association between the outcome and the explanatory variables, including the primary exposure of interest. And, all our efforts is to eliminate or reduce the bias introduced by the measurement errors in the estimates of these regression parameters. So, monitoring convergence of the MCMC chains for $\beta$ is even more important. For monitoring the convergence of the MCMC chains for the component of the parameter vector $\beta$ similar technique used in the previous subsection is employed. The convergence plots for the components of $\beta$ are presented in Figures 3.6–3.10.

Again, good mixing and convergence of MCMC chains for the components of $\beta$ are demonstrated by the convergence plots in each case of the simulation designs considered. In each case, two MCMC chains become indistinguishable after few iterations for all the components of $\beta$. The acceptance rates are found to be around 30% for all the data sets under each of the simulation
Two independent MCMC chains for $\beta_0$:
(a): For substantial measurement error
(b): For small measurement error

Two independent MCMC chains for $\beta_1$:
(a): For substantial measurement error
(b): For small measurement error

Two independent MCMC chains for $\beta_2$:
(a): For substantial measurement error
(b): For small measurement error

Figure 3.6: Convergence plots of $\beta$ under simulation set-up 1
Figure 3.7: Convergence plots of $\beta$ under simulation set-up 2

Two independent MCMC chains for $\beta_0$ (a): For substantial measurement error

Two independent MCMC chains for $\beta_2$ (b): For small measurement error

Two independent MCMC chains for $\beta_1$ (a): For substantial measurement error

Two independent MCMC chains for $\beta_1$ (b): For small measurement error

Two independent MCMC chains for $\beta_2$ (a): For substantial measurement error

Two independent MCMC chains for $\beta_2$ (b): For small measurement error
Figure 3.8: Convergence plots of $\beta$ under simulation set-up 3
Figure 3.9: Convergence plots of $\beta$ under simulation set-up 4
Figure 3.10: Convergence plots of $\beta$ under simulation set-up 5
settings considered.

3.5 Results and Discussion

For each data set, to have a posterior estimate of each of the regression parameters a posterior sample of size 7500 is generated through an appropriate MCMC algorithm. The convergence plots confirm that in each of the cases considered MCMC sampler converges well before 5000 iterations. However, to make sure that only the posterior draws after convergence are used for inferential purposes we discard first 5000 MCMC draws as burn–in out of 7500 draws. The remaining 2500 draws are used to estimate the model parameters. For the parameter estimate posterior mean is used. To compare the bias correction performance of the misspecified normal model and the misspecified skew–normal model we construct the kernel density plots of the parameter estimates obtained for the 50 generated data sets under each of the simulation designs considered. The kernel density plots of the of estimates of the regression parameters $\beta$ are presented in Figures 3.11- 3.15.

In each figure panel (a) corresponds to the case of high measurement errors and panel (b) corresponds to the case of low measurement errors. For instance, Figures in panel 3.11 (a) and 3.11 (b) represent the empirical kernel density estimates of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ for the cases of low and substantial measurement errors, respectively, for simulation design 1.

Under all the simulation designs, the naive estimates of $\beta_0$ and $\beta_1$ are seriously biased. In both cases, the empirical kernel densities of the naive estimates are found to be shifted away from the corresponding kernel densities of the estimates based on the true parameter values. As an aside, it is seen from the kernel density plots that the estimates of $\beta_2$, the coefficient associated with the precisely measured covariate, becomes biased only if the precisely measured covariate is correlated with the true but unobserved exposure. The stronger the correlation between the two the larger the degree of bias in the estimates of $\beta_2$.

In simulation design 1, a case with a true exposure distribution which is moderately skewed, both the misspecified normal exposure model and misspecified SN exposure model perform similarly in correcting the bias in the estimates of the regression parameters. Figure 3.11 shows that the Kernel density curves for the estimates of the regression parameters almost coincide and both are pretty close to the corresponding kernel density curve of the estimates based on the true exposure values in both high and low measurement error situations. This indicates
Figure 3.11: Kernel densities of the estimates of regression parameters under simulation design 1. Solid curve represents the kernel density of the naive estimates, dashed curve represents the kernel density of the estimates based on true exposure values, dotted curve gives the kernel density of the estimates based on normal exposure model, and the bold solid curve represents the kernel density of the estimates based on SN exposure model.
Figure 3.12: Kernel densities of the estimates of regression parameters under simulation design 2. Solid curve represents the kernel density of the naive estimates, dashed curve represents the kernel density of the estimates based on true exposure values, dotted curve gives the kernel density of the estimates based on normal exposure model, and the bold solid curve represents the kernel density of the estimates based on SN exposure model.
Figure 3.13: Kernel densities of the estimates of regression parameters under simulation design 3. Solid curve represents the kernel density of the naive estimates, dashed curve represents the kernel density of the estimates based on true exposure values, dotted curve gives the kernel density of the estimates based on normal exposure model, and the bold solid curve represents the kernel density of the estimates based on SN exposure model.
Figure 3.14: Kernel densities of the estimates of regression parameters under simulation design 4. Solid curve represents the kernel density of the naive estimates, dashed curve represents the kernel density of the estimates based on true exposure values, dotted curve gives the kernel density of the estimates based on normal exposure model, and the bold solid curve represents the kernel density of the estimates based on SN exposure model.
Figure 3.15: Kernel densities of the estimates of regression parameters under simulation design 5. Solid curve represents the kernel density of the naive estimates, dashed curve represents the kernel density of the estimates based on true exposure values, dotted curve gives the kernel density of the estimates based on normal exposure model, and the bold solid curve represents the kernel density of the estimates based on SN exposure model.
that the Bayesian adjustment based on misspecified normal or misspecified SN exposure model seems to be adequate in both cases of low and substantial measurement errors. For simulation design 2, a scenario with a true exposure distribution similar to that in scenario 1 in terms of the degree of skewness but with a strong correlation (0.76) between true exposure and the other covariate which is precisely measured, similar kind of performance of both the misspecified exposure models are observed. So, from the analysis of data under simulation designs 1 and 2 it is observed that the bias correction performance of the misspecified exposure models seems not to be affected by the degree of correlation between the true but unobserved exposure and the precisely measured covariate.

Simulation study results under design 3 reveal that though both the misspecified exposure models can correct some bias, none of them performs satisfactorily. However, the SN exposure model does a much better job than the normal exposure model. This is the case of the simulation designs where the true exposure distribution is heavy-tailed with moderate correlation between the unobserved and the precisely measured covariate.

Finally, the contrasting situations under simulation designs 4 and 5 reveal that in the former case of heavy-tailed exposure distribution the performance of neither exposure model is satisfactory in correcting the bias, though SN exposure model is observed to perform better than the normal exposure model. However, in the later case of a less skewed true exposure distribution and very high correlation between the unobserved and the precisely measured covariates both the exposure models perform equally well, and the performance seems to be adequate.

In brief, simulation results discover that for well behaved (less skewed) true exposure distribution both the normal and SN exposure models perform satisfactorily in correcting the measurement error bias in the estimates of the regression parameters. But, in the case of heavy-tailed true exposure distributions the performance of both the exposure models are found to be inadequate though SN exposure model is found to perform much better than the normal exposure model in those cases. Simulation studies also reveal that both the high degree of skewness of the true exposure distribution and the presence of high correlation between the unobserved and precisely measured covariates introduce more bias in the regression parameter estimates in presence of measurement errors. However, only the degree of skewness of the true exposure distribution, not the correlation between the unobserved and the precisely measured covariates, seems to hinder the bias correction performances of the misspecified exposure models.
Chapter 4

More Flexible Exposure Models: Flexible Extensions of Skew–normal Distribution

4.1 Introduction

Simulation studies in the previous chapter reveal that though the SN distribution provides more flexibility in modeling the unobserved exposure, still its performance is not satisfactory in the cases where the true exposure distribution has heavy tail. This may be due to the fact that the degree of skewness that can be modelled by the SN distribution is limited. This fact is supported by the boundedness of the coefficient of skewness of the SN distribution. By SN distribution limited degree of skewness can be captured (Azzalini, 1985). Furthermore, SN distribution can not accommodate more than one mode. But in reality, the distribution of the unobserved exposure may have more than one mode as well as higher degree of skewness and heavy–tailedness than can be accommodated by SN distribution. So, a more flexible exposure model, which can incorporate skewness, heavy–tailedness as well as multimodality, is needed. One obvious choice to this end is the mixture of normal distribution as considered by Carrol, Roeder and Wasserman (1999), Muller and Roeder (1997), and Richardson, Leblond, Jaussent and Green (2002). But using the normal mixture becomes computationally cumbersome. More importantly, as demonstrated by the simulation studies in Richardson, Leblond, Jaussent and Green (2002), mixture of normal can not perform satisfactorily in the cases where the distribution of the true but unobserved exposure is skewed and has heavier tail. An alternative choice that can handle both multimodality skewness and heavy–tailedness, and can offer computational advantage is to represent the distribution of the unobserved exposure by flexible generalized skew–normal (FGSN) distribution. This distribution is a member of the family called the flexible general-
ized skew–elliptical (FGSE) class of distributions introduced by Genton and Loperfido (2002). Apart from modelling skewness, FGSN distribution can also accommodate multimodality, and heavy–tailedness, and hence offering higher degree of flexibility in capturing the feature of the distribution of the unobserved exposure. Elaboration of the FGSN distribution including applications is provided in Ma and Genton (2004). Ma, Genton and Davidian (2004) examine the suitability of the multivariate version of the FGSN distribution to model the random effects distribution in linear mixed effects model. They mention that FGSN distributions are able to capture multimodality and are well suited to situations where the density is suspected to be highly skewed. Considering all the reasonable flexibility of the FGSN distribution in capturing the unknown exposure distribution, it is worth studying the applicability of it in comparison to SN distribution in adjusting the measurement error bias in the estimates of the regression parameters of logistic outcome model. Section 4.2 provides a brief sketch of the characterization of FGSN distribution.

4.2 Flexible Generalized Skew–normal Distribution

A random variable $X$ is said to have a univariate FGSN distribution if its density can be written as

$$f(x) = \frac{2}{\lambda} \phi \left( \frac{x - \mu}{\lambda} \right) \Phi \left[ \sum_{i=1}^{K} \omega_i \left( \frac{x - \mu}{\lambda} \right)^{2i-1} \right],$$

(4.1)

where $i = 1, 2, \ldots, \frac{K+1}{2}$, $K \in \mathbb{N}$ is an odd number, and $\omega_i \in \mathbb{R}$.

In (4.1), $\mu$ is the location parameter, $\lambda$ is the scale parameter and $\omega_i$’s are the shape parameters. The number of terms in $\Phi [\cdot]$ of (4.1) determines the number of modes of the distribution, and the values of $\omega_i$’s determine the degree of skewness. It can be shown that univariate FGSN distribution with order $K = 3$ may have at most two modes, and is unimodal for $K = 1$ (Ma and Genton, 2004). It is clear from (4.1) that

(i) if $\omega_i = 0, \forall i$ then (4.1) reduces to the pdf of a normal distribution

(ii) if $\omega_i = 0$ for $i = 2, 3, \ldots, K$, but $\omega_1 \neq 0$ then (4.1) reduces to the pdf of a SN distribution.

As regards to using FGSN distribution to model the unobserved exposure, it is necessary to decide how many terms (what value of $K$) to use in $\Phi [\cdot]$ of (4.1). Increasing the value of $K$
will provide more flexibility. However, there is always a price to pay in terms of efficiency of estimates for allowing extra flexibility. Hence, the value of $K$ should be so chosen that the FGSN distribution can give just adequate flexibility to capture the major features of the unobserved exposure distribution. According to Ma and Genton (2004), $K = 3$ can provide enough flexibility to approximate a wide variety of densities. So, throughout this study we investigate the suitability of an FGSN distribution with $K = 3$ as an exposure model in correcting the measurement error bias in the estimate of the regression parameters of the logistic outcome model.

### 4.2.1 Joint and Conditional Posteriors and the MCMC Algorithm under FGSN Exposure Model

The FGSN distribution is very similar to the SN distribution except the skewing function $\Phi(\cdot)$. For SN model the skewing function is given by $\Phi(\frac{\xi - \mu}{\sigma})$, whereas for FGSN model it is given as $\Phi\left(\sum_{t=1}^{K} \omega_t \left(\frac{\xi - \mu}{\sigma}\right)^{2t-1}\right)$. Since the FGSN exposure model has the same functional form as the SN exposure model, the joint and conditional posteriors under FGSN model have the similar functional forms as those obtained under SN exposure model in Section 3.4.1 of Chapter 3. The only exception is the argument of the skewing function $\Phi(\cdot)$. Note that the prior specifications for the model parameters under FGSN exposure model are same as those used under SN exposure model. That is, we consider flat priors for $\alpha$ and $\beta$, unit information inverted Gamma prior for the scale parameter $\lambda^2$, and vague normal prior with mean zero and variance $10^2$ for the shape parameters $\omega_1$ and $\omega_2$.

Regarding the posterior simulations of the model parameters, RWMH algorithm is used for all the unknown parameters except the unobserved covariate $X$. For $X$, independence sampler MH algorithm is used. Despite similar functional forms of the posteriors for $\lambda^2$ and $\alpha$, independence sampler MH algorithm is found to be inefficient under FGSN exposure model though the algorithm works well under SN exposure model. Hence, from the perspective of MCMC implementation using FGSN exposure model is a bit more demanding than using SN exposure model as we need to tune up the MCMC jumps for all the model parameters under FGSN exposure model. Furthermore, convergence monitoring for the shape parameters $\omega_1$ and $\omega_2$ demonstrates that under FGSN exposure model longer chains are required as compared to those under SN exposure model. So, for the purpose of posterior inference under FGSN exposure model we generate MCMC chains of length 12,500 iterations (as opposed to 7500 under SN exposure model) with a burn-in of length 7500 iterations (as opposed to 5000 under SN
exposure model). That is, 5000 iterations after the burn-in run are used for inference. Figures 4.1-4.5 represent the kernel densities of the parameter estimates under different simulation designs that are so far considered in this study. Panel (a) in each figure represents the case of high measurement errors, whereas panel (b) represents the case of low measurement errors.

Simulation results confirm our speculation about the comparative bias correction performances of SN and FGSN exposure models. As demonstrated in Figures 4.1, 4.2 and 4.5 the performance of both the misspecified exposure models (SN and FGSN models) seems to be very similar and adequate. These three Figures represent the cases where the distribution of the true but unobserved exposure is well-behaved (not heavy-tailed). On the other hand, Figures 4.3 and 4.4 demonstrate that for heavy-tailed distribution of the unobserved exposure FGSN model exhibits better bias correction performance than SN exposure model. In both cases, the empirical kernel densities of $\beta_0$ and $\beta_1$ under FGSN exposure model are seen to be shifted closer to the kernel densities for the estimates of the corresponding parameters obtained from the true exposure values.

Regarding the efficiency of the estimates under more flexible FGSN exposure model, the kernel density plots under different simulation designs do not show apparent loss of efficiency. Under all the simulation designs considered here the spread of the kernel densities of the estimates under SN and FGSN exposure models are found to be very similar. So, allowing extra flexibility by switching to FGSN exposure model from SN exposure model for the unobserved exposure exhibits better bias correction performance without apparent loss of efficiency. This is particularly true when the distribution of the true but unobserved exposure is heavy-tailed. Even in the cases of well-behaved exposure distribution the use of more flexible FGSN model does not exhibit any worse performance than SN exposure model in terms of the efficiency of the estimated regression parameters. Ma, Genton and Davidain (2004) observe similar findings while modelling random effects distribution by FGSN distribution. Their simulations studies demonstrate that allowing more flexibility while modelling the random effects distribution by FGSN distribution does not impose skewness or multimodality when these features are not present.

### 4.3 Flexible Generalized Skew–t Distribution

Flexible Generalized Skew–t (FGST) distribution is more flexible than flexible generalized skew–normal (FGSN) distribution in the line of flexible parametric exposure models we can consider to
Figure 4.1: Kernel densities of the estimates of regression parameters under simulation design 1. Solid curve represents the kernel density of the estimates based on true exposure values, dashed curve represents the kernel density of the estimates based on SN exposure model, and dotted curve gives the kernel density of the estimates based on FGSN exposure model.
Figure 4.2: Kernel densities of the estimates of regression parameters under simulation design 2. Solid curve represents the kernel density of the estimates based on true exposure values, dashed curve represents the kernel density of the estimates based on SN exposure model, and dotted curve gives the kernel density of the estimates based on FGSN exposure model.
Figure 4.3: Kernel densities of the estimates of regression parameters under simulation design 3. Solid curve represents the kernel density of the estimates based on true exposure values, dashed curve represents the kernel density of the estimates based on SN exposure model, and dotted curve gives the kernel density of the estimates based on FGSN exposure model.
Figure 4.4: Kernel densities of the estimates of regression parameters under simulation design 4. Solid curve represents the kernel density of the estimates based on true exposure values, dashed curve represents the kernel density of the estimates based on SN exposure model, and dotted curve gives the kernel density of the estimates based on FGSN exposure model.
Figure 4.5: Kernel densities of the estimates of regression parameters under simulation design 5. Solid curve represents the kernel density of the estimates based on true exposure values, dashed curve represents the kernel density of the estimates based on SN exposure model, and dotted curve gives the kernel density of the estimates based on FGSN exposure model.
adjust the bias induced by the covariate measurement errors in the estimates of the parameters of the logistic outcome model. It belongs to the same family of distributions as the FGSN distribution does. That is, it is also a member of the generalized skew-elliptical (GSE) class of distributions introduced in Genton and Loperfido (2002). This distribution can provide more flexibility than FGSN distribution in modeling heavy-tailedness and leptokurtic behavior of the underlying distribution of the true but unobserved exposure. It is already demonstrated by the simulation results in the previous section that the FGSN exposure model cannot adjust the measurement error bias completely. This is especially the fact when the true exposure distribution has heavier tail and/or high peakedness. This happens probably due to the fact that the FGSN distribution fails to capture the extreme heavy-tailedness and/or leptokurtic behavior of the true exposure distribution. So, further flexibility in terms of heavy-tailedness and/or extra peakedness of the assumed exposure distribution is sought. In this line, the most obvious choice is the FGST distribution as FGST distribution can provide extra flexibility to capture the heavy-tailedness and leptokurtic behavior of the underlying true exposure distribution. Furthermore, since FGST distribution generalizes the FGSN distribution, it retains all the other capabilities, like modeling multimodality and skewness, of the FGSN distribution. For the elaboration about the GSE class of distributions and its different candidates, including the FGST distribution, readers are referred to Ma and Genton (2004), and Genton (2004). Furthermore, Ma, Genton and Davidian (2004) demonstrate the applicability of FGSN and FGST distributions to model the random effects distribution in linear mixed effects models.

4.3.1 Definition and the Basic Characteristics of FGST Distribution

A random variable $X$ is said to have an FGST distribution if its density function can be written as

$$f(x) = 2t(x; \mu, \lambda^2, v) \Phi \left[ \sum_{i=1}^{K} \omega_i \left( \frac{x - \mu}{\lambda} \right)^{2i-1} \right], \quad (4.2)$$

where $i = 1, 2, 3, \ldots, \frac{K+1}{2}$, $K$ is an odd integer, and $\omega_i \in \mathbb{R}$. In (4.2), $t(\cdot)$ denotes the pdf of a univariate student $t$ distribution with location parameter $\mu$, scale parameter $\lambda$, and degrees of freedom $v$. The skewing function $\Phi[\cdot]$ is the standard normal distribution function. In (4.2), we could use the cdf of a univariate $t$ distribution instead of the cdf of a standard normal distribution as the skewing function. However, in practice, the choice of standard normal cdf as a skewing function is for its simplicity over the cdf of $t$ distribution (Ma and Genton, 2004). Like FGSN distribution, the number of terms in the skewing function $\Phi[\cdot]$ determines the
number of modes of the distribution. The values of \( \omega_i \)'s determine the shape of the underlying distribution. Since the FGSN distribution with order \( K = 3 \) (number of terms in \( \Phi[\cdot] \)) can have at most two modes (Ma and Genton, 2004), similar phenomenon holds for the FGST distribution, because both the FGSN and FGST distributions have similar shape characteristics except that the FGST distribution has extra flexibility of capturing heavy-tailedness and high peakedness. Furthermore, like FGSN distribution, the FGST distribution also holds similar kind of nice properties. For instance, the FGST distribution reduces to skew-t distribution when \( \omega_1 \neq 0 \), and \( \omega_i = 0 \) for \( i = 2, 3, \ldots, (K + 1)/2 \). Again, if \( \omega_i = 0 \) for all \( i \), the FGST distribution reduces to a student \( t \) distribution. So, this distribution is an extension of skew-t and student \( t \) distributions. Like the \( t \)-distribution approximates the normal distribution for large degree-of-freedom, the FGST distribution converges to the FGSN distribution as \( v \), the degree-of-freedom, becomes large.

In this investigation, to model true exposure distribution in presence of covariate measurement errors we consider an FGST distribution with order \( K = 3 \). Because, according to the simulation studies in Ma and Genton (2004), FGST or FGSN distribution of order \( K = 3 \) has enough flexibility to capture a wide variety of distributions. However, bigger values of \( K \) may not be too damaging as mentioned by Ma, Genton and Davidian (2004). Their simulation studies demonstrate that the use of FGSN or FGST distribution with more number of terms in the skewing function usually does not impose skewness and/or multimodality if these features are, in fact, absent.

### 4.3.2 Parameter Estimation with FGST Distribution

The difficulties of carrying out the likelihood-type inference in measurement error scenario are already discussed in Section 1.4 of Chapter 1, and in Section 3.2 of Chapter 3. Furthermore, as discussed by Ma, Genton and Davidian (2004), the likelihood type of inference becomes more cumbersome under FGST exposure model. Hence, as usual, we take the Bayesian approach to carry out the inference about the model parameters. Armed with the development of scale mixture normal distribution representation of the student's \( t \) distribution (Racine-Poon, 1992), the inference about the model parameters using FGST exposure model can be implemented by Bayesian MCMC algorithm in a very convenient way. In fact, the implementation of MCMC algorithm with FGST exposure model requires almost no extra effort as compared to that with FGSN exposure model.
In order to carry out Bayesian inference we need the conditional posterior distributions of the model parameters. Using the similar notations that are used in Section 3.4.1, under FGST exposure model, the joint posterior distribution of the model parameters in the case of the logistic outcome model can be defined as

\[
\pi(x, \theta | y, x^*, z) = \prod_{i=1}^{n} \frac{f(y_i | x_i, z_i, \beta) f(x_i^* | x_i) f(x_i | z_i, \alpha, \lambda^2, \omega, v)}{\pi(\alpha, \beta, \lambda^2, \omega, v)} \times \prod_{i=1}^{n} \exp \left[ \frac{y_i (\beta_0 + \beta_1 x_i + \beta_2 z_i)}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_2 z_i)} \right] \left[ \prod_{i=1}^{n} e^{-\frac{1}{2\sigma^2} (x_i^* - z_i)^2} \right]
\]

\[
\left[ \prod_{i=1}^{n} \frac{\Gamma \left( \frac{v+1}{2} \right)}{\Gamma \left( \frac{v}{2} \right) \sqrt{\pi v \lambda^2}} \left[ 1 + \frac{1}{v} \left( \frac{x_i - \alpha_0 - \alpha_1 z_i}{\lambda} \right)^2 \right]^{-(v+1)/2} \right] \times \left[ \prod_{i=1}^{n} \Phi \left( \frac{\omega_0 (x_i - \alpha_0 - \alpha_1 z_i) + \omega_1 (x_i - \alpha_0 - \alpha_1 z_i)^3}{\lambda^3} \right) \right]
\]

\[\pi(\alpha) \pi(\beta) \pi(\lambda^2) \pi(\omega) \pi(v). \] 

(4.3)

As described in Ma, Genton and Davidian (2004), Racine-Poon (1992), Wakefield et al. (1994), and Wakefield (1996), a \( t \) distribution with \( v \) degrees of freedom can be represented as a scale mixture of normal distribution as

\[
t(x; \mu, \sigma^2, v) = \left[ \frac{\Gamma \left( \frac{v+1}{2} \right)}{\Gamma \left( \frac{v}{2} \right) \sqrt{\pi v \sigma^2}} \left[ 1 + \frac{1}{v} \left( \frac{x - \mu}{\sigma} \right)^2 \right]^{-(v+1)/2} \right]
\]

\[
= \int \frac{\gamma^{1/2}}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2} \chi^2_v(\gamma) d\gamma,
\]

(4.4)

where \( \chi^2_v \) is a chi-square distribution with \( v \) degrees of freedom. With this scale-mixture normal representation of the \( t \) distribution, the joint posterior distribution of the model parameters can
The parameter vector $\theta$ now contains the vector of latent parameters $\gamma_1, \gamma_2, \ldots, \gamma_n$ in addition to the usual parameters $\beta, \alpha, \lambda^2$ and $\omega$. The latent parameters $\gamma_1, \gamma_2, \ldots, \gamma_n$ come into play due to mixture representation of $t$-distribution in the FGST model.

As in the case of FGSN exposure model, we adopt same priors for $\alpha, \beta, \lambda^2$ and $\omega$. For $v$, a very diffuse proper prior with a large mean is considered. We consider the proper prior in order to ensure the propriety of the resulting posterior distribution of the model parameters. We choose very diffuse prior because of lack of adequate prior knowledge about the distribution of $v$. Without adequate prior knowledge it is not reasonable to prefer some values of $v$ over the others. Again, our choice of a prior with a large prior mean is guided by the reasoning that without any prior indication about the heavy-tailedness and/or high peakedness of the unobserved exposure distribution it is reasonable to center our belief in favor of the FGSN distribution. By allowing a large prior variance we let the data to take the role in posterior inference about $v$, and hence about the shape of the unobserved exposure distribution. So, for $v$, we choose a gamma prior with shape parameter 4 and scale parameter $0.031125$. For this prior distribution, the prior mean of $v$ is 128 and the prior standard deviation is 64. With these choice of priors the conditional posterior distributions of the model parameters are obtained as

\begin{align*}
\pi(x, \theta \mid y, x^*, z) &= \prod_{i=1}^{n} f(y_i \mid x_i, z_i, \beta) f(x_i^* \mid x_i) f(x_i \mid z_i, \alpha, \lambda^2, \omega, v) \\
&\pi(\alpha, \beta, \lambda^2, \omega, v) \times \prod_{i=1}^{n} \exp \left[ \frac{\gamma_i}{2} \left( \frac{x_i - \alpha_0 - \gamma_1 z_i}{\lambda} \right)^2 \right] \prod_{i=1}^{n} \frac{1}{\Gamma\left(\frac{\nu}{2}\right)^{n/2}} \exp \left( -\frac{\gamma_i^2}{2} \right) \prod_{i=1}^{n} \Phi \left( \omega_0 (x_i - \alpha_0 - \gamma_1 z_i) + \omega_1 (x_i - \alpha_0 - \gamma_1 z_i)^3 \right) \pi(\alpha) \pi(\beta) \pi(\lambda^2) \pi(\omega) \pi(v). 
\end{align*}

(4.5)

(i) for $\beta, \pi(\beta \mid y, x, x^*, z, \theta_{-\beta}) \propto \prod_{i=1}^{n} \frac{\exp(y_i (\beta_0 + \beta_1 x_i + \beta_2 z_i))}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_2 z_i)},$

where $\theta_{-\beta}$ is the vector of all model parameters other than $\beta$

(ii) for $X, \pi(x \mid y, x^*, z, \theta) \propto \prod_{i=1}^{n} \frac{\exp(y_i (\beta_0 + \beta_1 x_i + \beta_2 z_i))}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_2 z_i)} \exp \left[ \frac{\gamma_i (x_i - \alpha_0 - \gamma_1 z_i)^2}{2} \right] \left( x - \frac{\lambda^2 x^* + \gamma^2 v w_{\alpha} z}{\lambda^2 + \gamma^2} \right) \Phi \left( \omega_1 (x - \alpha_0 - \gamma_1 z_{\alpha}) + \omega_1 (x - \alpha_0 - \gamma_1 z_{\beta})^3 \right),$

where $w = (1, z), \alpha = (\alpha_0, \alpha_1)$ and $1$ is a column vector of 1s of length $n$, the sample size.
(iii) for $\alpha$, $\pi(\alpha | y, x, x^*, z, \theta_{-\alpha}) \propto \exp \left( -\frac{1}{2 \lambda^2} \left[ \alpha - (d'd)^{-1}d'u \right] \right) \left[ \prod_{i=1}^{n} \Phi \left( \frac{\omega_1(z_i-a_0-a_1z_i)}{\lambda} + \frac{\omega_2(z_i-a_0-a_1z_i)}{\lambda^2} \right) \right],
where $u = x, \sqrt{\gamma}$ and $d = w, \sqrt{\gamma}$

(iv) for $\lambda^2$, $\pi(\lambda^2 | y, x, x^*, z, \theta_{-\lambda^2}) \propto \left( \frac{1}{\lambda^2} \right)^{n+1+1} \exp \left( -\frac{0.5+0.5(u-\alpha a)^2}{\lambda^2}+\alpha a z_i \right) \prod_{i=1}^{n} \Phi \left( \frac{\omega_1(z_i-a_0-a_1z_i)}{\lambda} + \frac{\omega_2(z_i-a_0-a_1z_i)}{\lambda^2} \right]

(v) for $\omega_k$, $\pi(\omega_k | y, x, x^*, z, \theta_{-\omega_k}) \propto e^{-\frac{1}{2\lambda^2} \omega_k^2} \prod_{i=1}^{n} \Phi \left( \frac{\omega_1(z_i-a_0-a_1z_i)}{\lambda} + \frac{\omega_2(z_i-a_0-a_1z_i)}{\lambda^2} \right]

(vi) the conditional distribution of $\gamma_i$ is a gamma distribution with parameters $\nu + 1$ and $\frac{1}{2} + \frac{1}{2} \left( \frac{y_i-a_0a}{\lambda} \right)^2$

(vii) for $\nu$, $\pi(\nu | y, x, x^*, z, \theta_{-\nu}) \propto (2^{\nu/2} \Gamma(\frac{\nu}{2})^{-n} \prod_{i=1}^{n} \gamma_i)^{n/2-1} \nu^3 \exp(-0.03125\nu)$

Note that the conditional posterior distributions for $\gamma_i$'s depend on the corresponding individual observations and the degree-of-freedom $\nu$, a parameter common to all study subjects. That is, while estimating the individual level latent parameter $\gamma_i$'s information is borrowed across the individuals through $\nu$.

Now, we see that the conditional posteriors of $\beta$ and $\omega$ remains the same as they are in the case of FGSN exposure model. The distributional form of the conditional posteriors for $X$, $\alpha$, $\lambda^2$ are similar to the corresponding posteriors found under the FGSN exposure model. For instance, the conditional posterior for $X$ under FGST exposure model is a product of a normal density, the logistic cdf and the skewing function $\Phi[\cdot]$, which is exactly the case under FGSN exposure model. Hence, like the FGSN case, the independence sampler MH algorithm is found to work reasonably well to generate posterior simulations for $X$ and $\lambda^2$. But for $\alpha$, random–walk Metropolis–Hastings (RWMH) algorithm is used because unlike the FGSN case the independence sampler MH algorithm exhibits poor mixing in this case. As it is seen from the conditional posterior distribution of $\alpha$, the first term of it is a bivariate normal distribution with variance $\lambda^2(d'd)^{-1}$, which is independent of $\alpha$. Therefore, for the implementation of RHMH algorithm for $\alpha$ we choose to use $\lambda^2(d'd)^{-1}$ or some multiple of it as the tuning parameter. In the RHMH algorithm we update $\alpha = (\alpha_0, \alpha_1)$ simultaneously through bivariate normal proposal equal with tuning parameter $\lambda^2(d'd)^{-1}$. That is, $\alpha$ is updated as

$$\alpha^{\text{new}} = \alpha^{\text{current}} + \epsilon,$$

where $\epsilon \sim N_2 \left( 0, k \lambda^2(d'd)^{-1} \right)$ with $k \in \mathbb{R}$. For most of the simulation settings $\lambda^2(d'd)^{-1}$ or $1.5 \lambda^2(d'd)^{-1}$ are found to work reasonably well as the tuning parameter. The acceptance rates
for $\alpha$ are found to be around 40%-50% under different simulation settings. Also, the MCMC chains for $\alpha$ exhibit good mixing and convergence.

For $\beta$, like the FGSN case, RWMH algorithm is used. The components of the parameter vector $\beta$ are updated simultaneously through multivariate normal proposal distribution with tunning parameter equal to estimated variance of $\hat{\beta}_{\text{mle(naive)}}$ or some multiple of it, depending on the acceptance rate and the mixing of the MCMC chains. Here, $\hat{\beta}_{\text{mle(naive)}}$ is the ML estimate of $\beta$ obtained by regressing the outcome $Y$ on $X^*$ and $Z$. With these specific tunning parameters the MCMC chains for the components of $\beta$ are found demonstrate very good mixing and convergence. The acceptance rate are found to around 30%-40% in different simulation settings. Since, the conditional distribution of $\gamma$ is gamma it is straightforward to generate them using Gibbs sampler. For $v$, we use RWMH algorithm through a log-normal proposal with tunning parameter equal to $0.10^2$. That is, we update $v$ as $v_{\text{new}} = v_{\text{current}} \times \exp(N(0,0.10^2))$.

With this tunning parameter good mixing and convergence are observed for the MCMC chains of $v$. The observed acceptance rates for $v$ vary from 20% to 30% under different simulation settings. Finally, the shape parameters $\omega_1$ and $\omega_2$ are updated in block by RWMH algorithm through independent normal proposals. For all the simulation settings, same tunning parameter equal to 5 is used for both $\omega_1$ and $\omega_2$. With this tunning parameter, the MCMC chains for both $\omega_1$ and $\omega_2$ demonstrate reasonable mixing with acceptance rates of around 25%.

### 4.4 Results

In order to compare the performance of FGSN and FGST exposure models in correcting the bias in the estimate of the regression coefficients of the logistic outcome model, we carry out simulation studies. We consider exactly the same simulation designs that we consider so far in this dissertation. For computing the posterior estimates of the model parameters 5000 MCMC draws after adequate burn-in draws are used. And as usual, the posterior mean is used as an estimate for each of the model parameters.

Figures 4.6–4.10 represent the empirical kernel density plots of the estimates of regression parameters of the logistic outcome model. In each case, solid curve represents the kernel density plot of the estimates based on true covariate values. The dashed curve represents the kernel density of the estimates adjusted for measurement error under FGSN exposure model, and the dotted curve represents the same under FGST exposure model. Again, panel (a) in each figure represents the case of substantial measurement errors, while panel (b) represents...
the case of low measurement errors.

Figure 4.6 represents a design where the true exposure distribution is not fat-tailed. In this case of well-behaved exposure distribution the kernel density plots under FGST model almost coincide with the corresponding ones under FGSN model. For each of the regression parameters, kernel densities of the estimates adjusted for the measurement errors under both the exposure models very closely follow the kernel density plot of the estimates based on the true exposure values. So, both the exposure models perform quite well in correcting the measurement error-induced bias under simulation design 1. Both the models seem to adjust the bias almost completely while estimating the regression parameters of the logistic outcome model considered in this study.

In simulation design 2, a case where the true exposure values are generated from an exposure distribution which is similar to that in design 1 but the correlation between the true exposure and the other precisely measured covariate is stronger, the kernel density plots under FGST model apparently track the corresponding true kernel densities more closely than those under FGSN model [Figure 4.7(a) and Figure 4.7(b)]. It is also observed that the spread of the kernel densities under FGST model are very similar to the corresponding kernel densities under FGSN model. These findings indicate that, for this simulation design, moving from FGSN exposure model to FGST exposure model to seek more flexibility does not result in any efficiency loss of the estimated regression parameters though the FGST model provides better adjustment for the measurement error induced bias.

Simulation designs 3 and 4 represent the cases where true exposure distributions have very thick tails. But the difference between the two designs is that in design 3 there is moderate correlation (0.39) between the true exposure and the other precisely measured covariate, whereas in design 4 there is no such correlation. Simulation studies under both the SN and FGSN exposure models demonstrate that these two situations represent the areas of main concern while adjusting the measurement error induced bias in the parameters of the outcome model. Both the SN and FGSN models are found to perform badly in these two situations. However, Figures 4.8 and 4.9 demonstrate a different picture about the performance of the FGST exposure model. In both figures, the centers of the kernel density plots under FGST model are very close to those of the corresponding kernel densities of the estimates based on the true exposure values. So, though neither the SN nor the FGSN exposure models can adjust the bias in the regression parameter estimates of the logistic outcome model completely, the FGST exposure model seems to produce almost unbiased estimates of these parameters.
Figure 4.6: Kernel density plots of the estimates of regression parameters under simulation design 1
Figure 4.7: Kernel density plots of the estimates of regression parameters under simulation design 2
Figure 4.8: Kernel density plots of the estimates of regression parameters under simulation design 3
Figure 4.9: Kernel density plots of the estimates of regression parameters under simulation design 4
Figure 4.10: Kernel density plots of the estimates of regression parameters under simulation design 5
Finally, simulation design 5 is represented by the Figures 4.10(a) and 4.10(b). This design represents a situation where the true exposure distribution is skewed but does not have heavy tail, and there is very strong correlation (0.785) between the true but unobserved exposure and the other precisely measured covariate. In this case, both the FGSN and FGST exposure models are found to adjust the bias adequately in the estimates of the outcome model parameters.

As already discussed so far, we consider FGST distribution to allow more flexibility so that unobserved exposure can be reasonably approximated and thus measurement bias can be eliminated. However, it is possible that by being more flexible the FGST exposure model may cause the estimates of the outcome model parameters to be more variable than those under FGSN model. Kernel density plots under simulation design 3 and 4 exhibits slightly higher variability of the estimates under FGST exposure model as compared to those under FGSN exposure model. Therefore, it is important to investigate how the competitive exposure models perform in terms bias-variance tradeoff. To make such comparison we compute the mean squared error (MSE) of the estimates of \( \beta_1 \), the coefficient of the unobserved true exposure \( X \), under the two competitive exposure models, along with MSE of the estimates based on true exposure values. The MSEs are reported in Table 4.1.

### Table 4.1: MSE of the estimates of \( \beta_1 \)

<table>
<thead>
<tr>
<th>Simulation Design</th>
<th>Substantial measurement error</th>
<th>Low measurement error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FGSN</td>
<td>FGST</td>
</tr>
<tr>
<td>1</td>
<td>0.095</td>
<td>0.094</td>
</tr>
<tr>
<td>2</td>
<td>0.153</td>
<td>0.112</td>
</tr>
<tr>
<td>3</td>
<td>0.144</td>
<td>0.089</td>
</tr>
<tr>
<td>4</td>
<td>0.098</td>
<td>0.066</td>
</tr>
<tr>
<td>5</td>
<td>0.129</td>
<td>0.116</td>
</tr>
</tbody>
</table>

* Bench–mark estimates are based on true \( X \)-values

Results presented in Table 4.1 reveal that in all the simulation settings the FGST exposure model beats the FGSN model in terms of bias-variance tradeoff. The performance of FGST exposure model is much better in the cases of heavier tailed exposure distribution as demonstrated by the MSEs under simulation settings 3 and 4. Note that simulation settings 3 and 4 represent situations with heavier tailed exposure distributions. In these two cases, though the estimates under FGST exposure model have slightly higher variability, the estimates under FGSN model are considerably biased. Thus slightly higher variability of the estimates under FGST exposure model is offset by much better bias correction performance, which is reflected
by the lower MSE of the estimates under FGST exposure model.

4.5 Discussion

In all, simulation studies reveal that making the exposure model more flexible provides complete adjustment of the measurement error induced bias in the estimates of the regression parameters of the logistic outcome model we consider in our investigation process. We start our investigation from normal exposure model, and subsequently consider skew-normal, flexible generalized skew-normal and flexible generalized skew-t distributions to model the unobserved exposure. Among all the exposure models we consider in this study, the FGST exposure model is the only one which is found to adjust the bias caused by the measurement errors in the estimates of the parameters of the assumed logistic outcome model completely in all the simulation situations we consider. However, in the process of correcting the bias we observe a little increase in the variability of the estimates under FGST exposure model in situations when the true exposure distributions have very thick tails. But, a slight larger variability of the estimates in those bad situations should not be too worrisome as long as the FGST exposure model is adjusting the bias in the estimates completely or almost completely and producing estimates with better bias–variance tradeoff. In any statistical inference bias is the primary concern ahead of variance. It is more so in measurement error scenario, because in the presence of covariate measurement errors the bias might be so damaging that the center of the empirical distribution of the estimates may be far away from what is true. Not only that, the presence of measurement errors in covariate can artificially deflate the variance of the estimate and thus producing a confidence interval which may not cover the true parameter value. So, estimates which are unbiased or almost unbiased with slightly bigger variance are preferable to those with substantial bias but with little smaller variance. Our simulation studies reveal that the FGST exposure model achieves the former goal even in the two worst simulation scenarios. Furthermore, this exposure model exhibits better bias-variance tradeoff in all the simulation settings. In the other scenarios the FGST model is found to perform very similarly as FGSN model. That is, simulation studies provide evidence that allowing more flexibility by considering the FGST model does not do any harm in terms of the efficiency of the estimates in situations with well-behaved true exposure distribution though, like FGSN model, it produces the unbiased estimates of the parameters of the outcome model. This finding is in conformity with the similar sort of findings by Ma, Genton and Davidian (2004) that allowing more flexibility in terms of heavy-tailedness and/or high peakedness does not impose these features if these features are, in fact, not present.
Chapter 5

Performance of the Proposed Flexible Parametric Approach Compared to other Flexible Parametric and Popular Nonparametric Approaches

5.1 Introduction

As already mentioned in the previous chapters, in the presence of covariate measurement error the main criticism of the parametric approach revolves around the robustness of the estimated parameters of the outcome model to the distributional assumption about the true but unobserved exposure. Therefore, we search for a flexible parametric approach that is robust against the possible exposure model misspecification. At the same time our goal is to achieve reasonable efficiency as compared to the parametric approaches. Simulation studies in the previous chapter reveal that the use of flexible generalized skew-t (FGST) distribution to model the true but unobserved exposure provides the necessary degree of robustness against misspecification of the exposure distribution. The reasons why we are advocating for the parametric approach to measurement error problems are its flexibility to handle complicated epidemiological design set-up, its applicability to most general problems including those with discrete covariates subject to misclassification, and its ability to produce more efficient estimates and more reliable confidence intervals. Therefore, if the potential sensitivity of the parametric approaches to exposure model misspecification in measurement error problems can be avoided by using a flexible parametric distribution for the true but unobserved exposure then the better calibrated inference of the outcome model parameters can be obtained. Furthermore, the wide flexibility of the parametric approach can be utilized to handle measurement error problems in many situations, such as GLM, highly non-linear models, longitudinal data setting etc., without much difficulty. It can
be mentioned that all the existing functional approaches are suitable for only simple scenarios. Moreover, most of the existing flexible parametric, semi-parametric and non-parametric methods are quite demanding to implement and execute (Gustafson, 2004).

In this thesis we explore a very flexible parametric distribution to model the unobserved exposure distribution while estimating the regression parameters of the logistic outcome model in the presence of measurement errors in covariates. Extensive simulation studies conducted so far reveal that the FGST distribution as a flexible exposure model is best in correcting the measurement error bias. Hence, throughout the rest of this dissertation we refer to the suggested parametric method with FGST exposure model as the flexible parametric method.

In this chapter, we investigate the performance of the suggested flexible parametric approach in comparison with the widely used flexible parametric and nonparametric approaches to measurement error problems. As usual, we investigate the performance of the competitive methods through simulation studies.

5.2 FGST Exposure Model versus Normal Mixture Exposure Model

Richardson et al. (2002) investigate the suitability of the mixture of normal distributions to model the unobserved exposure in the case of logistic regression model in the presence of covariate measurement error. Mixture of normal distributions is a reasonable choice to model the unobserved exposure since normal–mixture is flexible to capture a wide variety of distributional shapes and thus can reduce the sensitivity to the modelling assumption related to the true but unobserved exposure. Richardson et al. (2002) investigate the performance of the normal–mixture exposure model through simulation studies. They also compare the performance of the normal–mixture exposure model with that of the semi–parametric likelihood approach (known as NPML approach) suggested by Roeder, Carol and Lindsay (1996). Their simulation studies reveal that the normal–mixture exposure model performs better in terms MSE than the NPML approach, though both the methods produce similar point estimates of the regression parameters of the logistic outcome model in the presence of covariate measurement error. Therefore, comparing the performance of the proposed flexible parametric approach with the approach based on normal–mixture exposure model can provide an indirect comparison of the proposed method with the NPML method. For investigating the comparative performance
we consider simulation studies similar to those described in Richardson et al. (2002).

5.2.1 Simulation Study Designs

Richardson et al. (2002) consider two simulation designs in their studies. In both cases, the true exposure $X$ is generated from mixture of normal distributions. The first design is referred to as "bimod" case which corresponds to generating $X$ from a well-separated symmetric bimodal mixture given by $0.5 \ N(-2, 1.0) + 0.5 \ N(2, 1)$. The surrogate $X^*$ is generated as $X^* \sim N(X, \tau^2)$ with $\tau^2 = 0.75, 1.49$ and $2.94$ corresponding to ratios, $R$, of the measurement error variance ($\tau^2$) to the variance of $X$ equal to $0.25$, $0.5$, and $1$, respectively. In this design, the sample size, $n$, is taken to be $240$ with a validation sample of size $n_1 = 60$. The outcome model is a logistic regression model defined as $\text{logit}(\pi_i) = \beta_0 + \beta_1 X$ with $\beta_0 = 0.5$ and $\beta_1 = 0.6$.

The second design is referred to as the "skew" case, where $X$ is simulated from an asymmetric normal mixture: $0.5 \ N(0.19, 0.08^2) + 0.2 \ N(1.05, 0.2^2) + 0.3 \ N(2.0, 0.48^2)$. The surrogate $X^*$ is generated from $N(X, \tau^2)$ with $\tau^2 = 0.25, 0.556$ and $1.11$ corresponding to variance ratios, $R$, of $0.25, 0.556$ and $1.11$, respectively. For this design, $n = 300$ and $n_1 = 50$. The outcome model is $\text{logit}(\pi_i) = \beta_0 + \beta_1 X$ with $\beta_0 = -0.8$ and $\beta_1 = 0.9$.

5.2.2 Implementation, and Investigation of the Performance of the MCMC Algorithms

Though the true exposure $X$ is generated from the mixture of normal distributions we assume FGST distribution to model the distribution of $X$ while estimating the regression parameters of the outcome model in the presence of measurement error. That is, in order to assess the robustness of FGST distribution as an exposure model, we misspecify the distribution of the true exposure $X$ as FGST distribution. For both the bimod and skew cases we consider an FGST distribution which can model at most two modes. Though in the second simulation design the true $X$ is simulated from a trimodal distribution, only one mode is pronounced. The presence of the other two smaller modes is not so pronounced, and hence the overall $X$ distribution looks skewed. So, the use of an FGST distribution with capability of capturing two modes may suffice to approximate the true $X$-distribution. In practice, the shape of the true exposure distribution can not be known as we do not observe the true exposure. Therefore, it is not possible to decide about the number of modes of FGST exposure model needed in a
particular situation to capture the shape of the true but unobserved exposure. However, in practice, the true exposure distributions do not usually have too many well-separated modes. Rather, they are either skewed or have a primary mode and few (one or two) less pronounced modes, in which cases the overall exposure distributions look almost skewed. Given the fact that to eliminate the measurement error induced bias in the estimates of the parameters of the outcome model a rough approximation of the exposure distribution is adequate, the FGST (or FGSN) distribution with capability of capturing at most two modes may suffice in most of the practical applications. However, one may consider more flexible FGST exposure model by allowing FGST distribution to accommodate more than two modes. In that case, a posterior check of the modal parameters of the FGST model can provide a guideline about whether to use more flexible FGST exposure model. For example, if the posterior estimate of any of the modal parameters is close to zero then it may be an indication of dropping that parameter from the FGST exposure model and using one with fewer number of modes.

With the proposed FGST exposure model and the same prior distributions used in Section 4.2.1 of the previous chapter for the model parameters, the conditional posterior distributions of the model parameters under the two simulation designs considered in Richardson et al. (2002) are obtained as

(i) for \( \beta \), \( \pi(\beta | y, x, x^* \theta_{-\beta}) \propto \left[ \prod_{i=1}^{n} \frac{\exp[y_i(\beta_0 + \beta_1 x_i)]}{1 + \exp[\beta_0 + \beta_1 x_i]} \right] \),

where \( \theta_{-\beta} \) is the vector of all model parameters other than \( \beta \).

(ii) for \( X \), \( \pi(x | y, x, x^*, \theta) \propto \left[ \prod_{i=1}^{n} \frac{\exp[y_i(\beta_0 + \beta_1 x_i)]}{1 + \exp[\beta_0 + \beta_1 x_i]} \right] \exp \left[ -\frac{(x - \frac{x_0^2 + \frac{x_1}{2}}{x_2 + \frac{x_3}{2}})^2}{2} \right] \Phi \left[ \frac{\omega_1 (x - \mu)}{\lambda} + \frac{\omega_2 (x - \mu)^3}{\lambda^3} \right], \)

(iii) for \( \mu \), \( \pi(\mu | y, x, x^* \theta_{-\mu}) \propto \exp \left[ -\frac{n \gamma}{2 \lambda^2} (\mu - \frac{m}{\gamma})^2 \right] \left[ \prod_{i=1}^{n} \Phi \left( \frac{\omega_1 (\mu - x_i)}{\lambda} + \frac{\omega_2 (x - \mu)^3}{\lambda^3} \right) \right], \)

where \( m = \sum_{i=1}^{n} x_i \) and \( \gamma = \sum_{i=1}^{n} \).

(iv) for \( \lambda^2 \), \( \pi(\lambda^2 | y, x, x^*, \theta_{-\lambda^2}) \propto \left( \frac{1}{\lambda^2} \right)^{n+1} e^{-\frac{n \gamma + 0.5 + 0.5 (u-d)(u-d)}{\lambda^2}} \left[ \prod_{i=1}^{n} \Phi \left( \frac{\omega_1 (x - \mu)}{\lambda} + \frac{\omega_2 (x - \mu)^3}{\lambda^3} \right) \right] \)

where \( u = x \sqrt{\gamma} \) and \( d = \mu \sqrt{\gamma} \).

(v) for \( \omega_i \), \( \pi(\omega_i | y, x, x^*, \theta_{-\omega_i}) \propto e^{-\frac{n \gamma + 1}{2 \lambda} \omega_i^2} \left[ \prod_{i=1}^{n} \Phi \left( \frac{\omega_i (x - \mu)}{\lambda} + \frac{\omega_2 (x - \mu)^3}{\lambda^3} \right) \right] \)

(vi) the conditional distribution of \( \gamma_i \) is a gamma distribution with parameters \( \frac{1}{2} + \frac{1}{2} (\frac{x_i - \mu}{\lambda})^2 \)

(vii) for \( v \), \( \pi(v | y, x, x^*, \theta_{-v}) \propto (2^{v/2} \Gamma_{1/2}^2)^{-n} \left( \prod_{i=1}^{n} \gamma_i^{1/2-1} \right) v^{y/2} \exp(-0.03125v) \)

79
(ix) for $\tau^2, \pi(\tau^2 \mid y, x, x^*, \theta_{-\tau^2}) \propto \left(\frac{1}{\tau^2}\right)^{\frac{n_y+1}{2}} \exp \left[-\frac{0.5+0.5(x^*-x)'(x^*-x)}{\tau^2}\right],$

which is an inverted Gamma distribution with shape parameter $\frac{n_y+1}{2}$ and scale parameter $0.5 + 0.5(x^*-x)'(x^*-x)$. For the posterior simulations of the respective model parameters we use the similar MCMC algorithms as described in Section 4.3.2 of Chapter 4. Since, the conditional posterior distribution of the measurement error variance $\tau^2$ is inverted Gamma we use Gibbs sampler for its posterior simulation.

For all the model parameters except $v$, the degree of freedom for FGST distribution, we observe good mixing and convergence of MCMC chains. Especially, the mixing and convergence for the regression parameter $\beta$ of the outcome model is found to be very satisfactory. Monitoring the convergence of the degree of freedom parameter $v$ warrants special attention while using the FGST exposure model. Since we choose FGST exposure model over FGSN exposure model to have more flexibility in terms of heavy tailedness and/or leptokurtic behavior, monitoring the posterior distribution of $v$ can provide a guideline about whether more flexibility in terms heavy tailedness and/or leptokurtic behavior is required. If the posterior estimate of $v$ is large then there is not so much point to go for FGST distribution, rather we can go for FGSN distribution. On the other hand, a small posterior estimate of $v$ supports the use of FGST distribution as an exposure model. For monitoring the convergence and mixing of MCMC chain of $v$, and for monitoring the posterior distribution of $v$ we plot the MCMC chain of $v$ for a single data set in each of the three measurement error scenarios under each of the two simulation designs in Figures 5.1 and 5.2.

Figure 5.1 represents the MCMC plots for the bimod case. The top panel represents the MCMC plot for the case when the ratio, $R$, of the measurement error variance to the variance of the true exposure is 0.25, a situation representing low measurement error; the middle panel represents the MCMC plot in the case of substantial measurement error ($R = 0.5$), and the bottom panel represents the situation of very high measurement error ($R = 1$).
Figure 5.1: MCMC plots for degree of freedom $v$ in the bimod case
Figure 5.2: MCMC plots for degree of freedom $v$ in the skew case
Though the MCMC plots exhibit poor mixing, after certain number of initial iterations (usually 5000) all of them move around a region of the posterior distribution which support very large values of $v$. In such situations, one may get skeptical about the utility of a heavier tailed exposure distribution. The type of MCMC plots we observe in Figure 5.1 can be an indication that poor mixing may not be because of the problem of convergence of the MCMC chains to the true parameter values. Rather it may be an indication that the data do not provide adequate support to FGST exposure model. The real issue here is that for $v$ large and substantial change in $v$ corresponds to almost no change in the distributional shape of the unobserved exposure $X$. Therefore, the posterior is rather flat for $v$, and hence the MCMC draws are showing slow mixing and convergence. In the cases like these, FGSN exposure model might be a better choice. Therefore, posterior plot of the degree-of-freedom parameter $v$ can be used as an ad—hoc way in choosing between FGST and FGSN distributions as exposure model. As an aside, Figure 5.1 corresponds to a simulation design where the true exposure distribution is a well-separated symmetric bimodal distribution. That is, there is no heavy-tailedness in the true exposure distribution. As indicated by movement of the MCMC plots of $v$ in the area of the posterior density which supports large values of $v$, the flexible parametric approach with FGST exposure model has been found to be able to identify that particular feature of the true exposure distribution.

Figure 5.2 represents the MCMC plots of the degree-of-freedom parameter, $v$, of the FGST exposure model in the skew case. The MCMC plots of $v$ exhibit very good mixing in all the three measurement error scenarios considered. As expected, all the MCMC plots of $v$ converge to the values that support heavy-tailedness of the true underlying exposure distribution, which confirms the true feature of the exposure distribution. In this simulation design, the true exposure is generated from an asymmetric mixture of three normal distributions. A histogram of the true exposure $X$ with an empirical kernel density estimate superimposed on it for a typical simulated data set under the skew case is presented in Figure 5.3. Though the true exposure distribution is theoretically a trimodal one, the minor two modes are so less pronounced compared to the major one that they simply introduce skewness and heavy tailedness, rather than multiple modes, in the exposure distribution. Accordingly, the small values of $v$ as demonstrated by the MCMC plots of $v$ are indications of the ability of FGST distribution to capture the true underlying exposure distribution.
Figure 5.3: Histogram and empirical density estimate (superimposed) of the true covariate $X$ in the skew case
5.2.3 FGST Exposure Model versus Normal-Mixture Exposure Model: Results of the Simulation Studies

For each data set under the simulation designs described in Section 5.2.1, to have posterior estimate of each of the regression parameters a posterior sample of size 5000 after a sufficient burn-in run is used. The convergence plots (not presented) confirm that in each of the cases considered MCMC samplers of the outcome model parameters converge well before 20000 iterations. However, to make sure that only the posterior draws after convergence are used for inferential purposes we discard first 25500 MCMC draws as burn-in out of 30500 draws. The remaining 5000 draws are used to estimate the regression parameters of the outcome model. For the point estimates of these parameters we use the posterior mean. Results corresponding to “bimod” and “skew” cases are presented in Table 5.1 and Table 5.2, respectively. As already mentioned, three values of the measurement error variance, \( \tau^2 \), are considered in each of the simulation designs. Three values of \( \tau^2 \) corresponds to ratios \( R \) of 0.25, 0.5 and 1 between the measurement error variance and the variance of the true covariate \( X \). Thus, \( R = 1 \) represents the worst case scenario of measurement error, where the size of the noise is as large as the underlying variability in the true exposure \( X \). In Tables 5.1 and 5.2, we report the mean and standard deviation of the estimates of \( \beta_1 \) obtained by using the proposed flexible parametric method to 50 simulated data sets. For the purpose of comparison, we report the results from Richardson et al. (2002), which are also based on 50 data sets simulated by them. We do not use the normal-mixture exposure model to the data sets we simulate for this study. So, the comparison made between FGST exposure model and the normal-mixture exposure model is subject to simulation error because our simulated realizations for the 50 data sets are different from the realizations of Richardson et al. (2002). However, provided that the results are averaged over 50 data sets in both the studies (our simulation study and that conducted by Richardson et al. (2002)) the effect of the simulation error should not bar us from making reasonable comparisons. We also report the average of \( \beta_1 \) estimates obtained by using true \( X \) values for the 50 data sets. Richardson et al. (2002) call this the “bench-mark” estimate. Finally, to quantify the loss of precision due to using noisy rather than the true values of the covariate we report the ratio of the mean-square error (MSE) of the estimates of \( \beta_1 \) obtained under the two measurement error models to that of the corresponding bench-mark estimates. For the comparison of MSE ratios under the FGST exposure model and the normal-mixture exposure model the MSE ratios under normal-mixture model are reported from Richardson et al. (2002).

Note that Richardson et al. (2002) report different bench-mark results for different measure-
Table 5.1: Sensitivity to exposure model misspecification: comparison of FGST and Gaussian-mixture exposure models in the bimod case

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimates under FGST exposure model and normal-mixture exposure model</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R = 0.25$</td>
<td>$R = 0.5$</td>
<td>$R = 1$</td>
</tr>
<tr>
<td>FGST</td>
<td>Mixture</td>
<td>Naive*</td>
<td>FGST</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.624</td>
<td>0.630</td>
<td>0.530</td>
</tr>
<tr>
<td>$sd(\beta_1)$</td>
<td>0.072</td>
<td>0.090</td>
<td>0.067</td>
</tr>
<tr>
<td>MSE ratio ($\beta_1$)</td>
<td>1.323</td>
<td>1.900</td>
<td>2.245</td>
</tr>
</tbody>
</table>

Benchmark estimates (no measurement error)

<table>
<thead>
<tr>
<th></th>
<th>Based on current study data</th>
<th>Based on Richardson et. al study data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R = 0.25$</td>
<td>$R = 0.5$</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.617</td>
<td>0.610</td>
</tr>
<tr>
<td>$sd(\beta_1)$</td>
<td>0.064</td>
<td>0.080</td>
</tr>
<tr>
<td>$MSE(\beta_1)$</td>
<td>0.004</td>
<td>0.006</td>
</tr>
</tbody>
</table>

* Naive estimates are based on noisy data

ment error magnitudes (Table 5.1). This is because they simulate three different realizations of the true exposure $X$ and add different measurement error to each of them to get the noisy versions of $X$. On the other hand, we simulate a single realization of $X$ and add three different degree of measurement errors to introduce low, medium and substantial degree of measurement error. Table 5.1 shows that the averages of point estimates of $\beta_1$ under both the exposure models are very close to the corresponding averages of the estimates obtained from true $X$. For example, magnitudes of differences between the average of the estimated $\beta_1$ under the FGST exposure model and the average of the corresponding benchmark estimates are found to be 0.007, 0.008 and 0.015 corresponding to low, substantial and very high measurement error situations, respectively. These differences under normal–mixture exposure model are 0.02, 0.02, and 0.03, respectively. This indicates that FGST exposure model exhibits very similar or slightly better bias correction performance as compared to the normal-mixture exposure model. Furthermore, FGST exposure model exhibits better bias-variance tradeoff as indicated by the MSE ratios. For FGST model, the MSE ratios vary from 1.323 to 2.220 as opposed to 1.830 to 2.940 under normal–mixture exposure model.

In the skew case, results presented in Table 5.2 demonstrate that the FGST exposure model outperform the normal mixture exposure model in eliminating bias in the estimated regression parameter $\beta_1$. In this case, the magnitude of differences between the mean of the estimated $\beta_1$ under FGST exposure model and the mean of the estimated $\beta_1$ based on the corresponding true $X$ values are 0.007, 0.010 and 0.087 under low, substantial and very high measurement error magnitudes (Table 5.1). This is because they simulate three different realizations of the true exposure $X$ and add different measurement error to each of them to get the noisy versions of $X$. On the other hand, we simulate a single realization of $X$ and add three different degree of measurement errors to introduce low, medium and substantial degree of measurement error. Table 5.1 shows that the averages of point estimates of $\beta_1$ under both the exposure models are very close to the corresponding averages of the estimates obtained from true $X$. For example, magnitudes of differences between the average of the estimated $\beta_1$ under the FGST exposure model and the average of the corresponding benchmark estimates are found to be 0.007, 0.008 and 0.015 corresponding to low, substantial and very high measurement error situations, respectively. These differences under normal–mixture exposure model are 0.02, 0.02, and 0.03, respectively. This indicates that FGST exposure model exhibits very similar or slightly better bias correction performance as compared to the normal-mixture exposure model. Furthermore, FGST exposure model exhibits better bias-variance tradeoff as indicated by the MSE ratios. For FGST model, the MSE ratios vary from 1.323 to 2.220 as opposed to 1.830 to 2.940 under normal–mixture exposure model.
Table 5.2: Sensitivity to exposure model misspecification: comparison of FGST and Gaussian-mixture exposure models in skew case

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimates under FGST exposure model and normal-mixture exposure model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R = 0.25$</td>
</tr>
<tr>
<td></td>
<td>FGST</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.915</td>
</tr>
<tr>
<td>sd($\beta_1$)</td>
<td>0.176</td>
</tr>
<tr>
<td>MSE ratio ($\beta_1$)</td>
<td>1.489</td>
</tr>
</tbody>
</table>

Bench-mark estimates (no measurement error)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Based on current study data</th>
<th>Based on Richardson et. al study data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R = 0.25$</td>
<td>$R = 0.5$</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.908</td>
<td>0.920</td>
</tr>
<tr>
<td>sd($\beta_1$)</td>
<td>0.145</td>
<td>0.160</td>
</tr>
<tr>
<td>MSE($\beta_1$)</td>
<td>0.021</td>
<td>0.021</td>
</tr>
</tbody>
</table>

$^*$ Naive estimates are based on noisy data

error scenarios, respectively. On the other hand, under normal–mixture exposure model, these magnitudes are found to be 0.05, 0.12 and 0.19, respectively–much higher magnitude of bias relative to the corresponding measurement error scenarios under FGST exposure model. Simulation studies in Richardson et al. (2002) reveal that in the presence of very large measurement error (e.g., $R = 1$), the bias correction performance of the normal mixture exposure model is, in fact, no better than that of normal exposure model. Furthermore, dealing with normal–mixture exposure distribution can be quite demanding and computationally burdensome when two or more covariates are measured with error. In such case, one will need to consider the mixture of multivariate normal densities. With multivariate normal–mixture distribution the computations become very demanding. Marin et al. (2005) discusses the difficulties of different orders while dealing with mixture densities. On the other hand, as demonstrated in Ma, Genton and Davidian (2004), with multivariate version of the flexible generalized skew–t or flexible generalized skew–normal distributions the Bayesian implementation can be carried out with no added difficulty as compared to univariate case. Finally, results presented in Table 5.2 demonstrate that in terms of bias–variance trade–off mixture of normal exposure model seems to perform even worse (than it does in terms of bias correction) as compared to the FGST exposure model. Under FGST exposure model the MSE ratios vary from 1.468 to 3.281, while these ratios vary from 1.930 to 8.380 under FGSN exposure model. Inadequate performance of normal mixture exposure model when the true exposure distribution is skewed is also acknowledged in Richardson et al. (2002).
5.2.4 Application to the Study of Cholesterol and Coronary Heart Disease

In order to compare the performance of the FGST exposure model with that of the normal–mixture exposure model in real data example we reanalyze the data concerning the risk of coronary heart disease (CHD) as function of blood cholesterol level. Roeder et al. (1996) first analyzed this data in a measurement error setting. Later on, Richardson et al. (2002) reanalyzed the same data set to compare the bias correction performance of the flexible parametric approach with normal–mixture exposure model with that of the NPML method proposed by Roeder et al. (1996). The part of the data set that is analyzed in Roeder et al. (1996), Richardson et al. (2002) and in this study is described in details in Roeder et al. (1996). In this study, a subject is recorded as having CHD if they have had a previous heart attack, an abnormal exercise electrocardiogram, a history of angina pectoris, and so forth. The main interest is the effect of low density lipoprotein cholesterol (LDL) on CHD. The low density lipoprotein is one of the components of total cholesterol (TC). The direct measurement of LDL level is more costly and technically demanding than that of the TC level. The question of interest here is thus whether TC can be used as a surrogate for LDL.

The sample that is analyzed consists of 256 subjects of which 113 are cases and 143 are controls. For this whole sample, CHD status, LDL and TC levels are available. As in Richardson et al. (2002), coronary heart disease status is taken as dependent variable $Y$, LDL/100 as exposure $X$ and TC/100 as the surrogate $X^*$. Note that the LDL level constitutes a part of TC level, i.e., $TC = LDL + K$, where, $K$ is the part of TC other than LDL with $E(k) \neq 0$. That is, assumption of unbiased measurement error does not hold in this case. Therefore, in terms of $X$ and $X^*$, we can write

$$X^* = a_0 + a_1 X + \epsilon,$$

with $E(\epsilon) = 0$. So, to accommodate biased measurement error we take $X^*|X \sim N(a_0 + a_1 X, \tau^2)$. Since the validation data are available, the parameters $a_0, a_1$ and $\tau^2$ of the measurement model are estimable and hence so are all the other parameters in the model. For validation sample, we randomly select 32 cases and 40 controls as are done in Roeder et al. (1996) and Richardson et al. (2002). The analysis of data is carried out by using flexible parametric approach with FGST distribution as exposure model. We compare the results with those reported in Richardson et al. (2002). Using the true covariate $X$ as the predictor, the prospective logistic regression estimate for $\beta_1$ is found to be 0.656 with posterior standard deviation of 0.336. The logistic regression analysis ignoring measurement error gives an estimated value of 0.545 for $\beta_1$ with a posterior standard deviation of 0.312. Analysis based on only the validation sample leads to an estimated
value of 1.441 with posterior standard deviation of 0.723. Results based on validation sample are the results obtained by fitting the logistic regression model to the subset of study sample for which the measurements on true exposure $X$ are available. As it is already mentioned, the validation sample in this study is a subset obtained by randomly selecting 32 cases and 40 controls out of a total of 113 cases and 143 controls, respectively. On the other hand, the measurement error adjusted estimates of $\beta_1$ are found to be 0.643 and 0.640 under FGST exposure model and normal–mixture exposure model, respectively with estimated posterior standard deviations of 0.402 and 0.370, respectively. Apparently, both the adjustment methods exhibit a good recovery of the estimated value of $\beta_1$ based on the complete set of error free data, with the estimate under FGST exposure model being closer to the benchmark estimate. Note that the measurement error adjusted results found in this study and in Richardson et al. (2002) study are subject to error that might arise due to different validation samples.

In short, the results demonstrate that estimation based on surrogate data is biased with inaccurate estimate of precision. With only small validation sample the estimate will always be less precise as it is expected because of smaller sample size. Moreover, estimate based on only validation sample may be very different (as the case in this study) from the true one because of sampling fluctuation. On other hand, the bias and the precision of the estimates improve substantially when adjusted for measurement errors.

5.3 Flexible Parametric Approach versus Parametric Correction (PC) and Non-Parametric Correction (NPC) Approaches

Parametric correction (PC) and non-parametric correction (NPC) approaches are proposed by Huang and Wang (2001). Both are functional approaches to measurement error problems. They are appealing in the sense that they do not require any assumption about the distribution of the true but unobserved exposure. The NPC method further relieves distributional assumption on the measurement error distribution. Huang and Wang (2001) show that the proposed PC and NPC estimators are consistent and asymptotically normal.

Correction–amenable estimating functions (Huang and Wang, 1999) are the building blocks of both the PC and NPC approaches. The correction amenable original estimating functions for logistic regression are obtained by applying weights to the normalized logistic score function in
Consider the postulated logistic regression model

\[ \Pr(Y = 1 \mid X) = \psi(\alpha_0 + \beta_0^T X) , \quad (5.1) \]

where \( \psi(t) = (1 + \exp(-t))^{-1} \). Given a sample of \( n \) iid realizations of \((Y, X)\), the normalized score function for the logistic regression can be written as

\[ S(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^{n} \left[ \left( Y - \psi(\alpha + \beta^T X) \right) \left( \frac{1}{X} \right) \right] , \quad (5.2) \]

where \( Y \) denotes the binary response, \( X \) denotes the vector of covariates. Obviously, \( S(\alpha, \beta) \) is not the only root-consistent estimating function. In fact, one can form a class of root-consistent estimating functions by applying weights to the score function given as

\[ S_w(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^{n} \left[ w(\alpha + \beta^T X) \left( Y - \psi(\alpha + \beta^T X) \right) \left( \frac{1}{X} \right) \right] , \quad (5.3) \]

for some positive weight function \( w \) (Huang and Wang, 1999, Lemma 2). Note that root-consistency of an estimating function can be defined to be that at least one zero-crossing exists in probability and every one is consistent. Huang and Wang (2001) choose specific weight functions \( 1 + \exp(-\alpha - \beta^T X) \) and \( 1 + \exp(\alpha + \beta^T X) \) to obtain the following pair of estimating functions:

\[ \tilde{\Phi}_-(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^{n} \left[ \left( Y - 1 + Y \exp(-\alpha - \beta^T X) \right) \left( \frac{1}{X} \right) \right] , \quad (5.4) \]

\[ \tilde{\Phi}_+(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^{n} \left[ \left( Y + (Y - 1) \exp(\alpha + \beta^T X) \right) \left( \frac{1}{X} \right) \right] . \quad (5.5) \]

These are the original estimating functions based on which the PC and NPC approaches are developed. Huang and Wang (2001) show that, under the logistic regression model (5.1) and the regularity conditions C1 and C2 in the appendix A in Huang and Wang (2001), the zero-crossings of the estimating functions \( \tilde{\Phi}_-(\alpha, \beta) \) and \( \tilde{\Phi}_+(\alpha, \beta) \), say, \( (\tilde{\alpha}_-, \tilde{\beta}_0^T) \) and \( (\tilde{\alpha}_+, \tilde{\beta}_+^T) \) respectively, in probability both exist and converge to \( (\alpha_0, \beta_0^T) \). Furthermore, asymptotically \( n^{1/2}((\tilde{\alpha}_-, \tilde{\beta}_0^T) - (\alpha_0, \beta_0^T)) \) and \( n^{1/2}((\tilde{\alpha}_+, \tilde{\beta}_+^T) - (\alpha_0, \beta_0^T)) \) are jointly normal with mean zero. Given the two sets of estimating functions in (5.4) and (5.5), a class of consistent estimators can be formed, and the remaining question is how to construct an asymptotically efficient one in the class. In this regard, Huang and Wang (2001) suggest synthesizing the estimating functions in an optimal fashion. The synthesis process and the necessary results are summarized in Lemma 2 in Huang and Wang (2001), which is stated below.
Lemma 2: Consider two normalized estimating functions \( \tilde{\Omega}_k(\gamma(k), \mu), k = 1, 2 \), each of which converges uniformly in \((\gamma(k)^T, \mu^T)^T\) in probability to a limit with a unique zero-crossing at \((\gamma_0(k)^T, \mu_0^T)^T\). Suppose that in a neighborhood of \((\gamma_0(k)^T, \mu_0^T)^T\), \( \tilde{\Omega}_k(\gamma(k), \mu) \) is differentiable, and the derivative converges uniformly in probability to a continuous and non-singular limit. In addition, asymptotically \( \tilde{\Omega}_k(\gamma_0(k), \mu_0), k = 1, 2 \), are jointly normal. Then, for any positive definite matrix \( \tilde{M} \) converging in probability, estimator

\[
\eta(\tilde{\Omega}_1, \tilde{\Omega}_2, \tilde{M}; \gamma^{(1)}, \gamma^{(2)}, \mu) = \arg \min_{\gamma^{(1)}, \gamma^{(2)}, \mu} \left[ \begin{array}{c} \tilde{\Omega}_1(\gamma^{(1)}, \mu) \\ \tilde{\Omega}_2(\gamma^{(2)}, \mu) \end{array} \right]^T \\
\times \tilde{M}^{-1} \left[ \begin{array}{c} \tilde{\Omega}_1(\gamma^{(1)}, \mu) \\ \tilde{\Omega}_2(\gamma^{(2)}, \mu) \end{array} \right]
\]

is consistent for \((\gamma_0^{(1)^T}, \gamma_0^{(2)^T}, \mu_0^T)^T\) and asymptotically normal. Further, suppose that the asymptotic variance of \( n^{1/2}\{\tilde{\Omega}_1(\gamma_0^{(1)}, \mu_0)^T, \tilde{\Omega}_2(\gamma_0^{(2)}, \mu_0)^T\}^T \) is consistently estimated by, say, positive definite \( \tilde{\Sigma} \). Then, \( \eta(\tilde{\Omega}_1, \tilde{\Omega}_2, \tilde{\Sigma}^T; \gamma^{(1)}, \gamma^{(2)}, \mu) \) is asymptotically efficient in the class of consistent estimators based on \( \tilde{\Omega}_1 \) and \( \tilde{\Omega}_2 \).

Now, let \( \tilde{\Lambda} \) be a consistent variance estimate of \( n^{1/2}\{\tilde{\Phi}_-(\alpha_0, \beta_0)^T, \tilde{\Phi}_+(\alpha_0, \beta_0)^T\}^T \). One example of \( \tilde{\Lambda} \) is given by

\[
\frac{1}{n} \sum_{i=1}^{n} \left[ \begin{array}{c} \{Y - 1 + Y \exp(-\alpha_0 - \beta_0^T X)\} \\ \{Y + (Y - 1) \exp(\alpha_0 + \beta_0^T X)\} \end{array} \right] \otimes \left[ \begin{array}{c} 1 \\ X \end{array} \right]
\]

with \((\alpha_0, \beta_0^T)^T\) replaced by a consistent estimator, say, \((\tilde{\alpha}_-, \tilde{\beta}_-^T)^T\) or \((\tilde{\alpha}_+, \tilde{\beta}_+^T)^T\), where \( v \otimes v^T \) for a vector \( v \). In the light of Lemma 2, Huang and Wang (2001) suggest using \((\tilde{\alpha}, \tilde{\beta}^T)^T = \eta(\tilde{\Phi}_-, \tilde{\Phi}_+, \tilde{\Lambda}; \alpha, \beta)\) as the estimator of the estimation procedure. By virtue of Lemma 2 in Huang and Wang (2001), this estimator is asymptotically normal and efficient among the consistent estimators based on \( \tilde{\Phi}_- \) and \( \tilde{\Phi}_+ \).

Both the PC and NPC approaches for adjusting for measurement error in covariates are based on the correction–amenable estimation procedure described above in this section. In the presence of measurement error, Huang and Wang (2001) formulate correction–amenable estimating functions for logistic regression. They show that the estimating functions they construct in the presence of measurement error have the same limits as the original root–consistent estimating functions (5.4) and (5.5). Then they synthesize the estimating functions according to Lemma 2 in Huang and Wang (2001) to find asymptotically efficient estimator. In PC approach they
utilize the known error distribution to construct the root-consistent estimating functions, while in NPC approach they assume that replicated surrogates are available. Details about the root-consistent estimating functions, as well as of the synthesized estimating functions under PC and NPC approaches are available in Huang and Wang (2001).

The fact of some estimators' holding optimal asymptotic properties is not the end of the story. First of all, their performance in finite sample scenarios is worth investigating. Simulation studies in Huang and Wang (2001) reveal that under substantial measurement error both the PC and NPC methods do not behave properly in finite sample cases. Both methods are asymptotically consistent under restrictive regularity conditions described in Huang and Wang (2001). Especially, regularity condition C2 in Huang and Wang (2001) requires that the true exposure distribution should not be very skewed and/or should not have a thick tail. To ensure the regularity condition C2 Huang and Wang (2001) simulate the true covariate $X$ from distributions which have very little skewness and do not have any heavy tailedness. That is, both the PC and NPC methods are consistent only when the true but unobserved exposure distribution is well-behaved in terms of skewness and heavy tailedness. Therefore, both of these functional methods are limited in their applicability.

Moreover, estimating equations of both the PC and NPC methods may not always have zero-crossing (Song and Huang, 2005). That is, both the methods may suffer from having no solution of the estimating equations used to find these estimators. The proposed flexible parametric approach coupled with Bayesian estimation method does not suffer from such a computational problem. None of the PC and NPC methods have the general applicability. It is not clear how to accommodate measurement error in longitudinal data setting in PC and NPC approaches. On the other hand, structural approaches (e.g. flexible parametric approach proposed in this study) have the flexibility to accommodate any complicated data and design set-up. With Bayesian computational methods such complicated set-up can be implemented accurately and without much difficulty. Again, both the PC and NPC methods are applicable only when replicated surrogate data on the unobserved exposure (covariate) are available for all the study subjects. On the other hand, flexible parametric approach coupled with Bayesian computational techniques is applicable even when replicated or validation data are available only on a subset of the study subjects.

In all, structural approaches have many advantages over functional approaches including the flexibility to accommodate complicated design and data set-up. However, concern about the structural approaches lies in the lack of robustness of structural approaches to the misspecifi-
cation of the true underlying exposure model. To reduce such sensitivity to exposure model misspecification we propose the FGST distribution to model the unobserved exposure distribution in the presence of exposure measurement error. In this section, we attempt to investigate the relative performance of proposed flexible parametric approach to measurement error modeling with FGST distribution as an exposure model. We compare the performance of the proposed method with those of the PC and NPC methods. We select PC and NPC approaches for our comparison because these two methods are found to perform better than the other existing non–parametric methods such as, regression calibration (RC) approach and convergent conditional score (CCS) approach (Huang and Wang, 2001).

5.3.1 Flexible Parametric Approach versus PC and NPC Approaches: Simulation Studies

In order to investigate the relative performance of the flexible parametric approach with that of PC and NPC approaches we consider simulation study designs similar to those considered in Huang and Wang (2001). In their studies Huang and Wang (2001) consider two different simulation designs. In the first design, a single and error prone covariate logistic regression model is considered with true parameter values \((\beta_0, \beta_1)^T = (0, 1)^T\). The true covariate \(X\) is generated from normal and modified chi-square distribution, respectively, with mean zero and standard deviation 1. According to Huang and Wang (2001), modified chi-square distribution is a standardized chi-square distribution with 1 degree of freedom after being truncated at 5. Truncation is required for the PC and NPC methods in order to fulfill the regularity condition \(C2\) given in the Appendix A of Huang and Wang (2001). The measurement error distribution is generated from normal and modified chi-square distribution, respectively, with mean zero. Two different values of the measurement error variance \(\tau^2\), 0.5\(^2\) and 1\(^2\), corresponding to moderate and heavy error-contamination scenarios are considered. In each case, two different sample sizes, 200 and 500, are considered. For assessing the performance of the PC and NPC methods Huang and Wang (2001) generate 1000 simulated samples in each of the simulation settings. However, in this study we will base our assessment on 200 simulated samples. However, the summary measures (median and interquartile distance divided by 1.349) that are used in Huang and Wang (2001) of the estimator distributions should be comparable even if the number of replications differs.

In the second design, a two covariate logistic regression model with parameters values \((\beta_0, \beta_1, \beta_2)^T = (0, 1, 1)^T\) is considered. The true covariates \(X\) have the standard bivariate normal distribution,
where only the second one is error prone. As in the first simulation design, normal and modified chi–square error distributions with mean 0 and standard deviation 0.75 are used. Two different correlation (0,0.5) between the two true covariates are considered. Two and three surrogate replicates, respectively, are used for error assessment. Under this simulation design, only one sample size (n = 500) is attempted.

5.3.2 Posterior Conditionals and Estimation of the Model Parameters

For estimating the model parameters with mismeasured covariates we use FGST distribution with capability of modeling at most two modes to model the unobserved exposure distribution. That is, we misspecify the true exposure distributions (normal and modified chi–square) as FGST distribution to investigate the robustness of the FGST distribution as an exposure model. Similar prior distributions, as described in Section 4.2.1 of Chapter 4 and in Section 5.2.2 of the current chapter, are considered for the respective model parameters. Note that, as already mentioned earlier, all the priors considered are close to being non–informative. Hence, we expect the results to be robust with respect the prior distributions. In the first simulation design, the case with a single covariate logistic regression model, the posterior conditionals are exactly similar to those obtained in Section 5.2.2 of this chapter. On the other hand, for simulation design 2 the posterior conditionals are similar to those derived in Section 4.3.2 of Chapter 4.

For computational purposes, similar MCMC algorithms, as described in Section 4.3.2 of Chapter 4, and in Section 5.2.2 of this chapter, are used for the respective model parameters. For example, for MCMC update of the true but unobserved covariate $X$, the independence sampler MH algorithm is used. For $\lambda^2$, the scale parameter of the FGST distribution, RWMH algorithm with proper tuning parameter is used. To ensure the positivity of the scale parameter $\lambda^2$ we use log–normal distribution as the proposal distribution for the posterior simulation of it. That is, in MCMC algorithm $\lambda^2$ is updated according to the following rule

$$\lambda_{new}^2 = \lambda_{old}^2 \times \exp(N(0, k^2)),$$

where $k^2$ is known as the tuning parameter or the jump size. This tuning parameter needs to be adjusted or tuned up to have good mixing and convergence of the MCMC chains of the respective parameters. In both the simulation designs, the tuning parameter for $\lambda^2$ is taken to be $k^2 = 0.35^2$. With this tuning parameter the MCMC chains for $\lambda^2$ exhibit good mixing and convergence in both the simulation designs. The acceptance rates for $\lambda^2$ are found to be around 25%-40%. For posterior simulation of $\beta$, RWMH algorithm is adopted. The parameter
vector $\beta$ is updated simultaneously by using multivariate normal proposal distribution with tuning parameter equal to $3 \times \text{var}(\hat{\beta}_{\text{mle(naive)}})$, where $\text{var}(\hat{\beta}_{\text{mle(naive)}})$ is the estimated variance–covariance matrix of the naive MLE of $\beta$. Naive MLE is the maximum likelihood estimate of $\beta$ based on the average of the replicated surrogates of $X$. Such simultaneous updating of the parameter vector can help in improving mixing of the MCMC algorithms when the components of the parameter vector are correlated. With this tuning parameter the acceptance rate for $\beta$ vector is found to be around 35%-44% under different simulation settings.

In the first simulation design, the conditional posterior distribution of location parameter $\mu$ of the distribution of the true covariate $X$ is the product of a normal distribution with mean $\bar{m}$ and variance $\frac{\lambda^2}{n\gamma}$, and the skewing function. The quantities $\bar{m}$ and $\gamma$ are defined under equation (iii) in Section 5.2.2. Since, the mean and variance of the first part of the conditional posterior distribution of $\mu$ is independent of $\mu$ we first consider using the independence sampler MH algorithm with proposal distribution equal to the first part of the posterior distribution of $\mu$ for posterior simulation of it. But, the MCMC chains of $\mu$ generated by the independence sampler MH algorithm exhibits poor mixing. Hence, we use MH algorithm with normal proposal distribution with tuning parameter $\frac{\lambda^2}{n\gamma}$, the variance of the first part of the posterior conditional of $\mu$. This tuning parameter depends on the model parameters other than $\mu$. Hence, at every iteration this tuning parameter is updated as each of the parameters in it is updated at every iteration. Therefore, we can call this an adaptive tuning parameter since it is updated or adapted as the iteration progresses. With this tuning parameter the MCMC chains of $\mu$ exhibit very good mixing and convergence with acceptance rates of around 40%. In the second simulation design, the main part of the posterior conditional of the location parameter $\alpha = (\alpha_0, \alpha_1)^T$ is a bivariate normal with mean and variance independent of $\alpha$. Therefore, for posterior simulation of $\alpha$ using the RWMH algorithm the tuning parameter for the bivariate normal proposal distribution is taken to be 3 times the variance matrix of the first part of the posterior conditional of $\alpha$. That is, we generate $\alpha$ from the bivariate normal proposal $N_2 [(d'd)^{-1}d'u, 3\lambda^2(d'd)^{-1}]$, where $N_2[., .]$ constitutes the first part of the posterior conditional $N_2[., .] \times \Phi(.)$ derived in Section 4.2.1 of Chapter 4. The quantities $d$ and $u$ in the mean and variance of $N_2[., .]$ are also defined in Section 4.3.2 of Chapter 4. With this tuning parameter the acceptance rates for $\alpha$ is found to be around 35%. The MCMC chains also exhibit good mixing and convergence.

For the shape parameters, $\omega_1$ and $\omega_2$, of the FGST exposure distribution the RWMH algorithm is considered. The shape parameters are updated in a block (simultaneously). Independent normal proposals with same tuning parameter are used. In the first simulation design, the values
of tuning parameters for $\omega_1$ and $\omega_2$ providing good mixing and convergence of the MCMC chains are 0.25 and 1 under sample sizes 500 and 200, respectively, when the true exposure distribution is modified chi-square. On the other hand, under normal exposure model a value 3 for both the tuning parameters is used irrespective of sample size. In all the simulation settings under the second simulation design a value of 3 as the tuning parameter has been found to work reasonably well. For the degree of freedom, $v$, of the FGST exposure model, again RWMH algorithm with log-normal proposal distribution is used. The tuning parameters used in the single and double covariate scenarios are $0.15^2$ and $0.05^2$, respectively. Very good mixing and convergence are exhibited by the MCMC chains under these tuning parameters. The acceptance rates observed are around 25%. Finally, for posterior simulation of $\gamma_k$s and the error variance $\tau^2$ Gibbs sampler is employed.

### 5.3.3 Results

In both cases of the two simulation designs, the comparison of results are subject to simulation error, and differences that may persist due to the fact that different number of simulated data sets are used in this study than the studies in Huang and Wang (2001). In each case, the bias and the variation of an estimator are summarized by median bias and the interquartile distance divided by 1.349, respectively, where 1.349 is the ratio of interquartile distance and standard deviation for normal distribution. We use these two particular summary statistics to summarize the bias and variability of an estimator because Huang and Wang (2001) use them for the same purposes in their studies. As mentioned in Huang and Wang (2001), these two measures have the advantage, over the conventional empirical mean bias and standard deviation, of being invariant under one-to-one transformation of parameter and also being more robust to outliers. Therefore, for making the results of our study comparable with those reported in Huang and Wang (2001), we report empirical median bias ($B$), estimated standard deviation ($D$) based on interquartile distance of the estimates, empirical coverage ($C$) of 95% credible interval. For empirical coverage ($C$), Huang and Wang (2001) report the empirical coverage of 95% bootstrap percentile confidence interval for PC and NPC approaches.

Results obtained under the first simulation design are presented in Table 5.3. As demonstrated by the simulation results presented in Table 5.3, the PC approach is found to be the worst performer in terms of bias correction when the true exposure distribution is normal. The flexible parametric method stays in the middle with NPC approach being the best performer. The performance of the PC approach deteriorates further under non-normal error distribution.
In the case of non-normal exposure distribution, the flexible parametric approach and the PC approach perform similarly and better than the NPC approach under normal error distribution. Under non-normal exposure and non-normal error the flexible parametric approach performs the best and reasonably well in all the simulation settings. The PC and NPC approach, on the other hand, produce substantial bias when measurement error is large.

In all, the performance of the flexible parametric approach is found to be consistently adequate in terms of bias correction irrespective of the true exposure and measurement error distribution, and the degree of measurement error. Whereas, the PC and NPC approaches perform adequately in some situations but produce substantial bias in some other situations. For example, PC approach is found to perform badly in several situations under non-normal error distribution. In those cases, the magnitudes of the estimated median bias are found to be 0.121, 0.142, 0.118 and 0.144 (true covariate value is 1.0). For NPC approach, bad performance are observed when sample size is small with both exposure and measurement error having modified chi-square distribution. In those scenarios, the magnitudes of the estimated median bias of the NPC estimator are found to be 0.086, 0.161, and 0.068. In contrast, in the worst case scenarios for the flexible parametric approach these magnitudes are 0.064, 0.073 and 0.067.

When it comes to the question of efficiency of the estimates, the flexible parametric approach outperform both the PC and NPC approach in all but two out of 16 simulation settings we consider in the case of single covariate logistic outcome model. In the two situations where the PC approach just beats the flexible parametric approach, the estimated relative efficiencies are found to be 1.03 and 1.56 in favor of the PC approach. However, in those two situations, the magnitudes of the estimated median bias of the PC estimator are found to be 0.142 and 0.117 against only 0.058 and 0.034, respectively, under flexible parametric approach. In all the other situations, the estimated relative efficiency varies from 1.02 to 1.66 in favor of the flexible parametric approach. As expected, the NPC approach is found to be the worst performer in terms of the efficiency of the estimates in all simulation settings. When compared to the flexible parametric approach, the relative efficiency of the estimates are found to vary from 1.05 to 2.21 in favor of the flexible parametric approach estimator. In the case, when the NPC estimator comes close to the flexible parametric approach estimator (i.e., when relative efficiency of flexible parametric approach estimator is 1.05) the magnitude of bias in the NPC estimator is found to much higher than that of the flexible parametric approach estimator (0.161 for NPC estimator against 0.034 for flexible parametric approach estimator). In a nutshell, comparison of results from our simulation studies with those obtained in the simulation studies conducted in Huang and Wang (2001) reveals that the proposed flexible parametric approach achieves
Table 5.3: Sensitivity to exposure model misspecification: comparison of FGST model with PC and NPC methods in single-covariate logistic regression model

<table>
<thead>
<tr>
<th>n</th>
<th>τ</th>
<th>NV</th>
<th>PC</th>
<th>NPC</th>
<th>FGST</th>
<th>NV</th>
<th>PC</th>
<th>NPC</th>
<th>FGST</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>X ~ Normal</td>
<td>X ~ Modified chi-square</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.5</td>
<td>F</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>-119.3</td>
<td>-28.4</td>
<td>-0.6</td>
<td>19.1</td>
<td>-167.7</td>
<td>-11.7</td>
<td>-58.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>177.5</td>
<td>230.7</td>
<td>262.5</td>
<td>139.0</td>
<td>193.6</td>
<td>296.1</td>
<td>334.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>90.2</td>
<td>92.9</td>
<td>95.2</td>
<td>95.0</td>
<td>88.2</td>
<td>92.6</td>
<td>96.2</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>F</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>-</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>-363.9</td>
<td>-58.7</td>
<td>-29.3</td>
<td>64.0</td>
<td>-435.2</td>
<td>-52.5</td>
<td>-54.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>115.3</td>
<td>266.4</td>
<td>335.3</td>
<td>262.3</td>
<td>130.3</td>
<td>328.4</td>
<td>482.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>46.3</td>
<td>94.5</td>
<td>97.2</td>
<td>97.0</td>
<td>30.5</td>
<td>97.4</td>
<td>97.4</td>
</tr>
<tr>
<td>500</td>
<td>0.5</td>
<td>F</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>-</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>-134.8</td>
<td>-28.4</td>
<td>-1.5</td>
<td>12.5</td>
<td>-174.7</td>
<td>-18.1</td>
<td>-28.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>115.3</td>
<td>146.4</td>
<td>160.4</td>
<td>134.2</td>
<td>113.3</td>
<td>182.6</td>
<td>196.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>80.8</td>
<td>94.4</td>
<td>94.6</td>
<td>95.5</td>
<td>75.1</td>
<td>92.6</td>
<td>94.9</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>F</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>-</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>-379.6</td>
<td>-57.4</td>
<td>-18.1</td>
<td>54.8</td>
<td>-433.0</td>
<td>-44.3</td>
<td>-1.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>95.6</td>
<td>174.3</td>
<td>222.3</td>
<td>150.1</td>
<td>81.5</td>
<td>203.5</td>
<td>309.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>7.2</td>
<td>93.6</td>
<td>95.5</td>
<td>94.5</td>
<td>2.7</td>
<td>95.4</td>
<td>96.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: True covariate X is distributed with mean zero and standard deviation 1. The measurement error is distributed with mean zero and respective standard deviations τ. NV: naive, PC: parametric correction, NPC: nonparametric correction, F: root finding failure (%), B: empirical median bias×1000, D: estimated standard deviation×1000 based on interquartile distance, C: empirical coverage(%) of 95% bootstrap percentile confidence interval for PC and NPC approaches and of 95% credible interval for flexible parametric approach.
better bias-efficiency trade-off compared to PC and NPC approaches.

Finally, regarding the confidence coverage, the flexible parametric approach and NPC approach are seen to achieve accurate confidence coverage, whereas the PC approach suffers from slightly poor confidence coverage when the sample size is small under non-normal error. Note that, as seen from the simulation results, the naive estimates suffer from serious bias, and extremely poor confidence coverage.

Results obtained under the second simulation design are presented in Table 5.4. As already mentioned, in two-covariate case the second covariate is measured with error. In this case of multiple covariate model, the measurement error causes bias not only in the regression parameter of the error-prone covariate but also in that of the accurately measured one when the true covariates are correlated. Simulation results presented in Table 5.4 demonstrate this fact. In this case, besides reporting the measurement-error adjusted estimates (obtained from simulated data in this study and from Huang and Wang (2001)), we also report the bench-mark estimates obtained from our simulated data. The bench-mark estimates will allow us to see the discrepancy of the measurement-error adjusted estimates under FGST exposure model to those obtained using the two covariate values.

Results presented in Table 5.4 demonstrate that all the three methods achieve reasonable bias correction performance in all the simulation settings while estimating the regression coefficient of the error-prone covariate. In estimating the coefficient of the error-prone covariate the NPC method seems to do a better job than the other two methods in terms of bias correction. However, the bias correction performance of the NPC estimator is considerably worse than the PC and the flexible parametric approach estimator for the regression coefficient of the accurately measured covariate. For accurately measured covariate, the flexible parametric approach demonstrates the best performance in terms of bias correction. As regards the relative efficiency of the estimates of the regression coefficient of the error-prone covariate, simulation results show that like the single-covariate case the flexible parametric approach performs the best in all but two settings with relative efficiency varying from 0.67 to 1.07. Note that the relative efficiency computed in this double covariate case is with respect to the corresponding bench-mark estimates based on true X simulated in this study. Though the PC estimator beats the flexible parametric approach estimator in two simulation settings, the relative efficiencies of the flexible parametric approach estimator in those two situations are very close to those of the PC estimator (0.79 and 0.67 in the case of flexible parametric approach versus 0.84 and 0.70 in the case of PC approach, respectively). In all other situations, the proposed flexible
Table 5.4: Sensitivity to exposure model misspecification: comparison of FGST model with PC and NPC methods in double-covariate logistic regression model

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>NV</th>
<th>PC</th>
<th>NPC</th>
<th>FGST</th>
<th>Benchmark</th>
<th>NV</th>
<th>PC</th>
<th>NPC</th>
<th>FGST</th>
<th>Benchmark</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2 replicates</td>
<td>3 replicates</td>
<td>3 replicates</td>
<td>3 replicates</td>
<td>3 replicates</td>
<td>3 replicates</td>
<td>3 replicates</td>
<td>3 replicates</td>
<td>3 replicates</td>
<td>3 replicates</td>
</tr>
<tr>
<td>0.5 $\beta_1$ F</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2</td>
<td>-</td>
<td>-</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>-100.9</td>
<td>-18.9</td>
<td>76.1</td>
<td>15.9</td>
<td>10.8</td>
<td>-87.2</td>
<td>-16.7</td>
<td>66.1</td>
<td>5.2</td>
<td>10.8</td>
</tr>
<tr>
<td>D</td>
<td>144.5</td>
<td>175.8</td>
<td>255.1</td>
<td>159.9</td>
<td>152.2</td>
<td>146.6</td>
<td>174.9</td>
<td>237.8</td>
<td>163.8</td>
<td>152.2</td>
</tr>
<tr>
<td>RE</td>
<td>-0.86</td>
<td>0.60</td>
<td>0.55</td>
<td>-</td>
<td>-</td>
<td>0.87</td>
<td>0.64</td>
<td>0.53</td>
<td>0.93</td>
<td>-</td>
</tr>
<tr>
<td>C</td>
<td>87.3</td>
<td>93.9</td>
<td>92.0</td>
<td>95.5</td>
<td>94.5</td>
<td>90.5</td>
<td>94.2</td>
<td>93.4</td>
<td>94.0</td>
<td>94.5</td>
</tr>
<tr>
<td>$\beta_2$ B</td>
<td>-297.1</td>
<td>-63.2</td>
<td>31.6</td>
<td>61.5</td>
<td>9.7</td>
<td>-217.8</td>
<td>-46.9</td>
<td>36.7</td>
<td>11.5</td>
<td>9.7</td>
</tr>
<tr>
<td>D</td>
<td>126.4</td>
<td>186.0</td>
<td>303.5</td>
<td>197.4</td>
<td>146.8</td>
<td>125.9</td>
<td>188.4</td>
<td>278.0</td>
<td>145.8</td>
<td>146.8</td>
</tr>
<tr>
<td>RE</td>
<td>0.79</td>
<td>0.48</td>
<td>0.74</td>
<td>-</td>
<td>-</td>
<td>0.78</td>
<td>0.53</td>
<td>1.01</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C</td>
<td>45.5</td>
<td>94.4</td>
<td>95.9</td>
<td>93.5</td>
<td>93.5</td>
<td>93.5</td>
<td>93.5</td>
<td>93.5</td>
<td>93.5</td>
<td>93.5</td>
</tr>
<tr>
<td>0.0 $\beta_1$ F</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>-</td>
<td>-</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>-41.4</td>
<td>-31.2</td>
<td>42.6</td>
<td>17.9</td>
<td>9.3</td>
<td>-27.9</td>
<td>-31.3</td>
<td>31.2</td>
<td>9.3</td>
<td>4.3</td>
</tr>
<tr>
<td>D</td>
<td>125.6</td>
<td>143.3</td>
<td>200.0</td>
<td>129.7</td>
<td>117.7</td>
<td>118.4</td>
<td>128.5</td>
<td>172.8</td>
<td>114.4</td>
<td>117.7</td>
</tr>
<tr>
<td>RE</td>
<td>-0.87</td>
<td>0.59</td>
<td>0.91</td>
<td>-</td>
<td>-</td>
<td>0.92</td>
<td>0.68</td>
<td>1.03</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C</td>
<td>95.6</td>
<td>94.8</td>
<td>93.6</td>
<td>96.5</td>
<td>94.5</td>
<td>96.6</td>
<td>95.1</td>
<td>94.7</td>
<td>94.5</td>
<td>94.5</td>
</tr>
<tr>
<td>$\beta_2$ B</td>
<td>-246.6</td>
<td>-41.2</td>
<td>21.6</td>
<td>39.9</td>
<td>-1.6</td>
<td>-217.8</td>
<td>-46.9</td>
<td>36.7</td>
<td>11.5</td>
<td>9.7</td>
</tr>
<tr>
<td>D</td>
<td>112.1</td>
<td>188.1</td>
<td>286.1</td>
<td>185.8</td>
<td>157.2</td>
<td>113.2</td>
<td>154.1</td>
<td>212.2</td>
<td>161.9</td>
<td>108.5</td>
</tr>
<tr>
<td>RE</td>
<td>0.78</td>
<td>0.53</td>
<td>1.01</td>
<td>-</td>
<td>-</td>
<td>0.78</td>
<td>0.53</td>
<td>1.01</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C</td>
<td>45.5</td>
<td>94.4</td>
<td>95.9</td>
<td>93.5</td>
<td>93.5</td>
<td>93.5</td>
<td>93.5</td>
<td>93.5</td>
<td>93.5</td>
<td>93.5</td>
</tr>
</tbody>
</table>

Note: True covariate $X$ is distributed with mean zero and standard deviation 1. The measurement error is distributed with mean zero and respective standard deviations $\tau$. NV: naive, PC: parametric correction, NPC: nonparametric correction, $F$: root finding failure (%), $B$: empirical median bias x 1000, $D$: estimated standard deviation x 1000 based on interquartile distance, $RE$: estimated relative efficiency with respect to corresponding benchmark estimates, $C$: empirical coverage (%)

$\epsilon \sim$ Normal

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>NV</th>
<th>PC</th>
<th>NPC</th>
<th>FGST</th>
<th>Benchmark</th>
<th>NV</th>
<th>PC</th>
<th>NPC</th>
<th>FGST</th>
<th>Benchmark</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5 $\beta_1$ F</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>-</td>
<td>-</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>98.5</td>
<td>-2.8</td>
<td>79.1</td>
<td>5.9</td>
<td>-4.9</td>
<td>80.8</td>
<td>-17.0</td>
<td>76.6</td>
<td>7.3</td>
<td>-4.9</td>
</tr>
<tr>
<td>D</td>
<td>145.8</td>
<td>176.0</td>
<td>257.8</td>
<td>158.1</td>
<td>129.6</td>
<td>121.1</td>
<td>173.3</td>
<td>256.7</td>
<td>123.7</td>
<td>129.6</td>
</tr>
<tr>
<td>RE</td>
<td>-0.74</td>
<td>0.50</td>
<td>0.82</td>
<td>-</td>
<td>-</td>
<td>0.75</td>
<td>0.50</td>
<td>1.05</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C</td>
<td>87.3</td>
<td>93.6</td>
<td>91.9</td>
<td>92.5</td>
<td>95.5</td>
<td>89.7</td>
<td>93.6</td>
<td>92.7</td>
<td>97.0</td>
<td>95.5</td>
</tr>
<tr>
<td>$\beta_2$ B</td>
<td>-297.2</td>
<td>-81.7</td>
<td>7.5</td>
<td>44.8</td>
<td>-1.6</td>
<td>-180.7</td>
<td>-37.3</td>
<td>36.7</td>
<td>11.5</td>
<td>9.7</td>
</tr>
<tr>
<td>D</td>
<td>112.1</td>
<td>188.1</td>
<td>286.1</td>
<td>185.8</td>
<td>157.2</td>
<td>113.2</td>
<td>154.1</td>
<td>212.2</td>
<td>161.9</td>
<td>108.5</td>
</tr>
<tr>
<td>RE</td>
<td>0.84</td>
<td>0.55</td>
<td>0.85</td>
<td>-</td>
<td>-</td>
<td>0.91</td>
<td>0.61</td>
<td>0.92</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C</td>
<td>45.3</td>
<td>93.9</td>
<td>97.0</td>
<td>93.0</td>
<td>95.5</td>
<td>70.5</td>
<td>93.7</td>
<td>95.3</td>
<td>95.0</td>
<td>95.5</td>
</tr>
<tr>
<td>0.0 $\beta_1$ F</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>-</td>
<td>-</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>-42.5</td>
<td>-29.0</td>
<td>48.5</td>
<td>29.5</td>
<td>18.4</td>
<td>-28.0</td>
<td>-28.6</td>
<td>31.6</td>
<td>31.1</td>
<td>18.4</td>
</tr>
<tr>
<td>D</td>
<td>127.9</td>
<td>142.0</td>
<td>183.2</td>
<td>122.7</td>
<td>131.8</td>
<td>118.2</td>
<td>128.7</td>
<td>184.1</td>
<td>133.5</td>
<td>131.8</td>
</tr>
<tr>
<td>RE</td>
<td>-0.93</td>
<td>0.72</td>
<td>1.07</td>
<td>-</td>
<td>-</td>
<td>1.02</td>
<td>0.77</td>
<td>0.99</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C</td>
<td>95.0</td>
<td>94.8</td>
<td>94.1</td>
<td>93.5</td>
<td>93.5</td>
<td>96.1</td>
<td>94.9</td>
<td>94.2</td>
<td>93.5</td>
<td>93.5</td>
</tr>
<tr>
<td>$\beta_2$ B</td>
<td>-247.2</td>
<td>-55.7</td>
<td>2.7</td>
<td>53.7</td>
<td>-1.6</td>
<td>-217.8</td>
<td>-46.9</td>
<td>36.7</td>
<td>11.5</td>
<td>9.7</td>
</tr>
<tr>
<td>D</td>
<td>101.8</td>
<td>165.4</td>
<td>203.7</td>
<td>157.3</td>
<td>105.1</td>
<td>122.3</td>
<td>166.8</td>
<td>213.2</td>
<td>130.0</td>
<td>105.1</td>
</tr>
<tr>
<td>RE</td>
<td>0.64</td>
<td>0.52</td>
<td>0.67</td>
<td>-</td>
<td>-</td>
<td>0.63</td>
<td>0.49</td>
<td>0.84</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C</td>
<td>52.6</td>
<td>94.6</td>
<td>96.4</td>
<td>96.5</td>
<td>93.0</td>
<td>73.0</td>
<td>93.3</td>
<td>93.8</td>
<td>95.0</td>
<td>93.0</td>
</tr>
</tbody>
</table>

Note: True covariate $X$ is distributed with mean zero and standard deviation 1. The measurement error is distributed with mean zero and respective standard deviations $\tau$. NV: naive, PC: parametric correction, NPC: nonparametric correction, $F$: root finding failure (%), $B$: empirical median bias x 1000, $D$: estimated standard deviation x 1000 based on interquartile distance, $RE$: estimated relative efficiency with respect to corresponding benchmark estimates, $C$: empirical coverage (%) of 95% bootstrap percentile confidence interval for PC and NPC approaches and of 95% credible interval for flexible parametric approach.
parametric approach has much better relative efficiency compared to that of PC approach. The relative efficiency of the PC estimator vary from 0.63 to 0.91 with respect to the corresponding bench-mark estimates. Again, like the single-covariate case, the NPC approach exhibits the worst performance as the question of efficiency of the estimates comes into account. For NPC approach, the relative efficiency ranges from only 0.47 to the lowest to only 0.61 to the highest.

Finally, all the three methods are seen to achieve very reasonable confidence coverage under all the simulation settings considered in this two-covariate case. Noticeably, all the three estimators improve in bias correction and precision as the replications of surrogate measurement for the error-prone covariate increases from two to three. This is usual because as the replication number increases for the surrogate measurement more information are available about the parameter, \( \tau \), of the measurement error model. Finally, in double-covariate case, true exposure \( X \) is generated from normal distribution. This may be the cause of PC and NPC approaches' showing almost similar bias correction performance as the flexible parametric approach. As demonstrated in the earlier chapters and those in Section 5.2.3 of this chapter, the proposed flexible parametric method exhibits much superior bias correction performance in the cases where the true exposure distribution has heavier tail. Since Huang and Wang (2001) does not apply the PC and NPC approaches to the situations with heavier tailed exposure distributions their performance relative to the flexible parametric approach can not be investigated in those situations. In fact, theoretically, the PC and NPC approaches are not applicable to heavy-tailed exposure distribution as stated by the regularity condition \( C2 \) in Huang and Wang (2001). However, in many medical studies the true exposure distribution is usually highly skewed or has heavier tail. Therefore, the applicability of PC and NPC approaches in practical situations seems to be limited.
Chapter 6

A New Computational Method to Evaluate Large Sample Bias in the Estimate of Regression Coefficients due to Exposure Model Misspecification in the Presence of Measurement Error

6.1 Introduction

So far, in the earlier chapters, we have studied the impact of misspecified exposure model in correcting the bias in the estimates of the regression parameters in the presence of measurement errors. We have studied the empirical distribution of the regression parameter estimates based on 50 data sets to demonstrate the degree of bias when the exposure model is misspecified. But the empirical distribution of the estimates based on only 50 data sets may not be sufficient to demonstrate the degree of bias that may occur in the estimates due to the misspecification of the exposure model. Because with few data sets sampling fluctuation, rather than bias, may be more pronounced. To get rid of this problem one way is to use infinitely many data sets and construct the empirical densities of the estimates. But this process will be very time consuming. For example, the implementation of an MCMC algorithm for 50 data sets generated in each of the simulation set-ups considered in this study requires several hours. The alternative to this end is to apply large sample asymptotic theory of estimation. The idea is to generate one large sample, instead of repeated samples, and apply the theory of large sample estimation to estimate the outcome model parameters. The large sample estimate can then be compared with the true parameter values to investigate the degree of bias. However, in the presence of
covariate measurement error, the large sample estimation procedure can not be implemented analytically except in the case of linear regression model. Therefore, we need to develop a computational technique to implement the large sample estimation process in measurement error scenario to investigate the bias arising due to exposure model misspecification. We apply the suggested computational method to study the large–sample bias in the estimates of the regression parameters of the logistic outcome model in the presence of exposure measurement error when the exposure model is misspecified. In the new method, we adopt a careful combination of numerical integration techniques (both quadrature and Monte Carlo based) to compute the large–sample bias of Bayes/ML estimators. It can be mentioned that both the Bayesian and the ML estimators hold the same large–sample properties. So, the large–sample bias we compute in this study is applicable to both the estimators.

Implementation of large–sample bias calculation for models without closed–form likelihoods is a research contribution of our investigation. In measurement error scenario, the computation of large sample limits of estimates depends on an estimating equation which does not have closed form solution except in the case of linear regression, and hence a numerical integration technique is adopted. By being thoughtful about how to do numerical integration, and by taking advantage of some closed–form integration, we show that this kind of computation is feasible in measurement error scenarios, and perhaps in related scenarios.

Suppose $\theta$ is the true value of the model parameter $\theta$, and $\ell(\theta)$ is the log–likelihood function. The large sample calculation we suggest is based on the well–known result, $E\{\ell(\hat{\theta})\} = E\{\frac{\partial}{\partial \theta} \ell(\theta)\}|_{\theta = \theta} = 0$, of estimation theory. By the law of large numbers we have, $\frac{1}{n} \sum_{i=1}^{n} \ell_i(\hat{\theta}) \to E\{\ell(\theta)\}$ in probability when $n$ is large. Therefore, for large sample estimation of $\theta$ we need to solve

$$\sum_{i=1}^{n} \ell_i(\hat{\theta}) = 0, \quad (6.1)$$

for a large sample of size $n$, for $\theta$. In (6.1), $\ell_i(\theta)$ is the log–likelihood of the $i$–th simulated data point. Equation (6.1) holds for $\hat{\theta}$ only if all the related model assumptions are true, and in that case the estimator of the parameter $\theta$ obtained by solving (6.1) will be a consistent estimator of $\theta$. Since, we are interested to assess the impact of the exposure model misspecification in the presence of covariate (exposure) measurement error we assume that only the exposure model is misspecified while estimating the model parameters $\theta$ by parametric method. Under misspecified exposure model, let $\theta^*$ solves the estimating equation (6.1) for large $n$, then $(\theta^* - \hat{\theta})$ provides an estimate of the large sample bias while estimating $\theta$. Mathematical details of the suggested computational method are provided in Section 6.2. In Section 6.3, application of the
suggested computational technique to simulated data is described. Finally, Section 6.4 provides an overall discussion on the possible application of the new computational method to assess the large sample bias in the case of outcome model other than logistic model and in some other scenarios where the question of model misspecification may come into play.

6.2 Mathematical Details of the New Computational Technique

Suppose, in the presence of exposure measurement error in the case of a two covariate logistic regression model, we observe the outcome $Y$, surrogate exposure $X^*$, and the accurately measured covariate $Z$. That is, instead of the true exposure $X$ we observe a noisy version, $X^*$, of $X$. Therefore, we have the observed data $(Y, X^*, Z)$, and hence we need to evaluate the observed data likelihood in order to be able to estimate the model parameter $\theta$.

Let $L(\theta; y, x^*, z)$ be the observed data likelihood in the measurement error scenario. In the measurement error terminology we have introduced in Sections 1.2 and 1.5 of Chapter 1, we can write

$$L(\theta; y, x^*, z) = \int f(y|x, z, \theta_R)f(x^*|x)f(x|z, \theta_E)dx,$$

where $\theta = (\theta_R, \theta_E)$. When all the three component models are correctly specified then the estimator solving the estimating equation (6.1) will be a consistent estimator of the true value of parameter $\theta$. Under exposure model misspecification, the estimator $\hat{\theta}^*$ solving the estimating equation (6.1) may not be unbiased. Therefore, magnitude of difference between $\hat{\theta}$ and $\theta^*$ under large sample size can be considered as an estimate of bias that arise because of exposure model misspecification.

Since we are interested in investigating the impact of exposure model misspecification, we need to introduce different symbol to distinguish between true and the assumed distribution. Let us denote the true distribution by the symbol $\hat{f}$, and the assumed distribution by $f$ as we have done so far. For example, we denote the true exposure distribution by $\hat{f}(x|z, \theta_R)$, and the assumed distribution for the exposure by $f(x|z, \theta_E)$. Note that the true outcome model and the true measurement model match the corresponding assumed models as we assume that they are not misspecified. That is, $f(y|x, z, \theta_R) = \hat{f}(y|x, z, \theta_R)$, and $f(x^*|x) = \hat{f}(x^*|x)$.

As already mentioned, for large sample estimate of $\theta$ we need to solve $\sum_{i=1}^{n} \ell_i(\theta) = 0$ for large $n$, where $\ell_i(\theta) = \ln L(\theta; y_i, x_i^*, z_i)$. Under the correct model, if we denote the observed data
likelihood by $f(\theta; y, x^*, z)$, then it can be expressed as a joint distribution of an observed data point $(Y, X^*)$ conditional on $Z$ as

$$f(\theta; y, x^*, z) = f(y, x^*|z) = \int f(y|x, z)f(x^*|x, y, z)f(x|z)dx$$

$$= \int f(y|x, z)f(x^*|x)f(x|z)dx$$

$$= E_{X|Z}\{f(y|x, z)f(x^*|x)\}$$

$$= E_{X|Z}\{g(\theta; y, x, x^*, z)\}, \hspace{1cm} (6.3)$$

where $g(\theta; y, x, x^*, z) = f(y|x, z)f(x^*|x) = f(y, x^*|x, z)$. The expectation $E_{X|Z}$ in (6.3) is with respect to the true exposure model $f(x|z)$. In (6.3), the 3rd equality is derived from the 2nd equality under the assumption of non-differential measurement errors. The evaluation of the expectation on the right hand side of equation (6.3) requires evaluation of an integral which has no closed form in the case of logistic regression. In fact, such integration can not be evaluated analytically except in the case of simple linear regression model. So, numerical integration is required for the evaluation of the expectation in (6.3). Again, to get the large sample estimates of the model parameters through simulation we need a big Monte-Carlo sample of the triplet $(Y, X^*, Z)$ given the true parameter values. As any Monte-Carlo sample is subject to Monte-Carlo simulation error, we attempt to reduce such simulation error while obtaining the large-sample estimates of $\theta$. Such error reduction is achieved by averaging over the outcome variable $Y$ analytically while evaluating the expectation in (6.3). Once the analytic averaging is done over the values of $Y$ we only need a Monte-Carlo sample of the pair $(X^*, Z)$, rather than of the triplet $(Y, X^*, Z)$ and thus reducing the Monte-Carlo simulation error related to $Y$. Since the outcome variable $Y$ is dichotomous in (6.3), such averaging over $Y$, conditional
on $X^*$ and $Z$, is very straightforward and is accomplished as

\[
f(\theta; x^*, z) = E_{Y|X^*,Z}[f(\theta; y, x^*, z)]
\]

\[
= E_{Y|X^*,Z}[E_{X|Z}\{g(\theta; y, x^*, z)\}]
\]

\[
= E_{X|Z}[E_{Y|X^*,Z}\{g(\theta; y, x^*, z)\}]
\]

\[
= E_{X|Z}[E_{Y|X^*,Z}\{f(y, x^* | x, z)\}]
\]

\[
= \int E_{Y|X^*,Z}\{f(y, x^* | x, z)\} f(x | z) dx
\]

\[
= \int E_{Y|X^*,Z}\{f(y, x^* | z)\} dx
\]

\[
= \int f(1, x, x^* | z) Pr.(Y = 1 | x^*, z) dx + \int f(0, x, x^* | z) Pr.(Y = 0 | x^*, z) dx
\]

\[
= Pr.(Y = 1 | x^*, z) \int f(1, x, x^* | z) dx + Pr.(Y = 0 | x^*, z) \int f(0, x, x^* | z) dx
\]

\[
= q(x^*, z) \int f(1, x, x^* | z) dx + [1 - q(x^*, z)] \int f(0, x, x^* | z) dx,
\]

(6.4)

where $q(x^*, z) = Pr.(Y = 1 | x^*, z)$. Again, the expectation $E_{Y|X^*,Z}$ in (6.4) is with respect to the true distribution of the outcome $Y$, given $X^*$ and $Z$, and $E_{X|Z}$ is with respect to the true exposure distribution. The integration in (6.4) over $X$ needs to be evaluated numerically because no closed form solution exists. After averaging over $X$ (by numerical integration) and $Y$ (by analytic integration) we get the function $\tilde{m}(\theta) = E_{X|Z}[E_{Y|X^*,Z}\{g(\theta; y, x^*, z)\}]$ to be utilized to get the estimating equation for $\theta$. According to large sample estimation theory, in order to find the large sample estimate of $\theta$ we need to solve

\[
\sum_{i=1}^{n} \frac{\delta}{\delta \theta} log E_{X|Z}[E_{Y|X^*,Z}\{g(\theta; y_i, x_i, x_i^*, z_i)\}] = 0,
\]

(6.5)

for $\theta$ when $n$ is large. That is, under true models in measurement error scenario, the estimating equation (6.1) turns out to be estimating equation (6.5). Now, for solving the estimating equation (6.5) we need to evaluate the function in (6.4), and its first and second derivatives, which requires a combination of Monte–Carlo integration (to generate sample of the pair $(X^*, Z)$), numerical integration (to integrate out the unobserved exposure $X$) and analytic integration (to average over the outcome variable $Y$) to evaluate.
In (6.4), we need to evaluate $q(x^*, z)$ and $\int f(y, x, x^* | z)dx$ numerically. Now, we can write

$$q(x^*, z) = Pr(Y = 1 | x^*, z)$$
$$= \int Pr(Y = 1, x | x^*, z)dx$$
$$= \int Pr(Y = 1 | x, x^*, z)f(x | x^*, z)dx$$
$$= \int Pr(Y = 1 | x, z)f(x | x^*, z)dx$$
$$= E_{X^* | Z} \{ Pr(Y = 1 | x, z) \}$$
$$= E_{X^* | Z} \left[ \frac{e^{\beta_0 + \beta_1 x + \beta_2 z}}{1 + e^{\beta_0 + \beta_1 x + \beta_2 z}} \right]$$  \hspace{1cm} (6.6)

The expectation in (6.6) does not have any closed form solution and hence needs to be evaluated numerically. For evaluating $q(x^*, z)$ numerically we write $q(x^*, z)$ as

$$q(x^*, z) = Pr(Y = 1 | x^*, z)$$
$$= f(Y = 1 | x^*, z)$$
$$= f(1, x^* | z)$$
$$= \frac{\int f(1 | x, z)f(x^* | x)\tilde{f}(x | z)dx}{\int f(x^* | x)\tilde{f}(x | z)dx}$$
$$= E_{X | Z} \{ f(1 | x, z)f(x^* | x) \} \left[ E_{X | Z} \{ f(x^* | x) \} \right]$$  \hspace{1cm} (6.7)

In (6.7), the expectation is taken with respect to the true exposure distribution given $Z$, and is evaluated at the true parameter values of the outcome, measurement and exposure models because the quantity $q(x^*, z)$ comes into play as a result of averaging over the true distribution of the outcome variable $Y$, given $X^*$ and $Z$. As an aside, taking analytic average over the outcome variable is not essential to implement the suggested computational technique. It is taken to reduce the Monte-Carlo simulation error related to $Y$. However, such reduction in Monte-Carlo error does not come without price. Analytic averaging over $Y$ adds another level of computational complication through the introduction of the quantity $q(x^*, z)$, which requires numerical integration to evaluate. Therefore, though averaging over $Y$ analytically can reduce Monte-Carlo error it may introduce another source of error related to numerical integration required to evaluate $q(x^*, z)$, a by-product introduced due to averaging over $Y$. In fact, error related to the evaluation of $q(x^*, z)$ may be more compared to the Monte-Carlo error, especially when the sample size is large. Furthermore, such analytic averaging is possible only in the case of logistic regression model because the outcome variable is dichotomous. Therefore, one may not choose to taking analytic averaging of $Y$, rather generate a Monte-Carlo sample of it and
carry out the rest of the computation through the combination of Monte-Carlo and numerical integrations. In our investigation process we choose to integrate out \( Y \) analytically simply because it is very straightforward in the case of dichotomous outcome.

Since, to evaluate \( q(x^*, z) \) we need to take integration with respect to the true exposure distribution, we have, according to our simulation set-up

\[
\begin{align*}
    f(1 \mid x, z) &= \frac{e^{\beta_0 + \beta_1 x + \beta_2 z}}{1 + e^{\beta_0 + \beta_1 x + \beta_2 z}} \\
    f(x^* \mid x) &= N(x, \tau^2) \\
    \text{and } \tilde{f}(x \mid z) &= \text{LN}(x; \alpha_0 + \alpha_1 z, \lambda^2)
\end{align*}
\]

as the outcome, measurement and exposure models, respectively. That is, in our simulation studies we generate the true exposure from log-normal distribution. Under the specified outcome, measurement and the true exposure models mentioned above, \( q(x^*, z) \) can be written as

\[
q(x^*, z) = \frac{\int \frac{e^{\beta_0 + \beta_1 x + \beta_2 z}}{1 + e^{\beta_0 + \beta_1 x + \beta_2 z}} \phi(x^*; x, \tau^2) \text{LN}(x; \alpha_0 + \alpha_1 z, \lambda^2) dx}{\int \phi(x^*; x, \tau^2) \text{LN}(x; \alpha_0 + \alpha_1 z, \lambda^2) dx}
\]

\[
= \frac{\int h_1(x, z) \phi(x^*; x, \tau^2) h_2(x, z) dx}{\int h_2(x, z) \phi(x^*; x, \tau^2) dx}
\]

\[
= \frac{\int h_1(x, z) h_2(x, z) \phi(x; x^*, \tau^2) dx}{\int h_2(x, z) \phi(x; x^*, \tau^2) dx}
\]

\[
= \frac{E_X [h_1(x, z) h_2(x, z)]}{E_X [h_2(x, z)]} \tag{6.8}
\]

where \( h_1(x, z) = \frac{e^{\beta_0 + \beta_1 x + \beta_2 z}}{1 + e^{\beta_0 + \beta_1 x + \beta_2 z}} \), \( h_2(x, z) = f(x \mid z) = \text{LN}(x; \alpha_0 + \alpha_1 z, \lambda^2) \) and \( \phi \) represents the density function of a normal random variate. In equation (6.8), the expectation is taken with respect to the distribution of \( X \) given \( X^* \) and \( \tau^2 \). Therefore, we can use this fact to define the \( X \)-grids for numerically evaluating the expectation in (6.8). Based on the distributional structure of \( X \), on which the expectation in (6.8) is defined, the \( X \)-grids are obtained as \( x = x^* + z_q \tau \), where \( z_q \) is the quantiles of standard normal distribution. Here, the quantiles \( z_q \) are obtained as \( z^{\frac{1}{m+1}}, z^{\frac{2}{m+1}}, \ldots, z^{\frac{m}{m+1}} \), where \( m \) is the number of grids taken for a particular \( X \)-value. That is, the \( X \)-grids are defined over the resulting distribution \( \phi(x; x^*, \tau^2) \) of \( X \) based on surrogate exposure \( X^* \). Note that \( \phi(x; x^*, \tau^2) \) is the most curved term among the terms involved in the integral (6.8), a condition required for grid–based integral to work properly. Once the \( X \)-grids are attained, the expectations in the numerator and in the denominator of (6.8) can easily be computed. Suppose, for each \( X \) value we have \( m \) grids, \( x^{(1)}, x^{(2)}, \ldots, x^{(m)} \), then \( E_X [h_1(x, z) h_2(x, z)] = \frac{1}{m} \sum_{j=1}^{m} h_1(x^{(j)}, z) h_2(x^{(j)}, z) \) and \( E_X [h_2(x, z)] = \frac{1}{m} \sum_{j=1}^{m} h_2(x^{(j)}, z) \).
Finally, in order to evaluate the function in (6.4), and its first and second derivatives we also require to evaluate the integral \( \int f(1, x, x^* | z)dx \). While evaluating the integral \( \int f(1, x, x^* | z)dx \) we need to misspecify the exposure distribution as normal distribution, because we want to investigate the performance of the assumed normal exposure model. Under misspecified exposure model this integral can be written as

\[
\begin{align*}
  f(y, x^* | z) &= \int f(y, x, x^* | z)dx \\
  &= \int f(y | x, x^*, z) f(x, x^* | z)dx \\
  &= \int f(y | x, z) f(x^* | x, z) f(x | z)dx \\
  &= \int f(y | x, z) f(x^* | x) f(x | z)dx,
\end{align*}
\]

(6.9)

where the \( f(x | z) \) is the assumed misspecified exposure distribution. Now, under misspecified exposure model we denote the integral in (6.4) by \( m(\theta) = f(\theta; x^*, z) \) to distinguish it from \( \hat{m}(\theta) \), which is obtained under true exposure model. In (6.9), we misspecify the true exposure distribution as normal distribution. We misspecify the exposure distribution as normal distribution mainly because the normal distribution is the most widely used distribution in the measurement error literature to model the unobserved continuous exposure. Furthermore, Gustafson (2004) shows that the use of normal exposure model provides unbiased and consistent estimates of the outcome model parameters in linear regression model in the presence of exposure measurement error irrespective of the true exposure distribution. In statistics literature, it is believed that phenomena that hold in the case of linear regression model also hold, at least approximately, for the other members of the GLM family.

Now, for the purpose of estimation, we misspecify the exposure distribution as normal distribution. According to our simulation studies, the measurement error distribution is also normal. Therefore, we can take the advantage of the structural similarity of the measurement model and the misspecified exposure model to form one normal model for \( X \) and base our \( X \)-grids on the quantiles of the resulting \( X \)-distribution to evaluate \( \int f(y, x^*, x | z)dx \). Mathematically,
under misspecified exposure model, we have

\[
\int f(y, x, x^* | z) dx = \int f(y | x, z) f(x^* | x) f(x | z) dx \\
= \int h(x, z) \phi(x^*; x, \tau^2) \phi(x; \alpha_0 + \alpha_1 z, \lambda^2) dx \\
= \int h(x, z) \phi(x; x^*, \tau^2) \phi(x; \alpha_0 + \alpha_1 z, \lambda^2) dx \\
= \int h(x, z) k(x^*, z) \phi(x; \mu_0, \sigma_0^2) dx \\
= k(x^*, z) \mathbb{E}_X [h(x, z)],
\]

(6.10)

where \( h(x, z) = f(y | x, z) = \frac{e^{\rho_0 + \beta_1 x + \beta_2 z}}{1 + e^{\rho_0 + \beta_1 x + \beta_2 z}}, \mu_0 = \lambda^2 x^* + \tau^2 \alpha_0 + \tau^2 \alpha_1 z, \sigma_0^2 = \frac{\tau^2 \lambda^2}{\lambda^2 + \tau^2}, \) and \( k(x^*, z) \) is a normalizing constant independent of \( x \). The expectation on the right hand side of (6.10) is with respect to the resulting distribution of \( X \), which is normal with mean \( \mu_0 \) and variance \( \sigma_0^2 \). Therefore, we take the \( X \)-grids to be the quantiles of \( \phi(x; \mu_0, \sigma_0^2) \) distribution, which are obtained as \( x = \mu_0 + q_z \sigma_0 \). The \( X \)-grids defined here are adaptive in the sense that at each iteration of the maximization algorithm these \( X \)-grids are updated as the quantities \( \lambda^2, \alpha_0 \) and \( \alpha_1 \) in \( \mu_0 \) and \( \sigma_0^2 \) are updated as the iteration progresses. Once the quantity \( q(x^*, z) \) and the integration in (6.10) are evaluated, the function in (6.4) can easily be evaluated by plugging in the values of \( q(x^*, z) \) and of the integral in (6.10) in the right hand side of (6.4). For the evaluation of the first and second derivatives of the function \( f(\theta; x^*, z) \) same \( X \)-grids are appropriate as the first and second derivatives of \( f(\theta; x^*, z) \) are some multiples of it. To solve the estimating equation (6.5) for \( \theta \) under misspecified exposure model we need to evaluate \( U(\theta) = \log \{ m(\theta) \} = \log f(\theta; x^*, z) \), and first and second derivatives, \( U'(\theta) \) and \( U''(\theta) \), respectively, of it for all \( X \) values. To accomplish these we have

\[
U'(\theta) = \frac{f'(\theta; x^*, z)}{f(\theta; x^*, z)} \\
and \quad U''(\theta) = \frac{f'(\theta; x^*, z) f''(\theta; x^*, z) - [f'(\theta; x^*, z)]^2}{[f(\theta; x^*, z)]^2},
\]

where \( f'(\theta; x^*, z) \) and \( f''(\theta; x^*, z) \) are the first and second derivatives of \( f(\theta; x^*, z) \) with respect to \( \theta \). Once \( q(x^*, z), f(\theta; x^*, z), U'(\theta) \) and \( U''(\theta) \) are evaluated through numerical integration described so far we can solve the estimating equation (6.5) for \( \theta \) by Newton–Raphson iteration method. The iterative method to obtain the solution of the estimating equation (6.5) proceeds as follows:

(a) given the values of \( \tau^2 \) and \( x^* \), define the \( X \)-grids for \( q(x^*, z) \) as \( x = x^* + q_z \tau \) and evaluate \( q(x^*, z) \)
(b) given the values of $\lambda^2$, $\tau^2$, $\alpha_0$, $\alpha_1$ and $Z$, evaluate $\mu_0$ and $\sigma_0^2$, compute the $X$-grids as $x = \mu_0 + q_2\sigma_0$ and evaluate $f(\theta; x^*, z)$, $U'(\theta)$ and $U''(\theta)$

(c) given the values of $q(x^*, z)$ in step (a) and, of $f(\theta; x^*, z)$, $U'(\theta)$ and $U''(\theta)$ in step (b), update the parameter estimates via Newton–Raphson algorithm as $\theta^{(i+1)} = \theta^{(i)} - \left[U''(\theta^{(i)})\right]^{-1} U'(\theta^{(i)}),$ where $\theta^{(i)}$ is the value of the parameter $\theta$ at the $i$-th iteration.

(d) the method is continued until convergence is achieved.

6.3 Application of the New Computational Technique: Simulation Studies

To provide a comparative picture of the new computational method with that based on the repeated samples we consider exactly the same simulation designs that are used under repeated sample technique in Section 2.2.1 of Chapter 2. For studying the large sample bias by the suggested computational method we use a sample of size 5000 under each of the simulation settings considered. As already mentioned earlier, like in the case repeated samples, we use normal distribution to model the true but unobserved exposure in the estimation process though the true exposure is generated from log-normal distributions. In the Newton–Raphson algorithm, the naive estimates of $\beta$ are used as initial guess for the regression parameters $\beta$. The naive estimates are the ML estimates obtained by fitting the logit model of $Y$ on $X^*$ and $Z$. For the location parameters, $\alpha$, of the exposure distribution the least-square estimate of $\alpha$ obtained by regressing $X^*$ on the $Z$ are used as initial guess. Finally, the residual variance of the linear regression fit of $X^*$ on $Z$ is taken to be the initial guess for the scale parameter, $\lambda^2$, of the exposure distribution. With these initial guess, the Newton–Raphson algorithm for the new technique is found to converge in 5 to 10 iterations under different simulation settings considered in this study.

6.3.1 Results

Results of the simulation studies are presented in Table 6.1. We report the estimates of the regression parameters obtained under repeated sample case and those obtained by the suggested computational technique. Along with the parameter estimates obtained by both the techniques we also report the bench-mark estimates to provide an idea of the magnitude of bias that
may arise due to exposure model misspecification. Note that, in the case of repeated samples, by benchmark estimate we mean the average of the 50 estimates obtained by fitting logistic regression of $Y$ on the true exposure $X$ and the precisely measured covariate $Z$ to 50 simulated data sets under each simulation setting. On the other hand, in the case of large sample, the benchmark estimates are estimates obtained by fitting logistic regression model of $Y$ on the true exposure $X$ and the precisely measured covariate $Z$ to a large data set. The measurement error adjusted estimates under repeated samples are computed by averaging the estimates obtained from 50 simulated data sets (50 samples), each of size $n = 500$. In this case, the parametric method based on conditional independence assumption (described in Chapter 1) for adjusting for covariate measurement error is adopted for parameter estimation. For estimation purpose, Bayesian method of inference is used, and is implemented through appropriate MCMC algorithm. Details of the estimation and computational technique for the repeated sample case are provided in Chapters 1 and 2. For large sample assessment via the new computational technique, again a sample of size $n = 5000$ is considered. As an aside, though both the repeated sample and the large sample methods provide the idea about the magnitude of bias arising due to exposure model misspecification the two techniques are slightly different conceptually. In repeated sample method, we use a $50 \times 500$ set-up, i.e. generate 50 data sets each of size $n = 500$, to approximate $E(\hat{\beta})$. The magnitude of bias is then given by $\left(E(\hat{\beta}) - \beta\right)$. On the other hand, in the large sample case, we use a $1 \times 5000$ set-up, i.e. generate one sample of size $n = 5000$, to approximate $\lim_{n\to\infty}\hat{\beta} = \lim_{n\to\infty}E(\hat{\beta})$. That is, we use a large sample to approximate the large sample limit of $\hat{\beta}$, which in turns gives the large sample limit of $E(\hat{\beta})$. In this case, the magnitude of bias is thus $\left(\lim_{n\to\infty}E(\hat{\beta}) - \beta\right)$.

For each of the simulation set-ups considered in this study, the large-sample method produces very similar results as are produced in the case of repeated samples. In each case, the mean of the estimates obtained in the repeated sample scenario is very close to the corresponding large-sample estimates. Therefore, the simulation studies carried out in the case of two-covariate logistic regression with measurement error in one covariate (first covariate) demonstrate that the suggested computational technique to assess the large sample impact of exposure model misspecification works very well.

Finally, simulation studies in both the repeated sample and the large-sample cases provide quite convincing evidence that the estimates of the regression parameters in the outcome models are sensitive to the choice of the exposure distribution in some cases, especially when the true exposure distribution has thicker tail and the true but unobserved exposure is strongly correlated with the error-free covariate.
Table 6.1: Sensitivity to exposure model misspecification: comparison of repeated sample results with those obtained by the suggested large sample method

<table>
<thead>
<tr>
<th></th>
<th>Measurement error adjusted Estimates</th>
<th>Bench-mark Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Repeated Samples</td>
<td>Large Sample</td>
</tr>
<tr>
<td></td>
<td>$\beta_0$</td>
<td>$\beta_1$</td>
</tr>
<tr>
<td>1</td>
<td>0.2</td>
<td>-0.2</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
<td>-0.2</td>
</tr>
<tr>
<td>3</td>
<td>0.6</td>
<td>-0.2</td>
</tr>
<tr>
<td>4</td>
<td>0.4</td>
<td>-0.2</td>
</tr>
<tr>
<td>5</td>
<td>0.6</td>
<td>-0.2</td>
</tr>
</tbody>
</table>

6.4 Discussion

In this study we investigate the performance of the suggested computational technique in the case of logistic regression model only. The method can well be adapted to the other members of the GLM family, e.g. Poisson regression model, to investigate the large sample bias arising due to exposure model misspecification in the estimates of the outcome model parameters in the presence of covariate measurement error. The suggested computational methodology can also be extended to study the large sample bias in many other misspecified hierarchical model scenarios. Such examples are misspecified frailty distributions in frailty models, misspecified random effects distributions in random effects models, etc. For example, in random effects model a standard assumption is that the random effects are normally distributed. But in many instances normality assumption for the random effects may be unrealistic, e.g., a common features in biological contexts is the skewness of underlying individual profiles (Ma, Genton Davidian, 2004). In such instances, it is necessary to study the extent to which the misspecified normal distribution for the random effects bias the estimates of the fixed effect regression parameters of interest. Such type of investigation is necessary to decide whether relaxation of normality assumption for the random effects is required.

Though one may indulge into the argument that the use of repeated samples can serve the same purpose then why this new rather complicated computational technique is required. The answer to this question can be provided from different perspectives. First, in the case of
repeated samples, to provide large sample justification to the results one needs to generate quite large number of samples, which may be very time consuming. Therefore, it may be a better alternative to use one large sample instead of using large number of smaller samples. Second, with small or moderate number of samples the empirical distribution of the estimates may be a reflection of sampling fluctuation rather than that of bias. Third, using two parallel computational methods provides more credence to the results if the results from the two methods are close to each other.
Chapter 7

Discussion and Scope of Future Research

7.1 Discussion

The work of this dissertation focuses on issues related to the problems of covariate measurement errors while modeling the relationship between a response variable and covariates of interest. Especially, we concentrate on eliminating the measurement error bias from the estimated regression parameters through a flexible parametric approach. We investigate the applicability of the proposed method with reference to logistic regression model. However, being a structural approach, the proposed method can easily be extended or adapted for the other member of GLM, and for other nonlinear regression models. The motivation for searching for a flexible parametric approach for adjusting the measurement error bias stems from the unavailability of widely applicable method in eliminating the measurement error bias. as already discussed in the earlier chapters, the existing methods in this regard can not offer wide range of flexibility either in terms of accommodating complicated epidemiological design set-up, efficiency of the estimates, or incorporating categorical exposure with misclassification. By adopting a flexible parametric approach we demonstrate that the sensitivity of the estimates to the exposure model misspecification can be eliminated, a virtue that is considered to be typical to non-parametric approaches. Extensive simulation studies we have conducted demonstrate that the proposed flexible parametric approach achieves the virtue of robustness to exposure model misspecification while estimating the underlying relationship between the outcome and the explanatory variables in the presence of measurement error. In addition to that, being parametric, the proposed method can offer efficiency gain and wider range of flexibility. In order to implement estimation and inference through the proposed method we adopt the Bayesian method
of inference. The advantages of Bayesian method over the competing methods in measurement error scenario are discussed in Chapter 1. In Chapter 1, we also discuss various issues related to the covariate measurement error and its impacts on the estimates of the response-covariate relationship, and review the related literature, including some mathematical details.

In Chapter 2, we investigate the sensitivity of the estimates of the regression parameters of the logistic outcome model to exposure model misspecification. Simulation studies demonstrate that presence of high skewness and/or heavy-tailedness in the distribution of the true exposure is the main cause of concern in measurement error scenario. In such cases, the misspecified normal exposure model can not eliminate the measurement error bias adequately in the case of logistic outcome model.

In Chapter 3, we discuss the potential of considering flexible parametric model for the unobserved exposure distribution. Consequently, we consider the applicability of skew-normal distribution through simulation studies. Simulation studies we conduct in this chapter show that though the SN exposure model can perform better than the normal exposure model its performance is also not adequate in eliminating the measurement error bias, especially in the cases with very skewed or heavy-tailed exposure distribution.

In Chapter 4, we continue searching for more flexible exposure models. In this line, we explore the applicability of FGSN and FGST distributions as flexible exposure models in correcting the measurement error bias. We find that the FGST exposure model performs adequately in the all the simulation settings considered. That is, simulation studies reveal that the FGST exposure model has enough flexibility to provide adequate degree of robustness to exposure model misspecification while estimating response-covariate relationship in the presence of measurement error.

In Chapter 5, we investigate the comparative bias correction performance of the proposed flexible parametric method with the other flexible parametric and the popular non-parametric approaches. Simulation studies reveal that the proposed method with FGST exposure model performs reasonably well irrespective of the shape of the true exposure distribution, while the competing approaches perform well in some situations but perform badly in some other situations.

Finally, in Chapter 6, we develop a new computational technique to approximate the large sample bias arising due to exposure model misspecification in the presence of covariate measurement error. We apply the new computational methodology to the simulated data to investigate its
applicability in approximating the large sample bias by comparing the results obtained by the suggested computational technique with those obtained by repeated sample technique. Simulation studies show that the suggested computational technique performs similarly compared to the repeated sample technique.

In short, on the basis of the results of the extensive simulation studies we have conducted throughout this dissertation we can conclude that the proposed flexible parametric method with FGST exposure model can work well in eliminating measurement error bias adequately. It can also provide better efficiency of the estimated response-covariate relationship in comparison to the competing methods. Coupled with Bayes-MCMC computational technique the proposed method can be implemented fairly easily. The proposed method is especially well suited to situations where the density of the unobserved exposure is highly skewed and/or has heavy tails, a typical situation in many epidemiological studies. In such situations the competing methods are found to perform very inadequately in correcting the measurement error bias. In the other well-behaved scenarios the proposed method is found to perform equally well both in terms of bias correction and the efficiency of the estimated response-covariate relationship.

7.2 Future Research Work

Following is the list of potential future research work.

1. Since the proposed flexible parametric approach is found to perform adequately in correcting measurement error bias in the case logistic regression model, we can think of extending or adapting it to

   (i) the other members of GLM, and to the other non-linear models, e.g. Cox regression model

   (ii) longitudinal studies with covariate measurement error

   (iii) though throughout the thesis X and Z are considered as univariate, the proposed method can be extended to the case where X or Z (or both) are vectors. In that case, we need to use multivariate FGST distribution as exposure model. It is demonstrated in Ma, Genton and Davidian (2004) in the case of linear mixed effects models that the implementation of Bayes MCMC algorithm with multivariate version of FGST distribution can be carried out without much difficulty
2. In this dissertation, we investigate the applicability of the proposed method under certain assumptions, e.g., non-differential measurement error, independence among the replicated surrogate measurements etc. By being parametric, the proposed method has the flexibility to be extended to

(i) the case of differential measurement error

(ii) accommodate correlated replicates of the surrogate measurements

3. For any model, diagnostics are helpful to indicate whether there is evidence of problem for a specific model. In the presence of measurement errors the true exposure is not observed and hence it is not clear whether there are meaningful residuals. In such situation some sort of diagnostics procedure may of great help in identifying appropriate exposure model.

4. The computational approach that we develop in Chapter 6 for studying the large sample bias of the misspecified exposure model in measurement error scenario might be extended to study the large sample bias in many other misspecified hierarchical model scenarios. Such examples are the misspecified frailty distribution in frailty model, misspecified random effects distribution in random effects models etc.
Bibliography


generalized skew–elliptical random effects. In Ma, Y., editors, *Skew–elliptical distributions


studies with errors in variables. *Biometrika*, 84: 523-537.


University Press, London.


[33] Richardson, S. and Leblond, L. (1997). Some comments on misspecification of priors in

[34] Richardson, S., Leblond, L., Jaussent, I. and Green, P.J. (2002). Mixture models in mea-
surement error problems with reference to epidemiological studies. *Journal of the Royal
Statistical Society A*, 165: 549-566.

to case–control studies with errors in covariates. *Journal of the American Statistical

relative risk estimates and confidence intervals for systematic within-person measurement


