BAYESIAN APPROACH TO ERRORS-IN-VARIABLES IN COUNT DATA REGRESSION MODELS

NUR AAINAA ROZLIMAN

FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

BAYESIAN APPROACH TO ERRORS-IN-VARIABLES IN COUNT DATA REGRESSION MODELS

NUR AAINAA ROZLIMAN

DISSERTATION SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

INSTITUTE OF MATHEMATICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Nur Aainaa bt Rozliman

Matric No: SGP150006

Name of Degree: Master of Science

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"): Bayesian Approach to Errors-in-Variables in Count Data Regression Models

Field of Study: Statistics

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:
Designation:

BAYESIAN APPROACH TO ERRORS-IN-VARIABLES IN COUNT DATA REGRESSION MODELS

ABSTRACT

In most practical applications, data sets are often contaminated with error or mismeasured covariates. When these errors-in-variables or measurement errors are not corrected, they will cause misleading statistical inferences and analysis. Therefore, we will focus on addressing errors-in-variables problems in count data regression models, specifically Poisson regression and negative binomial regression models. To remain useful in realistic situations, we utilize the Bayesian approach where the variance is estimated instead of assumed as known. We relax the distributional assumption of the exposure model by intentionally misspecifying the model with a flexible distribution. Following this, we shall also compare the performance between two different flexible distributions in modelling the exposure, namely the flexible generalized skew-normal distribution and flexible skew-generalized normal distribution. We also conduct simulation studies on synthetic data sets using Markov Chain Monte Carlo simulation techniques to investigate the performance of the flexible Bayesian approach. The results of our findings show that the flexible Bayesian approach is able to estimate the values of the true regression parameters consistently and accurately with a significant bias reduction.

Keywords: Count data regression, errors-in-variables, Bayesian, Markov chain Monte Carlo.

PENDEKATAN BAYESAN DALAM MODEL RALAT-DALAM-PEMBOLEHUBAH DALAM MODEL REGRESI DATA BILANG

ABSTRAK

Dalam kebanyakan aplikasi praktikal, set data sering terkontaminasi dengan ralat atau kesilapan sukatan pada kovariat. Apabila ralat-dalam-pembolehubah atau ralat sukatan tidak diperbetulkan, mereka akan menyebabkan kesimpulan dan analisis statistik yang mengelirukan. Oleh itu, kami akan memberi tumpuan dalam menangani masalah ralat-dalam-pembolehubah dalam model regresi data bilang, khususnya regresi Poisson dan model regresi binomial negatif. Untuk terus berguna dalam situasi yang realistik, kami menggunakan pendekatan Bayesan di mana varians dianggarkan dan bukannya dianggap sebagai tercerap. Kami melonggarkan andaian taburan model tak bersandar dengan menggantikannya dengan model fleksibel yang salah secara sengaja. Berikutan ini, kami juga membandingkan prestasi dua taburan fleksibel yang berbeza dalam memodelkan pembolehubah tak bersandar, iaitu taburan pencong-normal teritlak yang fleksible dan taburan normal pencong-teritlak yang fleksibel. Kami juga menjalankan kajian simulasi pada set data sintetik menggunakan teknik simulasi rantai Markov Monte Carlo untuk menyiasat prestasi pendekatan Bayesan yang fleksibel. Hasil penemuan kami menunjukkan bahawa pendekatan Bayesan yang fleksibel dapat menganggarkan nilai-nilai parameter regresi sebenar secara konsisten dan jitu dengan pengurangan pincang yang signifikan.

Kata Kunci: Regresi data bilang, ralat-dalam-pembolehubah, Bayesan, rantai Markov Monte Carlo.

ACKNOWLEDGEMENTS

Alhamdulillah to the Most Merciful for His countless gifts and to Whom I owe it all.

I am eternally grateful to my parents, Ayah and Mama for their endless support and lending me their strength to complete this thesis. Without them, I would not have been able to endure the trials and tribulations faced during my research work. My heartfelt appreciation for my grandmother; always keen to know what I was doing and how I was proceeding, although it is most likely that most of my explanations are lost in translation. Their tremendous love and prayers are things that I will forever be thankful about.

It is with great pleasure to express my gratitude to Dr. Adriana Irawati Nur bt Ibrahim for her unwavering support and guidance throughout my Masters degree. It is an honour to work under her supervision. The same goes to my second supervisor, Dr Rossita bt Mohamad Yunus.

Special thanks to Yayasan Khazanah for funding my studies, with special mention to Mdm. Intan, Ms. Hidayah and Mr. Kamarul Bahrain.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

R	:	ratio of measurement error variance to the variance of true exposure variable.
X	:	unobserved exposure variable.
X^*	:	observed surrogate variable.
Y	:	outcome variable.
$\boldsymbol{\beta}$	:	regression parameter vector.
$\boldsymbol{\theta}$	:	parameter vector.
$\boldsymbol{\theta}_{NBRM}$	:	parameter vector of negative binomial regression model.
$\boldsymbol{\theta}_{PRM}$	:	parameter vector of Poisson regression model.
ϵ	:	errors-in-variables.
$\pi(\boldsymbol{\theta})$	:	prior distribution.
τ^2	:	variance of errors-in-variables.
r	:	dispersion parameter.
EIV	:	errors-in-variables.
EIVM	:	errors-in-variables model.
FGSE	:	flexible generalized skew-elliptical.
FGSN	:	flexible generalized skew-normal.
FGST	:	flexible generalized skew-*t*.
FSGN	:	flexible skew-generalized normal.
GLM	:	generalized linear model.
IG	:	inverse-Gamma.
MCMC	:	Markov chain Monte Carlo.
ME	:	measurement error.
MH	:	Metropolis-Hastings.
MSE	:	mean squared error.
NBRM	:	negative binomial regression model.
pdf	:	probability density function.
pmf	:	probability mass function.
PRM	:	Poisson regression model.
RCAL	:	regression calibration.
RWMH	:	random walk Metropolis-Hastings.
SIMEX	:	simulation extrapolation.
SN	:	skew-normal.
SQS	:	structural quasi score.
ST	:	skew-*t*.

CHAPTER 1: INTRODUCTION

1.1 Background of Study

Count data consist of non-negative integers that have many applications in various fields of studies. Poisson regression model (PRM) is mostly used to model for this type of data. However, PRM requires count data to have the property of equal mean and variance. This property is referred to as equidispersion. Although some count data could fulfill this property, realistically overdispersion may occur. So, as to model for count data with overdispersion, negative binomial regression model (NBRM) is another model that is regularly employed to model for overdispersed count data. In addition to this, the covariates of these count data regressions are usually riddled with error. When the independent variables of these count data models are contaminated with error, we use the term errors-in-variables (EIVs) to describe it. EIV occurs when instead of observing the true values of the independent variables, their incorrect proxy values which has EIV are instead observed and taken as true. There are various reasons on why EIV emerges (e.g. human blunder, machine error, expensive or impossible to measure exposure variables directly). When EIV is ignored or not addressed, there will be serious drawbacks, especially when estimating the parameters in a model that has this type of error contamination. By not addressing EIVs, researchers may reach the wrong statistical conclusions as parameter that is estimated in a non-corrected model will be biased.

To date, there is a significant amount of literature on methods to solve EIV problems. Whilst most research has been carried out on EIV for other types of regression (i.e., logistic regression), only a few have investigated EIV issues around count data regression models, which shall be discussed in detail in Chapter 2. Approaches on handling EIV models can be widely classified into two conceptual frameworks; Bayesian and frequentist (non-Bayesian) approaches. Corrected score (Stefanski, 1989; Nakamura, 1990), structural quasi score (Carroll et al., 2006; Thamerus, 1998) and conditional score (Stefanski & Carroll, 1987) are examples of non-Bayesian methods. As for Bayesian approach to EIV problems, it was introduced by Richardson and Gilks (1993) in the context of epidemiology study. Dellaportas and Stephens (1995) and Mallick and Gelfand (1996) analysed EIV models in the fully Bayesian framework for nonlinear regression models and generalized linear models, respectively.

In the Bayesian paradigm, there will be a need to specify the distribution of the independent variables, but since in EIV model the observed independent variables are incorrect, then the specification of the distribution might lead to misspecification bias (Richardson et al., 2002). Following this, most researchers explore the usage of functional approaches where there is no specification of model; nevertheless in comparison to Bayesian approaches, the former may lead to a loss in efficiency (Hossain & Gustafson, 2009). To reduce distributional assumptions, researchers in the Bayesian paradigm consider flexible models where the exposure model is intentionally misspecified with a flexible model. Carroll et al. (1999) demonstrated the use of mixtures of normals as flexible exposure model for linear EIV models. Later, Richardson et al. (2002) extended the use of mixtures of normals as misspecified exposure model to EIV logistic regression. However, in these studies, they reported that the performance of the mixtures of normals model deteriorated when the true exposure distribution is skewed and/or heavy-tailed. Huang et al. (2006) implemented a second-order nonparametric density but they did not investigate its robustness for exposure distribution with skewness and heavy-tailedness. Hossain and Gustafson (2009) utilized flexible generalized skew-normal (FGSN) and flexible generalized skew-*t* (FGST) as misspecified exposure distribution. They investigated the robustness of both FGSN and FGST to model exposure distribution that exhibits different levels of skewness and heavy-tailedness.

1.2 Problem Statement

It is imperative to stress that the vast majority of investigations carried in Bayesian EIV models focused on other types of regression models such as logistic regression and probit regression; much less attention is given to correcting EIV in PRM and NBRM despite their importance in modeling for count data. This is especially true for PRM in the Bayesian paradigm and even more so for NBRM in general. To the best of our knowledge, researches that were done on the subject of fixing EIV in NBRM are by El-Basyouny and Sayed (2010) and Yang et al. (2013) where both papers addressed EIV in NBRM using Bayesian approach and applied it to safety performance analysis. Nevertheless, they assumed the true exposure distribution as known such that it follows either normal or log-normal distributions. Thus, any departures from normality and log-normality may lead to extra bias caused by exposure model misspecification.

Throughout the years, most EIV correction studies in count data regression models have focused on the use of classical methods (non-Bayesian methods). However, non-Bayesian methods faced problems such as inconsistent roots especially when the distribution of EIV is non-normal. Furthermore, some of these methods also show pathological behaviours and when the contamination level of EIV is high, multiple roots, estimate-finding failure, as well as skewness, are also found. In addition to this, non-Bayesian methods are unrealistic in general practices since in these methods, they often assume the distribution of the variance of EIV as known.

In our research, we propose the use of flexible Bayesian approach which is the Bayesian approach with flexible independent variables distribution. This type of approach could offer compensations on the shortcomings of the non-Bayesian approach in solving EIV problems mentioned in the previous paragraph as in the Bayesian paradigm, one does not deal with estimating functions which therefore will not lead to any roots problem. In this study, the flexible Bayesian approach is introduced to count data regression models with EIV, particularly the PRM and NBRM.

1.3 Objective of Research

The main objectives of this research are

1. To implement the Bayesian framework to EIV in count data regression models, particularly the Poisson regression and negative binomial regression models.

2. To introduce the flexible parametric approach to account for different types of true unobserved exposure distributions for the count data regression models with EIV and compare the performance of two flexible distributions, i.e., flexible generalized skew-normal (FGSN) and flexible skew generalized-normal (FSGN) as an intentionally misspecified distribution of the unobserved independent variables distribution.

3. To apply the Markov chain Monte Carlo sampling methods when estimating the regression parameters of these EIV count data regression models while reducing bias in parameter estimations caused by EIV.

4. To investigate the performance of the flexible Bayesian approach using simulation studies.

1.4 Significance of Research

The significance and benefits of this research are

1. When most studies have been focused on using frequentist methods in the context of count data models, we employ the Bayesian approach to correct bias due to EIV in parameter estimations for count data models that have better efficiency according to Hossain and Gustafson (2009).

2. True exposure distribution is considered as unknown unlike existing researches in EIV correction of PRM and NBRM.

3. We adapt the flexible parametric approach such that the exposure model is misspecified with a flexible distribution, hence our approach remains robust against any departures from normality in its true underlying exposure distribution.

4. Current non-Bayesian approaches to correcting EIV assume the variance of EIV as known, but in this thesis, since the Bayesian approach is used, we spare the assumption that the EIV variance is known and instead it is estimated aided with validation data in order to achieve model identifiability.

1.5 Outline of Research

Our research applies the Bayesian method with an intentionally misspecified flexible exposure distribution to correct EIV in count data regression models, namely the PRM and NBRM. The outline of our research is as follows,

Chapter 2 of this thesis contains the literature review of this study where any existing academic literature that is significantly related to our study is discussed. In the first part, we discuss the development of count data regression models and their usage. Following this, we examine all significant literature on EIVMs in any regression models. This chapter also contains the different techniques used in correcting EIVs which is separated into two; Non-Bayesian methods and Bayesian methods. Next, we also discuss on the basic understanding of the Bayesian paradigm and a brief review of the Markov chain Monte Carlo (MCMC) algorithm.

Chapter 3 presents the framework in which the Bayesian approach that is utilized to address EIV in regression models. The formulation of the posterior distribution in the presence of EIV is also presented here. This is followed by a discussion on the impact of misspecification of outcome and exposure models and how the implementation of an intentionally misspecified flexible model can mitigate misspecification bias. We also provide a brief introduction to the flexible models considered in our research.

Chapter 4 contains our implementation of the Bayesian approach to EIV in PRM. We modify current flexible Bayesian approach in correcting EIVs to Poisson regression. This chapter is separated into two main parts, that is when flexible generalized skew-normal (FGSN) is used and when flexible skewed generalized-normal (FSGN) is used. The prior distributions, posterior distributions and conditional posterior densities of all the parameters in question are given in this chapter as well as the MCMC that is implemented. The results of the simulation studies done for PRM outcome model are also given, the first part of the results are when the error is normal and second part of results is when the error is non-normal.

Chapter 5 focuses on our usage of Bayesian approach to EIV for NBRM. Similarly, this chapter is made up of two parts; the first part is when FGSN is considered as the intentionally misspecified exposure model and the next part is when FSGN is considered as the intentionally misspecified exposure model. The prior distributions, posterior distributions and conditional posterior densities of all the parameters in question are given in this chapter as well as the MCMC that is implemented. The results of the simulation studies are also presented here. The results are also separated into two parts, that is when the distribution of error is normal and when the distribution of error is non-normal.

Chapter 6 discusses the overall results of the simulation studies conducted and explains the main findings of our research.

Chapter 7 provides the concluding remarks as well as suggestions on extending the studies done in this thesis.

CHAPTER 2: LITERATURE REVIEW

2.1 Count Data Regression Models

Data with non-negative discrete count outcomes, denoted by Y, are usually referred to as count data. Count data can be found in most if not all industries and fields of research, which is why in this dissertation we shall focus on regressions that can be used to model count data. To illustrate their wide implementation, we give examples of count data usages found in literature. Schwalbach and Zimmermann (1991) used a data set on the number of patents of German companies registered at the German patent office in 1982, then Dionne et al. (1997) studied the frequency of airline accidents by a carrier in Canada on a quarterly basis between 1974 and 1988. Kawanishi and Sunquist (2004) used photographic capture data in Taman Negara National Park, Malaysia to provide a reliable density estimate of tigers across 600-km^2 study sites. Much recently, Ahmed et al. (2014) studied number of traffic accidents occurrence and their causes. These examples are only a small fraction of count data implementations in literature. To handle count data, there are various statistical models that can be employed corresponding to the properties of the count data studied. This is further discussed in the coming subsections. In our study, we shall focus more on Poisson regression model (PRM) and negative binomial regression model (NBRM).

2.1.1 Poisson Regression Model

Generally, PRM is the most popular regression employed in modelling count data as its main advantage is that it clearly recognizes non-negative integers as independent variables. Poisson distribution originated from the work by Simeon Poisson (Poisson, 1837). Using Poisson distribution as basis, the PRM is developed where explanatory variables X_i are explicitly taken into account in its vital component, that is the mean parameter. Unlike Poisson distribution, where its mean parameter is a non-negative constant, the PRM specifies its mean parameter, μ_i, as a function such that, $\mu_i = \exp(\beta_0 + \beta_1 X_i)$ for

$i = 1, 2, \ldots, n$; or in simple vector form, $\mu_i = \exp(\boldsymbol{X}_i'\boldsymbol{\beta})$ where $\boldsymbol{\beta}$ is the regression parameters vector and $\boldsymbol{X}_i$ denotes the vector of exposure variables.

Note that, the exponential function ensures that the mean function of PRM remains non-negative. For a discrete response, $Y_i = 0, 1, 2, \ldots$ such that, $Y_i \sim Poisson(\mu_i)$ the probability mass function (pmf) of PRM is given by,

$$f(Y_i|\mu_i) = \frac{\exp(-\mu_i)\mu_i^{Y_i}}{Y_i!}. \tag{2.1}$$

PRM has expected value

$$E(Y_i|\mu_i) = \mu_i;$$

and variance

$$Var(Y_i|\mu_i) = \mu_i.$$

As is clearly seen above, as $E(Y_i|\mu_i) = Var(Y_i|\mu_i)$, PRM requests for equidispersion in count data. Due to this restricted property of PRM, more flexible count data regression models are developed to account for overdispersion (where the value of the variance is larger than the value of mean) and underdispersion (where the value of the variance is smaller than the value of mean). Following this, we will also consider the negative binomial regression model which is another commonly used count data regression model when dealing with extra variability.

2.1.2 Negative Binomial Regression Model

As mentioned in the previous subsection, when count data shows evidence of overdispersion, PRM is no longer appropriate and therefore, NBRM shall be used as an alternative which allows the variance to be larger than the mean. Using NBRM, Campbell et al. (2002) conducted a case-control study on a sample of women enrollees in a metropolitan health maintenance organization to identify the significances of physically and/or sexually abused women, meanwhile Makary et al. (2010) used frailty in 594 patients between 2005 and 2006 as a measure of predictor for surgical outcomes. NBRM was also used by Lozano et al. (2013) to study data on causes of death across 187 countries from the year 1980 to the year 2010.

The modelling of data with overdispersed counts using the NBRM is made possible with the introduction of a dispersion parameter, $r > 0$. Introduced by Consul and Jain (1973), using similar notations as in subsection 2.1.1 where, $\mu_i = \exp(\boldsymbol{X}_i'\boldsymbol{\beta})$, let $Y_i \sim NB(r, \mu_i)$, where its pmf is defined by,

$$f(Y_i|\mu_i) = \frac{\Gamma(Y_i + r)}{Y_i!\Gamma(r)}\left(\frac{r}{r+\mu_i}\right)^r\left(\frac{\mu_i}{r+\mu_i}\right)^{Y_i}, \tag{2.2}$$

such that, $\Gamma(.)$ is the gamma function. NBRM has following mean and variance,

$$E(Y_i|\mu_i) = \mu_i, \quad \text{and}$$
$$Var(Y_i|\mu_i) = \mu_i\left(1 + \frac{\mu_i}{r}\right).$$

It is clear that since $\mu_i > 0$, and the variance is the product of mean, μ_i, and positive dispersion factor, $1 + (\mu_i/r)$, thus NBRM can be used to model overdispersed count data. As noted in Winkelmann (2008) when r approaches infinity, NBRM converges to PRM with parameter μ_i.

2.1.3 Overview of Other Count Data Regression Models

PRM and NBRM are the two most commonly utilized regression models to analyse count data. However, there are other regression models that are developed to accommodate different properties or problems that may arise when considering count data such as inflated number of zero counts. For this, zero-inflated models are used to model the zero counts by considering the binary and count processes separately, that is, the model estimates zero counts using a different type of distribution than the non-zero counts. According to Winkelmann (2008), there are two main reasons why addressing excess zeros in count data is important. The first reason is that from an empirical point of view the ratio of the number of zeros to the number of non-zeros is often too high to be compatible with a standard underlying count data regression models. The second reason is that zeros often reflect corner solution outcomes in economic choice models.

The zero-inflated Poisson (ZIP) model is a model that can be used to address zero-inflation or non-occurrences in equidispersed count data. In literature, ZIP are implemented in various applications, including manufacturing defects (Lambert, 1992), road safety (Miaou, 1994) and health care utilizations (Gurmu, 1997).

Another model that can be used to model zero-inflated count data is the zero-inflated negative binomial regression model. The zero-inflated negative binomial model is an extension from zero-inflated Poisson but with the relaxation on the restriction for equidispersion assumption. Its applications in literature include, modeling accident frequencies (Shankar et al., 1997), consumption of cigarettes (Sheu et al., 2004) and marijuana-related problems among college students (Simons et al., 2006).

2.2 Errors-in-Variables Model

The earliest literature that could be found to the best of our knowledge on the discussion of error in measurement is by Pearson (1902). In the epidemiology field of studies, Wong et al. (1999) conducted research to eliminate bias caused by errors-in-variables (EIV) in linear models. Fuller (2009) provided an extensive review on linear models with EIV and its effects on causing bias in parameter estimations.

Meanwhile, for non-linear models which are measured with error, a comprehensive account of literature are discussed in Carroll et al. (2006), where the authors discussed various methods on estimating regression coefficients with bias reduction in non-linear models with EIV.

There are many issues that may contribute to the situation in which the exposure variables are measured with error. An instance of which has contributed to the rise of measurement error is due to instrument/human error. To elaborate, in a self-reported dietary intake study, participants are asked to report their intake which is inaccurate, according to Schoeller (1990). This is because they depend on the recall method which is prone to human error. Errors-in-variables may also arise when it is impossible or expensive to measure the true exposure variables directly. Pridemore (2011) described an investigation on the relationship between poverty and homicide rates. In their investigation, there is no physical instrument that can measure the actual value of poverty. Therefore, they take surrogate values that might indicate deprivation in place of the true poverty values and consequently, biased regression parameters are estimated.

Carroll et al. (2006) gave two types of EIV classification such that for EIV, ϵ, where ϵ is independent and identically distributed,

$$X^* = X + \epsilon, \tag{2.3}$$

$$X = X^* + \epsilon. \tag{2.4}$$

Note that, the true unobserved exposure is denoted by X, and its corresponding surrogate exposure is denoted by X^*. Equation (2.3) refers to the classical EIV model meanwhile

Equation (2.4) refers to Berkson or non-classical EIV model. Classical EIV model is used to model the conditional distribution of the observed with error surrogate exposure variables given the unobserved true exposure variables.

In classical EIV model as given in Equation (2.3), its true exposure, X, is unobserved and instead its surrogate measures, X^*, are observed with contamination of error, ϵ. ϵ is independent of outcome and true exposure variables. Although many studies assumed that ϵ is normally distributed, this is not always the case especially if the data exhibit skewness. According to Verbeke and Lesaffre (1996) and Ghosh et al. (2007), the normal assumption lacks robustness against departures from normality. Following this, Huang and Dagne (2011) investigated the performance of skew-normal distribution in modeling both random error and random effects under the non-linear mixed-effects model. In the same vein, Fu et al. (2015) considered skew-normal and skew-t distributions for random errors and random effects for zero-inflated Poisson with measurement error in its covariates.

As given in Equation (2.4), Berkson error model, X is equal to the sum ofits corresponding surrogate, X^* and measurement error. One example of Berkson error is in most typical ecological experiments, where the amount of nutrients given to a certain plant is recorded. However, the real value of nutrients uptake by the plant is unknown.

The stark difference between classical measurement error model and Berkson error model is that in the former, ϵ is independent of X. Meanwhile, in the latter, its ϵ is independent of X^*. These independence properties imply that for classical measurement error, $Var(X^*) > Var(X)$ and for Berkson error, $Var(X) > Var(X^*)$.

For our study, we assume the EIV follows the classical model. This is because, according to Carroll (1989), Berkson error suggests that there is little to no bias in log-linear regression coefficients. In addition to this, most studies assume the variance of the measurement error as known, however, in our study, we will estimate its value. We will discuss this further in Chapters 4 and 5.

There are many effects of EIV if not addressed. A comprehensive account discussing the impact of EIV is provided by Carroll et al. (2006). If EIV is not corrected, one of the consequences includes attenuation where the error causes bias to the slope estimate in the direction of zero. Bias caused by EIV often leads to more serious problems. As mentioned in Gustafson (2003), the regression relationship between outcome and accurately measured covariates becomes distorted and will also produce biased regression estimates if not addressed. In addition to this, the confidence limits of the regression estimates would also be artificially narrow. According to Carroll et al. (2006), the effects of EIV depend on the type of regression model; if the mismeasured variable is univariate, then the magnitude of bias present in the measurement will be smaller in comparison to the magnitude of bias in multivariate mismeasured variables. Nevertheless, bias in both should be addressed in order to diverge from false statistical inferences.

2.3 Techniques to Correcting Errors-in-Variables Problem

There is a considerable amount of research done on methods of mitigating bias caused by measurement error. Two broad classifications of addressing EIV model (EIVM) are Bayesian and frequentist (non-Bayesian) approaches. In non-Bayesian (classical) or frequentist paradigm, there is a number of estimators that can be employed to reduce bias when estimating regression parameters in the presence of measurement error. Meanwhile, in the Bayesian paradigm, a general and unified framework can be employed to accommodate different types of models and scenarios.

2.3.1 Non-Bayesian Techniques to Correcting Errors-in-Variables Problem

In this section, we will discuss the basics of the frequentist methods and their strengths and weaknesses which will show the reason why Bayesian should be the preferred method.

The structural quasi-score (SQS) is a method used to address measurement error which was first proposed by Wedderburn (1976) in generalized linear model (GLM).

Kukush et al. (2004) demonstrated the implementation of SQS to PRM. Instead of depending on the whole distribution of outcome variable given the surrogate exposures, SQS is only dependent on its conditional mean and variance. The SQS function for Poisson regression is given by Carroll et al. (2006), subsequently the solution to the function is solved using iteratively reweighted least square method. However, in terms of bias-variance tradeoffs, other methods (e.g. regression calibration (RCAL) and Bayesian) show better values in comparison with SQS method (Carroll & Stefanski, 1990). The discussion for the usage of SQS for NBRM in the presence of EIV is presented in Yang (2012). However, according to the author, adjusted MLE achieved higher efficiency than SQS.

For PRM, two most prominent methods for reducing bias caused by measurement error are conditional score and corrected score. The conditional score was first introduced by Lindsay (1982). The unobserved true covariates are treated as unknown parameters and their sufficient statistics are obtained. Conditional on the sufficient statistics, the conditional score function is constructed from the mean and variance of the outcome variables.

Meanwhile, corrected score was first developed by Stefanski (1989) and later, Nakamura (1990) improved the score function with its implementation focused on Poisson regression. The corrected score function is built on the basis that the expectation of the corrected function is equal to the expectation of the usual score function conditional on the unknown true exposure variables. By maximizing or finding the zero-crossing to the derivatives of the corrected score function, one may solve the function and thus the estimated parameters are obtained.

Conditional score performed better than corrected score when error distribution is normal. On the other hand, when the error distribution is non-normal, conditional score yields inconsistent roots. Nonetheless, the corrected score has not been widely adopted in practice due to its pathological behaviors. When the measurement error is high, corrected score reveals multiple roots, estimate finding failure as well as skewness even when the sample size is large. Therefore, Huang (2014) attempted to fix these behaviours in his paper by imposing trend constraints on the score. Nevertheless, the corrected score is still disadvantaged as it assumes that the parameters in the measurement error distribution as known which is unrealistic in practice.

Regression calibration (RCAL) is one of the most straightforward approaches, introduced by Carroll and Stefanski (1990). RCAL is also known as linear imputation method. The RCAL method addresses the error by transforming the observed covariate with the conditional mean of its estimated true covariate given its respective surrogate covariate. It is obtained by imputing the estimated true covariate for each observation, given the value of the surrogates. There are a few disadvantages of the RCAL method. One of the disadvantages is that the surrogate values are only considered during the first iteration of estimating the true covariates and on later estimating iterations, the imputed values are used as the regressors in the outcome model which encourages the propagation of uncertainty. Moreover, although RCAL yields consistent estimate for slope parameters, the same could not be said for the intercept parameters.

Maximum likelihood estimator (MLE) is a non-Bayesian method which was first implemented in the EIVM context by Fuller (2009) for linear models. In the same vein, Carroll et al. (1993) suggested the use of maximum likelihood and least square to covariates measured with error in generalized linear models which can be applied to PRM and NBRM. However, the score function for MLE in PRM and NBRM are complicated (Yang, 2012). Thus, Yang (2012) proposed an adjusted MLE which can be applied to approximate the MLE for both PRM and NBRM in the presence of EIV.

Simulation extrapolation (SIMEX) is a straightforward and simple method for reducing bias caused by errors-in-variables in count data regression models. The usage of SIMEX is first developed by Cook and Stefanski (1994) where the measurement error variance is either estimated or assumed as known. The algorithm of SIMEX method is as follows. Given the original dataset, independent measurement error is added to create a new dataset. Using direct regression, the naive estimates are obtained in the new dataset. Further measurement error is added and estimating the parameters are repeated a large number of times. A smooth line or curve is then fitted to the mean of these estimated parameters. Finally, SIMEX estimates with bias correction are obtained by finding back the extrapolated estimates in the case where the measurement error variance is zero. Whilst the advantage of SIMEX method in comparison to the Bayesian method is that it can be simply implemented with no exposure distribution specification, it does not yield good estimation especially when measurement error is high even when its sample size is big. In addition to this, SIMEX also risks poor extrapolation bias (Küchenhoff & Carroll, 1997).

2.3.2 Bayesian Techniques to Correcting Errors-in-Variables Problem

In count data regression models, most of the research dealing with errors in covariates implemented frequentist methods due to the complexity of integral imposed when using Bayesian approach. However, over the last decade, the availability of Markov chain Monte Carlo sampling has provided a path for the complex integrals problem in Bayesian method to be dealt with implicitly and therefore, has greatly simplified the difficulties faced in the Bayesian paradigm.

Bayesian treatment of errors-in-variables in epidemiology study was introduced by Richardson and Gilks (1993) to the logistic regression model. Meanwhile, Dellaportas and Stephens (1995) and Mallick and Gelfand (1996) used the Bayesian formulation for EIV in nonlinear regression and GLM, respectively. The latter study featured Poisson regression as an example.

Up to now, much less attention has been given to address EIV using Bayesian methods in PRM and NBRM compared to other types of regression despite its importance in modeling count data. This is especially true for PRM in the Bayesian paradigm and even more so for NBRM in general. To the best of our knowledge, studies that have focused on fixing EIV in NBRM are El-Basyouny and Sayed (2010) and Yang et al. (2013). Both papers dealt with errors in covariates in NBRM using Bayesian methods and applied them to safety analysis which is important in road safety applications. It is important to note that, in their studies, the normality and log-normality assumption were imposed on the true exposure distribution. Thus any departures from normality and/or log-normality will result in an added misspecification bias.

The frequentist methods in reducing bias caused when estimating count data regression estimates show a few serious drawbacks. This can be easily avoided if one uses the Bayesian approach instead. There are many advantages to using the Bayesian approach compared to frequentist approaches. Following a study done by Gustafson (2003), our study therefore uses Bayesian approach based on many grounds, such as,

1. larger gain in efficiency in comparison to frequentist approach,
2. parameters of the measurement error distribution is estimated instead of assumed as known,
3. construction of likelihood based credible intervals that have coverage probabilities closer to the minimal level and,
4. applicable to a wide range of problems with a unified framework.

However, the Bayesian approach is often attacked with the fact that it requires the specification of exposure model and therefore will have the risk of model misspecification. Thus, to counter this, many studies have proposed intentionally misspecifying the exposure model with a flexible distribution. The mixture of normal distributions as flexible exposure model were attempted by Carroll et al. (1999) in the linear regression model. Later, mixtures of normal distribution were extended to be implemented in EIV logistic models (Richardson et al., 2002). However, these authors reported that the performance of the mixture of normals as flexible exposure model deteriorated in the case of skewed and heavy-tailed true exposure distribution. Seeing these weaknesses, Hossain and Gustafson (2009) studied skew-normal (SN) distribution and its more flexible variants, namely the flexible generalized skew-normal (FGSN) and flexible generalized skew-*t* (FGST) as exposure model in the case of logistic outcome regression model where problems such as the detection of artifactual modes when using normal mixtures and semi-nonparametric density are solved.

SN distribution which was introduced by Azzalini (1985) provides more flexibility in modeling the unobserved exposures, however when the exposures have a heavy-tailed distribution the performance is unsatisfactory. Similarly, for FGSN, the flexible distribution showed adequate performance in correcting bias and had reasonable bias-variance trade-off, but when the true unobserved quantities have heavy-tailedness property, FGSN lacked robustness in capturing the shape of the distribution. Therefore, Hossain and Gustafson (2009) advocated the usage of FGST. In our research, we focused on the implementation of FGSN as the flexible model as FGST is considered redundant in the case of count data outcome regression models which will be discussed in Section 3.3.2.

Subsequently, in addition to all the advantages listed above, another advantage of the flexible Bayesian approach is that by misspecifying the true unobserved exposure distributions as flexible, we are able to capture any skewness, heavy-tailedness, and bimodality in the distribution of count data exposure covariates that are contaminated with error. Finally, the approach is more appealing as it has a larger gain in efficiency (Roeder et al., 1996) and general applicability (Richardson & Gilks, 1993).

To the best of our knowledge, there is yet any study conducted in correcting EIV in PRM and NBRM using the Bayesian approach with intentionally misspecified flexible exposure distribution. Besides that, current frequentist approach in correcting EIV in both PRM and NBRM assumes the variance of EIV as known. Using the flexible Bayesian approach, the assumption of variance as unknown is allowed and could be estimated aided with validation data.

2.4 Bayesian Inference

In this section, the basic ideas of the Bayesian approach are briefly discussed. We base our explanation of the following subsections from Gilks et al. (1996) and Gelman et al. (2014).

2.4.1 Likelihood Distribution

Let $f(y|\boldsymbol{\theta})$ be the density function of observable quantities, $Y = y$, that depends on a set of parameters vector, $\boldsymbol{\theta}$, which is usually referred to as the likelihood function such that Y only affects the posterior through $f(Y|\boldsymbol{\theta})$. In Bayesian inference, parameter $\boldsymbol{\theta}$ is assumed to be random with prior distribution $\pi(\boldsymbol{\theta})$. Bayesian inference follows the likelihood principle which expresses that the inferences on the value of $\boldsymbol{\theta}$ is found in the equivalence class to which $f(y|\boldsymbol{\theta})$ belongs.

2.4.2 Prior Distribution

The prior distribution of $\boldsymbol{\theta}$ characterizes the ‘prior beliefs’ or ‘prior information’ of $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ could be a set of parameters vector or latent variable. Before choosing a prior distribution, the distribution must be able to cover the range of all the possible values of the unknown quantity. For example, if $\theta \in (0, \infty)$, then the distribution of prior must not have the range of $(-\infty, \infty)$. Most applications prefer the usage of conjugate prior (if it is available), but not all likelihood distribution will have its corresponding conjugate prior distribution. A conjugate prior is when the prior probability distribution has the same family as the posterior distributions.

Informative prior could be used if conjugate prior is not available. Also known as the subjective prior, informative prior is specified when there is a presence of prior information. The information may come from either expert opinions or from previous experiments and applications. If there is a lack of prior information, then non-informative prior can be used. However, even though the prior information is available, an investigator might also prefer to specify the prior distribution where such prior is referred to as uninformative prior (also called as ‘flat’ prior). A reason for this is to ‘let the data speak for themselves’. An example of uninformative prior is the normal distribution with large variance, i.e., $N(\mu, 100^2)$, or uniform distribution, $U(0, 1)$. As an alternative, one may also use diffuse priors (or weakly informative prior) where only a little information is included in the prior but not enough to hugely to be able to influence the posterior.

To choose between the different types of prior is based on two major issues; the ‘deepness’ of information of θ that is chosen to be included and properties of posterior density. Gelman (2006) and Gelman et al. (2014) provided thorough discussion on prior distributions.

2.4.2.1 Posterior Density of Bayesian Model

Using Bayes theorem, the posterior distribution of $\boldsymbol{\theta}$ is as follows,

$$f(\boldsymbol{\theta}|y) = \frac{f(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(y)} \tag{2.5}$$

$$= \frac{f(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{2.6}$$

$$\propto f(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{2.7}$$

The general posterior distribution as shown in Equation (2.6), is not an analytically and numerically tractable function, which is why before the introduction of Markov chain Monte Carlo (MCMC) researchers refused to use the Bayesian approach in their studies. However, with the introduction of Markov chain Monte Carlo algorithm, instead of solving the complex Equation (2.6), its proportional and simpler counterpart, Equation (2.7) can be used to generate approximate samples of the posterior distribution. Thus, the samples can be utilized to approximate the desired summary of the posterior distribution (e.g., posterior mean, mode etc.).

2.4.3 Posterior Density of Bayesian Hierarchical Model

Bayesian hierarchical model is also known as the Bayesian multilevel model. There are many reasons on why hierarchical models are important in the Bayesian paradigm. According to Efron and Morris (1975) and Morris (1983), theoretically, hierarchical models estimate the parameters of the prior distribution from the data rather than specifying them manually which is a more objective approach. In the Bayesian hierarchical model, the hyperparameter, $\boldsymbol{\phi}$, is assumed as unknown and therefore has its own prior distribution which shall be labeled as $\pi(\boldsymbol{\phi})$ and is known as hyperprior. From this, the joint prior distribution is given as,

$$\pi(\boldsymbol{\theta}, \boldsymbol{\phi}) = \pi(\boldsymbol{\theta}|\boldsymbol{\phi})\pi(\boldsymbol{\phi}),$$

and now the posterior distribution is given as follows,

$$\begin{aligned} f(\boldsymbol{\theta}, \boldsymbol{\phi}|y) &\propto f(y|\boldsymbol{\theta}, \boldsymbol{\phi})\pi(\boldsymbol{\theta}|\boldsymbol{\phi})\pi(\boldsymbol{\phi}) \qquad (2.8) \\ &= f(y|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\phi})\pi(\boldsymbol{\phi}), \qquad (2.9) \end{aligned}$$

which the simplification of Equation (2.8) to Equation (2.9) holds as the data distribution, $f(\boldsymbol{\theta}, \boldsymbol{\phi}|y)$ depends only on $\boldsymbol{\theta}$ and the hyperparameter $\boldsymbol{\phi}$ affects y only through $\boldsymbol{\theta}$.

Bayesian hierarchical models are able to accommodate very complicated structures from a succession of relatively simple components, yielding better flexibility (Ntzoufras, 2011). Other advantages include good performance as well as ease of computation. We shall discuss this further in Chapter 4.

2.5 Markov chain Monte Carlo Algorithm

In this section, we briefly discuss on Markov chain Monte Carlo (MCMC) sampling. A more detailed explanation is provided by Gilks et al. (1996). Markov chain, named after Andrey Markov, is a random process where a memoryless transition from one state to another state takes place and the transition probabilities for its next state only depend on the current state and not on the previous states (Gilks et al., 1996). To illustrate this in a mathematical notation, let $X^{(t)}$ be the random variable at state t and $x^{(t)}$ denotes the observed value of $X^{(t)}$ at state t, such that,

$$P(X^{(t+1)} = x|X^{(1)} = x^{(1)}, X^{(2)} = x^{(2)}, \dots, X^{(n)} = x^{(n)}) = P(X^{(t+1)} = x|X^{(t)} = x^{(t)}).$$

Monte Carlo method (Metropolis & Ulam, 1949) is a method of drawing independent and identically distributed samples from a target distribution. The distribution that is desired can be approximated by the simulated samples and once the Markov chain converges to the stationary distribution, the Markov chain will be able to estimate the quantities of interest (e.g. posterior mean, posterior mode, etc). As mentioned before, Monte Carlo can be used to solve integration problems. This is possible by the law of large numbers, such that,

$$\frac{1}{N}\sum_{t=1}^{N} g(x_i) \xrightarrow[N\to\infty]{\text{a.s.}} \int_X g(x)f(x)dx,$$

where x_i is the i^{th} sample from the target distribution, $f(.)$, N is the total number of draws and $g(.)$ is a measureable function of X. Therefore, to reiterate, by the law of large numbers, the integral estimate is unbiased and will converge to the value of the solved integral.

Thus, MCMC is able to randomly sample from a probability distribution that is too complex to simulate from directly. Before the introduction of MCMC, practitioners have avoided the usage of Bayesian methods. The Metropolis algorithm, which was first developed by Metropolis et al. (1953) is a MCMC method that can be used when the full conditional posterior distribution does not take a known form. The Metropolis algorithm is later modified by Hastings (1970) to not require symmetry in the proposal function which is now known as the Metropolis-Hastings (MH) algorithm. Gibbs sampler, which was first used by Geman and Geman (1984) for Bayesian image restoration, drew random samples from the target posterior distribution without solving Equation (2.6), which may consists of an integration that is computationally intractable.

2.5.1 Metropolis-Hastings Algorithm

As discussed extensively in Gilks et al. (1996), for the MH algorithm, at current state, t, the next state value, $X^{(t+1)}$, is chosen by sampling a candidate value, $X^{(cand)}$, from a proposal distribution, $q(.|X^{(t)})$, where the proposal distribution may depend on the current value, $X^{(t)}$. The candidate value, $X^{(cand)}$, is accepted with probability $\alpha(X^{(cand)}|X^{(t)})$, such that

$$\alpha(X^{(cand)}|X^{(t)}) = min\left(1, \frac{f(X^{(cand)})q(X^{(t)}|X^{(cand)})}{f(X^{(t)})q(X^{(cand)}|X^{(t)})}\right),$$

where $f(.)$ is the target density. If the candidate value is accepted, then, let $X^{(t+1)} = X^{(cand)}$. If the candidate value is rejected, then, let $X^{(t+1)} = X^{(t)}$.

Tierney (1994) introduced the usage of autoregressive chains when estimating parameters using MH algorithm. According to the study conducted, these chains can be used to induce negative autocorrelation between successive elements of the chain by letting

$$X^{(cand)} = a + B(X^{(t)} - a) + z,$$

and

$$q(X^{(t)}|X^{(cand)}) = q(X^{(cand)} - a - B(X^{(t)} - a)),$$

where a is a vector and B is a matrix such that both are conformable with $X^{(t)}$, $q(.)$ is a symmetric proposal distribution and z has $q(.)$ its density. If B is set to be the negative of identity matrix, $-\boldsymbol{I}$, then the chains produced will be reflected about the point a thus, the chains induced will have a negative autocorrelation. A simpler method of MH with autoregressive chains is by generating a candidate step that is reflected around the current value, $X^{(t)}$, about the point, a, to produce $X^{(cand)} = 2a - X^{(t)}$. Now, the probablity of acceptance is

$$\alpha(X^{(cand)}|X^{(t)}) = min\left(1, \frac{f(2a - X^{(t)})}{f(X^{(t)})}\right),$$

where $f(.)$ is the target density.

2.5.2 Random Walk Metropolis Hastings Algorithm

As shown in detail by Gilks et al. (1996), in random walk Metropolis Hastings (RWMH) algorithm, the proposal distribution is symmetric such that it is in the form of the following,

$$q(X^{(cand)}|X^{(t)}) = q(X^{(t)}|X^{(cand)}) = q(|X^{(cand)} - X^{(t)}|).$$

Thus, the the acceptance probability is simplified just the ratio of the target densities,

$$\alpha(X^{(cand)}|X^{(t)}) = min\left(1, \frac{f(X^{(cand)})}{f(X^{(t)})}\right).$$

The algorithm for the RWMH is the same as the MH algorithm, with the acceptance probability shown above. In RWMH, the variance of the proposal distribution can be tuned using tuning parameter to make the variance higher or lower. When the variance of the proposal distribution increases, the acceptance rate decreases. When the variance of the proposal distribution decreases, the acceptance rate increases. Therefore, tuning parameter can be used to control the acceptance rate of a RWMH algorithm (Chib & Greenberg, 1995). According to Roberts et al. (1997), the recommended acceptance rate is in the range of 30% to 60%.

2.5.3 Gibbs Sampler

If $\boldsymbol{X}$ is n-dimensional, instead of updating the whole of $\boldsymbol{X}$ by block, it is more convenient and computationally efficient to divide $\boldsymbol{X}$ into components, $\{X_1, X_2, \ldots, X_n\}$, of possibly differing dimensions and update these components one by one as proposed by Metropolis et al. (1953). Let X_i be the i^{th} component and, let X_{-i} be the set of all components except X_i, Gibbs sampling is a special case of single-component MH where the values are sampled exactly from the conditional distributions as the conditional distributions are in a closed form of known distributions. To clarify, the proposal density is the target density, i.e.,

$$q(X_i^{(cand)}|X_{-i}^{(t)}) = f(X_i^{(cand)}|X_{-i}^{(t)}),$$

such that, $f(X_i^{(cand)}|X_{-i}^{(t)})$ is the target density. The result of this, is that, the acceptance probability will always equal to one, i.e., the Gibbs sampler candidates are always accepted. The following shows that the acceptance probability is always equal to 1 for Gibbs sampling:

$$\begin{aligned}
&\alpha(X_i^{(cand)}, X_{-i}^{(t)}|X_i^{(t)}, X_{-i}^{(t)}) \\
&= min\left(1, \frac{q(X_i^{(t)}, X_{-i}^{(t)}|X_i^{(cand)}, X_{-i}^{(t)})f(X_i^{(cand)}, X_{-i}^{(t)})}{q(X_i^{(cand)}, X_{-i}^{(t)}|X_i^{(t)}, X_{-i}^{(t)})f(X_i^{(t)}, X_{-i}^{(t)})}\right) \\
&= min\left(1, \frac{f(X_i^{(t)}|X_{-i}^{(t)})f(X_i^{(cand)}, X_{-i}^{(t)})}{f(X_i^{(cand)}|X_{-i}^{(t)})f(X_i^{(t)}, X_{-i}^{(t)})}\right) \\
&= min\left(1, \frac{f(X_i^{(t)}|X_i^{(t)})f(X_i^{(cand)}|X_{-i}^{(t)})f(X_{-i}^{(t)})}{f(X_i^{(cand)}|X_{-i}^{(t)})f(X_i^{(t)}|X_{-i}^{(t)})f(X_{-i}^{(t)})}\right) \\
&= 1.
\end{aligned}$$

According to Banerjee et al. (2014), Gibbs sampler generates new values at each iteration slower than the MH sampler. However, its convergence is much faster.

CHAPTER 3: BAYESIAN FRAMEWORK TO CORRECTING ERRORS-IN-VARIABLES IN REGRESSION MODELS

3.1 Conditional Independence Model for Errors-in-Variables Scenario

In this chapter, we shall discuss the framework where the Bayesian approach is utilized to address EIV in regression models. The Bayesian approach is constructed using conditional independence model that was first introduced by Richardson and Gilks (1993). Based on their paper, three submodels need to be specified. For $i, \ldots, n$, let the outcome variable be Y_i, X_i as the true but unobserved covariate and X_i^* is its corresponding surrogate of X_i which is observed with error. Therefore, according to Richardson and Gilks (1993) the three submodels are distinguished as the following,

1. *Outcome model* with density denoted by $f(Y_i|X_i, \boldsymbol{\theta}_O)$, which expresses the relationship between outcome Y and X with parameter vector $\boldsymbol{\theta}_O$.

2. *Measurement model* with density denoted by $f(X_i^*|X_i, \boldsymbol{\theta}_M)$, which expresses the relationship between the surrogate X^* and true covariate X with parameter vector $\boldsymbol{\theta}_M$.

3. *Exposure model* with density denoted by $f(X_i|\boldsymbol{\theta}_E)$, which describes the distribution of true X with parameter vector $\boldsymbol{\theta}_E$.

From the three submodels, the joint distribution of our model in the presence of EIV can be written as

$$f(Y_i, X_i^*, X_i|\boldsymbol{\theta}_O, \boldsymbol{\theta}_M, \boldsymbol{\theta}_E) = f(Y_i|X_i, \boldsymbol{\theta}_O) f(X_i^*|X_i, \boldsymbol{\theta}_M) f(X_i|\boldsymbol{\theta}_E). \tag{3.1}$$

As seen in Equation (3.1), a distribution is specified for each of the submodels and each involving their respective unknown parameters $\boldsymbol{\theta}_O, \boldsymbol{\theta}_M$ and $\boldsymbol{\theta}_E$. Thus, Equation (3.1) could be used in constructing the likelihood function of the unknown parameters, if the observed quantities is given by (X_i^*, Y_i, X_i). Realistically, only $\boldsymbol{S} = (\boldsymbol{X}^*, \boldsymbol{Y})$ is observed, therefore the density,

$$\begin{aligned} f(\boldsymbol{X}^*, \boldsymbol{Y}|\boldsymbol{\theta}_O, \boldsymbol{\theta}_M, \boldsymbol{\theta}_E) &= \int f(X_i^*, Y_i, X_i|\boldsymbol{\theta}_O, \boldsymbol{\theta}_M, \boldsymbol{\theta}_E) dX_i \\ &= \int f(Y_i|X_i, \boldsymbol{\theta}_O) f(X_i^*|X_i, Y_i, \boldsymbol{\theta}_M) f(X_i|\boldsymbol{\theta}_E) dX_i, \end{aligned} \tag{3.2}$$

is needed to form the likelihood function for our model. In some problems such as in binary outcome variable, the integral shown above is intractable. Nonetheless, we can evaluate the integral using Markov chain Monte Carlo (MCMC) methods. As mentioned in Section 2.5, the strength of MCMC is that it has provided an easier path for evaluating complex integrals problem in Bayesian paradigm. So, the integral in Equation (3.2) can be dealt with implicitly and Equation (3.1) is evaluated instead.

In our study, we assume a non-differential EIV such that given the true exposure variable, the surrogate exposure variable does not depend on the outcome variable, i.e., $f(X_i^*|X_i, Y_i, \boldsymbol{\theta}_M) = f(X_i^*|X_i, \boldsymbol{\theta}_M)$; EIV is differential if otherwise. Many problems can plausibly be classified as having a non-differential error, especially when the X_i and X_i^* occur at a fixed point of time and Y_i measured at a later time (Carroll et al., 2006). In addition to this, to ensure parameter identifiability, additional data is needed for the parameter $\boldsymbol{\theta}_M$ of the measurement model. According to Richardson and Gilks (1993) these additional data that help ensure identifiability in EIV analysis can be categorised as the following,

1. *Data from previous studies*, such that Y_i and X_i^* are the variables and the parameter $\boldsymbol{\theta}_M$ is measurable.

2. *Validation data* in which the true exposure variable, X_i, is measured directly (also known as ‘gold standard’ data).

3. *Replication data* in which repeated measurements of X_i^* are available.

The type of additional data used in EIV scenarios must be inspected upon its practicality. If data from previous studies are used as additional data and parameter $\boldsymbol{\theta}_M$ is known, then one must investigate if the value of $\boldsymbol{\theta}_M$ is transportable across different study populations. In some cases, accurately measured X_i may also be available for a subset of the study and is referred to as 'validation sample' or 'gold-standard sample'. Greenland (1988) and Spiegelman et al. (1994) studied the relationship between cost-information tradeoffs and the size of the gold-standard sample to the main study sample. However, in reality, the gold-standard data/sample are often unavailable or expensive. Thus, additional data with replicated measures of X_i^* is preferred in study applications. In our research, in order to maintain realistic approaches to correcting EIV problems, we use replication data to ensure parameter identifiability.

3.2 Formulation of Posterior Distribution in the Presence of Errors-in-Variables

Assume *n* study subjects with exposure and outcome variables independent of each other, the joint distribution of all the relevant quantities is written as

$$f(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{X}^*, \boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ f(Y_i|X_i, \boldsymbol{\theta_O}) f(X_i^*|X_i, \boldsymbol{\theta}_M) f(X_i|\boldsymbol{\theta}_E) \right\} \times \pi(\boldsymbol{\theta}_O, \boldsymbol{\theta}_M, \boldsymbol{\theta}_E), \qquad (3.3)$$

such that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$, $\boldsymbol{X}^* = (X_1^*, X_2^*, \ldots, X_n^*)$, $\boldsymbol{\theta}$ denotes the parameter vector of the model that contains $\boldsymbol{\theta}_O, \boldsymbol{\theta}_M$ and $\boldsymbol{\theta}_E$ and $\pi(.)$ denotes the prior distribution of the model parameters.

As given by Bayes theorem, the density of unobserved quantities, $\boldsymbol{U} = (\boldsymbol{X}, \boldsymbol{\theta})$, given the density of observed quantity, $\boldsymbol{S} = (\boldsymbol{X}^*, \boldsymbol{Y})$, is proportional to the joint density of $\boldsymbol{U}$ and

$\boldsymbol{S}$. So, the posterior density is proportional to the joint density of $\boldsymbol{U}$ and $\boldsymbol{S}$, such that

$$f(\boldsymbol{X}, \boldsymbol{\theta} | \boldsymbol{X}^*, \boldsymbol{Y}) \propto \prod_{i=1}^{n} \left\{ f(Y_i | X_i, \boldsymbol{\theta}_O) f(X_i^* | X_i, \boldsymbol{\theta}_M) f(X_i | \boldsymbol{\theta}_E) \right\} \times \pi(\boldsymbol{\theta}_O, \boldsymbol{\theta}_M, \boldsymbol{\theta}_E). \qquad (3.4)$$

To find the actual normalized posterior density of the unobserved quantities $\boldsymbol{U}$, given the observed quantities $\boldsymbol{S}$, the integration of Equation (3.4) over $\boldsymbol{U}$ given fixed $\boldsymbol{S}$ must be calculated. Solving the integration of Equation (3.4) is impossible unless it is in closed form, which can only be achieved in EIV problems if the regression model is linear. Nevertheless, as alluded in Section 3.1, MCMC does not need one to solve the normalized integral of posterior density and therefore, Equation (3.4) is enough when we want to carry out analysis on the model parameter $\boldsymbol{\theta}$ (Gustafson, 2003). To elaborate, MCMC algorithm can be implemented to draw samples from the distribution of the unobserved quantities given the observed quantities. Furthermore, samples from the density $f(\boldsymbol{X}, \boldsymbol{\theta} | \boldsymbol{X}^*, \boldsymbol{Y})$ trivially lead to samples from density $f(\boldsymbol{\theta} | \boldsymbol{X}^*, \boldsymbol{Y})$ upon ignoring the sampled $\boldsymbol{X}$ values, i.e., MCMC algorithm samples from the distribution of the unobserved parameters given all the observed data. Therefore, all inferences on the model parameters and their respective distributions can be obtained from the MCMC samples. This is the greatest computational advantages of MCMC inference in scenarios involving mismeasurements, missing data or censored data over maximum likelihood and other classical approaches.

In our study, priori independence is assumed and thus joint distribution of all our priori can be written in the form,

$$\pi(\boldsymbol{\theta}_O, \boldsymbol{\theta}_M, \boldsymbol{\theta}_E) = \pi(\boldsymbol{\theta}_O)\pi(\boldsymbol{\theta}_M)\pi(\boldsymbol{\theta}_E).$$

3.2.1 Posterior Distribution with Additional Data for Measurement Model

In this section, we construct the posterior density for EIV scenarios where the measurement model has additional data to ensure parameter identifiability. As mentioned in Section 3.1, there are three types of additional data (Richardson & Gilks, 1993); data available from previous studies where $\boldsymbol{\theta}_M$ can be measured, validation data and replication data. $\boldsymbol{\theta}_M$ usually are non-transportable across different studies especially when the independent variables are measured with error, therefore realistically $\boldsymbol{\theta}_M$ observed in data from previous studies are very rarely considered. The posterior construction of this type of additional data is trivial and will not be discussed here. Validation data in measurement error scenarios refer to the availability of gold-standard measurements and they are usually expensive, therefore to reduce cost, instead of observing the gold-standard measurements for the entire study sample, only a subsample of the data is observed. Since the gold-standard sample is not pragmatic in real life situations, in our study, repeated measurements of surrogate exposures are used to extract extra information for identifiability. However, for the sake of discussion, we shall construct a posterior density in the presence of the gold-standard sample.

Validation Data

Let $\boldsymbol{X}_c$ denotes true and observed exposure $\boldsymbol{X}_c$, and $\boldsymbol{X}_r$ denotes the true but unobserved exposure $\boldsymbol{X}$ such that for the entire study sample, $\boldsymbol{X} = (\boldsymbol{X}_c, \boldsymbol{X}_r)$. Therefore, the posterior density is of the form

$$
\begin{aligned}
&f(\boldsymbol{X}_r, \boldsymbol{\theta}_O, \boldsymbol{\theta}_M, \boldsymbol{\theta}_E | \boldsymbol{X}^*, \boldsymbol{Y}, \boldsymbol{X}_c) \\
&\quad \propto \left[\prod_{i=1}^{n} f(Y_i | X_i, \boldsymbol{\theta}_O)\right] \times \left[\prod_{i=1}^{n} f(X_i^* | X_i, \boldsymbol{\theta}_M)\right] \times \left[\prod_{i=1}^{n} f(X_i | \boldsymbol{\theta}_E)\right] \times \pi(\boldsymbol{\theta}_O, \boldsymbol{\theta}_M, \boldsymbol{\theta}_E).
\end{aligned}
\tag{3.5}
$$

Even though the right hand-side of Equation (3.5) does not differ from that of the posterior density in the absence of validation sample, the MCMC algorithm for Equation (3.5) will provide a principled way to make simultaneous inferences about $\boldsymbol{\theta}_O, \boldsymbol{\theta}_M$ and $\boldsymbol{\theta}_E$.

Replication Data

Replication data are validation data that for at least some study subjects X_i^*, repeated measurements are available. Let m denotes the number of replicated measurements and assuming that replicated measurements of X_i^* are conditionally independent given the true value X_i, then the posterior density of the unobserved quantities, $(\boldsymbol{X}, \boldsymbol{\theta})$ given observed quantities, $(\boldsymbol{X}^*, \boldsymbol{Y})$ takes the following form,

$$\begin{aligned} f(\boldsymbol{X}, &\boldsymbol{\theta}_O, \boldsymbol{\theta}_M, \boldsymbol{\theta}_E | \boldsymbol{X}^*, \boldsymbol{Y}) \\ &\propto \prod_{i=1}^{n} f(Y_i|X_i, \boldsymbol{\theta}_O) \prod_{i=1}^{n} \prod_{j=1}^{m} f(X_{ij}^*|X_i, \boldsymbol{\theta}_M) \prod_{i=1}^{n} f(X_i|\boldsymbol{\theta}_E) \times \pi(\boldsymbol{\theta}_O, \boldsymbol{\theta}_M, \boldsymbol{\theta}_E). \end{aligned} \tag{3.6}$$

Note that, X_{ij}^* is the j^{th} replicate of surrogate X_i^* for the i^{th} study subject.

3.3 Misspecification of Outcome and Exposure Models

Bayesian formulation requires the specification of models, in which the distributional assumptions on outcome, Y_i, and exposure, X_i, are important for parameter estimation. Misspecification in the distribution of both exposure variables as well as outcome variables may lead to serious bias in estimation (Richardson et al., 2002). In this section, we shall discuss the misspecification of outcome and exposure models in EIV count data regression models.

3.3.1 Misspecification of Outcome Model

Both PRM and NBRM are commonly used for modeling count data outcomes. It is important to apply the correct regression models according to the characteristics of the count data in question to avoid any outcome misspecification bias. PRM assumes the equidispersion property where the mean shall be equal to the variance. If this property is violated, it would be wise to use an alternative model, such as NBRM. However, note that the violation of equidispersion may be caused by the presence of EIV. As shown in Guo and Li (2002), in PRM where $\boldsymbol{X}$ is unobservable, when using its surrogate $\boldsymbol{X}^*$ as proxy, equidispersion of mean, $E(\boldsymbol{Y}|\boldsymbol{X}^*)$, and variance, $var(\boldsymbol{Y}|\boldsymbol{X}^*)$, only holds when $E(\boldsymbol{Y}|\boldsymbol{X}^*) = 1$ or when the conditional density of $\boldsymbol{Y}$ given $\boldsymbol{X}$ is almost everywhere zero. If EIV is not the cause of overdispersion, then one must specify NBRM to model the count data instead of PRM.

3.3.2 Misspecification of Exposure Model

The exposure model is unknown which is a subsequent result of the unobservable nature of the true independent variable, $\boldsymbol{X}$, and therefore is exposed to the risk of misspecification. To avoid any distributional assumption, some researchers explore the use of functional approaches where no model specification is required; however, this may lead to a loss in efficiency in comparison to structural approaches (Huang, 2014).

To relax modeling assumptions, researchers that utilize structural approaches consider using flexible parametric models which were first utilized by Carroll et al. (1999). Carroll et al. (1999) demonstrated the use of mixtures of normals as flexible exposure model for linear EIV models. Meanwhile, Richardson et al. (2002) extended the use of mixtures of normals to EIV logistic model. However, these authors reported that the performance of the flexible model deteriorated in the case of skewed and heavy-tailed true exposure distributions. Huang et al. (2006) utilized second-order nonparametric density but the study did not investigate its robustness for exposure distribution with skewness and heavy-tailedness. Hossain and Gustafson (2009) implemented the flexible generalized skew-elliptical class of distributions, specifically they utilized flexible generalized skew-normal (FGSN) and flexible generalized skew-*t* (FGST) as the misspecified exposure distribution. They investigated the robustness of both FGSN and FGST to model exposure distribution that exhibited different levels of skewness and heavy-tailedness. In summary, they advocated the implementation of FGST as FGST showed better regression parameter estimations in comparison to FGSN. In our study, we focus on the implementation of FGSN which is described in detail in the next section. This is because FGST is considered as redundant in the case of count data regression models. Our simulated estimate of the degree of freedom parameter, v, of FGST is large and since FGST converges to FGSN when v goes to infinity, the implementation of FGSN is adequate. Moreover, computation time is decreased when using FGSN as there are fewer parameters that need to be updated in the simulation algorithm.

It is important to note that most studies only investigated the use of flexible distribution on models with logistic outcomes. Richardson et al. (2002), Huang et al. (2006) and Hossain and Gustafson (2009) investigated the usage of flexible models to reduce model misspecification sensitivity in logistic regression with EIV. A few other researchers also attempted the flexible parametric model on other types of outcome distribution; for example, Bolfarine and Lachos (2007) made use of skew-normal as the exposure model for probit regression. To date, there is no study that utilizes flexible parametric exposure model in the Bayesian paradigm for EIV in both PR and NBR models. Therefore, in our study, we shall investigate the performance of the implementation of intentionally misspecified flexible exposure model in reducing modeling assumptions.

(a) Flexible Distributions as Intentionally Misspecified Exposure Model

In this subsection, we discuss the flexible distributions that are considered in this dissertation. In typical studies of correcting for EIV, the normal distribution is used to model the true but unobserved exposures. However, if the distribution departed from normality, it is obvious that an added misspecification bias will decrease the accuracy in estimating the regression parameters. In our dissertation, we search for the most suitable flexible distribution that can be used to model the exposure distribution for both PRM and NBRM. It is important to note that, the exposure model is intentionally misspecified by a flexible model as realistically in EIV problems, the exposure distribution cannot be observed. To our knowledge, there is yet literature that contributed to the implementation of the Bayesian method with flexible exposure model for PRM and NBRM. Thus, we consider using FGSN to model for the unobserved quantities. Furthermore, we also study the performance of newer flexible models which are variants of the skew-normal (SN) distribution (Azzalini, 1985) that have been developed over the years, namely the flexible skew-generalized normal distribution (Nekoukhou et al., 2013) and the extended skew generalized normal distribution (Choudhury & Matin, 2011).

(i) Flexible Generalized Skew-Normal Distribution

According to Hossain and Gustafson (2009), an alternative choice of flexible model that can be used to handle both bimodality skewness and heavy-tailedness, and can offer a computational advantage is the flexible generalized skew-normal (FGSN) distribution. Genton and Loperfido (2005) developed this distribution under a class of distribution called the flexible generalize skew-elliptical. Since FGSN can accommodate bimodality, heavy-tailedness, and skewness, a higher degree of flexibility is offered when trying to capture the distribution of unobserved quantities. A thorough discussion on FGSN is provided in Ma and Genton (2004).

Let $\phi(.)$ denote the standard normal density and $\Phi(.)$ denote the standard normal distribution functions, respectively, then let the distribution of a random variable X, be a univariate FGSN with the density given as,

$$f(x) = \frac{2}{\lambda}\phi\left(\frac{x-\alpha}{\lambda}\right)\Phi\left[\sum_{h=1}^{H}\omega_h\left(\frac{x-\alpha}{\lambda}\right)^{2h-1}\right], \tag{3.7}$$

where $\alpha \in \Re$ is the location parameter, and $\omega_h \in \Re$ and $\lambda > 0$ are the shape and scale parameters, respectively and $h = 1, 2, \ldots, H$, such that, $K = 2H - 1$ signifies the order of the polynomial. FGSN is unimodal if $K = 1$ and if $K = 3$ FGSN may have at most two modes (Ma & Genton, 2004). From Equation (3.7), it can be seen that

1. If $\omega_h = 0$, for all h, then Equation (3.7) reduces to a normal distribution.

2. If $\omega_h = 0$ for $h = 2, 3, \ldots, H$, but $\omega_1 \neq 0$ then Equation (3.7) reduces to a SN distribution.

A higher value of K will offer more flexibility, but efficiency is sacrificed. In our study, we use $K = 3$ as according Ma and Genton (2004) this value of K would offer enough flexibility to capture the properties of the unobserved exposure model.

(ii) Flexible Skew-Generalized Normal Distribution

Flexible skew-generalized normal (FSGN) distribution is developed by Nekoukhou et al. (2013) that stems from a skew generalized-normal (SGN) introduced by Arellano-Valle et al. (2004) which is the generalization of Azzalini's SN distribution . The flexibility of FSGN is introduced by adding more parameters to model for the modes in the distribution.

Let the distribution of a random variable X to be a univariate FSGN with the density given as,

$$f(x) = \frac{2}{\lambda_1}\phi\left(\frac{x-\alpha}{\lambda_1}\right)\Phi\left(\frac{\omega_1(x-\alpha) + \omega_2(x-\alpha)^3/\lambda_1^2}{\sqrt{\lambda_1^2 + \lambda_2(x-\alpha)^2}}\right), \qquad x \in \Re, \tag{3.8}$$

where $\alpha \in \Re$ is the location parameter and $\lambda_1 > 0$ is the scale parameter. Also, $\omega_1, \omega_2 \in \Re$ and $\lambda_2 \geq 0$ are constants. From Equation (3.8), it is clear that

1. If $\omega_h = 0$ for $h = 1, 2$ but $\lambda_2 \neq 0$, then Equation (3.8) reduces to the a normal distribution for all $\lambda_2 > 0$.

2. If $\lambda_2 = \omega_2 = 0$ but $\omega_1 \neq 0$, then Equation (3.8) reduces to a SN distribution for all $\omega_1 \in \Re$.

3. If $\omega_2 = 0$, for all $\omega_1 \in \Re$ and $\lambda_2 > 0$, then Equation (3.8) reduces SGN distribution.

4. If $\lambda_2 = 0$, for all $\omega_1, \omega_2 \in \Re$, then Equation (3.8) coincides with FGSN of $K = 3$.

FSGN distribution is more flexible than FGSN as it may be reduced to the latter distribution.

CHAPTER 4: BAYESIAN APPROACH TO ERRORS-IN-VARIABLES IN POISSON REGRESSION MODEL

4.1 Introduction

Estimating parameters of PRM often leads to bias as the data collected are prone to EIV problems. There are many existing non-Bayesian methods proposed to address this problem, however, most of them require the variance of the measurement error (ME) distribution to be known. This rarely happens in practice and even with this assumption, these existing estimators exhibit pathological behaviour, inconsistent root problems as well as estimate-finding failure. Thus, we utilized the Bayesian approach to address EIV PRM such that the variance parameter of the ME distribution is estimated instead.

We also intentionally misspecify the exposure model with a flexible distribution, in order to relax distributional assumption and therefore decrease the impact of model misspecification bias. Since most studies done in correcting bias in EIV for parameter estimations often impose a normal assumption on the true exposure distribution, in our study we conducted extensive simulation studies for different properties of underlying exposure model (i.e, skewness, bimodality and heavy-tailedness). So, we shall study the performance of two flexible distributions, flexible generalized skew-normal (FGSN) and flexible skew-generalized normal (FSGN), in relaxing the distributional assumptions of the exposure model. To the best of our knowledge, there is yet any study conducted in correcting EIV in PRM using the Bayesian approach with intentionally misspecified flexible exposure distribution.

As mentioned in Chapter 3, the underlying structure of the joint distribution is a product of the probability density function (pdf) of the three different submodels which was provided by Richardson and Gilks (1993). Thus, we shall specify the outcome model, measurement model, and exposure model to apply the Bayesian approach to EIV in PRM.

Throughout this chapter, we shall consider independent count data with outcome, Y_i, $i = 1, \ldots, n$ where n is the sample size and their corresponding accurately measured but unobserved variables X_i. Let X_i^* be their respective surrogate covariate that was measured with error.

4.2 Poisson Regression Outcome Model

Suppose Y_i follows a PRM distribution, such that its probability mass function (pmf) is written as

$$f(Y_i|X_i, \boldsymbol{\theta}_{PRM}) = \frac{\mu_i^{Y_i} \exp(-\mu_i)}{Y_i!}, \tag{4.1}$$

where

$$\mu_i = \exp(\beta_0 + \beta_1 X_i), \tag{4.2}$$

such that, the vector of parameters, $\boldsymbol{\theta}_{PRM} = (\beta_0, \beta_1)$, is our main inferential focus and the main parameter vector that we want to estimate with accuracy in the presence of EIV.

4.3 Measurement Model

In this study, we choose normal distribution as the measurement model distribution as the distribution shows robustness in modelling EIV even when the EIV distribution is non-normal (refer to Section 4.7). Its pdf is given by,

$$f(X_{ij}^*|X_i, \boldsymbol{\theta}_M) = \left(\frac{1}{2\pi\tau^2}\right)^{1/2} \exp\left(-\frac{1}{2\tau^2}\left(X_{ij}^* - X_i\right)^2\right), \tag{4.3}$$

such that $\boldsymbol{\theta}_M = \tau^2$ and X_{ij}^* signifies the j^{th} replicated surrogate of i^{th} observation of $\boldsymbol{X}^*$ for $j = 1, \ldots, m$ and $\boldsymbol{X}^*$ is the observed surrogate of $\boldsymbol{X}$. To ensure identifiability and in order to successfully estimate the measurement error (ME) variance, τ^2, additional data or error assessment data are necessary. There are several types of error assessment data (as discussed in Chapter 3); however, to closely follow a realistic approach, the data are available in forms of *m* replicated surrogates.

4.4 Bayesian Approach using Flexible Exposure Model

The third model required to form the joint distribution given by Richardson and Gilks (1993) is the exposure model; therefore the specification of exposure model is required.

To obtain the adjusted estimated regression parameters in the presence of EIV with flexible misspecification of exposure model, the true exposure X_i is generated from different types of distribution according to their respective simulation settings. However, we misspecify the exposure distribution as a flexible distribution to relax modeling assumptions. Therefore, we investigate the performance of various misspecified flexible exposure models, i.e., FGSN and FSGN in the Bayesian EIV model and test their robustness in simulation studies using synthetic data sets.

4.4.1 Flexible exposure model − FGSN

We shall let X_i follow the FGSN distribution such that

$$f(X_i|\boldsymbol{\theta}_{FGSN}) = \frac{2}{\lambda}\phi\left(\frac{X_i-\alpha}{\lambda}\right)\Phi\left[\omega_1\left(\frac{X_i-\alpha}{\lambda}\right)+\omega_2\left(\frac{X_i-\alpha}{\lambda}\right)^3\right], \tag{4.4}$$

where $\boldsymbol{\theta}_{FGSN} = (\alpha, \lambda, \omega_1, \omega_2)$, $\Phi(.)$ is the standard normal distribution function and $\phi(.)$ is the standard normal density. In our study, we use polynomial of order $K = 3$ following Ma and Genton (2004), as polynomial of that particular order offers enough flexibility; a higher number of K will offer more flexibility, however, efficiency will be sacrificed.

4.4.2 Flexible Exposure Model – FSGN

We let X_i follow the FSGN distribution such that

$$f(X_i|\boldsymbol{\theta}_{FSGN}) = \frac{2}{\lambda_1}\phi\left(\frac{X_i-\alpha}{\lambda_1}\right)\Phi\left(\frac{\omega_1(X_i-\alpha)+\omega_2(X_i-\alpha)^3/\lambda_1^2}{\sqrt{\lambda_1^2+\lambda_2(X_i-\alpha)^2}}\right), \tag{4.5}$$

where $\boldsymbol{\theta}_{FSGN} = (\alpha, \lambda_1, \lambda_2, \omega_1, \omega_2)$.

4.5 Joint Posterior Density

4.5.1 Flexible Bayesian Approach under FGSN exposure model

In this section, we use FGSN as the flexible exposure model. The same models are utilized here for both outcome and measurement models stated in Sections 4.2 and 4.3.

Following Equation (3.6), the joint posterior density of all the relevant variables, can be written as

$$f(\boldsymbol{X}, \boldsymbol{\theta}|\boldsymbol{X}^*, \boldsymbol{Y}) \propto \prod_{i=1}^{n} f(Y_i|X_i, \boldsymbol{\theta}_{PRM}) \prod_{i=1}^{n}\prod_{j=1}^{m} f(X_{ij}^*|X_i, \boldsymbol{\theta}_M) \prod_{i=1}^{n} f(X_i|\boldsymbol{\theta}_{FGSN}) \times \pi(\boldsymbol{\theta}), \tag{4.6}$$

where $\boldsymbol{\theta}$ is the parameter vector of the model that contains $\boldsymbol{\theta}_{PRM}, \boldsymbol{\theta}_M$ and $\boldsymbol{\theta}_{FGSN}$ which denote vectors of parameters for outcome, measurement and exposure model, respectively.

In the case of Poisson outcome model, we introduce latent variable, $\eta_i = \beta_0 + \beta_1 X_i$ to ease computational complexity and achieve faster convergence rate (Asfaw Dagne, 1999). By introducing η_i, we will show that $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ can be updated using Gibbs sampling.

Let $\pi(\boldsymbol{\theta})$ denote the prior distribution for $\boldsymbol{\theta} = (\boldsymbol{\eta}, \boldsymbol{\beta}, \alpha, \tau^2, \sigma^2, \lambda^2, \omega_1, \omega_2)$ where $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ is an unknown vector of parameters and the main parameter vector that we want to estimate. Assuming priori independence, the joint distribution for all our priori is given by

$$\pi(\boldsymbol{\eta},\boldsymbol{\beta},\tau^2,\sigma^2,\alpha,\lambda^2,\omega_1,\omega_2) = \left\{\prod_{i=1}^{n}\pi(\eta_i|\boldsymbol{\beta},\sigma^2)\right\}\pi(\boldsymbol{\beta})\pi(\alpha)\pi(\tau^2)\pi(\sigma^2)\pi(\lambda^2)\pi(\omega_1)\pi(\omega_2), \tag{4.7}$$

where $\boldsymbol{\beta}$ and σ^2 are set to be the hyperparameters for hyperprior of η_i.

We assign an informative prior for the latent variables η_i introduced in PRM where it follows normal with mean and variance $\beta_0+\beta_1 X_i$ and σ^2, respectively. The hyperparameter $\boldsymbol{\beta}$ and location parameter α are set to have a flat prior with locally uniform distribution, $U(1)$ as suggested by Box and Tiao (2011). The prior distributions for ω_1 and ω_2 are assigned to be a normal distribution with high variance as to ensure that the priori are as close to non-informative as possible. The reasoning behind this is to let the data be the main role in estimating these parameters. We set the distribution of prior for scale parameters τ^2 and λ^2 to be $IG(0.5, 0.5)$ where IG stands for inverse-Gamma distribution. According to Gelman et al. (2014), the centre of $IG(0.5, 0.5)$ is equal to one and thus, the prior guesses for both τ^2 and λ^2 are one which shows that the prior has a unit information for its variance components. This implies that the information relayed using the prior is worth a single data point about the variance components and therefore the data will steer the estimation of τ^2 and λ^2. Following this information, it is safe to say that $IG(0.5, 0.5)$ is a non-informative prior distribution. Similarly, the hyperprior for σ^2 is also set to be $IG(0.5, 0.5)$.

Rewriting Equation (4.6) in a more detailed manner, we obtain the following joint posterior density:

$$
\begin{aligned}
f(\boldsymbol{X},\boldsymbol{\theta}|\boldsymbol{X}^{*},\boldsymbol{Y}) \propto \prod_{i=1}^{n} \Bigg\{ & \left[\frac{\exp(Y_i\eta_i)\exp(-\exp(\eta_i))}{Y_i!}\right]\left[\prod_{j=1}^{m}\left(\frac{1}{\tau^2}\right)^{1/2}\exp\left(-\frac{1}{2\tau^2}(X_{ij}^{*}-X_i)^2\right)\right] \\
& \times\left[\left(\frac{1}{\lambda^2}\right)^{1/2}\exp\left(-\frac{1}{2\lambda^2}(X_i-\alpha)^2\right)\right]\Phi\left[\left(\frac{\omega_1(X_i-\alpha)}{\lambda}\right)+\left(\frac{\omega_2(X_i-\alpha)^3}{\lambda^3}\right)\right]\Bigg\} \\
& \times\left\{\prod_{i=1}^{n}\pi(\eta_i|\boldsymbol{\beta},\sigma^2)\right\}\pi(\boldsymbol{\beta})\pi(\tau^2)\pi(\lambda^2)\pi(\alpha)\pi(\sigma^2)\pi(\omega_1)\pi(\omega_2),
\end{aligned}
\tag{4.8}
$$

Conditional Posterior Density

In this subsection, the conditional posterior density for each of the parameters studied is now derived from Equation (4.8). The derivation of the conditional posterior density for all the parameters are reparameterised into closed forms (if possible). We then estimate each parameters using MCMC sampling method. Let us denote A^C as the vector of all model parameters except A.

MCMC Implementation

i. For η_i,

$$f(\eta_i|\eta_i^C) \propto \exp\left\{y_i\eta_i - \exp(\eta_i) - \frac{1}{2\sigma^2}\left[\eta_i - (\beta_0+\beta_1 X_i)\right]^2\right\}$$

We introduce η_i in PRM for parameter $\boldsymbol{\beta}$ as the parameter shows slow convergence rate. It is clear from the conditional posterior that part of it is a normal distribution with mean $\beta_0 + \beta_1 X_i$ and variance σ^2. This latent variable is updated by component using random walk Metropolis-Hastings (RWMH) with autoregressive chain. Its proposal distribution is univariate normal with the aforementioned mean and variance. The algorithm for this type of RWMH is described in Section 2.5.1.

ii. For $\boldsymbol{\beta}$,

$$f(\boldsymbol{\beta}|\boldsymbol{\beta}^C) \propto \exp\big(-\frac{1}{2\sigma^2}(\boldsymbol{\eta} - \boldsymbol{X\beta})'(\boldsymbol{\eta} - \boldsymbol{X\beta})\big).$$

$\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)'$ is $n \times 1$ matrix and $\boldsymbol{X}$ is $n \times 2$ matrix with the i^{th} row equals to $(1, X_i)$. The conditional posterior of $\boldsymbol{\beta}$ follows normal distribution, which is possible after the latent variable η_i is obtained. Applying linear transformation and completing of squares on the above conditional posterior as suggested by Gelman et al. (2014), starting with,

$$(\boldsymbol{\eta} - \boldsymbol{X\beta})'(\boldsymbol{\eta} - \boldsymbol{X\beta}) = \boldsymbol{\eta}'\boldsymbol{\eta} - 2\boldsymbol{\eta}'\boldsymbol{X\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta} \tag{4.9}$$

and differentiating (4.9) with respect to $\boldsymbol{\beta}$,

$$\begin{aligned} -2\boldsymbol{\eta}'\boldsymbol{X} + 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X} &= 0 \\ \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X} &= \boldsymbol{\eta}'\boldsymbol{X} \\ \boldsymbol{\beta}' &= \boldsymbol{\eta}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ \boldsymbol{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\eta}. \end{aligned}$$

Now, since $\boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\eta}$ and $var(\boldsymbol{\eta}) = \sigma^2 \cdot \boldsymbol{I}_n$ where $\boldsymbol{I}_n$ denotes identity matrix of order n then,

$$\begin{aligned} var(\boldsymbol{\beta}) &= var((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\eta}) \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'var(\boldsymbol{\eta})[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']' \\ &= \sigma^2 \cdot \boldsymbol{I}_n(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}((\boldsymbol{X}'\boldsymbol{X})^{-1})' \\ &= \sigma^2 \cdot \boldsymbol{I}_n(\boldsymbol{X}'\boldsymbol{X})^{-1}. \end{aligned}$$

Thus, the conditional posterior density for $\boldsymbol{\beta}$ now follows a multivariate normal distribution with mean $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\eta}$ and covariance matrix $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Therefore, $\boldsymbol{\beta}$ is updated using Gibbs sampling.

iii. For X_i,

$$f(X_i|X_i^C) \propto \exp\left\{-\frac{1}{2\sigma_X^2}(X_i-\mu_X)^2\right\}\left\{\Phi\left(\frac{\omega_1(X_i-\alpha)}{\lambda}+\frac{\omega_2(X_i-\alpha)^3}{\lambda^3}\right)\right\},$$

where

$\sigma_X^2 = \tau^2\sigma^2\lambda^2/(m\lambda^2\sigma^2+\tau^2\sigma^2+\beta_1^2\lambda^2\tau^2)$,

$\mu_X = (m\lambda^2\sigma^2\bar{X}_i^* + \alpha\tau^2\sigma^2 + \beta_1(\eta_i-\beta_0)\tau^2\lambda^2)/(m\lambda^2\sigma^2+\tau^2\sigma^2+\tau^2\lambda^2\beta_1^2)$,

$\bar{X}_i^* = \sum_{j=1}^m X_{ij}^*/m.$

Note that the main part of this conditional posterior has a normal distribution, with mean μ_X and variance σ_X^2. Hence, $X_i, i = 1,\dots,n$ are component-wise updated using independent normal with mean μ_X and variance σ_X^2 as proposals via the Metropolis-Hastings (MH) algorithm.

iv. For α,

$$f(\alpha|\alpha^C) \propto \exp\left\{-\frac{n}{2\lambda^2}(\alpha-\bar{X})^2\right\}\left\{\prod_{i=1}^n \Phi\left(\frac{\omega_1(X_i-\alpha)}{\lambda}+\frac{\omega_2(X_i-\alpha)^3}{\lambda^3}\right)\right\},$$

where the first component of the above conditional distribution is normal with mean $\bar{X}$ and variance λ^2/n where $\bar{X} = \sum_{i=1}^n X_i/n.$ To have good mixing and acceptance rate when updating α, we use RWMH scheme with $N(0, k_\alpha^2\lambda^2/n)$ as proposal distribution where k_α is the tuning parameter. We set $k_\alpha = 0.75$ so that the algorithm exhibits acceptance rate between 30% and 40%.

v. For ω_h where $h = 1, 2$,

$$f(\omega_h|\omega_h^C) \propto \left\{ \prod_{i=1}^{n} \Phi\left(\frac{\omega_1(X_i - \alpha)}{\lambda} + \frac{\omega_2(X_i - \alpha)^3}{\lambda^3} \right) \right\} \exp\left\{ - \frac{\omega_h^2}{2 \times 100} \right\}.$$

These shape parameters both have $N(0, k_\omega^2)$ as their proposal distributions and are sampled using RWMH sampling method. For both parameters, we set the tuning parameter, k_ω as 0.5 which yield acceptance rate between 30% and 40%.

vi. For τ^2,

$$f(\tau^2|\tau^{2^C}) \propto \left(\frac{1}{\tau^2}\right)^{\frac{mn+1}{2}+1} \exp\left[- \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(X_{ij}^* - X_i)^2 + 1}{2\tau^2} \right],$$

which is IG with shape and scale parameter $(mn+1)/2$ and $\sum_{i=1}^{n} \sum_{j=1}^{m} 0.5(X_{ij}^* - X_i)^2 + 0.5$, respectively. Therefore, to update τ^2, the Gibbs sampler is used.

vii. For λ^2,

$$f(\lambda^2|\lambda^{2^C}) \propto \left(\frac{1}{\lambda^2}\right)^{\frac{n+1}{2}+1} \exp\left[- \frac{0.5}{\lambda^2} \left(\sum_{i=1}^{n} (X_i - \alpha)^2 + 1 \right) \right] \\ \times \left\{ \prod_{i=1}^{n} \Phi\left(\frac{\omega_1(X_i - \alpha)}{\lambda} + \frac{\omega_2(X_i - \alpha)^3}{\lambda^3} \right) \right\},$$

where the main part of the conditional posterior is IG with shape $(n + 1)/2$ and scale $\sum_{i=1}^{n} 0.5(X_i - \alpha)^2 + 0.5$. Hence, we use MH algorithm to update this scale parameter with proposal distribution $IG((n + 1)/2, \sum_{i=1}^{n} 0.5(X_i - \alpha)^2 + 0.5)$.

viii. For σ^2,

$$f(\sigma^2|\sigma^{2^C}) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n+1}{2}+1} \exp\left[-\frac{1}{2\sigma^2}((\boldsymbol{\eta} - \boldsymbol{X\beta})'(\boldsymbol{\eta} - \boldsymbol{X\beta}) + 1)\right],$$

which is IG with shape $0.5(n+1)$ and scale $0.5(\boldsymbol{\eta} - \boldsymbol{X\beta})'(\boldsymbol{\eta} - \boldsymbol{X\beta}) + 0.5$. Thus, we update σ^2 using Gibbs sampler.

4.5.2 Flexible Bayesian Approach under FSGN Exposure Model

We also study the effectiveness of a newer flexible model, that is the FSGN distribution in modeling the unobserved exposures. We shall use the same model for both the outcome and measurement models specified in Sections 4.2 and 4.3; in this section, instead of FGSN, we specify FSGN as the flexible exposure model.

Thus, when using FSGN as the exposure model, the joint posterior density following Equation (3.6) is given by

$$f(\boldsymbol{X}, \boldsymbol{\theta}|\boldsymbol{X}^*, \boldsymbol{Y}) \propto \prod_{i=1}^{n} f(Y_i|X_i, \boldsymbol{\theta}_{PRM}) \prod_{i=1}^{n} \prod_{j=1}^{m} f(X_{ij}^*|X_i, \boldsymbol{\theta}_M) \prod_{i=1}^{n} f(X_i|\boldsymbol{\theta}_{FSGN}) \times \pi(\boldsymbol{\theta}), \tag{4.10}$$

where $\boldsymbol{\theta}$ is the parameter vector of the model that contains $\boldsymbol{\theta}_{PRM}, \boldsymbol{\theta}_M$ and $\boldsymbol{\theta}_{FSGN}$ which denote vectors of parameters for outcome, measurement and FSGN exposure model respectively.

Again, we introduce $\eta_i = \beta_0 + \beta_1 X_i$ as the latent variable as MCMC sampling for parameter $\boldsymbol{\beta}$ shows slow convergence when updating using MH. Assuming priori independence, the joint distribution for all priori is given as

$$\begin{aligned}\pi(\boldsymbol{\eta}, \boldsymbol{\beta}, &\tau^2, \lambda_1^2, \lambda_2, \alpha, \sigma^2, \omega_1, \omega_2)\\ &= \left\{\prod_{i=1}^{n} \pi(\eta_i|\boldsymbol{\beta}, \sigma^2)\right\}\pi(\boldsymbol{\beta})\pi(\tau^2)\pi(\lambda_1^2)\pi(\lambda_2)\pi(\alpha)\pi(\sigma^2)\pi(\omega_1)\pi(\omega_2).\end{aligned} \tag{4.11}$$

As in the case of FGSN exposure model, the same prior distributions are adopted for $\eta_i, \boldsymbol{\beta}, \sigma^2, \tau^2, \alpha$ and ω_h for $h = 1, 2$. For scale parameter λ_1^2 of FSGN distribution, $IG(0.5, 0.5)$ is used which follows the same reasoning as the parameters that are set to have $IG(0.5, 0.5)$ as their parameter distribution; that is to let the data be the commandeer of the parameter estimation. As for λ_2, we use half-normal distribution with scale parameter 1, centered around 0 as its prior distribution (Gelman, 2006). Half-normal distribution as the prior for λ_2 is appropriate as the distribution has a positive support.

Now, Equation (4.10) can be written as

$$\begin{aligned}f(\boldsymbol{X}, \boldsymbol{\theta}|\boldsymbol{X}^*, \boldsymbol{Y}) \propto \prod_{i=1}^{n} &\left\{\left[\frac{\exp(Y_i\eta_i)\exp(-\exp(\eta_i))}{Y_i!}\right]\left[\prod_{j=1}^{m}\left(\frac{1}{\tau^2}\right)^{1/2}\exp\left(-\frac{1}{2\tau^2}\left(X_{ij}^* - X_i\right)^2\right)\right]\right.\\ &\times\left.\left[\left(\frac{1}{\lambda_1^2}\right)^{1/2}\exp\left(-\frac{1}{2\lambda_1^2}(X_i-\alpha)^2\right)\right]\Phi\left(\frac{\omega_1(X_i-\alpha)+\omega_2(X_i-\alpha)^3/\lambda_1^2}{\sqrt{\lambda_1^2+\lambda_2(X_i-\alpha)^2}}\right)\right\}\\ &\times\left\{\prod_{i=1}^{n}\pi(\eta_i|\boldsymbol{\beta}, \sigma^2)\right\}\pi(\boldsymbol{\beta})\pi(\sigma^2)\pi(\tau^2)\pi(\lambda_1^2)\pi(\lambda_2)\pi(\alpha)\pi(\omega_1)\pi(\omega_2).\end{aligned} \tag{4.12}$$

The conditional density of each of the parameters in question can now be obtained from Equation (4.12). The details on the derivation and MCMC sampling methods are discussed in the next subsection.

4.5.3 Conditional Posterior Density

We consider the conditional posterior density for all the parameters used in our flexible Bayesian approach with FSGN as the exposure model and find the possible reparametrisation of the densities into closed forms. Note that, conditional posterior densities for latent variable, η_i and parameters $\boldsymbol{\beta}$, σ^2 and τ^2 have the same densities as the ones in Subsection 4.5.1. Therefore, their implemented MCMC methods are also the same for the aforementioned parameters. The conditional densities for $\alpha, \lambda_1^2, \lambda_2, \omega_1$ and ω_2 are described and using MCMC sampling method, the estimation of these parameters are also done in this subsection.

MCMC Implementation

i. For X_i,

$$f(X_i|X_i^C) \propto \exp\left\{-\frac{1}{2\sigma_X^2}(X_i-\mu_X)^2\right\}\left\{\Phi\left(\frac{\omega_1(X_i-\alpha)+\omega_2(X_i-\alpha)^3/\lambda_1^2}{\sqrt{\lambda_1^2+\lambda_2(X_i-\alpha)^2}}\right)\right\},$$

such that,

$$\sigma_X^2 = \tau^2\sigma^2\lambda_1^2/(m\lambda_1^2\sigma^2+\tau^2\sigma^2+\beta_1^2\lambda_1^2\tau^2),$$

$$\mu_X = (m\lambda_1^2\sigma^2\bar{X}_i^* + \alpha\tau^2\sigma^2 + \beta_1(\eta_i-\beta_0)\tau^2\lambda_1^2)/(m\lambda_1^2\sigma^2+\tau^2\sigma^2+\tau^2\lambda_1^2\beta_1^2),$$

$$\bar{X}_i^* = \textstyle\sum_{j=1}^m X_{ij}^*/m.$$

Since the main part of the above conditional posterior has a normal distribution of mean μ_X and variance σ_X^2, then we shall use this as a proposal to update X_i independently for $i = 1, 2, \ldots, n$ using MH algorithm.

ii. For α,

$$f(\alpha|\alpha^C) \propto \exp\left\{-\frac{n}{2\lambda_1^2}(\alpha-\bar{X})^2\right\}\left\{\prod_{i=1}^n \Phi\left(\frac{\omega_1(X_i-\alpha)+\omega_2(X_i-\alpha)^3/\lambda_1^2}{\sqrt{\lambda_1^2+\lambda_2(X_i-\alpha)^2}}\right)\right\},$$

where the first component of the conditional distribution is normal with mean and variance $\bar{X} = \sum_{i=1}^n X_i/n$ and λ_1^2/n, respectively. To update the parameter α with good mixing and acceptance rate, we use RWMH with normal proposal distribution, $N(0, k_\alpha^2\lambda_1^2/n)$ where k_α is the tuning parameter. We choose $k_\alpha = 1$ which so that the acceptance rate is between 35% and 40%.

iii. For λ_1^2,

$$f(\lambda_1^2|\lambda_1^{2^C}) \propto \left(\frac{1}{\lambda_1^2}\right)^{\frac{n+1}{2}+1} \exp\left[-\frac{0.5}{\lambda_1^2}\Big(\sum_{i=1}^{n}(X_i-\alpha)^2+1\Big)\right]$$
$$\left\{\prod_{i=1}^{n}\Phi\left(\frac{\omega_1(X_i-\alpha)+\omega_2(X_i-\alpha)^3/\lambda_1^2}{\sqrt{\lambda_1^2+\lambda_2(X_i-\alpha)^2}}\right)\right\},$$

where the main part of the condtional posterior is IG. Utilizing MH algorithm, the proposal distribution for λ_1^2 is $IG((n+1)/2, 0.5(\sum_{i=1}^{n}(X_i-\alpha)^2+1)$.

iv. For λ_2,

$$f(\lambda_2|\lambda_2^C) \propto \exp\left(-\frac{\lambda_2^2}{2}\right)\left\{\prod_{i=1}^{n}\Phi\left(\frac{\omega_1(X_i-\alpha)+\omega_2(X_i-\alpha)^3/\lambda_1^2}{\sqrt{\lambda_1^2+\lambda_2(X_i-\alpha)^2}}\right)\right\},$$

such that $\lambda_2 > 0$ and the first component on the right-handside of the conditional posterior is the half-normal distribution. Thus, λ_2 is updated using RWMH with *Half-Normal*$(0, k_{\lambda_2}^2)$ as its proposal distribution and tuning parameter, $k_{\lambda_2} = 0.1$, yields acceptance rate between 10% and 30%.

v. For ω_h where $h = 1, 2$,

$$f(\omega_h|\omega_h^C) \propto \left\{\prod_{i=1}^{n}\Phi\left(\frac{\omega_1(X_i-\alpha)+\omega_2(X_i-\alpha)^3/\lambda_1^2}{\sqrt{\lambda_1^2+\lambda_2(X_i-\alpha)^2}}\right)\right\}\exp\left\{-\frac{\omega_h^2}{2\times 100}\right\}.$$

These shape parameters have independent $N(0, k_\omega^2)$ as their proposal distribution and are updated using RWMH sampling method. $k_\omega = 0.5$ is chosen as the tuning parameter which exhibits acceptance rate between 25% and 40%.

4.6 Simulation Studies

In this section, we conduct extensive simulation studies to investigate the performance of the proposed technique under various different true unobserved X_i distributions for the count data regression outcome models discussed in Section 4.2. To thoroughly confirm the robustness of the Bayesian approach with misspecified flexible exposure model, we check its bias correction mechanism when the distribution of X shows evidence of departures from normality, that is, skewness, bimodality, and heavy-tailedness in various simulation settings. We also compare our findings against different levels of error contamination denoted as R such that $R = 0.25$, 0.5 and 1.0 indicating low, medium and high magnitude of error, respectively. Note that, R here is the ratio of ME variance to the variance of true X, i.e., $R = \tau^2/(var(X))$.

4.6.1 Simulation Set-ups

Let Y_i denote non-negative count integers; PRM are denoted by $Y_i \sim$ Poisson$(\exp(\beta_0 + \beta_1 X_i))$. The true regression parameters take values of $(\beta_0, \beta_1) = (0.5, 1.0)$. As previously stated, the surrogate X^*_{ij} follows classical ME model such that, $X^*_{ij} = X_i + \epsilon_j$ for $j = 1, \ldots, m$ where $\epsilon_j \sim N(0, \tau^2)$ and m denotes the number of repeated measurements. Later on, we will also simulate ϵ_j from non-normal distributions, namely skew-normal and skew-t distributions. For the sake of simulating data that is similar to real life research situations, the number of replicated surrogates is limited to $m = 2$. Values of τ^2 are estimated instead of assumed as known like many other studies done on EIV in count data models. The following are the simulation set-ups for the distribution of synthetic data sets for true exposure distribution:

Simulation setting 1: $X_i \sim 0.5N(0.19, 0.08^2) + 0.2N(1.05, 0.2^2) + 0.3N(2, 0.48^2)$

Simulation setting 2: $X_i \sim 0.5N(-2, 1) + 0.5N(2, 1)$

Simulation setting 3: $X_i \sim Gamma(2, 2^{-1})$

Simulation setting 4: $X_i \sim LN(0, 1)$

Simulation settings 1 and 2 follow similar configuration as Richardson et al. (Richardson et al., 2002). The first configuration follows an asymmetric mixture of normal which corresponds to a skewed true exposure distribution. $\tau^2 = 0.25$ signifies low ME. Meanwhile, $\tau^2 = 0.556$ and $\tau^2 = 1.11$ correspond to medium and high error contamination, respectively. Simulation setting 2 represents symmetric but bimodal mixture of normal. To generate low, medium and high ME contamination in the case where the true exposure has a bimodal distribution, let $\tau^2 = 0.75$, 1.49 and 2.94, respectively.

To generate true exposure distribution with high skewness and heavy tail, we consider simulation setting 3 where X_i is generated from Gamma with shape and scale parameter of 2. $\tau^2 = 2, 4$ and 8 will generate low to high error contamination for this simulation setting.

Finally, in simulation setting 4, true exposure is generated from log-normal distribution to study the effectiveness of the proposed flexible model to capture skewness and even heavier tail relative to simulation setting 3 in the exposures of count data regression. We set the ME variance to be $\tau^2 = 1.1675, 2.335$ and 4.67 for low, medium and high ME, respectively. Under each simulation setting, 50 data sets are generated for two different sample sizes ($n = 50, 100$).

4.7 Results

In this section, the performance of our proposed flexible Bayesian approach to correct EIV in PRM are presented for each simulation settings discussed in Section 4.6 under two flexible distributions, i.e., FGSN and FSGN. For each of the 50 data sets, we run MCMC chains of length $300,000$ and the first $100,000$ MCMC iterations are discarded. For each data set, we compute the posterior estimates of each of the model parameters with sample size $200,000$ which is the remainder of the MCMC iterations after burn-in. The mean of these posterior estimates is taken as our model parameter estimates for each data set. The convergence of the chains are diagnosed by constructing trace plots, and the plots show that our simulation study has good mixing and have achieved convergence with the given iteration length. Example of the trace plots for our parameter estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$ from a randomly selected dataset and simulation study are presented in Figure 4.1.

Table 4.1 contains the results of various analyses for FGSN exposure model while, Table 4.2 contains the results for FSGN exposure model with labels:

1. M as the mean of the model parameter estimates obtained based on the 50 different data sets,

$$\frac{\sum_{t=1}^{50} \hat{\beta}_k^{(t)}}{50} \text{ for } k = 0, 1;$$

2. B as bias with respect to the mean of the true covariate values of the 50 data sets,

$$\frac{\sum_{t=1}^{50}(|\hat{\beta}_k^{(t)} - \beta_k|)}{50};$$

3. MSE as mean squared error of the mean estimates,

$$\frac{\sum_{t=1}^{50}(\hat{\beta}_k^{(t)} - \beta_k)^2}{50}.$$

Table 4.1: Accuracy and sensitivity of estimated parameters, β_0 and β_1 under different true and unobserved distributions of X Poisson regression model with FGSN as misspecified exposure model

Sample size n = 50

			$R = 0.25$		$R = 0.5$		$R = 1$		
Simulation setting	Parameter		Naive	Flexible	Naive	Flexible	Naive	Flexible	Benchmark
1	β_0	M	0.70519	0.31260	0.85549	0.33742	1.04232	0.35440	0.53171
		B	0.20519	0.18740	0.35549	0.16258	0.54232	0.14560	0.03171
		MSE	0.06877	0.07676	0.15480	0.07661	0.32303	0.08634	0.02186
	β_1	M	0.84786	1.04355	0.72825	1.01214	0.57743	0.98768	0.98292
		B	0.15214	0.04355	0.27175	0.01214	0.42257	0.01232	0.01708
		MSE	0.03277	0.01695	0.08454	0.02229	0.18908	0.03713	0.00620
2	β_0	M	0.97564	0.51338	1.19514	0.51631	1.43278	0.60145	0.51605
		B	0.47564	0.01338	0.69689	0.01631	0.93278	0.10145	0.01605
		MSE	0.27722	0.04078	0.53430	0.05722	0.90996	0.09056	0.03027
	β_1	M	0.77464	0.97630	0.66106	0.97612	0.53003	0.95171	0.99408
		B	0.22536	0.02370	0.33894	0.02388	0.46997	0.04829	0.00592
		MSE	0.06169	0.00921	0.12560	0.01458	0.22989	0.02365	0.00437
3	β_0	M	0.49727	0.46975	0.26295	0.44434	1.91242	0.32435	0.50125
		B	0.00273	0.03025	0.23705	0.05566	1.41242	0.17565	0.00125
		MSE	7.44074	0.07140	26.7085	0.14898	16.1708	0.41036	0.00023
	β_1	M	0.99619	0.99696	0.99972	0.99998	0.82313	1.02473	0.99980
		B	0.00381	0.00304	0.00028	0.00002	0.17687	0.02473	0.00020
		MSE	0.07038	0.00292	0.23499	0.00612	0.15807	0.01671	$2.23e$-6
4	β_0	M	0.14322	0.41165	0.56923	0.40196	0.69690	0.36098	0.49864
		B	0.35678	0.08835	0.06923	0.09804	0.19690	0.13902	0.00136
		MSE	11.8566	0.03353	1.90997	0.05993	3.83394	0.12809	0.00279
	β_1	M	0.98455	0.99754	0.91333	0.99517	0.85350	1.01420	1.00220
		B	0.01545	0.00246	0.08667	0.00483	0.14650	0.01420	0.00220
		MSE	0.13688	0.00424	0.05664	0.00889	0.10391	0.01927	0.00016

Sample size n = 100

			$R = 0.25$		$R = 0.5$		$R = 1$		
Simulation setting	Parameter		Naive	Flexible	Naive	Flexible	Naive	Flexible	Benchmark
1	β_0	M	0.67436	0.48483	0.82081	0.51815	1.00180	0.55411	0.49799
		B	0.17436	0.01517	0.32081	0.01815	0.50180	0.05411	0.00201
		MSE	0.03920	0.01013	0.11450	0.01299	0.26571	0.01984	0.00803
	β_1	M	0.85970	0.97135	0.74449	0.93699	0.36957	0.91118	0.99901
		B	0.14030	0.02865	0.25551	0.06301	0.63043	0.08882	0.00099
		MSE	0.02309	0.00541	0.06957	0.01067	0.66862	0.01103	0.00278
2	β_0	M	0.96523	0.47569	1.19689	0.45695	1.44294	0.42036	0.49679
		B	0.46523	0.02431	0.69689	0.04305	0.94294	0.07964	0.00321
		MSE	0.24201	0.01779	0.52145	0.02795	0.93603	0.05141	0.00866
	β_1	M	0.78208	1.00591	0.66665	1.02926	0.53498	1.09507	0.99966
		B	0.21792	0.00591	0.33335	0.02926	0.46502	0.09507	0.00034
		MSE	0.05254	0.004533	0.11808	0.010219	0.225018	0.03237	0.000973
3	β_0	M	0.96253	0.51677	0.89460	0.52493	-1.32639	0.49710	0.49929
		B	0.46253	0.01677	0.39460	0.02493	1.82639	0.00290	0.00071
		MSE	9.21236	0.02860	24.22310	0.05602	394.31439	0.11380	0.00005
	β_1	M	0.94979	0.99035	0.94093	0.98762	1.07001	0.99369	1.00005
		B	0.05021	0.00965	0.05907	0.01238	0.07001	0.00631	0.00005
		MSE	0.06516	0.00159	0.16105	0.00312	1.99271	0.00625	0.0000003
4	β_0	M	0.54943	0.45419	0.64043	0.45594	0.63387	0.42702	0.50781
		B	0.04943	0.04581	0.14043	0.04406	0.13387	0.07298	0.00781
		MSE	3.69577	0.02795	4.11726	0.04596	7.47841	0.09137	0.00141
	β_1	M	0.96577	0.99674	0.93503	0.99668	0.89668	1.02188	0.99930
		B	0.03423	0.00326	0.06497	0.00332	0.10332	0.02188	0.00070
		MSE	0.039523	0.005321	0.055759	0.010717	0.108898	0.02443	$2.22e$-05

Table 4.2: Accuracy and sensitivity of estimated parameters, β_0 and β_1 under different true and unobserved distributions of X for Poisson regression model with FSGN as misspecified exposure model

Sample size n = 50

			R = 0.25		R = 0.5		R = 1		
Simulation setting	Parameter		Naive	Flexible	Naive	Flexible	Naive	Flexible	Benchmark
1	β_0	M	0.70519	0.48850	0.85549	0.51497	1.04232	0.53904	0.53171
		B	0.20519	0.01150	0.35549	0.01497	0.54232	0.03904	0.03171
		MSE	0.06877	0.06877	0.15480	0.03714	0.32303	0.04881	0.02186
	β_1	M	0.84786	0.97271	0.72825	0.94794	0.57743	0.92383	0.98292
		B	0.15214	0.02729	0.27175	0.05206	0.42257	0.07617	0.01708
		MSE	0.03277	0.03277	0.08454	0.01949	0.18908	0.03174	0.00620
2	β_0	M	0.97564	0.50587	1.19514	0.50468	1.43278	0.48038	0.51605
		B	0.47564	0.00587	0.69689	0.00468	0.93278	0.01962	0.01605
		MSE	0.27722	0.04063	0.53430	0.04874	0.90996	0.06465	0.03027
	β_1	M	0.77464	0.97875	0.66106	0.98216	0.53003	1.00596	0.99408
		B	0.22536	0.02125	0.33894	0.01784	0.46997	0.00596	0.00592
		MSE	0.06169	0.00769	0.12560	0.01204	0.22989	0.02364	0.00437
3	β_0	M	0.49727	0.51279	0.26295	0.50127	1.91242	0.41281	0.50125
		B	0.00273	0.01279	0.23705	0.00127	1.41242	0.08719	0.00125
		MSE	7.44074	0.07302	26.7085	0.15161	16.1708	0.32193	0.00023
	β_1	M	0.99619	0.99143	0.99972	0.99288	0.82313	1.01339	0.99980
		B	0.00381	0.00857	0.00028	0.00712	0.17687	0.01339	0.00020
		MSE	0.07038	0.01749	0.23499	0.02503	0.15807	0.02771	2.23e-6
4	β_0	M	0.14322	0.41364	0.56923	0.40559	0.69690	0.37327	0.49864
		B	0.35678	0.08636	0.06923	0.09441	0.19690	0.12673	0.00136
		MSE	11.8566	0.04720	1.90997	0.06488	3.83394	0.12661	0.00279
	β_1	M	0.98455	0.99748	0.91333	0.99356	0.85350	1.00620	1.00220
		B	0.01545	0.00252	0.08667	0.00644	0.14650	0.00620	0.00220
		MSE	0.13688	0.01760	0.05664	0.018619	0.10391	0.02477	0.00016

Sample size n = 100

			R = 0.25		R = 0.5		R = 1		
Simulation setting	Parameter		Naive	Flexible	Naive	Flexible	Naive	Flexible	Benchmark
1	β_0	M	0.67436	0.48475	0.82081	0.51549	1.00180	0.54884	0.49799
		B	0.17436	0.01525	0.32081	0.01549	0.50180	0.04884	0.00201
		MSE	0.03920	0.01026	0.11450	0.01276	0.26571	0.02008	0.00803
	β_1	M	0.85970	0.97146	0.74449	0.93873	0.60215	0.90109	0.99901
		B	0.14030	0.02854	0.25551	0.06127	0.39785	0.09891	0.00099
		MSE	0.02309	0.00549	0.06957	0.01073	0.16275	0.02041	0.00278
2	β_0	M	0.96523	0.51688	1.19689	0.54484	1.44294	0.52276	0.49679
		B	0.46523	0.01688	0.69689	0.04484	0.94294	0.02276	0.00321
		MSE	0.24201	0.01720	0.52145	0.01763	0.93603	0.04047	0.00866
	β_1	M	0.78208	0.96978	0.66665	0.95076	0.53498	0.96525	0.99966
		B	0.21792	0.03022	0.33335	0.04924	0.46502	0.03475	0.00034
		MSE	0.05254	0.00488	0.11808	0.00517	0.225018	0.01015	0.000973
3	β_0	M	0.96253	0.51442	0.89460	0.51978	-1.32639	0.49203	0.49929
		B	0.46253	0.01442	0.39460	0.01978	1.82639	0.00797	0.00071
		MSE	9.21236	0.02869	24.22310	0.05677	394.31439	0.11628	0.00005
	β_1	M	0.94979	0.99085	0.94093	0.98762	1.07001	0.99464	1.00005
		B	0.05021	0.00915	0.05907	0.01129	0.07001	0.00536	0.00005
		MSE	0.06516	0.00159	0.00315	0.00312	1.99271	0.00639	0.0000003
4	β_0	M	0.54943	0.42863	0.64043	0.43064	0.63387	0.42045	0.50781
		B	0.04943	0.07137	0.14043	0.06936	0.13387	0.07955	0.00781
		MSE	3.69577	0.01964	4.11726	0.02740	7.47841	0.04899	0.00141
	β_1	M	0.96577	1.00375	0.93503	0.99887	0.89668	1.00417	0.99930
		B	0.03423	0.00375	0.06497	0.00113	0.10332	0.00417	0.00070
		MSE	0.039523	0.00298	0.05576	0.00531	0.108898	0.01060	2.22e-05

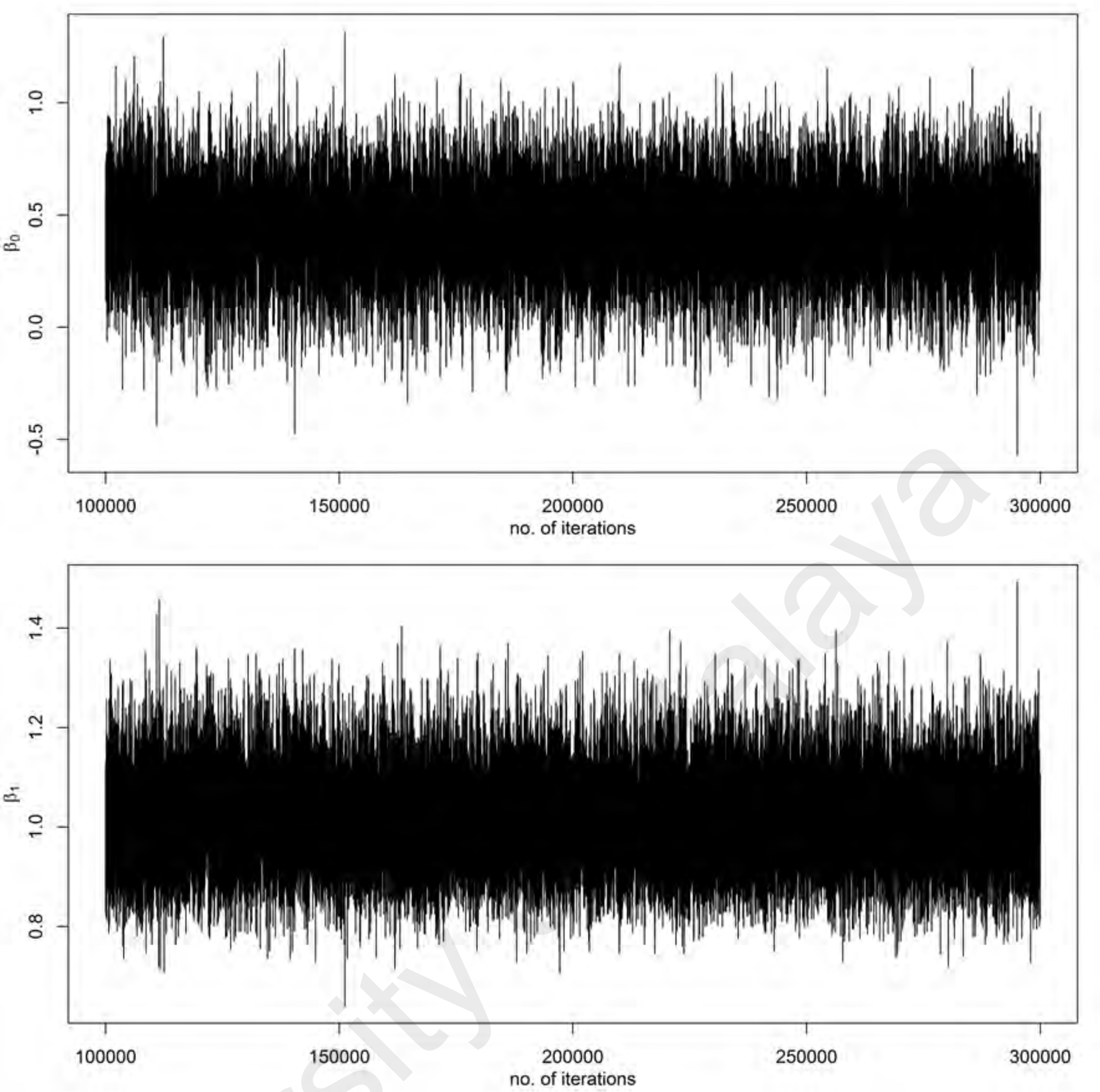

Figure 4.1: Trace plots for estimated regression parameters, $\hat{\beta_0}$ and $\hat{\beta_1}$ in one of the simulation studies

The MSE and bias values depicted in the tables are to demonstrate the bias-variance tradeoff where low values in both bias and MSE are indicators for a good performance in estimating the parameters (Geman et al., 1992). The posterior summaries of our approach after ME correction is labelled as *flexible*. To highlight the performance of our model, we also present the *naive* and *benchmark* estimates. Regression parameters estimates for each data set drawn from a *naive* analysis are obtained when direct regression are applied on the mean between m surrogates, $\bar{X}_i^* = \sum_{j=1}^{m=2} X_{ij}/m$, are taken to be as precisely measured. Meanwhile, in the *benchmark* analysis, we assume that the unobserved true values X_i as known and similarly, apply direct regression to estimate the regression parameters.

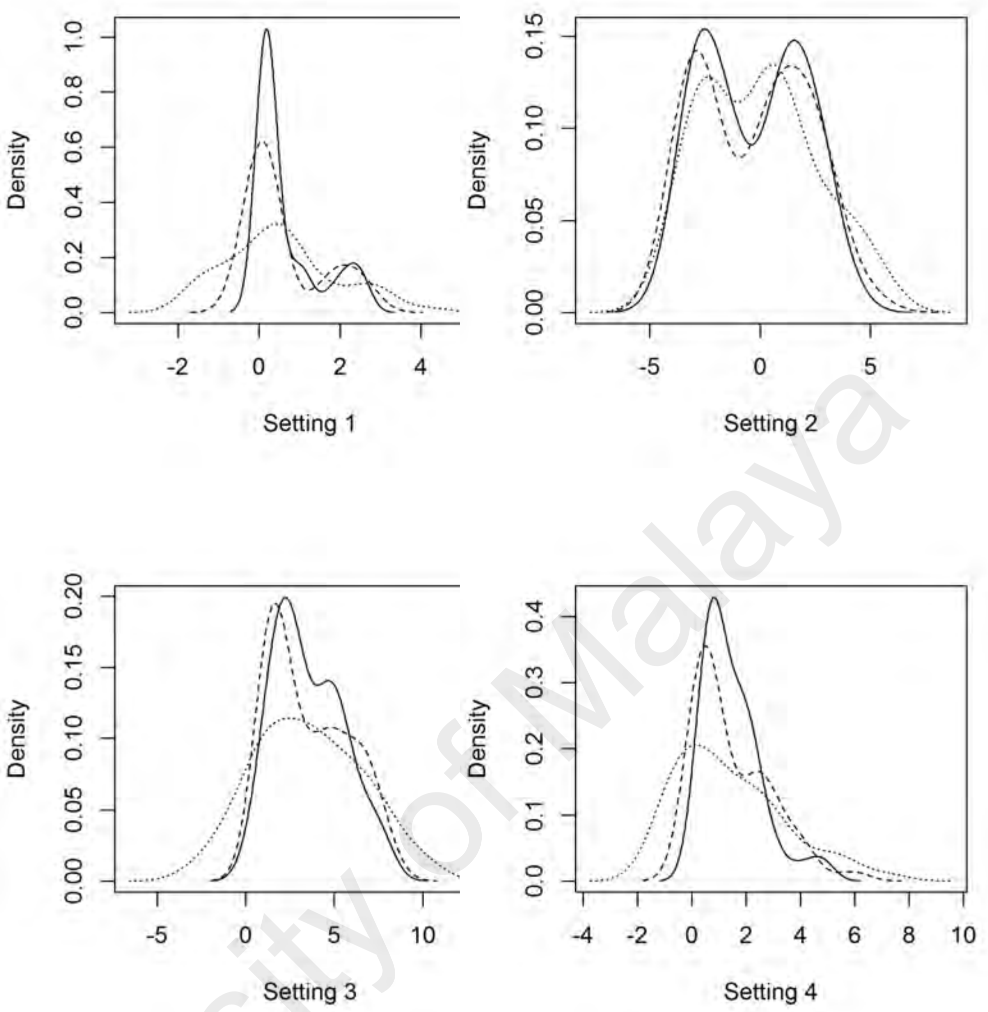

Figure 4.2: Kernel density estimates for settings 1-4 in the case of misspecified FGSN exposure model for EIV in PRM: true exposure X_i (solid curve); estimated X_i under flexible Bayesian approach (dashed curve); mean proxy $\bar{X}_i^*$ (dotted curve).

This is to illustrate how closely our approach performs in terms of bias correction and efficiency in comparison with the ideal (benchmark) situation and how in the absence of bias correction, non-credible estimated values will be reached.

To clearly visualize the effects of our EIV correction using the Bayesian approach, we plot the posterior kernel densities of estimated X_i, that have been corrected for ME using the flexible Bayesian approach from a randomly selected data with FGSN and FSGN exposure model, respectively, as shown in Figures 4.2 and 4.3. As a comparison, we also construct the kernel densities of their corresponding true exposure variables X_i and mean

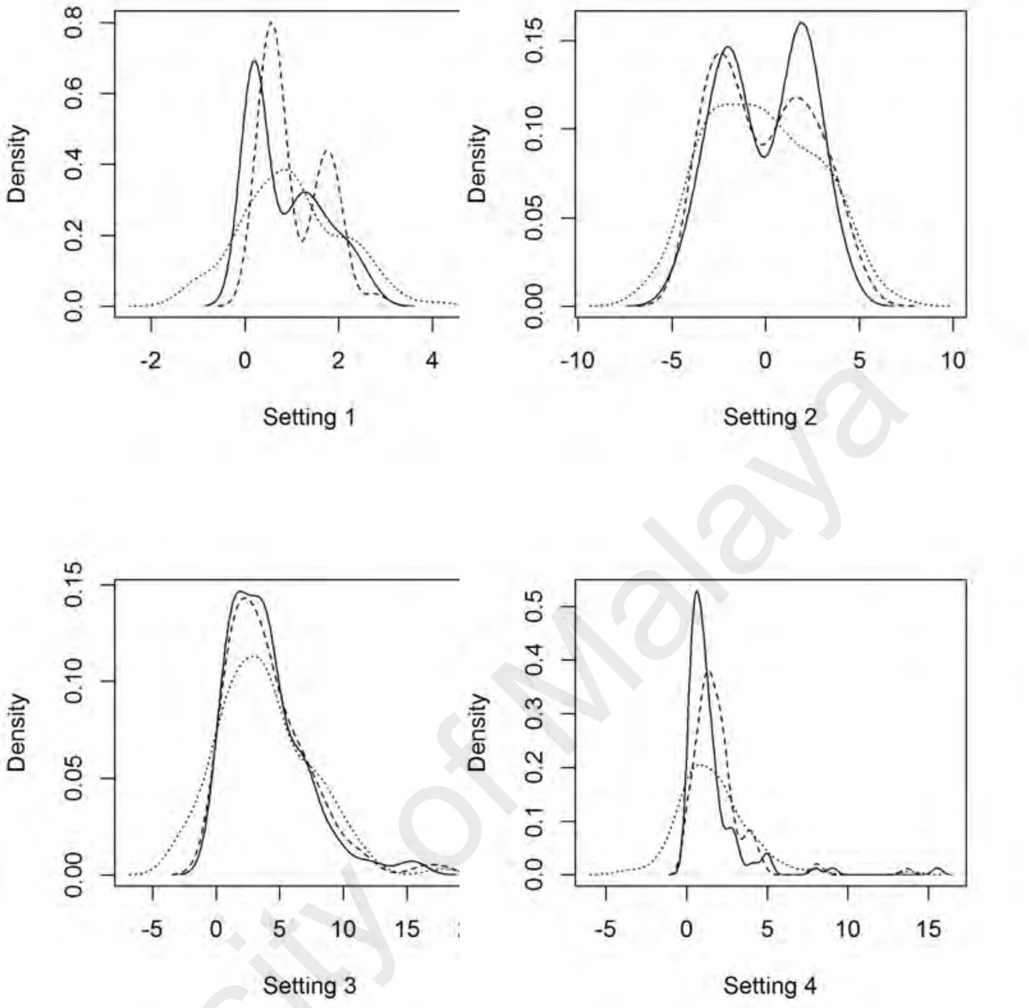

Figure 4.3: Kernel density estimates for settings 1-4 in the case of misspecified FSGN exposure model for EIV in PRM: true exposure X_i (solid curve); estimated X_i under flexible Bayesian approach (dashed curve); mean proxy $\bar{X}_i^*$ (dotted curve).

proxy $\bar{X}_i^*$. The randomly selected data set has sample size $n = 100$ and $R = 1.0$ level of ME contamination. This comparison is to further highlight the performance of our model and also to illustrate the ability of the flexible Bayesian approach with the usage of FGSN and FSGN as the misspecified exposure model to capture the shape of the unknown exposure distribution in each simulation configurations.

The discussion of the results are separated into two subsections; Subsection 4.7.1 discusses the results when using FGSN exposure model, meanwhile Subsection 4.7.2 discusses the results when using FSGN exposure model.

4.7.1 Using FGSN Exposure Model

The results presented in Table 4.1 demonstrate that the flexible Bayesian approach with misspecified exposure model using FGSN distribution does very well in attenuating bias when estimating the unknown true regression parameters under distributions that exhibit skewness, bimodality, heavy-tailedness and even in the case of both skewed and heavy-tailed exposures; their values follow closely to the values of the *benchmark* estimates. The *naive* estimates under every simulation settings and sample sizes have significantly heavy bias and do poorly in terms of estimating the correct values of β_0 and β_1. Under certain simulation settings, when comparing in terms of ME contamination level, R, the larger R may yield smaller mean bias. This is most probably due to simulation error.

When estimating parameters using MCMC, a good measure of performance would be the bias-variance tradeoff where the two sources of error; bias and variance need to be minimized. In Table 4.1 under sample size $n = 50$, we see that for the 3rd simulation setting in the case of FGSN exposure model, the mean bias of *naive* β_0 estimate for $R = 0.25$ is smaller than the mean bias of our *flexible* β_0. However, not surprisingly *naive* estimate reports substantially high MSE value which implies that using the *naive* approach yield highly inconsistent values between the 50 data sets and therefore performs very poorly in terms of bias-variance trade-off. In this case, even though our approach shows slightly bigger bias than that of the *naive* estimates, the MSE values suggest that our proposed approach still yields better performance as higher flexibility may sacrifice accuracy according to Ma and Genton (2004).

As shown in Figure 4.2, we see that in each setting: skewed (setting 1), bimodal (setting 2) and skewness paired with heavy-tailedness (settings 3 and 4), the kernel density of our corrected X_i follows closely to the kernel density shape of true X_i distribution. Meanwhile, $\bar{X}_i^*$ gives a very blurred kernel density shape under every simulation setting.

4.7.2 Using FSGN Exposure Model

Similar results are reported for when the exposure model is misspecified using FSGN as shown in Table 4.2. There is also a significant difference between the *naive* estimates and the *flexible* estimates such that the latter have closer values to the *benchmark* estimates. This show that using FSGN, the approach is also successful in estimating the values of the unknown true regression distributions at every simulation settings. Here, we also see that in certain simulation settings, the mean bias of *flexible* β_0 estimate for smaller R is larger than that of the bigger R which is also may be the result from simulation error. The low MSEs also imply that the flexible Bayesian approach with FSGN as its exposure model has a good bias-variance tradeoff despite the model being more flexible than FGSN.

Figure 4.3 shows that under each setting, the kernel density of our corrected X_i follows closely to the kernel density shape of true X_i distribution. Meanwhile, $\bar{X}_i^*$ gives a very blurred kernel density shape under every simulation setting.

4.7.3 Comparing the Performance between FGSN and FSGN as the Misspecified Exposure Model for EIV PRM.

Using the same exact simulation settings and the same exact number of iterations and burn-ins, the results of parameter regression estimates, β_0 and β_1, with adjustment to bias report similar results under both FGSN and FSGN exposure model as represented in Table 4.1 and Table 4.2 where both perform well in reducing bias caused by EIV. However, we shall compare the performance of FGSN and FSGN as the misspecified exposure model to find which of the two flexible models yield better bias reduction.

To paint a clearer picture the difference of performance between FGSN and FSGN, we provide a visual comparison. In Figures 4.4 to 4.7, the kernel densities of estimated β_0 and β_1 with $R = 1.0$ and $n = 100$ under each simulation setting for FGSN and FSGN are compared. We let the solid curve to depict the kernel density of *benchmark* estimates, the dashed curve to depict the kernel density plot of estimates under FGSN exposure model

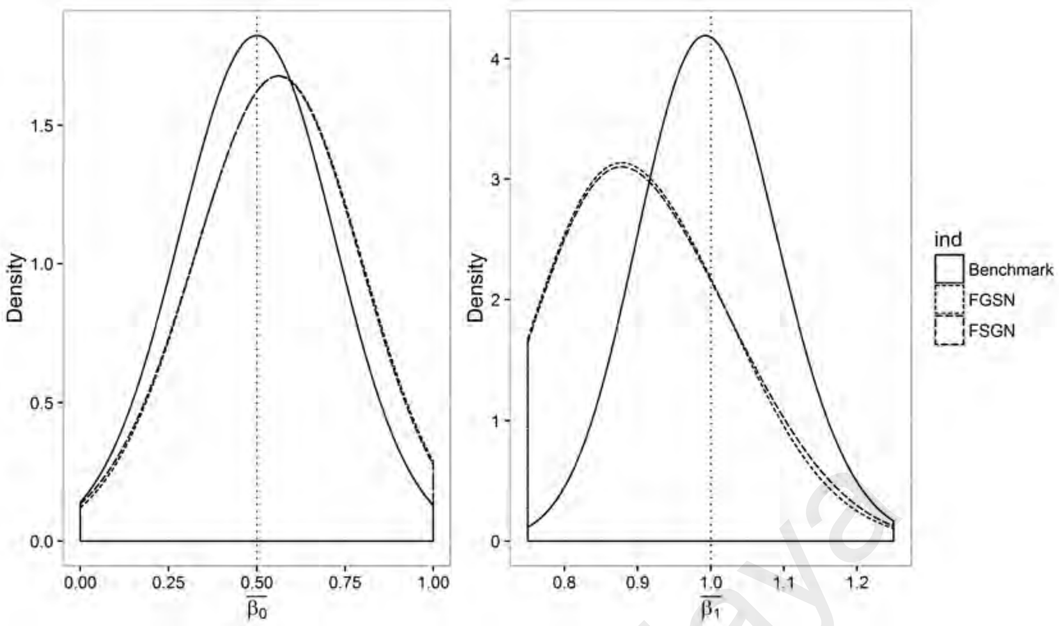

Figure 4.4: Kernel density of estimated regression parameters under simulation setting 1 - Skewed mixture of normal distribution: Benchmark (solid curve); FGSN (dashed curve); FSGN (long-dashed curve).

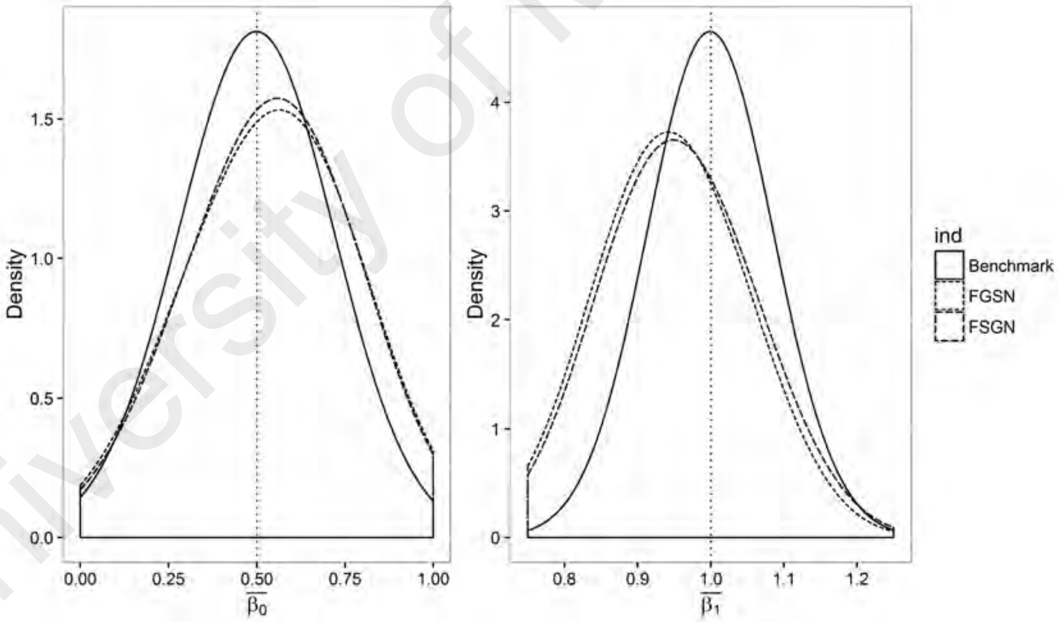

Figure 4.5: Kernel density of estimated regression parameters under simulation setting 2 - Bimodal mixture of normal distribution: Benchmark (solid curve); FGSN (dashed curve); FSGN (long-dashed curve).

and the long-dashed curve to depict the kernel density plot of estimates under FSGN exposure model.

Figure 4.4 depicts the parameter estimates under simulation settings 1 for when the true exposures are generated from the skewed mixture of normal distribution. In this

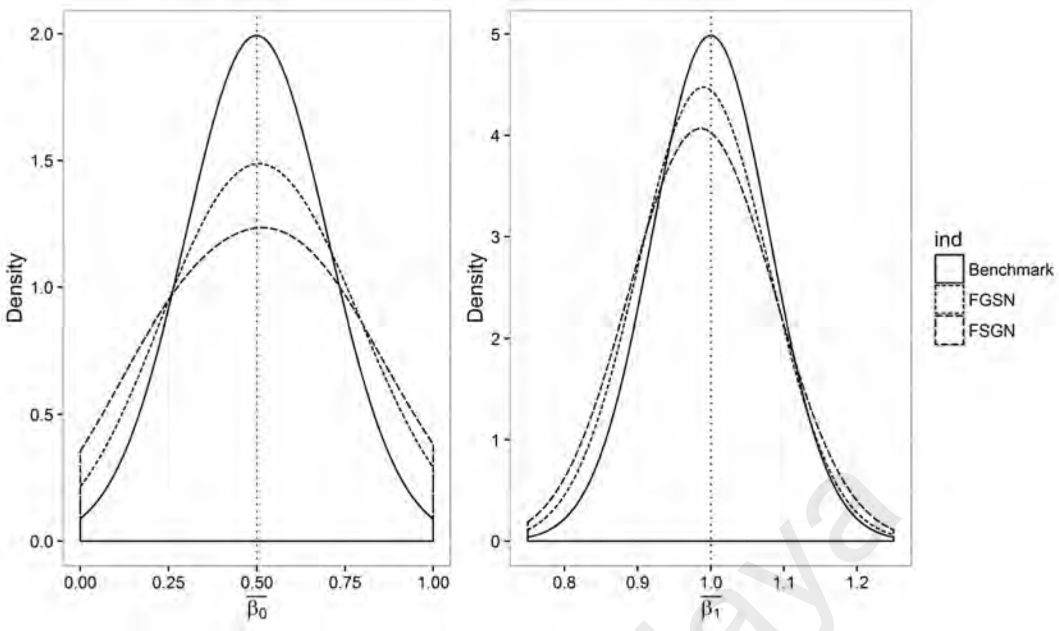

Figure 4.6: Kernel density of estimated regression parameters under simulation setting 3 - Gamma distribution: Benchmark (solid curve); FGSN (dashed curve); FSGN (long-dashed curve).

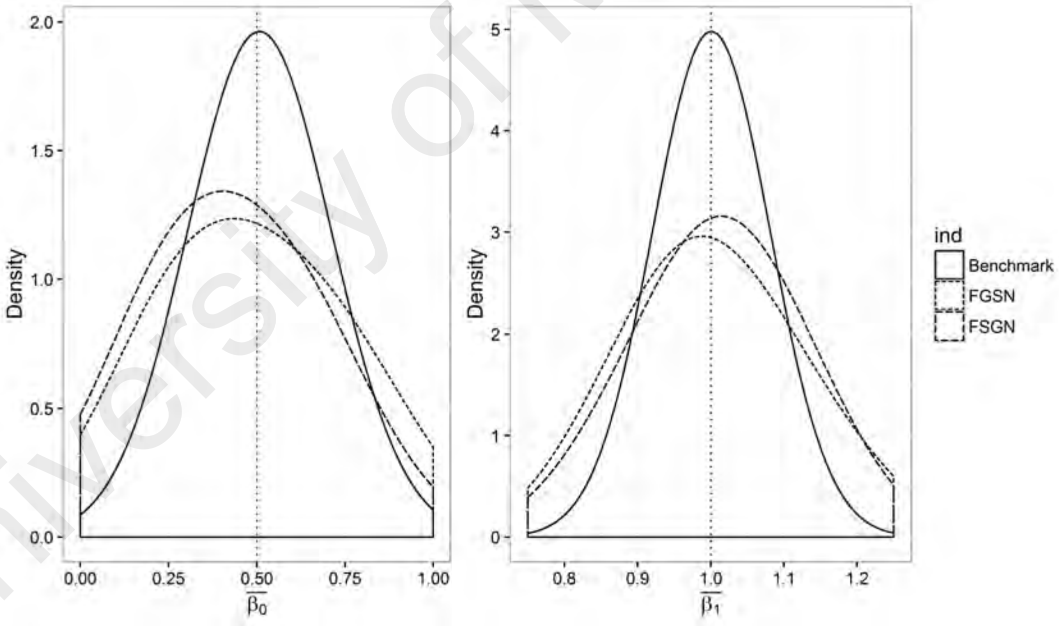

Figure 4.7: Kernel density of estimated regression parameters under simulation setting 4 - Log-normal distribution: Benchmark (solid curve); FGSN (dashed curve); FSGN (long-dashed curve).

setting, the kernel density plot estimates, β_0 and β_1 for FGSN and FSGN models have very similar kernel densities such that both of the densities are almost perfectly overlapping with each other. Therefore, both flexible exposure models have almost the same performance in correcting bias for skewed true exposures.

Almost the same result is reported in kernel density estimate plots for simulation setting 2 where the true exposures are generated from the bimodal mixture of normal distribution. As shown in Figure 4.5, the flexible models FGSN and FSGN have similar shapes such that none actually showed that it has a significantly better performance than the other.

As for the kernel densities depicted in Figure 4.6 for both parameters β_0 and β_1, FGSN exposure model shows better performance than that of FSGN exposure model. As can clearly be seen in the figure, FGSN model tracks better kernel density shape of the *benchmark* estimates than the corresponding kernel density estimates of FSGN model. The position of the peak under FGSN model is much closer to the true value of β_0 and β_1 and the spread of the kernel densities also follows much closer to the *benchmark* estimates than under the FSGN model for sample size $n = 100$. However, for $n = 50$ if we compare the values given in the tables above, for simulation setting 3, FSGN shows a slightly better performance; but the difference in bias and MSE between the two models does not really have a profound difference.

In Figure 4.7, the true exposures are generated from simulation setting 4 which is a log-normal distribution that has a heavy-tail. In the figure shown in this simulation setting for parameter β_0, FGSN exposure model shows better performance than its corresponding FSGN exposure model. As can be seen in the kernel density plots, the peak for FGSN model is much closer to the true value, 0.5 in comparison to the peak of FSGN exposure model. As for the spread, in our observation, both flexible exposure models do not report any significant difference in their kernel density plots. For kernel density plot estimates of the parameter β_1, the performance of both FGSN and FSGN models are very similar although one can argue that the peak of the kernel density plot for FGSN model is closer to the true value, 1.0, than the kernel density plot of FSGN model.

Therefore, from the findings shown in Table 4.1 and Table 4.2, as well as the comparison between the kernel density plots of FGSN and FSGN misspecified exposure model for the 4 simulation settings, our approach when using both models show similar performance. However, FGSN should be the preferred flexible model as it shows more efficiency than the FSGN model. This is because the extra parameter in FSGN makes for a slower and longer MCMC simulation time. Although FSGN offers more flexibility than FGSN, its performance, however, showed no significant increase. Therefore, for Poisson regression outcome model, we advocate the usage of FGSN as the misspecified exposure model.

4.7.4 Non-normal Distribution of EIV

Now, to test for the robustness of normal distribution as the measurement model, we generate the ME, ϵ_j, from two types non-normal distributions, the skew-normal (SN) and skew-*t* (ST) distribution. In technical terms, the first non-normal error is generated from $\epsilon_j \sim SN(0,1)$ and the second non-normal error is generated from $\epsilon_j \sim ST(0,1)$. Also, X_i is generated from skewed mixture of normal and the contamination of error is taken as $R = 1$ indicates a high and substantial ME. Since FGSN is the preferred model as discussed earlier, we use FGSN as the misspecified flexible exposure model.

Table 4.3: Estimated values of β_0 and β_1 of EIV PRM where EIV is generated from skew-normal and skew-*t* distributions.

Distribution of EIV	Parameter		Naive	Flexible	Benchmark
Skew-normal	β_0	M	0.98672	0.54489	0.49799
		B	0.48873	0.04690	0.00201
		MSE	0.24804	0.01666	0.00820
	β_1	M	0.63134	0.93795	0.99901
		B	0.36767	0.06107	0.00099
		MSE	0.13783	0.01462	0.0028
Skew-*t*	β_0	M	1.14211	0.54443	0.49799
		B	0.64412	0.04644	0.00201
		MSE	0.43381	0.02427	0.00820
	β_1	M	0.50778	0.92632	0.99901
		B	0.49123	0.07269	0.00099
		MSE	0.24864	0.02222	0.0028

The results are shown in Table 4.3. As depicted in the table, even when EIV departed from normality, normal distribution as the measurement model still provides robustness and there is no deterioration in bias correction for the Poisson regression outcome model. So, even though there are some studies that suggested the use of flexible distribution not only for the exposure model but also the measurement model, we, however, considered it as redundant following from the results of our simulation studies.

CHAPTER 5: BAYESIAN APPROACH TO ERRORS-IN-VARIABLES IN NEGATIVE BINOMIAL REGRESSION MODEL

5.1 Introduction

In the previous chapter, we have discussed and investigated the flexible Bayesian method to correct errors-in-variables (EIV) in Poisson regression. Although Poisson is the most popular model for count data, sometimes the data are overdispersed in which Poisson regression may no longer be used to model the data. In a count data set where the variance is larger than the mean, negative binomial regression model (NBRM) should be employed to model it. In current studies, there were no usage of flexible distributions such that the exposure model assumes a flexible distribution. Therefore, in this chapter we propose the usage of Bayesian approach to address bias caused by EIV in an overdispersed count data regression model, that is NBRM. By intentionally misspecifying the flexible models as the exposure model, we are able to implement a general framework even when the non-normal distribution used in every simulation settings are different (i.e, skewness, bimodality and heavy-tailedness).

5.2 Negative Binomial Regression Outcome Model

Using similar notations as in the previous chapter, we denote the outcome variable as Y_i, the true but unobserved exposure variable as X_i and its corresponding observed with error exposure variable as X_i^*. NBRM denoted by $Y_i \sim NB(r, \exp(\beta_0 + \beta_1 X_i))$ has a dispersion parameter $r > 0$ and mean parameter $\exp(\beta_0 + \beta_1 X_i)$. Thus, we shall specify the outcome model as NBRM with the following pmf,

$$f(Y_i|X_i, \boldsymbol{\theta}_{NBRM}) = \frac{\Gamma(Y_i + r)}{Y_i!\Gamma(r)} \left(\frac{r}{r + \exp(\beta_0 + \beta_1 X_i)} \right)^r \left(\frac{\exp(\beta_0 + \beta_1 X_i)}{r + \exp(\beta_0 + \beta_1 X_i)} \right)^{Y_i}, \qquad (5.1)$$

where $\boldsymbol{\theta}_{NBRM} = (\beta_0, \beta_1, r)$ and it follows that

$$E(Y_i|\boldsymbol{\theta}_{NBRM}) = \exp(\beta_0 + \beta_1 X_i), \quad \text{and}$$
$$Var(Y_i|\boldsymbol{\theta}_{NBRM}) = \exp(\beta_0 + \beta_1 X_i)\left(1 + \frac{\exp(\beta_0 + \beta_1 X_i)}{r}\right).$$

It is clear that since $\exp(\beta_0 + \beta_1 X_i) > 0$, then overdispersed count data can be modelled by NBRM.

5.3 Measurement Model

In this chapter, we also specify a normal distribution as the measurement model. The extensive simulation studies conducted in the previous chapter suggest that normal distribution is robust enough to be specified as the measurement model distribution even when the distribution of ME has departures from normality. So, the pdf is given by

$$f(X_{ij}^*|X_i, \boldsymbol{\theta}_M) = \left(\frac{1}{2\pi\tau^2}\right)^{1/2} \exp\left(-\frac{1}{2\tau^2}\left(X_{ij}^* - X_i\right)^2\right), \qquad (5.2)$$

such that $\boldsymbol{\theta}_M = \tau^2$ and X_{ij}^* signifies the j^{th} replicated surrogate of i^{th} observation of $\boldsymbol{X}^*$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

5.4 Bayesian Approach using Flexible Exposure Model

In the next section, we shall describe the usage of flexible Bayesian approach to correct EIV in NBRM with the exposure model misspecified with a flexible model. Again, even though the distribution of the true exposures X_i is generated according to its simulation setting, we will intentionally misspecify the exposure model with FGSN distribution such that, $X_i|\boldsymbol{\theta}_{FGSN} \sim FGSN(\alpha, \lambda^2, \omega_1, \omega_2)$. The pdf of the FGSN is the same as the one given in Section 4.4.

Besides that, we shall also thoroughly describe our study on correcting EIV in

NBRM with FSGN as its misspecified model. In technical terms, we set $X_i|\boldsymbol{\theta}_{FSGN} \sim FSGN(\alpha, \lambda_1^2, \lambda_2, \omega_1, \omega_2)$ such that its pdf is given in Section 4.4.2.

5.5 Joint Posterior Density

5.5.1 Flexible Bayesian Approach under FGSN exposure model

With NBRM as the outcome model, normal distribution as the measurement model and FGSN as the misspecified exposure model, we can now construct the joint posterior which is the product of these three submodels. Using Richardson and Gilks (1993) framework of the Bayesian approach to correct EIV, we can write the joint posterior density as,

$$f(\boldsymbol{X}, \boldsymbol{\theta}|\boldsymbol{X}^*, \boldsymbol{Y}) \propto \prod_{i=1}^{n} f(Y_i|X_i, \boldsymbol{\theta}_{NBRM}) \prod_{i=1}^{n} \prod_{j=1}^{m} f(X_{ij}^*|X_i, \boldsymbol{\theta}_M) \prod_{i=1}^{n} f(X_i|\boldsymbol{\theta}_{FGSN}) \times \pi(\boldsymbol{\theta}). \tag{5.3}$$

Let $\boldsymbol{\theta}$ be the parameter vector of the model that contains $\boldsymbol{\theta}_{NBRM}, \boldsymbol{\theta}_M$ and $\boldsymbol{\theta}_{FGSN}$ which denote vectors of parameters for outcome, measurement and FGSN exposure model, respectively.

Unlike when the outcome model is PRM, in the case of NBRM we do not have to introduce a latent variable as it already have a quite fast convergence rate and low bias for the parameter $\boldsymbol{\beta}$ as observed in our simulation studies. Using similar notations, we let $\pi(\boldsymbol{\theta})$ represent the prior distribution of our parameter vector, where $\boldsymbol{\theta}$ contains $\boldsymbol{\beta}, r, \tau^2, \alpha, \lambda^2, \omega_1, \omega_2$ such that $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ is the main parameter vector that we want to estimate. We assume priori independence and thus, the joint distribution for all of the priori is given as,

$$\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\beta})\pi(r)\pi(\tau^2)\pi(\alpha)\pi(\lambda^2)\pi(\omega_1)\pi(\omega_2).$$

We assign a weakly informative prior for the parameter $\boldsymbol{\beta}$ such that it follows a normal distribution with high variance, $N(\mathbf{0}, 10^2 \cdot \boldsymbol{I}_2)$ where $\boldsymbol{I}_2$ denotes identity matrix of order 2. The parameter r, which is the dispersion parameter of NBRM, needs to maintain its positive support. So, taking this into account, we set its prior distribution as $IG(0.5, 0.5)$. As alluded in Section 4.5.1, the reason why IG is chosen with its shape and scale parameter are both 0.5 is to ensure that the prior that we use is as close to non-informative as possible. This is because, without enough knowledge on the values of r, it is unreasonable to set a prior that will have an influence on its construction. In other words, we want the data to take the main role in the posterior distribution. As for the parameter α, we assign a common choice of flat prior distribution, that is, one (Box & Tiao, 2011). The prior distribution for parameters λ^2 and τ^2 is also $IG(0.5, 0.5)$, recommended by Gelman et al. (2014). The choice of prior follows the same logic as when we assign the same prior distribution to r, which is to stay as close to non-informative as possible (Gelman et al., 2014). For parameters ω_1, ω_2 we let both of their prior distributions to be $N(0, 10^2)$.

We rewrite Equation (5.3) and the posterior density is now written as the following,

$$
\begin{aligned}
f(\boldsymbol{X}, \boldsymbol{\theta} | \boldsymbol{X}^*, \boldsymbol{Y}) \propto \prod_{i=1}^{n} & \frac{\Gamma(Y_i + r)}{Y_i! \Gamma(r)} \left(\frac{r}{r + \exp(\beta_0 + \beta_1 X_i)} \right)^r \left(\frac{\exp(\beta_0 + \beta_1 X_i)}{r + \exp(\beta_0 + \beta_1 X_i)} \right)^{Y_i} \\
& \times \left[\prod_{j=1}^{m} \left(\frac{1}{\tau^2} \right)^{1/2} \exp\left(- \frac{1}{2\tau^2} (X_{ij}^* - X_i)^2 \right) \right] \\
& \times \left[\left(\frac{1}{\lambda^2} \right)^{1/2} \exp\left(- \frac{1}{2\lambda^2} (X_i - \alpha) \right) \right] \Phi \left[\left(\frac{\omega_1 (X_i - \alpha)}{\lambda} \right) + \left(\frac{\omega_2 (X_i - \alpha)^3}{\lambda^3} \right) \right] \\
& \times \pi(\boldsymbol{\beta}) \pi(r) \pi(\tau^2) \pi(\alpha) \pi(\lambda^2) \pi(\omega_1) \pi(\omega_2),
\end{aligned}
\tag{5.4}
$$

where $\Phi(.)$ is the standard normal distribution function.

Conditional Posterior Density

In the this subsection, we shall use Equation (5.4) to derive the conditional posterior density for each of the parameters in our model. The conditional posterior density of the parameters are then reparametrised into closed forms (if possible). Using MCMC sampling method, we shall estimate the parameters. Let A^C be the complement of the parameter A.

MCMC Implementation

i. For $\boldsymbol{\beta}$,

$$f(\boldsymbol{\beta}|\boldsymbol{\beta}^C) \propto \prod_{i=1}^{n} \left\{[\exp(\beta_0 + \beta_1 X_i)]^{Y_i}(\exp(\beta_0 + \beta_1 X_i) + r)^{-(r+Y_i)}\right\} \times \prod_{k=0}^{1} \exp\big(- \frac{\beta_k^2}{2 \times 10^2}\big).$$

For updating $\boldsymbol{\beta}$ in NBR outcome model and FGSN exposure model, since the posterior distribution does not follow any known distribution, we propose $\boldsymbol{\beta}$ to be sampled using RWMH sampling method with normal distribution as its proposal distribution, $N(0, k_\beta^2)$ such that k_β is the tuning parameter. We choose $k_\beta = 0.02$ such that the tuning parameter will yield acceptance rate between 25% and 30%.

ii. For X_i,

$$\begin{aligned} f(X_i|X_i^C) \propto &\left\{[\exp(\beta_0 + \beta_1 X_i)]^{Y_i}(\exp(\beta_0 + \beta_1 X_i) + r)^{-(r+Y_i)}\right\} \\ &\exp\left\{ - \frac{1}{2\sigma_X^2}(X_i - \mu_X)^2\right\}\left\{\Phi\left(\frac{\omega_1(X_i - \alpha)}{\lambda} + \frac{\omega_2(X_i - \alpha)^3}{\lambda^3}\right)\right\}, \end{aligned}$$

where the first component of the conditional posterior follows normal distribution with mean $\mu_X = (\alpha\tau^2 + m\lambda^2\bar{X})/(\tau^2 + m\lambda^2)$ and variance $\sigma_X^2 = \lambda^2\tau^2/(\tau^2 + m\lambda^2)$. Hence, we update X_i by component using MH algorithm with univariate normal proposal distribution of mean μ_X and variance σ_X^2.

iii. For α,

$$f(\alpha|\alpha^C) \propto \exp\left\{-\frac{n}{2\lambda^2}(\alpha-\bar{X})^2\right\}\left\{\prod_{i=1}^{n}\Phi\left(\frac{\omega_1(X_i-\alpha)}{\lambda}+\frac{\omega_2(X_i-\alpha)^3}{\lambda^3}\right)\right\},$$

where the first component of the conditional posterior distribution is a normal distribution with mean $\bar{X}$ and variance λ^2/n. In our simulation studies, we use RWMH scheme to update α with tuning parameter k_α where $k_\alpha = 1$, and the proposal distribution is $N(0, k_\alpha^2\lambda^2/n)$. The choice of tuning parameter will give us acceptance rate between 25% and 40%.

iv. For ω_h where $h = 1, 2$,

$$f(\omega_h|\omega_h^C) \propto \left\{\prod_{i=1}^{n}\Phi\left(\frac{\omega_1(X_i-\alpha)}{\lambda}+\frac{\omega_2(X_i-\alpha)^3}{\lambda^3}\right)\right\}\exp\left\{-\frac{\omega_h^2}{2\times 100}\right\}.$$

For ω_h where $h = 1, 2$, we sample using RWMH method with $N(0, k_\omega^2)$ as the proposal distribution and we set the tuning parameter k_ω as 0.09 which will yield acceptance rate between 25% and 30%.

v. For τ^2,

$$f(\tau^2|\tau^{2^C}) \propto \left(\frac{1}{\tau^2}\right)^{\frac{mn+1}{2}+1}\exp\left[-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(X_{ij}^*-X_i)^2+1}{2\tau^2}\right],$$

is a closed form distribution, namely IG with shape and scale parameter $(mn+1)/2$ and $\sum_{i=1}^{n}\sum_{j=1}^{m}0.5(X_{ij}^*-X_i)^2+0.5$, respectively. Therefore, we use Gibbs sampler to update τ^2.

vi. For λ^2,

$$f(\lambda^2|\lambda^{2^C})$$

$$\propto \left(\frac{1}{\lambda^2}\right)^{\frac{n+1}{2}+1} \exp\left[-\frac{0.5}{\lambda^2}\Big(\sum_{i=1}^{n}(X_i-\alpha)^2+\Big)\right]\left\{\prod_{i=1}^{n}\Phi\Big(\frac{\omega_1(X_i-\alpha)}{\lambda}+\frac{\omega_2(X_i-\alpha)^3}{\lambda^3}\Big)\right\},$$

where as we can see above, the first component of the conditional posterior is IG with shape $(n+1)/2$ and scale $\sum_{i=1}^{n} 0.5(X_i-\alpha)^2+0.5$. Hence for both count data regression models, using MH algorithm, λ^2, we use proposal from $IG((n+1)/2, \sum_{i=1}^{n} 0.5(X_i-\alpha)^2+0.5)$.

vii. For r,

$$f(r|r^C) \propto \left(\frac{r^r}{\Gamma(r)}\right)^n \exp(-0.5r)\prod_{i=1}^{n}\left[\Gamma(Y_i+r)(\exp(\beta_0+\beta_1 X_i)+r)^{-(r+Y_i)}\right].$$

Since the conditional posterior for r, as shown above, does not follow any known distribution, we apply the MH algorithm and use the exponential distribution with rate 0.5 as the proposal distribution.

5.5.2 Flexible Bayesian Approach under FSGN Exposure Model

Using NBRM as the outcome model, normal distribution as the measurement model and FSGN as the intentionally misspecified exposure model, we construct the joint posterior density of EIV NBRM which is the product of all the three models mentioned before. The joint posterior density may be observed as the following,

$$f(\boldsymbol{X}, \boldsymbol{\theta}|\boldsymbol{X}^{*}, bmY) \propto \prod_{i=1}^{n} f(Y_i|X_i, \boldsymbol{\theta}_{NBRM}) \prod_{i=1}^{n} \prod_{j=1}^{m} f(X_{ij}^{*}|X_i, \boldsymbol{\theta}_M) \prod_{i=1}^{n} f(X_i|\boldsymbol{\theta}_{FSGN}) \times \pi(\boldsymbol{\theta}). \tag{5.5}$$

where $\boldsymbol{\theta}$ is the parameter vector of the model that contains $\boldsymbol{\theta}_{NBRM}, \boldsymbol{\theta}_M$ and $\boldsymbol{\theta}_{FSGN}$ which denote vectors of parameters for outcome, measurement and FSGN exposure model, respectively.

Letting the prior distribution denoted as $\pi(\boldsymbol{\theta})$ be independent and $\boldsymbol{\theta} = (\boldsymbol{\beta}, r, \tau^2, \lambda_1^2, \lambda_2, \omega_1, \omega_2)$, the joint distribution of all the priori on the parameters considered is,

$$\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\beta})\pi(r)\pi(\tau^2)\pi(\alpha)\pi(\lambda_1^2)\pi(\lambda^2)\pi(\omega_1)\pi(\omega_2).$$

We set the prior for parameters $\boldsymbol{\beta}, \omega_1$ and ω_2 to be a normal distribution with mean 0 and variance, 10^2. Meanwhile, α has a flat prior distribution. Since r has a positive support, we let its prior to be exponential with rate one. Following Gelman et al. (2014) where for the scale parameters, IG is proposed as prior, the prior distribution for λ_1^2 and τ^2 both follow $IG(0.5, 0.5)$. As for the scale parameter λ_2, its prior distribution is given by half-normal with variance one (Gelman, 2006).

We rewrite Equation (5.5) as the following:

$$
\begin{aligned}
f(X,\boldsymbol{\theta}|X^{*},Y) \propto \prod_{i=1}^{n} &\left\{ \frac{\Gamma(Y_i+r)}{Y_i!\Gamma(r)} \left(\frac{r}{r+\exp(\beta_0+\beta_1 X_i)} \right)^{r} \left(\frac{\exp(\beta_0+\beta_1 X_i)}{r+\exp(\beta_0+\beta_1 X_i)} \right)^{Y_i} \right. \\
&\times \left[\prod_{j=1}^{m} \left(\frac{1}{\tau^2} \right)^{1/2} \exp\left(-\frac{1}{2\tau^2} (X_{ij}^{*}-X_i)^2 \right) \right] \\
&\times \left. \left[\left(\frac{1}{\lambda_1^2} \right)^{1/2} \exp\left(-\frac{1}{2\lambda_1^2} (X_i-\alpha) \right) \right] \Phi\left(\frac{\omega_1(X_i-\alpha)+\omega_2(X_i-\alpha)^3/\lambda_1^2}{\sqrt{\lambda_1^2+\lambda_2(X_i-\alpha)^2}} \right) \right\} \\
&\times \pi(\boldsymbol{\beta})\pi(\tau^2)\pi(\lambda_1^2)\pi(\lambda_2)\pi(\alpha)\pi(\sigma^2)\pi(\omega_1)\pi(\omega_2),
\end{aligned}
\tag{5.6}
$$

We construct the conditional posterior density of all the parameters in the model from Equation (5.6). If possible, we shall provide the conditional distributions in closed form.

5.5.3 Conditional Posterior Density

In the case of NBR outcome model, under the FSGN exposure model, the parameters $\boldsymbol{\beta}$, τ^2 and r has the same posterior conditional densities as the ones in Subsection 5.5.1 under the FGSN exposure model. Therefore, in this section, we will elaborate on the conditional posterior density derived from Equation (5.6) and the MCMC methods used to update the parameters $\alpha, \lambda_1^2, \lambda_2, \omega_1, \omega_2$ which have different condtional posterior densities than in Subsection 5.5.1.

MCMC Implementation

i. For X_i,

$$
\begin{aligned}
f(X_i|X_i^{C}) \propto \exp &\left\{ [\exp(\beta_0+\beta_1 X_i)]^{Y_i} (\exp(\beta_0+\beta_1 X_i)+r)^{-(r+Y_i)} \right\} \\
&\left\{ -\frac{1}{2\sigma_X^2}(X_i-\mu_X)^2 \right\} \left\{ \Phi\left(\frac{\omega_1(X_i-\alpha)+\omega_2(X_i-\alpha)^3/\lambda_1^2}{\sqrt{\lambda_1^2+\lambda_2(X_i-\alpha)^2}} \right) \right\},
\end{aligned}
$$

such that,

$$\sigma_X^2 = \tau^2\lambda_1^2/(\tau^2 + m\lambda_1^2),$$

$$\mu_X = (\alpha\tau^2 + m\bar{X}_i^*\lambda_1^2)/(\tau^2 + m\lambda_1^2),$$

$$\bar{X}_i^* = \textstyle\sum_{j=1}^m X_{ij}^*/m.$$

Using proposal normal distribution of mean μ_X and variance σ_X^2, we update X_i independently for $i = 1, 2, \ldots, n$ using MH algorithm.

ii. For α,

$$f(\alpha|\alpha^C) \propto \exp\left\{ -\frac{n}{2\lambda_1^2}(\alpha - \bar{X})^2 \right\}\left\{ \prod_{i=1}^n \Phi\left(\frac{\omega_1(X_i - \alpha) + \omega_2(X_i - \alpha)^3/\lambda_1^2}{\sqrt{\lambda_1^2 + \lambda_2(X_i - \alpha)^2}} \right) \right\},$$

where the first component of the above conditional distribution is normal with mean and variance $\bar{X}$ and λ_1^2/n, respectively. So, the parameter α is updated using RWMH $N(0, k_\alpha^2\lambda_1^2/n)$ where k_α is the tuning parameter and we set $k_\alpha = \sqrt{0.8}$ so that the algorithm has acceptance rate between 25% and 40%.

iii. For λ_1^2,

$$f(\lambda_1^2|\lambda_1^{2^C})$$
$$\propto \left(\frac{1}{\lambda_1^2}\right)^{\frac{n+1}{2}+1} \exp\left[-\frac{0.5}{\lambda_1^2}\Big(\sum_{i=1}^n (X_i - \alpha)^2 + \Big) \right]\left\{ \prod_{i=1}^n \Phi\left(\frac{\omega_1(X_i - \alpha) + \omega_2(X_i - \alpha)^3/\lambda_1^2}{\sqrt{\lambda_1^2 + \lambda_2(X_i - \alpha)^2}} \right) \right\},$$

where the first component of the condtional posterior is IG, thus we choose to use MH algorithm to update this parameter such that, λ_1^2 is sampled using the proposal distribution, $IG((n + 1)/2, 0.5(\sum_{i=1}^n (X_i - \alpha)^2 + 1)$.

iv. For λ_2

$$f(\lambda_2|\lambda_2^C) \propto \exp\left(-\frac{\lambda_2^2}{2} \right)\left\{ \prod_{i=1}^n \Phi\left(\frac{\omega_1(X_i - \alpha) + \omega_2(X_i - \alpha)^3/\lambda_1^2}{\sqrt{\lambda_1^2 + \lambda_2(X_i - \alpha)^2}} \right) \right\},$$

such that $\lambda_2 > 0$ and the first component on the right-handside of the conditional posterior is half- normal distribution which is constructed from the half-normal prior distribution specified earlier. Thus, λ_2 is updated using RWMH with *Half-Normal*$(0, k_{\lambda_2})$ as its proposal distribution and we let the tuning parameter be $k_{\lambda_2} = 1$ which yields acceptance rate between 25% and 30%.

v. For ω_h where $h = 1, 2$,

$$f(\omega_h|\omega_h^C) \propto \left\{ \prod_{i=1}^{n} \Phi\left(\frac{\omega_1(X_i - \alpha) + \omega_2(X_i - \alpha)^3/\lambda_1^2}{\sqrt{\lambda_1^2 + \lambda_2(X_i - \alpha)^2}} \right) \right\} \exp\left\{ -\frac{\omega_h^2}{2 \times 100} \right\}.$$

So, we propose ω_1 and ω_2 to be updated from independent $N(0, k_\omega^2)$ using RWMH sampling method. To have acceptance rate between 25% and 40%, we set k_ω to be 0.5.

5.6 Simulation Studies

The same simulation studies conducted in the parameter estimation for EIV PRM are carried out here in order to examine the performance of our approach when the outcome is NBRM where the true values $(\beta_0, \beta_1) = (0.5, 1.0)$, $X_{ij}^* = X_i + \epsilon_j$ for $j = 1, \ldots, m$, and $\epsilon_j \sim N(0, \tau^2)$ is the distribution of EIV. For the sake of simulating data that is similar to real life research situations, the number of replicated surrogates is limited to $m = 2$. We also will simulate EIV using non-normal distribution, which will be discussed in detail later in Subsection 5.7.4 Similarly, in this chapter R also denotes the level of error contamination such that $R = 0.25$ signifies low EIV, $R = 0.5$ signifies medium EIV, meanwhile $R = 1.0$ signifies high EIV. However, now the outcome variable, Y_i is generated from $Y_i \sim NB(r, \exp(\beta_0 + \beta_1 X_i))$ and r is set to be 1.0, which indicates a high dispersion happening in the count data. We again consider the four simulation settings,

Simulation setting 1: $X_i \sim 0.5N(0.19, 0.08^2) + 0.2N(1.05, 0.2^2) + 0.3N(2, 0.48^2)$

Simulation setting 2: $X_i \sim 0.5N(-2, 1) + 0.5N(2, 1)$

Simulation setting 3: $X_i \sim Gamma(2, 2^{-1})$

Simulation setting 4: $X_i \sim LN(0, 1)$

Simulation setting 1 represents true exposure X_i distribution that is a skewed mixture of normal meanwhile simulation setting 2 represents a distribution that is a bimodal mixture of normal. X_i that are simulated from simulation setting 3 will have a skewed distribution and heavy-tailedness. Finally, we also study the case in which X_i is generated from log-normal distribution in simulation setting 4 and hence will have both skewness and heavy-tailedness. The difference between simulation setting 3 and 4 is that the latter will have an even heavier tail in its distributional shape. 50 datasets are generated under each simulation setting and the sample sizes used are $n = 50$ and $n = 100$.

5.7 Results

We present the results of our simulation studies and the performance of our flexible Bayesian approach to correct EIV in NBRM in this section. For both flexible distributions that are studied, FGSN and FSGN, we run MCMC chains of length $300,000$ and $100,000$ length of burn-ins. For each of the 50 data sets, we shall have posterior estimates of each of the model parameters with sample size $200,000$ which is the remainder of the MCMC iterations after burn-in. The mean of these posterior estimates is taken as our model parameter estimates in each data set. To confirm the convergence of these MCMC chains, we construct trace plots and based on the visual, we see that these chains have good mixing and have achieved convergence with the given iteration length. See Figure

5.1 for traceplots of β_0 and β_1 estimates from a randomly selected simulation study.

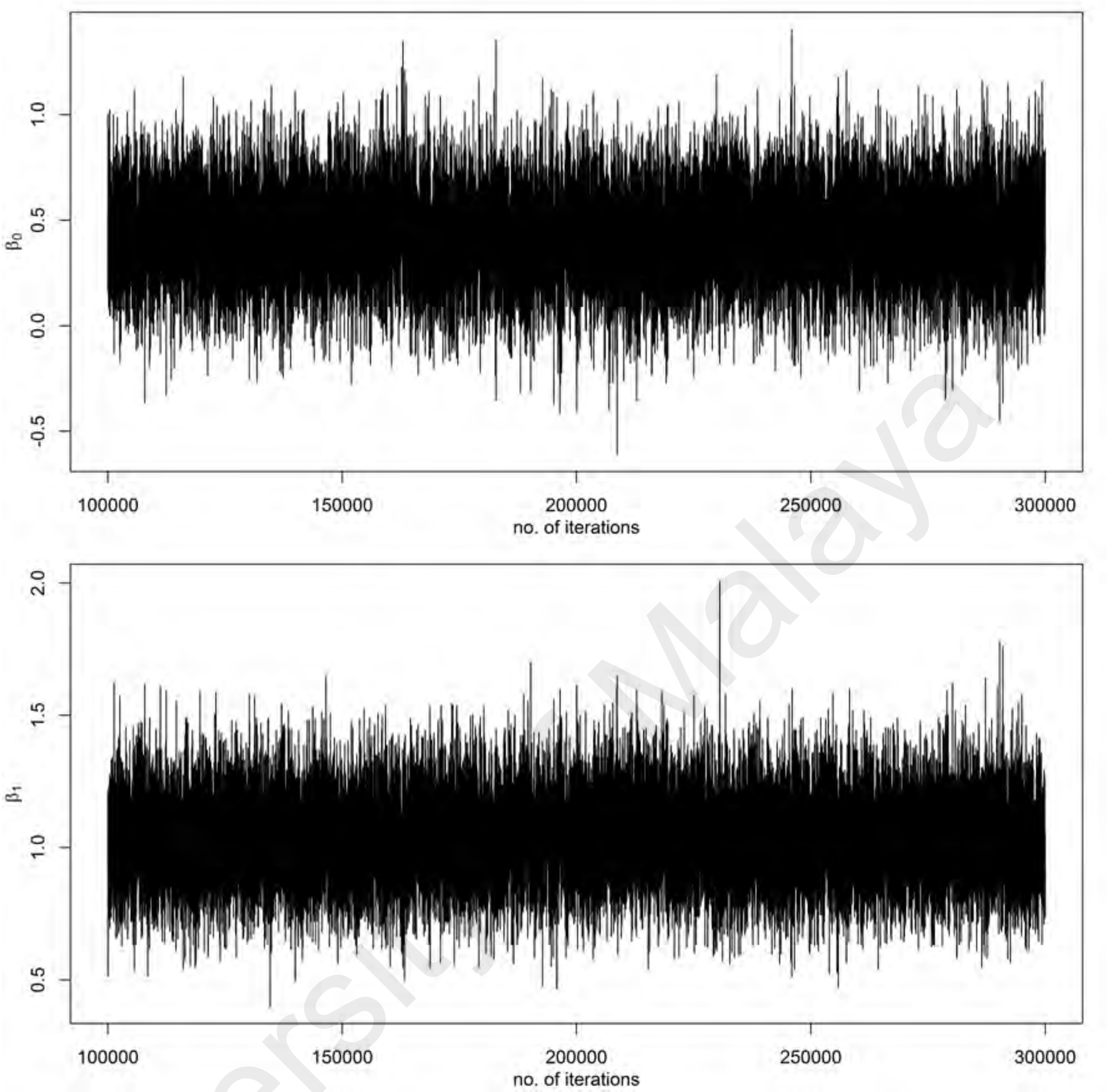

Figure 5.1: Trace plots for estimated β_0 and β_1 in one of the simulation studies

The results of various analysis for NBR outcome model is shown in Tables 5.1 and 5.2, where the former contains the results for FGSN misspecified exposure model, meanwhile the latter contains the result for FSGN misspecified exposure model. We shall use the same criteria as explained in Section 4.7. To provide visualisations of the performance of the flexible Bayesian approach in correcting EIV for NBRM, we plot the kernel posterior densities of the adjusted X_i against its corresponding true exposures, X_i and mean proxy $\bar{X}_i^*$ from a randomly selected dataset with $R = 1.0$ and $n = 100$, as given in Figures 5.2 and 5.3.

In this chapter, we also separate the results into two subsections; Subsection 5.7.1 presents the results when using FGSN exposure model, and Subsection 5.7.2 presents the results when using FSGN exposure model.

Similarly as in the previous chapter, we also compare the performance of the two flexible models by constructing their kernel posterior densities of estimated β_0 and β_1 and choose the best of the two (where the better flexible model shall have kernel posterior densities that have shapes which will closely follow to the *benchmark* kernel densities shape). This is discussed and shown in Section 5.7.3. After choosing the preferred flexible model, we use it to find the performance of our approach when the EIV distribution is non-normal.

5.7.1 Using FGSN Exposure Model

In Table 5.1, FGSN as the flexible misspecified exposure model does well in attenuating bias caused by EIV in NBRM, such that our *flexible* estimations of parameters β_0 and β_1 has better values than that of the *naive* estimates. There is a significant decrease in bias and the flexible Bayesian approach shows good bias-variance trade-off as seen in the MSE values. We also would argue that the values of the *flexible* estimates follow closely to their corresponding *benchmark* estimates.

As depicted in Figure 5.2, under every setting, the kernel densities of our corrected X_i closely follow the shapes of the kernel densities of their respective unknown but true exposure. Under simulation setting 1, where the shape is skewed, the skewness and tail of our corrected exposure kernel density is similar to that of the true exposure kernel density under FGSN exposure model. The two peaks for simulation setting 2, where the true exposures are generated from the bimodal mixture of normal are clearly shown Figure 5.2, such that the usage of misspecified exposure FGSN has clear bimodal shape and adequately follow the *benchmark* (true) kernel density shape. For simulation settings 3

Table 5.1: Accuracy and sensitivity of estimated parameters, β_0 and β_1 under different true and unobserved distributions of X for negative binomial regression model with FGSN as misspecified exposure model

Sample size n = 50

			R = 0.25		R = 0.5		R = 1		
Simulation setting	Parameter		Naive	Flexible	Naive	Flexible	Naive	Flexible	Benchmark
1	β_0	M	0.68213	0.47880	0.83455	0.40340	1.02503	0.23577	0.52233
		B	0.18213	0.02120	0.33455	0.09660	0.52503	0.26423	0.02233
		MSE	0.09505	0.07884	0.17542	0.12858	0.35081	0.28688	0.06837
	β_1	M	0.85842	1.04731	0.73488	1.11389	0.57964	1.27784	0.98915
		B	0.14158	0.04731	0.26512	0.11389	0.42036	0.27784	0.01085
		MSE	0.04774	0.04277	0.09385	0.07747	0.19966	0.20264	0.03216
2	β_0	M	0.66790	0.42713	0.83158	0.37680	1.06919	0.30528	0.46091
		B	0.16790	0.07287	0.33158	0.12320	0.56919	0.19472	0.03909
		MSE	0.15002	0.15867	0.25319	0.22264	0.50836	0.27689	0.09477
	β_1	M	0.93987	1.06880	0.86459	1.11399	0.74497	1.19951	1.02752
		B	0.06013	0.06880	0.13541	0.11399	0.25503	0.19951	0.02752
		MSE	0.01696	0.02544	0.03275	0.04502	0.08145	0.12964	0.01212
3	β_0	M	1.32666	0.42409	1.91196	0.29135	2.79529	0.07632	0.56901
		B	0.82666	0.07591	1.41196	0.20865	2.29529	0.42368	0.06901
		MSE	0.87072	0.19463	2.25498	0.39505	5.63384	1.11944	0.06787
	β_1	M	0.88832	1.01816	0.82078	1.04937	0.75124	1.10956	0.98107
		B	0.11168	0.01816	0.17922	0.04937	0.24876	0.10956	0.01893
		MSE	0.02211	0.00824	0.04712	0.01771	0.11720	0.05753	0.00311
4	β_0	M	0.86229	0.40966	1.13186	0.33512	1.51641	0.14351	0.48305
		B	0.36229	0.09034	0.63186	0.16488	1.01641	0.35649	0.01695
		MSE	0.22097	0.12807	0.5112	0.2278	1.19051	0.61547	0.0563
	β_1	M	0.88197	1.06187	0.81130	1.11847	0.73504	1.25062	1.00255
		B	0.11803	0.06187	0.18870	0.11847	0.26496	0.25062	0.00255
		MSE	0.02977	0.02524	0.05844	0.05741	0.10786	0.20184	0.00651

Sample size n = 100

			R = 0.25		R = 0.5		R = 1		
Simulation setting	Parameter		Naive	Flexible	Naive	Flexible	Naive	Flexible	Benchmark
1	β_0	M	0.65768	0.47461	0.81472	0.43998	1.01402	0.27497	0.49121
		B	0.15768	0.02539	0.31472	0.06002	0.51402	0.22503	0.00879
		MSE	0.05746	0.04416	0.13473	0.07228	0.30439	0.20866	0.02936
	β_1	M	0.86337	1.02045	0.74295	1.05469	0.58961	1.21028	0.98647
		B	0.13663	0.02045	0.25705	0.05469	0.41039	0.21028	0.01353
		MSE	0.03884	0.02542	0.08744	0.04287	0.19054	0.15646	0.01738
2	β_0	M	0.65969	0.47731	0.81822	0.49507	1.05603	0.50448	0.43531
		B	0.15969	0.02269	0.31822	0.00493	0.55603	0.00448	0.06469
		MSE	0.05637	0.03736	0.14026	0.04244	0.36381	0.05583	0.03225
	β_1	M	0.93116	1.01195	0.86468	1.01149	0.75507	1.02612	1.01665
		B	0.06884	0.01195	0.13532	0.01149	0.24493	0.02612	0.01665
		MSE	0.01050	0.00764	0.02562	0.00962	0.07030	0.02256	0.00615
3	β_0	M	1.25485	0.44826	1.88449	0.35972	4.24647	0.20287	0.51224
		B	0.75485	0.05174	1.38449	0.14028	3.74647	0.29713	0.01224
		MSE	0.68772	0.13687	2.20742	0.24426	56.8338	0.54457	0.05173
	β_1	M	0.91205	1.01225	0.84895	1.02594	0.76438	1.05299	1.00160
		B	0.08795	0.01225	0.15105	0.02594	0.23562	0.05299	0.00160
		MSE	0.01219	0.00424	0.02904	0.00761	2.11603	0.01762	0.00146
4	β_0	M	0.99346	0.44236	1.13939	0.40007	1.53714	0.30854	0.50449
		B	0.49346	0.05764	0.63939	0.09993	1.03714	0.19146	0.00449
		MSE	0.18288	0.07516	0.49320	0.12497	1.21163	0.20921	0.03074
	β_1	M	0.89234	1.02914	0.82815	1.05487	0.75814	1.10663	0.98804
		B	0.10766	0.02914	0.17185	0.05487	0.24186	0.10663	0.01196
		MSE	0.02180	0.01125	0.04857	0.01965	0.09200	0.03939	0.00440

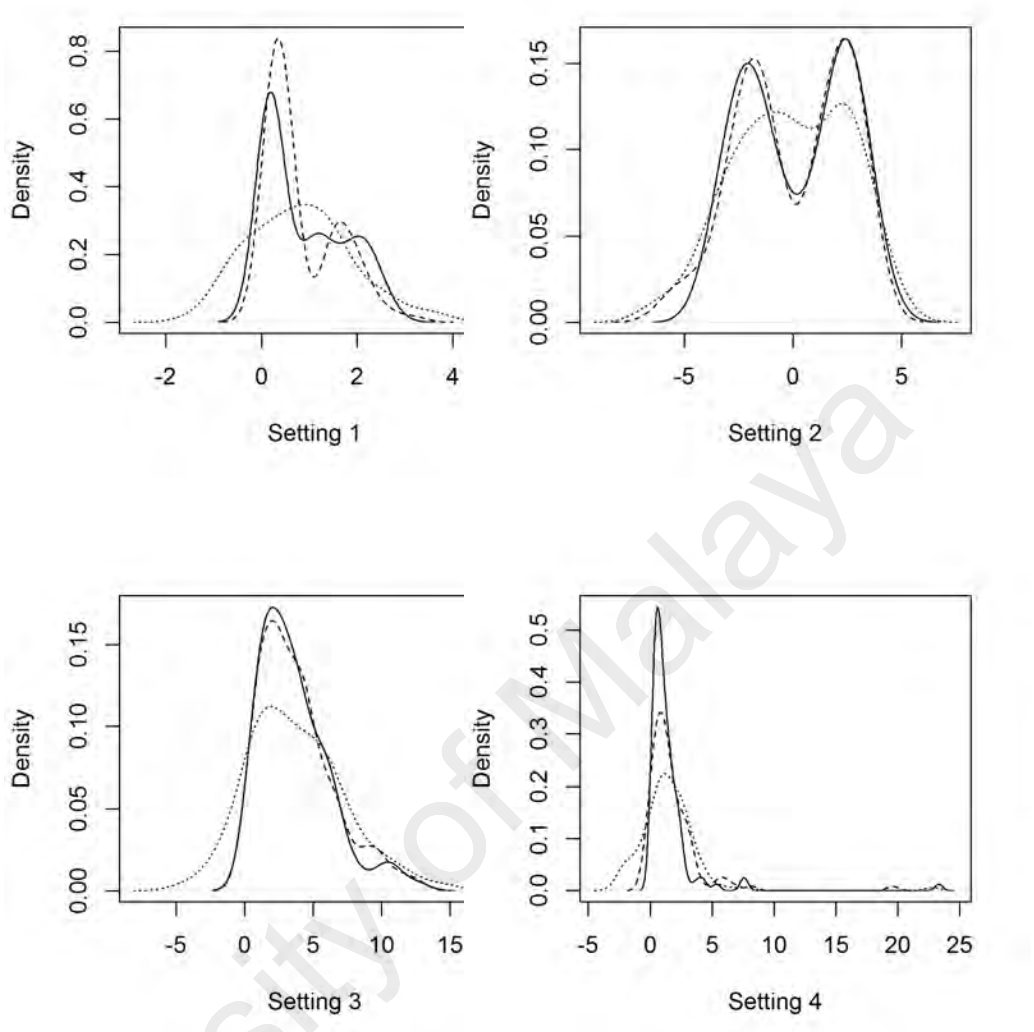

Figure 5.2: Kernel density estimates for settings 1-4 in the case of misspecified FGSN exposure model for EIV in NBRM: true exposure X_i (solid curve); estimated X_i under flexible Bayesian approach (dashed curve); mean proxy $\bar{X}_i^*$ (dotted curve).

and 4, where there exist heavy-tailedness for both of the simulation settings, our corrected X_i still manage to have better kernel density shapes in comparison to the non-adjusted exposure kernel density $\bar{X}_i^*$. On the other hand, the kernel densities of $\bar{X}_i^*$ have blurry shapes under every simulation settings which further prove that if EIV is not corrected, wrong statistical analysis and conclusions might be made.

5.7.2 Using FSGN Exposure Model

The results presented in Table 5.2 show that the Bayesian approach with FSGN as the flexible exposure model is found to adjust the bias adequately in estimating the NBRM

Table 5.2: Accuracy and sensitivity of estimated parameters, β_0 and β_1 under different true and unobserved distributions of X for negative binomial regression model with FSGN as misspecified exposure model

Sample size $n = 50$

			$R = 0.25$		$R = 0.5$		$R = 1$		
Simulation setting	Parameter		Naive	Flexible	Naive	Flexible	Naive	Flexible	Benchmark
1	β_0	M	0.68213	0.47567	0.83455	0.38623	1.02503	0.17000	0.52233
		B	0.18213	0.02433	0.33455	0.11377	0.52503	0.33000	0.02233
		MSE	0.09505	0.10153	0.17542	0.16408	0.35081	0.41854	0.06837
	β_1	M	0.85842	1.04707	0.73488	1.12527	0.57964	1.32311	0.98915
		B	0.14158	0.04707	0.26512	0.12527	0.42036	0.32311	0.01085
		MSE	0.04774	0.06708	0.09385	0.09915	0.19966	0.23859	0.03216
2	β_0	M	0.66790	0.42217	0.83158	0.39964	1.06919	0.27270	0.46091
		B	0.16790	0.07783	0.33158	0.10036	0.56919	0.22730	0.03909
		MSE	0.15002	0.17010	0.25319	0.21606	0.50836	0.37255	0.09477
	β_1	M	0.93987	1.07228	0.86459	1.10234	0.74497	1.21679	1.02752
		B	0.06013	0.07228	0.13541	0.10234	0.25503	0.21679	0.02752
		MSE	0.01696	0.04773	0.03275	0.06676	0.08145	0.16603	0.01212
3	β_0	M	1.32666	0.47609	1.91196	0.40720	2.79529	0.21735	0.56901
		B	0.82666	0.02391	1.41196	0.09280	2.29529	0.28265	0.06901
		MSE	0.87072	0.21444	2.25498	0.37837	5.63384	0.91939	0.06787
	β_1	M	0.88832	1.01354	1.03299	1.04937	0.75124	1.07921	0.98107
		B	0.11168	0.01354	0.03299	0.04937	0.24876	0.07921	0.01893
		MSE	0.02211	0.04073	0.04712	0.03948	0.11720	0.04453	0.00311
4	β_0	M	0.86229	0.42150	1.13186	0.28712	1.51641	0.08506	0.48305
		B	0.36229	0.07850	0.63186	0.21288	1.01641	0.41494	0.01695
		MSE	0.22097	0.14984	0.5112	0.34589	1.19051	1.04609	0.0563
	β_1	M	0.88197	1.07227	0.81130	1.14372	0.73504	1.22925	1.00255
		B	0.11803	0.07227	0.18870	0.14372	0.26496	0.22925	0.00255
		MSE	0.02977	0.06515	0.05844	0.14374	0.10786	0.34020	0.00651

Sample size $n = 100$

			$R = 0.25$		$R = 0.5$		$R = 1$		
Simulation setting	Parameter		Naive	Flexible	Naive	Flexible	Naive	Flexible	Benchmark
1	β_0	M	0.65768	0.47461	0.81472	0.43998	1.01402	0.27497	0.49121
		B	0.15768	0.02539	0.31472	0.06002	0.51402	0.22503	0.00879
		MSE	0.05746	0.04416	0.13473	0.07228	0.30439	0.20866	0.02936
	β_1	M	0.86337	1.02045	0.74295	1.05469	0.58961	1.21028	0.98647
		B	0.13663	0.02045	0.25705	0.05469	0.41039	0.21028	0.01353
		MSE	0.03884	0.02542	0.08744	0.04287	0.19054	0.15646	0.01738
2	β_0	M	0.65969	0.47731	0.81822	0.49507	1.05603	0.50448	0.43531
		B	0.15969	0.02269	0.31822	0.00493	0.55603	0.00448	0.06469
		MSE	0.05637	0.03736	0.14026	0.04244	0.36381	0.05583	0.03225
	β_1	M	0.93116	1.01195	0.86468	1.01149	0.75507	1.02612	1.01665
		B	0.06884	0.01195	0.13532	0.01149	0.24493	0.02612	0.01665
		MSE	0.01050	0.00764	0.02562	0.00962	0.07030	0.02256	0.00615
3	β_0	M	1.25485	0.44826	1.88449	0.35972	4.24647	0.20287	0.51224
		B	0.75485	0.05174	1.38449	0.14028	3.74647	0.29713	0.01224
		MSE	0.68772	0.13687	2.20742	0.24426	56.8338	0.54457	0.05173
	β_1	M	0.91205	1.01225	0.84895	1.02594	0.76438	1.05299	1.00160
		B	0.08795	0.01225	0.15105	0.02594	0.23562	0.05299	0.00160
		MSE	0.01219	0.00424	0.02904	0.00761	2.11603	0.01762	0.00146
4	β_0	M	0.99346	0.44236	1.13939	0.40007	1.53714	0.30854	0.50449
		B	0.49346	0.05764	0.63939	0.09993	1.03714	0.19146	0.00449
		MSE	0.18288	0.07516	0.49320	0.12497	1.21163	0.20921	0.03074
	β_1	M	0.89234	1.02914	0.82815	1.05487	0.75814	1.10663	0.98804
		B	0.10766	0.02914	0.17185	0.05487	0.24186	0.10663	0.01196
		MSE	0.02180	0.01125	0.04857	0.01965	0.09200	0.03939	0.00440

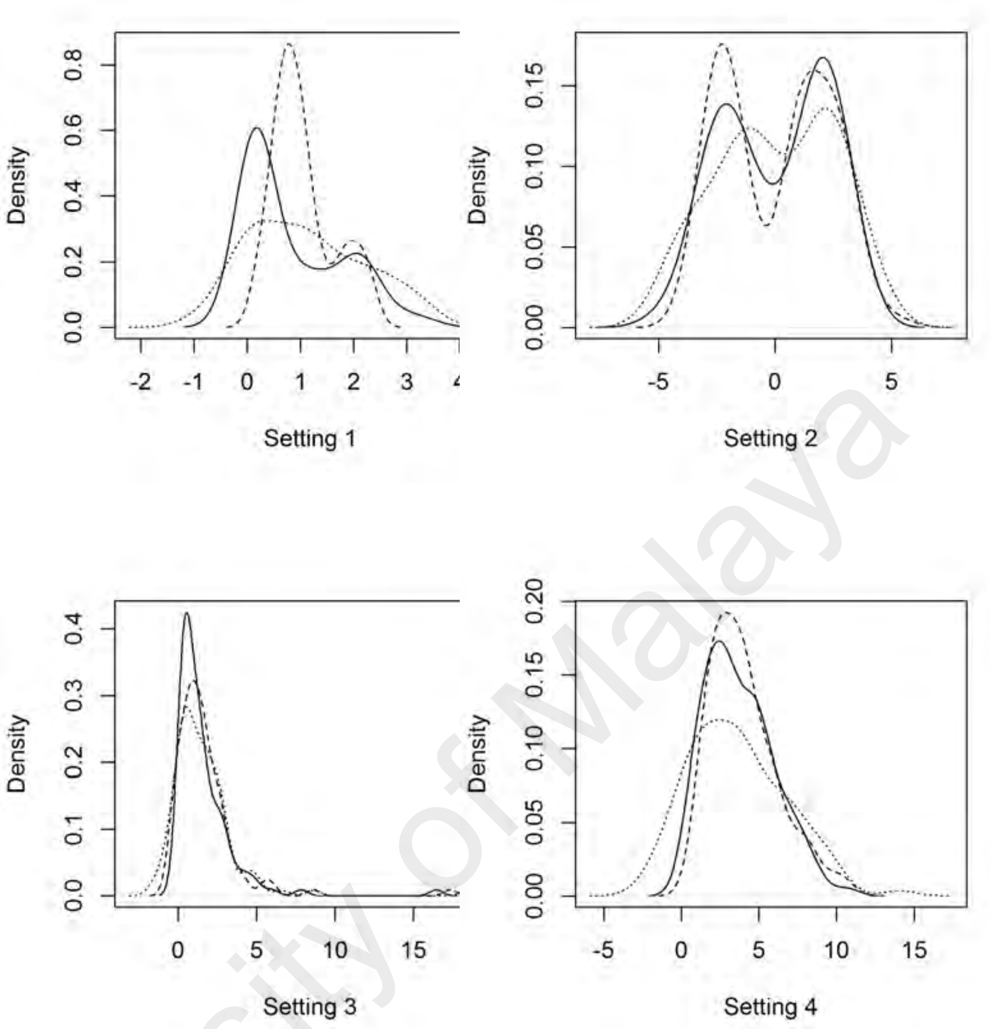

Figure 5.3: Kernel density estimates for settings 1-4 in the case of misspecified FSGN exposure model for EIV in NBRM: true exposure X_i (solid curve); estimated X_i under flexible Bayesian approach (dashed curve); mean proxy $\bar{X}_i^*$ (dotted curve).

parameter estimates in every simulation settings and follow closely to the *benchmark* estimate values. This is including the MSE and bias of the *flexible* parameter estimates even when the error contamination is substantial. Meanwhile, the parameter estimates under non-corrected estimates, i.e., the *naive* estimates have poor values in each simulation settings for all the error contamination levels.

Figure 5.3 shows that the kernel densities of our corrected exposure X_i for FSGN misspecified exposure models have shapes that closely follow the shapes of the kernel densities of unknown but true exposure X_i under simulation settings 1-4. Under simulation setting 1, where the shape is skewed, there is a deterioration in the shape of the kernel density but do note that it still has a better shape than its corresponding $\bar{X}_i^*$ kernel density. The kernel density of the corrected exposure in simulation setting 2, has clear bimodal shape and adequately follows the *benchmark* kernel density shape. Under simulation settings 3 and 4, where there exist heavy-tailedness for both of the simulation settings, our corrected X_i still manage to have better kernel density shapes in comparison to the non-adjusted exposure kernel density $\bar{X}_i^*$. Under every simulation setting, the kernel densities of $\bar{X}_i^*$ have blurry shapes under every simulation settings which further prove that if EIV is not corrected, wrong statistical analysis and conclusions might be made.

5.7.3 Comparing the Performance between FGSN and FSGN as the Misspecified Exposure Model for EIV NBRM.

In general, the results shown in the tables provide proof that the Bayesian approach with FGSN and FSGN as the misspecified exposure model are robust in estimating the values of NBRM parameter estimates in the presence of EIV. The approach shows good bias correction under different error contamination levels as well as under different simulation settings. In addition to this, the low values of their MSEs also imply that the flexible Bayesian approach has adequate bias-variance trade-offs in comparison to *benchmark* estimates. Even when the true exposure distribution has departures from normality, the approach shows no deterioration in performance and still strikes better result than that of the *naive* estimates where no bias correction is done. Therefore, in comparison with the *naive* estimates, our approach using both FGSN and FSGN, shows superior performance in terms of accuracy and consistency.

Now, to cross-compare the robustness of FGSN and FSGN as the misspecified exposure model, we provide the kernel empirical density plots of β_0 and β_1 for every simulation settings in Figures 5.4 to 5.7. The solid curve represents the *benchmark* estimates, dashed curve is the estimates under FGSN model and long-dashed curve

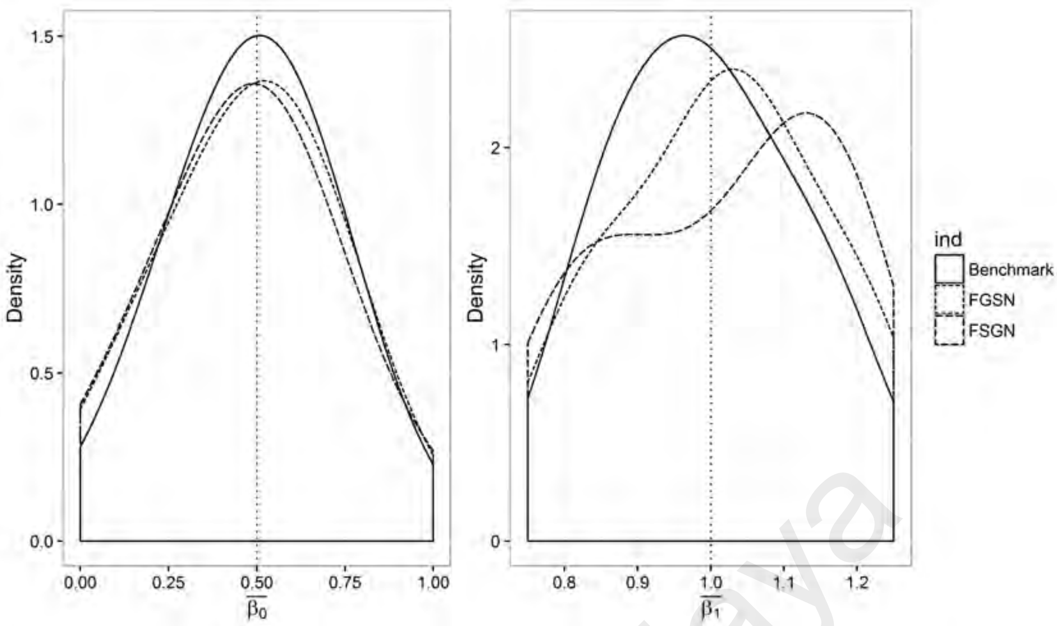

Figure 5.4: Kernel density of estimated regression parameters, $\bar{\beta}_0$ and $\bar{\beta}_1$ under simulation setting 1: Benchmark (solid curve); FGSN (dashed curve); FSGN (long-dashed curve).

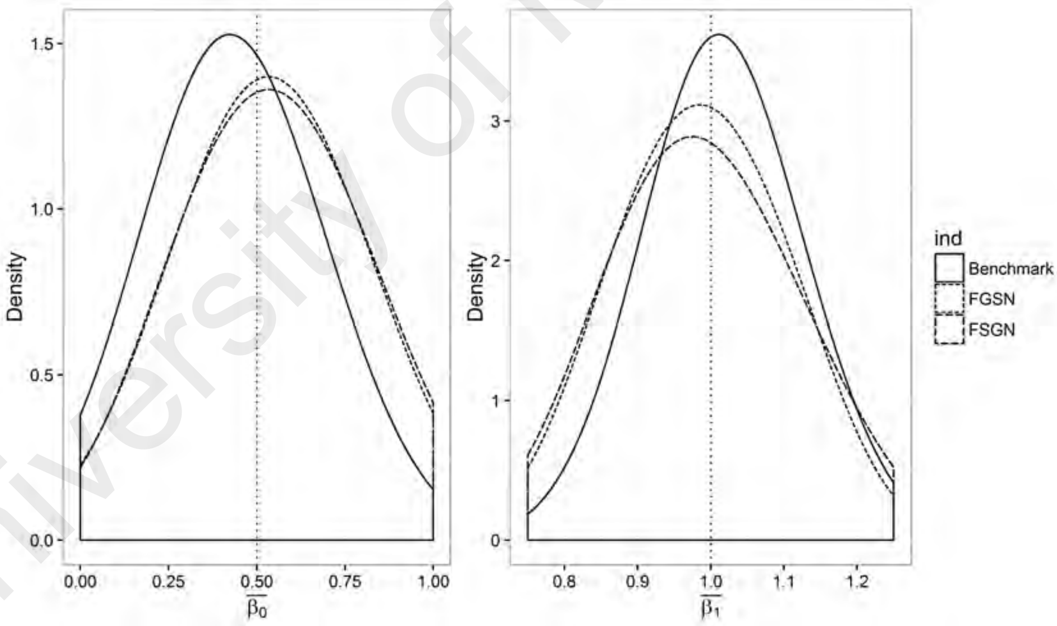

Figure 5.5: Kernel density of estimated regression parameters, $\bar{\beta}_0$ and $\bar{\beta}_1$ under simulation setting 2: Benchmark (solid curve); FGSN (dashed curve); FSGN (long-dashed curve).

represents the estimates under FSGN exposure model. They are all plotted estimates of β_0 and β_1 from simulation studies with $R = 1.0$ EIV contamination ratio and sample size of $n = 100$.

In Figure 5.4 which follows simulation setting 1, the kernel density estimates for β_0

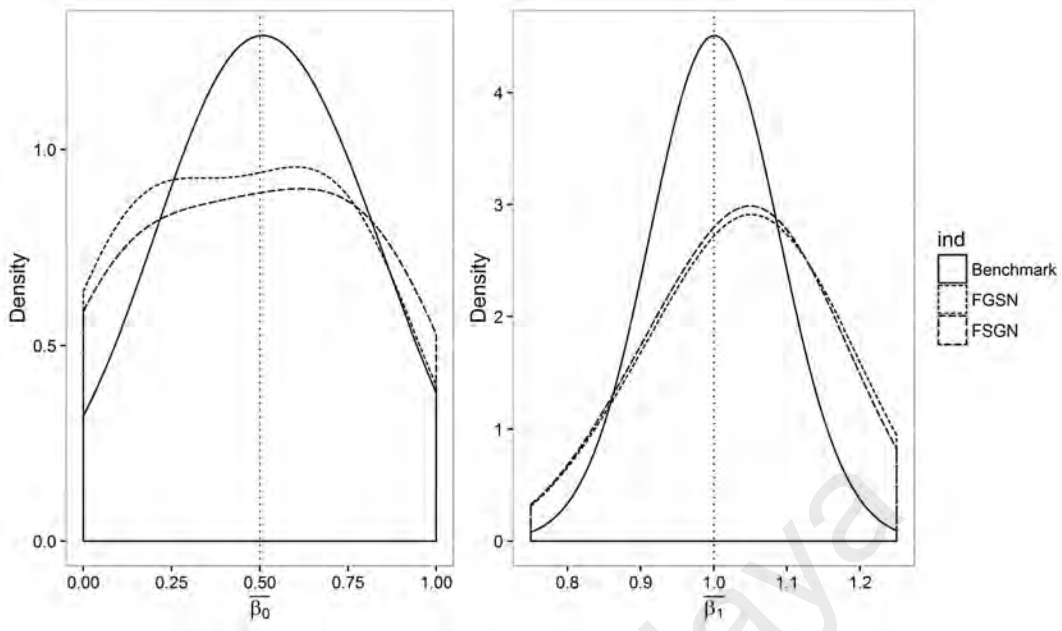

Figure 5.6: Kernel density of estimated regression parameters, $\bar{\beta_0}$ and $\bar{\beta_1}$ under simulation setting 3: Benchmark (solid curve); FGSN (dashed curve); FSGN (long-dashed curve).

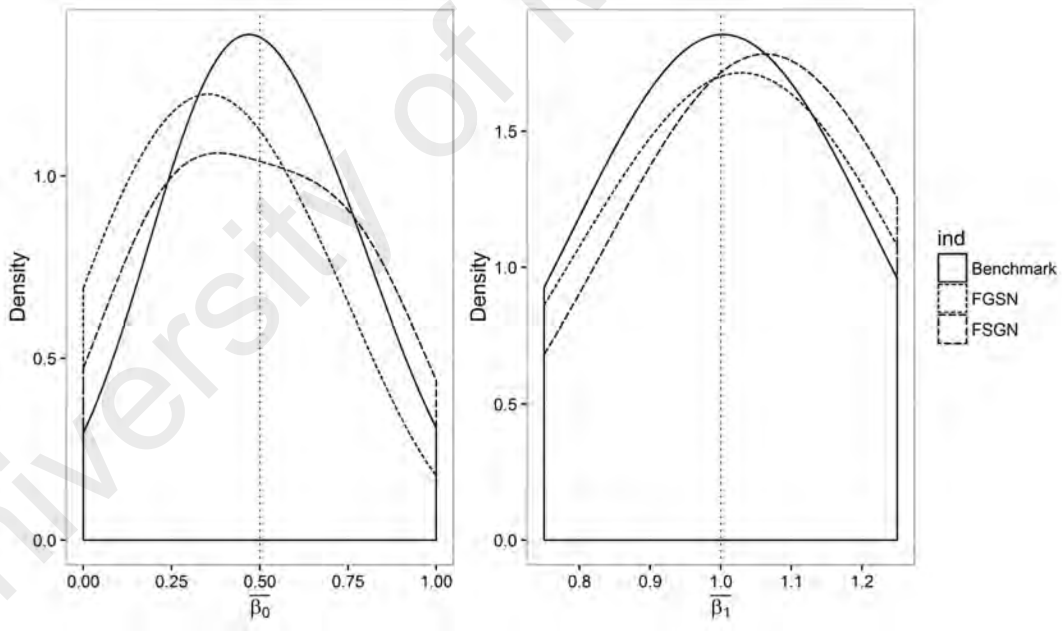

Figure 5.7: Kernel density of estimated regression parameters, $\bar{\beta_0}$ and $\bar{\beta_1}$ under simulation setting 4: Benchmark (solid curve); FGSN (dashed curve); FSGN (long-dashed curve).

has no significant difference between FGSN and FSGN. However, for β_1 kernel density estimates, FSGN is shown to have departed quite far away than the *benchmark* estimates, and therefore we can say that FGSN performs better parameter estimation for EIV NBRM under skewed true exposure distribution. As for when the true exposures follow the distribution stated in simulation setting 2, not much difference can be seen between the performance of FGSN and FSGN as the misspecified exposure model although one could argue that the peak of the FGSN kernel empirical estimates is better than the peak for FSGN exposure model. Again, under simulation setting 3, Figure 5.6 suggests that there is no significant difference between the usage of FGSN and FSGN. On the other hand, when true exposures are generated from simulation setting 4, FGSN shows greater performance than FSGN such that its plots are better for both β_0 and β_1 when compared to FSGN. This can clearly be seen by the peaks of the two kernel empirical estimates.

Thus, from the observations above, we choose FGSN as the misspecified exposure model as it shows better performance than FSGN. We shall continue our research on the correction of EIV when the errors are non-normal using FGSN.

5.7.4 Non-normal Distribution of EIV

As mentioned before, we now conduct a study where the distribution of EIV is no longer normal. Here, the measurement error, ϵ is generated from $SN(0, 1)$ and $ST(0, 1)$, meanwhile true exposures X_i are generated from skewed mixture of normal with substantial EIV ($R = 1.0$). Since we chose FGSN as the better flexible distribution in the previous subsection, here we shall only use FGSN as the misspecified exposure model.

The results are provided in Table 5.3. According to the results of our simulation studies, when using the normal distribution as the measurement model, even when the distributions of EIV are non-normal, the choice of our model still show robustness. Following this, we reach the conclusion that to specify a flexible model also on the measurement model is redundant and unnecessary. It might even reduce the effectiveness of our model as when using flexible models excessively, efficiency is sacrificed (Ma & Genton, 2004).

Table 5.3: Estimated values of β_0 and β_1 of EIV NBRM where EIV is generated from skew-normal and skew-*t* distributions.

Distribution of EIV	Parameter		Naive	Flexible	Benchmark
Skew-normal	β_0	M	0.98897	0.34248	0.48536
		B	0.48897	0.15752	0.01464
		MSE	0.27572	0.14453	0.02812
	β_1	M	0.58843	1.12313	0.98901
		B	0.41157	0.12313	0.01099
		MSE	0.18679	0.11375	0.01602
Skew-*t*	β_0	M	1.13862	0.16015	0.48536
		B	0.63862	0.33985	0.01464
		MSE	0.46186	0.32586	0.02812
	β_1	M	0.47338	1.29900	0.98901
		B	0.52662	0.29900	0.01099
		MSE	0.29911	0.24265	0.01602

CHAPTER 6: DISCUSSION

6.1 Bayesian Approach to Errors-in-Variables in Poisson Regression Model

From the simulation studies done, it is reported that the use of flexible Bayesian approach results in a significant bias reduction caused by EIV when estimating the regression parameters of PRM in comparison to when the EIV is not addressed. The results are shown in Tables 4.1 and 4.2. In addition to that, the proposed approach also has very low MSEs which implies that we have a good bias-variance tradeoff. We consider two different flexible distributions, which are FGSN and FSGN. The latter distribution offers more flexibility than that of the preceding one. However, FGSN still shows more significant bias reduction than FSGN especially when the ratio of error contamination *R* is large. From the kernel density plots of the exposures, we can see more clearly that for both flexible models, there are not much difference in bias reduction and bias-variance tradeoffs.

FSGN has an extra parameter which offers more flexibility but in return, deteriorates in terms of efficiency as the computation time for FSGN in comparison to FGSN is much longer. We also investigated the use of the extended skew generalized-normal model as the misspecified exposure model, but similarly, as FSGN, the performance shows a little deterioration as it is more flexible and has even more extra parameters. The same simulation studies are conducted for FGST, since the degree of freedom for FGST that is estimated in EIV PRM is large, FGST converges to FGSN. Therefore, the implementation of FGSN is adequate. In addition to this, since FGST has more parameters, its computation time is significantly more than FGSN.

To summarize, in our study for estimating biased parameters of EIV PRM, FGSN should be the preferred flexible exposure model.

Using the advocated model, FGSN, we also study the case where EIV is generated from SN and ST distributions. The justification behind this is to investigate if the normal distribution which we specified as the measurement model shows robustness in estimating the parameters accurately when EIV distributions are non-normal. From our simulation studies, the normal distribution is adequate and to specify a flexible distribution also in the measurement model would be redundant.

6.2 Bayesian Approach to Errors-in-Variables in Negative Binomial Regression Model

In our search for literature on studies done in correcting EIV in NBRM, we came across very few of them. Current studies on EIV correction in NBRM used the Bayesian approach but the exposure model distribution is considered as known and is either normal or log-normal. After acknowledging this observation, we use the Bayesian approach to correct bias in parameter estimations caused by EIV when the exposures have departures from normality. By intentionally misspecifying the flexible models as the exposure model, we are able to implement a general framework even when the non-normal distribution used in every simulation setting is different (i.e., skewed, bimodal and heavy-tailed distributions).

Results from simulation settings 1 to 4 as shown in Tables 5.1 and 5.2, report that our approach successfully reduces bias caused by EIV when estimating the regression parameters of NBRM. The values of the *flexible* MSEs also suggest that the approach has a good bias-variance trade-off in comparison to the values of MSE reported in *naive* estimates. Both FGSN and FSGN flexible models show good bias attenuation, however again in this chapter, FGSN is preferred. The reasoning is the same as in Chapter 4, such that although FSGN offers more flexibility, the difference in performances between the two flexible models is not significant. Since FSGN has more parameters, then the MCMC algorithms will take a longer time than that of when FGSN is utilized. Here, we also investigated extended skew generalized-normal distribution but similarly, the bias reduction deteriorated when this distribution is implemented, not to mention that the flexible distribution also has more parameters to be estimated, and thus is computationally more expensive. Therefore, FGSN still holds to be the superior misspecified flexible model.

Therefore, using FGSN we study the effects of our approach when EIV distributions are non-normal. We then see that normal distribution as the measurement model is adequate and there is no need to specify another flexible distribution for the measurement model.

As a summary, the flexible Bayesian approach is advocated as the method to reduce bias in estimating parameters for EIV NBRM.

CHAPTER 7: CONCLUDING REMARKS AND FUTURE RESEARCH

7.1 Concluding Remarks

The research in this thesis focuses on reducing the impact of bias caused by EIV when estimating count data regression parameters. While existing researches main focus is on addressing EIV in logistic regression, we study on mitigating the impact of bias caused by EIV in count data regression models, namely the PRM and NBRM. Utilizing the framework provided by Richardson and Gilks (1993), we adapted the Bayesian approach to count for EIVs in these two models. To reduce the sensitivity of the estimates to potential misspecification bias, we demonstrate the usage of flexible distributions, FGSN and FSGN in modeling for the distribution of the true exposures. Extensive simulation studies are carried out to illustrate that the flexible Bayesian approach is robust to exposure model misspecification while estimating the PRM and NBRM regression parameters in the presence of EIV. The regression parameters are estimated with a wide implementation of the MCMC algorithms. The advantages of the flexible Bayesian approach in comparison to competing methods in EIV count data regression models are that the Bayesian approach provides more efficiency (Hossain & Gustafson, 2009) as well as the fact that we consider the true exposure distribution as unknown and has departures from normality which is more realistic and applicable in practice. Besides that, existing methods also assume the EIV variance as known, in this thesis however it is estimated and considered as unknown.

In our research, we looked into estimating parameters in EIV PRM and NBRM using Bayesian approach and found the best flexible models between FGSN and FSGN to minimize model misspecification bias. From the results reported in this thesis using simulation studies, the flexible Bayesian approach works well in eliminating EIV bias adequately while providing consistent and accurate regression parameter estimates. This is shown in Tables 4.1 and 4.2 for PRM and Tables 5.1 and 5.2 for NBRM as there is a significant bias reduction from the *naive* parameter estimates and *flexible* parameter estimates. This is also shown when the MSEs for *naive* estimates are much larger than that of the *flexible* estimates. Following this, we also compare the performance between FGSN and FSGN as the intentionally misspecified exposure model. Under Poisson regression outcome model, FGSN and FSGN shows similar performance in terms of EIV bias reduction. The distinction between the two, however is that, FSGN has slower and longer MCMC simulation time due to its extra parameter. Thus, for Poisson regression outcome model, the usage of FGSN is advocated. As for negative binomial outcome model, FGSN shows better performance than FSGN which could be seen clearly from the kernel empirical density plots of the parameter estimates in Figures 5.4 - 5.7.

7.2 Future Research

Following are suggestions for further research in this area:

1. adapt the flexible Bayesian approach to other count data regression models such as zero-inflated Poisson regression model.
2. adapt the flexible Bayesian approach to panel count data or longitudinal count data.
3. extend to the case where there are more than one covariate vectors that are measured with error.
4. extend to the case where the replicates of surrogate exposures are correlated with each other.

REFERENCES

Ahmed, A., Sadullah, A. F. M., & Shukri Yahya, A. (2014). Accident analysis using count data for unsignalized intersections in Malaysia. *Procedia Engineering*, *77*, 45–52.

Arellano-Valle, R. B., Gómez, H. W., & Quintana, F. A. (2004). A new class of skew-normal distributions. *Communications in Statistics-Theory and Methods*, *33*(7), 1465–1480.

Asfaw Dagne, G. (1999). Bayesian analysis of hierarchical Poisson models with latent variables. *Communications in Statistics-Theory and Methods*, *28*(1), 119–136.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 171–178.

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. Boca Raton, FL: CRC Press.

Bolfarine, H., & Lachos, V. H. (2007). Skew-probit measurement error models. *Statistical Methodology*, *4*(1), 1–12.

Box, G. E., & Tiao, G. C. (2011). *Bayesian inference in statistical analysis* (Vol. 40). New York, NY: John Wiley & Sons.

Campbell, J., Jones, A. S., Dienemann, J., Kub, J., Schollenberger, J., O'campo, P., . . . Wynne, C. (2002). Intimate partner violence and physical health consequences. *Archives of Internal Medicine*, *162*(10), 1157–1163.

Carroll, R. J. (1989). Covariance analysis in generalized linear measurement error models. *Statistics in Medicine*, *8*(9), 1075–1093.

Carroll, R. J., Gail, M. H., & Lubin, J. H. (1993). Case-control studies with errors in covariates. *Journal of the American Statistical Association*, *88*(421), 185–199.

Carroll, R. J., Roeder, K., & Wasserman, L. (1999). Flexible parametric measurement error models. *Biometrics*, *55*(1), 44–54.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement* *error in nonlinear models: a modern perspective* (2nd ed.). Boca Raton, FL: CRC Press.

Carroll, R. J., & Stefanski, L. A. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*, *85*(411), 652–663.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, *49*(4), 327–335.

Choudhury, K., & Matin, M. A. (2011). Extended skew generalized normal distribution. *Metron*, *69*(3), 265–278.

Consul, P. C., & Jain, G. C. (1973). A generalization of the Poisson distribution. *Technometrics*, *15*(4), 791–799.

Cook, J. R., & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, *89*(428), 1314–1328.

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

Dionne, G., Gagné, R., Gagnon, F., & Vanasse, C. (1997). Debt, moral hazard and airline safety an empirical evidence. *Journal of Econometrics*, *79*(2), 379–402.

Efron, B., & Morris, C. (1975). Data analysis using Stein's estimator and its generalization. *Journal of American Statistical Association*, *70*(350), 311–319.

El-Basyouny, K., & Sayed, T. (2010). Safety performance functions with measurement errors in traffic volume. *Safety Science*, *48*(10), 1339–1344.

Fu, Y., Chu, P., & Lu, L. (2015). A Bayesian approach of joint models for clustered zero-inflated count data with skewness and measurement errors. *Journal of Applied Statistics*, *42*(4), 745–761.

Fuller, W. A. (2009). *Measurement error models* (Vol. 305). New York, NY: John Wiley & Sons.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*(6), 721–741.

Genton, M. G., & Loperfido, N. M. (2005). Generalized skew-elliptical distributions and their quadratic forms. *Annals of the Institute of Statistical Mathematics*, *57*(2), 389–401.

Ghosh, P., Branco, M. D., & Chakraborty, H. (2007). Bivariate random effect model using skew-normal distribution with application to HIV-RNA. *Statistics in Medicine*, *26*(6), 1255–1267.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Introducing Markov chain Monte Carlo* (Vol. 1). Boca Raton, FL: CRC Press.

Greenland, S. (1988). Statistical uncertainty due to misclassification: implications for validation substudies. *Journal of Clinical Epidemiology*, *41*(12), 1167–1174.

Guo, J. Q., & Li, T. (2002). Poisson regression models with errors-in-variables: implication and treatment. *Journal of Statistical Planning and Inference*, *104*(2), 391–401.

Gurmu, S. (1997). Semi-parametric estimation of hurdle regression models with an application to medicaid utilization. *Journal of Applied Econometrics*, 225–242.

Gustafson, P. (2003). *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. Boca Raton, FL: CRC Press.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Hossain, S., & Gustafson, P. (2009). Bayesian adjustment for covariate measurement errors: a flexible parametric approach. *Statistics in Medicine*, *28*(11), 1580–1600.

Huang, Y. (2014). Corrected score with sizable covariate measurement error: pathology and remedy. *Statistica Sinica*, *24*(1), 357.

Kawanishi, K., & Sunquist, M. E. (2004). Conservation status of tigers in a primary rainforest of Peninsular Malaysia. *Biological Conservation*, *120*(3), 329–344.

Küchenhoff, H., & Carroll, R. (1997). Segmented regression with errors in predictors: Semi-parametric and parametric methods. *Statistics in Medicine*, *16*(2), 169–188.

Kukush, A., Schneeweis, H., & Wolf, R. (2004). Three estimators for the Poisson regression model with measurement errors. *Statistical Papers*, *45*(3), 351–368.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1–14.

Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., . . . Ahn, S. Y. (2013). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010. *The Lancet*, *380*(9859), 2095–2128.

Ma, Y., & Genton, M. G. (2004). Flexible class of skew-symmetric distributions. *Scandinavian Journal of Statistics*, *31*(3), 459–468.

Makary, M. A., Segev, D. L., Pronovost, P. J., Syin, D., Bandeen-Roche, K., Patel, P., . . . Tian, J. (2010). Frailty as a predictor of surgical outcomes in older patients. *Journal of the American College of Surgeons*, *210*(6), 901–908.

Mallick, B. K., & Gelfand, A. E. (1996). Semiparametric errors-in-variables models a Bayesian approach. *Journal of Statistical Planning and Inference*, *52*(3), 307–321.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.

Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, *44*(247), 335–341.

Miaou, S.-P. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis & Prevention*, *26*(4), 471–482.

Morris, C. N. (1983). Parametric empirical Bayes inference: theory and applications. , *78*(381), 47–55.

Nakamura, T. (1990). Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika*, *77*(1), 127–137.

Nekoukhou, V., Alamatsaz, M., & Aghajani, A. (2013). A flexible skew-generalized normal distribution. *Communications in Statistics-Theory and Methods*, *42*(13), 2324–2334.

Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS* (Vol. 698). New York, NY: John Wiley & Sons.

Pearson, K. (1902). On the mathematical theory of errors of judgment, with special reference to the personal equation. *Philosophical Transactions of the Royal Society of London*, *198*, 235–299.

Poisson, S. (1837). Research on the probability of judgments in criminal and civil matters. *Paris, France: Bachelier*.

Pridemore, W. A. (2011). Poverty matters: A reassessment of the inequality–homicide relationship in cross-national studies. *The British Journal of Criminology*, *51*(5), 739–772.

Richardson, S., & Gilks, W. R. (1993). Conditional independence models for epidemiological studies with covariate measurement error. *Statistics in Medicine*, *12*(18), 1703–1722.

Richardson, S., Leblond, L., Jaussent, I., & Green, P. J. (2002). Mixture models in measurement error problems, with reference to epidemiological studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *165*(3), 549–566.

Roberts, G. O., Gelman, A., & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, *7*(1), 110–120.

Roeder, K., Carroll, R. J., & Lindsay, B. G. (1996). A semiparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association*, *91*(434), 722–732.

Schoeller, D. A. (1990). How accurate is self-reported dietary energy intake? *Nutrition Reviews*, *48*(10), 373–379.

Schwalbach, J., & Zimmermann, K. F. (1991). A Poisson model of patenting and firm structure in Germany. *Innovation and Technological Change: An International Comparison*, 109–120.

Shankar, V., Milton, J., & Mannering, F. (1997). Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. *Accident Analysis & Prevention*, *29*(6), 829–837.

Sheu, M. L., Hu, T. W., Keeler, T. E., Ong, M., & Sung, H. Y. (2004). The effect of a major cigarette price change on smoking behavior in california: a zero-inflated negative binomial model. *Health Economics*, *13*(8), 781–791.

Simons, J. S., Neal, D. J., & Gaher, R. M. (2006). Risk for marijuana-related problems among college students: An application of zero-inflated negative binomial regression. *The American Journal of Drug and Alcohol Abuse*, *32*(1), 41–53.

Spiegelman, D., Colditz, G. A., Hunter, D., & Hertzmark, E. (1994). Validation of the gail et al. model for predicting individual breast cancer risk. *JNCI: Journal of the National Cancer Institute*, *86*(8), 600–607.

Stefanski, L. A. (1989). Unbiased estimation of a nonlinear function a normal mean with application to measurement error models. *Communications in Statistics-Theory and Methods*, *18*(12), 4335–4358.

Stefanski, L. A., & Carroll, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika*, *74*(4), 703–716.

Thamerus, M. (1998). Different nonlinear regression models with incorrectly observed covariates. In *Econometrics in theory and practice* (pp. 31–44). Springer.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 1701–1728.

Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, *91*(433), 217–221.

Wedderburn, R. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, *63*(1), 27–32.

Winkelmann, R. (2008). *Econometric analysis of count data*. New York, NY: Springer Science & Business Media.

Wong, M., Day, N., Bashir, S., & Duffy, S. (1999). Measurement error in epidemiology: the design of validation studies I: univariate situation. *Statistics in Medicine*, *18*(21), 2815–2829.

Yang, H., Ozbay, K., Ozturk, O., & Yildirimoglu, M. (2013). Modeling work zone crash frequency by quantifying measurement errors in work zone length. *Accident Analysis & Prevention*, *55*, 192–201.

Yang, Y. (2012). *Poisson regression with measurement error in covariates* (Unpublished doctoral dissertation). Hong Kong University of Science and Technology.

LIST OF PUBLICATIONS AND PAPERS PRESENTED

Rozliman, N. A., Ibrahim, A. I. N., & Yunus, R. M. (2017). Bayesian approach to errors-in-variables in regression models. In *AIP Conference Proceedings* (Vol. 1842, p. 030018).

Rozliman, N. A., Ibrahim, A. I. N., & Yunus, R. M. (2018). Bayesian approach to errors-in-variables in count data regression models with departures from normality and overdispersion. *Journal of Statistical Computation and Simulation*, *88*(2), 203–220.