Assessing model stability and sensitivity

O O Onamusi

PhD

March 2022

Assessing model stability and sensitivity

Oluwasikemi Oluwadamilola Onamusi

A thesis submitted in partial fulfilment of the requirements of the University of Northumbria at Newcastle for the degree of Doctor of Philosophy

Research undertaken in the Faculty of Engineering and Environment

March 2022

Abstract

Statistical inferences from observed studies with error prone measurements are often biased. The bias is a consequence of the deviation of the probability distribution that generates the observed data from that which generates the true unobserved data. For example, in binary data where measurement error is a misclassification problem, an observation with a true value of 0 is observed as 1 or vice versa. Past research in this framework often focuses on the use of a validation study to account for measurement error in the main study. A shortcoming of this approach is a lack of validation data to inform the correction of measurement error in the main study. Another challenge is the non-availability of ready to use statistical software in implementation. To overcome some of the challenges of current approach to the analysis of binary data with measurement error, we investigate the performance of the naive logistic regression model, which we refer to as the assumed model, against a modified model. By modified model we mean an extended logistic regression model, where we introduced the probability of measurement error as a modification weight. The modification weight introduced is in the direction of the nondifferential and differential misclassification pattern. Following Cook's (1986) normal curvature approach, we derive an influence measure for the special cases of when the presence of binary outcome $Y_i^* = 1$ is error prone, and also for the absence of the outcome $Y_i^* = 0$. The method is applied to a dataset from the rehabilitation programme study for juvenile offenders, where $Y_i^* = 0$ is measured with error. As different compositions of measured values of binary outcomes often exist in real studies, hence, we further conducted a simulation study for different scenarios for the special cases $Y_i^* = 1$ and $Y_i^* = 0$. Our theoretical results show that when there is no information about the error size, the assumed model appears to be the most stable model compared to the modified models. But, the assumed model estimates could be biased when measurement error is present. Thus, it is important to investigate model stability and how model estimates behave within a plausible range of error size, and report all the findings.

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to thank my mum and dad (Maami and Oldman as I fondly call them) for all the love, support and prayers that they have given me over the years. Without God and you, I would not be the lady that I am today. I would like to appreciate Dr Emmanuel Ogundimu and family, it has been an absolute privilege to work with you. Your eagerness to help throughout the programme has been gratefully appreciated. Your kindness, guidance, encouragement and support is priceless. I look forward to future collaborations. I would also like to thank my siblings and family, and Dr Ojubolamo for their prayers, kindness, love and support. To K my Oga, thank you for your love, support and patience.

I would like to thank Dr Nan Lin for his comments, support and guidance. It has been an absolute privilege to work with you. I also acknowledge the PhD studentship from Northumbria University.

I would also like to thank my fellow PGR students for the support. Finally, I would like to thank Trinity Church Members, the Aromokuns, Akejus, Ayeitigbos, Rajis, Peter Linton, and Jerry Patterson for all the love and support. God Bless.

Declaration

I declare that the work contained in this thesis has not been submitted for any other award and that it is all my own work. I also confirm that this work fully acknowledges opinions, ideas and contributions from the work of others.

Any ethical clearance for the research presented in this thesis has been approved. Approval has been sought and granted by the *Faculty Ethics Committee* on *1st February, 2018.*

I declare that the Word Count of this thesis is 28,256 words (excluding acknowledgements and bibliography).

Name: Oluwasikemi Oluwadamilola Onamusi

Date: 23 March 2022

Chapter 1

Introduction

Problems of model uncertainty and incomplete data arising from measurement error, confounding and missing data are fundamental problems in statistical sciences. As a result, studies in many areas are prone to potential biases which threaten the validity of results. In practice, studies will lead to misleading information if the impact of potential biases are not considered in statistical techniques. However, most researchers normally assume that the statistical model being considered is a true realisation from which the observations or data are generated. Sometimes, a model is chosen for reasons of mathematical conveniences or because the model has been used by previous researchers (Copas and Egushi, 2010). Although it is never possible to completely capture all the potential biases, if a perturbation to the assumed model significantly affects the inference drawn from the analysis, then this may indicates a further investigation into the model's assumptions.

Most of the literature addresses each source of bias separately to achieve better understanding. For example, it is well known that missing data and confounding are frequent issues in observational studies, surveys, and randomised controlled experiments. In the case of missing data, the missing data mechanism can generally be characterised into three types: missing at random (MAR), missing not at random (MNAR), and missing completely at random (MCAR) (Little and Rubin, 2002). Therefore understanding and addressing the reasons, mechanism of missingness, and methods to handle missing data is essential for the methods used for inference. Hence, several methods have been developed for handling incomplete data (see Diggle and Kenward, 1994; Little and Rubin, 2002; Ibrahim et al., 2005; Molenberghs et al., 2008; Little and Zhang, 2011 for a review).

While incomplete (or missing) data and confounding are now increasingly taken into account by means of study design or modification in data analysis, the potential for hidden bias due to measurement error, which can lead to inaccurate estimates of the parameter(s) of interest, is more routinely disregarded. Many researchers incorporate the unknown misclassification probability in the likelihood function of an assumed model, and they mostly performed sensitivity analysis in order to account for measurement error (for example, Cook, 1986; Copas and Li, 1997; Copas and Shi, 2000; Molenberghs et al., 2001). In this way, they were able to examine the influence of uncertainty both from the data and model. This has been used in different ways for the same purposes in assessing the influence of potential biases in an assumed model. Copas and Eguchi, 2005; 2001 discussed local sensitivity analysis when inference is based on missing observations. Zhu and Lee (2001) considered an alternative to likelihood displacement based on the power and wide applicability of the EM algorithm (for full detail on EM algorithm, see Dempster et al., 1977), and proposed a method to assess the influence of a deviation in a model with missing data. The proposed method was extended to generalized linear mixed models and a variety of complicated models. Lin et al. (2012) extended Copas and Eguchi's work and discussed the marginal model bias in confounder problems for generalized linear mixed model with nonlinear link functions.

On the other hand, most literature on measurement error is based on methods used to account for either error prone outcomes and/or covariate(s) which might be present in observed data. Very little has been done to assess the local influence in a minor perturbation of a binary model with outcome misclassification. Motivated by recent advances in assessing local influence relating to incomplete data, the rest of this chapter presents the overview of binary outcome misclassification, and an approach to account for the error is highlighted in Section 1.1. In Section 1.2, the underlying existing methods of influence analysis are illustrated. Followed by the aims and objectives in Section 1.3. The thesis structure is presented in Section 1.4.

1.1 Binary Outcome Missclassification

There are inevitable problems in measurements where an observation(s) is classified or measured in error for a variety of reasons. For example, in medical sciences, precise classification may only rarely be possible. It has been suggested in the literature to construct a mental health scale (Bross, 1954) with some critical points for classification of whether an individual falls below or above the scale. For some diseases some classifications may contain considerable risk of error, while some may involve almost no risk of error. Some researchers are of the opinion that the effect of misclassfication would cancel out and the usual (commonly referred to as naive) analysis can be done. On the other hand, some argued that no conclusion can be drawn from naive analysis since data are prone to error.

Suppose Y_i is the unobserved true response variable, for $i = 1, \cdots, n$ with n being the number of subjects and X_i are the true covariates. The unobserved true response variable takes one of only two possible values, that is, $y_i = 1$ represents presence of an attribute of interest and $y_i = 0$ otherwise. Suppose the true distribution for Y_i is $Y_i | X_i \sim Bin(1, \pi_i)$, and

$$\pi_i = P(Y_i = 1|X_i) = \frac{\exp(\alpha + \theta X_i)}{1 + \exp(\alpha + \theta X_i)}$$

where θ is the unknown true parameter of interest, α is the nuisance parameter. In the case of the observed data, the observed response variable is assumed to be measured with error. When measurement error is not considered, the assumed logistic model (commonly referred to as naive analysis) is $Y_i^*|X_i \sim Bin(1, \pi_i^*)$, and

$$\pi_i^* = P(Y_i^* = 1|X_i) = \frac{\exp(\alpha^* + \theta^* X_i)}{1 + \exp(\alpha^* + \theta^* X_i)}$$

where θ^* is the observed parameter of interest for Y_i^*. The corresponding log-likelihood function is

$$\ell(\theta^*) = \sum_{i=1}^{n} y_i^* \log(\pi_i^*) + (1 - y_i^*) \log(1 - \pi_i^*) \tag{1.1.1}$$

However, it has been suggested in the literature that estimates of θ^* obtained from naive analysis (1.1.1) are biased for θ, since errors in the outcome variable can lead to loss of essential information about covariate effects (see Baron, 1977; Greenland, 1980; Neuhaus, 1999; Carrol et al., 2006; Lyles et al., 2011).

Hence, there are several different approaches in the literature used to account for measurement error. Some of these are regression calibration (see Carroll and Stefanski, 1990; Thurston et al., 2003; Freedman et al., 2008), and multiple imputation (Cole et al., 2006). Another approach used in literature to account for measurement error is maximum likelihood estimation for the main study data with additional validation data (Karen and Loki, 2008; Lyles et al., 2011). A validation study can be internal or external. An internal validation study is a portion of observed (commonly referred to as main study in literature) data. For example, suppose $y_1^*, ..., y_k^*, ...y_{k+m}^*, \cdots, y_n^*, x_1, \cdots, x_k, \cdots, x_{k+m}, \cdots, x_n, y_k, \cdots, y_{k+m}$ is the observed data. Then $y_k^*, ...y_{k+m}^*, x_k, ..., x_{k+m}, y_k, ..., y_{k+m}$ is the internal validation data contained in the observed/main data. External validation data is a set of data outside of the main study data to help correct for measurement error, which contains for example, $y_1^*, \cdots, y_v^*, y_1, \cdots, y_v, x_1, \cdots, x_v$. In Chapter Two, we will show how to use the internal validation subgroup combined with the main study to account for measurement error.

However, when validation substudy data is not available, the most common approach in literature is to vary the value of probability of measurement error between 0 and 1. Here, we denote probabilities of measurement error to be $\zeta = P(Y_i^* = 0|Y_i = 1)$ and $\gamma = P(Y_i^* = 1|Y_i = 0)$. Thus, sensitivity is the probability of being observed to be true (that is, $Y_i^* = 1$) given that the unknown outcome is true $Y_i = 1$. Similarly, the probability of being observed to be false (that is, $Y_i^* = 0$) given that the unknown outcome is false $Y_i = 0$, is referred to as specificity. The mechanism of measurement error of a binary outcome is shown in Figure 1.1. An overview of some of the methods used in accounting for measurement error is described in Chapter Two.

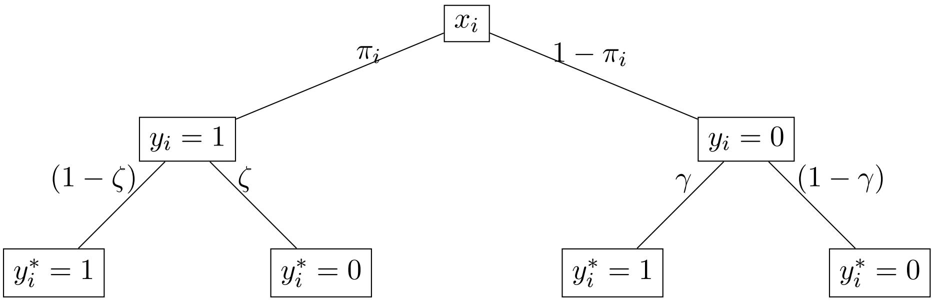


Figure 1.1: Mechanism of measurement error of binary outcome

Broadly, there are several papers in the literature where researchers have described frequentist and Bayesian methods to account for measurement error in the response and/or covariate(s). However, it appears there are very few approaches to assess the discrepancy between naive (assumed) analysis and the underlying model from which the data was generated using the above mentioned methods for accounting for measurement error. Therefore, it is important to investigate how stable an assumed or any other models are under the influence of measurement error, especially when there is no validation subgroup.

1.2 Influence Analysis

There is cause for concern if a slight modification to an assumed model influences key results or conclusions drawn from an analysis. Otherwise the assumed model is said to be robust if such modification will do no harm (Barnard, 1980). More confidence can be put in an assumed model which is relatively stable under slight modification as highlighted by (Cook, 1986). Hence, it becomes imperative to examine a sensitivity study or an influential analysis of model assumptions as discussed by Critchley et al., (2001), since a statistical model is more or less a realisation of a scientific process from which the data is generated (Cook, 1986).

1.2.1 Case Deletion

Initially, measures such as studentized residuals, estimated variances of the predicted values as illustrated by Behnken and Draper, 1972; and sum of squares of the variances of the predicted values, a robust design approach developed by Box and Draper, 1975 are used for detecting data point(s) that contributed or exert undue influence on summary statistics arising from data analysis, especially in linear regression models. These measures were believed to contain useful information such that, when a potential influential data point has been detected, it follows that the analysis of the effects of deleting the data point is next (Cook, 1977). However, one fundamental issue is the question of which data point(s) is to be deleted from an observed data. As a result, a new measure that incorporates information from both studentized residuals and variances of the residuals and predicted values was developed by (Cook, 1977). The measure is used to determine the degree of influence each data point has on the estimated parameter of interest.

An example of influence analysis in case deletion of linear regression is as follows: Suppose observed data containing n observations $(y_1, x_1), \cdots, (y_n, x_n)$. Let $\hat{\theta}$ represent the maximum likelihood estimate obtained from a linear regression model based on the full observed data. Next, delete the $i'th$ subject and recalculate the maximum likelihood estimate, denoted as $\hat{\theta}_i$. If the removal of the $i'th$ subject causes substantial change in the results of the analysis, it may be necessary to inspect the effect or impact of the corresponding subject. This suggests that the assumed linear regression model might be sensitive to the modification. Hence, Cook (1977) provides a measure to judge the contribution of each observed data point, which can be assessed by examining the difference between $\hat{\theta}$ obtained from the full data, and $\hat{\theta}_i$ a maximum likelihood estimate calculated from the data without the $i'th$ subject. This is commonly known as the Cook distance D_i, defined as

$$D_i \equiv \frac{(\hat{\theta}_i - \hat{\theta})' X' X (\hat{\theta}_i - \hat{\theta})}{ps^2}$$

where D_i is obtained for each subject ith, p is the dimension of θ (that is, the number of covariates for each subject), X is an $n \times p$ full rank design matrix including a constant, with n being the number of subjects, and s^2 is the mean squared error of the assumed linear regression model. D_i provides varying information in relation to the effect of the ith subject on the fitted model, which can be used to determine observations not well explained by the fitted model.

Furthermore, in the case of a logistic regression model, Pregibon (1981) highlighted that asymptotic arguments suggest that the deviance and chi-squared statistics should identify data point(s) dominating part of the fitted model and observations not well explained by the assumed model. However, these statistics cannot measure the individual effects on the many components of the fitted model, such as the estimated parameter(s), estimated covariance matrix of the parameters, and individual standard errors. Hence, he suggested two approaches for assessing the impact or effect of each data point on a fitted model. One is assessment by deletion as described above, and the second is assessment by small modification. Examples of modifications include: A weight is attached to an individual subject generally referred to as case weight, or changes in the individual parameter estimates. Other modifications are possible.

However, the case deletion idea described above has major concerns since it allow only one of two possibilities; either a case is retained or it is removed, which may cause a considerable amount of change in the results obtained from an analysis. Rather than deleting a case from the full observed data, we can perturb it by incorporating modification in an assumed fitted model, where the ith subject becomes 0 when the case is deleted and

1 otherwise.

1.2.2 Case Weight Modification

With modification, the case deletion idea was extended to accommodate varying weights using sensitivity analysis rather than deleting cases. This is commonly referred to as a case weight diagnostic procedure, where a modification parameter, say γ, is attached to an assumed model, such that a case or the i'th subject is allowed to have possible values of γ while the remaining cases have $\gamma = 1$. Hence, we can investigate the i'th subject over the entire range of γ, using the more general influence statistics D_γ as proposed by Cook, where D_γ is

$$D_\gamma = \frac{\|\hat{Y} - \hat{Y}_\gamma\|^2}{p\sigma^2}$$

where $\hat{Y}$ and $\hat{Y}_\gamma$ are the predicted values obtained from the assumed model and a model over the specified range of values for weight γ respectively.

Since the modification parameter γ can be any well-defined $q \times 1$ perturbation vector, Cook (1986) developed a new approach for handling a variety of model modification. This approach, generally known as local influence, is used for investigating how the results of an analysis change when a statistical model is under minor modification. A general discussion on using a likelihood approach to evaluate the consequences of model deviation was highlighted by (Barnard, 1980). Thus, Cook's concept of a local influence approach is developed following the idea of likelihood displacement which is based on comparing the likelihood function of an assumed model $\ell(\theta^*)$ to the likelihood function $\ell(\theta_\gamma)$ of modified model (that is, a model where modification, or weights γ are introduced in the assumed model, varying throughout space Ω, but bound to some $\Omega \subset \mathbb{R}^q$).

In influence analysis, curvature of likelihood displacement is used to study the behaviour of the log-likelihood function of a model over the entire range of the modification parameter γ. This provides different information from that of Cook's distance. Hence, the concept of curvature of likelihood displacement uses differentiation (see Chapter Three of this thesis) rather than the method of differencing in subject deletion analysis. Likelihood displacement is generally expressed as

$$LD(\gamma) = 2[\ell(\theta^*) - \ell(\theta|\gamma)]$$

In addition, the method uses normal curvature from differential geometry from its use of an influence graph (that is a graph of LD(γ) versus γ) to measure the effect of influence. An overview of Cook's local influence is detailed in Section 3.1 of Chapter Three.

Poon and Poon (1999) pointed out that the computation of eigenvalues and eigenvectors in Cook's approach may be difficult in problems with large dimensions, where there may be no objective criterion to judge the magnitude of normal curvatures and the relative size of the components of the corresponding directions. As a result, they developed another measure (named conformal curvature) said to be a one-to-one function of normal curvature, ranging between 0 and 1. The conformal curvature method is an effective measure of local influence, which could be used as a criterion or benchmark to judge magnitude of curvature and could enhance the applicability in a theoretical framework. Hence the computation of eigenvectors is no longer necessary (Poon and Poon, 1999).

On the other hand, Cook's concept of normal curvature has become a key means for researchers to develop influence measures used to evaluate the difference between an assumed model and a perturbed model. There is extensive research activity where normal curvature has been used to detect influential data points. Cook's (1986) concept of normal curvature has a feature that relies on log-likelihood and differential geometry which provides a general framework for assessing influence of deviation in a statistical model. This has been used to describe and propose diagnostic measures in a variety of ways. For example: detection of influential observations when a weight is introduced in linear regression (Cook, 1986); modification of the missing-at-random model assumption in the direction of a non-random dropout model, this method provides a framework for non-random dropout for incomplete longitudinal data (Verbeke et al., 2001); a statistic based on local influence measures was proposed to identify influential data points and for checking the fit of a model (Zhu and Zhang, 2004).

Verbeke et al. (2001) developed a measure to assess the influence of nonrandom dropout on the parameter of interest in incomplete repeated continous data assumed to be under the MAR (Missing-at-random) mechanism. A dropout in the analysis of a longitudinal study occurs when some sequences of responses are missing or terminate early from the study. The three major missingness techniques are Missing at random (MAR), Missing not at random (MNAR) and Missing completely at random (MCAR). Comprehensive information about missing data mechanisms can be found in Rubin, (1987). Verbeke et al. (2001) illustrated how a missing at random model can be modified in the direction of non-random dropout, in order to investigate subjects that have considerable influence on the model parameters. The proposed influence measure is based on normal curvature, an overview of which is presented in Chapter Three of this thesis. Verbeke et al. (2001) emphasized to consider modification of the missing not-at- random process, and that the study of local influence measures can provide essential information into which subject influences model parameters.

Steen et al. (2001) followed Verbeke's approach to propose a measure to assess the influence of the parameter describing non-random dropout and the measurement process parameter of incomplete repeated ordinal data. They used normal curvatures at the point $\gamma = 0$ (an assumed model before perturbation) for each direction h_j of $\gamma_1, \cdots, \gamma_q$ as a measure to assess influence.

Zhu and Zhang (2004) stated that most influence measures based on normal curvature proposed in the literature only identify influential subjects or observations, but that the approach seemingly could not be used to assess the difference between the underlying model and an assumed model which does not belong to the true model. Hence they proposed a measure based on normal curvature for judging the adequacy of an assumed model for a case weight modification scheme. Zhu and Zhang (2004) highlighted the measure as a diagnostic approach for checking overall fit of the postulated model. Therefore it can be adapted to generalised linear models.

Selecting an inappropriate modification vector to a statistical model may lead to misleading inferences about the effect of the modification, since the perturbation vector is central to the development of inference measures (Zhu et al., 2007). However, choice of a modification vector has been widely neglected, and development of most influence measures is limited to normal curvature of an objective function at a point with a zero first derivative (critical point). This is another major drawback with conformal curvatures and Cook's normal curvature. Hence, Zhu et al.(2007) proposed a measure (named perturbation manifold) based on the definitive clarity of metric tensors, and finely-adjusted connections to choose an appropriate perturbation to a statistical model. In addition, the measure can assess the local influence of minor modification to a statistical model by using the first and second derivatives of likelihood displacement. Chen et al.(2010) applied the perturbation selection method of Zhu et al, (2007) in the context of Generalized Linear Mixed Models. Chen et al.(2010) considered four different modification scenarios but used metric tensors to determine the appropriate perturbation, with second order measure in local influence analysis (Chen et al.,2010).

Frequentist methods are not the only means used to assess influence of modification of a statistical model. For example, Cho et al.(2009) developed case-deletion influence based on the Kullback–Leibler (KL) divergence, to assess the influence of a case on joint and marginal posterior distributions for Bayesian survival models with continuous survival time data. The 3 key elements of Bayesian analysis are the data $y = y_1, \cdots, y_n$, the sampling distribution $P(y|\theta)$, where $\theta = (\theta_1, \cdots, \theta_p)^T$ are the parameters of interest, and the prior distribution $p(\theta)$ of the parameters. In order to identify outliers and influential observation(s), the influence of an individual y_i or a set of observations is often assessed by deleting observation(s), and then comparing the predictive and/or posterior distribution $p(\theta|y)$ based on the full data $y_1, \cdots, y_n$ compared to that of posterior and/or predictive distribution of the deleted observation(s). With regard to the above-mentioned three key elements of Bayesian analysis, there are three major formal influence methods for measuring the degree of dependence to which posterior distributions are sensitive. These three are case influence, global robustness and local robustness techniques (Berger, 1990, 1994). The most popular case influence measures, including posterior probabilities of an outlying set of observations, posterior outlier quantities, predictive diagnostics, and posterior diagnostics for identifying outliers and influential points, are developed by using either modification or deletion of observation(s) (Zhu et al., 2010). In the Bayesian literature, little research has been done on developing unifying influence measures based on the Bayesian approach for assessing the effects of perturbing the prior, sampling distribution, or the data (Berger et al. (2000)). Hence, a unifying geometric framework of influence measures based on the Bayesian perturbation manifold is developed by Zhu et al., (2011). Their proposed measure can be used to quantify the effect of various modification schemes on statistical models, and is very much applicable for many complex Bayesian models.

Zhu et al. (2012) highlighted that Bayesian case influence measures using case deletion are difficult to implement computationally in most models with missing data. Hence, they suggest using Markov chain Monte Carlo (MCMC) samples from the posterior distribution $p(\theta|y)$ based on the full data to compute first-order approximations to the Bayesian case influence measures. This was examined and illustrated for latent variable models with missing data (see Zhu et al., 2012).

1.3 Aims and objectives

In this thesis we aim to determine the sensitivity or robustness of a logistic regression model to the impact of measurement error. The primary focus will be the use of normal curvature from the differential geometry to provide essential information about any given model, modified to account for outcome measurement error. A modification of a logistic model in the direction of nondifferential and differential misclassification patterns (see detail in Section 4 of Chapter Three, and Chapter Four) present in observed data with binary outcomes will be explored. This will permit the introduction of measurement error parameters in (1.1.1) (that is, the log-likelihood likelihood function of a logistic model, assumed to be without error). Although several researchers accounted for misclassification through the use of main/external or main/internal validation study data, we will assume values from 0 to 1, since obtaining accurate values for misclassification probabilities might be difficult. Each assumed value is a distinct model. Thus, assuming values from 0 to 1 will yield several models corresponding to each of the combinations of the assumed varying values of ζ and/or γ.

1.3.1 Motivation

We consider a motivating example taken from a rehabilitation programme study for juvenile offenders. There are 2 groups of juvenile offenders who participated in the study, one group joined the rehabilitation programme, the other did not. A cross-sectional analysis of the data was considered, where rehabilitation programme status is the exposure X and reoffending is the outcome of interest Y^*. The primary objective of the study was to examine whether the rehabilitation programme can reduce the reoffending rate amongst juvenile offenders within 2 years. A naive analysis suggests a positive relationship between reoffending and the rehabilitation programme. However, the problem of outcome misclassification arising from undetected crime would distort the validity of the estimated odds ratios (ORs) and might lead to misleading conclusions, since the naive analysis ignores the problem of measurement error. Figure 1.2 depicts the structure of the misclassification rate present in the rehabilitation study data where;

- y_i is the true response variable (0 if an offence has occurred, 1: otherwise)
- y_i^* is the observed response variable (0 observed re-offence, 1: otherwise: non-offending)
- X_i is a binary treatment variable (1; joined rehabilitation programme, 0 did not join the programme)
- π_i is the unknown true probability of not reoffending
- γ is the probability of undetected crime

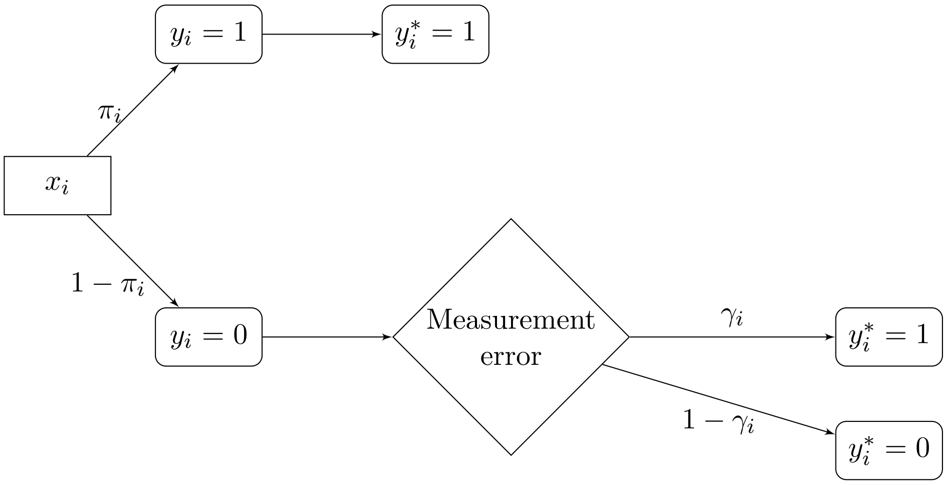


Figure 1.2: Undetected Crime: Misclassification of occurrence of observed re-offence $y_i^* = 0$.

In addition, there is a fundamental issue about the choice of appropriate assumptions surrounding the potential misclassification rate that may be present in the rehabilitation programme data. The misclassfication pattern with respect to the outcome variable (that is, reoffending) in our motivating example may assume nondifferential or a differential approach. A nondifferential misclassification scheme assumes the unknown probability of measurement error is constant based on a participant's specific variable, while a differential misclassification pattern assumes the error-prone variable is dependent on other covariate(s) of interest. As a result, it is important to examine into the different measurement error patterns in outcome Y^* and assess its impact in the stability of the assumed fitted model or modified model. In our example, exploration into the misclassification rate of reoffending may be invariant to exposure status as in the case of nondifferential misclassification or may depend on whether an offender joined $x_i = 1$ or did not join the programme $x_i = 0$.

Therefore, to investigate the robustness of the model fitted from the naive analysis (assumed model) in adjusting for undetected crime, the likelihood function of the assumed model will be modified to account for potential measurement error present in the rehabilitation study data. This can be achieved by incorporating the unknown probability of measurement error (that is, probability of undetected crime) in the standard logistic regression model. In the literature there are existing methods used in accounting for response measurement error. We adopted the maximum likelihood estimator approach since it provides easy numerical optimisation with statistical software R used for our analysis. Further, the maximum likelihood technique has been proven to provide consistent estimates compared to other methods such as regression calibration, SIMEX-Simulation extrapolation method, moment reconstruction, or multiple imputation - MI, (see Karen and Loki, 2008).

Each misclassification pattern assumed gives separate, or several, models since the probability of undetected crime varies under individual assumptions. Hence there is need to examine how each model changes when misclassification is altered compared to the naive or standard logistic regression fitted when no measurement error (undetected crime) is considered. Cook's normal curvature approach assesses how an assumed model changes around its small neighbourhood when an observation(s) or subjects are deleted. However, we extended the normal curvature approach to examine each model obtained under the individual misclassification patterns.

1.3.2 Methodology

We first illustrated Cook's influence approach using a simulation scenario from the motivating example data. Cook's influence method follows the idea of likelihood displacement which is based on comparing between models. The illustration was carried out using a simulation method where the binary response variable is mismeasured. In the illustration; we provided an assumed model (commonly referred to as naive the model) which does not account for measurement error, possible modified models, and an influence graph (likelihood displacement versus γ, see Section 3.1). Each point on the graph represents different models based on the value of misclassfication rate γ.

Next, an overview of curvature in differential geometry is introduced because our proposed method is based on curvature. Cook used normal curvature of an influence curve or surface to investigate if a modification to an assumed model would change results or conclusions drawn from such analysis, and, most importantly, to find the cause of the change if necessary. Cook's concept of normal curvature has become a key means for researchers in developing influence measures. In Section 3.2 of Chapter 3, we describe how to calculate curvature of a plane curve. This is because a plane curve can be shown from an analysis that incorporated a scalar modification parameter in an assumed model. However, when more than one modification parameter is introduced in an assumed model, we showed a surface and not a plane curve. There are infinitesimal curves passing through a point on a surface. As a result, there are infinitesimal curvatures associated with the curves. In this case, it becomes necessary to present how to calculate normal curvatures to study the properties of the curves. Hence, an overview of normal curvature (see Section 3.2) is presented.

From the curvature formula presented, we derived an influence measure for misclassified binary outcome. The influence measure developed is based on normal curvature. There are two fundamental misclassification patterns that may potentially invalidate estimates obtained from an analysis.

1. Nondifferential Mislassification: This occurs when the unknown misclassification probability of a variable in the study data is constant and does not depend on other variables
2. Differential Misclassification: In this case, a differential misclassification pattern assumes the potential probability of measurement error of a variable depends on other variables, hence misclassification rates vary.

Next, an influence measure for nondifferential misclassification pattern (Section 3.4) present in our motivating example is derived. Similarly, an influence measure based on normal curvature under potential independent differential misclassification rates present in the rehabilitation programme study data is described.

The influence method derived for the measurement error problems present in the motivating example can be extended to other measurement error problem which may occur in practice. Hence, a general framework based on our method is derived.

Furthermore, since there is no magic process telling us which model is correct as highlighted by Copas (2005), we assessed the influence of measurement error in any given model since a model is meant to account for measurement error. Most especially, the derived influence measure is used to investigate the stability of any possible model.

1.4 Thesis structure

The rest of the thesis is structured as follows. Chapter Two reviews key existing literature of methods used in accounting for measurement error, including occasions when a binary response variable and/or covariate(s) is mis-measured. From the frequentist paradigm, the likelihood functions developed under most of these methods that account for misclassification are modifications of an assumed model. Chapters Three - Four of this thesis focus on influence measures developed based on normal curvature for a logistic regression model under the impact of measurement error. An overview of curvature both in R^2 and higher dimensions is presented. More specifically, in Chapter Three an influence measure is developed under the impact of potential non-differential and differential misclassification patterns that may be present in the absence attribute (that is $Y_i^* = 0$) of an observed binary outcome with measurement error. In Chapter Four, we extend the method to when only the presence of an observed binary outcome of interest is subject to error (that is, $Y_i^* = 1$). In this case, we derive curvature, and the analytical expressions for the influence measure methods are formulated. In Chapter Five, the method is applied to our motivating example, and a simulation study is performed. For both cases $Y_i^* = 0$ and $Y_i^* = 1$ presented in Chapter Three and Four, we investigated how stable an assumed or modified model is under the impact of nondifferential and differential measurement error. By assuming values for the probability of measurement error, we examine the behaviour of each distinct model. In Chapter Six a conclusion for the work is presented, and possible future work is highlighted.

Chapter 2

Methods that account for measurement error

There exists a breadth of literature on different approaches used to account for measurement error. Cole et al. (2006) highlighted that all of these methods use information that maps observed measurements to true values based on a validation study as proposed by Spiegelman et al., (2001) or sensitivity analysis in which the mapping may be based on prior information or speculation as highlighted by Greenland, (1996). Some of these methods include regression calibration, maximum likelihood estimation for main study data with internal or external validation data, multiple imputation and Bayesian method where prior information is provided for sensitivity and specificity parameters. They can be grouped into two, that is, maximum likelihood and Bayesian approach, depending on their method of estimation. For example, in adjusting for covariate(s) misclassification, validation-based design was used to compare maximum likelihood, multiple imputation methods and regression calibration methods (Karen and Loki, 2008); Carroll and Stefanski, (1990) proposed regression calibration method for adjusting for measurement error. Cole et al. (2006) illustrate a multiple imputation approach to account for measurement error in survival analysis. Lyles et al. (2011) demonstrate a likelihood-based approach that makes use of main study, external, and internal validation data to account for measurement error in response variable. From the Bayesian perspective, Gerlach et al. (2007) and McInturff et al. (2004) incorporate prior assumption of misclassification probabilities to augment mismeasured data; Richardson and Gilks (1993) used Gibbs sampling for studies with continous measurement error; Stamey et al. (2008) adjusted for misclassification in poisson regression using Bayesian method. Although the methods for correcting measurement error are based on both the freqentist and the Bayesian paradigm, the method that is proposed in this thesis is based on maximum likelihood estimation principles. In addition, the focus will be on measurement error/misclassification problem in binary data. Consequently, we provide a review of the maximum likelihood method for correcting measurement error in binary data.

While some of the methods for correcting measurement error will not directly influence the formulation of the proposed method presented in Chapter Three, their relevance in the field of modifying a binary regression model in the direction of misclassification warrants their inclusion. In this chapter, the aim is to provide the background of deviating an assumed logistic model in the direction of a model that accounts for measurement error. The structure of the rest of the chapter is as follows. In Section 2.1, some of the methods used in adjusting for measurement error in exposure variables are illustrated. Section 2.2 shows how to account for misclassification when outcome variable is measured with error. In real application, both exposure and response variables are measured with error. Thus, in Section 2.3, measurement error adjustment based on different types of misclassification patterns present in both error-prone outcome and exposure variables are demonstrated. Bayesian perspectives of accounting for measurement error is highlighted in Section 2.4. The structure of this chapter is summarised in Figure 2.1. The acronyms for the methods to be discussed are listed below:

- RC - Regression Calibration
- MR - Moment Reconstruction
- MI- Multiple Imputation
- ND - Maximum likelihood based on Nondifferential misclassification
- IND - Maximum likelihood based on Independent Differential misclassification
- DD- Maximum likelihood based on Dependent Differential misclassification

2.1 Measurement Error in Covariates

When referring to measurement error, it has been noted that it does not only arise or identify in response variable, but the problem of misclassification is common to covariates as well in applied research. For instance, in an epidemiology setting where the primary interest of analysis centers on the effect of exposure on disease outcome, exposure misclassification may occur due to an imperfect condition of the measurement tool characterising the observed exposure variable, or participants not wanting to give too much information. In this case the relationship or association between the exposure factor and the disease response is affected if measurement error is ignored, such that the estimated parameters linking the exposure factor to the disease response are biased. There are a variety of proposed methods which can be used to adjust for measurement error in covariates. In this Section, five different approaches are illustrated. They are regression calibration, moment reconstruction, multiple imputation, maximum likelihood based on non-differential covariate misclassification, and maximum likelihood based approach of differential mis-

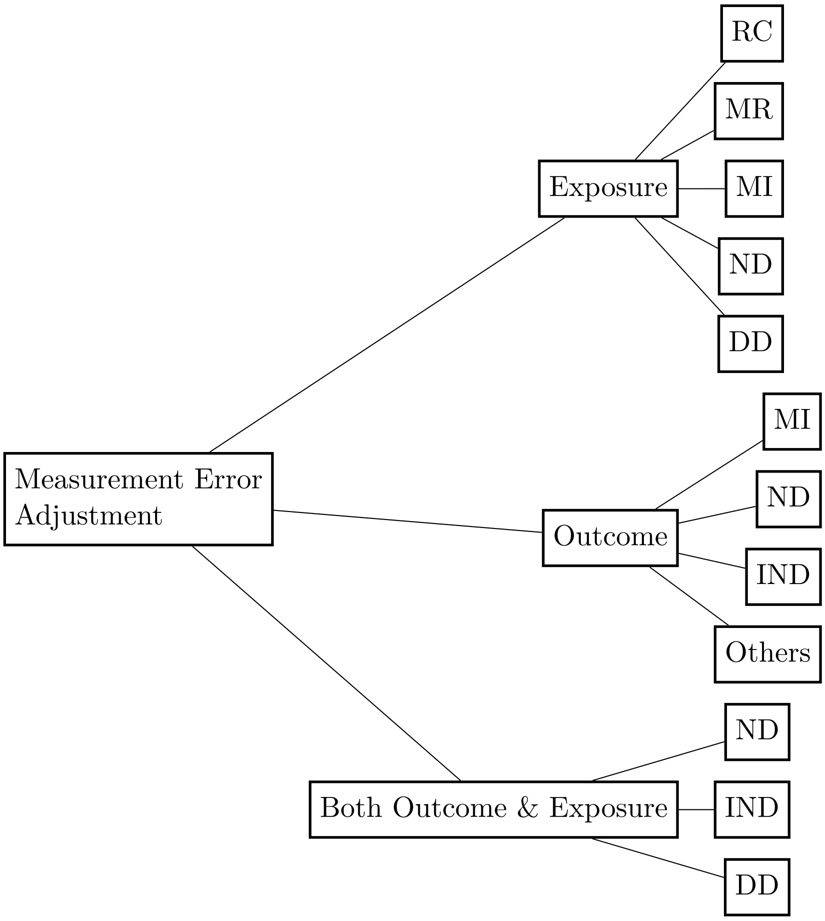


Figure 2.1: Structure of Chapter Two. The methods discussed are RC - Regression Calibration, MR - Moment Reconstruction, MI- Multiple Imputation, ND - Maximum likelihood based on Nondifferential misclassification, IND - Maximum likelihood based on Independent Differential misclassification, DD- Maximum likelihood based on Dependent Differential misclassification

classification patterns associated with error-prone exposure variable(s). It is important to know that some of these methods used information from a validation subgroup study in estimating parameters of interest. A validation study can either be a subset from observed data (internal validation) or from external sources.

2.1.1 Regression Calibration

One of the most commonly used method for accounting for measurement error in explanatory variables is Regression calibration (Carroll and Stenfaski, 1990; Gleser, 1990). The main idea is to replace mismeasured measurements with substitute values and then perform analysis as would be used with an error-free measurement. The substituted value is the expectation of the true measured value conditional on the observed or error-prone measurements. For instance, when an observed response variable Y is associated by regression model with an unknown error-free covariate X, but imprecise/mis-measured covariate X^* is observed, then the main idea of regression calibration is to replace the unknown true covariate X with an adjusted/estimated value in order to estimate the parameter of interest(s). In applied research, the effect of bias is underestimated when one variable is mis-measured. However, when one or more variables measured with error are included in an analysis of data, estimates of the effect may be overestimated or underestimated for error-free measured covariates (see Greenland, 1980; Rosner et al., 1990).

Suppose observed data y_i, x_i^* and z_i contain misclassified covariate(s), where x_i^* is assumed to be misclassified or error-prone $p_1 \times 1$ covariates, z_i is some other $p_2 \times 1$ covariate(s) assumed to be error-free, and y_i is a binary outcome (with $y_i = 0$; presence of an attribute, and $y_i = 1$ otherwise) of interest measured without error as well. Let $\mathbf{x_i}$ be the $p_1 \times 1$ corresponding vector of unknown true covariates. Suppose a logistic regression is assumed for y_i, x_i and z_i where y_i is associated with z_i, x_i such that $Y_i \sim Bin(1, \pi_i)$ and

$$\pi_i = \frac{\exp(\alpha + \theta_1 x_i + \theta_2 z_i)}{1 + \exp(\alpha + \theta_1 x_i + \theta_2 z_i)} \tag{2.1.1}$$

where α is a nuisance parameter, and θ_1, and θ_2 are the $1 \times p_1$ and $1 \times p_2$ vectors of parameters of interest for x and z respectively.

Since x^* is prone to error, the distribution of the outcome variable Y_i is $Y_i \sim Bin(1, \pi_i^*)$, and

$$\pi_i^* = \frac{\exp(\alpha^* + \theta_1^* x_i^* + \theta_2^* z_i)}{1 + \exp(\alpha^* + \theta_1^* x_i^* + \theta_2^* z_i)} \tag{2.1.2}$$

is fitted on the observed data y_i, x_i^* and z_i, where the superscript * in θ_1^*, and θ_2^* in (2.1.2) denotes the parameters are from observed data. Thus, the corresponding regression coefficient θ_1^* will be imprecise. Consequently the corresponding true regression coefficients θ_1 in (2.1.1) will be underestimated (Rosner et al., 1990).

It follows that if there are more than one error prone covariates, each of the regression coefficients will be imprecise and the precision of the covariates measured without errors will also be affected. As a result, Rosner et al. (1990) developed their regression calibration method to correct measurement error when one or more covariates are imprecise or measured with error. This method is suitable and can be used to estimate the true regression coefficients or the parameters of interest from cohort studies with rare disease outcomes and mismeasured continous covariates. The practical implementation of the regression calibration method requires a validation study to include a subset of the main/observed study commonly referred to as internal validation subgroup study or a distinct (or external) study design. It is worth noting that the validation subgroup study contains both x_i and x_i^*, where x_i is the true covariate measured at the same time with other covariates x_i^* and z_i. Some of the assumptions made by Rosner et al., (1990) for cohort design are

1. The conditional distribution of true covariate(s) x_i given the observed covariates x_i^* and z_i is assumed to be the same for the observed study and that of validation study

2. In addition, the distribution of observed covariates x_i^* given true covariates x_i is the same for both cases. That is, $y_i = 1$, and that of control group $y_i = 0$. In otherwords, $P(x_i^*|x_i, y_i = 1) = P(x_i^*|x_i, y_i = 0)$

And based on the assumption of rare disease outcome, Rosner et al. (1990) proposed the following procedure to correct for measurement error present in multiple covariates. These are described as follows;

- Fit a logistic regression model (2.1.2) to the observed study data (y_i, x_i^*, z_i), and obtain estimates of the regression coefficients $\hat{\theta}_1^*, \hat{\theta}_2^*$ which are assumed to be biased coefficients.

- Suppose there exist a multivarate linear relationship between true covariates x_i and some other covariates x_i^*, z_i of the validation study data (y_i, x_i, x_i^*, z_i), such that

$$x_i = \alpha_0 + \alpha_1 x_i^* + \alpha_2 z_i + \epsilon \tag{2.1.3}$$

  where $\alpha_0, \alpha_1, \alpha_2$ corresponds to the $p_1 \times 1$, $p_1 \times p_1$, $p_1 \times p_2$ matrices of regression coefficients. And $\epsilon \sim MVN$ (that is, ϵ follow multivariate normal distribution with $p_1 \times 1$ mean vector 0 and $p_1 \times p_1$ covariance matrix Σ). Hence, fit a multivariate linear model to the validation study, so as to obtain estimates from (2.1.3).

- The parameters of interest are the unknown true regression coefficients $\theta = (\theta_1, \theta_2)$, where θ_1 is the $1 \times p_1$ vector with respect to the unknown true log odds ratio of outcome y on covariates x after adjusting for z. And θ_2 is the $1 \times p_2$ vector with respect to the unknown true log odds ratio of outcome y on covariate z after adjusting for x. The next step is to estimate the true parameters of interest which is the vector of unknown regression coefficients $\theta = (\theta_1, \theta_2)$ from (2.1.1). The true estimated regression coefficients as shown in Rosner et al, (1990) is expressed as

$$\hat{\theta} = \hat{\alpha}^{-1}\hat{\theta}^* \tag{2.1.4}$$

where $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$, $\hat{\theta}^* = (\hat{\theta}_1^*, \hat{\theta}_2^*)$,

$$\hat{\alpha} = \begin{pmatrix} \hat{\alpha}_1 & \hat{\alpha}_2 \\ 0_{p_2 \times p_1} & I_{p_2 \times p_2} \end{pmatrix}$$

where $0_{p_2 \times p_1}$ is a $p_2 \times p_1$ zero matrix (all entries are zero), and $I_{p_2 \times p_2}$ is a $p_2 \times p_2$ matrix with 1 on the diagonal and 0 elsewhere commonly referred to as identity matrix.

When the probability of the response variable is small, then

$$P(y|x, z) = \frac{\exp(\alpha + \theta_1 x_i + \theta_2 z_i)}{1 + \exp(\alpha + \theta_1 x_i + \theta_2 z_i)} \cong \exp(\alpha + \theta_1 x_i + \theta_2 z_i)$$

And

$$P(y|x^*, z) \cong \exp(c + \alpha_1 \theta_1 x_i^* + (\theta_2 + \theta_1 \alpha_2) z_i).$$

Thus, the estimate $\hat{\theta}_1^*$ calculated from the naive analysis in (2.1.2) will be a consistent estimate of $\theta_1 \alpha_1$, and $\hat{\theta}_2^*$; a consistent estimate of $\theta_2 + \theta_1 \alpha_2$. This leads to the multivariate linear approximation estimator θ in (2.1.4).

- Next is to calculate the variance covariance matrix $Cov(\hat{\theta}_1, \hat{\theta}_2)$. Following the multivariate delta method described by Rao, (1973), from (2.1.4)

$$cov(\hat{\theta}_{k_1}, \hat{\theta}_{k_2}) \cong (V' \sum_{\theta^*} V)_{k_1,k_2} + \hat{\theta}^* \sum_{V_{k_1,k_2}} \hat{\theta}^{*'} \tag{2.1.5}$$

where

- $V = \hat{\alpha}^{-1}$ as described in (2.1.4)
- k_1, k_2 are $k_1 th$ and $k_2 th$ columns of V, where $k_1, k_2 = 1, \cdots, p$, and $p = p_1 + p_2$
- $\hat{\theta}_{k_1}$ and $\hat{\theta}_{k_2}$ are the unknown $k_1 th$ and $k_2 th$ true parameters of interest
- $\sum_{\theta^*}$ is the $p \times p$ variance covariance matrix of estimate $\hat{\theta}^*$ obtained from naive analysis model fitted to observed data in (2.1.2)
- $\sum_{V_{k_1,k_2}}$ is the $p \times p$ variance-covariance matrix based on the entries in the $k_1 th$ and $k_2 th$ columns of V. This can be obtained by fitting a multivariate regression model to the validation subgroup study as shown in (2.1.3)

In the following, as highlighted by Rosner et al,(1990) the derivative of $\sum_{V_{k_1,k_2}}$ can be obtained. From multivariate delta method (Rao, 1973), it follows that

$$cov(V_{h_1,k_1}, V_{h_2,k_2}) \cong \sum_{r=1}^{p}\sum_{s=1}^{p}\sum_{t=1}^{p}\sum_{u=1}^{p}\left(\frac{\partial V_{h_1,k_1}}{\partial \hat{\alpha}_{rs}}\right)\left(\frac{\partial V_{h_2,k_2}}{\partial \hat{\alpha}_{tu}}\right)cov\Big(\hat{\alpha}_{rs}, \hat{\alpha}_{tu}\Big) \qquad (2.1.6)$$

where V_{h_r,k_s} is the $(h_r, k_s)th$ entries of matrix V from (2.1.5)

From Searle (1982), for any parameter say w, the $p \times p$ matrices $\frac{\partial V}{\partial w}$, $\frac{\partial \hat{\alpha}}{\partial w}$ are related by

$$\frac{\partial V}{\partial w} = -V\frac{\partial \hat{\alpha}}{\partial w}V \qquad (2.1.7)$$

Using (2.1.7), (2.1.6) can be rewritten as

$$cov(V_{h_1,k_1}, V_{h_2,k_2}) \cong \sum_{r=1}^{p}\sum_{s=1}^{p}\sum_{t=1}^{p}\sum_{u=1}^{p} V_{h_1r}V_{k_1s}V_{h_2t}V_{k_2u}cov\Big(\hat{\alpha}_{rs}, \hat{\alpha}_{tu}\Big) \qquad (2.1.8)$$

Substituting (2.1.8) in (2.1.5) yields the result obtained from solution of (2.1.6).

To obtain (2.1.5), Rosner et al. (1990) assumed that $\hat{\theta}^*$ and $\hat{\alpha}$ are independent (see Rosner et al. 1990 for detailed description), because the internal validation subgroup and observed study are mutually exclusive subset from the same data, or independent study data if in the case of an external validation and observe study.

Regression calibration is simple both in concept and in practical implementation. One of the advantages attributed to the method is that, when obtaining standard errors and confidence limits for the corrected regression parameters, error in the validation subgroup estimates is accounted for (Rosner et al., 1990). However, there are some limitations associated with this method. Some of which are as follows;

1. Regression calibration is closely consistent only in the case of non-linear regression models. For example, Carroll et al. (1995*b*) shows that in logistics regression, regression calibration is inconsistent.

2. The method is valid only when nondifferential misclassification pattern is assumed. A nondifferential misclassification under this case is when the distribution of X^* conditional on X equals the distribution of X^* conditional on X and Y.

3. This method can only be used when validation subgroup study data is available.

4. Regression calibration is not appropriate for estimating properties of a regression model (to include residual variance, classification error rate) except for regression coefficients (see Freedman et al., 2004).

2.1.2 Moment Reconstruction

Moment reconstruction is another method for correcting for measurement error especially when the covariates are error-prone. It is simple both in concept and in the use of standard software in implementation. Another advantage of moment reconstruction is that the method is valid when considering certain or specific types of differential misclassification patterns. A differential misclassification pattern occurs when the probability of measurement error of a variable depends on any other variables. Most importantly, the method provides consistent estimates even when the residual variances differs by outcome y_i, and it is appropriate for estimating other properties of regression models, to include residual variance. The idea of moment reconstruction involves substituting values for exposure variable X similar to that of regression calibration. In moment reconstruction, the first and second moment of the substituted values are the same as the first and second moment of exposure variable X. Also, Freedman et al. (2004) proved that under linear regression analysis, results obtained from moment reconstruction are the same as those obtained from the regression calibration method (see Freedman et al., (2004) for the detail of the test). However, one advantage of moment reconstruction over regression calibration is that, estimates obtained from moment reconstruction procedures are consistent with logistic regression model fitted to case control study data, compared to regression calibration (see full detail in Freedman et al., (2004)).

Suppose $(n \times 1)$ vector of response variable Y is associated with a $(n \times p)$ matrix of true covariate X such that, for some unknown parameter of interest θ

$$E(Y|X, \theta) = f(X, \theta)$$

In real application, an error-prone X^* is observed, but the idea of the moment reconstruction method is to substitute values in place of the error-prone covariate x_i^*. Let

$$X_{mm}(X^*, Y) = X_{mm}$$

be the substituted values. To obtain consistent parameter estimates, it is important that;

1. The substituted values have the same distribution as that of the true covariates x,
2. The joint distribution of substituted value with outcome y is the same as the joint distribution of exposure variable X with outcome Y

The difficulty of getting the correct substitution exists, hence, Freedman et al. (2004) proposed to match the first and second moments of the joint distribution of substituted value X_{mm} with observed Y. Assuming measurement error in X is such that

$$E(X^*|Y) = E(X|Y)$$

It follows that the first moment is

$$\begin{aligned} E(X_{mm}|Y) &= E(X^*|Y)(I-D) + E(X^*|Y)D \\ &= E(X^*|Y) = E(X|Y) \end{aligned} \tag{2.1.9}$$

where

$$D = D(Y) = (cov(X^*|Y)^{\frac{1}{2}})^{-1}cov(X|Y)^{\frac{1}{2}} \tag{2.1.10}$$

And from (2.1.9), the second moment is

$$Cov(X_{mm}|Y) = D^T cov(X^*|Y)D$$

Using (2.1.10),

$$\begin{aligned} Cov(X_{mm}|Y) &= \big(\big(cov(X^*|Y)^{\frac{1}{2}}\big)^{-1}\big)^T \big(cov(X|Y)^{\frac{1}{2}}\big)^T \times cov(X^*|Y) \times \big(cov(X^*|Y)^{\frac{1}{2}}\big)^{-1} cov(X|Y)^{\frac{1}{2}} \\ &= \big(cov(X|Y)^{\frac{1}{2}}\big)^T cov(X|Y)^{\frac{1}{2}} \\ &= cov(X|Y). \end{aligned} \tag{2.1.11}$$

It is obvious that from (2.1.11) and the assumptions stated above, the unconditional second moment of substituted value X_{mm} is the same as unobserved X. Likewise, it follows that the covariance of joint distribution of X_{mm} and response variable Y is the same as covariance of joint distribution of X an Y. As a result, if distributions of unobserved (X, Y) and that of observed (X^*, Y) are multivariate normal, it implies that both substituted values (X_{mm}, Y) and (X, Y) have multivariate normal distributions, based on the first and second moment of the defined distribution.

Freedman et al. (2004) highlighted that substituting estimate for $E(X^*|Y)$ and $D = D(Y)$ in (2.1.9) would produce estimated $\hat{X}_{mm}$. Thus, estimating $E(X^*|Y)$ and $D = D(Y)$ depends on

1. Regression model for (Y, X) data
2. Validation sub-study containing information about the measurement error in X^*
3. Measurement error model

Freedman et al. (2004) suggested that when $\hat{X}_{mm}$ is consistent, the distribution of $(\hat{X}_{mm}, Y)$ will keep the asymptotic properties of the first and second moment of the distribution of (X, Y). As a result, if the normality assumption of $(X^*|Y)$ and $(X|Y)$ hold, intuitively similar function of $(\hat{X}_{mm}, Y)$ and (X, Y) will have same asymptotic properties (see detailed example in Section 6 of Freedman et al., 2004).

Next is the illustration on how to obtain consistent estimates of θ. Here, for step 1 we are interested in the logistic regression model of a binary response variable Y and covariate X (X is not a matrix in this illustration). Freedman et al, (2004) assume that covariate

X has a normal distribution with mean μ_0 and covariance matrix $\sum_{xx}$ among subjects in the control group (that is, $X = \mathbf{N}(\mu_0, \sum_{xx}\theta)$ for $y_i = 0$). And among subjects in the case group ($y_i = 1$), covariate $X = \mathbf{N}(\mu_0 + \sum_{xx}\theta, \sum_{xx})$; has a normal distribution with mean $\mu_0 + \sum_{xx}\theta$ and covariance matrix $\sum_{xx}$ (see detail in Carroll et al., 1995*b*). Therefore, $(X|Y = y) = \mathbf{N}(\mu_0 + y\sum_{xx}\theta, \sum_{xx})$.

In terms of the validation substudy carried out, Freedman et al. (2004) highlighted that

- In a case control study design, participants are randomly sampled separately from the populations of cases ($Y_i = 1$) and controls ($Y_i = 0$).
- In the cohort design, subjects are randomly sampled from a single population containing both cases and control groups, and response variable Y is observed on these subjects.

In both settings, when unobserved X and observed exposure variable X^* given outcome variable Y are multivariate normal, moment reconstruction produces consistent estimates of θ.

For step 3, Freedman et al. (2004) assume for differential misclassification pattern where the variance of measurement error is assumed to depend on the outcome Y, and classical measurement error process. In the case of classical measurement error model,

$$X^* = X + W \tag{2.1.12}$$

where

- W has mean 0 and variance $\sum_{ww}$
- X has mean μ_x and covariance matrix $\sum_{xx}$

As a result, from (2.1.12), the covariance matrix of X^* is given as

$$\sum_{x^*x^*} = \sum_{xx} + \sum_{ww}$$

In the case of a differential misclassification pattern, $(W|Y = y)$ follows a multivariate normal distribution with mean vector of zeros and covariance $\sum_{ww,y}$, (that is, $\mathbf{N}(0, \sum_{ww,y})$. Hence, the classical error model, $(X^*|Y = y)$ follows $\mathbf{N}(\mu_0 + \sum_{xx}\theta y, \sum_{xx} + \sum_{ww,y})$; multivariate normal with mean $\mu_0 + \sum_{xx}\theta y$ and $\sum_{xx} + \sum_{ww,y}$.

Let

$$\hat{\sum}_{xx} \text{ be a consistent estimate of variance of } X \text{ conditional on } Y$$

Under the case control study design, estimate of $\hat{\sum}_{xx}$ can be obtained by

1. calculating sampling covariance matrix of x_i^* within the case and control group
2. Subtract $\hat{\sum}_{ww,y}$ from the covariance matrix obtained from Step 1, where $\hat{\sum}_{ww,y}$ is obtained from repeated measurements of X^* observed from independent or separate substudy
3. Based on the number of cases and controls, compute a weighted average of the resulting estimates.

Similarly, in the case of cohort setting, let $\hat{E}(X^*|Y)$ be the mean of X^* within cases and controls group, and $\hat{D}_y$ is defined as,

$$\hat{D}_y = \left(\hat{\sum}_{xx} + \sum_{ww,y} \right)^{-\frac{1}{2}} \hat{\sum}_{xx}^{\frac{1}{2}}$$

Hence,

$$\hat{X}_{mm} = \hat{E}(X^*|Y)(I - \hat{D}_y) + \hat{D}_y$$

since the joint distribution of (X_{mm}, Y) is the same as the joint distribution of (X, Y), and $\hat{X}_{mm} = \hat{X}_{mm}(Y, X^*)$ is consistent estimate of $X_{mm} = X_{mm}(Y, X^*)$. Therefore, the parameter estimate resulting from the moment reconstruction method is shown to be consistent (see Section 4.2 of Freedman et al., 2004 for a simulation study comparing the performance of moment reconstruction to regression calibration case).

2.1.3 Multiple Imputation with Misclassified Covariates

In this section, we illustrate the use of multiple imputation to account for measurement error in exposure variable. Treating bias as a missing data issue leads to valid and easy methods for correcting misclassification (Carroll et al., 1995). The missingness mechanism adopted mostly are Missing At Random (MAR) and Missing Completely At Random (MCAR), but there are three main missing data patterns or assumptions identified in the literature. These are briefly described as follows;

- Missing Completely At Random (MCAR): This implies the probability of a data point being missing is independent of the observed and/or unobserved data values of the variable(s) considered. Missing Completely At Random pattern is a special case of Missing At Random assumptions (Little and Rubin, 2002)
- Missing At Random (MAR): Here, missingness of data value is independent of the unobserved data, but conditionally dependent on observed values. For clarity, suppose a subset of values is missing from measurements (say variable T_i) obtained from a participant in a study. The probability of values missing depends on the partially observed data points, but is unrelated to the unobserved values.
- Missing Not At Random (MNAR): In this case, the missing pattern is neither missing

at random (MAR) nor missing completely at random (MCAR). This is a bit complex because the probability of a value(s) missing is dependent on the unobserved observation(s).

Although the pattern of missingness can rarely be identified, the observed data can be used to frame consistent assumptions about the pattern of the missing data. Rubin (1987) suggested exploring the range of plausible missingness patterns, this provides essential information on the different methods used for parameter estimate(s) which has an impact on the conclusion drawn. Moreover it can be used to examine the robustness of model for inference purposes.

Given that, from validation subgroup study, observed measures can be mapped to true values, Cole et al. (2006) proposed multiple imputation, based on internal validation study data, as a method for correcting measurement error in covariate(s). If validation subsample participants are internal, then the missingness technique on the unknown observation on covariate (say X) is missing completely at random (MCAR). However, if the validation substudy participants are stratified on any other exposure variable (say Z), then the missingness pattern will be missing at random (MAR) (Rubin, 1976; Little and Rubin, 2002). Cole et al. (2006) assumed the MAR pattern since subjects from the internal validation study were random sample from the main/observed study group. Irrespective of the missingness assumption made, the conditional distribution of $X|Y, X^*, Z$ will be similar for both validation, for whom exposure X is observed for each participant, and main study subjects, for whom exposure X is missing (Rubin, 1976; Little and Rubin, 2002).

Suppose a logistic model is considered for a binary response variable Y which is related to a vector of covariates $\mathbf{X}$ and other covariate variables Z such that $Y \sim Bin(1, \pi)$ and π is

$$\pi = P(Y = 1|x, z) = \frac{\exp(\alpha + \theta_1 x + \theta_2 z)}{1 + \exp(\alpha + \theta_1 x + \theta_2 z)}$$

where α is nuisance parameter and θ are the parameters of interest.

However, in real application, error prone exposure variable X^* is observed. Thus, the main/observed study data contain response variable Y and X^*. The error prone X^* is related to X by a multivariate normal linear regression model commonly referred to as measurement error (Karen and Loki, 2008). The measurement error model is defined as

$$X^* = X\delta_x + Z\delta_z + \epsilon_{x^*}$$

where δ are regression coefficients, $\epsilon_{x^*} \sim \mathbf{N}(0, \sigma^2_{x^*} I)$ and is said to be independent of any other variables except X^*.

Since true measure X is unobserved in the main study, in literature, observed X^* is mapped to true exposure variable from validation subgroup, so as to estimate the conditional distribution of X and X^* and then use X^* to predict unobserved exposure variable X. Here, Karen and Loki (2008) assume vector of exposure variable **X** is $X \sim MVN(0, \sum_x)$; multivariate normal with mean vector zero and covariance matrix $\sum_x$. Thus, by using standard normal theory (see standard normal theory in Anderson, 2003), it follows that the conditional distribution of X given X^* is multivariate normal. Hence, the exposure variable X can be expressed as

$$X = X^* \omega_{x^*} + Z\omega_z + \epsilon_x$$

where $\epsilon_x \sim N(0, \sum_{x|x^*})$ is dependent on X, but independent of other variables and ω are parameters. These coefficients of measurement error (δ, $\sum_x$, $\sigma^2_{x^*}$ and ω $\sum_{x|x^*}$) can be found from either the internal or external validation subgroup study.

Karen and Loki (2008) aim to compare maximum likelihood with multiple imputation and regression calibrator estimator in accounting for measurement error in exposure variables. Hence, they employed the frequentist approach of the multiple imputation procedure (where at each iteration, there are two steps; the imputation and parameter estimation step) because it eases comparison between the estimators (see the detail of the comparison and example in Karen and Loki, 2008). They used Gibbs sampling for the parameter and imputation stages of the algorithm in developing the proposed method. Gibbs sampling is one of the Markov Chain Monte Carlo (MCMC) techniques for generating samples from posterior distribution (see Rubin, 1987; Schafer, 1997; Wang and Robin, 1998 for full description of Gibbs sampling). Karen and Loki (2008) focused on the use of multiple imputation in accounting for measurement error in exposure variables by combining internal substudy and main/observed study. In this case, the implementation of multiple imputation is demonstrated as follows;

- Based on the main study data (Y_i, X_i), at iteration $(r-1)$, let

  $$\left(\theta, \omega. \sum_{x|x^*}\right)^{(r-1)}, \quad and \quad X(Y_i, X_i^*)^{(r-1)}$$

  be the current parameters and imputed data for an *ith* participant.

- Next is iteration r: First, Markov chain algorithm calculate $\left(\theta, \omega. \sum_{x|x^*}\right)^{(r)}$; estimate parameter from combination of main and internal study likelihood function defined as

$$\sum_{i \in observed} l_1(\theta; Y_i, \tilde{X}(Y_i, X_i^*)^{(r-1)}) + l_2(\omega, \sum_{x|x^*}; \tilde{X}(Y_i, X_i^*)^{(r-1)}, X_i^*) \qquad (2.1.13)$$
$$+ \sum_{i \in internalval.} l_1(\theta; Y_i, X_i) + l_2(\omega, \sum_{x|x^*}; X_i, X_i^*)$$

where l_1 and l_2 are from logistic and normal regression models respectively. It follows that an updated parameter $(\theta, \omega, \sum_{x|x^*})^r$ is drawn from the normal distribution with mean $(\hat{\theta}, \hat{\omega}, \hat{\sum}_{x|x^*})^r$ and variance of (2.1.13) estimated at $(\hat{\theta}, \hat{\omega}, \hat{\sum}_{x|x^*})^r$ by using standard software.

The associated imputed data for iteration r: From the conditional distribution of unobserved X given the observed Y_i and X_i^*, Markov chain algorithmn randomly imputes $X(Y_i, X_i^*)$ for participants in the observed study at evaluated $(\theta, \omega, \sum_{x|x^*})^r$. The conditional distribution of unobserved X given the observed Y_i and X_i^* is proportional to

$$f(x) = \frac{\exp{(x\theta^r Y_i)}}{1 + \exp{(x\theta^r)}} \exp\left(\frac{-1}{2}(x - X_i^* \omega^r)' \frac{1}{\sum_{x|x^*}^r}(x - X_i^* \omega^r) \right)$$

- After an adequate number of iterations, D imputed completed data are used for estimation purposes. For individual complete data, maximum likelihood estimate $\tilde{\theta}^i$ is obtained from (2.1.13).
- The multiple imputation estimate $\hat{\tilde{\theta}}$ is obtained by averaging the individual estimate $\tilde{\theta}^i$ from each of the D complete dataset. Therefore, $\hat{\tilde{\theta}}$ is,

$$\tilde{\hat{\theta}} = \frac{1}{D} \sum_{i=1}^{D} \tilde{\theta}^i$$

Rather than estimating variance from the imputation process described above, Karen and Loki (2008) highlighted the use of the bootstrap estimate of variance because a multiple imputation based variance estimator may be sensitive to the approximation of distribution. Hence, they did not demonstrate how to obtain a multiple imputation estimate of variance.

One of the limitations of this approach is the difficulty of having a ready to-use software that samples from the conditional distribution of the unobserved exposure variable X given the observed Y_i and X_i^* evaluated at $(\hat{\theta}, \hat{\omega}, \hat{\sum}_{x|x^*})^r$ (Karen and Loki, 2008). Hence, implementing the multiple imputation approach using observed/internal subgroup study data is computationally challenging in real studies.

2.1.4 Maximum Likelihood with Nondifferential Misclassified Covariates

In this section, we give an illustration of an observed study data said to have followed a nondifferential misclassification pattern. When information about measurement error or misclassification parameters of a variable is independent of any other variables, then the observed study data show nondifferential measurement error pattern. Prior to Lyles et al. (2007) work, methods that account for measurement error suggested the use of a closed form weighted average estimator of log odds ratio (see full description in Greenland, 1988; Spiegelman et al., 2001; and Thurston, 2003). These methods combined information from both external/internal validation substudies in adjusting for nondifferential pattern of measurement error when relating continous or binary response variables with continuos mismeasured exposure variables. Although external data is cost effective (Lyles et al. 2007), a fundamental issue with any external validation subgroup is its lack of transportability, generally believed to be inefficient, thereby resulting to potential bias in estimating parameters. Lack of transportability occurs when external validation studies are obtained under conditions or a design different from the main study. Therefore, when using the observed/main and external substudy, the design of the external validation data should be looked into so as to ensure valid transportability.

Lyles et al. (2007) correct for nondifferential measurement error patterns using the main study and a combination of internal/external validation design. Their focus is on binary response Y and binary covariate X. In this case the covariate is univariate. They associate the observed exposure variable X_i^* (assumed to be measured with error) in place of the true measure covariate X_i to the response variable Y_i. The observed covariate variable is said to be binary with $x_i^* = 1$; indicating the presence of exposure, and $x_i^* = 0$ otherwise. The observed study design considered in this section has a case control setting taken from Lyles et al., (2007). Generally, the probabilities of no measurement error associated with observed study data are commonly referred to as sensitivity and specificity. In this chapter, in the case of measurement error in covariates, sensitivity is denoted as $(1-\zeta_X) = P(X_i^* = 1|X_i = 1)$; the probability of observing the presence of mismeasured exposure $X_i^* = 1$ given the presence of the true measure exposure $X_i = 1$, and specificity is denoted as $(1 - \gamma_X) = P(X_i^* = 0|X_i = 0)$; the probability of observing the absence of error-prone exposure $X_i^* = 0$ given the absence of true measure exposure. The subscript X in ζ_X and γ_X indicates that probabilities of measurement error of exposure variable X are independent of other variables' information.

In the following, we illustrate the method of combining main and internal and external validation studies in a case control design to obtain the correct estimate of a parameter when covariate variable X is misclasified. We begin with the data structure. Figure 2.2 shows the data structure of the main, external and internal substudies respectively. From Figure 2.2, the main study data layout is the first figure containing (Y_i, X^*) where Y_i is a binary outcome assumed to be error-free and X_i^* is error-prone binary covariates with $X_i^* = 1$ representing the presence of exposure, and $X_i^* = 0$ otherwise. It can be seen that there are four groups that make up the observe study. The external validation substudy is the middle figure containing (X_i, X_i^*) with four groups as well, where X_i is the true covariate. The internal validation subgroup has eight groups containing (Y_i, X_i, X_i^*) as shown in the third figure.

Lyles et al. (2007) highlighted that the estimation of $(1 - \zeta_X)$ and $(1 - \gamma_X)$ can be obtained from internal validation subgroup data, which can be used as a substitute for true covariate X_i. Note that this is carried out separately for cases $X_i = 1$ and controls $X_i = 0$ groups. When the estimation is carried out separately, it allows for correction of differential measurement error pattern of covariates variables (Thomas et al., 1993). When the misclassification rates of a variable depend on other variable(s), it is said to be a differential measurement error process.

Lyles et al. (2007) focused attention on a nondifferential misclassification process of exposure variable X and assumed both probabilities of measurement error from external subgroups to be the same for internal validation substudies. Lyles et al. (2007) used the combination of internal and external validation studies to estimate sensitivity and specificity.

Next, in Lyles et al. (2007) a standard model that relates unobserved exposure X to response variable Y is

$$P(X|Y) = \pi_X$$

However, in real studies, observed data contain potential measurement errors. Hence, they considered the following analysis;

1. Naive analysis: A naive model often known for ignoring measurement error is fitted

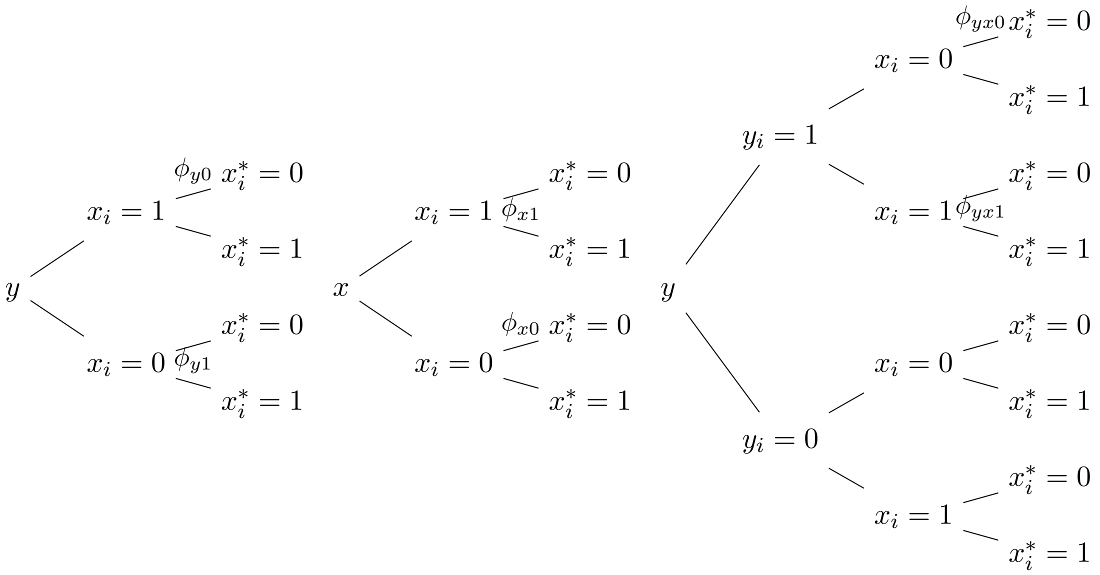


Figure 2.2: Data Structure: Main, External and Internal Validation Study Data for a case control setting with Misclassified Covariate variable with corresponding probabilities, where ϕ_{y0} is the probability when $(Y_i, X_i^* = 0)$ for *ith* participant from main study, ϕ_{x0} is the probability when $(X_i, X_i^* = 0)$ for *ith* participant from external substudy, and ϕ_{yx0} is the probability when $(Y_i, X_i, X_i^* = 0)$ for *ith* participant from internal study

to the observed study data such that

$$P(X^*|X) = \pi^*_X \tag{2.1.14}$$

Estimate of parameter from (2.1.14) could be bias.

2. Combination of observe and internal substudies:

   (a) First, they explore using the combination of observed study data and the pair of (Y, X^*) only from internal validation subgroup, but the estimate is said to be biased (Lyles et al., 2007)

   (b) Furthermore, they combined the pair of true measures (Y, X) from internal validation substudy and observed study. The estimate of parameters from the analysis is valid, but inefficient (Lyles et al., 2007)

3. Combination of observe and internal and external substudies: Since the result of the analysis carried out in step (2) is inefficient, they combined observed and internal and external validation studies so as to obtain efficient estimates of parameters. This was achieved by using the joint log-likelihood function of the combined study groups. The overall log-likelihood function of the full data is defined as

$$\ell = \sum_{y=0}^{1}\sum_{x^*=0}^{1} n_{yx^*}\log(\phi_{yx^*}) + \sum_{y=0}^{1}\sum_{x^*=0}^{1}\sum_{x=0}^{1} n_{yx^*x}\log(\phi_{yx^*x}) + \sum_{x^*=0}^{1}\sum_{x=0}^{1} n_{x^*x}\log(\phi_{x^*x}) \tag{2.1.15}$$

   From (2.1.15), the first term is the log-likelihood of the observed study, followed by the log-likelihood of the internal validation subgroup, and the last term is log-likelihood of external substudy, where

   - n_{yx^*}: Number of observations in each group from the observed study
   - (ϕ_{yx^*}): The corresponding probabilities with regards to n_{yx^*} from the observed study such that,

     When $X^*_i = 1$; Case group

     $$\begin{aligned}\phi_{y1^*} =& P(X^*_i = 1|x_i = 1)P(X_i = 1|y) + P(X^*_i = 1|x_i = 0)(1 - P(x_i = 1|y))\\ &= (1 - \zeta_X)\pi_{Xi} + \gamma_X(1 - \pi_{Xi})\end{aligned}$$

     And for the control group $X^*_i = 0$

     $$\begin{aligned}\phi_{y0^*} =& P(X^*_i = 0|x_i = 1)P(X_i = 1|y) + P(X^*_i = 0|x_i = 0)(1 - P(x_i = 1|y))\\ &= \zeta_X\pi_{Xi} + (1 - \gamma_X)(1 - \pi_{Xi})\end{aligned}$$

   - n_{yx^*x}: Number of observations in each group from internal subgroup

- (ϕ_{yx^*x}): The corresponding probabilities with regards to n_{yx^*x} from the internal substudy such that

$$\begin{aligned}\phi_{y1^*1} &= (1-\zeta_X)P(X_i=1|y) = (1-\zeta_X)\pi_{Xi}\\ \phi_{y1^*0} &= \gamma_X(1-P(X_i=1|y)) = \gamma_X(1-\pi_{Xi})\\ \phi_{y0^*1} &= \zeta_X P(X_i=1|y) = \zeta_X\pi_{Xi}\\ \phi_{y0^*0} &= (1-\gamma_X)P(X_i=1|y) = (1-\gamma_X)(1-\pi_{Xi})\end{aligned}$$

- n_{x^*x}: number of observations in each group in external study
- (ϕ_{x^*x}): The corresponding probabilities in relation to n_{x^*x} in the external validation substudy. Each group in the external subsample has

$$\begin{aligned}\phi_{1^*1} &= (1-\zeta_{X\,ext})\pi_x,\\ \phi_{1^*0} &= \gamma_{X\,ext}(1-\pi_x),\\ \phi_{0^*1} &= \zeta_{X\,ext}\pi_x,\\ \phi_{0^*0} &= (1-\gamma_{X\,ext})(1-\pi_x).\end{aligned}$$

where π_x is classified as nuisance parameter in the external subgroup, and the subscript *ext* indicates the probabilities of measurement error are from the external subsample. In the case of transportability, misclassificate rates from external substudies equal those from observed and internal validation such that $\zeta_X = \zeta_{X\,ext}$ and $\gamma_X = \gamma_{X\,ext}$ (Lyles et al., 2007).

2.1.5 Maximum Likelihood with Differential Misclassified Covariate

When the probabilities of measurement error of a variable depend on other variable(s), it is said to be differentially misclassified. An example can be seen in epidemiology research where exposure variables are often prone to error. Here, we illustrate the maximum likelihood approach of correcting the differential misclassification process for mismeasured exposure. Morrissey and Spiegelman (1999) investigated the performance of matrix, inverse matrix and maximum likelihood estimator under a differential misclassification pattern of covariate measured with error. They used observed and internal validation substudies to correct for measurement error in covariate variables, and suggested that the maximum likelihood approach is optimal and often to be recommended when deciding which estimator to apply when there is differential misclassification. However, there is a limitation to the use of the maximum likelihood approach, namely its software implementation. This is because the maximum likelihood approach under differential misclassification requires iterative method which is inaccessible in standard software, then Lyles, 2002 modified Morrissey and Spiegelman, 1999 work. Lyles (2002) suggested that in the case of differential measurement error process, the inverse method is similar to the maximum likelihood approach when an observed study is combined with interval validation subsample in a case-control study. In this way, the limitation may be avoided, and the maximum likelihood approach can be carried out without requiring sophisticated software.

In the following, we highlight the illustration in Lyles (2002) on how to correct for measurement error in exposure variable X when the probabilities of measurement error are dependent on response variable Y, using a combination of observe/main and internal validation study.

Suppose a case control study is observed with n being the number of participants and covariate X_i^* is mismeasured and the response variable Y_i is assumed to be error-free. Lyles (2002) assumes sampling is conditional on outcome variable Y_i under the case of differential measurement error process. If there are no measurement errors,

$$P(X_i = 1|y) = \pi_{Xi} = \frac{\exp{(\alpha + \theta y_i)}}{1 + \exp{(\alpha + \theta y_i)}},$$

where θ is the unknown parameter of interest, and α is nuisance parameter.

However, when a naive analysis is performed on the observed study only, $P(X_i^* = 1|y)$ is

$$P(X_i^* = 1|y) = \pi^*_{Xi} = \frac{\exp{(\alpha^* + \theta^* y_i)}}{1 + \exp{(\alpha^* + \theta^* y_i)}}$$

where θ^* is the parameter from the naive model expected to be biased, and α^* is the nuisance parameter.

For differential measurement error in covariates, we defined both classification rates; sensitivity as $(1-\zeta_{X_y}) = P(X_i^* = 1|X_i = 1, Y)$ and specificity as $(1-\gamma_{X_y}) = P(X_i^* = 0|X_i = 0, Y)$. The subscript X_y is to indicate the measurement error of exposure variable. This depends on the response variable Y. When both probabilities of measurement error ζ_{X_y} and γ_{X_y} are introduced in the naive model, $P(X_i^* = 1|y)$ becomes

$$\begin{aligned}\pi^*_{X_i} &= P(X_i^* = 1|y) = P(X_i^* = 1|X_i = 1)P(X_i = 1|y_i) + P(X_i^* = 0|X_i = 0)P(X_i = 0|y_i)\\ &= \zeta_{X_y}\pi_{X_i} + \gamma_{X_y}(1 - \pi_{X_i}) \qquad (2.1.16)\end{aligned}$$

In Lyles (2002), (2.1.16) is the same as the matrix method. However, Lyles (2002) intention is to show the similarity between the inverse matrix method and the maximum likelihood approach. Hence, under a differential measurement error process, the inverse matrix method is developed in terms of classification parameters known as positive predictive values PPV, and negative predictive values NPV (Marshall, 1990; and Morrissey an Spiegelman 1999). Let

$$PPV_y = P(X_i = 1|X_i^* = 1, Y = y) NPV_y = P(X_i = 0|X_i^* = 0, Y = y)$$

It follows that

$$\pi_{Xi} = PPV_y \pi^*_{Xi} + (1 - NPV_y)(1 - \pi^*_{Xi})$$

The joint likelihood of observed and internal studies can be used to obtain an estimate of parameters under the maximum likelihood approach. Hence, the joint log-likelihood is

$$\ell \propto \sum_{i=1}^{n} \log\big(P(X^*_i|y_i)\big) + \sum_{i=1}^{n_v} \log\big(P(X^*_i|x_i, y_i) \times P(X_i|y_i)\big) \tag{2.1.17}$$

In a case control study design,

- An observed/main study data (Y, X^*) has four groups. As shown in the first figure of Figure 2.2, this includes $(1,1),(1,0),(0,1),(0,0)$. The number of participants in each group is denoted as n_{11}, n_{12}, n_{01} and n_{02} respectively.
- The third figure from Figure 2.2 shows that the internal validation data (Y, X, X^*) has eight groups; $(1,1,1),(1,0,1),(1,1,0),(1,0,0),(0,1,1),(0,0,1),(0,1,0),(0,0,0)$. The respective number of participants in each group in the internal subgroup are n_{13}, n_{14}, n_{15}, n_{16}, n_{03}, n_{04}, n_0 and n_{06}

Although the maximum likelihood estimator is optimal, it is computationally intense. Hence, Lyles et al. (2002) provided an alternative form of (2.1.17) in terms of $\pi^*_{X_i}$, PPV_y, and NPV_y given as

$$\begin{aligned}\ell = \sum_{y=0}^{1} \Big(& n_{y1}\log(\pi^*_{X_i}) + n_{y2}\log(1-\pi^*_{X_i}) + n_{y3}\log(PPV_y\pi^*_{X_i}) \\ & + n_{y4}\log((1-PPV_y)\pi^*_{X_i}) + n_{y5}\log\big((1-NPV_y)(1-\pi^*_{X_i})\big) + n_{y6}\log\big(NPV_y(1-\pi^*_{X_i})\big)\Big)\end{aligned} \tag{2.1.18}$$

Lyles et al, (2002) highlighted that (2.1.18) is tractable, and when the algebra is excluded, the maximum likelihood estimates can be obtained where

$$\begin{aligned}\widehat{PPV}_y &= \frac{n_{y3}}{(n_{y3}+n_{y4})}, \\ \widehat{NPV}_y &= \frac{n_{y6}}{(n_{y5}+n_{y6})}, \\ \hat{\pi}^*_{X_i} &= \frac{(n_{y1}+n_{y3}+n_{y4})}{n_d}, \\ \hat{\pi}_{X_i} &= \widehat{PPV}_y\hat{\pi}^*_{X_i} + (1-\widehat{NPV}_y)(1-\hat{\pi}^*_{X_i}),\end{aligned}$$

2.2 Measurement Error in Outcome

In Section 2.1, we considered current methods used for accounting for measurement error in exposure variables. In this section, we consider some of the proposed methods that account for measurement error in outcome variable. In real studies, observed data are error-prone. Instead of obtaining true measures of outcome of interest Y (that is error-free outcome), an alternative Y^* is sometimes observed. That is, in practice we could have an observed data where only the response Y_i^* is misclassified, but the exposure variable X is without error.

Here, we described multiple imputation and maximum likelihood approaches that account for measurement error in response variable. More specifically, the maximum likelihood method illustrated are nondifferential and differential patterns of measurement error.

2.2.1 Multiple Imputation with Misclassified Outcome

Multiple imputation is one of the techniques used in handling missing data (Rubin, 1987; Little and Rubin, 2002). The method is well established for dealing with uncertainty or bias associated with unobserved observations. It has also been used by researchers to develop methods that account for measurement error. Multiple imputation involves three stages namely; imputation, estimation and lastly pooling. The approach repeatedly imputes values over time (say R times) for observation(s) missing from a dataset, which result in R complete datasets. This method is developed purposely to recover (not ignore) missing information or variability of imputed values for analysis, not necessarily to obtain correct values for the unobserved. Each complete data is analysed, results are summarised within and across each complete data over the r number of imputations using rules developed by Rubin, (1987). During implementation, the imputed values must reflect the missingness present in the observed data in order to obtain an accurate estimates.

We begin with the general overview of a multiple imputation procedure for missing data. Suppose the primary objective is to obtain a single parameter estimate $\hat{\theta}$ from partially observed data. Then, values are imputed for missing observation(s) in R iterations yielding R complete datasets. An estimate $\hat{\theta}_r$ is obtained from each of the R dataset. Based on the R iterated imputed datasets, the multiple imputation estimate $\bar{\hat{\theta}}$ is the average of the individual estimates $\hat{\theta}_r$ from each of the R complete dataset. Hence, $\bar{\hat{\theta}}$ is defined as

$$\bar{\hat{\theta}} = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r. \tag{2.2.1}$$

Let $Var(\hat{\theta}_r)$ denote the estimated variance of the individual complete dataset. It is important to consider the mean of the variance from individual complete data. This is commonly referred to as within-imputation variance defined as

$$V_{wn} = \frac{1}{R}\sum_{r=1}^{R}Var(\hat{\theta}_r). \tag{2.2.2}$$

In addition, variance V_{bn} across the individual estimate is calculated. In practice, V_{bn} is commonly referred to as between-imputation variance, defined as

$$V_{bn} = \frac{1}{R-1}\sum_{r=1}^{R}(\hat{\theta}_r - \bar{\hat{\theta}})^2. \tag{2.2.3}$$

Therefore, the variance associated with multiple imputation estimate $\bar{\hat{\theta}}$ in (2.2.1) is denoted as $Var(\bar{\hat{\theta}})$ which can be defined as

$$\begin{aligned} Var(\bar{\hat{\theta}}) &= \frac{1}{R}\sum_{r=1}^{R}Var(\hat{\theta}_r) + \frac{1}{R-1}\sum_{r=1}^{R}(\hat{\theta}_r - \bar{\hat{\theta}})^2(1+\frac{1}{R}) \\ &= V_{wn} + V_{bn}(1+\frac{1}{R}), \end{aligned} \tag{2.2.4}$$

where $(1+\frac{1}{R})$ is the factor that corrects bias.

In the following, we describe the implementation of the multiple imputation method in accounting for measurement error. It follows that mismeasured response is treated as a missing data problem, where the unobserved response variable Y is said to be missing for participants.

In our study, we used multiple imputation to account for misclasification rate in an outcome variable, using our motivating example. It follows that,

- A random subset (that is, internal validation subgroup data) is drawn from a simulated data (y_i, x_i), where true outcome y_i is available for subjects in the validation subgroup study, but missing for those subjects not included in the internal validation study
- A logistic model was fitted(applied) to the validation subgroup
- We assume the fitted parameter and covariance matrix obtained from the validation data follow a multivariate distribution, and we draw coefficients for each number of imputation (R = 40)
- We introduced an indicator variable t_i: ($t_i = 1$ if the subject is included in validation study, 0 otherwise). We retain the true outcome y_i where $t_i = 1$ and imputed true outcome values where $t_i = 0$, using the regression coeffcients drawn from the fitted logistic regression model
- Next, a model is fitted to each imputed response variable and the exposure variable
- To investigate multiple imputation under different scenarios, we performed sensitivity analysis for different percentage of validation subgroup data and for different $\gamma = 0, 0.1, 0.3, 0.5, 0.7, 0.9)$

Edwards et al. (2013) proposed a method based on a multiple imputation procedure to account for measurement error in outcome. Since observed response variable can be mapped to unobserved response from a validation substudy, Edwards et al., (2013) illustrated a multiple imputation method using observed study with an internal validation study group. The proposed method assumed the observed (sometimes referred to as main) study contained misclassified or error-prone outcome Y^*, covariate or treatment variable X and some other covariate(s) Z. In addition, each subject from the validation group is assumed to have the unknown true outcome Y_v, misclassified binary outcome Y_v^*, treatment variable X_v and other covariate(s) Z_v. The subscript **v** being used to indicate that the subject/participant is from the validation subgroup. Hence, a validation subgroup study is assumed to contain a full or complete data, since subjects from the validation study have true response variable measure, whereas these are missing for some subsets of subjects from the observed study.

Furthermore, Edwards et al. (2013) assumed a nondiffrential and differential measurement error pattern in their analysis. When the probability of measurement error is independent of one or more variables, the misclassification is said to be nondifferential, and otherwise for differential pattern. Edwards et al. (2013) examined the relationship between y_v^*, y_v, x_v, and z_v amongst subjects in the validation study. They relied on the information from the validation study to impute values for other subjects in the observed study. The algorithm explored is describe as follows;

- First, logistic regression was used to model the relationship between Y_v and other variables Y_v^*, X_v, Z_v contained in the validation data. The likelihood function for the validation substudy denoted as L_v is

  $$\mathcal{L}_v = \prod_{v=1}^{n_v} \pi_v^{y_v}(1-\pi_v)^{1-y_v},$$

  where n_v is the number of participants from the validation subgroup, and $\pi_v = P(y_v = 1 \mid y_v^*, x_v, z_v)$ is the probability of the outcome of interest in the validation subgroup, defined as

  $$\pi_v = \frac{\exp(\alpha + \beta_1 y_v^* + \beta_2 x_v + \beta_3 y_v^* x_v + \beta_4 z_v)}{[1 + \exp(\alpha + \beta_1 y_v^* + \beta_2 x_v + \beta_3 y_v^* x_v + \beta_4 z_v)]}, \tag{2.2.5}$$

  where β are the parameters from the validation subgroup, α is the nuisance parameter.

- From the posterior predictive distribution of the parameter values, a set of regression coefficients was drawn so as to impute values for the unknown true outcome missing from the main study data. There is an assumption that the parameter values follow multivariate gaussian distribution with mean vector $(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$ and covariance matrix $\hat{\Sigma}$ obtained from the model (2.2.5).

- For imputation purposes, a new variable relative to true outcome, say y_r, is created. It follows that y_r values for subjects in the main study are imputed based on the regression coefficients drawn from the above step. Since validation study data contain the true outcome y_v, the supposed imputed outcome y_r for subjects in the validation group is y_v. In otherwords, let y_r denote imputed true outcome, subjects in the main study would have imputed y_r while validation group retains y_v. This step is repeated for each of R iterations.

- For each iteration, y_r is assigned by a random draw with probability π_r defined as

$$\hat{\pi}_r = \frac{\exp(\hat{\alpha}^r + \hat{\beta}_1^r y^* + \hat{\beta}_2^r x + \hat{\beta}_3^r y^* x + \hat{\beta}_4^r z)}{\Big(1 + \exp(\hat{\alpha} + \hat{\beta}_1^r y^* + \hat{\beta}_2^r x + \hat{\beta}_3^r y^* x + \hat{\beta}_4^r z)\Big)}.$$

- The last stage is the analysis of the imputed values given the other variables of interest. Edwards et al. (2013) obtain estimates for each imputation using the standard logistic model that should have been applied originally if the observed study had complete data.

  It follows that from (2.2.1), (2.2.2), (2.2.3) and (2.2.4) (that is, the general procedure for multiple imputation described above), the estimate for R imputed values can be derived. From the example provided by Edwards et al. (2013), they used logistic regression to estimate odd ratio for the imputed response variable Y_r given exposure variable X and other covariates Z where

$$P(Y_r = 1|x, z) = \frac{\exp\left(\theta_0^r + \theta_1^r x + \theta_2^r z\right)}{1 + \exp\left(\theta_0^r + \theta_1^r x + \theta_2^r z\right)}.$$

  For all the R imputation, the average of the estimated odd ratio of interest θ_1 is defined as

$$\exp\left(\bar{\theta}_1\right) = \exp\left(\frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_1^r\right),$$

  where $\hat{\theta}_1^r$ is the individual natural log of the estimated odd ratio from each of the R complete datasets.

  And the variance of $\bar{\theta}_1$ is given as

$$Var(\bar{\theta}_1) = \frac{1}{R}\sum_{r=1}^{R}\widehat{\mathrm{Var}}(\hat{\theta}_1^r) + \frac{1}{R-1}\sum_{r=1}^{R}(\hat{\theta}_1^r - \bar{\theta}_1)^2(1 + \frac{1}{R}).$$

2.2.2 Maximum Likelihood with Nondifferential Misclassified Outcome

In this section, we demonstrate the maximum likelihood approach of accounting for measurement error in response variable Y_i. More specifically, we considered a non differential pattern where the misclassified rates of Y_i is independent on true measurements of exposure X. Suppose a realization of true unknown binary response variable Y_i assume two possible attributes; with $y_i = 0$ as the absence or failure of an attribute and $y_i = 1$ otherwise, for $i = 1, \cdots, n$ and n being the sample size. The study has two groups; the exposure indicator variable is $X_i = 1$ for the treatment group and $X_i = 0$ for control group. The true distribution for response variable is $Y_i|X_i \sim Bin(1, \pi_i)$, and π_i is

$$\pi_i = P(Y_i = 1|X_i) = \frac{\exp(\alpha + \theta X_i)}{1 + \exp(\alpha + \theta X_i)}$$

where θ is the unknown true parameter of interest, α is the nuisance parameters.

For the observed data, the assumed logistic model is

$$Y_i^*|X_i \sim Bin(1, \pi_i^*), \quad \log\left(\frac{\pi_i^*}{1 - \pi_i^*}\right) = \alpha^* + \theta^* X_i, \tag{2.2.6}$$

where Y_i^* is the observed response variable measured with error, and θ^* is the observed parameter. The corresponding log-likelihood function is

$$\ell(\theta^*) = \sum_{i=1}^{n} y_i^* \log \pi_i^* + (1 - y_i^*) \log(1 - \pi_i^*), \tag{2.2.7}$$

However, the conclusion from observed study based on θ^* might be misleading as the problem of measurement error was not considered. Therefore, we need to investigate how the inference can be impacted by measurement error.

In this section, we denote sensitivity and specificity of response variable associated with an observe study data as

$$(1 - \zeta_Y) = Pr(Y_i^* = 1|Y_i = 1)$$
$$(1 - \gamma_Y) = Pr(Y_i^* = 0|Y_i = 0),$$

where

- The subscript Y indicate that probabilities of measurement error is with regards to response variable Y.
- We denote sensitivity as $(1 - \zeta_Y)$: The probability of observing the presence of an attribute $y_i^* = 1$, given that a true measure response attribute is present $y_i = 1$
- Specificity is denoted as $(1 - \gamma_Y)$: The probability of observing the absence of an attribute $y_i^* = 0$, given that a true attribute is absent $y_i = 0$

By assuming a nondifferential independent pattern for a main/observe study wherein only the response variable is misclassified,

$$\begin{aligned} P(Y_i^* = 1|X_i) &= P(Y_i^* = 1|Y_i = 1)P(Y_i = 1|X_i) + P(Y_i^* = 1|Y_i = 0)P(Y_i = 0|X_i) \\ &= (1 - \zeta_Y)\pi_i + \gamma_Y(1 - \pi_i) \end{aligned}$$

The distribution of the observed outcome variable conditional on ζ_Y and γ_Y is

$$Y_i^*|X_i \sim Bin\Big(1, (1 - \zeta_Y)\pi_i + \gamma_Y(1 - \pi_i)\Big) \tag{2.2.8}$$

The corresponding log-likelihood for a given ζ_Y and γ_Y is

$$\ell(\theta \mid \zeta_Y, \gamma_Y) = \sum_{i=1}^{n} y_i^* \log\big((1 - \zeta_Y)\pi_i + \gamma_Y(1 - \pi_i)\big) + (1 - y_i^*)\log\big(\zeta_Y\pi_i + (1 - \gamma_Y)(1 - \pi_i)\big) \tag{2.2.9}$$

Here, the probabilities of measurement error are ζ_Y and γ_Y, assume to be independent of X_i and therefore remain constant. Thus, to consider the measurement error problem, both probabilities of measurement error are introduced into (2.2.7) resulting to model (2.2.9). Hence, from (2.2.9), we can obtain estimate of parameter of interest.

Note that it is practically impossible to calculate estimate of parameter in (2.2.9). However, both γ_Y and ζ_Y can be obtain either from

- Validation substudy: If there is validation data, information from validation subgroup study can be use to map observed measures to true value for parameter estimation (Lyles et al., 2011)
- Pre-specified values: By assuming correct values for ζ_Y and γ_Y would result to unbiased estimate of the parameter of interest (Magder and Hughes, 1997)

2.2.3 Maximum Likelihood with Independent Differential Misclassified Outcome

In this section, we consider a case of independent differential misclassification. The assumption here is that the probability of measurement error is dependent on error-free measurement of binary exposure variable x_i. In this case, both sensitivity $(1 - \zeta_{Y_{X_i}})$ and specificity $(1 - \gamma_{Y_{X_i}})$ are dependent on true covariate X_i. The subscript Y_X indicates the measurement error of response Y is dependent on X. Each combination of $\zeta_{Y_{X_i}}$ and $\gamma_{Y_{X_i}}$ result to varying values for both probabilities of misclassification. The mechanism of independent differential misclassification of response variable Y is shown in Figure 2.3.

Suppose the probabilities of measurement error are

$$P(Y_i^* = 0|Y_i = 1, x_i) = \zeta_{Y_{X0}} - \zeta_{Y_{X0}}x_i + \zeta_{Y_{X1}}x_i \tag{2.2.10}$$

$$P(Y_i^* = 1|Y_i = 0, x_i) = \gamma_{Y_{X0}} - \gamma_{Y_{X0}}x_i + \gamma_{Y_{X1}}x_i \tag{2.2.11}$$

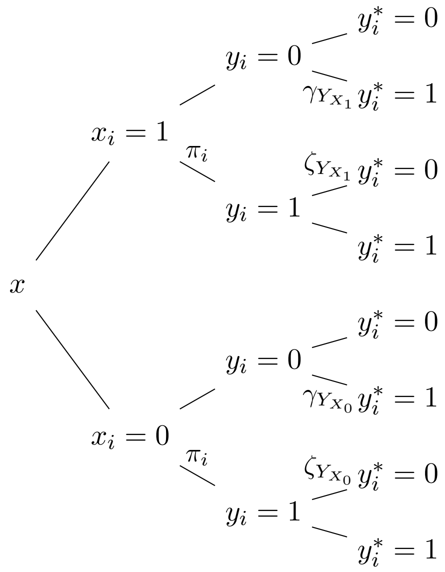

$y_i^* = 0$
$y_i = 0$
$\gamma_{Y_{X_1}}$ $y_i^* = 1$
$x_i = 1$
π_i
$\zeta_{Y_{X_1}}$ $y_i^* = 0$
$y_i = 1$
$y_i^* = 1$
x
$y_i^* = 0$
$y_i = 0$
$\gamma_{Y_{X_0}}$ $y_i^* = 1$
$x_i = 0$
π_i
$\zeta_{Y_{X_0}}$ $y_i^* = 0$
$y_i = 1$
$y_i^* = 1$

Figure 2.3: Mechanism of Independent Differential Misclassification of observed outcome

such that

$$P(Y_i^* = 1|Y_i = 0, x_i) = \begin{cases} \gamma_{Y_{X_0}}, & \text{when } x_i = 0 \\ \gamma_{Y_{X_1}}, & \text{when } x_i = 1 \end{cases}$$

And,

$$P(Y_i^* = 0|Y_i = 1, x_i) = \begin{cases} \zeta_{Y_{X_0}}, & \text{when } x_i = 0 \\ \zeta_{Y_{X_1}}, & \text{when } x_i = 1 \end{cases}$$

By assuming (2.2.10) and (2.2.11)

$$\begin{aligned} P(Y_i^*|x_i) &= P(Y_i^* = 1|Y_i = 1, x_i)P(Y_i = 1|x_i) + P(Y_i^* = 1|Y_i = 0, x_i)P(Y_i = 0|x_i) \\ &= (1 - \zeta_{Y_{X0}} + \zeta_{Y_{X0}}x_i - \zeta_{Y_{X1}}x_i)\pi_i + (\gamma_{Y_{X0}} - \gamma_{Y_{X0}}x_i + \gamma_{Y_{X1}}x_i)(1 - \pi_i) \end{aligned}$$

Hence, the distribution of the observed outcome variable conditional on $\zeta_{Y_{X0}}$ and $\gamma_{Y_{X0}}$ is

$$Y_i^*|X_i \sim Bin\bigg(1, (1 - \zeta_{Y_{X0}} + \zeta_{Y_{X0}}x_i - \zeta_{Y_{X1}}x_i)\pi_i + (\gamma_{Y_{X0}} - \gamma_{Y_{X0}}x_i + \gamma_{Y_{X1}}x_i)(1 - \pi_i)\bigg)$$

To obtain an unbiased estimate of parameters under the case of independent differential measurement error pattern, both misclassification rate $\zeta_{Y_{Xi}}$ and $\gamma_{Y_{X_i}}$ are introduced in

(2.2.7). Therefore, the log-likelihood is

$$\begin{aligned}\ell(\theta|\zeta_{Y_X},\gamma_{Y_X}) = \sum_{i=1}^{n} y_i^* \log\Big(& (1-\zeta_{Y_{X0}}+\zeta_{Y_{X0}}x_i-\zeta_{Y_{X1}}x_i)\pi_i \\ & + (\gamma_{Y_{X0}}-\gamma_{Y_{X0}}x_i+\gamma_{Y_{X1}}x_i)(1-\pi_i)\Big) \\ & + (1-y_i^*)\log\Big((\zeta_{Y_{X0}}-\zeta_{Y_{X0}}x_i+\zeta_{Y_{X1}}x_i)\pi_i \\ & + (1-\gamma_{Y_{X0}}+\gamma_{Y_{X0}}x_i-\gamma_{Y_{X1}}x_i)(1-\pi_i)\Big) \qquad (2.2.12)\end{aligned}$$

For simplicity of expression, we can further rewrite (2.2.12) as

$$\begin{aligned}\ell(\theta \mid \zeta_{Y_X},\gamma_{Y_X}) = \sum_{i=1}^{n} y_i^* \log\Big((1-\zeta_{Y_{X_i}})\pi_i + \gamma_{Y_{X_i}}(1-\pi_i)\Big) \qquad (2.2.13) \\ + (1-y_i^*)\log\Big(\zeta_{Y_{X_i}}\pi_i + (1-\gamma_{Y_{X_i}})(1-\pi_i)\Big)\end{aligned}$$

Again, estimation of (2.2.13) can be achieved in two ways;

1. When there is lack of validation subgroup study, an assumed value can be prespecified for both misclassification rates. An assume correct values for $\zeta_{Y_{X_i}}$ and $\gamma_{Y_{X_i}}$ would yield unbiased estimated parameter of interest and correct inferences can be made from the analysis.

2. In addition, information about probabilities of measurement error $\zeta_{Y_{X_i}}$ and γ_{Y_X} can be obtained from validation (internal or external) subgroup study.

In the following, we described how to correct for measurement error in a binary response variable assumed to be differentialy misclassified using observed and internal validation substudy data. The implementation is illustrated as follows;

True outcome Y_v measures are recorded or available for each participants from internal subgroup (Lyles et al., 2011). As a result, participants from the validation substudy have complete data (x_v, y_v, y_v^*). By using the complete data from the validation study, Lyles et al. (2011) defined the likelihood function of the internal validation subgroup data as

$$\begin{aligned}\mathcal{L}_v = \prod_{v=1}^{n_v} \Bigg(&\Big(P(y_v^* = 1 | y_v = 1, x_v) P(y_v = 1 | x_v) \Big)^{y_v^* y_v} \\ &\Big(P(y_v^* = 1 | y_v = 0, x_v) P(y_v = 0 | x_v) \Big)^{y_v^*(1-y_v)} \\ &\Big(P(y_v^* = 0 | y_v = 1, x_v) P(y_v = 1 | x_v) \Big)^{(1-y_v^*) y_v} \\ &\Big(P(y_v^* = 0 | y_v = 0, x_v) P(y_v = 0 | x_v) \Big)^{(1-y_v^*)(1-y_v)} \Bigg)\end{aligned}$$

The corresponding log-likelihood is

$$\begin{aligned}\ell_v(\theta \mid \zeta_{Y_{X_v}}, \gamma_{Y_{X_v}}) = \sum_{v=1}^{n_v} & y_v^* y_v \log \Big((1 - \zeta_{Y_{X_v}}) \pi_v \Big) + y_v^*(1 - y_v) \log \Big(\gamma_{Y_{X_v}} (1 - \pi_v) \Big) \\ &+ y_v (1 - y_v^*) \log \Big(\zeta_{Y_{X_v}} \pi_v \Big) + (1 - y_v^*)(1 - y_v) \log \Big((1 - \gamma_{Y_{X_v}})(1 - \pi_v) \Big)\end{aligned}$$

where $\zeta_{Y_{X_v}}$ and $\gamma_{Y_{X_v}}$ are the probabilities of measurement error in the validation subgroup. Both are assumed to be dependent on the true covariate X_v, and π_v is from the validation study.

Therefore, when using both observed and internal validation substudy, the full log-likelihood can be use to account for measurement error in observed response variable. The full log-likelihood is

$$\begin{aligned}\ell = &\sum_{i=1}^{n} \Big(y_i^* \log \big((1 - \zeta_{Y_{X_i}}) \pi_i + \gamma_{Y_{X_i}} (1 - \pi_i) \big) + (1 - y_i^*) \log \big(\zeta_{Y_{X_i}} \pi_i + (1 - \gamma_{Y_{X_i}})(1 - \pi_i) \big) \Big) \\ &+ \sum_{v=1}^{n_v} \Big(y_v^* y_v \log \big((1 - \zeta_{Y_{X_v}}) \pi_v \big) + y_v^* (1 - y_v) \log \big(\gamma_{Y_{X_v}} (1 - \pi_v) \big) \\ &+ y_v (1 - y_v^*) \log \big(\zeta_{Y_{X_v}} \pi_v \big) + (1 - y_v^*)(1 - y_v) \log \big((1 - \gamma_{Y_{X_v}})(1 - \pi_v) \big) \Big)\end{aligned}$$

An example is demonstrated by Lyles et al., (2011). They used observed and internal validation data with one continous and binary covariate(s), to account for measurement error where a response variable is assumed to be differentially misclassified.

2.2.4 Other Possible Misclassification Patterns

In practice, other different measurement error patterns exist. For instance, another example of misclasification pattern could be that the probabilities of measurement error in Y are independent on some subsets of the covariates, and dependent on others. It is nearly impossible to list all the misclassification patterns potentially present in an observed data. However, the likelihood specification would depend on the assumptions about the chosen misclasification pattern, and the data structure of the covariate(s) from the study.

2.3 Measurement error in Outcome and Covariates

In this section, we considered a case where both the response and exposure variable are measured with error. When this happen, the validity of the estimates of parameters are impacted by the probabilities of measurement error present in both response and exposure variables. In this case, the pattern of misclassification rates appears to be complex (as shown in Figure 2.4), but Tang et al., (2015) described a maximum likelihood framework on how to account for measurement error using observed and internal validation substudy.

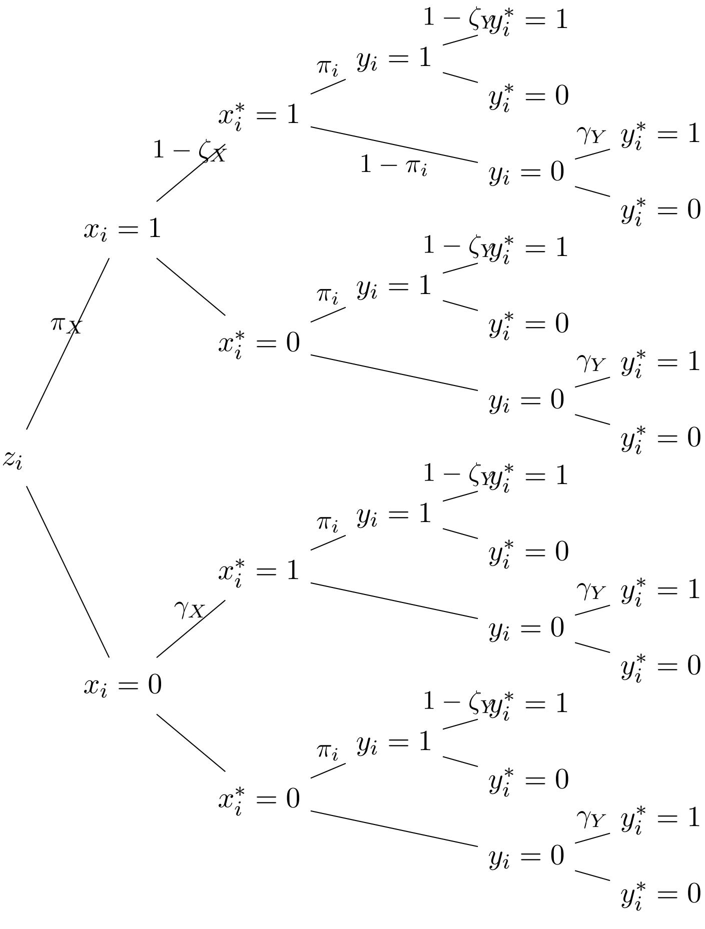


Figure 2.4: Mechanism of misclassification of both outcome and covariate

Suppose a standard logistic regression model fitted to a realization of true measures (Y_i, X_i, Z_i), for $i = 1, \cdots, n$ and n being the number of sample size. Both the response Y_i and covariate variable X_i are binary, and Z_i are some other covariate(s). The true distribution for response variable is $Y_i|X_i, Z_i \sim Bin(1, \pi_i)$, and π_i is

$$\pi_i = P(Y_i = 1|X_i, Z_i) = \frac{\exp(\alpha + \theta X_i + \psi Z_i)}{1 + \exp(\alpha + \theta X_i + \psi Z_i)}, \tag{2.3.1}$$

where θ and ψ are unknown true parameters, α is nuisance parameter.

Here, observed study contain (Y_i^*, X_i^*, Z_i) where both outcome Y_i^* and exposure variable X_i^* are potentially measured with error, but Z_i is assumed to be some other error free covariate(s).

For the observed study, the assumed logistic model is $Y_i^*|X_i, Z_i \sim Bin(1, \pi_i^*)$, and π_i^* is

$$\pi_i^* = P(Y_i^* = 1|X_i^*, Z_i) = \frac{\exp(\alpha^* + \theta^* X_i^* + \psi^* Z_i)}{1 + \exp(\alpha^* + \theta^* X_i^* + \psi^* Z_i)}$$

where θ^* and ψ^* are parameters. The corresponding log-likelihood function is

$$\ell(\theta^*) = \sum_{i=1}^{n} y_i^* \log \pi_i^* + (1 - y_i^*) \log(1 - \pi_i^*), \tag{2.3.2}$$

However, the conclusion of the study based on (2.3.2) might be misleading as the problem of measurement error of both outcome Y_i^* and covariate X_i^* was ignored. Therefore, there is need to investigate how the conclusion can be impacted by misclassification. Tang et al. (2015) used observed and internal validation subgroup in assessing complex measurement error process arising from mismeasured exposure and outcome variables. They incorporate probabilities of measurement error arising from both response and exposure variables in the likelihood function of an observed/internal validation design. In the following sections, we consider three types of potential misclassification in observed study. These are nondifferential, independent differential and dependent differential measurement error patterns.

2.3.1 Independent Nondifferential Misclasssification in both Outcome and Covariate

In Sections 2.1.4 and 2.2.2, sensitivity and specificity of response Y^* is $(1-\zeta_Y)$ and $(1-\gamma_Y)$ respectively, while sensitivity and specificity of exposure X^* is $(1 - \zeta_X)$ and $(1 - \gamma_X)$. In this section, the assumption is that both sensitivity and specificity of Y_i^* and X_i^* are independent of Z_i. By assuming probabilities of measurement error for both response and covariate, it follows that in the case of independent nondifferential measurement error mechanism,

$$\begin{aligned} P(Y^*, X^*, Y, X|Z) &= P(Y^*|Y, X^*, X, Z) \times P(X^*|X, Y, Z) \times P(Y|X, Z) \times P(X|Z) \\ &= P(Y^*|Y) \times P(X^*|X) \times P(Y|X, Z) \times P(X|Z) \end{aligned}$$

where

- $P(Y^*|Y)$ is sensitivity and specificity of outcome variable Y, that is, $(1 - \zeta_Y)$ and $(1 - \gamma_Y)$
- $P(X^*|X)$ is sensitivity and specificity of exposure X, that is, $(1 - \zeta_X)$ and $(1 - \gamma_X)$

- $P(Y|X, Z)$: model (2.3.1)
- $P(X|Z)$: Relates exposure variable X with some other covariate Z

The corresponding likelihood function for the observed study is

$$
\begin{aligned}
\ell = \prod_{i=1}^{n} \Bigg(\Bigg(& P(y_i^* = 1|y_i = 1)P(y_i = 1|x_i^*)P(x_i^* = 1|x_i = 1)P(x_i|z_i) \\
& + P(y_i^* = 1|y_i = 0)P(y_i = 0|x_i^*)P(x_i^* = 1|x_i = 1)P(x_i|z_i) \\
& + P(y_i^* = 1|y_i = 1)P(y_i = 1|x_i^*)P(x_i^* = 0|x_i = 1)P(x_i|z_i) \\
& + P(y_i^* = 1|y_i = 0)P(y_i = 0|x_i^*)P(x_i^* = 0|x_i = 1)P(x_i|z_i) \\
& + P(y_i^* = 1|y_i = 1)P(y_i = 1|x_i^*)P(x_i^* = 1|x_i = 0)P(x_i|z_i) \\
& + P(y_i^* = 1|y_i = 0)P(y_i = 0|x_i^*)P(x_i^* = 1|x_i = 0)P(x_i|z_i) \\
& + P(y_i^* = 1|y_i = 1)P(y_i = 1|x_i^*)P(x_i^* = 0|x_i = 0)P(x_i|z_i) \\
& + P(y_i^* = 1|y_i = 0)P(y_i = 0|x_i^*)P(x_i^* = 0|x_i = 0)P(x_i|z_i) \Bigg)^{y_i^*} \\
& \times \Bigg(P(y_i^* = 0|y_i = 1)P(y_i = 1|x_i^*)P(x_i^* = 1|x_i = 1)P(x_i|z_i) \\
& + P(y_i^* = 0|y_i = 0)P(y_i = 0|x_i^*)P(x_i^* = 1|x_i = 1)P(x_i|z_i) \\
& + P(y_i^* = 0|y_i = 1)P(y_i = 1|x_i^*)P(x_i^* = 0|x_i = 1)P(x_i|z_i) \\
& + P(y_i^* = 0|y_i = 0)P(y_i = 0|x_i^*)P(x_i^* = 0|x_i = 1)P(x_i|z_i) \\
& + P(y_i^* = 0|y_i = 1)P(y_i = 1|x_i^*)P(x_i^* = 1|x_i = 0)P(x_i|z_i) \\
& + P(y_i^* = 0|y_i = 0)P(y_i = 0|x_i^*)P(x_i^* = 1|x_i = 0)P(x_i|z_i) \\
& + P(y_i^* = 0|y_i = 1)P(y_i = 1|x_i^*)P(x_i^* = 0|x_i = 0)P(x_i|z_i) \\
& + P(y_i^* = 0|y_i = 0)P(y_i = 0|x_i^*)P(x_i^* = 0|x_i = 0)P(x_i|z_i) \Bigg)^{(1-y_i^*)} \Bigg)
\end{aligned}
$$

Using probabilities of measurement error ζ_Y, γ_Y, ζ_X, and γ_X,

$$
\begin{aligned}
\mathcal{L} = \prod_{i=1}^{n} \Bigg(\Bigg(& (1-\zeta_Y)\pi_i(1-\zeta_X)\pi_X + \gamma_Y(1-\pi_i)(1-\zeta_X)\pi_X + (1-\zeta_Y)\pi_i\zeta_X\pi_X + \gamma_Y(1-\pi_i)\zeta_X\pi_X \\
& + (1-\zeta_Y)\pi_i\gamma_Y\pi_X + \gamma_Y(1-\pi_i)\gamma_X\pi_X + (1-\zeta_X)\pi_i(1-\gamma_X)\pi_X + \gamma_Y(1-\pi_i)\gamma_X\pi_X \Bigg)^{y_i^*} \\
& \times \Bigg(\zeta_Y\pi_i(1-\zeta_X(1-\pi_X) + (1-\gamma_Y)(1-\pi_i)(1-\zeta_X\pi_X + (1-\gamma_Y)\pi_i(1-\gamma_X)\pi_X \\
& + (1-\gamma_X)(1-\pi_i)\zeta_X\pi_X + \zeta_Y\pi_i\gamma_X\pi_X + (1-\gamma_Y)(1-\pi_i)\gamma_X\pi_X + \zeta_Y\pi_i(1-\gamma_X)\pi_X \\
& + (1-\gamma_Y)(1-\pi_i)(1-\gamma_X)\pi_X \Bigg)^{(1-y_i^*)} \Bigg)
\end{aligned}
$$

To avoid complexity, Tang et al. (2015) assume logit link of $P(X|Z)$;

$$\text{logit}(P(X = 1|Z_1, \cdots, Z_g)) = \omega_0 + \sum_{g=1}^{G} \omega_g Z_g,$$

where ω_0 and ω_g are parameters that relates X to Z. Note that $P(X|Z)$ should be specified correctly, otherwise it could lead to biased results.

Hence, by using the total probability rule, the likelihood term of individual participant from observed study is

$$\begin{aligned} P(Y^* = y^*, X^* = x^*|Z = z) &= \sum_{y=0}^{1}\sum_{x=0}^{1} \big(P(y^*|y, x^*, x, z) \times P(x^*|x, y, z) \times P(y|x, z) \times P(x|z)\big) \\ &= \sum_{y=0}^{1}\sum_{x=0}^{1} \big(P(y^*|y) \times P(y|x, z) \times P(x|z) \times P(x^*|x)\big) \quad (2.3.1) \end{aligned}$$

The general likelihood for the observe study under independent nondifferential measurement error pattern is

$$\mathcal{L} = \prod_{i=1}^{n} \Big(\sum_{y_i=0}^{1}\sum_{x_i=0}^{1} P(y_i^*|y_i) \times P(y_i|x_i, z_{iy}) \times P(x_i|z_{ix}) \times P(x_i^*|x_i) \Big) \tag{2.3.3}$$

The subscript ix and iy represent version of covariates Z vector which may vary across models so far as z_{ix} is a subset of z_{iy} (Tang et al., 2015).

Prespecified values can be assumed for ζ_Y, γ_Y, ζ_X and γ_X, but Tang et al., (2015) combined observed and internal validation substudy for estimation purposes. Each participants from the internal validation subgroup is assumed to have complete data, (y_v^*, y_v, x_v^*, x_v). Hence, the likelihood function of the validation subgroup is

$$\ell_v = \prod_{v=1}^{n_v} P(y_v^*|y_v) \times P(y_v|x_v, z_v) \times P(x_v|z_v) \times P(x_v^*|x_v) \tag{2.3.4}$$

Therefore, the overall likelihood is $\ell \times \ell_v$, that is, the product of (2.3.3) and (2.3.4).

2.3.2 Independent Differential Misclasssification in both Outcome and Covariate

In this section, the assumption is based on differential misclassification mechanism where measurement error probabilities depend on true measures of other variables. In the following, we illustrate the method from Tang et al, (2015). Let

$$P(Y^* = 1|Y = 1, X = x, Z = z) = (1 - \zeta_{Y_Z}) \tag{2.3.5}$$

$$P(Y_i^* = 0|Y_i = 0, X = x, Z = z) = (1 - \gamma_{Y_Z}) \tag{2.3.6}$$

$$P(X^* = 1|X = 1, Y = y, Z = z) = (1 - \zeta_{X_Z}) \tag{2.3.7}$$

$$P(X^* = 0|X = 0, Y = y, Z = z) = (1 - \gamma_{X_Z}), \tag{2.3.8}$$

be the sensitivity and specificity for both variables, where the subscript used in (2.3.5) to (2.3.8) indicate that probabilities of measurement error of $Y$and X depend on true measures of other covariates Z. Tang et al. (2015) highlighted that these probabilities are functions of exposure X and some other covariates Z. Hence classification probabilities of observed outcome variable y_i^* is

$$(1 - \zeta_{Y_Z}) = \frac{\exp\left(\beta_0 + \sum_{g=1}^{G} \beta_g Z_g + \beta_{R+1} Y + \beta_{R+2} X\right)}{1 + \exp\left(\beta_0 + \sum_{g=1}^{G} \beta_g Z_g + \beta_{R+1} Y + \beta_{R+2} X\right)} \tag{2.3.9}$$

$$(1 - \gamma_{Y_Z}) = \frac{1}{1 + \exp\left(\beta_0 + \sum_{g=1}^{G} \beta_g Z_g + \beta_{R+2} X\right)}, \tag{2.3.10}$$

where β's are parameters obtained from modelling both of the measurement error probabilities of mis-measured outcome y_i^*. Similarly, it follows that

$$(1 - \zeta_{X_Z}) = \frac{\exp\left(\psi_0 + \sum_{d=1}^{D} \psi_d Z_d + \psi_{D+1} X + \psi_{D+2} Y\right)}{1 + \exp\left(\psi_0 + \sum_{d=1}^{D} \psi_d Z_d + \psi_{D+1} X + \psi_{D+2} Y\right)} \tag{2.3.11}$$

$$(1 - \gamma_{X_Z}) = \frac{1}{1 + \exp\left(\psi_0 + \sum_{d=1}^{D} \psi_d Z_d + \psi_{D+1} X + \psi_{D+2} Y\right)}, \tag{2.3.12}$$

where ψ's are parameters obtained from modelling both of the measurement error probabilities of mis-measured covariate x_i^*.

Therefore, the likelihood function of the observed study when both response and exposure is independently differentially misclassified is

$$\ell = \prod_{i=1}^{n} \Big(\sum_{y_i=0}^{1} \sum_{x_i=0}^{1} P(y_i^*|y_i, x_i, z_{iy^*}) \times P(y_i|x_i, z_{iy}) \times P(x_i|z_{ix}) \times P(x_i^*|x_i, y_i, z_{ix^*}) \Big) \tag{2.3.13}$$

Also, the likelihood function of the validation subgroup is

$$\ell_v = \prod_{v=1}^{n_v} \big(P(y_v^*|y_v, x_v, z_{vy^*}) \times P(y_v|x_v, z_{vy}) \times P(x_v|z_{vx}) \times P(x_v^*|x_v, y_v, z_{vx^*}) \big) \tag{2.3.14}$$

where the term z_{iy^*} and z_{vy^*} in both observed and internal validation likelihood function indicate that the misclassification process of the observed response variable y_* is conditional on true measure of some other covariates. Likewise, z_{ix^*} and z_{vx^*} indicate that the misclassification process of the observed covariate x_* is conditional on some true measure of covariate(s).

The overall likelihood is the product of (2.3.13) and (2.3.14). The difference between the likelihood function in this Section and that of the nondifferential likelihood in Section 2.3.1 is as a result of the modelling of the probabilities of measurement error defined in (2.3.9) to (2.3.12) (Tang et al., 2015).

2.3.3 Dependent Differential Misclasssification in both Outcome and Covariates

In Section 2.3.2, sensitivity and specificity of response and exposure variables depend on true measures of other variables. However, in this section, we consider the case in which sensitivity and specificity of one variable may be dependent on other mismeasured variables. The pattern under this case is referred to as dependent differential misclassification. Here, the assumption is that sensitivity and specificity of response Y^* is dependent on X^* and vice versa.

Let

$$\begin{aligned} P(Y^* = y^*, X^* = x^* | Z = z) &= \sum_{y=0}^{1} \sum_{x=0}^{1} P(y^*, x^*, y, x|z) \\ &= \sum_{y=0}^{1} \sum_{x=0}^{1} P(y^*|y, x, x^*, z) \times P(x^*|x, y, z) \times P(y|x, z) \times P(x|z) \end{aligned} \tag{2.3.15}$$

where

- From (2.3.15), $P(y^*, x^*, y, x|z)$ is the sensitivity or specificity of Y_i^* dependent on error-prone exposure X_i^* conditional on (Y_i, X_i, Z_i).

From Tang et al. (2015), the likelihood function of individual participants from observed study is

$$\ell = \sum_{y_i=0}^{1} \sum_{x_i=0}^{1} \Big(P(y_i^*|y_i, x_i, x_i^*, z_{iy^*}) \times P(y_i|x_i, z_{iy}) \times P(x_i|z_{ix}) \times P(x_i^*|x_i, y_i, z_{ix^*}) \Big) \tag{2.3.16}$$

It follows that

$$\ell_v = P(y_v^*|y_v, x_v^*, x_v, z_{vy^*}) \times P(y_v|x_v, z_{vy}) \times P(x_v|z_{vx}) \times P(x_v^*|x_v, y_v, z_{vx*}) \tag{2.3.17}$$

where (2.3.17) is the likelihood of individual participants from internal validation substudy.

Hence, the overall likelihood is product of (2.3.16) and (2.3.17). The overall likelihood can be simplify further when correct model is selected for $P(X|Z)$ or measurement error models (Tang et al., 2015). This is because when exposure variables involve in $X|Z$ or measurement error models are excluded from an outcome model, the resulting estimate could be biased. For example, suppose variable U is included in the $X|Z$, and Y is dependent on U conditional on X and all other covariates in a real study. If U is excluded from the outcome model, the validity of the analytic result is subject to bias (Tang et al., 2015). Likewise, all variables of measurement error models should be accounted for in the model (2.3.1), except there is evidence suggesting an assumption about the independency of Y conditional on X and other covariates in the outcome variable model (2.3.1) (see an empirical detail in Section 4.2 of Tang et al., 2015). Hence, it is important to account for all variables involved in the measurement error models and the covariate(s) in the possible relationship $P(X|Z)$.

2.4 Bayesian Method

This research is placed within the Frequentist framework, hence the reason for the extensive review illustrated in the previous sections. However, a brief review of Bayesian approach in accounting for measurement error is highlighted in this section.

Bayesian concept of probability differs from Frequentists approach. In Bayesian context, model parameters are treated as random variables, and probability distribution of parameter of a model is interpreted based on the degrees of believe about the parameter values. Meanwhile in the Frequentist approach, parameters of model are fixed and probabilities are treated as frequencies observed from an experiment. As a result, Bayesian methods on accounting for measurement error is different from the methods illustrated in Section 2.1 to 2.3. Although Bayesian methods can be formulated in measurement error adjustment, but the methods have not been widely used in many application (Breslow, 1990). This might be due to computational and statistical difficulty in analyzing a particular case of measurement error.

In Bayesian context, some assumptions are made when dealing with measurement error of exposure variable. It is assume that at the beginning of a study, information about the distribution of the unobserved exposure is available from the population (Richardson and Gilks, 1993). This possible information leads to specification of prior distribution of true X. However, if the study is from a different population, the prior distribution of true X is imprecise. In addition, during the study it is assume that information obtained from the observed exposure X^* reduces uncertainty in the unobserved (Richardson and Gilks, 1993). A bayesian perspective of measurement error is illustrated by Richardson & Gilks (1993). They developed a parametric method in terms of conditional independence modelling where different sources of information can be integrated for parameter estimation using gibbs sampling (see Richardson & Gilks 1993). In the following we illustrate the layout/design of measurement error using the approach from Richardson and Gilks, (1993).

Let Y_i denote the outcome variable, X_i is unknown true covariate, X_i^* is error prone exposure. There are two stages to conditional independence models of measurement error; the structural and functional stage. First, is the structural stage of the model, followed by functional part. One of the building blocks of structural part is conditional independence assumptions, where all the uncertainties about X is accounted for, such that each variable in the model is assumed to be related conditionally to a subset of other variables (Richardson and Gilks, 1993). It is important to note that correct specification of conditional independence assumptions would strengthen inference drawn from an analysis. And in the functional part, all the required distribution are stated.

In the case of measurement error of X, Richardson and Gilks (1993) broke down the conditional independence assumptions into two simpler components: a disease model and measurement error model. The disease model relate X, and some other known covariates U to Y, while measurement error model relates observed X^* and the true measure X. Thus, the following model conditions are the components of the structural part considered by Richardson and Gilks, (1993)

$$\text{outcome or disease model} = Y_i|X_i, U_i, \theta \tag{2.4.1}$$

$$\text{measurement error model} = X_i^*|X_i, \psi \tag{2.4.2}$$

$$\text{covariate model } = X_i|U_i, \pi \tag{2.4.3}$$

where θ, π, and ψ are the prior distributions of the parameters (that is, the respective information available from the outset of the study).

Therefore, the product of all the conditional distribution models (commonly referred to as joint distribution) can be used as the structural part of conditional independence model when dealing with measurement error in bayesian context (Richardson and Gilks, 1993). The joint distribution is given as

$$\text{Joint distribution } = [\theta][\pi][\psi]\prod_{i=1}^{n}(X_i|U_i, \pi)\prod_{i=1}^{n}(X_i^*|X_i, \psi)\prod_{i=1}^{n}(Y_i|X_i, U_i, \theta) \tag{2.4.4}$$

Richardson and Gilks (1993) highlighted that in general 2.4.4 can be said to have no local dependencies other than those stated by the model (2.4.1), (2.4.2) and (2.4.3). In otherwords, (2.4.4) implies there are other additional conditional independence assumptions (see Section model conditionals of Richardson and Gilks, (1993) for more detail).

Next, is the functional part of the conditional independence model. The functional part is the specification of the parametric form of all the conditional distributions. For instance, in this example, logistic regression can be defined for the response variable model (2.4.1). Description of the parametric distribution mostly depends on the area of study or investigators choice, but the distribution need to be define before estimation. Richardson and Gilks (1993) suggested the use of sensitivity analysis when there is conflict about the choice of parametrization.

Bayesian estimator is based on the posterior distribution of the parameters (that is, θ, ψ,π and unobserved exposure X) given the data (Richardson and Gilks, 1993). They are interested in the marginal posterior distribution of θ given the data, (that is, the distribution of θ when other parameters are unknown), but computation was intractable. Hence, Richardson and Gilks (1993) generated samples from the joint posterior distribution of all of θ, ψ,π and unobserved exposure X using Gibbs sampling method. Inferences on parameter θ can be made from samples drawn from the joint posterior distribution of parameters given the study data.

McInturff et al. (2004) proposed a Bayesian method that can be used to account for a nondifferential measurement error in a binary outcome. McInturff et al. (2004) account for unknown sensitivity and specificity, while they include expert prior knowledge in the form of a conditional mean prior in a model. They explored probit, logit and complimentary log log link functions, and Bayes factor (see Kass and Raftery, (1995) for more detail on Bayes factor) was used for comparison and model selection. The proposed method is as follows;

Suppose a logistic model is considered for a binary response variable Y which is related to a vector of covariates $\mathbf{X}$ such that $Y \sim Bin(1, \pi)$ and π is

$$\pi \equiv P(Y = 1|x) = \frac{\exp(\alpha + \theta x)}{1 + \exp(\alpha + \theta x)}$$

where α is a nuisance parameter, and θ is a vector of regression coefficients. Alternative link functions include

$$\begin{aligned} \text{probit;} \quad & \pi \equiv \Phi(\theta x) \\ \text{complimentary log log;} \quad & \pi \equiv 1 - \exp\left(-\exp\left(\theta x\right)\right) \end{aligned}$$

Generally,

$$g(\pi) = \theta x$$

By using generalized linear model notation

$$\pi = g^{-1}(\theta x) \quad \text{for monotone g(.)}$$

where $g(\,\cdot\,)$ is the link function.

However, for an observed data (y_i^*, x_i), $i = 1, \cdots, n$, y_i^* is misclassified outcome and $\mathbf{x}_i$ is row vector of exposure assumed to be error-free. For simplicity, McInturff et al. (2004) assume $y_i|x_i \sim Bernoulli(\pi_i^*)$, with probability of success calculated from

$$\pi^* = P(Y^* = 1|x) = \pi(1 - \zeta_Y) + (1 - \pi)\gamma_Y$$

The likelihood function is given as

$$\ell(\theta, (1 - \zeta_Y), (1 - \zeta_Y)) = \prod_{i=1}^{n} (\pi_i(1 - \zeta_Y) + (1 - \pi_i)\gamma_Y)^{y_i} (\pi_i\zeta_Y + (1 - \pi_i)(1 - \gamma_Y))^{1-y_i}$$

An independent normal distributions with zero mean and a large variance is a common choice for specifying prior distribution for θ (McInturff et al., (2004). However, McInturff et al. (2004) used conditional means prior approach illustrated by Bedrick et al. (1997) to induce prior distribution for θ irrespective of the link function. Thus, the joint posterior is given as

$$p((1 - \zeta_Y), (1 - \zeta_Y), \theta|X, Y^*) \propto L((1 - \zeta_Y), (1 - \zeta_Y), \theta) \times p(\theta) \times p((1 - \zeta_Y)) \times p((1 - \gamma_Y)) \tag{2.4.5}$$

McInturff et al. (2004) highlighted that distribution in (2.4.5) is not identifiable, but sampling is possible through Gibbs sampling. A common choice for modelling uncertainty of a probability is to use Beta distribution $Be(a, b)$. McInturff et al. (2004), and Bedrick et al. (1997) highlighted that $Be(a, b)$ distribution can be done by indicating parameters a and b from an imaginary binomial trial in which $a - 1$ successes out of $a + b - 2$ trials are observed. Following Bedrick et al. (1997) approach, McInturff et al. (2004) assumed independent beta prior $Be(a_{(1-\zeta_Y)}, b_{(1-\zeta_Y)})$ and $Be(a_{(1-\gamma_Y)}, b_{(1-\gamma_Y)})$ for sensitivity and specificity respectively. Then, Bayes factor was used to select which of the links function is most appropriate.

Gerlach and Stamey (2007) highlighted that when misclassification of a variable depends on covariate(s), and differential misclassification is assumed, the individual model involved is large. Hence, Gerlach and Stamey (2007) proposed a Bayesian model selection to determine what set of covariates to be included in such model.

Suppose an observed sample $y^* = y_1^*, y_2^*, \cdot, y_s^*$, where s is the number of distinct error-free covariate patterns. For the observed data, the assumed logistic model

$$Y_i^* \sim Bin(n_i, \pi_i^*)$$

where $\pi_i^* = \pi_i(1 - \zeta_Y) + (1 - \pi_i)\gamma_Y$, and π_i is assumed to depend on a set of p exposure variables $\mathbf{X}$, through an assumed logistic model,

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \theta_1\mathbf{X}_i, \quad i = 1, \cdots, s$$

Gerlach and Stamey (2007) assume independent beta priors for ζ_Y and γ_Y. And in the case of differential misclassification,

$$\pi_i^* = \pi_i(1 - \zeta_{Y_{X_i}}) + (1 - \pi_i)\gamma_{Y X_i}$$

where

$$\begin{cases} \zeta_{Y_{X_i}} & \log\left(\frac{\zeta_{Y_{X_i}}}{1-\zeta_{Y_{X_i}}}\right) = \theta_2 \mathbf{X}_i, \quad i = 1, \cdots, s \\ \gamma_{Y_{X_i}}, & \log\left(\frac{\gamma_{Y_{X_i}}}{1-\gamma_{Y_{X_i}}}\right) = \theta_3 \mathbf{X}_i, \quad i = 1, \cdots, s \end{cases}$$

Gerlach and Stamey (2007) assume validation data is available for the *ith* covariate pattern, such that

$$\begin{aligned} z_i &\sim Bin(n_{v_i}, \pi_i) \\ r_i | z_i &\sim Bin(z_i, \zeta_{Y_{X_i}}) \\ k_i | z_i &\sim Bin(n_{v_i} - z_i, \gamma_{Y_{X_i}}) \end{aligned}$$

where z_i is the total number of actual occurrences in covariate pattern i, n_{v_i} is the sample size of a validation data, r_i is the total number of actual occurrences labelled as non-occurrences and k_i is the number of actual non-occurrences labelled as occurrences (see more detail in Section 2; Gerlach and Stamey, 2007).

Hogg et al, (2017) proposed a simple Bayesian approach for the analysis of pair-matched study subject to response misclassification. They considered the association between a binary exposure variable X and an outcome Y in a pair-matched case-control setting. Here, Y is true, but not observable, X is true, and Y_i^* is observed such that $Y_i^* = 1$ is the case, and $Y_i^* = 0$ is the control. In the case of matched sampling, $Y_i^* = 1$ are chosen from the population of interest, and $Y_i^* = 0$ are matched to cases on a set of error-free covariates or factors that might influences the association of X and Y_i. Under matched case control sampling, the model incorporates positive $P(Y = 1|Y_i^* = 1)$ and negative $P(Y = 0|Y_i^* = 0)$ predictive values as a measure of misclassification, rather than sensitivity and specificity into the analysis.

2.5 Choice of Correct Model

These existing approaches of dealing with measurement error appears to have something in common, and that is, introducing misclassification probabilities in a standard logistic model. The process of modification yield many models based on the potential measurement error pattern assume. The choice of correct model is rather arbitrary (Copas, 2010), and there is no magical method of knowing the correct model. Hence, in Chapter Three we study the stability of an assumed and modified models. The approach is based on normal curvature, and using the motivating example described in Chapter One, we derived influence measure.

Chapter 3

Influence Measure for Misclassified Absence of Binary Outcome

This Chapter is the core of our research, which focuses on using curvature to study local influence and model stability. The Chapter is structured as follows. Section 3.1 introduces the general theory of local influence approach. In Section 3.2, curvature is introduced, followed by the development of an influence measure when a perturbation is incorporated in the general likelihood function of an assumed model in Section 3.3. More specifically, in Section 3.4, the approach is extended to a binary model where response variable $Y_i^* = 0$ is subject to misclassification.

3.1 Motivating Example

Here is an example we use to illustrate the concept of Cook's local influence approach. Consider a study of assessing the effectiveness of rehabilitation programme on juvenile offenders. We take the outcome for the ith subject in the study to be $Y_i^* = 1$ if there is no reconviction observed, otherwise $Y_i^* = 0$. The study has two groups. The exposure indicator variable is $X_i = 1$ for the rehabilitation group and $X_i = 0$ for the control group. For the observed data, the assumed logistic model is

$$Y_i^*|X_i \sim Bin(1, \pi_i^*), \quad \log\left(\frac{\pi_i^*}{1-\pi_i^*}\right) = \alpha^* + \theta^* X_i, \tag{3.1.1}$$

where α^* is the intercept and θ^* is the observed effect of rehabilitation programmes in reducing the odds of juvenile offence. The log-likelihood function is

$$\ell(\theta^*) = \sum_{i=1}^{n} y_i^* \log \pi_i^* + (1 - y_i^*) \log(1 - \pi_i^*), \tag{3.1.2}$$

where n is the sample size. The estimate of θ^* with its p-value or confidence interval is used to judge whether the rehabilitation programme can reduce the juvenile reoffending rate or not.

However, the conclusion of the study based on $\hat{\theta}^*$ might be misleading as the problem of undetected crimes/measurement error was not considered. If there are some undetected offences in the study, the true effect of the rehabilitation programme can be over or under estimated. Therefore, we need to investigate how the inference can be impacted by undetected crimes.

Under the problem of undetected crimes, there is measurement error in the observed response variable Y_i^*. Let Y_i be the unknown true outcome variable, which equals 0 if offence have occurred, and 1 otherwise. The true distribution for Y_i is

$$Y_i|X_i \sim Bin(1, \pi_i), \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \theta X_i, \tag{3.1.3}$$

where θ is the unknown true effectiveness of the rehabilitation programmes of interest. We make the reasonable assumption that there is no reconviction if there is no offence, so $Y_i^* = 1$ if $Y_i = 1$, but an actual offence may not lead to detection and reconviction, i.e. when $Y_i = 0$, Y_i^* is either 0 or 1. In other words, there is no measurement error if there is no offence, but there may be measurement error if there is an offence.

We assume the probability of non-detection to be $\gamma = P(Y_i^* = 1|Y_i = 0)$. The mechanism of measurement error is shown in Figure 3.1.

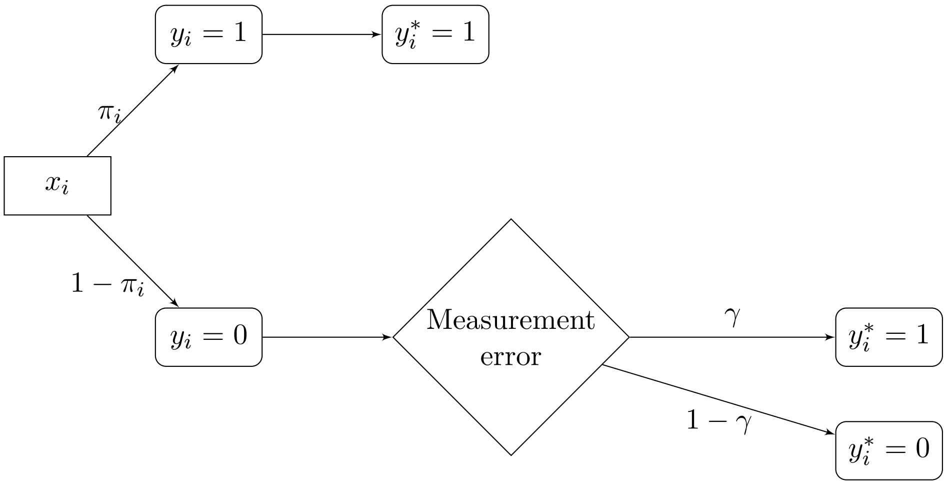


Figure 3.1: Mechanism of measurement error for the motivating example

By assuming the probability of non-detection γ,

$$\begin{aligned} P(Y_i^* = 1|X_i) =& P(Y_i^* = 1|Y_i = 1)P(Y_i = 1|X_i) + P(Y_i^* = 1|Y_i = 0)P(Y_i = 0|X_i) \\ =& \pi_i + \gamma(1 - \pi_i) \end{aligned}$$

The distribution of the observed outcome variable conditional on γ is

$$Y_i^*|X_i \sim Bin\left(1, \pi_i + \gamma(1-\pi_i)\right). \tag{3.1.4}$$

The corresponding log-likelihood for a given γ is

$$\ell(\theta|\gamma) = \sum_{i=1}^{n} y_i^* \log\Big(\pi_i + \gamma(1-\pi_i)\Big) + (1-y_i^*)\log\Big(1 - \big(\pi_i + \gamma(1-\pi_i)\big)\Big) \tag{3.1.5}$$

We call (3.1.1) the assumed model, where the superscript * indicates the parameters/variables are for the observed data only which are prone to measurement error. To consider the measurement error problem, we introduce the probability of measurement error γ as a modification weight into (3.1.1) and this results to the modified model (3.1.4). The true value of γ is unknown, but we can vary the value of γ between 0 and 1 to assess the impact of measurement error by comparing the assumed and modified models, likelihoods or estimates. When $\gamma = 0$, (3.1.4) is equal to (3.1.1), $\ell(\theta^*) = \ell(\theta|\gamma)$ and $\theta^* = \theta$. When $\gamma > 0$, the modified model is deviated from the assumed model and $\ell(\theta^*) \neq \ell(\theta|\gamma)$. Each value assigned to γ corresponds to a distinct model. Different values of $\gamma > 0$ perturb the assumed model (3.1.2) to be different modified models.

Here, γ is a scalar modification weight. But later, we will show that γ can be extended to a vector of modification weights, say $\gamma = (\gamma_1, ..., \gamma_q)$ bound to some $\Omega \subset \mathbb{R}^q$, with the assumption that there is a vector of weights, say γ_0, in Ω such that $\ell(\theta^*) = \ell(\theta|\gamma)$. An example is when γ depends on all or subset of covariates, we will have a vector of modification weights rather than a scalar.

3.1.1 Cook's Local Influence Approach

In this section, we now consider to use the Cook's likelihood displacement (Cook, 1986) to measure the difference between assume model (3.1.2) under the impact of measurement error and the modified model (3.1.5). Cook's concept of local influence approach is developed following the idea of likelihood displacement given in (3.1.6) which is based on comparing the maximum likelihood with the estimate of parameter $\hat{\theta}^*$ obtained from an assumed model to the maximum likelihood with the estimate $\hat{\theta}$ obtained from when a modification weights γ is incorporated in assumed model. The likelihood displacement studies the behaviour of the log-likelihood function of a model over the entire range of modification parameter γ given as

$$LD(\gamma) = 2[\ell(\hat{\theta}^*) - \ell(\hat{\theta}|\gamma)] \tag{3.1.6}$$

Note that the extension of (3.1.6) to include all quantities that affect LD depends on the modification of interest. Table 3.1 shows few rows of likelihood displacement $LD(\gamma)$ obtained from a simulated data under model (3.1.5). Data were simulated under conditions following the motivating example.

Table 3.1: Extract of Likelihood displacement results from a simulated data with γ ranges from 0 to 1

γ	LD(γ)
0	0
0.14	9.22
0.39	128.79
0.55	470.90
0.63	1051.67
0.67	2111.70
0.71	60749.91
0.75	79272.82
0.82	109246.20
0.87	128956.00
0.94	154457.70
0.99	171335.30

The response variable Y_i was simulated from model (3.1.3), with the true value $\theta = 1$ and $\alpha = -0.5$ with $n = 1000$ being the number of participants. The exposure indicator variable is $X_i \sim Bin(1000, 1, 0.5)$. The parameters used are one of the possible range of values for this study. Based on the assumption that there is no measurement error if there is no offence, the observed response variable Y_i^* was generated as $Y_i^* = 1$ if $Y_i = 1$, but there may be measurement error if there is an offence, thus $Y_i^*|Y_i = 0 \sim Bin(1, 0.5)$.

Modified model were generated from varying the value of γ from 0 to 1 using model (3.1.5). The assumed model estimate was calculated using log-likelihood function (3.1.2), and modified model is from log-likelihood (3.1.5). Hence, likelihood displacement (3.1.6) was evaluated from the simulated data. As shown in Table 3.1, there is a break point from when $\gamma = 0.68$. This could be a cut off value for the range of error size.

The graph of likelihood displacement $LD(\gamma)$ versus γ conventionally referred to as influence graph is illustrated in Figure 3.2. A point on the curve in Figure 3.2 is the difference between the assumed model (where $\gamma = 0$) and a modified model with a distinct $\gamma \neq 0$. The influence graph gives us how large the difference between the likelihoods of the assumed and modified models will change as the value of γ increases from 0 to 1. Discussion (and Figures) on how model estimates behave when a slight change in γ is introduced in an assumed model is shown in Chapter Five.

3.2 Introduction of Curvature

When more than one modification weights, $\gamma = (\gamma_1, \cdots, \gamma_q)$ for $q > 1$, is introduced in an assumed model, the plot of $LD(\gamma)$ against $(\gamma_1, \cdots, \gamma_q)$ is a q-dimensional object in a $\mathbb{R}^{q+1}$ space (e.g. a surface in a $\mathbb{R}^3$ space). As a results, in the cases $q > 1$, we are not able to plot

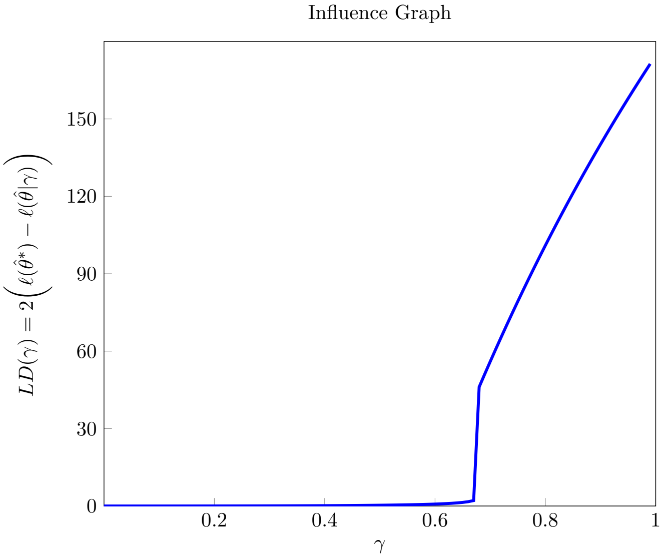


Figure 3.2: Influence Graph: Likelihood Displacement vs γ for simulation study mimicking our motivating example

an influence graph like the one in Figure 3.2, and it is difficult to describe the influence of varying γ on the models. To solve this problem, Cook (1986) proposed to use curvature as a measure which characterizes the behaviour of an influence graph around a selected point on a surface or higher dimensional object in space parameterised in $(\gamma_1, \cdots, \gamma_q, LD(\gamma))$. In our setting, each point on the surface of $(\gamma_1, \cdots, \gamma_q, LD(\gamma))$ represents a distinct model different from the assumed model, while the assumed model is represented by the origin of the surface at $LD(\gamma) = \gamma_1 = \gamma_2 = \cdots = \gamma_q = 0$. Curvature can be used to measure how curvy a surface is at a point with respect to slight change in $\gamma = (\gamma_1, \cdots, \gamma_q)$. This can be used to measure how sensitive or stable a distinct model on the surface is under the impact of slight modification in $\gamma = (\gamma_1, \cdots, \gamma_q)$.

In differential geometry, curvature, denoted by κ, is the rate of change of a unit tangent vector at any given point, turning in direction along a curve with respect to an arc length (that is, the size of a tiny step along a curve). As the example shown in Figure 3.3, consider a circle with radius R is inscribed to a curve at point p, the curvature at the point p is then equals to $\frac{1}{R}$. However, this definition is difficult to manipulate and to express in formulas. A more general form of curvature κ at a point along a curve in $\mathbb{R}^2$ is expressed as

$$\kappa = \left\| \frac{\partial T}{\partial s} \right\|, \tag{3.2.1}$$

where $T(t)$ represents a unit tangent vector along a curve in the arc length s, as detailed as follows, and $||.||$ is the notation of the norm or magnitude of $\frac{\partial T}{\partial s}$. Suppose a curve is parameterized by a function $\vec{r}(t)$ with each respective components dependent on parameter t, where t lies within an interval $a \leq t \leq b$. For example, the curve in Figure 3.2 is parameterized by the function $\vec{r}(\gamma) = (\gamma, LD(\gamma))$, where γ lies within an interval $0 \leq \gamma \leq 1$.

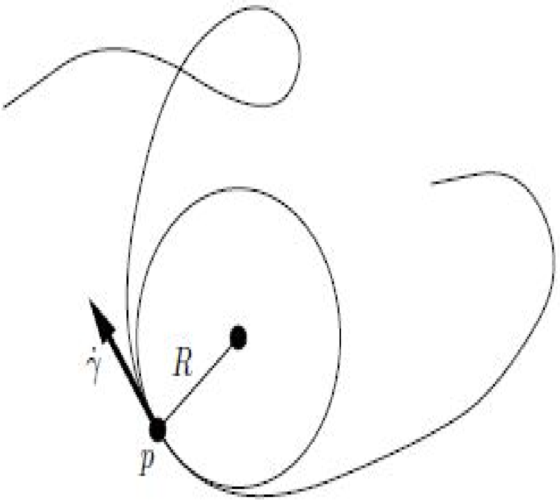


Figure 3.3: Osculating Circle: Curvature at a point: A circle with radius R inscribed to a curve at a given point p. The curvature at point p is equivalent to the curvature of the inscribed circle define as $\frac{1}{R}$, and $\dot{\gamma}$ is a unit tangent vector at point p. Figure from Lee, (1997).

A tangent or velocity vector at any given point t is the derivative of the function $\vec{r}(t)$ with respect to t at that point. Again, for the curve in Figure 3.2, the tangent vector $\frac{\partial \vec{r}(\gamma)}{\partial \gamma} = \left(1, \frac{\partial LD(\gamma)}{\partial \gamma}\right)$.

Normalising a tangent vector does not change the direction, but changes its length to 1 resulting to a unit tangent vector defined as

$$T(t) = \frac{\vec{r'}(t)}{\|\vec{r'}(t)\|}, \tag{3.2.2}$$

where $\|\vec{r'}(t)\|$ is the magnitude or length of $\vec{r'}(t)$, which can be obtain by taking the square-root of the sum of the squared of components of $\frac{\partial \vec{r}(\gamma)}{\partial \gamma}$. In our example, for instance, the components of tangent vector $\frac{\partial \vec{r}(\gamma)}{\partial \gamma}$ are 1 and $\frac{\partial LD(\gamma)}{\partial \gamma}$, hence, the magnitude of tangent vector $\frac{\partial \vec{r}(\gamma)}{\partial \gamma}$ is

$$\|\vec{r'}(t)\| = \left\|\frac{\partial \vec{r}(\gamma)}{\partial \gamma}\right\| = \sqrt{1 + \left(\frac{\partial LD(\gamma)}{\partial \gamma}\right)^2}.$$

The unit normal vector $N(t)$ points in the direction (up or down) the curve is turning, and $N(t)$ is defined as

$$N(t) = \frac{T'(t)}{\|T'(t)\|}. \tag{3.2.3}$$

The cross product of tangent and normal vector is referred to as binormal $B(t)$ defined as

$$B(t) = T(t) \times N(t). \tag{3.2.4}$$

The arc length s in (3.2.1) is the size of a tiny step or distance travelled along a curve within an interval $t_1 \leq t \leq t_2$ is defined as

$$s = \int_{t_1}^{t_2} \|\vec{r'}(t)\| \partial t. \tag{3.2.5}$$

3.2.1 Plane Curve

In this section, we will give an example of curvature for a plane curve. Let

$$\vec{r}(t) = (x(t), y(t)),$$

be a plane curve in a two dimension $\mathbb{R}^2$ space parameterized by t for any general function $\vec{r}(t)$, such that the x component and y component is given as $x = x(t)$, $y = y(t)$ respectively. Differentiating $\vec{r}(t)$ with respect to t gives a tangent vector

$$\vec{r'}(t) = (x'(t), y'(t)).$$

And the magnitude of $\vec{r'}(t)$ is

$$\|\vec{r'}(t)\| = \sqrt{(x'(t))^2 + (y'(t))^2}.$$

From (3.2.2), the unit tangent vector T(t) is given by

$$\vec{T}(t) = \left(\frac{x'(t)}{\sqrt{(x'(t))^2 + (y'(t))^2}}, \frac{y'(t)}{\sqrt{(x'(t))^2 + (y'(t))^2}} \right).$$

As a result, $\vec{T}(t)$ can be rewritten as

$$\vec{T}(t) = \frac{\vec{r'}(t)}{||\vec{r'}(t)||}. \tag{3.2.6}$$

Within interval $t_1 \leq t \leq t_2$, the arc length s is defined by

$$s = \int_{t_1}^{t_2} \sqrt{(x'(t))^2 + (y'(t))^2} \partial t.$$

From (3.2.5),

$$\frac{\partial s}{\partial t} = ||\vec{r'}(t)||. \tag{3.2.7}$$

Using (3.2.7), $T(t)$ in (3.2.6) can be rewritten as

$$\vec{T}(t) = \frac{\vec{r'}(t)}{\frac{\partial s}{\partial t}}. \tag{3.2.8}$$

Thus, the curvature κ of a plane curve in two dimension $\mathbb{R}^2$ given in (3.2.1) can be rewritten in terms of arc length as

$$\kappa = \left\| \frac{\vec{T}'(t)}{\frac{\partial s}{\partial t}} \right\|. \tag{3.2.9}$$

Using (3.2.7), curvature κ can also be expressed as

$$\kappa = \frac{||\vec{T}'(t)||}{||\vec{r}'(t)||}.$$

However, in practice, tangent vectors $\vec{T}(t)$ and its first derivative $\vec{T}'(t)$ are expressed in terms of t. Hence, calculating curvature κ in terms of arc length as defined in (3.2.9) may be inconvenient. From (3.2.8), $\vec{r}'(t)$ can be rewritten as

$$\vec{r}'(t) = \vec{T}(t)\frac{\partial s}{\partial t}.$$

With product rule, differentiating $\vec{r}'$(t) gives

$$\vec{r}''(t) = \frac{\partial^2 s}{\partial t^2}\vec{T}(t) + \vec{T}'(t)\frac{\partial s}{\partial t}.$$

Multiplying $\vec{r}'(t) \times \vec{r}''(t)$ gives

$$\vec{r}'(t) \times \vec{r}''(t) = \left[\frac{\partial s}{\partial t}\frac{\partial^2 s}{\partial t^2}\vec{T}(t) \times \vec{T}(t)\right] + \left[\left(\frac{\partial s}{\partial t}\right)^2 \vec{T}'(t) \times \vec{T}(t)\right]. \tag{3.2.10}$$

Since the cross product $\vec{T}(t) \times \vec{T}(t) = 0$, (3.2.10) reduces to

$$\vec{r}'(t) \times \vec{r}''(t) = \left[\left(\frac{\partial s}{\partial t}\right)^2 \vec{T}'(t) \times \vec{T}(t)\right]. \tag{3.2.11}$$

From (3.2.3)

$$T'(t) = N(t)||T'(t)||. \tag{3.2.12}$$

With chain rule, differentiating $T(t)$ gives

$$T'(t) = \frac{\partial T}{\partial t} = \frac{\partial s}{\partial t}\frac{\partial T}{\partial s}. \tag{3.2.13}$$

As a result, using (3.2.13), $T'(t)$ in (3.2.12) can be rewritten as

$$T'(t) = N(t)\frac{\partial s}{\partial t}\left\|\frac{\partial T}{\partial s}\right\|. \tag{3.2.14}$$

Since $\kappa = \|\frac{\partial T}{\partial s}\|$ in (3.2.1), $T'(t)$ in (3.2.14) is

$$T'(t) = N(t)\frac{\partial s}{\partial t}\kappa$$

Hence, from (3.2.11), $\vec{r}'(t) \times \vec{r}''(t)$ can be rewritten as

$$\begin{aligned}\vec{r}'(t) \times \vec{r}''(t) &= \left[\left(\frac{\partial s}{\partial t}\right)^2 N(t)\frac{\partial s}{\partial t}\kappa \times \vec{T}(t)\right] \\ &= \left(\frac{\partial s}{\partial t}\right)^3 \vec{T}(t)N(t)\kappa.\end{aligned}$$

From (3.2.4), $B(t) = \vec{T}(t)N(t)$. As $\frac{\partial s}{\partial t} = ||\vec{r}'(t)||$ and curvature κ are both magnitudes,

$$\begin{aligned}||\vec{r}'(t) \times \vec{r}''(t)|| &= \left\|\left(\frac{\partial s}{\partial t}\right)^3 B(t)\kappa\right\| \\ &= \left(\frac{\partial s}{\partial t}\right)^3 ||B(t)||\kappa. \qquad (3.2.16)\end{aligned}$$

Since both $T(t)$ and $N(t)$ have unit length, $B(t)$ has a unit length, as a result (3.2.16) is,

$$||\vec{r}'(t) \times \vec{r}''(t)|| = \left(\frac{\partial s}{\partial t}\right)^3 \kappa. \qquad (3.2.17)$$

From (3.2.7), $\frac{\partial s}{\partial t} = ||\vec{r}'(t)||$, such that $\left(\frac{\partial s}{\partial t}\right)^3 = ||r'(t)||^3$. Dividing both sides of (3.2.17) by $||\vec{r}'(t)||^3$ gives the curvature κ of a plane curve in two dimension $\mathbb{R}^2$ as

$$\kappa = \frac{||r'(t) \times r''(t)||}{||r'(t)||^3} \qquad (3.2.18)$$

3.2.2 Normal Curvature

In the case of $\mathbb{R}^3$ (or higher dimension), we will have a surface (or a higher dimensional object) rather than a plane curve. Curvature in these cases is called normal curvature, whose idea is a bit complicated. We use Figure 3.4 below to give a brief introduction of normal curvature. Consider a surface S and an arbitrary curve C through a point P of the surface. The curve C has the tangent vector v at the point P. Let N be the normal vector (often simply called the normal), which is the vector perpendicular to the surface S at the point P. Let π be the normal plane formed by the tangent vector v and the normal vector N. The intersection of the normal plane π with the surface S is a curve C_n named normal section. The curvature of C_n at the point P is called normal curvature.

When $q > 1$, the influence graph is a surface S in R^{q+1} formed by modification weights $\gamma = (\gamma_1, ..., \gamma_q)$ and the likelihood displacement, i.e. $S = (\gamma_1, ..., \gamma_q, LD(\gamma))$. To calculate

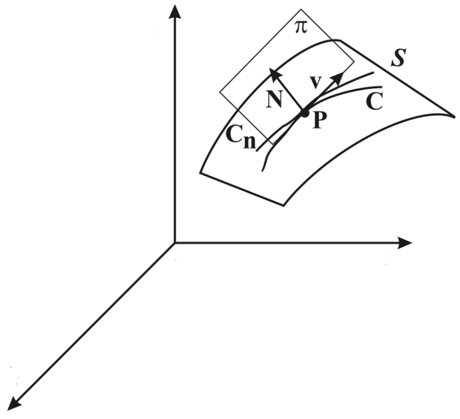


Figure 3.4: Normal Section. Figure from Ioan, (2010)

a normal curvature at a point, say $P = (\gamma_1^P, ..., \gamma_q^P, LD(\gamma^P))$, on the influence graph, consider a straight line in the range of $\gamma = (\gamma_1, ..., \gamma_q)$, i.e. Ω, passing through the point $\gamma^P = (\gamma_1^P, ..., \gamma_q^P)$. The straight line is represented by $\gamma^P + ah$, where a is a constant magnitude of the line and $h \in R^q$ is a vector of unit length, which can be thought of as the direction of the line. The straight line $\gamma^P + ah$ generated a lifted line on the influence graph S passing through the point P. The lifted line has a tangent vector at the point on the surface and hence corresponds to a normal plane, a normal section and a normal curvature. Following Cook (1986), the normal curvature κ_h of the normal section from the lifted line in the direction h can be obtained by

$$\kappa_h = \frac{\left| h \frac{\partial^2 LD(\gamma)}{\partial \gamma \partial \gamma^T} h^T \right|}{\left\{ 1 + \left| \frac{\partial LD(\gamma)}{\partial \gamma} \right|^2 \right\}^{\frac{1}{2}} h \left[I + \frac{\partial LD(\gamma)}{\partial \gamma} \left\{ \frac{\partial LD(\gamma)}{\partial \gamma} \right\}^T \right] h^T}.$$

Since $LD(\gamma) = 2 \left\{ l(\hat{\theta}^*) - l(\hat{\theta}|\gamma) \right\}$ and $l(\hat{\theta}^*)$ does not contain γ, the normal curvature κ_h associated with the direction h can be written as

$$\kappa_h = \frac{2 \left| h \frac{\partial^2 l(\hat{\theta}|\gamma)}{\partial \gamma \partial \gamma^T} h^T \right|}{\left\{ 1 + 4 \left| \frac{\partial l(\hat{\theta}|\gamma)}{\partial \gamma} \right|^2 \right\}^{\frac{1}{2}} h \left[I + 4 \frac{\partial l(\hat{\theta}|\gamma)}{\partial \gamma} \left\{ \frac{\partial l(\hat{\theta}|\gamma)}{\partial \gamma} \right\}^T \right] h^T}. \tag{3.2.19}$$

At a point P, the curvature is evaluated at $\gamma^P = (\gamma_1^P, ..., \gamma_q^P)$ and the corresponding $\hat{\theta}$.

It can be understood that there are infinitely many directions h range over all unit vectors in R^q and therefore at a point on the surface, there are infinitely many curvatures associated with the infinitely many directions. To study the characterization of the normal curvatures along different directions, in this thesis, we follow Steen et al., (2001) and use the sum of curvatures along the q directions of the modified weights $\gamma_1, ..., \gamma_q$.

3.3 Influence Measure using curvature

In Section 3.1, likelihood displacement was introduced to explain the concept of comparing two models using an influence graph (Figure 3.2) in two dimensional space $\mathbb{R}^2$. More specifically, likelihood displacement compares the difference in the likelihoods between an assumed model and a modified model. This is shown in Table 3.1 where the assumed model (at $\gamma = 0$) is compared with a modified model (when $\gamma \neq 0$).

However, in the following sections, we will focus on influence measures derived from the curvature of likelihood displacement rather than the likelihood displacement itself. This is because for $q > 1$, the influence graph is a surface (or a q-dimensional object in a $\mathbb{R}^{q+1}$ space), where influence analysis based on the plane curve of likelihood displacement (Figure 3.2) is not applicable.

It should be noted that using curvature as an influence measure leads to different interpretations. First, as shown in (3.2.19), the calculation of curvature depends on the first and second derivatives of the likelihood displacement with respect to γ. As a result, the curvature only relates to the modified model and is independent of the choice of assumed model. Second, each point on the influence surface (or the q-dimensional object) represents a distinct assumed or modified model. The curvature measures how curvy the surface is at a point (or a distinct model). Therefore, when using curvature as an influence measure, it can be used to interpret how stable or sensitive a model is with respect to slight changes in $(\gamma_1, \cdots, \gamma_q)$ (or say the slight change in the model assumption of measurement error), while the likelihood displacement compares the difference between an assumed model and a modified model. This is the difference when using curvature rather than the likelihood displacement as an influence measure.

3.3.1 One modification weight

We first consider the case when a scalar of modification weight γ is incorporated in an assumed model. The influence graph is a plane curve parameterised by γ for a given function $\vec{r}(\gamma) = (\gamma, LD(\gamma))$ in two dimensional $\mathbb{R}^2$. It follows that the tangent vector of $\vec{r}'(\gamma)$ is

$$\vec{r}'(\gamma) = \left(1, \frac{\partial LD(\gamma)}{\partial \gamma}\right).$$

And the corresponding second order derivative with respect to γ is

$$\vec{r}''(\gamma) = \left(0, \frac{\partial^2 LD(\gamma)}{\partial \gamma^2}\right).$$

The magnitude (commonly referred to as norm) of $\vec{r}'(\gamma)$ is

$$||\vec{r}'(\gamma)|| = \left(1 + \left(\frac{\partial LD(\gamma)}{\partial \gamma}\right)^2\right)^{\frac{1}{2}}.$$

Therefore, from (3.2.18), the curvature κ at any given point on the curve $r(\gamma)$ can be expressed as

$$\kappa = \frac{\left|\left(1, \frac{\partial LD(\gamma)}{\partial \gamma}\right) \times \left(0, \frac{\partial^2 LD(\gamma)}{\partial \gamma^2}\right)\right|}{\left(1 + \left(\frac{\partial LD(\gamma)}{\partial \gamma}\right)^2\right)^{\frac{3}{2}}} = \frac{\left|\frac{\partial^2 LD(\gamma)}{\partial \gamma^2}\right|}{\left(1 + \left(\frac{\partial LD(\gamma)}{\partial \gamma}\right)^2\right)^{\frac{3}{2}}}.$$

Since $LD(\gamma) = 2(\ell(\hat{\theta}^*) - \ell(\hat{\theta}|\gamma))$, and $\ell(\hat{\theta}^*)$ does not contain γ, the curvature κ at any given point on the curve $r(\gamma)$ can be written as

$$\kappa = \frac{2\left|\frac{\partial^2 \ell(\hat{\theta}|\gamma)}{\partial \gamma^2}\right|}{\left(1 + 4\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma}\right)^2\right)^{\frac{3}{2}}} \tag{3.3.1}$$

Later in Section 4, we will use this formula (3.3.1) of curvature to study how an assumed or modified logistic regression model is stable under measurement error.

3.3.2 More than one modification weights

We now consider the case when a q dimensional modification vector $\gamma = (\gamma_1, \cdots, \gamma_q)$ is introduced in an assumed model. In this case, the influence graph is no longer a plane curve, but rather a surface (or a q-dimensional object in a $q+1$ space). In otherwords, when $q > 1$, Figure 3.2 is not applicable. In this case, there are infinitely many curvatures from the infinitely many curves passing through a given point on the surface as explained in Section 3.2. Hence, a new measure or quantity to summarise the information from all the curvatures is needed to access the stability of a model.

Let's first consider two modification weights, say $\gamma = (\gamma_0, \gamma_1)$, are introduced in an assumed model. Let γ_0, γ_1 and $LD(\gamma_0, \gamma_1)$ be the three axes of a $\mathbb{R}^3$ space to embed the 2-dimensional influence surface $S = (\gamma_0, \gamma_1, LD(\gamma_0, \gamma_1))$. The curvatures along the directions of γ_0 and γ_1 can be derived as follows. Let the unit vectors along the directions of γ_0 and γ_1 respectively be $h_{\gamma_0} = [1, 0]$ and $h_{\gamma_1} = [0, 1]$. From (3.2.19), the curvature along the direction of γ_0 is

$$\begin{aligned}
\kappa_{h_{\gamma_0}} &= \frac{2\left|\begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}\frac{\partial^2 \ell(\hat{\theta}|\gamma)}{\partial \gamma_0^2} & \frac{\partial^2 \ell(\hat{\theta}|\gamma)}{\partial \gamma_0 \partial \gamma_1} \\ \frac{\partial^2 \ell(\hat{\theta}|\gamma)}{\partial \gamma_1 \partial \gamma_0} & \frac{\partial^2 \ell(\hat{\theta}|\gamma)}{\partial \gamma_1^2}\end{bmatrix}\begin{bmatrix}1 \\ 0\end{bmatrix}\right|}{\left[1 + 4\left\{\sqrt{\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_0}\right)^2 + \left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_1}\right)^2}\right\}^2\right]^{\frac{1}{2}}\begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}1 + 4\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_0}\right)^2 & 4\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_0}\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_1} \\ 4\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_1}\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_0} & 1 + 4\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_1}\right)^2\end{bmatrix}\begin{bmatrix}1 \\ 0\end{bmatrix}} \\
&= \frac{2\left|\frac{\partial^2 \ell(\hat{\theta}|\gamma)}{\partial \gamma_0^2}\right|}{\sqrt{1 + 4\sum_{j=0}^{1}\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_j}\right)^2}\left\{1 + 4\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_0}\right)^2\right\}} \qquad (3.3.2)
\end{aligned}$$

Similarly, the curvature along the direction of γ_1 is

$$
\begin{aligned}
\kappa_{h_{\gamma_1}} &= \frac{2\left|\begin{bmatrix}0 & 1\end{bmatrix}\begin{bmatrix}\frac{\partial^2 \ell(\hat{\theta}|\gamma)}{\partial \gamma_0^2} & \frac{\partial^2 \ell(\hat{\theta}|\gamma)}{\partial \gamma_0 \partial \gamma_1} \\ \frac{\partial^2 \ell(\hat{\theta}|\gamma)}{\partial \gamma_1 \partial \gamma_0} & \frac{\partial^2 \ell(\hat{\theta}|\gamma)}{\partial \gamma_1^2}\end{bmatrix}\begin{bmatrix}0 \\ 1\end{bmatrix}\right|}{\left[1+4\left\{\sqrt{\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_0}\right)^2+\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_1}\right)^2}\right\}^2\right]^{\frac{1}{2}}\begin{bmatrix}0 & 1\end{bmatrix}\begin{bmatrix}1+4\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_0}\right)^2 & 4\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_0}\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_1} \\ 4\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_1}\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_0} & 1+4\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_1}\right)^2\end{bmatrix}\begin{bmatrix}0 \\ 1\end{bmatrix}} \\
&= \frac{2\left|\frac{\partial^2 \ell(\hat{\theta}|\gamma)}{\partial \gamma_1^2}\right|}{\sqrt{1+4\sum_{j=0}^{1}\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_j}\right)^2}\left\{1+4\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_1}\right)^2\right\}} \qquad (3.3.3)
\end{aligned}
$$

The sum of normal curvatures along the directions γ_0 and γ_1 is

$$
\begin{aligned}
\sum_{j=0}^{1} \kappa_{h_{\gamma_j}} &= \kappa_{h\gamma_0} + \kappa_{h_{\gamma_1}} \\
&= \frac{2\left|\frac{\partial^2 \ell(\hat{\theta}|\gamma)}{\partial \gamma_0^2}\right|}{\sqrt{1+4\sum_{j=0}^{1}\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_j}\right)^2}\left\{1+4\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_0}\right)^2\right\}} \\
&+ \frac{2\left|\frac{\partial^2 \ell(\hat{\theta}|\gamma)}{\partial \gamma_1^2}\right|}{\sqrt{1+4\sum_{j=0}^{1}\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_j}\right)^2}\left\{1+4\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_1}\right)^2\right\}} \\
&= \frac{2\left|\frac{\partial^2 \ell(\hat{\theta}|\gamma)}{\partial \gamma_0^2}\right| / \left\{1+4\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_0}\right)^2\right\}}{\sqrt{1+4\sum_{j=0}^{1}\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_j}\right)^2}} + \frac{2\left|\frac{\partial^2 \ell(\hat{\theta}|\gamma)}{\partial \gamma_1^2}\right| / \left\{1+4\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_1}\right)^2\right\}}{\sqrt{1+4\sum_{j=0}^{1}\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_j}\right)^2}} \\
&= \frac{2\left[\left|\frac{\partial^2 \ell(\hat{\theta}|\gamma)}{\partial \gamma_0^2}\right| / \left\{1+4\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_0}\right)^2\right\} + \left|\frac{\partial^2 \ell(\hat{\theta}|\gamma)}{\partial \gamma_1^2}\right| / \left\{1+4\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_1}\right)^2\right\}\right]}{\sqrt{1+4\sum_{j=0}^{1}\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_j}\right)^2}} \\
&= \frac{2\sum_{j=0}^{1}\left\{\frac{\left|\frac{\partial^2 \ell(\hat{\theta}|\gamma)}{\partial \gamma_j^2}\right|}{1+4\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_j}\right)^2}\right\}}{\sqrt{1+4\sum_{j=0}^{1}\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_j}\right)^2}} \qquad (3.3.4)
\end{aligned}
$$

For a q dimensional modification vector $\gamma = (\gamma_1, \cdots, \gamma_q)$, it can be shown that the sum of curvatures along the directions $\gamma_1, \cdots, \gamma_q$ is

$$\sum_{j=1}^{q} \kappa_{h_{\gamma_j}} = \frac{2\sum_{j=1}^{q}\left\{\frac{\left|\frac{\partial^2 \ell(\hat{\theta}|\gamma)}{\partial \gamma_j^2}\right|}{1+4\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_j}\right)^2}\right\}}{\sqrt{1+4\sum_{j=1}^{q}\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_j}\right)^2}} \tag{3.3.5}$$

Later in Section 3.4, we will show how to use the sums of curvatures, (3.3.3) and (3.3.4), to study how stable an assumed or modified model is under the impact of a potential differential measurement error in binary response variable.

3.4 Stability under Misclassification of absence of an attribute

Our main interest is to use the influence measure shown in Section 3.2, to investigate the stability of logistic regression model fitted to an observed study, in which the observed outcome variable is under the problem of measurement error. An observed binary response variable Y_i^* has two attributes; with $y_i^* = 0$ as the absence or failure of an attribute and $y_i^* = 1$ otherwise. There are three possible scenarios where an observed binary response variable may be subject to measurement error. These are when;

- Case 1: Both attributes of response variable (that is, $y_i^* = 1$ and $y_i^* = 0$) are subject to measurement error.
- Case 2: Only the absence of an attribute; $y_i^* = 0$ may potentially be error-prone. The mechanism of this case is shown in Figure 3.1.
- Case 3: The presence of an attribute; $y_i^* = 1$ may be subject to measurement error.

In this section, we use the influence measure first to assess how stable a logistic model is under Case 2. For instance, an example under this case is our motivating example highlighted in Section 3.1 above. As stated in Section 3.1, inferences from assumed model (3.1.1) may be subject to potential bias, since measurement error is not considered. Thus, to assess the impact of measurement error, assume model is modified. Here, we considered modification based on nondifferential and differential misclassification process:

- Nondifferential Misclassification: Under Case 2, a nondifferential misclassification process assume the probability of measurement error γ does not depend on the covariate variable X_i.
- Differential Misclassification: By assuming that the probability of measurement error γ depends on true exposure variable X_i, we say the observed study is under differential misclassification pattern

This section is structured as follows. In Section 3.4.1, we derived influence measure under nondifferential measurement error pattern. In Section 3.4.2, we derived influence measure under a special case of differential misclassification process based on a binary covariate variable. More generally, in Section 3.4.3, the approach is extended to an influence measure under a general framework. Note that the model stability investigated below is not only applicable to our motivating example, but a general method of assessing how robust a logistic model is when observed study has an error-prone binary response variable $Y_i^* = 0$.

3.4.1 NonDifferential Misclassification

Here, we considered the case of when the misclassification pattern is nondifferential. To account for nondifferential measurement error, the probability of measurement error γ is incorporated in an assumed model (3.1.1). The resulting model is a modified model (3.1.4) where γ is assume to be independent of X_i, thus γ is a scalar modification weight introduced in an assumed model. As a result, from (3.3.1), an influence measure based on curvature can be derived under nondifferential measurement error pattern. Hence, the graph of curvature κ versus γ is a plane curve which can be used to study assume model (that is, the curvature around $\gamma = 0$) under a slight modification to γ caused by the impact of potential nondifferential measurement error. Beyond this, the measure derived can be used to investigate other modified models (that is, when $\gamma \neq 0$).

From (3.1.5), $\ell(\theta|\gamma)$ can be rewritten as

$$\ell(\theta|\gamma) = \sum_{i=1}^{n} \ell_i, \tag{3.4.1}$$

where ℓ_i is the log-likelihood of the modified model for individual subjects from the observed study, defined as

$$\begin{aligned} \ell_i &= y_i^* \log(\pi_i^*) + (1 - y_i^*) \log(1 - \pi_i^*) \\ \pi_i^* &= \pi_i + \gamma(1 - \pi_i) \end{aligned}$$

such that,

$$1 - \pi_i = \left(\frac{1}{1-\gamma}\right)(1 - \pi_i^*). \tag{3.4.2}$$

Next, is the derivative of ℓ_i with respect to γ

1. The first derivative of ℓ_i is given as

$$\frac{\partial \ell_i}{\partial \gamma} = \frac{\partial \ell_i}{\partial \pi^*} \frac{\partial \pi_i^*}{\partial \gamma} = \left(\frac{y_i^*}{\pi_i^*} - \frac{1 - y_i^*}{1 - \pi_i^*}\right)(1 - \pi_i)$$

It follows that from (3.4.1),

$$\frac{\partial}{\partial\gamma}\ell(\theta|\gamma) = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \pi_i^*}\frac{\partial \pi_i^*}{\partial \gamma} = \sum_{i=1}^{n}\left(\frac{y_i^*}{\pi_i^*} - \frac{1-y_i^*}{1-\pi_i^*}\right)(1-\pi_i)$$

Using (3.4.2),

$$\frac{\partial}{\partial\gamma}\ell(\theta|\gamma) = \sum_{i=1}^{n}\left(\frac{1}{1-\gamma}\right)(1-\pi_i^*)\left(\frac{y_i^*}{\pi_i^*} - \frac{1-y_i^*}{(1-\hat{\pi}_i^*)}\right) \tag{3.4.3}$$

2. The second derivative

$$\frac{\partial^2 \ell_i}{\partial\gamma^2} = -(1-\pi_i)^2\left(\frac{y_i^*}{\pi_i^{*2}} + \frac{1-y_i^*}{(1-\pi_i^*)^2}\right)$$

similarly, it follows that

$$\frac{\partial^2 \ell(\theta|\gamma)}{\partial\gamma^2} = \sum_{i=1}^{n}(1-\pi_i)^2\left(\frac{y_i^*}{\pi_i^{*2}} + \frac{1-y_i^*}{(1-\pi_i^*)^2}\right)$$

and using (3.4.2),

$$\frac{\partial^2 \ell(\theta|\gamma)}{\partial\gamma^2} = \sum_{i=1}^{n}\left(\frac{1}{1-\gamma}\right)^2(1-\pi_i^*)^2\left(\frac{y_i^*}{\pi_i^{*2}} + \frac{1-y_i^*}{(1-\pi_i^*)^2}\right) \tag{3.4.4}$$

Using (3.4.3) and (3.4.4), curvature κ in (3.3.1) is

$$\begin{aligned}\kappa &= \frac{\left|2\sum_{i=1}^{n}\left(\frac{1}{1-\gamma}\right)^2(1-\hat{\pi}_i^*)^2\left(\frac{y_i^*}{\hat{\pi}_i^{*2}} + \frac{1-y_i^*}{(1-\hat{\pi}_i^*)^2}\right)\right|}{\left(1+4\left(\sum_{i=1}^{n}(\frac{1}{1-\gamma})(1-\hat{\pi}_i^*)\left(\frac{y_i^*}{\hat{\pi}_i^*} - \frac{1-y_i^*}{(1-\hat{\pi}_i^*)}\right)\right)^2\right)^{\frac{3}{2}}} \\ &= \frac{2(\frac{1}{1-\gamma})^2\sum_{i=1}^{n}(1-\hat{\pi}_i^*)^2\left(\frac{y_i^*}{\hat{\pi}_i^{*2}} + \frac{1-y_i^*}{(1-\hat{\pi}_i^*)^2}\right)}{\left(1+(\frac{1}{1-\gamma})^2 4\left(\sum_{i=1}^{n}(1-\hat{\pi}_i^*)\left(\frac{y_i}{\hat{\pi}_i^*} - \frac{1-y_i^*}{(1-\hat{\pi}_i^*)}\right)\right)^2\right)^{\frac{3}{2}}}\end{aligned} \tag{3.4.5}$$

This curvature can be used to assess the stability of each distinct model (assumed and modified models) under a slight change in γ. The larger a curvature is, the faster the curve is turning, which means the model with large curvature is sensitive to measurement error. A plot of curvature against γ for a random simulated data shows that the assumed model has the lowest curvature among all the models (as shown in Figure 3.5 and Table 3.2).

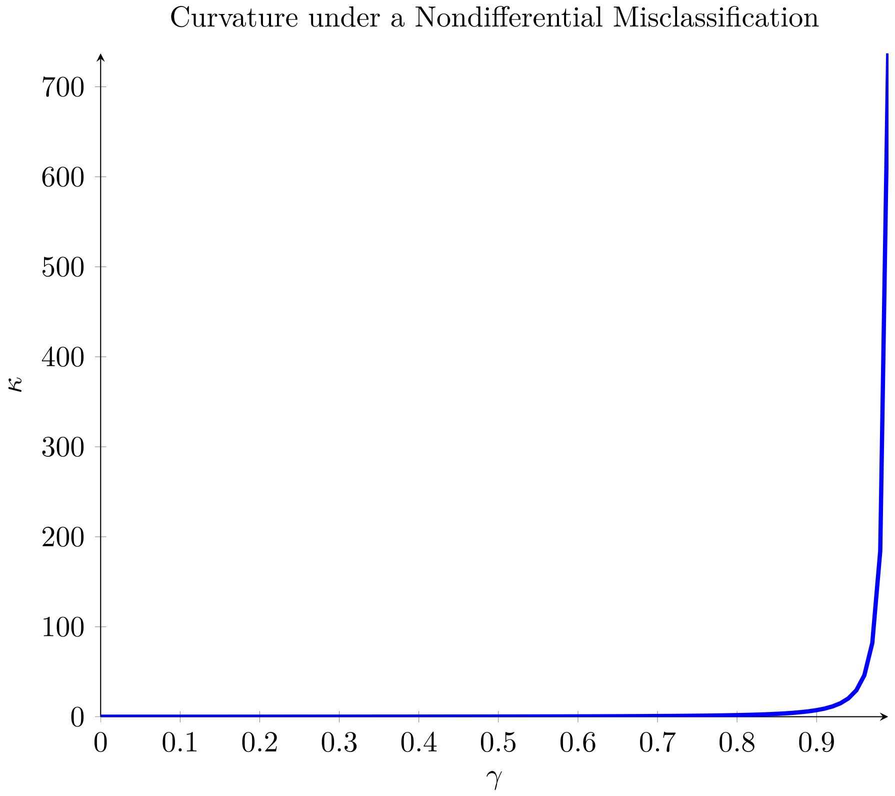


Figure 3.5: Influence Measure based on Curvature results from a simulated data under a Nondifferential Misclassification with varying γ ranges from 0 to 1

Table 3.2: Extract of curvature results from a simulated data under a Nondifferential Misclassification mimicking motivating example (see detail in Section 3.1) with γ ranges from 0 to 1. Each γ represents a distinct model with assumed model at $\gamma = 0$, and modified models when $\gamma \neq 0$

γ	Curvature
0	737.13
0.14	996.66
0.39	1981.01
0.55	3640.16
0.63	5384.46
0.67	6768.89
0.71	8764.95
0.75	11794.12
0.82	22751.00
0.87	43617.30
0.94	204759.01
0.99	7371324.35

According to the random simulation result, it indicates that the assumed mode with $\gamma = 0$

is the most stable model compared with any other modified models with $\gamma \neq 0$.

We can also prove this result mathematically. Let

$$\begin{aligned} A &= 2\sum_{i=1}^{n}(1-\hat{\pi}_i^*)^2\left(\frac{y_i^*}{\hat{\pi}_i^{*2}}+\frac{1-y_i^*}{(1-\hat{\pi}_i^*)^2}\right) \\ B &= 4\left[\sum_{i=1}^{n}(1-\hat{\pi}_i^*)\left(\frac{y_i^*}{\hat{\pi}_i^*}-\frac{1-y_i^*}{(1-\hat{\pi}_i^*)}\right)\right]^2 \\ D(\gamma) &= \left(\frac{1}{1-\gamma}\right)^2, \end{aligned}$$

such that curvature κ in (3.4.5) can be rewritten as

$$\kappa = \frac{D(\gamma)A}{[1+D(\gamma)B]^{\frac{3}{2}}} = AD(\gamma)\left[1+BD(\gamma)\right]^{-\frac{3}{2}}. \tag{3.4.6}$$

The derivative of κ in (3.4.6) with respect to γ is

$$\begin{aligned} \frac{\partial\kappa}{\partial\gamma} &= A\left[\frac{\partial D(\gamma)}{\partial\gamma}(1+BD(\gamma))^{-\frac{3}{2}}+D(\gamma)\left(-\frac{3}{2}\right)(1+D(\gamma)B)^{-\frac{5}{2}}B\frac{\partial D(\gamma)}{\partial\gamma}\right] \\ &= A\frac{\partial D(\gamma)}{\partial\gamma}\left\{\frac{[1+D(\gamma)B]^2-\frac{3}{2}D(\gamma)B}{[1+D(\gamma)B]^{\frac{5}{2}}}\right\} \\ &= A\frac{\partial D(\gamma)}{\partial\gamma}\left\{\frac{1+[D(\gamma)B]^2+2D(\gamma)B-\frac{3}{2}D(\gamma)B}{[1+D(\gamma)B]^{\frac{5}{2}}}\right\} \\ &= A\frac{\partial D(\gamma)}{\partial\gamma}\left\{\frac{1+[D(\gamma)B]^2+\frac{1}{2}D(\gamma)B}{[1+D(\gamma)B]^{\frac{5}{2}}}\right\} \end{aligned}$$

$$A = \begin{cases} 2\sum_{i=1}^{n}\left(\frac{1-\hat{\pi}_i^*}{\hat{\pi}_i^*}\right)^2, & \text{when } y_i^*=1 \\ 2\sum_{i=1}^{n}\left(\frac{1-\hat{\pi}_i^*}{1-\hat{\pi}_i^*}\right)^2 = 2n, & \text{when } y_i^*=0 \end{cases}$$

And,

$$B = \begin{cases} 4\left(\sum_{i=1}^{n}\left(\frac{1-\hat{\pi}_i^*}{\hat{\pi}_i^*}\right)\right)^2, & \text{when } y_i^*=1 \\ 4\left(\sum_{i=1}^{n}\left(\frac{1-\hat{\pi}_i^*}{1-\hat{\pi}_i^*}\right)\right)^2 = 4n^2, & \text{when } y_i^*=0 \end{cases}$$

Since $0 \leq \gamma \leq 1$, $A \geq 0$. Similarly, $B \geq 0$. Also $D(\gamma) = \left(\frac{1}{1-\gamma}\right)^2 \geq 0$ and

$$\frac{\partial D(\gamma)}{\partial\gamma} = \frac{2}{(1-\gamma)^3} \geq 0.$$

Therefore, $\frac{\partial\kappa}{\partial\gamma} \geq 0$ for $0 \leq \gamma \leq 1$. It means κ is a monotonically increasing function of γ, for γ between 0 and 1 (like it shows in Figure 3.5). Hence, κ is minimized at $\gamma = 0$.

This proves that an assumed model (at $\gamma = 0$) under a nondifferential misclassification is the most stable model under the problem of error prone Y_i^*. In otherwords, the assumed model with $\gamma = 0$ is the most stable model compared with other modified models with $\gamma \neq 0$. Although, Figure 3.5 and Table 3.2 was based on a simulated random data, our proof here shows that κ is minimize at $\gamma = 0$ for any random data under the problem of mismeasured response $Y_i^* = 0$.

3.4.2 Differential Misclassification based on binary covariate

In this section we consider a special case of differential misclassification. The assumption here is that the probability of measurement error is dependent on the error-free measurement of binary exposure variable x_i.

Suppose the modification weights introduced are $\gamma = (\gamma_0, \gamma_1)$ and the probability of measurement error is

$$\begin{aligned} P(Y_i^* = 1|Y_i = 0, x_i) =& \gamma_0(1 - x_i) + \gamma_1 x_i \\ =& \gamma_0 - \gamma_0 x_i + \gamma_1 x_i \end{aligned} \tag{3.4.11}$$

such that the probability depends on the binary exposure variable x_i, i.e.

$$P(Y_i^* = 1|Y_i = 0, x_i) = \begin{cases} \gamma_0, & \text{when } x_i = 0 \\ \gamma_1, & \text{when } x_i = 1 \end{cases}$$

By assuming (3.4.11),

$$\begin{aligned} P(Y_i^*|x_i) &= P(Y_i^* = 1|Y_i = 1, x_i)P(Y_i = 1|x_i) + P(Y_i^* = 1|Y_i = 0, x_i)P(Y_i = 0|x_i) \\ &= \pi_i + (\gamma_0 - \gamma_0 x_i + \gamma_1 x_i)(1 - \pi_i). \end{aligned} \tag{3.4.12}$$

The distribution of the observed outcome variable conditional on γ_0 and γ_1 is hence

$$Y_i^*|X_i \sim Bin(1, \pi_i + (\gamma_0 - \gamma_0 x_i + \gamma_1 x_i)(1 - \pi_i)), \tag{3.4.13}$$

and the log-likelihood for the modified model with the weights $\gamma = (\gamma_0, \gamma_1)$ is

$$\begin{aligned} \ell(\theta|\gamma) = \sum_{i=1}^{n} & y_i^* \log\Big(\pi_i + (\gamma_0 - \gamma_0 x_i + \gamma_1 x_i)(1 - \pi_i)\Big) \\ & + (1 - y_i^*) \log\Big((1 - \gamma_0 + \gamma_0 x_i - \gamma_1 x_i)(1 - \pi_i)\Big) \end{aligned} \tag{3.4.14}$$

When we assume $(\gamma_0, \gamma_1) = (0, 0)$, there is no measurement error, (3.4.12) is equal to the assumed model (3.1.1), $\ell(\theta^*) = \ell(\theta|\gamma)$ and $\theta^* = \theta$. When we assume $(\gamma_0, \gamma_1) \neq (0, 0)$, the modified model is deviated from the assumed model and $\ell(\theta^*) \neq \ell(\theta|\gamma)$. The modification weights $\gamma = (\gamma_0, \gamma_1)$ are bounded by $\Omega = [0, 1]^2$. The assumed model corresponds to the point at $\gamma = (0, 0)$, and all the other points in Ω correspond to the infinitely many distinct modified models.

In this special case, the sum of normal curvatures (3.3.4) along the directions of γ_0 and γ_1 can be used to assess how stable a model (assumed or modified) is under the impact of differential misclassification as specified in (3.4.11). Figure 3.6 shows the influence surface $S = (\gamma_0, \gamma_1, LD(\gamma))$ of the sum of normal curvatures (3.3.4) against γ_0 and γ_1 respectively.

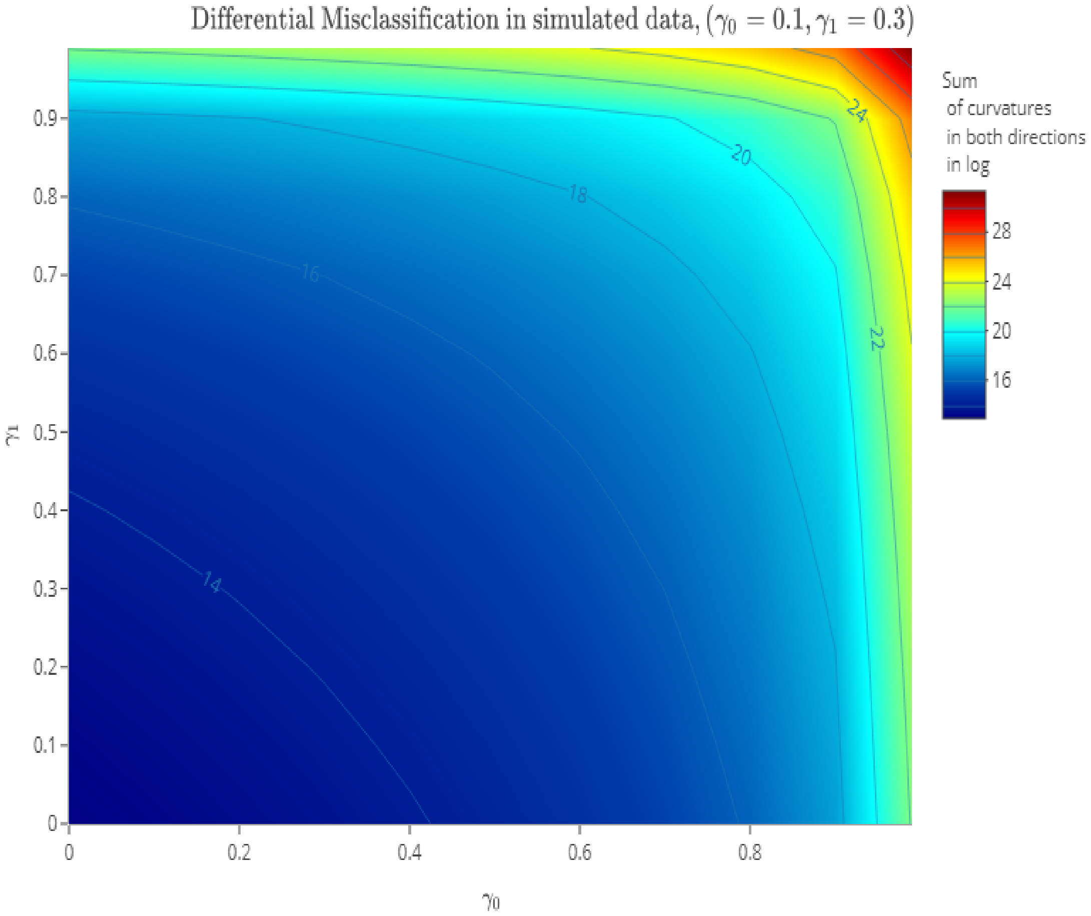


Figure 3.6: Influence Surface under a Differential Misclassification pattern. Sum of curvatures for both directions of γ_0 and γ_1 from a simulation study with $\gamma_0 = 0.1$ and $\gamma_1 = 0.3$ under the impact of a Differential Misclassification pattern

Similar to Figure 3.5, Figure 3.6 shows that the assumed model has the lowest sum of curvatures among all the models. This indicates that for the special case of Section 3.2, when there is no information about the error size, the assumed model is still the most stable model under the potential impact of differential misclassification. Again, we can prove this rigorously.

Proof. We will use convergence in probability to prove that the assumed model is asymptotically the most stable model under the special case of Section 3.2. The proof requires the use of average log-likelihood function, therefore, we need to re-write (3.4.13) as

$$\ell(\hat{\theta}|\gamma) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(\hat{\theta}|\gamma) \tag{3.4.15}$$

where $\ell_i(\hat{\theta}|\gamma)$ is the likelihood function of individual subjects under the probability of measurement error (γ_0, γ_1) defined as

$$\begin{aligned} \ell_i(\hat{\theta}|\gamma) =& y_i^* \log\Bigg(\hat{\pi}_i + (\gamma_0 - \gamma_0 x_i + \gamma_1 x_i)(1-\hat{\pi}_i)\Bigg) \\ &+ (1-y_i^*)\log\Bigg(\big(1-\hat{\pi}_i\big)\big(1-\gamma_0 + \gamma_0 x_i - \gamma_1 x_i\big)\Bigg) \end{aligned} \tag{3.4.16}$$

Note that using the average log-likelihood does not change the maximum likelihood estimates.

For simplicity of expression, we can further rewrite (3.4.16) as

$$\ell_i(\hat{\theta}|\gamma) = y_i^* \log(\hat{\pi}_i^*) + (1-y_i^*)\log(1-\hat{\pi}_i^*) \tag{3.4.17}$$

and now with assumed (γ_0, γ_1), $\hat{\pi}_i^*$ can also be written as

$$\hat{\pi}_i^* = \hat{\pi}_i + (\gamma_0 - \gamma_0 x_i + \gamma_1 x_i)(1-\hat{\pi}_i). \tag{3.4.18}$$

The reason we can re-write $\hat{\pi}_i^*$ as (3.4.18) is because from (3.4.12)

$$\hat{\pi}_i + (\gamma_0 - \gamma_0 x_i + \gamma_1 x_i)(1-\hat{\pi}_i) = \hat{P}(Y_i^*|x_i) = \hat{\pi}_i^*$$

It should be noted that the estimated $\hat{\alpha}^*$, $\hat{\theta}^*$ and

$$\hat{\pi}_i^* = \frac{\exp(\hat{\alpha}^* + \hat{\theta}^* x_i)}{1+\exp(\hat{\alpha}^* + \hat{\theta}^* x_i)}$$

are based on the observed y_i^* and x_i and therefore do not vary with different assumptions of γ_0 and γ_1. Thus in (3.4.18), the value of $\hat{\pi}_i^*$ is constant when y_i^* and x_i are observed, but the value of $\hat{\pi}_i$ can change with different values of (γ_0, γ_1).

Due to the consistency of MLEs,

$$\begin{gathered} \hat{\alpha}^* \xrightarrow{P} \alpha^* \\ \hat{\theta}^* \xrightarrow{P} \theta^* \end{gathered}$$

and using Slutsky's and continuous mapping theorems, we have

$$\hat{\pi}_i^* = \frac{\exp(\hat{\alpha}^* + \hat{\theta}^* x_i)}{1 + \exp(\hat{\alpha}^* + \hat{\theta}^* x_i)} \overset{P}{\to} \frac{\exp(\alpha^* + \theta^* x_i)}{1 + \exp(\alpha^* + \theta^* x_i)} = \pi^* \tag{3.4.19}$$

To obtain the asymptotic property of the sum of curvatures (3.3.4), we first need to use (3.4.19) to derive the limits of the first two derivatives of $l(\hat{\theta}|\gamma)$ with respect to γ_0 and γ_1. Again using Slutsky's and continuous mapping theorems, it follows

$$\begin{aligned}
\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_0} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \hat{\pi}_i^*}\frac{\partial \hat{\pi}_i^*}{\partial \gamma_0} = \frac{1}{n}\sum_{i=1}^{n} \frac{1-\hat{\pi}_i^*}{1-\gamma_0+\gamma_0 x_i - \gamma_1 x_i}\left(\frac{y_i^*}{\hat{\pi}_i^*} - \frac{1-y_i^*}{(1-\hat{\pi}_i^*)}\right)(1-x_i) \\
&\overset{P}{\to} E\left[\frac{1-\pi_i^*}{1-\gamma_0+\gamma_0 X_i - \gamma_1 X_i}\left(\frac{Y_i^*}{\pi_i^*} - \frac{1-Y_i^*}{(1-\pi_i^*)}\right)(1-X_i)\right] \\
&= E\left[\frac{1-\pi_i^*}{1-\gamma_0+\gamma_0 X_i - \gamma_1 X_i}\left(\frac{E(Y_i^*|X_i)}{\pi_i^*} - \frac{1-E(Y_i^*|X_i)}{(1-\pi_i^*)}\right)(1-X_i)\right] \\
&= E\left[\frac{1-\pi_i^*}{1-\gamma_0+\gamma_0 X_i - \gamma_1 X_i}\left(\frac{\pi_i^*}{\pi_i^*} - \frac{1-\pi_i^*}{(1-\pi_i^*)}\right)(1-X_i)\right] \\
&= 0
\end{aligned} \tag{3.4.20}$$

and

$$\begin{aligned}
\frac{\partial^2 \ell(\hat{\theta}|\gamma)}{\partial \gamma_0^2} &= -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1-\hat{\pi}_i^*}{1-\gamma_0+\gamma_0 x_i - \gamma_1 x_i}\right)^2\left(\frac{y_i^*}{\hat{\pi}_i^{*2}} + \frac{1-y_i^*}{(1-\hat{\pi}_i^*)^2}\right)(1-x_i)^2 \\
&\overset{P}{\to} -E\left[\left(\frac{1-\pi_i^*}{1-\gamma_0+\gamma_0 X_i - \gamma_1 X_i}\right)^2\left(\frac{Y_i^*}{\pi_i^{*2}} + \frac{1-Y_i^*}{(1-\pi_i^*)^2}\right)(1-X_i)^2\right] \\
&= -E\left[\left(\frac{1-\pi_i^*}{1-\gamma_0+\gamma_0 X_i - \gamma_1 X_i}\right)^2\left(\frac{E(Y_i^*|X_i)}{\pi_i^{*2}} + \frac{1-E(Y_i^*|X_i)}{(1-\pi_i^*)^2}\right)(1-X_i)^2\right] \\
&= -E\left[\left(\frac{1-\pi_i^*}{1-\gamma_0+\gamma_0 X_i - \gamma_1 X_i}\right)^2\left(\frac{\pi_i^*}{\pi_i^{*2}} + \frac{1-\pi_i^*}{(1-\pi_i^*)^2}\right)(1-X_i)^2\right] \\
&= -E\left[\left(\frac{1}{1-\gamma_0+\gamma_0 X_i - \gamma_1 X_i}\right)^2\left(\frac{1-\pi_i^*}{\pi_i^*}\right)(1-X_i)^2\right]
\end{aligned} \tag{3.4.21}$$

Similarly,

$$\begin{aligned}
\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_1} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \hat{\pi}_i^*}\frac{\partial \hat{\pi}_i^*}{\partial \gamma_1} = \frac{1}{n}\sum_{i=1}^{n} \frac{1-\hat{\pi}_i^*}{1-\gamma_0+\gamma_0 x_i-\gamma_1 x_i}\left(\frac{y_i^*}{\hat{\pi}_i^*}-\frac{1-y_i^*}{(1-\hat{\pi}_i^*)}\right)x_i \\
&\xrightarrow{P} E\left[\frac{1-\pi_i^*}{1-\gamma_0+\gamma_0 X_i-\gamma_1 X_i}\left(\frac{Y_i^*}{\pi_i^*}-\frac{1-Y_i^*}{(1-\pi_i^*)}\right)X_i\right] \\
&= E\left[\frac{1-\pi_i^*}{1-\gamma_0+\gamma_0 X_i-\gamma_1 X_i}\left(\frac{E(Y_i^*|X_i)}{\pi_i^*}-\frac{1-E(Y_i^*|X_i)}{(1-\pi_i^*)}\right)X_i\right] \\
&= E\left[\frac{1-\pi_i^*}{1-\gamma_0+\gamma_0 X_i-\gamma_1 X_i}\left(\frac{\pi_i^*}{\pi_i^*}-\frac{1-\pi_i^*}{(1-\pi_i^*)}\right)X_i\right] \\
&= 0 \qquad (3.4.22)
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial^2 \ell(\hat{\theta}|\gamma)}{\partial \gamma_1^2} &= -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1-\hat{\pi}_i^*}{1-\gamma_0+\gamma_0 x_i-\gamma_1 x_i}\right)^2\left(\frac{y_i^*}{\hat{\pi}_i^{*2}}+\frac{1-y_i^*}{(1-\hat{\pi}_i^*)^2}\right)x_i^2 \\
&\xrightarrow{P} -E\left[\left(\frac{1-\pi_i^*}{1-\gamma_0+\gamma_0 X_i-\gamma_1 X_i}\right)^2\left(\frac{Y_i^*}{\pi_i^{*2}}+\frac{1-Y_i^*}{(1-\pi_i^*)^2}\right)X_i^2\right] \\
&= -E\left[\left(\frac{1-\pi_i^*}{1-\gamma_0+\gamma_0 X_i-\gamma_1 X_i}\right)^2\left(\frac{E(Y_i^*|X_i)}{\pi_i^{*2}}+\frac{1-E(Y_i^*|X_i)}{(1-\pi_i^*)^2}\right)X_i^2\right] \\
&= -E\left[\left(\frac{1-\pi_i^*}{1-\gamma_0+\gamma_0 X_i-\gamma_1 X_i}\right)^2\left(\frac{\pi_i^*}{\pi_i^{*2}}+\frac{1-\pi_i^*}{(1-\pi_i^*)^2}\right)^2 X_i^2\right] \\
&= -E\left[\left(\frac{1}{1-\gamma_0+\gamma_0 X_i-\gamma_1 X_i}\right)^2\left(\frac{1-\pi_i^*}{\pi_i^*}\right)X_i^2\right] \qquad (3.4.23)
\end{aligned}$$

Using (3.4.20) and (3.4.22), we have in (3.3.4)

$$1+4\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_j}\right)^2 \xrightarrow{P} 1 \quad \text{for } j=0 \text{ and } 1 \qquad (3.4.24)$$

$$\sqrt{1+4\sum_{j=0}^{1}\left(\frac{\partial \ell(\hat{\theta}|\gamma)}{\partial \gamma_j}\right)^2} \xrightarrow{P} 1 \qquad (3.4.25)$$

With (3.4.21), (3.4.23), (3.4.24) and (3.4.25), the limit of the sum of curvatures (3.3.4) is

$$\begin{aligned}\sum_{j=0}^{1}\kappa_{h_{\gamma_j}} = \kappa_{h\gamma_0}+\kappa_{h_{\gamma_1}} &= \frac{2\sum_{j=0}^{1}\left\{\frac{\left|\frac{\partial^2\ell(\hat{\theta}|\gamma)}{\partial\gamma_j^2}\right|}{1+4\left(\frac{\partial\ell(\hat{\theta}|\gamma)}{\partial\gamma_j}\right)^2}\right\}}{\sqrt{1+4\sum_{j=0}^{1}\left(\frac{\partial\ell(\hat{\theta}|\gamma)}{\partial\gamma_j}\right)^2}} \\ &\xrightarrow{P} 2E\left[\left(\frac{1}{1-\gamma_0+\gamma_0X_i-\gamma_1X_i}\right)^2\left(\frac{1-\pi_i^*}{\pi_i^*}\right)(1-X_i)^2\right. \\ &\qquad\left.+\left(\frac{1}{1-\gamma_0+\gamma_0X_i-\gamma_1X_i}\right)^2\left(\frac{1-\pi_i^*}{\pi_i^*}\right)X_i^2\right] \\ &= 2E\left[\left(\frac{1}{1-\gamma_0+\gamma_0X_i-\gamma_1X_i}\right)^2\left(\frac{1-\pi_i^*}{\pi_i^*}\right)\left((1-X_i)^2+X_i^2\right)\right]. \qquad (3.4.26)\end{aligned}$$

When $X_i=0$, the limit (3.4.26) is

$$\sum_{j=0}^{1}\kappa_{h\gamma_j}\xrightarrow{P}2\left[\left(\frac{1}{1-\gamma_0}\right)^2\left(\frac{1-\frac{e^{\alpha^*}}{1+e^{\alpha^*}}}{\frac{e^{\alpha^*}}{1+e^{\alpha^*}}}\right)\right]=2\left[\left(\frac{1}{1-\gamma_0}\right)^2e^{-\alpha^*}\right], \qquad (3.4.27)$$

which is minimized at $\gamma_0=0$ for $0\le\gamma_0\le1$.

When $X_i=1$,

$$\sum_{j=0}^{1}\kappa_{h\gamma_j}\xrightarrow{P}2\left[\left(\frac{1}{1-\gamma_1}\right)^2\left(\frac{1-\frac{e^{\alpha^*+\theta^*}}{1+e^{\alpha^*+\theta^*}}}{\frac{e^{\alpha^*+\theta^*}}{1+e^{\alpha^*+\theta^*}}}\right)\right]=2\left[\left(\frac{1}{1-\gamma_1}\right)^2e^{-\alpha^*-\theta^*}\right], \qquad (3.4.28)$$

which is minimized at $\gamma_1=0$ for $0\le\gamma_1\le1$.

The limit of the sum of curvatures (3.4.26) is then

$$\sum_{j=0}^{1}\kappa_{h_{\gamma_j}}\xrightarrow{P}2e^{-\alpha^*}\left(\{1-P(X_i=1)\}\left[\left(\frac{1}{1-\gamma_0}\right)^2\right]+P(X_i=1)\left[\left(\frac{1}{1-\gamma_1}\right)^2e^{-\theta^*}\right]\right),$$

which is minimized at $(\gamma_0,\gamma_1)=(0,0)$ for $(\gamma_0,\gamma_1)\in[0,1]^2$. □

The proof above shows that the sum of curvatures, $\sum_{j=0}^{1}\kappa_{h_{\gamma_j}}$, is asymptotically minimized at $(\gamma_0,\gamma_1)=(0,0)$. This means that the assumed model with $(\gamma_0,\gamma_1)=(0,0)$ is asymptotically more stable than the modified models with $(\gamma_0,\gamma_1)\neq(0,0)$.

3.4.3 General case of Misclassification

In Section 3.4.1 and 3.4.2 respectively, we considered the nondifferential misclassification and a special case of differential misclassification which depends on a binary covariate.

In this section, we consider the setting below, which is general for both the nondifferential misclassification and the differential misclassification conditional on an arbitrary (continuous or/and discrete) vector of covariates, say $\mathbf{x_i} = (x_{i1}, x_{i2}, ..., x_{ip})^{tr}$.

To set up the general misclassification, let

$$\gamma_i = P(Y_i^* = 1 | Y_i = 0, x_i)$$

be the probability of measurement error for the ith subject conditional on $\mathbf{x_i} = (x_{i1}, x_{i2}, ..., x_{ip})^{tr}$. For a sample size of n, the probabilities of measurement error are $(\gamma_1, ..., \gamma_n)$. As the notation γ_i has been used for the probability of measurement error, to avoid confusion, we need now to introduce the new notations $\omega = (\omega_1, ..., \omega_q)$ as the modification weights, which influence the individual probability of measurement error γ_i through a function g, i.e.

$$P(Y_i^* = 1 | Y_i = 0, x_i) = \gamma_i = g(\omega; x_i) = g(\omega_1, ..., \omega_q; x_i) \qquad (3.4.29)$$

As shown below, the setting (3.4.29) is general to the case of nondifferential misclassification in Section 3.4.1 and the special case of differential misclassification in Section 3.4.2:

(a) In Section 3.4.1, all the individuals have the same probability of measurement error γ. To see how the general setting (3.4.29) can cover the case of nondifferential misclassification, consider $q = 1$ and the modification weight is $\omega = \omega_1$, which can influence the probability of measurement error by

$$\gamma_1 = ... = \gamma_n = \gamma = g(\omega; x_i) = g(\omega_1) = \omega_1.$$

In this way, the setting (3.4.29) is general to the case of nondifferential misclassification by letting $g(\omega_1) = \omega_1 = \gamma$. Here the modification weight $\omega = \omega_1$ is bound by $[0, 1] \in R$.

(b) In Section 3.4.2, misclassification is differential based on a binary x_i. In this special case, $q = 2$ and the modification weights $\omega = (\omega_1, \omega_2)$ influence the probability of measurement error through the function g:

$$\gamma_i = g(\omega; x_i) = g(\omega_1, \omega_2; x_i) = \omega_1(1 - x_i) + \omega_2 x_i \quad \text{for } i = 1, ..., n.$$

Under this function g, $\omega_1 = \gamma_0$, $\omega_2 = \gamma_1$ and the vector of weights $\omega = (\omega_1, \omega_2)$ is bound by $[0, 1]^2 \in R^2$.

Under the general setting of misclassification (3.4.29), the average log-likelihood function based on the assumed modification weights $\omega = (\omega_1, ..., \omega_q)$ is

$$l(\hat{\theta}|\omega) = \frac{1}{n}\sum_{i=1}^{n} l_i(\hat{\theta}|\omega) = \frac{1}{n}\sum_{i=1}^{n} y_i^* \log\{\hat{\pi}_i + \gamma_i(1-\hat{\pi}_i)\} + (1-y_i^*)\log\{1-\hat{\pi}_i-\gamma_i(1-\hat{\pi}_i)\}, \tag{3.4.30}$$

where

$$\hat{\pi}_i = \frac{\exp(\hat{\alpha}+\hat{\theta}x_i)}{1+\exp(\hat{\alpha}+\hat{\theta}x_i)}$$

and $\hat{\alpha}$ and $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_p)$ are estimated under the assumed $\omega = (\omega_1, ..., \omega_q)$.

Different assumed values of ω lead to different values of $\hat{\alpha}$, $\hat{\theta}$, $\hat{\pi}_i$ and the log-likelihood function (3.4.30). The different values of $(q+1)\times 1$ vector

$$\begin{pmatrix} \omega_1 \\ \vdots \\ \omega_q \\ LD(\omega) = 2[l(\hat{\theta}^*) - l(\hat{\theta}|\omega)] \end{pmatrix}$$

as ω varies throughout some $\Omega \in R^q$ form the influence surface. There is one point or a vector of weights in Ω, say

$$\omega^* = \{\omega : g(\omega; x_i) = \gamma_i = 0 \quad \text{for} \quad i = 1, ..., n\}$$

corresponds to the assumed model, $l(\hat{\theta}^*) = l(\hat{\theta}|\omega = \omega^*)$. All the other points that $\omega \neq \omega^*$ in Ω correspond to the infinitely many distinct modified models, $l(\hat{\theta}|\omega \neq \omega^*)$.

The sum of curvatures of the influence surface along the directions of $\omega = (\omega_1, ..., \omega_q)$ can be used to assess how stable a model is around each particular point in Ω. By substituting γ in (3.3.5) with ω, the sum of curvatures is

$$\sum_{j=1}^{q} \kappa_{h_{\omega_j}} = \sum_{j=1}^{q} \frac{2\left\{\frac{\left|\frac{\partial^2 \ell(\hat{\theta}|\omega)}{\partial \omega_j^2}\right|}{1+4\left(\frac{\partial \ell(\hat{\theta}|\omega)}{\partial \omega_j}\right)^2}\right\}}{\sqrt{1+4\sum_{j=1}^{q}\left(\frac{\partial \ell(\hat{\theta}|\omega)}{\partial \omega_j}\right)^2}} \tag{3.4.31}$$

To understand the asymptotic property of (3.4.31), again we need to derive the limits of the first two derivatives of $l(\hat{\theta}|\omega)$ with respect to ω_j. Using multivariate and higher derivatives chain rule, it follows

$$\frac{\partial l(\hat{\theta}|\omega)}{\partial \omega_j} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial l_i(\hat{\theta}|\omega)}{\partial \gamma_i}\frac{\partial \gamma_i}{\partial \omega_j} \tag{3.4.32}$$

$$\frac{\partial^2 l(\hat{\theta}|\omega)}{\partial \omega_j^2} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 l_i(\hat{\theta}|\omega)}{\partial \gamma_i^2}\left(\frac{\partial \gamma_i}{\partial \omega_j}\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial l_i(\hat{\theta}|\omega)}{\partial \gamma_i}\frac{\partial^2 \gamma_i}{\partial \omega_j^2}, \tag{3.4.33}$$

where

$$\frac{\partial l_i(\hat{\theta}|\omega)}{\partial \gamma_i} = \{1-\hat{\pi}_i\}\left[\frac{y_i^*}{\hat{\pi}_i + \gamma_i\{1-\hat{\pi}_i\}} - \frac{1-y_i^*}{1-\hat{\pi}_i-\gamma_i\{1-\hat{\pi}_i\}}\right] \tag{3.4.34}$$

$$\frac{\partial^2 l_i(\hat{\theta}|\omega)}{\partial \gamma_i^2} = -\{1-\hat{\pi}_i\}^2\left(\frac{y_i^*}{[\hat{\pi}_i+\gamma_i(1-\hat{\pi}_i)]^2} + \frac{1-y_i^*}{[1-\hat{\pi}_i-\gamma_i\{1-\hat{\pi}_i\}]^2}\right) \tag{3.4.35}$$

and $\partial\gamma_i/\partial\omega_j = \partial g(\omega_1, ..., \omega_q; x_i)/\partial\omega_j$ and $\partial^2\gamma_i/\partial\omega_j^2 = \partial^2 g(\omega_1, ..., \omega_q; x_i)/\partial\omega_j^2$.

Since

$$\hat{\pi}_i^* = \hat{P}(Y_i^* = 1|x_i) = \hat{\pi}_i + \gamma_i\{1-\hat{\pi}_i\} \quad \text{and} \quad 1-\hat{\pi}_i = \frac{1-\hat{\pi}_i^*}{1-\gamma_i},$$

(3.4.34) and (3.4.35) can be simplified as

$$\frac{\partial l_i(\hat{\theta}|\omega)}{\partial \gamma_i} = \left(\frac{1-\hat{\pi}_i^*}{1-\gamma_i}\right)\left(\frac{y_i^*}{\hat{\pi}_i^*} - \frac{1-y_i^*}{1-\hat{\pi}_i^*}\right) \tag{3.4.36}$$

$$\frac{\partial^2 l_i(\hat{\theta}|\omega)}{\partial \gamma_i^2} = -\left(\frac{1-\hat{\pi}_i^*}{1-\gamma_i}\right)^2\left(\frac{y_i^*}{\hat{\pi}_i^{*2}} + \frac{1-y_i^*}{(1-\hat{\pi}_i^*)^2}\right) \tag{3.4.37}$$

From (3.4.32) and (3.4.36), and using Slutsky's and continuous mapping theorems

$$\begin{aligned}
\frac{\partial l(\hat{\theta}|\omega)}{\partial \omega_j} =& \frac{1}{n}\sum_{i=1}^{n}\frac{\partial l_i(\hat{\theta}|\omega)}{\partial \gamma_i}\frac{\partial \gamma_i}{\partial \omega_j}\\
=& \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1-\hat{\pi}_i^*}{1-\gamma_i}\right)\left(\frac{y_i^*}{\hat{\pi}_i^*} - \frac{1-y_i^*}{1-\hat{\pi}_i^*}\right)\frac{\partial \gamma_i}{\partial \omega_j}\\
\xrightarrow{P}& E\left\{\left(\frac{1-\pi_i^*}{1-\gamma_i}\right)\left(\frac{Y_i^*}{\pi_i^*} - \frac{1-Y_i^*}{1-\pi_i^*}\right)\frac{\partial \gamma_i}{\partial \omega_j}\right\}\\
=& E\left\{\left(\frac{1-\pi_i^*}{1-\gamma_i}\right)\left(\frac{E(Y_i^*|X_i)}{\pi_i^*} - \frac{1-E(Y_i^*|X_i)}{1-\pi_i^*}\right)\frac{\partial \gamma_i}{\partial \omega_j}\right\}\\
=& E\left\{\left(\frac{1-\pi_i^*}{1-\gamma_i}\right)\left(\frac{\pi_i^*}{\pi_i^*} - \frac{1-\pi_i^*}{1-\pi_i^*}\right)\frac{\partial \gamma_i}{\partial \omega_j}\right\}\\
=& 0
\end{aligned} \tag{3.4.38}$$

Similarly, in (3.4.33), the first component on the right hand side converges in probability to

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 l_i(\hat{\theta}|\omega)}{\partial\gamma_i^2}\left(\frac{\partial\gamma_i}{\partial\omega_j}\right)^2 =& -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1-\hat{\pi}_i^*}{1-\gamma_i}\right)^2\left(\frac{y_i^*}{\hat{\pi}_i^{*2}}+\frac{1-y_i^*}{(1-\hat{\pi}_i^*)^2}\right)\left(\frac{\partial\gamma_i}{\partial\omega_j}\right)^2\\
\xrightarrow{P}& -E\left\{\left(\frac{1-\pi_i^*}{1-\gamma_i}\right)^2\left(\frac{Y_i^*}{\pi_i^{*2}}+\frac{1-Y_i^*}{(1-\pi_i^*)^2}\right)\left(\frac{\partial\gamma_i}{\partial\omega_j}\right)^2\right\}\\
=& -E\left\{\left(\frac{1-\pi_i^*}{1-\gamma_i}\right)^2\left(\frac{E(Y_i^*|X_i)}{\pi_i^{*2}}+\frac{1-E(Y_i^*|X_i)}{(1-\pi_i^*)^2}\right)\left(\frac{\partial\gamma_i}{\partial\omega_j}\right)^2\right\}\\
=& -E\left\{\left(\frac{1-\pi_i^*}{1-\gamma_i}\right)^2\left(\frac{\pi_i^*}{\pi_i^{*2}}+\frac{1-\pi_i^*}{(1-\pi_i^*)^2}\right)\left(\frac{\partial\gamma_i}{\partial\omega_j}\right)^2\right\}\\
=& -E\left\{\left(\frac{1-\pi_i^*}{1-\gamma_i}\right)^2\left(\frac{1}{\pi_i^*}+\frac{1}{(1-\pi_i^*)}\right)\left(\frac{\partial\gamma_i}{\partial\omega_j}\right)^2\right\}\\
=& -E\left\{\left(\frac{1-\pi_i^*}{1-\gamma_i}\right)^2\frac{1}{\pi_i^*(1-\pi_i^*)}\left(\frac{\partial\gamma_i}{\partial\omega_j}\right)^2\right\}\\
=& -E\left\{\left(\frac{1}{1-\gamma_i}\right)^2\left(\frac{1-\pi_i^*}{\pi_i^*}\right)\left(\frac{\partial\gamma_i}{\partial\omega_j}\right)^2\right\} \qquad (3.4.39)
\end{aligned}$$

and the second component on the right hand side converges in probability to

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\frac{\partial l_i(\hat{\theta}|\omega)}{\partial\gamma_i}\frac{\partial^2\gamma_i}{\partial\omega_j^2} =& \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1-\hat{\pi}_i^*}{1-\gamma_i}\right)\left(\frac{y_i^*}{\hat{\pi}_i^*}-\frac{1-y_i^*}{1-\hat{\pi}_i^*}\right)\frac{\partial^2\gamma_i}{\partial\omega_j^2}\\
\xrightarrow{P}& E\left\{\left(\frac{1-\pi_i^*}{1-\gamma_i}\right)\left(\frac{Y_i^*}{\pi_i^*}-\frac{1-Y_i^*}{1-\pi_i^*}\right)\frac{\partial^2\gamma_i}{\partial\omega_j^2}\right\}\\
=& E\left\{\left(\frac{1-\pi_i^*}{1-\gamma_i}\right)\left(\frac{E(Y_i^*|X_i)}{\pi_i^*}-\frac{1-E(Y_i^*|X_i)}{1-\pi_i^*}\right)\frac{\partial^2\gamma_i}{\partial\omega_j^2}\right\}\\
=& E\left\{\left(\frac{1-\pi_i^*}{1-\gamma_i}\right)\left(\frac{\pi_i^*}{\pi_i^*}-\frac{1-\pi_i^*}{1-\pi_i^*}\right)\frac{\partial^2\gamma_i}{\partial\omega_j^2}\right\}\\
=& 0 \qquad (3.4.40)
\end{aligned}$$

Therefore, the components in (3.4.31) converge in probability respectively to

$$\frac{\partial^2 l(\hat{\theta}|\omega)}{\partial\omega_j^2} \xrightarrow{P} -E\left\{\left(\frac{1}{1-\gamma_i}\right)^2\left(\frac{1-\pi_i^*}{\pi_i^*}\right)\left(\frac{\partial\gamma_i}{\partial\omega_j}\right)^2\right\} \qquad (3.4.41)$$

based on the results (3.4.33), (3.4.36), (3.4.37), (3.4.39) and (3.4.40), and

$$1 + 4\left(\frac{\partial \ell(\hat{\theta}|\omega)}{\partial \omega_j}\right)^2 \xrightarrow{P} 1 \tag{3.4.42}$$

$$\sqrt{1 + 4\sum_{j=1}^{q}\left(\frac{\partial \ell(\hat{\theta}|\omega)}{\partial \omega_j}\right)^2} \xrightarrow{P} 1 \tag{3.4.43}$$

based on the results (3.4.32), (3.4.36) and (3.4.38).

The results (3.4.41)-(3.4.43) together lead to the limit of the sum of curvatures:

$$\begin{aligned}\sum_{j=1}^{q}\kappa_{h_{\omega_j}} \xrightarrow{P} & 2\sum_{j=1}^{q}E\left\{\left(\frac{1}{1-\gamma_i}\right)^2\left(\frac{1-\pi_i^*}{\pi_i^*}\right)\left(\frac{\partial\gamma_i}{\partial\omega_j}\right)^2\right\} \\ =& 2E\left\{\left(\frac{1-\pi_i^*}{\pi_i^*}\right)\left(\frac{1}{1-\gamma_i}\right)^2\sum_{j=1}^{q}\left(\frac{\partial\gamma_i}{\partial\omega_j}\right)^2\right\}\end{aligned} \tag{3.4.44}$$

Note that in (3.4.44), $\pi_i^* = P(Y_i^* = 1|X_i)$ only depends on the true distribution of Y_i^* conditional on X_i, and therefore is invariant with respect to the assumed ω and γ_i. The sum of curvatures are minimized when $\left(\frac{1}{1-\gamma_i}\right)^2\sum_{j=1}^{q}\left(\frac{\partial\gamma_i}{\partial\omega_j}\right)^2$ are minimized. By given the function g, one can derive $\sum_{j=1}^{q}\left(\frac{\partial\gamma_i}{\partial\omega_j}\right)^2$ and use (3.4.44) to find the most stable model with the lowest sum of curvatures.

Again using the special cases in Section 3.4.1 and 3.4.2 as examples:

(a) In Section 3.4.1, $\gamma_1 = ... = \gamma_n = \gamma = \omega_1$. It follows that

$$\frac{\partial\gamma_i}{\partial\omega_j} = \frac{\partial\gamma}{\partial\omega_1} = 1 \quad \text{and} \quad \sum_{j=1}^{q}\left(\frac{\partial\gamma_i}{\partial\omega_j}\right)^2 = 1$$

Therefore,

$$\left(\frac{1}{1-\gamma_i}\right)^2\sum_{j=1}^{q}\left(\frac{\partial\gamma_i}{\partial\omega_j}\right)^2 = \left(\frac{1}{1-\gamma}\right)^2$$

and hence the curvature are minimized at $\gamma = 0$ for $0 \leq \gamma \leq 1$. As already proved in Section 3.4.1, the assumed model with $\gamma = 0$ is the most stable model as it has the lowest curvature.

(b) In Section 3.4.2, $\gamma_i = \omega_1(1 - X_i) + \omega_2 X_i$. It follows that

$$\frac{\partial\gamma_i}{\partial\omega_1} = (1 - X_i), \quad \frac{\partial\gamma_i}{\partial\omega_2} = X_i, \quad \text{and} \quad \sum_{j=1}^{q}\left(\frac{\partial\gamma_i}{\partial\omega_j}\right)^2 = \{(1 - X_i)X_i\}^2$$

Therefore,

$$\left(\frac{1}{1-\gamma_i}\right)^2 \sum_{j=1}^{q} \left(\frac{\partial \gamma_i}{\partial \omega_j}\right)^2 = \left(\frac{1}{1-\gamma_i}\right)^2 \{(1-X_i)X_i\}^2$$

and hence the sum of curvatures are minimized at $\omega_1 = \omega_2 = 0$ and $\gamma_i = 0$ for $0 \leq \gamma_i \leq 1$. Although, the assumed model has the lowest curvature when there is no information about the error size, it is important to assess stability and the robustness of model estimates within a plausible range of error size.

Chapter 4

Influence Measure for Misclassified Presence of Binary Outcome

In Section 3.4 of Chapter Three, we derived influence measures that can be used to investigate how stable an assumed or modified model is under the impact of error-prone $Y_i^* = 0$. This Chapter shows a special case of a binary outcome where only the presence of an attribute; $Y_i^* = 1$ is subject to measurement error. Following the general case of misclassification setting introduced in Section 3.4.3 of Chapter Three, we assess how stable an assume or modified model is under nondifferential and differential misclassified $Y_i^* = 1$. The final results will be largely similar to the description provided in Section 3.4.3, but there are important differences along the line of the derivation.

The chapter is structured as follows. In Section 4.1, we provide the study design. Following the general setting in Section 3.4.3, we described how to derive influence measure for nondifferential and differential misclassification in Section 4.2.

4.1 Study Design

In Chapter Three, we introduced the log-likelihood function of a true response variable Y_i as

$$\ell(\theta) = \sum_{i=1}^{n} y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i), \tag{4.1.1}$$

The assumed logistic model for an observed response Y_i^* is

$$Y_i^* | X_i \sim Bin(1, \pi_i^*), \quad \log\left(\frac{\pi_i^*}{1 - \pi_i^*}\right) = \alpha^* + \theta^* X_i. \tag{4.1.2}$$

Here, the focus is on error-prone $Y_i^* = 1$. In this case, the probability of measurement error $\zeta = Pr(Y_i^* = 0 | Y_i = 1)$. The mechanism of measurement error of $Y^* = 1$ is shown in Figure 4.1.

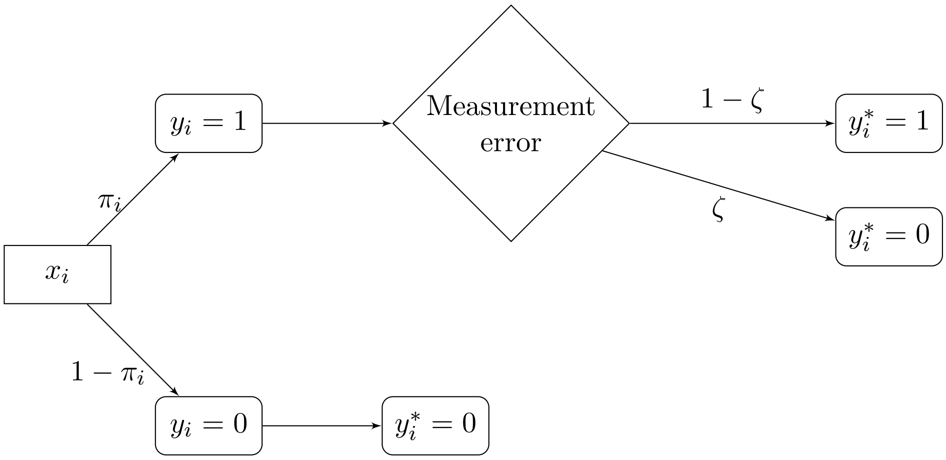


Figure 4.1: Mechanism of Misclassification of the presence of an outcome attribute

By assuming the probability of measurement error to be ζ,

$$\begin{aligned} P(Y_i^* = 1|X_i) =& P(Y_i^* = 1|Y_i = 1)P(Y_i = 1|X_i) + P(Y_i^* = 1|Y_i = 0)P(Y_i = 0|X_i) \\ =& (1-\zeta)\pi_i \end{aligned}$$

The distribution of the observed response variable conditional on ζ is

$$Y_i^*|X_i \sim Bin\big(1, (1-\zeta)\pi_i\big), \tag{4.1.3}$$

And the corresponding log-likelihood for a given ζ is

$$\ell(\theta|\zeta) = \sum_{i=1}^{n} y_i^* \log\Big((1-\zeta)\pi_i\Big) + (1-y_i^*)\log\Big(1 - \big((1-\zeta)\pi_i\big)\Big) \tag{4.1.4}$$

To consider a measurement error problem in this special case, we introduced the probability of measurement error ζ as a modification weight into (4.1.2) and this results to the modified model (4.1.3). The true value of ζ is unknown, but we can vary the value of ζ between 0 and 1 to assess the impact of measurement error, by comparing the assumed and modified models, likelihoods or estimates. When $\zeta = 0$, (4.1.3) is equal to (4.1.2), $\ell(\theta^*) = \ell(\theta|\zeta)$ and $\theta^* = \theta$. When $\zeta > 0$, the modified model is deviated from the assumed model and $\ell(\theta^*) \neq \ell(\theta|\zeta)$. Each value assigned to ζ corresponds to a distinct model. Different values of $\zeta > 0$ perturb the assumed model (4.1.2) to be different modified models.

In the following, we derive influence measure for the special case of $Y_i^* = 1$ under the impact of nondifferential and differential misclasification patterns.

4.2 Influence Measure using General case of Misclassification

Following the notations introduced in Section 3.4.3, we consider modification weights $\omega = (\omega_1, \cdots, \omega_q)$ which influence individual probability of measurement error ζ_i. Here, the general setting (3.4.29) to the case of misclassified $Y_i^* = 1$ is

$$P(Y_i^* = 0|Y_i = 1, x_i) = \zeta_i = g(\omega; x_i) = g(\omega_1, ..., \omega_q; x_i) \tag{4.2.1}$$

where

- The modification weight $\omega = \omega_1$ which can influence the probability of measurement error in the case of nondifferential misclassification is

$$\zeta_1 = \cdots = \zeta_n = \zeta = g(\omega; x_i) = g(\omega_1) = \omega_1$$

- And in the case of differential misclassification, the vector of modification weights $\omega = (\omega_1, \omega_2)$ which can influence the probability of measurement error through function g:

$$\zeta_i = g(\omega; x_i) = g(\omega_1, \omega_2; x_i) = \omega_1(1 - x_i) + \omega_2 x_i$$

Therefore, the average log-likelihood function is

$$l(\hat{\theta}|\omega) = \frac{1}{n}\sum_{i=1}^{n} l_i(\hat{\theta}|\omega) = \frac{1}{n}\sum_{i=1}^{n} y_i^* \log\{\hat{\pi}_i(1-\zeta_i)\} + (1-y_i^*)\log\{1-((1-\zeta_i)\hat{\pi}_i)\}, \tag{4.2.2}$$

where

$$\hat{\pi}_i = \frac{\exp(\hat{\alpha} + \hat{\theta}x_i)}{1 + \exp(\hat{\alpha} + \hat{\theta}x_i)}$$

and $\hat{\alpha}$ and $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_p)$ are estimated under the assumed $\omega = (\omega_1, ..., \omega_q)$.

Again, to assess how stable a model is around each particular point in Ω, we follow the asymptotic property of the sum of curvatures, i.e

$$\sum_{j=1}^{q} \kappa_{h_{\omega_j}} = \sum_{j=1}^{q} \frac{2\left\{\frac{\left|\frac{\partial^2 \ell(\hat{\theta}|\omega)}{\partial \omega_j^2}\right|}{1+4\left(\frac{\partial \ell(\hat{\theta}|\omega)}{\partial \omega_j}\right)^2}\right\}}{\sqrt{1+4\sum_{j=1}^{q}\left(\frac{\partial \ell(\hat{\theta}|\omega)}{\partial \omega_j}\right)^2}} \tag{4.2.3}$$

In the case of misclassified $Y_i^* = 1$, the final results of the limits of the first two derivatives of $l(\hat{\theta}|\omega)$ with respect to ω_j is largely similar to proof in Section 3.4.3, but there are important differences along the line of the derivation. Using multivariate and higher derivatives chain rule, it follows

$$\frac{\partial l(\hat{\theta}|\omega)}{\partial \omega_j} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial l_i(\hat{\theta}|\omega)}{\partial \zeta_i}\frac{\partial \zeta_i}{\partial \omega_j} \tag{4.2.4}$$

$$\frac{\partial^2 l(\hat{\theta}|\omega)}{\partial \omega_j^2} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 l_i(\hat{\theta}|\omega)}{\partial \zeta_i^2}\left(\frac{\partial \zeta_i}{\partial \omega_j}\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial l_i(\hat{\theta}|\omega)}{\partial \zeta_i}\frac{\partial^2 \zeta_i}{\partial \omega_j^2}, \tag{4.2.5}$$

where

$$\frac{\partial l_i(\hat{\theta}|\omega)}{\partial \zeta_i} = -\hat{\pi}_i\left[\frac{y_i^*}{(1-\zeta_i)\hat{\pi}_i} - \frac{1-y_i^*}{1-(1-\zeta_i)\hat{\pi}_i}\right] \tag{4.2.6}$$

$$\frac{\partial^2 l_i(\hat{\theta}|\omega)}{\partial \zeta_i^2} = -\hat{\pi}_i^2\left(\frac{y_i^*}{[(1-\zeta_i)\hat{\pi}_i]^2} + \frac{1-y_i^*}{[1-(1-\zeta_i)\hat{\pi}_i]^2}\right) \tag{4.2.7}$$

and $\partial\zeta_i/\partial\omega_j = \partial g(\omega_1, ..., \omega_q; x_i)/\partial\omega_j$ and $\partial^2\zeta_i/\partial\omega_j^2 = \partial^2 g(\omega_1, ..., \omega_q; x_i)/\partial\omega_j^2$.

Since

$$\hat{\pi}_i^* = \hat{P}(Y_i^* = 1|x_i) = (1-\zeta_i)\hat{\pi}_i \quad \text{and} \quad \hat{\pi}_i = \frac{\hat{\pi}_i^*}{1-\zeta_i},$$

(4.2.6) and (4.2.7) can be simplified as

$$\frac{\partial l_i(\hat{\theta}|\omega)}{\partial \zeta_i} = -\left(\frac{\hat{\pi}_i^*}{1-\zeta_i}\right)\left(\frac{y_i^*}{\hat{\pi}_i^*} - \frac{1-y_i^*}{1-\hat{\pi}_i^*}\right) \tag{4.2.8}$$

$$\frac{\partial^2 l_i(\hat{\theta}|\omega)}{\partial \zeta_i^2} = -\left(\frac{\hat{\pi}_i^*}{1-\zeta_i}\right)^2\left(\frac{y_i^*}{\hat{\pi}_i^{*2}} + \frac{1-y_i^*}{(1-\hat{\pi}_i^*)^2}\right) \tag{4.2.9}$$

From (4.2.4) and (4.2.8), and using Slutsky's and continuous mapping theorems

$$\begin{aligned}
\frac{\partial l(\hat{\theta}|\omega)}{\partial \omega_j} =& \frac{1}{n}\sum_{i=1}^{n}\frac{\partial l_i(\hat{\theta}|\omega)}{\partial \zeta_i}\frac{\partial \zeta_i}{\partial \omega_j} \\
=& -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{\pi}_i^*}{1-\zeta_i}\right)\left(\frac{y_i^*}{\hat{\pi}_i^*} - \frac{1-y_i^*}{1-\hat{\pi}_i^*}\right)\frac{\partial \zeta_i}{\partial \omega_j} \\
\xrightarrow{P}& -E\left\{\left(\frac{\pi_i^*}{1-\zeta_i}\right)\left(\frac{Y_i^*}{\pi_i^*} - \frac{1-Y_i^*}{1-\pi_i^*}\right)\frac{\partial \zeta_i}{\partial \omega_j}\right\} \\
=& -E\left\{\left(\frac{\pi_i^*}{1-\zeta_i}\right)\left(\frac{E(Y_i^*|X_i)}{\pi_i^*} - \frac{1-E(Y_i^*|X_i)}{1-\pi_i^*}\right)\frac{\partial \zeta_i}{\partial \omega_j}\right\} \\
=& -E\left\{\left(\frac{\pi_i^*}{1-\zeta_i}\right)\left(\frac{\pi_i^*}{\pi_i^*} - \frac{1-\pi_i^*}{1-\pi_i^*}\right)\frac{\partial \zeta_i}{\partial \omega_j}\right\} \\
=& 0
\end{aligned} \tag{4.2.10}$$

Similarly, in (4.2.5), the first component on the right hand side converges in probability to

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 l_i(\hat{\theta}|\omega)}{\partial\zeta_i^2}\left(\frac{\partial\zeta_i}{\partial\omega_j}\right)^2 =& -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{\pi}_i^*}{1-\zeta_i}\right)^2\left(\frac{y_i^*}{\hat{\pi}_i^{*2}}+\frac{1-y_i^*}{(1-\hat{\pi}_i^*)^2}\right)\left(\frac{\partial\zeta_i}{\partial\omega_j}\right)^2\\
\xrightarrow{P}& -E\left\{\left(\frac{\pi_i^*}{1-\zeta_i}\right)^2\left(\frac{Y_i^*}{\pi_i^{*2}}+\frac{1-Y_i^*}{(1-\pi_i^*)^2}\right)\left(\frac{\partial\zeta_i}{\partial\omega_j}\right)^2\right\}\\
=& -E\left\{\left(\frac{\pi_i^*}{1-\zeta_i}\right)^2\left(\frac{E(Y_i^*|X_i)}{\pi_i^{*2}}+\frac{1-E(Y_i^*|X_i)}{(1-\pi_i^*)^2}\right)\left(\frac{\partial\zeta_i}{\partial\omega_j}\right)^2\right\}\\
=& -E\left\{\left(\frac{\pi_i^*}{1-\zeta_i}\right)^2\left(\frac{\pi_i^*}{\pi_i^{*2}}+\frac{1-\pi_i^*}{(1-\pi_i^*)^2}\right)\left(\frac{\partial\zeta_i}{\partial\omega_j}\right)^2\right\}\\
=& -E\left\{\left(\frac{\pi_i^*}{1-\zeta_i}\right)^2\left(\frac{1}{\pi_i^*}+\frac{1}{(1-\pi_i^*)}\right)\left(\frac{\partial\zeta_i}{\partial\omega_j}\right)^2\right\}\\
=& -E\left\{\left(\frac{\pi_i^*}{1-\zeta_i}\right)^2\frac{1}{\pi_i^*(1-\pi_i^*)}\left(\frac{\partial\zeta_i}{\partial\omega_j}\right)^2\right\}\\
=& -E\left\{\left(\frac{1}{1-\zeta_i}\right)^2\left(\frac{\pi_i^*}{1-\pi_i^*}\right)\left(\frac{\partial\zeta_i}{\partial\omega_j}\right)^2\right\} \qquad (4.2.11)
\end{aligned}$$

and the second component on the right hand side converges in probability to

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\frac{\partial l_i(\hat{\theta}|\omega)}{\partial\zeta_i}\frac{\partial^2\zeta_i}{\partial\omega_j^2} =& -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{\pi}_i^*}{1-\zeta_i}\right)\left(\frac{y_i^*}{\hat{\pi}_i^*}-\frac{1-y_i^*}{1-\hat{\pi}_i^*}\right)\frac{\partial^2\zeta_i}{\partial\omega_j^2}\\
\xrightarrow{P}& -E\left\{\left(\frac{\pi_i^*}{1-\zeta_i}\right)\left(\frac{Y_i^*}{\pi_i^*}-\frac{1-Y_i^*}{1-\pi_i^*}\right)\frac{\partial^2\zeta_i}{\partial\omega_j^2}\right\}\\
=& -E\left\{\left(\frac{\pi_i^*}{1-\zeta_i}\right)\left(\frac{E(Y_i^*|X_i)}{\pi_i^*}-\frac{1-E(Y_i^*|X_i)}{1-\pi_i^*}\right)\frac{\partial^2\zeta_i}{\partial\omega_j^2}\right\}\\
=& -E\left\{\left(\frac{\pi_i^*}{1-\zeta_i}\right)\left(\frac{\pi_i^*}{\pi_i^*}-\frac{1-\pi_i^*}{1-\pi_i^*}\right)\frac{\partial^2\zeta_i}{\partial\omega_j^2}\right\}\\
=&0 \qquad (4.2.12)
\end{aligned}$$

Therefore, the components in (4.2.5) converge in probability respectively to

$$\frac{\partial^2 l(\hat{\theta}|\omega)}{\partial\omega_j^2}\xrightarrow{P} -E\left\{\left(\frac{1}{1-\zeta_i}\right)^2\left(\frac{\pi_i^*}{1-\pi_i^*}\right)\left(\frac{\partial\zeta_i}{\partial\omega_j}\right)^2\right\} \qquad (4.2.13)$$

based on the results (4.2.5), (4.2.8), (4.2.9), (4.2.11) and (4.2.12), and

$$1+4\left(\frac{\partial\ell(\hat{\theta}|\omega)}{\partial\omega_j}\right)^2 \xrightarrow{P} 1 \tag{4.2.14}$$

$$\sqrt{1+4\sum_{j=1}^{q}\left(\frac{\partial\ell(\hat{\theta}|\omega)}{\partial\omega_j}\right)^2} \xrightarrow{P} 1 \tag{4.2.15}$$

based on the results (4.2.4), (4.2.8) and (4.2.10).

The results (4.2.13)-(4.2.15) together lead to the limit of the sum of curvatures:

$$\begin{aligned}\sum_{j=1}^{q}\kappa_{h_{\omega_j}} \xrightarrow{P} & 2\sum_{j=1}^{q}E\left\{\left(\frac{1}{1-\zeta_i}\right)^2\left(\frac{\pi_i^*}{1-\pi_i^*}\right)\left(\frac{\partial\zeta_i}{\partial\omega_j}\right)^2\right\}\\ =& 2E\left\{\left(\frac{\pi_i^*}{1-\pi_i^*}\right)\left(\frac{1}{1-\zeta_i}\right)^2\sum_{j=1}^{q}\left(\frac{\partial\zeta_i}{\partial\omega_j}\right)^2\right\}\end{aligned} \tag{4.2.16}$$

Note that in (4.2.16), $\pi_i^* = P(Y_i^* = 1|X_i)$ only depends on the true distribution of Y_i^* conditional on X_i, and therefore is invariant with respect to the assumed ω and ζ_i. The sum of curvatures is minimized when $\left(\frac{1}{1-\zeta_i}\right)^2\sum_{j=1}^{q}\left(\frac{\partial\zeta_i}{\partial\omega_j}\right)^2$ is minimized.

In the following, we show how to use the sums of curvatures, (4.2.16) to study how stable an assumed or modified model is under the impact of potential measurement error ζ.

4.2.1 Stability under Nondifferential Misclassification

In this section, we consider the case of nondifferential misclassification where all the individuals have the same probability of measurement error ζ. Here, $q = 1$ and the modification weight is $\omega = \omega_1$, which can influence the probability of measurement error by

$$\zeta_1 = ... = \zeta_n = \zeta = g(\omega; x_i) = g(\omega_1) = \omega_1.$$

Given function g, we can derive $\sum_{j=1}^{q}\left(\frac{\partial\zeta_i}{\partial\omega_j}\right)^2$ and use (4.2.16) to find the most stable model with the lowest sum of curvatures.

It follows that

$$\frac{\partial\zeta_i}{\partial\omega_j} = \frac{\partial\zeta}{\partial\omega_1} = 1 \quad \text{and} \quad \sum_{j=1}^{q}\left(\frac{\partial\zeta_i}{\partial\omega_j}\right)^2 = 1$$

Therefore,

$$\left(\frac{1}{1-\zeta_i}\right)^2\sum_{j=1}^{q}\left(\frac{\partial\zeta_i}{\partial\omega_j}\right)^2 = \left(\frac{1}{1-\zeta}\right)^2$$

and hence the curvature are minimized at $\zeta = 0$ for $0 \leq \zeta \leq 1$.

Although, the result suggests that when there is no information about the measurement error, the assumed model with $\zeta = 0$ has the lowest curvature, it is important to assess the robustness of model estimate(s) or conclusion(s) within a plausible range of ζ, and report all the findings.

4.2.2 Stability under Differential Misclassification

Differential misclassifcation is based on a binary x_i where the modification weights $\omega = (\omega_1, \omega_2)$ influence the probability of measurement error through the function g:

$$\zeta_i = g(\omega; x_i) = g(\omega_1, \omega_2; x_i) = \omega_1(1 - x_i) + \omega_2 x_i \quad \text{for } i = 1, ..., n.$$

In differential case, $q = 2$ and through function g, $\omega_1 = \zeta_0$, $\omega_2 = \zeta_1$ such that, the vector of weights $\omega = (\omega_1, \omega_2)$ is bound by $[0, 1]^2 \in R^2$.

Again we can use (4.2.16) to find the most stable model with the lowest sum of curvatures. In this special case, $\zeta_i = \omega_1(1 - X_i) + \omega_2 X_i$. It follows that

$$\frac{\partial \zeta_i}{\partial \omega_1} = (1 - X_i), \quad \frac{\partial \zeta_i}{\partial \omega_2} = X_i, \quad \text{and} \quad \sum_{j=1}^{q} \left(\frac{\partial \zeta_i}{\partial \omega_j} \right)^2 = \{(1 - X_i) X_i\}^2$$

Therefore,

$$\left(\frac{1}{1 - \zeta_i} \right)^2 \sum_{j=1}^{q} \left(\frac{\partial \zeta_i}{\partial \omega_j} \right)^2 = \left(\frac{1}{1 - \zeta_i} \right)^2 \{(1 - X_i) X_i\}^2$$

and hence the sum of curvatures are minimized at $\omega_1 = \omega_2 = 0$ and $\zeta_i = 0$ for $0 \leq_i \leq 1$.

Again, it is important to assess model stability and quality of model estimates within plausible range of error size.

Chapter 5

Application and Results

This Chapter shows the illustrative example with binary data. The influence measurement method from our research is demonstrated in the analysis of our motivating example, designed to study the effect of a rehabilitation programme on juvenile offenders. Given that different compositions of measured values of binary outcomes often exist in real studies, we conducted simulation studies for different scenarios for the special cases of binary outcome investigated in both Chapters Three and Four.

In our study, there is no prior information about the error size. Hence, the true value of measurement error (either γ or ζ) is unknown in the analyses illustrated. Thus, for model development and in order to empirically assess model stability under the impact of measurement error, the values of the error size range from 0 to 1. Note that in a specific context, extreme values of the error size may not be plausible. In that case, if measurement error is suspected, a set of plausible values can be incorporated in the model, otherwise, one possible solution is to consider range of values between 0 and 1.

The Chapter is structured as follows. In Section 5.1, we present the analysis and results from the rehabilitation programme study. We next consider four different simulation studies. In Section 5.2, we first present analysis and results (of the influence measure methods in Section 3.4.1 in Chapter Three) from simulation study I, where the absence attribute of outcome $Y_i^* = 0$ is based on nondifferential misclassification assumption. This is followed by simulation study II, which is based on a differential measurement error assumption of $Y_i^* = 0$, the absence of outcome in Section 5.3. In Section 5.4, we consider simulation III, where the underlining assumption is that error-prone $Y_i^* = 1$ is nondifferentially misclassified. Analysis and results from simulation IV is presented in Section 5.5. The Simulation IV setting follows a differential misclassification of $Y_i^* = 1$, the presence of attribute of outcome.

5.1 Application to Rehabilitation Programme study

The first numerical implementation is as follows. To illustrate the stability of an assumed model under the impact of a measurement error of $Y_i^* = 0$, we apply the influence measure method to the rehabilitation programme study data. The motivating example study is taken from a rehabilitation programme designed for juvenile offenders. This study was conducted as a result of an increase in the number of crimes being committed by juveniles, which leads to the research question of how effective a rehabilitation programme can be. The study was used to examine if a rehabilitation programme can reduce the reoffending rate amongst juvenile offenders. Prior to the study, the participants had an episode of committed crimes or they were something of a risk to public security. The outcome variable of interest is a binary attribute of participants either reoffending with $Y_i^* = 1$ in the case of no reconviction observed, or $Y_i^* = 0$ otherwise. We make the reasonable assumption that there is no reconviction if there is no offence, so $Y_i^* = 1$ if $Y_i = 1$, but an actual offence may not lead to detection and reconviction, that is, when $Y_i = 0$, Y_i^* is either 0 or 1. In other words, there is no measurement error if there is no offence, but there may be measurement error if there is an offence. The treatment and control group profiles of the rehabilitation programme study are shown in Table 5.1, where exposure variable X is binary with $x_i = 1$ being those participants who joined the rehabilitation programme, and $x_i = 0$ otherwise. As shown in Table 5.1, hypothetical data relating to a total of 2000 participants was used in the study. There are two groups (either control or treatment) of juvenile offenders in the study. The treatment group has 1000 participants who joined the rehabilitation programme. The remaining 1000 did not join. Amongst the treatment group who underwent the programme, 300 out of 1000 participants reoffended. Meanwhile in the control group, 400 participants did reoffend, while 600 who had not joined the programme did not reoffend.

Table 5.1: Hypothetical Data: Rehabilitation Programme

	Treatment group	Control group
No Re-offence	700	600
Re-offence	300	400

The goal of the rehabilitation study is to compare reoffending between the group that joined the programme and those who did not join. Note that predictors such as gender, race, and education could play a role, but for the purpose of illustrating the influence measure method from our research we performed a univariable analysis. For the motivating data, we sought to test the effect of the rehabilitation programme on juvenile offenders by using a logistic model to perform the significance test(s).

5.1.1 Naive Analysis

First, a standard logistic model is fitted to the study data, and the results of the analysis are shown in Table 5.2. The estimate of the parameter of interest $\hat{\theta}^*$ is the log odd ratio comparing reoffending amongst participants that joined the rehabilitation programme to reoffending amongst those who did not join. From the naive analysis results, it can be seen that the effect of the rehabilitation programme is significant (95% confidence interval (CI): 0.26, 0.63), but the response variable $Y_i^* = 0$ could be error-prone, and the naive analysis assumes no measurement error. Hence, this suggests that $\hat{\theta}^*$ (the observed effectiveness of the rehabilitation programme) may be inaccurate due to undetected crime or measurement error. As a result, we can say that the conclusion drawn from the assumed model may be imprecise due to potential undetected crimes in the outcome. Therefore, we need to investigate how the inference can be impacted by undetected crimes.

Table 5.2: Naive analysis results from the Rehabilitation Programme

	$\hat{\theta}^*$(Std.Err)	PValue	95%
Naive Analysis			
	0.44(0.09)	< 0.01	$(0.26, 0.63)$

We assumed that there is no measurement error if there is no offence, but there may be measurement error if there is an offence. By assuming the probability of non-detection to be γ, we introduced γ in the assumed model, and the resulting model is such that, the distribution of the observed outcome variable conditional on γ is $Y_i^*|X_i \sim Bin(1, \pi_i + \gamma(1 - \pi_i))$. The true value of γ is unknown and there is no prior information about γ. Hence, for model development we vary the value of γ between 0 and 1 to empirically assess model stability and biasedness of model estimates under the impact of undetected crimes or measurement error.

To do this, we applied the influence method under both the nondifferential and differential misclassification patterns described in Chapter Three to the rehabilitation data. We calculated model estimates and curvature under the individual measurement error pattern. The model estimates and curvature results are used as a measure to illustrate how the stability evaluation and parameter estimation behave under the impact of slight modification to γ.

5.1.2 Influence Methods Analysis: Nondifferential Misclassification

In this section, we applied the influence measure under nondifferential misclassification of $Y_i^* = 0$ to the rehabilitation programme data. To illustrate the stability and parameter evaluation of an assumed or modified model under a nondifferential mechanism, we obtained the model estimates and curvature for each distinct (assumed and modified) model. Note that in the case of a nondifferential measurement error pattern, the modification weight introduced in the assume model is scalar. Thus, we used the curvature derived in (3.4.5) from Chapter Three to calculate curvature. Values ranging from 0 to 1 are assumed for γ during the analysis. The model estimates and curvature (in log transformation) results are shown in Table 5.3 and Figure 5.1. We use the logarithmic transformation because the plain curve shown in Figure 5.1(b) is more visible in log transformation than the original values. Table 5.3(a) shows an extract of the model estimates under a nondifferential misclassification. It can be seen that the model estimate increases as γ increases from 0 to about when $\gamma = 0.6$, but then breaks. Each γ represents a distinct model with an assumed model at $\gamma = 0$ and modified models when $\gamma \neq 0$. Table 5.3(b) shows an extract of curvature results from the influence measure under nondifferential misclassification of $Y_i^* = 0$. In general, curvature κ increases as the value of probability of non-detection γ increases, with the lowest curvature at $\gamma = 0$ and highest curvature at $\gamma = 1$. In otherwords, as specificity $(1 - \gamma) = P(Y_i^* = 0|Y_i = 0)$ decreases, curvature increases. It follows that at assumed model $\gamma = 0$, specificity is 100% with the lowest curvature value 7.692. However, at modified model $\gamma = 0.67$, specificity is 33% with higher curvature 9.715.

Table 5.3: Application to rehabilitation study: Extract of model estimates and curvature results of distinct models under a nondifferential misclassification pattern. The probability of non-detection γ is allowed to range from 0 to 1. Each γ represents a distinct model with the assume model at $\gamma = 0$ and modified models when $\gamma \neq 0$. (a) Extract of Model Estimates. (b)Extract of Curvatures.

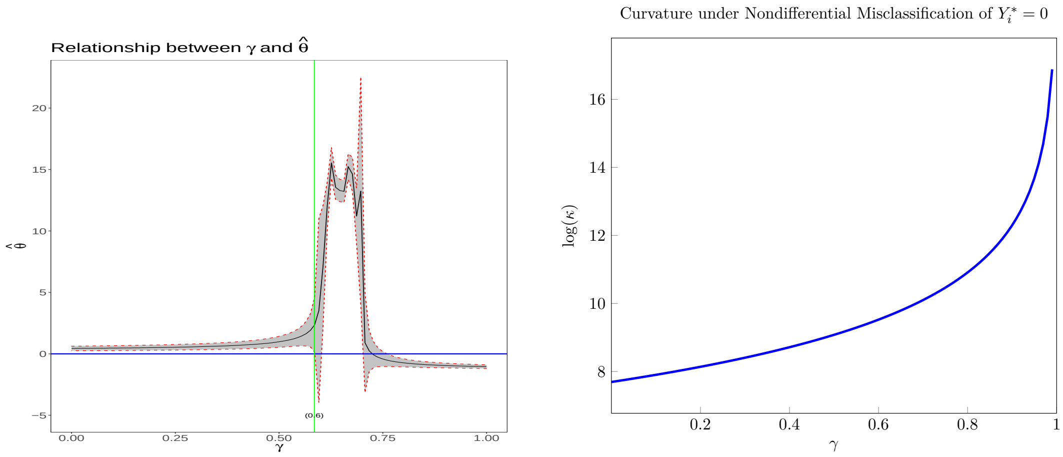
γ	$\hat{\theta}\|\gamma$	95%C.I	P-value
0	0.44	(0.26,0.63)	<0.001
0.14	0.48	(0.28,0.69)	<0.001
0.39	0.68	(0.39,0.97)	<0.001
0.55	1.47	(0.65,2.28)	<0.001
0.63	13.55	(12.55,14.55)	<0.001
0.67	14.63	(13.25,16)	<0.001
0.71	0.25	(-1.44,1.93)	0.77
0.75	-0.49	(-1.03,0.05)	0.07
0.82	-0.82	(-1.09,-0.55)	<0.001
0.87	-0.92	(-1.13,-0.71)	<0.001
0.94	-1	(-1.17,-0.83)	<0.001
0.99	-1.04	(-1.19,-0.89)	<0.001

(a) Model Estimates. CI;Confidence Interval.

γ	$\log(\kappa)$
0	7.692
0.14	7.997
0.39	8.693
0.55	9.314
0.63	9.715
0.67	9.951
0.71	10.218
0.75	10.526
0.82	11.216
0.87	11.912
0.94	13.663
0.99	16.882

(b) Curvature under a Nondifferential Misclassification.

The left plot in Figure 5.1 depicts the model estimates. The solid black line is the estimates, while the red dashed lines indicate the 95% confidence interval. The blue horizontal line is at when $\hat{theta} = 0$, while the green line indicates the intersection of the blue line and the confidence interval. Although there are at least one intersection, there is a break at $\gamma = 0.6$ and wide confidence interval between $\gamma = 0.6$ and $\gamma = 0.8$. The intersection point (the green line) might suggests a cut off value for the unknown γ. In otherwords, the plausible range of error size might be from 0 to 0.6.



(a) Model estimates

(b) Curvature

Figure 5.1: Application to rehabilitation study: Model estimates and curvature under a nondifferential misclassification of $Y_i^* = 0$. The assumed model is at $\gamma = 0$ and modified models when $\gamma \neq 0$

The right plot (Figure 5.1b) is the influence measure (that is, a plane curve of curvature κ against γ) under the nondifferential miclassification of Y_i^*. A point on the curve represents a distinct model with the assumed logistic model at $\gamma = 0$, and modified models otherwise. The plot shows how curvy the plane curve is at each distinct model. The larger a curvature is, the faster the curve is turning, which means the model with large curvature appears to be sensitive to measurement error. From Figure 5.1, we can see that there is a clear increasing pattern as γ increases.

Table 5.3 shows model estimates from $\gamma = 0$ to $\gamma = 0.6$ are unbiased, but the curve of the curvature increases within these values. Although the lowest curvature is at when $\gamma = 0$, it is important to validate or find prior information about γ when measurement error is suspected. On the other hand, if there is no prior information about γ, based on these results, the standard logistic regression model seems more stable compared to modified models.

5.1.3 Influence Methods Analysis: Differential Misclassification

Here, we use the rehabilitation programme study data to illustrate the influence measure under a differential misclassification of $Y_i^* = 0$. In this section, we obtained model estimates, and used the sum of curvatures (3.4.26) from Chapter Three, with assumed values for γ_0 and γ_1, to calculate the normal curvatures along the direction of γ_0 and γ_1. The model estimates are shown in Table 5.4. The normal curvatures results in log transformation for each of the directions $\kappa_{h\gamma_0}$ and $\kappa_{h\gamma_1}$, and that of the sum of the normal curvatures for both directions $\sum_{j=0}^{1} \kappa_{h\gamma_j}$ are depicted in Table 5.5.

Table 5.4: Application to rehabilitation study: Extract of model estimates results of distinct model at different values of γ under a nondifferential undetected crime or measurement error process. The probability of non-detection γ is allowed to range from 0 to 1. Each γ represents a distinct model with assumed model at $\gamma = 0$ and modified models when $\gamma \neq 0$.

[Misc. rate]				
γ_0	γ_1	$\hat{\theta}\|\gamma_0, \gamma_1$	95%CI	P-value
0	0	0.4	(0.26,0.63)	<0.001
	0.1	0.29	(0.10,0.48)	0.003
	0.3	-0.12	(-0.33,0.09)	0.27
	0.5	-0.81	(-1.08,-0.54)	<0.001
	0.7	-8.78	(-298.49,280.93)	0.95
	0.9	-12.35	(-64.06,39.35)	0.64
0.1	0	0.62	(0.43,0.82)	<0.001
	0.1	0.47	(0.27,0.67)	<0.001
	0.3	0.06	(-0.15,0.28)	0.56
	0.5	-0.63	(-0.90,-0.36)	<0.001
	0.7	-8.39	(-244.06,227.28)	0.94
	0.9	-12.60	(-76.52,51.33)	0.69
0.3	0	1.13	(0.91,1.36)	<0.001
	0.1	0.98	(0.75,1.21)	<0.001
	0.3	0.58	(0.33,0.82)	<0.001
	0.5	-0.12	(-0.41,0.18)	0.44
	0.7	-6.47	(-64.16,51.23)	0.83
	0.9	-17.83	(-1149.41,1113.74)	0.98
0.5	0	2.23	(1.83,2.64)	<0.001
	0.1	2.08	(1.67,2.48)	<0.001
	0.3	1.67	(1.26,2.09)	<0.001
	0.5	0.98	(0.53,1.43)	<0.001
	0.7	-5.93	(-106.39,94.54)	0.91
	0.9	-14.88	(-461.96,432.21)	0.95
0.7	0	11.07	(-18.08,40.23)	0.46
	0.1	11	(-18.66,40.65)	0.47
	0.3	10.79	(-17.45,39.04)	0.45
	0.5	10.45	(-27.61,48.5)	0.59
	0.7	7.22	(-3594.28,3608.71)	0.99
	0.9	1.69	(-3392.54,3395.92)	0.99
0.9	0	11.17	(-8.21,30.56)	0.26
	0.1	11.10	(-8.85,31.04)	0.28
	0.3	10.89	(-9.51,31.3)	0.29
	0.5	10.55	(-15.38,36.47)	0.43
	0.7	7.88	(-4375.01,4390.77)	0.99
	0.9	1.71	(-2902.27,2905.69)	0.99

Table 5.5: Application to rehabilitation study: Influence measure under differential measurement error pattern. Each combination of γ_0 and γ_1 is a distinct model. (a)Extract of normal curvatures for direction γ_0 and γ_1. (b)Extract of sum of normal curvatures arranged from lowest to highest.

	[Misc. rate]				
Model	γ_0	γ_1	$\log(\kappa_{h\gamma_0})$	$\log(\kappa_{h\gamma_1})$	$\sum_{j=0}^{1}\log(\kappa_{h\gamma_j})$
1	0	0	7.195	6.754	13.949
2		0.1	7.195	6.964	14.159
3		0.3	7.195	7.467	14.662
4		0.5	7.195	8.139	15.335
5		0.7	7.195	9.162	16.357
6		0.9	7.195	11.359	18.554
7	0.1	0	7.406	6.754	14.159
8		0.1	7.406	6.964	14.371
9		0.3	7.406	7.467	14.873
10		0.5	7.406	8.139	15.546
11		0.7	7.406	9.162	16.568
12		0.9	7.406	11.359	18.765
13	0.3	0	7.909	6.754	14.662
14		0.1	7.909	6.964	14.873
15		0.3	7.909	7.467	15.376
16		0.5	7.909	8.139	16.049
17		0.7	7.909	9.162	17.070
18		0.9	7.909	11.359	19.268
19	0.5	0	8.582	6.754	15.335
20		0.1	8.582	6.964	15.546
21		0.3	8.582	7.467	16.049
22		0.5	8.582	8.139	16.722
23		0.7	8.582	9.162	17.743
24		0.9	8.582	11.359	19.941
25	0.7	0	9.603	6.754	16.357
26		0.1	9.603	6.964	16.568
27		0.3	9.603	7.467	17.070
28		0.5	9.603	8.139	17.743
29		0.7	9.603	9.162	18.765
30		0.9	9.603	11.359	20.962
31	0.9	0	11.801	6.754	18.554
32		0.1	11.801	6.964	18.765
33		0.3	11.801	7.467	19.268
34		0.5	11.801	8.139	19.941
35		0.7	11.801	9.162	20.962
36		0.9	11.801	11.359	23.159

(a) Normal curvatures for direction γ_0 and γ_1.

[Misc. rate]		
γ_0	γ_1	$\sum_{j=0}^{1}\log(\kappa_{h\gamma_j})$
0	0	13.949
	0.1	14.159
0.1	0	14.159
	0.1	14.371
0	0.3	14.662
0.3	0	14.662
0.1	0.3	14.873
0.3	0.1	14.873
0	0.5	15.335
0.5	0	15.335
0.3	0.3	15.378
0.1	0.5	15.546
0.5	0.1	15.546
0.3	0.5	16.049
0.5	0.3	16.049
0	0.7	16.357
0.7	0	16.357
0.1	0.7	16.568
0.7	0.1	16.568
0.5	0.5	16.722
0.3	0.7	17.07
0.7	0.3	17.07
0.5	0.7	17.743
0.7	0.5	17.743
0	0.9	18.554
0.9	0	18.554
0.1	0.9	18.765
0.7	0.7	18.765
0.9	0.1	18.765
0.3	0.9	19.268
0.9	0.3	19.268
0.5	0.9	19.941
0.9	0.5	19.941
0.7	0.9	20.962
0.9	0.7	20.962
0.9	0.9	23.159

(b) Sum of normal curvatures arranged from lowest to highest.

Table 5.4 shows an extract of model estimates under a differential misclassification. We can identify significant models with unbiased estimates, to include (γ_0, γ_1)=(0,0), (0,0.1), (0.1,0), (0.1,0.1)). Meanwhile model estimates from when (γ_0, γ_1)=(0.5,0.7) are insignificant. Table 5.5*a* (left Table) displays extracts of the results of normal curvatures for each of the direction γ_0 and γ_1 and their sums under the differential measurement error mechanism of $Y_i^* = 0$. Each combination of γ_0 and γ_1 (represented under Miscl rate column) is a distinct model. Model 1 is the assumed model when $\gamma_0 = 0$ and $\gamma_1 = 0$, while the modified models are from model 2 to 36. The normal curvature at each direction helps explain which of the directions dictate a substantial change to the modification introduced into a distinct model. It can be seen that a slight change in γ_1 of a distinct model either increases or decreases in $\kappa_{h\gamma_1}$ (the normal curvature in the direction of γ_1), while $\kappa_{h\gamma_0}$ remains constant. For example, an increase in γ_1 of model 5 increases in $\kappa_{h\gamma_1}$, and it follows that a decrease in γ_1 of model 5 decreases in $\kappa_{h\gamma_1}$. In other words, $\kappa_{h\gamma_1}$; the normal curvature in the direction of γ_1 of model 5:$(\gamma_0 = 0, \gamma_1 = 0.7)$ is 9.162, it increases to 11.359 when γ_1 increases to 0.9 (model 6), but decreases to 7.467 when γ_1 decreases to 0.3 (model 3). Similarly, an increase/decrease in γ_0 of a model either increases or decreases in $\kappa_{h\gamma_0}$. In general, the results shown in Table 5.5*a* reveal that each distinct model is influenced by slight changes in the direction of γ_0 and γ_1. As a result, an increment in γ_0 and γ_1 increases normal curvature for both directions. Likewise, a decrement in γ_0 and γ_1 decreases normal curvature for both directions.

Table 5.5*b* (right Table) is the arrangement of each of the distinct models shown in Table 5.5*a*, based on their sum of normal curvatures $\sum_{j=0}^{1} \log(\kappa_{h\gamma_j})$ in ascending order. From Table 5.5*b*, when direction γ_0 alternates with direction γ_1, the sum of curvatures is equal. For example, the sum of normal curvatures; $\sum_{j=0}^{1} \log(\kappa_{h\gamma_j}) = 14.873$ for model 9 is equal to the sum of curvatures for model 14; $(\gamma_0 = 0.3, \gamma_1 = 0.1)$.

Figure 5.2(a) shows the combination of γ_0 and γ_1 in relation to model estimates. The γ_0 and γ_1 values are displayed along the X and Y axes respectively. The extreme model estimates are in the top left and bottom right corner. The range of error (γ_0, γ_1) values that produce these extremes might not be plausible in the criminology context. It can be seen that the contour lines become tightly spaced from the middle near contour line 2 (from $\gamma_1 = 0.5$) and contour line 2 (from $\gamma_0 = 0.5$). This means the model estimates changes more quickly in relation to changes in γ_0 and γ_1. Hence, the plausible error range to consider could be combinations of γ_0 and γ_1 that fall between contour lines 0 and 4.

Figure 5.2(b) shows the influence measure of the sum of normal curvatures $\sum_{j=0}^{1} \log(\kappa_{h\gamma_j})$ under a differential misclassification pattern. The plot displays the sum of curvatures in logarithmic transformation under different combinations of γ_0 and γ_1 ranging from 0 and 1. The assumed model is at the origin of the plot, otherwise the modified model.

The contour lines join models (combinations of γ_0 and γ_1) with the same sum of curvatures.

In other words, we can identify models that produce equal values of sum of curvatures using the contour lines. For instance, from Figure 5.2b, we can see that model ($\gamma_0 = 0, \gamma_1 = 0.4$) has the same $\sum_{j=0}^{1} \log(\kappa_{h\gamma_j}) = 15$ with model ($\gamma_0 = 0.4, \gamma_1 = 0$). And we can see that tighter spaced lines occur from around model ($\gamma_0 = 0.76, \gamma_1 = 0.78$), which indicate that the sum of curvatures changes more quickly with changes in γ_0 and γ_1. The contour bands indicate ranges of $\sum_{j=0}^{1} \log(\kappa_{h\gamma_j})$ in both directions of γ_0 and γ_1. Figure 5.2b shows that the highest sum of curvatures occurs near model ($\gamma_0 = 0.9, \gamma_1 = 0.9$), while the origin ($\gamma_0 = 0, \gamma_1 = 0$) has the lowest sum of curvatures.

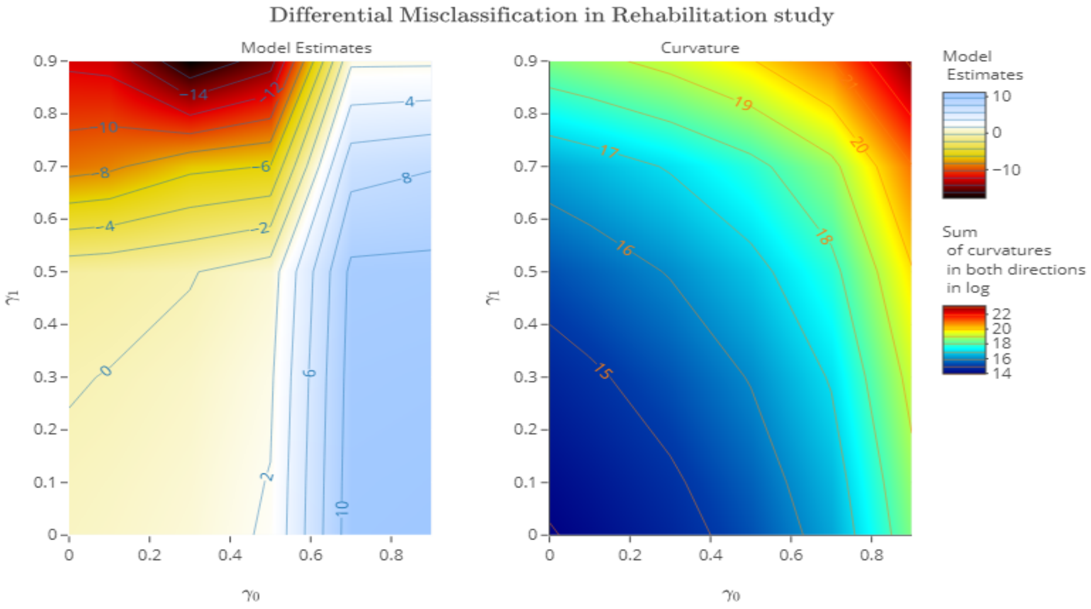


Figure 5.2: Application to rehabilitation study: Model estimates and Sum of curvatures for both directions of γ_0 and γ_1 under the influence of Differential Misclassification of $Y_i^* = 0$

As shown in Table 5.4 and Figure 5.2b, the estimates from the models that fall between contour lines 0 and 4 appear to be unbiased. Although the assumed model has the lowest sum of curvatures, it is important to investigate how the conclusion changes along these values, and report all findings regardless of prior information about γ_0 and γ_1.

5.2 Simulation I: Absence of the Outcome under Nondifferential Misclassification

In this section, we used a simulation study to illustrate model stability under the impact of a nondifferential misclassification assumption of $Y_i^* = 0$. We considered six different scenarios under nondifferential settings, with varying values of $\gamma = (0, 0.1, 0.3, 0.5, 0.7, 0.9)$. During the analysis, the probability of measurement error γ remains constant (is invariant) for the individual subject in each simulation.

This section is structured as follows. First, we provide the data generation and carry out (conducted) naive and modified analysis. The simulation study is used to demonstrate the finite sample properties of estimation under a nondifferential misclassification mechanism of $Y_i^* = 0$. Then we apply to the generated data the influence measure method under the nondifferential misclassification in Section 3.4.1 shown in Chapter Three.

5.2.1 Data Generation I

One exposure variable X is generated by distribution $X_i \sim Bin(n, 1, 0.5)$, of two sample sizes $n = 1000$ and $n = 5000$, and X is assumed to be without error. True measures of the response variable Y is generated from a logistic regression model $Y_i|X_i \sim Bin(1, \pi_i)$, with $\text{logit}(\theta_i) = \alpha + \theta X_i$, with $\alpha = -0.5$ and $\theta = 1$. Values $\alpha = -0.5$ and $\theta = 1$ are one of the possible range of parameters that could accomplish the aim of the simulation. .

Based on the assumed probability of measurement error γ, and the true measure response Y_i; which will have a misclassified response variable Y_i^*, we then simulated Y_i^* as $Y_i^* = 1$ if $Y_i = 1$, but there may be measurement error of $Y_i^* = 0$, thus $Y_i^*|Y_i = 0 \sim Bin\left(1, \pi_i + \gamma(1 - \pi_i)\right)$. Each datasets scenario is generated based on Y, Y^*, X and the assumed value of γ. In other words, the settings of the six generated datasets are based on nondifferential misclassication, thus each dataset varies.

5.2.2 Naive Analysis I

First, a naive model is applied (fitted) to each of the simulated datasets, and the extract of the summaries of results is shown in Table 5.6. The Simdata γ column represents the probability of measurement error γ present in each of the datasets, in which each row corresponds to a distinct dataset.

For each simulated dataset, we obtain the point estimate $\hat{\theta}^*$, and demonstrate the finite sample properties. We also calculate $E(\hat{\theta}^*)$; the average estimated parameter, $E(\text{Std Err}(\hat{\theta}^*))$; the average standard error, and the 95% confidence interval coverage. The estimated parameter $\hat{\theta}^*$ results are summarised over 1000 iterations with $E(\hat{\theta}^*) = \frac{1}{R}\sum_{r=1}^{R} \hat{\theta}_r^*$. The interval estimate provides a range of plausible values for the unknown true parameter θ with a degree of confidence, thus we obtained 95% confidence interval coverage so as to consider the variation in $\hat{\theta}^*$ from one run to another run, that is from one iteration to another. As a result the proportion of simulations, in which the estimated Wald-type Interval endpoints included the true value, gives the confidence interval coverage. For each dataset, and sample size $n = 1000$, $n = 5000$, each result averages out (average) across 1000 iterations. In general, it can be seen that as the probability of measurement error γ increases in a study, the effects of estimates of parameters are biased when measurement error is not considered. As γ increases, the spread of the estimated parameter tends to increase for sample size $n = 1000$ and $n = 5000$. As expected, the spread of parameter estimates for larger sample sizes $= 5000$ is smaller compared to that of $n = 1000$. The confidence interval coverage obtained for each dataset scenario in the simulation is

Table 5.6: Naive Analysis I: Extract of results from the simulation study under nondifferential misclassification of $Y_i^* = 0$. Results summarised over 1000 iterations, with $E(\hat{\theta}^*) = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r^*$, and $E(\text{Std.Err}(\hat{\theta}^*)) = \frac{1}{R}\sum_{r=1}^{R}\text{Std.Err}(\hat{\theta}_r^*)$, where R represents the number of iterations.

SIMdata		$n = 1000$			$n = 5000$	
γ	E($\hat{\theta}_r^*$)	E(Std.Err($\hat{\theta}_r^*$))	95%	E($\hat{\theta}_r^*$)	E(Std.Err($\hat{\theta}_r^*$))	95%
0	1.0006	0.1307	94.7	0.9986	0.0584	94.9
0.1	0.9062	0.1433	91	0.9051	0.064	69.7
0.3	0.7628	0.182	80.7	0.7646	0.0813	10.1
0.5	0.6623	0.2643	87.9	0.6655	0.1178	6.4
0.7	0.5947	0.4956	99.7	0.5865	0.2205	57.8
0.9	0.5396	2.2558	100	0.5291	0.9859	100

$< 95\%$, except the dataset with $\gamma = 0$ which is approximately the closest to 95% coverage for both sample sizes. Although the coverage obtained for the other dataset scenarios for sample size $N = 1000$ is better than $N = 5000$, overall the coverage is poor. Therefore, from Table 5.6, it can be seen that the assumed logistic model returns biased estimates of parameters when the potential probability of measurement error γ contained in observed data is ignored during analysis.

5.2.3 Modified Model Analysis I

Due to the biased results from the naive model, we investigate how inference can be impacted by measurement error. Hence, for each simulated dataset, with an assumed value of γ the modified estimate $\hat{\theta}|\gamma$ was calculated using (3.1.5) from Chapter Three.

Table 5.7 shows the results of the modified model under the nondifferential measurement error mechanism of $Y_i^* = 0$. The Miscl rate column indicates possible distinct models that could be applied to each of the datasets generated in column one (SIMdata). For each distinct model applied, the estimated parameter and its coverage was calculated for sample size $n = 1000$ and $n = 5000$. From Table 5.7, it can be seen that for each of the six datasets, as γ increases, the estimated parameter increases, but then breaks at $\gamma = 0.9$ when $n = 5000$, likewise the spread of the parameters increases. In general, when an incorrect model is applied during analysis, the estimate and the coverage is poor, otherwise the results look good. For example, when Miscl rate $\gamma \neq 0.3$ is applied (fitted) to dataset with SIMdata $\gamma = 0.3$, the results are biased.

5.2.4 Influence Method Analysis I

Here, we apply the influence measure under nondifferential misclassification of $Y_i = 0$ to the simulated datasets. Also, we establish the link between model stability and parameter estimation. Table 5.8 shows extract of model estimates and curvature results from the simulated datasets. The model estimates and curvature κ of the assumed model is at

Table 5.7: Modified model I: Extract of results from simulation study I under a Nondifferential Misclassification of $Y_i^* = 0$. Results are summarised over 1000 iterations, with $E(\hat{\theta}|\gamma) = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r^*$, and $E(\text{Std.Err}(\hat{\theta}_r|\gamma)) = \frac{1}{R}\sum_{r=1}^{R}\text{Std.Err}(\hat{\theta}_r|\gamma)$

SIMdata	Miscl		$n = 1000$			$n = 5000$	
	rate						
γ	γ	$\text{E}(\hat{\theta}_r\|\gamma)$	$\text{E}(\text{Std.Err}(\hat{\theta}_r\|\gamma))$	95%	$\text{E}(\hat{\theta}_r\|\gamma)$	$\text{E}(\text{Std.Err}(\hat{\theta}_r\|\gamma))$	95%
0	0.1	1.1342	0.1366	81.1	1.131	0.0609	42.5
	0.2	1.3726	0.1605	36.2	1.365	0.0711	0.4
0.1	0.1	1.0005	0.1454	95	0.9988	0.0649	94.6
	0.3	1.4569	0.316	74.5	1.4455	0.1169	0.7
	0.4	2.6878	0.949	21.4	2.4028	0.1296	0.3
0.3	0.1	0.8117	0.1749	85.8	0.8134	0.0781	29.3
	0.3	0.9979	0.1811	95.5	0.9989	0.0807	94.5
	0.4	1.2159	0.2501	93.8	1.2138	0.1086	48.4
0.5	0.1	0.6875	0.2648	86.6	0.6909	0.11	9.1
	0.3	0.7722	0.2184	87.7	0.7759	0.0974	34.6
	0.5	0.9998	0.2323	95.2	1.0029	0.1033	94.8
	0.6	1.3868	1.4326	98.2	1.3739	0.304	98.9
0.7	0.1	0.6064	0.4549	98.9	0.598	0.2025	47.1
	0.5	0.7209	0.3072	92.5	0.7105	0.1369	42.5
	0.7	1.0225	0.3247	95.2	1.0005	0.1429	96.1
	0.8	3.3528	0.822	36.1	2.5811	0.1947	2.6
0.9	0.1	0.5427	2.0431	100	0.5321	0.8932	100
	0.7	0.6127	0.7844	100	0.6013	0.3445	96.2
	0.9	1.0613	0.6557	96.8	1.008	0.2646	95.6
	0.95	-2.0424	0.8664	6.4	-2.6133	0.1498	3.3

when Miscl rates $\gamma = 0$, otherwise modified models (that is, when Miscl rates $\gamma \neq 0$). From Table 5.8a, for each dataset, the model estimates increase as the error size increases, except for when Miscl rates $\gamma = 0.95$. In addition, these estimates are not close to the assigned true parameter ($\theta = 1$) which was used as input to the simulator. But when correct model is applied, the estimate appears to be closer to 1 and unbiased. Table 5.8b shows curvature for each simulated dataset. It can be seen that as γ increases, the curvature increases.

Table 5.8: Model Estimates and Curvature results from simulation I under the nondifferential misclassification of $Y_i^* = 0$. Each SIMdata γ represents a distinct dataset where $n = 1000$, and Miscl rates γ represents potential models that could be fitted to each datasets.(a)Extract of model estimates (b)Extract of curvatures

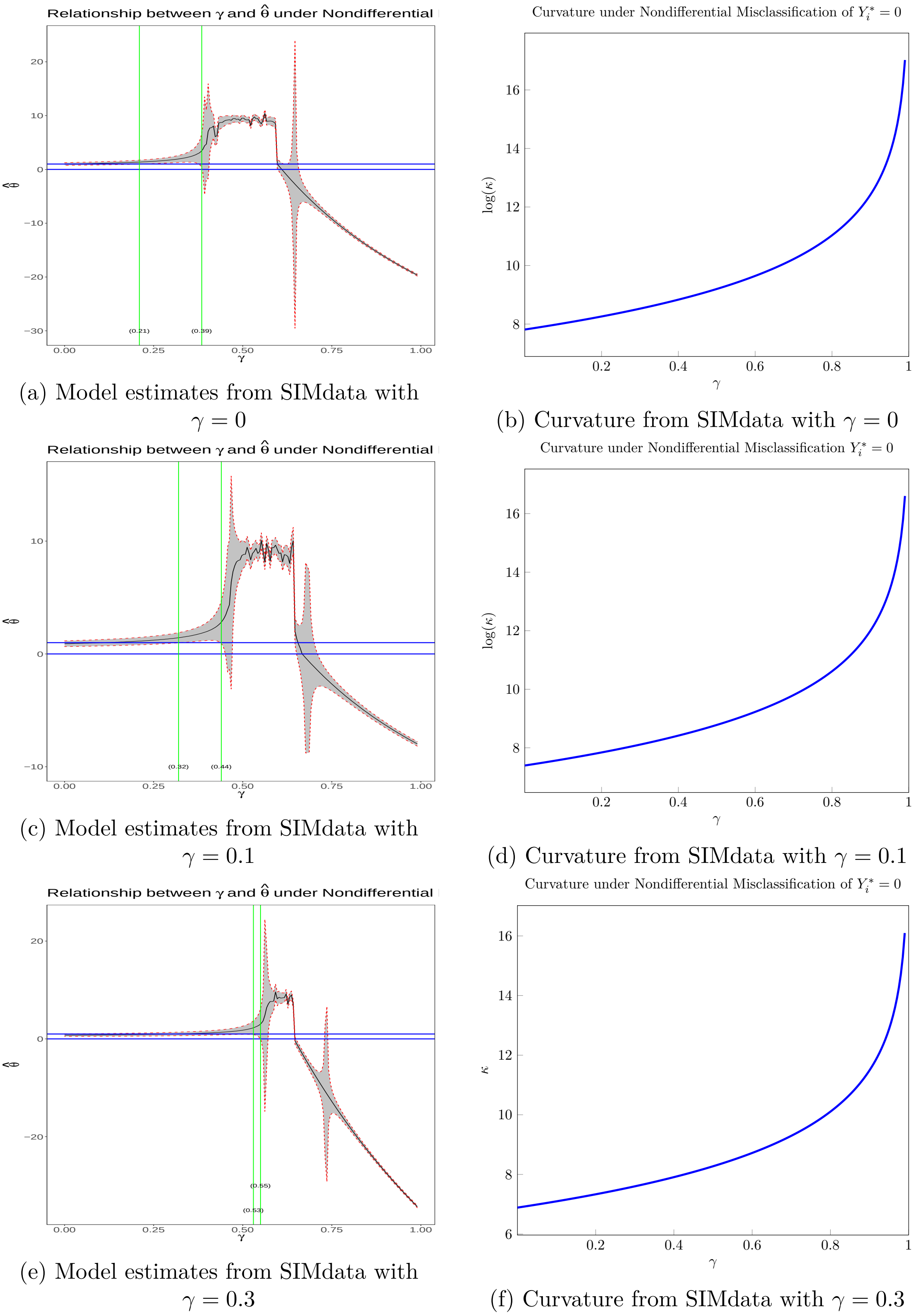
SIMdata	Miscl rates			
γ	γ	$\hat{\theta} \mid \gamma$	95%CI	PValue
0	0	1.01	(0.75,1.26)	<0.001
	0.1	1.14	(0.84,1.43)	<0.001
	0.2	1.34	(0.98,1.71)	<0.001
0.1	0	0.9	(0.65,1.16)	<0.001
	0.1	0.99	(0.71,1.28)	<0.001
	0.3	1.37	(0.95,1.8)	<0.001
	0.4	2.03	(1.19,2.87)	<0.001
0.3	0	0.76	(0.49,1.02)	<0.001
	0.1	0.81	(0.52,1.09)	<0.001
	0.3	0.99	(0.64,1.35)	<0.001
	0.4	1.21	(0.76,1.66)	<0.001
0 .5	0	0.69	(0.39,0.98)	<0.001
	0.1	0.71	(0.41,1.02)	<0.001
	0.3	0.8	(0.46,1.14)	<0.001
	0.5	1.03	(0.58,1.48)	<0.001
	0.6	1.39	(0.72,2.07)	<0.001
0 .7	0	0.59	(0.24,0.94)	<0.01
	0.1	0.6	(0.24,0.96)	<0.01
	0.5	0.72	(0.29,1.15)	<0.01
	0.7	1.04	(0.39,1.68)	<0.01
	0.8	2.89	(-1.89,7.67)	0.24
0 .9	0	0.62	(0.03,1.21)	0.04
	0.1	0.62	(0.03,1.22)	0.04
	0.7	0.71	(0.04,1.38)	0.04
	0.9	1.15	(0.02,2.27)	0.05
	0.95	-0.6	(-2.44,1.25)	0.5

(a) Model Estimates.CI;Confidence Interval.

SIMdata	Miscl rates	
γ	γ	$\log(\kappa)$
0	0	7.633
	0.1	7.854
	0.2	8.089
0.1	0	7.394
	0.1	7.615
	0.3	8.118
	0.4	8.426
0.3	0	7.03
	0.1	7.251
	0.3	7.754
	0.4	8.062
0 .5	0	6.509
	0.1	6.729
	0.3	7.232
	0.5	7.905
	0.6	8.351
0 .7	0	5.903
	0.1	6.123
	0.5	7.299
	0.7	8.32
	0.8	9.131
0.9	0	4.682
	0.1	4.892
	0.7	7.099
	0.9	9.199
	0.95	10.682

(b) Curvature under a Nondifferential Misclassification.

Figures 5.3 and 5.4 depict the model estimates and curvature from the first three (SIMdata $\gamma = 0$, SIMdata $\gamma = 0.1$, and SIMdata $\gamma = 0.3$) and the last three (SIMdata $\gamma = 0.5$, SIMdata $\gamma = 0.7$, and SIMdata $\gamma = 0.9$) datasets respectively. Each of the left graph



(a) Model estimates from SIMdata with $\gamma = 0$

(b) Curvature from SIMdata with $\gamma = 0$

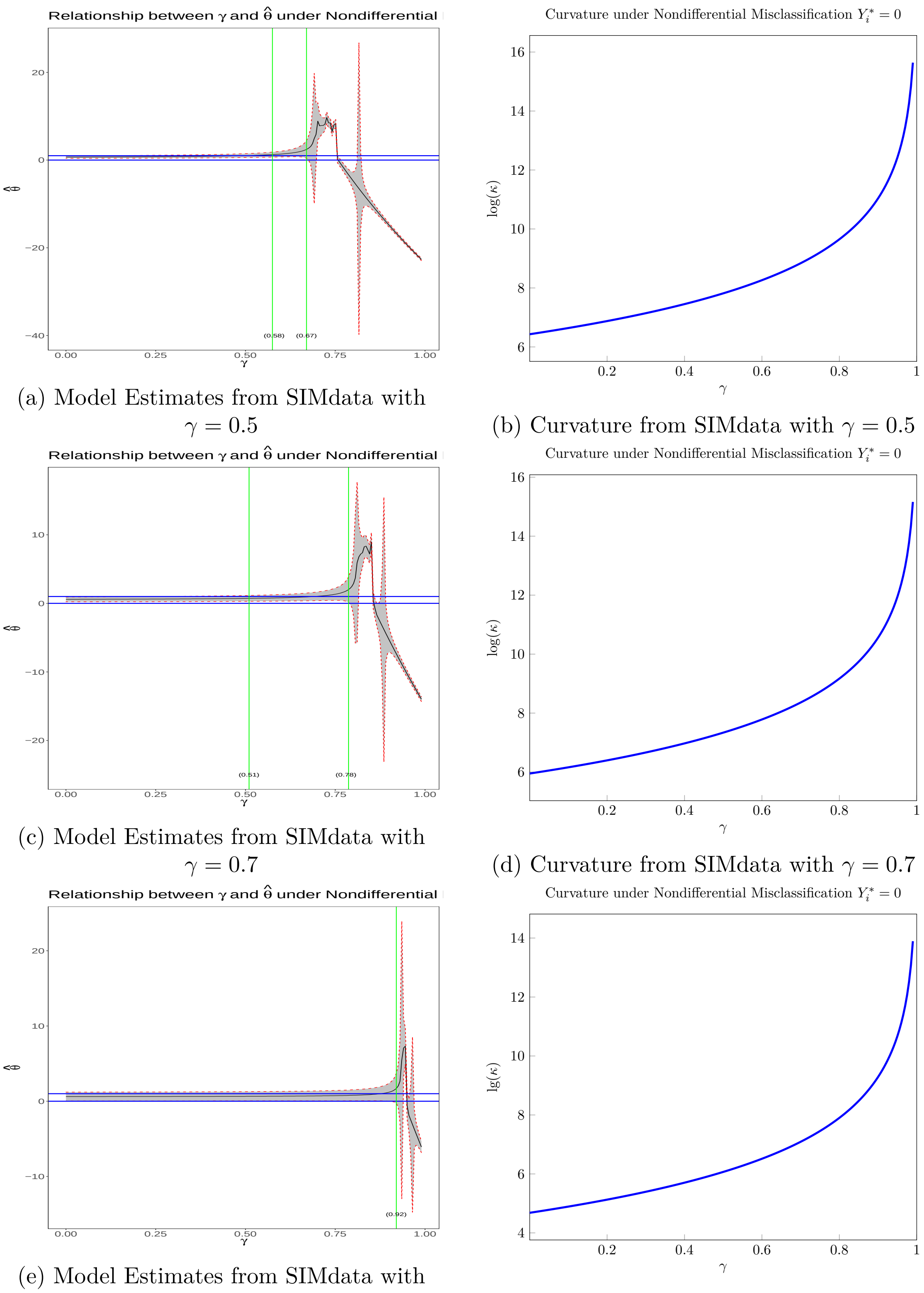
(c) Model estimates from SIMdata with $\gamma = 0.1$

(d) Curvature from SIMdata with $\gamma = 0.1$

(e) Model estimates from SIMdata with $\gamma = 0.3$

(f) Curvature from SIMdata with $\gamma = 0.3$

Figure 5.3: Simulation Ia: Model Estimates and Influence Measure under a Nondifferential Misclassification of $Y_i^* = 0$ of the first three datasets. Figures (a)-(b); Model Estimates and Curvature from SIMdata $\gamma = 0$, (c)-(d); Model Estimates and Curvature from SIMdata $\gamma = 0.1$, and (e)-(f); Model Estimates and Curvature from SIMdata $\gamma = 0.3$



(a) Model Estimates from SIMdata with $\gamma = 0.5$

(b) Curvature from SIMdata with $\gamma = 0.5$

(c) Model Estimates from SIMdata with $\gamma = 0.7$

(d) Curvature from SIMdata with $\gamma = 0.7$

(e) Model Estimates from SIMdata with $\gamma = 0.9$

(f) Curvature from SIMdata with $\gamma = 0.9$

Figure 5.4: Simulation Ib: Model Estimates and Influence Measure under a Nondifferential Misclassification of $Y_i^* = 0$ of the last three datasets. Figures (a)-(b); Model Estimates and Curvature from SIMdata $\gamma = 0.5$, (c)-(d); Model Estimates and Curvature from SIMdata $\gamma = 0.7$, and (e)-(f); Model Estimates and Curvature from SIMdata $\gamma = 0.9$

depicts the model estimates, while the right graph is the curvature. From the left graph, model estimates are represented as solid black line, and the red dashed lines are the corresponding 95%CI confidence interval. The top blue horizontal line is the assigned true parameter $\hat{\theta} = 1$ used as input to the simulator, while the bottom blue horizontal line is at when $\gamma = 0$. The first green line indicates the intersection of the top blue line and the confidence interval. And the second green line is the intersection of the bottom blue line and the confidence interval. It can be seen that there are more than one intersection points as γ increases, but there is a break point after the second green line. Hence the intersection points after the break can be ignored.

For the purpose of this simulation, the assigned true parameter is at when $\gamma = 1$. Hence, the focus is on the point at which the first green vertical line is drawn. The point suggests a cut off value for the unknown γ. In other words, the plausible range of error size could be from $\gamma = 0$ to $\gamma = 0.21$ as in the case of dataset SIMdata $\gamma = 0$, or from $\gamma = 0$ to $\gamma = 0.53$ for dataset SIMdata $\gamma = 0.3$. On the other hand, in real practice where true parameter is unknown, the cut off of the error size would be the second green line.

Each of the right graph of Figure 5.3 and 5.4 are the corresponding curvature κ (in log transformation) against γ for the six datasets. The plane curve reveals the curvature for multiple models (assumed and modified models) as γ increases from 0 to 1.

Although, Figures 5.3 and 5.4 show that the assumed model at $\gamma = 0$ has the lowest curvature amongst all the models if there are no prior information about γ, a step further would be to investigate how sensitive the conclusion and stability assessment is along those range of plausible values.

5.3 Simulation II: Absence of the Outcome under Differential Misclassification

In this section, we generate a data driven example to illustrate the use of the influence measure method under the differential misclassification described in Section 3.4.2 in Chapter Three.

In the following subsections we present data generation under the differential measurement error setting. Then we demonstrate empirical performance of an estimator based on naive and modified analyses under a differential measurement error of $Y_i^* = 0$. Most specifically, we apply the influence method to the generated datasets.

5.3.1 Data Generation II

Here, the simulation settings are under a differential misclassification of $Y_i^* = 0$. The exposure variable X is assumed to be error-free, generated as $X_i \sim Bin(n, 1, \pi_i)$ with $\pi_i = 0.5$. We simulated six different datasets with varying values of γ_0 and γ_1, each of size $n = 1000$ an $n = 5000$. Each value assigned to γ_0 and γ_1 corresponds to a distinct (assumed and modified) model. The selected (γ_0, γ_1) include $(0.1, 0.3)$, $(0.4, 0.2)$, $(0.3, 0)$, $(0.4, 0.6)$, $(0.7, 0.8)$, $(0.9.0.6)$. In other words, each simulated dataset has different combinations of γ_0 and γ_1.

Then from logit link function we simulated the true outcome Y_i as $Y_i|X_i \sim Bin(1, \pi_i)$, based on exposure variable X, with α and θ as -0.5 and 1 respectively. By assuming differential misclassification, such that when $X_i = 1$ the probability of measurement error is γ_1, otherwise γ_0, the error-prone Y_i^* version of the response variable is simulated as $Y_i^* = 1$ if $Y_i = 1$, but $Y_i^* = 0$ may be misclassified such that, $Y_i^*|Y_i = 0 \sim Bin\,(1, \pi_i + (\gamma_0 - \gamma_0 x_i + \gamma_1 x_i)(1 - \pi_i))$.

5.3.2 Naive Analysis II

For each dataset generated, we conducted naive analysis using the logit links of the logistic model. We calculated the average estimated parameter, the spread of the results and the 95% confidence interval coverage over 1000 replicates.

Table 5.9 provides the results of the naive analysis for the six simulated datasets under simulation II. The combination of γ_0 and γ_1 under Simdata are the six datasets. It can be seen that when measurement error is ignored, the estimated parameter appears to be biased except for model $(0.7, 0.8)$ whose $\hat{\theta}^*$ and coverage are approximately close to 1 and 95%, but overall the estimates and coverage for both sample sizes are poor.

Table 5.9: Extract of naive analysis II from the simulation study II under a Differential Misclassification of $Y_i^* = 0$: Results are summarised over 1000 iterations, with $E(\hat{\theta}^*) = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r^*$, and $E(\text{Std.Err}(\hat{\theta}_r^*)) = \frac{1}{R}\sum_{r=1}^{R}\text{Std.Err}(\hat{\theta}^*)$

SIMdata			$n = 1000$			$n = 5000$	
γ_0	γ_1	E($\hat{\theta}_r^*$)	E(Std.Err($\hat{\theta}_r^*$))	95%	E($\hat{\theta}_r^*$)	E(Std.Err($\hat{\theta}_r^*$))	95%
0.1	0.3	1.2672	0.1359	48.9	1.2658	0.0607	1.2
0.4	0.2	0.3240	0.1345	0.2	0.3206	0.0601	0
0.3	0	0.2376	0.1292	0	0.2404	0.0577	0
0.4	0.6	1.2145	0.1558	73.5	1.2095	0.0695	14
0.7	0.8	1.0343	0.2062	95.6	1.0343	0.0917	94.6
0.9	0.6	-0.9977	0.2255	0	-0.9933	0.1002	0

Again, in the case of the differential misclassification of $Y_i^* = 0$, the conclusions from naive analysis could be misleading as measurement error was ignored. Therefore, we investigate how inference can be impacted by slight changes to γ_0 and γ_1.

5.3.3 Modified Model Analysis II

We evaluated the estimate of the modified models under the assumption of differential misclassifiation of $Y_i^* = 0$ for each of the generated datasets. Table 5.10 shows an extract of the results (estimated coefficient with its standard error and 95% coverage) summarised

Table 5.10: Extract of modified analysis from simulation II under a Differential Misclassification $Y_i^* = 0$: Results are summarised over 1000 iterations, with $E(\hat{\theta}|\gamma_0, \gamma_1) = \frac{1}{R}\sum_{r=1}^{R} \hat{\theta}_r|\gamma_0, \gamma_1$, and $E(\text{Std.Err}(\hat{\theta}_r|\gamma_0, \gamma_1)) = \frac{1}{R}\sum_{r=1}^{R} \text{Std.Err}(\hat{\theta}|\gamma_0, \gamma_1)$

SIMdata		Miscl rates		$n = 1000$			$n = 5000$		
					E(Std.Err)			E(Std.Err)	
γ_0	γ_1	γ_0	γ_1	E($\hat{\theta}_r\|\gamma_0, \gamma_1$)	($\hat{\theta}_r\|\gamma_0, \gamma_1$)	95%	E($\hat{\theta}_r\|\gamma_0, \gamma_1$)	($\hat{\theta}_r\|\gamma_0, \gamma_1$)	95%
0.1	0.3								
		0.1	0.2	1.2083	0.1535	73.8	1.2066	0.0685	14.9
		0.1	0.3	1.0012	0.1596	95.3	0.9999	0.0713	94
		0.2	0.4	1.0904	0.1887	92.3	1.0879	0.0840	82.7
0.4	0.2								
		0.2	0.3	0.1469	0.1620	0	0.1433	0.0723	0
		0.4	0.2	1.0101	0.1898	95.1	1.0011	0.0843	94.7
		0.4	0.3	0.7853	0.1960	79.4	0.7768	0.0871	26.9
0.3	0								
		0	0.1	0.0621	0.1341	0	0.0653	0.0599	0
		0.3	0	0.9975	0.1639	95.5	0.9988	0.0731	94.4
		0.2	0.3	0.0157	0.1680	0	0.02	0.0749	0
0.4	0.6								
		0.5	0.1	2.7108	0.2696	0	2.6871	0.1177	0
		0.4	0.6	1.0089	0.2310	95	1.0004	0.1027	95.7
		0.3	0.5	0.9786	0.1970	94.9	0.9722	0.0878	93.9
0.7	0.8								
		0.5	0.4	1.4206	0.2346	57.7	1.4212	0.1043	2
		0.7	0.8	0.9995	0.3577	95.3	1.0003	0.1580	95.8
		0.1	0.4	0.5983	0.2160	52.5	0.5987	0.0960	1.5
0.9	0.6								
		0.2	0.1	-0.8833	0.2295	0	-0.8789	0.1020	0
		0.9	0.6	1.0306	0.5340	96.5	0.9937	0.2213	95.2
		0.4	0.3	-0.8780	0.2384	0	-0.8735	0.1060	0

over 1000 replicates for each dataset. The combination of γ_0 and γ_1 under Miscl rates are some of the distinct models, while the remaining columns are the results of the two sample sizes. It appears that a model where Miscl rates $\gamma_0 <$ Miscl rates γ_1, the estimate and the spread are less. If otherwise, the estimate and spread are greater. For example, for dataset $(0.7, 0.8)$, the parameter estimate from the modified model $(0.1, 0.4)$ is less compared to the modified model $(0.5, 0.4)$. Furthermore, when the correct model is applied, the estimates and coverage appear to be good. Overall, from Table 5.10, it can be seen that when an incorrect model is applied, the estimate is biased and the coverage is poor.

5.3.4 Influence Method Analysis II

We now consider influence measure based on the sum of normal curvatures. To demonstrate how stable or sensitive a distinct model is when there is a slight modification to (γ_0, γ_1), we calculated model estimates and normal curvature in the direction of γ_0 and γ_1 using the equation in Section 3.4.2 of Chapter Three.

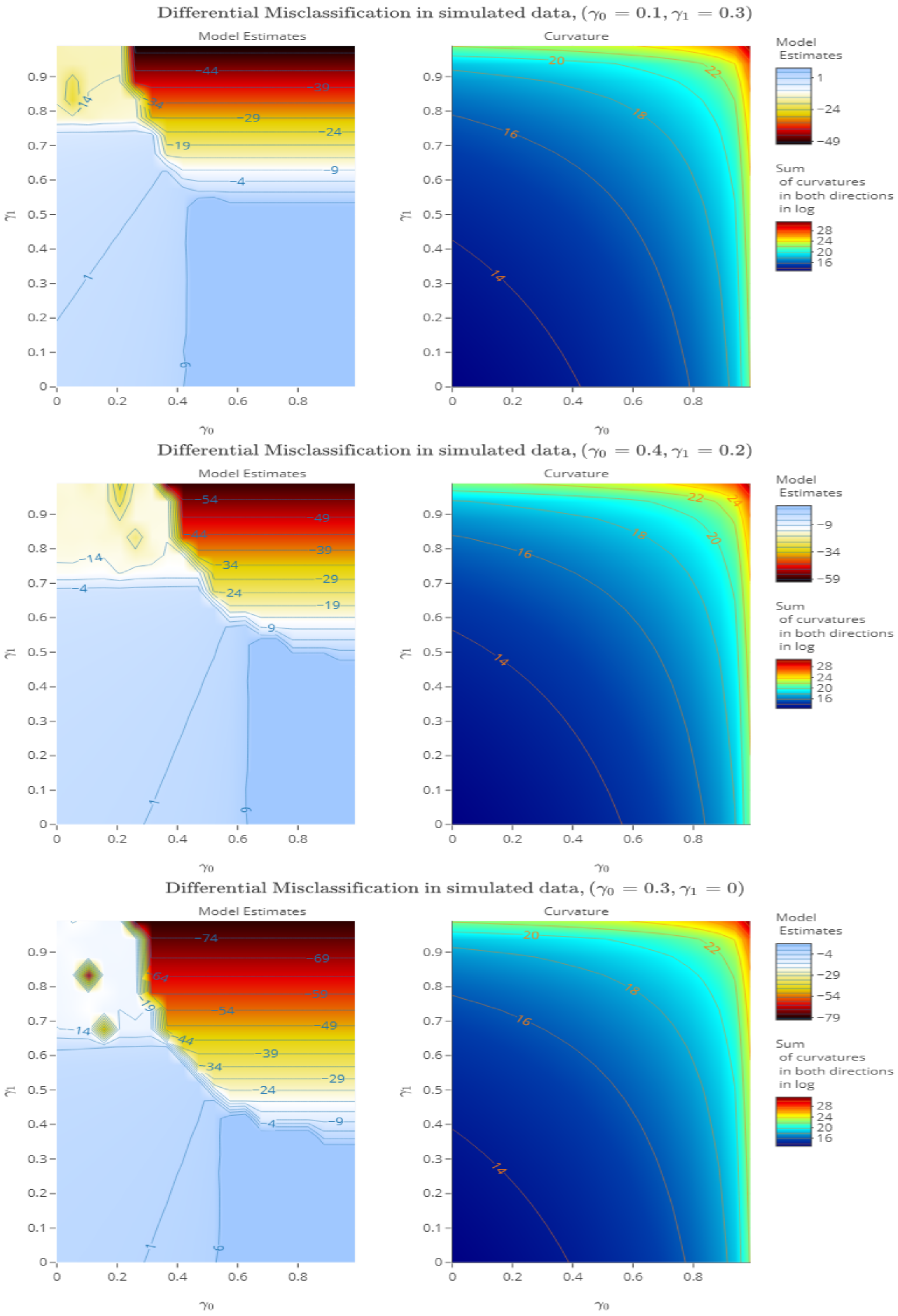
Table 5.11 shows an extract of model estimates and curvatures under a differential misclassification of $Y_i^* = 0$. We can identify significant models with unbiased estimates. It can be seen that when a correct model is fitted(applied), the estimate seems to be unbiased and significant. Also, when model Miscl rates $(\gamma_0 = 0.3, \gamma_1 = 0.5)$ is fitted to SIMdata $(\gamma_0 = 0.4, \gamma_1 = 0.6)$, the estimate seems to be unbiased. Similarly, the assumed model estimate of SIMdata $(\gamma_0 = 0.4, \gamma_1 = 0.6)$ is unbiased. In addition, Table 5.11 depicts results of normal curvatures in log transformation for each of the direction γ_0 and γ_1 and their sums. In general, the assumed model has the smallest values compared to other models, under the differential misclassification pattern of $Y_i^* = 0$.

Figure 5.5 and 5.6 depict the model estimates and sum of curvatures for both directions γ_0 and γ_1 under a differential misclassifiation of $Y_i^* = 0$. The left graph is the model estimates while the right graph is the sum of curvatures. Each contour line joins model estimates of the same value. Hence, we can identify models with estimated parameter 1 using contour line 1. These combinations (contour line 1) of γ_0, γ_1 might be the likely value of error size to consider in the analysis. In general, the contour lines in the right graphs become tightly spaced from $(\gamma_0 = 0.9, \gamma_1 = 0.9)$ which indicates that the sum of curvature changes more quickly. For each simulated data, the assumed model is at the origin. It follows that the smallest curvature occurs at the origin $(\gamma_0 = 0, \gamma_1 = 0)$, while the highest curvature is near $(\gamma_0 = 1, \gamma_1 = 1)$.

Although the assumed model has the lowest curvature, when measurement error is suspected, the recommendation would be to validate or find good information about the error size. If the conclusion does not change much across a range of plausible values γ_0 and γ_1, then it is robust. On the other hand, if the conclusion changes as the error size changes and the curvature of the curve are so much, thereby the model gives a very different result, it is important to report the findings. In other words, regardless of prior information of the error size, a further step would be to investigate how sensitive the conclusion and stability assessment is along the error values.

Table 5.11: Extract of results of Model Estimates and Sum of Curvatures under a Differential Misclassification from six simulation study. The six datasets selected are $(0.1, 0.3), (0.4, 0.2), (0.3, 0), (0.4, 0.6), (0.7, 0.8), (0.9, 0.6)$

SIMdata		Miscl rates		Estimates			Curvature		
γ_0	γ_1	γ_0	γ_1	$\hat{\theta}\|\gamma_0, \gamma_1$	95%CI	Pval	$\log(\kappa_{h\gamma_0})$	$\log(\kappa_{h\gamma_1})$	$\sum_{j=0}^{1} \log(\kappa_{h_{\gamma_j}})$
0.1	0.3	0	0	1.29	(1.02,1.56)	<0.001	7.099	5.79	12.889
		0.1	0.2	1.23	(0.93,1.53)	<0.001	7.309	6.237	13.546
		0.1	0.3	1.03	(0.72,1.34)	<0.001	7.309	6.504	13.813
		0.2	0.4	1.11	(0.75,1.48)	<0.001	7.545	6.812	14.357
0.4	0.2	0	0	0.39	(0.12,0.65)	0.004	6.371	5.966	12.337
		0.2	0.3	0.23	(-0.09,0.54)	0.16	6.817	6.679	13.497
		0.4	0.2	1.06	(0.70,1.43)	<0.001	7.393	6.679	14.072
		0.4	0.3	0.85	(0.47,1.22)	<0.001	7.393	6.679	14.072
0.3	0	0	0	0.28	(0.02,0.53)	0.03	6.480	6.302	12.782
		0	0.1	0.10	(-0.16,0.37)	0.4	6.480	6.513	12.993
		0.3	0	1.03	(0.71,1.36)	<0.001	7.194	6.302	13.496
		0.2	0.3	0.39	(0.03,0.75)	0.03	6.927	7.015	13.942
0.4	0.6	0	0	1.26	(0.95,1.57)	<0.001	6.371	5.094	11.465
		0.5	0.1	2.70	(2.20,3.20)	<0.001	7.757	5.305	13.062
		0.4	0.6	1.06	(0.61,1.51)	<0.001	7.393	6.926	14.319
		0.3	0.5	1.03	(0.64,1.42)	<0.001	7.084	6.480	13.564
0.7	0.8	0	0	1.06	(0.65,1.47)	<0.001	5.429	4.352	9.782
		0.5	0.4	1.45	(0.99,1.91)	<0.001	6.816	5.374	12.189
		0.7	0.8	1.03	(0.35,1.72)	0.003	7.838	7.571	15.409
		0.1	0.4	0.63	(0.20,1.05)	0.004	5.641	5.374	11.014
0.9	0.6	0	0	-0.99	(-1.44,-0.55)	<0.001	4.083	5.094	9.176
		0.2	0.1	-0.88	(-1.33,-0.43)	<0.001	4.529	5.305	9.833
		0.9	0.6	0.96	(0.02,1.9)	0.05	8.688	6.926	15.614
		0.4	0.3	-0.87	(-1.34,-0.41)	<0.001	5.104	5.807	10.911



Differential Misclassification in simulated data, $(\gamma_0 = 0.1, \gamma_1 = 0.3)$

Differential Misclassification in simulated data, $(\gamma_0 = 0.4, \gamma_1 = 0.2)$

Differential Misclassification in simulated data, $(\gamma_0 = 0.3, \gamma_1 = 0)$

Figure 5.5: Simulation IIa: Model Estimates and Sum of Curvatures for both directions γ_0 and γ_1 under Differential Misclassification of $Y_i^* = 0$ of the first three datasets. First row; Model Estimates and Sum of Curvatures from SIMdata $(\gamma_0 = 0.1, \gamma_1 = 0.3)$, Middle row; Model Estimates and Sum of Curvatures from SIMdata $(\gamma_0 = 0.4, \gamma_1 = 0.2)$, and Last row; Model Estimates and Sum of Curvatures from SIMdata $(\gamma_0 = 0.3, \gamma_1 = 0)$

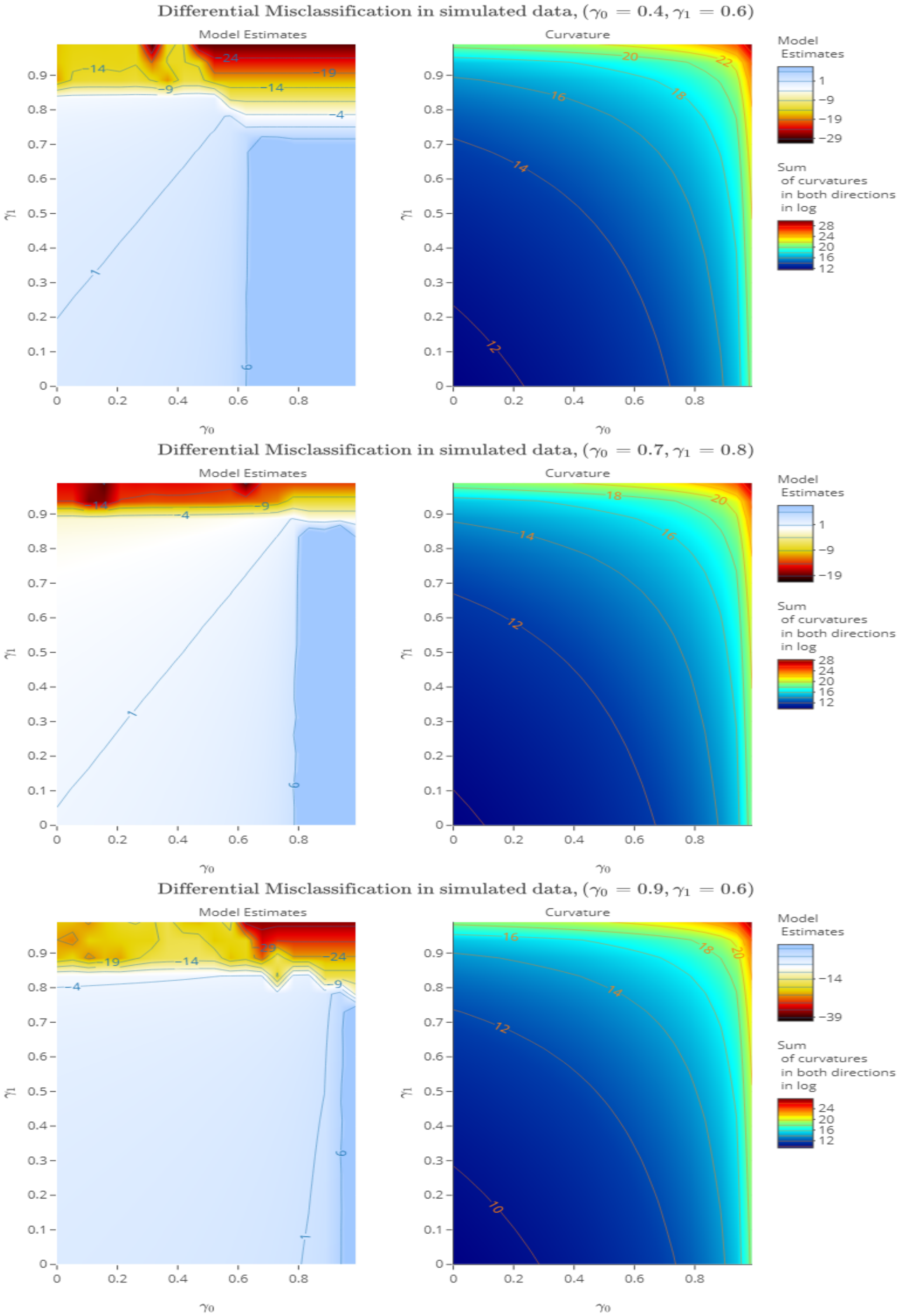


Figure 5.6: Simulation IIb: Model Estimates and Sum of Curvatures for both directions γ_0 and γ_1 under a Differential Misclassification of $Y_i^* = 0$ of the last three datasets.First row; Model Estimates and Sum of Curvatures from SIMdata ($\gamma_0 = 0.4, \gamma_1 = 0.6$), Middle row; Model Estimates and Sum of Curvatures from SIMdata ($\gamma_0 = 0.7, \gamma_1 = 0.8$), and Last row; Model Estimates and Sum of Curvatures from SIMdata ($\gamma_0 = 0.9, \gamma_1 = 0.6$)

5.4 Simulation III: Presence of the Outcome under Nondifferential Misclassification

In this section, we considered a simulation study to demonstrate the influence measure under the nondifferential measurement error mechanism described in Section 4.2.1 in Chapter Four. By assuming that the observed response $Y_i^* = 1$ is nondifferentially misclassified, we used simulation to assess the finite sample properties of both the naive and modified models. In addition, from the simulation we calculate model estimates and curvature and show the influence plot under nondifferential misclassification of $Y_i^* = 1$. The rest of this section is as follows. In Section 5.4.1, we describe how the data is generated. A naive analysis was conducted in Section 5.4.2. In Section 5.4.3, we demonstrate estimation based on modified model analysis. The influence measure under the impact of nondifferential misclassification of $Y_i^* = 1$ is illustrated in Section 5.4.4.

5.4.1 Data Generation III

We simulated an exposure indicator variable X_i by distribution $X_i \sim Bin(n, 1, 0.5)$, for n number of participants. The exposure variable generated is assumed to be true measurements, and we considered two sample sizes $n = 1000$, and $n = 5000$. True measures of the response variable Y_i were generated from the logistic regression model with $\alpha = -0.5$ and $\theta = 1$. Based on the assumed probability of measurement error ζ, and the true measure response Y_i, a response variable measured with error Y_i^* was generated as $Y_i^* = 0$ if $Y_i = 0$, but there may be a measurement error of $Y_i^* = 1$, thus $Y_i^*|Y_i = 1 \sim Bin(1, (1-\zeta)\pi_i)$.

The simulation was repeated for six different datasets with varying values of ζ. The assumed values selected are $(\zeta = 0, 0.2, 0.4, 0.6, 0.8, 0.9)$. It follows that each simulation is based on Y, Y^*, X and an assumed value of the probability of measurement error ζ. For instance, simulated data with $\zeta = 0.2$ is assumed to have 80% sensitivity, while simulated data with $\zeta = 0.6$ has 40% sensitivity and so on. As a result, each generated result varies. The analysis conducted for each simulated scenario is described in the following sections.

5.4.2 Naive Analysis III

Here, we show the summaries of results obtained from the assumed model in Table 5.12. The Simdata ζ column contains selected values of ζ. In other words, each row corresponds to a distinct dataset with an assumed probability of measurement error of $Y_i^* = 1$. We performed naive analysis and a study of the finite properties of the estimates for each of the six datasets. The estimated parameter $\hat{\theta}^*$ with its standard error and 95% are summarised over 1000 iterations.

From Table 5.12 it can be seen that, in general, the estimate of parameter decreases as ζ increases. The naive model produces biased estimates since measurement error is not

Table 5.12: naive Analysis II: Extract of results from simulation III under Nondifferential Misclassification of $Y_i^* = 1$: Results are summarised over 1000. iterations

SIMdata	$n = 1000$			$n = 5000$		
ζ	$\mathrm{E}(\hat{\theta}_r^*)$	$\mathrm{E}(\mathrm{Std.Err}(\hat{\theta}_r^*))$	95%	$\mathrm{E}(\hat{\theta}_r^*)$	$\mathrm{E}(\mathrm{Std.Err}(\hat{\theta}_r^*))$	95%
0	1.001	0.1307	94.7	0.9986	0.0584	94.9
0.2	0.8309	0.1325	76.1	0.8292	0.0592	17.2
0.4	0.7156	0.1417	48.9	0.7108	0.0632	0.2
0.6	0.6199	0.1628	0.36	0.6202	0.0725	0
0.8	0.5597	0.2191	45.9	0.5516	0.097	0.6
0.95	0.5457	0.4382	77.5	0.5116	0.188	26.2

considered. As ζ increases, the spread of the estimated parameter tends to increase for both sample sizes. As expected, the spread of parameter estimates for the larger sample size $n = 5000$ is smaller compared to $n = 1000$. The confidence interval coverage obtained for each dataset in the simulation is $< 95\%$, except the dataset with $\zeta = 0$ which is approximately the closest to 95% coverage for both sample sizes. Although the coverage obtained for the other dataset scenarios for sample size $n = 1000$ is better than $n = 5000$, the overall coverage is poor. Hence, any conclusion based on point estimate $\hat{\theta}^*$ might be misleading, given that the problem of measurement error was not considered. Therefore, for each dataset, we assess the impact of measurement error by analysing the modified model under nondifferential pattern.

5.4.3 Modified Model Analysis III

Here the modified model follows a nondifferential misclassification pattern of $Y_i^* = 1$. Thus, for each simulated data, by assuming values fr ζ, an estimate of parameters of the modified model $\hat{\theta}|\zeta$ was calculated using Section 4.2 of Chapter Four.

Table 5.13 shows results of the modified model under the non-differential measurement error of $Y_i^* = 1$. The first column (SIMdata ζ) represents six different datasets scenarios assumed to be under the problem of measurement error of $Y_i^* = 1$. In other words, each simulated dataset is based on each assumed value of ζ.

The Miscl rate column is for different models that could be applied to any of the datasets. For each distinct applied model the estimated parameters and coverage were calculated for both sample sizes $n = 1000$ and $n = 5000$. From Table 5.13, it can be seen that for each of the datasets, as ζ increases the estimated parameter increases, likewise the spread of the parameter increases. For example, for the dataset with 80% sensitivity, that is $\zeta = 0.2$, when an incorrect model ($\zeta \neq 0.2$) is applied, the estimated parameter and coverage is poor. When the correct model is applied, the model estimate 1.0027 is close to $\theta = 1$ with 95.7% coverage. In general, when an incorrect model is applied during analysis, the estimated parameter with its coverage is poor, otherwise the results look good.

Table 5.13: Extract of modified model results from simulation study under a Nondifferential Misclassification of $Y_i^* = 1$. Results are summarised over 1000.

SIMdata	Miscl rate		$n = 1000$			$n = 5000$	
ζ	ζ	$\text{E}(\hat{\theta}_r\|\zeta)$	$\text{E}(\text{Std.Err}(\hat{\theta}_r\|\zeta))$	95%	$\text{E}(\hat{\theta}_r\|\zeta)$	$\text{E}(\text{Std.Err}(\hat{\theta}_r\|\zeta))$	95%
0	0.1	1.134	0.1505	85.8	1.1314	0.0672	50.3
	0.2	1.372	0.1924	50.9	1.3664	0.0855	1.1
0.2	0.1	0.8988	0.1437	88.8	0.8967	0.0642	62.8
	0.2	1.0027	0.162	95.7	0.9998	0.0723	95.7
	0.3	1.1835	0.1976	86.7	1.1784	0.0879	45.6
0.4	0.1	0.7514	0.1487	63.1	0.7463	0.0664	3.8
	0.3	0.8783	0.1748	88.4	0.8719	0.078	63.3
	0.4	1.0093	0.2042	94.9	1.007	0.0908	95
0.6	0.1	0.6375	0.1673	42.6	0.6376	0.0745	0.2
	0.3	0.6935	0.1818	59	0.6939	0.081	4.5
	0.5	0.8265	0.2178	86.3	0.8268	0.097	57.1
	0.6	0.9993	0.2698	94.9	0.9974	0.1197	93.6
0.8	0.1	0.5664	0.2217	48.4	0.5584	0.0982	0.8
	0.5	0.6276	0.2449	65.4	0.6195	0.1086	7.4
	0.8	1.0156	0.4127	95.7	0.9982	0.1809	94.9
0.95	0.1	0.5472	0.4392	77.7	0.5131	0.1885	26.6
	0.7	0.5781	0.4626	81.6	0.5435	0.1992	36.7
	0.9	0.7016	0.5579	89.7	0.6628	0.2419	70
	0.95	1.2249	4.6891	97.4	0.1977	0.3748	39.2

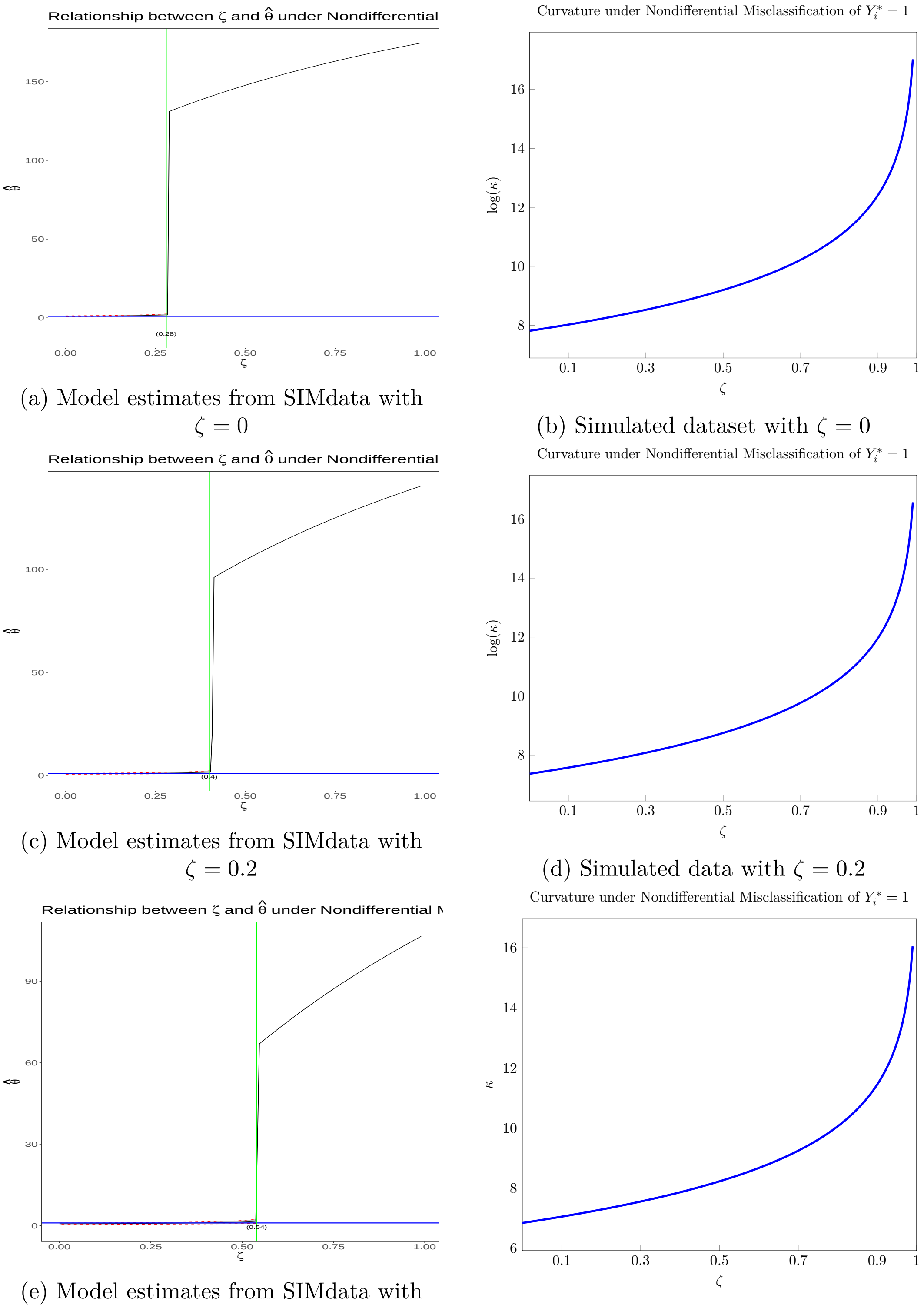
5.4.4 Influence Method Analysis III

In this section, we show model estimates and curvature results in Table 5.14, and an influence graph of curvature against ζ for the case of a nondifferential misclassification of $Y_i^* = 1$.

Table 5.14: Influence Method Analysis III: Extract of Model Estimates and Curvature results of simulation III under a Nondifferential Misclassification of $Y_i^* = 1$.

SIMdata	Miscl rate		Estimates		Curvature
ζ	ζ	$\hat{\theta}\vert\zeta_0,\zeta_1$	95%CI	Pval	$\log(\kappa)$
0	0	1.04	(0.78,1.3)	<0.001	7.813
	0.1	1.18	(0.89,1.48)	<0.001	8.034
	0.2	1.44	(1.05 ,1.83)	<0.001	8.184
0.2	0	0.84	(0.59, 1.1)	<0.001	7.362
	0.1	0.92	(0.63, 1.2)	<0.001	7.583
	0.2	1.03	(0.71 1.35)	<0.001	7.737
	0.3	1.22	(0.83, 1.62)	<0.001	8.005
0.4	0	0.71	(0.44, 0.99)	<0.001	6.838
	0.1	0.75	(0.46, 1.04)	<0.001	7.059
	0.3	0.89	(0.54, 1.23)	<0.001	7.548
	0.4	1.03	(0.62, 1.44)	<0.001	7.856
0.6	0	0.65	(0.32, 0.98)	<0.001	6.296
	0.1	0.67	(0.33, 1)	<0.001	6.517
	0.3	0.72	(0.36, 1.08)	<0.001	6.850
	0.5	0.84	(0.42, 1.27)	<0.001	7.692
	0.6	0.99	(0.48, 1.5)	<0.001	7.97
0.8	0	0.51	(0.09,0.92)	0.02	5.518
	0.1	0.51	(0.09 ,0.93)	0.02	5.739
	0.5	0.57	(0.1, 1.04)	0.02	6.914
	0.8	0.95	(0.15, 1.75)	0.02	8.683
0 .95	0	0.36	(0.04, 0.67)	0.03	4.055
	0.1	0.36	(0.04, 0.68)	0.03	4.276
	0.7	0.39	(0.05, 0.73)	0.03	8.242
	0.9	0.52	(0.07, 0.97)	0.03	8.669
	0.95	1.04	(0.04, 2.05)	0.04	11.825

Table 5.14 shows an extract of model estimates and curvature results for the six simulated datasets. The curvature results are in log transformation. Each value of Miscl rate ζ column is a distinct model (assumed and modified models). The model estimates and the curvature κ of the assumed model is at Miscl rates $\zeta = 0$, otherwise modified models when Miscl rates $\zeta \neq 0$. From Table 5.14, for each dataset, the model estimates increase as the error size increases. It can be seen that when correct model is applied, the estimate appears



(a) Model estimates from SIMdata with $\zeta = 0$

(b) Simulated dataset with $\zeta = 0$

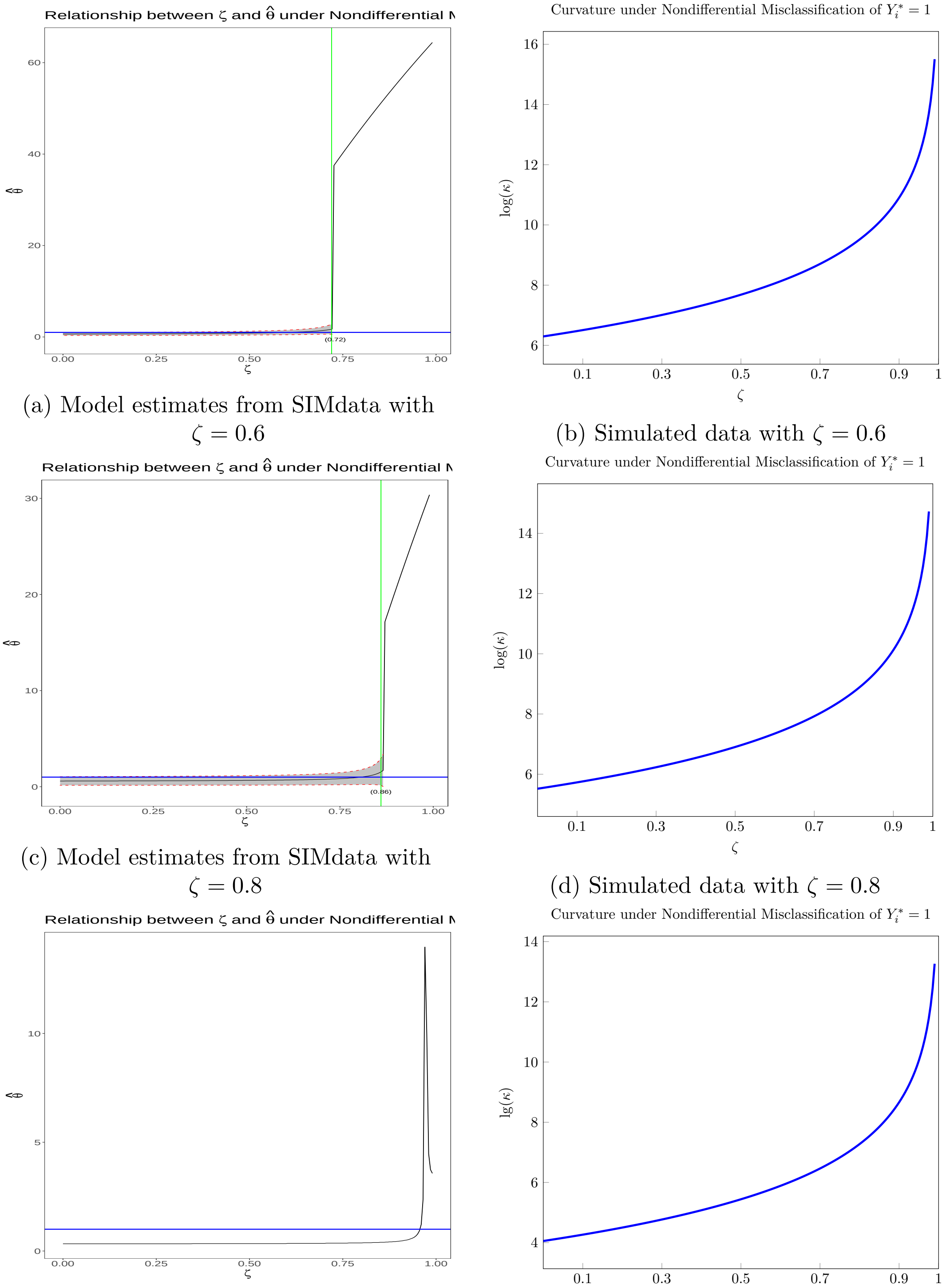
(c) Model estimates from SIMdata with $\zeta = 0.2$

(d) Simulated data with $\zeta = 0.2$

(e) Model estimates from SIMdata with $\zeta = 0.4$

(f) Simulated data with $\zeta = 0.4$

Figure 5.7: Simulation IIIa: Model Estimates and Curvature under Nondifferential Misclassification of $Y_i^* = 1$. Figure (a), (c) and (e) are the model estimates from SIMdata $\zeta = 0$, SIMdata $\zeta = 0.2$, and SIMdata $\zeta = 0.4$ respectively. Figure (b), (d) and (f) are the corresponding curvature



(a) Model estimates from SIMdata with $\zeta = 0.6$

(b) Simulated data with $\zeta = 0.6$

(c) Model estimates from SIMdata with $\zeta = 0.8$

(d) Simulated data with $\zeta = 0.8$

(e) Model estimates from SIMdata with $\zeta = 0.95$

(f) Simulated dataset with $\zeta = 0.95$

Figure 5.8: Simulation IIIb: Model Estimates and Curvature under Nondifferential Misclassification of $Y_i^* = 1$. Figure (a), (c) and (e) are the model estimates from SIMdata $\zeta = 0.6$, SIMdata $\zeta = 0.8$, and SIMdata $\zeta = 0.95$ respectively. Figure (b), (d) and (f) are the corresponding curvature

to be closer to 1 and unbiased. Again, the curvature increases, as ζ increases,.

Figures 5.7 and 5.8 depict the model estimates and curvature from the first three (SIMdata $\zeta = 0$, SIMdata $\zeta = 0.2$, and SIMdata $\zeta = 0.4$) and the last three (SIMdata $\zeta = 0.6$, SIMdata $\zeta = 0.8$, and SIMdata $\zeta = 0.95$) datasets respectively. Each of the left graphs depicts the model estimates, while the right graph is the curvature. Again, from the left graph, it can be seen that there are more than one intersection points as ζ increases, but there is a break point after the green line. Here, the plausible range of error size could be from $\zeta = 0$ to the break point. For instance, the likely range of error size for dataset SIMdata $\zeta = 0.2$ is from $\zeta = 0$ to $\zeta = 0.4$. Except for SIMdata $\zeta = 0.95$ which suggests that extreme cases may be impossible.

Each of the right graph of Figure 5.7 and 5.8 shows the influence graph of curvature for the six simulated datasets under a nondifferential misclassfication of $Y_i^* = 1$. The plane curve reveals the curvature for multiple models (assumed and modified models) as ζ increases from 0 1, where each point on the curve is the curvature of a distinct model.

Again, it is important to investigate how sensitive the conclusion and stability assessment is along that range of plausible values of ζ, and report all the findings.

5.5 Simulation IV: Presence of the Outcome under Differential Misclassification

In this section, based on simulation, we provide empirical results of the influence measure under the differential misclassification pattern investigated in Chapter Four. This section is structured as follows. In section 5.5.1, we describe data driven generation. A naive analysis was conducted in Section 5.5.2 with its finite sample properties. In Section 5.5.3, we assess the performance of an estimation based on modified model analysis. In Section 5.5.4 we show the influence surface of the sum of curvatures in the direction of ζ_0 and ζ_1.

5.5.1 Data Generation IV

In this section, a complete dataset is simulated based on the assumption of a differential measurement error mechanism of $Y_i^* = 1$. The exposure variable X_i is generated by distribution $X_i \sim Bin(1, 0.5)$, and X_i is assumed to be without error. The true response variable Y_i is generated from logistic model discussed in Chapter Four, where $\alpha =?0.5$ and $\theta = 1$. The misclassified response variable Y_i^* is generated as $Y_i^* = 0$ if $Y_i = 0$, but $Y_i^* = 1$ may be measured with error such that $Y_i^*|Y_i = 1 \sim Bin(1, (\zeta_0 - \zeta_0 x_i + \zeta_1 x_i)\pi_i)$. We considered two sample sizes, $n = 1000$ and $n = 5000$, and by assuming the probability of measurement error to be ζ_0 and ζ_1, six different combinations of (ζ_0, ζ_1) were selected to initiate differential misclassification mechanisms in the dataset. The six chosen (ζ_0, ζ_1) are $(0.1, 0.3)$, $(0.4, 0.2)$, $(0.3, 0)$, $(0.4, 0.6)$, $(0.7, 0.8)$, and $(0.9, 0.6)$. Thus, a dataset with

$(\zeta_0, \zeta_i) = (0.1, 0.3)$ would have 70% sensitivity conditional on X_1, and 90% sensitivity conditional on X_0. Hence, each generated dataset varied based on the assumed value of (ζ_0, ζ_i). The analysis conducted on each simulated dataset is presented in the following sections.

5.5.2 Naive Analysis IV

For each generated dataset with a combination of ζ_0 and ζ_1, we performed naive analysis using the logit links of logistic model from Chapter Four. We calculated the average estimated parameter, the spread of the results and the 95% confidence interval coverage over 1000 replicates for each dataset.

Table 5.15 shows the results of the naive model for each of the datasets. It can be seen that when measurement error is ignored, the estimated parameter appears to be biased, and the coverage for both sample sizes are very poor.

Table 5.15: Naive Analysis IV: Extract of naive analysis from the simulation study IV under a Differential Misclassification of $Y_i^* = 1$: Results are summarised over 1000.

SIMdata			$n = 1000$			$n = 5000$	
ζ_0	ζ_1	$\mathrm{E}(\hat{\theta}_r^*)$	$\mathrm{E}(\mathrm{Std.Err}(\hat{\theta}_r^*))$	95%	$\mathrm{E}(\hat{\theta}_r^*)$	$\mathrm{E}(\mathrm{Std.Err}(\hat{\theta}_r^*))$	95%
0.1	0.3	0.406	0.1308	0.4	0.4054	0.0584	0
0.4	0.2	1.2267	0.1398	63.8	1.2195	0.0623	5
0.3	0	1.5258	0.137	3.1	1.5247	0.0613	0
0.4	0.6	0.1246	0.1493	0	0.1228	0.0665	0
0.7	0.8	0.1005	0.1968	0.3	0.1041	0.0875	0
0.9	0.6	2.1554	0.2615	0.1	2.1348	0.1149	0

Again, this suggests that conclusions from the simulation study based on naive analysis could be misleading as the problem of measurement error is not considered. Therefore, we assess the impact of the differential misclassification pattern of $Y_i^* = 1$.

5.5.3 Modified Model Analysis IV

Here, for each dataset we demonstrate the empirical performance of the modified model under the assumption of a differential misclassification of $Y_i^* = 1$. Table 5.16 shows an extract of the results for each dataset. The combinations of (ζ_0, ζ_1) under SIMdata are the six datasets, while the combinations under Miscl rates are the modified models. When a correct model is applied, the estimates and coverage appear to be good. Overall, from Table 5.16, it can be seen that when an incorrect model is applied, the estimate could be biased and the coverage is poor.

Table 5.16: Extract from the simulation study under a Differential Misclassification of $Y_i^* = 1$: Results are summarised over 1000 iterations.

SIMdata		Miscl rates			$n = 1000$			$n = 5000$	
					E(Std.Err)			E(Std.Err)	
ζ_0	ζ_1	ζ_0	ζ_1	$\mathrm{E}(\hat{\theta}_r \mid \zeta_0, \zeta_1)$	$(\hat{\theta}_r \mid \zeta_0, \zeta_1)$	95%	$\mathrm{E}(\hat{\theta}_r \mid \zeta_0, \zeta_1)$	$(\hat{\theta}_r \mid \zeta_0, \zeta_1)$	95%
0.1	0.3								
		0.1	0.2	0.6801	0.1504	44.3	0.6792	0.067	0
		0.1	0.3	1.0025	0.1686	95.6	1.0007	0.0752	95.9
		0.2	0.4	1.2864	0.2176	76.3	1.281	0.0965	15.4
0.4	0.2								
		0.2	0.3	1.8431	0.1951	0.1	1.8316	0.0867	0
		0.4	0.2	1.0084	0.1789	95.8	0.9995	0.0798	95
		0.4	0.3	1.4140	0.206	48.2	1.4024	0.0916	0.6
0.3	0								
		0	0.1	1.834	0.1523	0.1	1.8328	0.0679	0
		0.3	0	1.0013	0.1517	95	1.0007	0.0677	96
		0.2	0.3	2.8341	0.367	0.3	2.7988	0.1513	0
0.4	0.6								
		0.5	0.1	-0.773	0.1861	0	-0.775	0.083	0
		0.4	0.6	1.0055	0.2472	95.2	0.9999	0.1097	95.4
		0.3	0.5	0.7308	0.1979	72.4	0.7282	0.0882	14.4
0.7	0.8								
		0.5	0.4	-0.1201	0.2216	0.1	-0.1158	0.0986	0
		0.7	0.8	1.0031	0.3825	96.4	0.9987	0.1677	95.3
		0.1	0.4	0.5912	0.2083	50.1	0.5949	0.0927	0.9
0.9	0.6								
		0.2	0.1	2.0652	0.2652	0.3	2.0447	0.1166	0
		0.9	0.6	1.0164	0.4266	95.4	0.9983	0.1874	95.2
		0.4	0.3	2.1279	0.2744	0.2	2.1074	0.1207	0

5.5.4 Influence Method Analysis IV

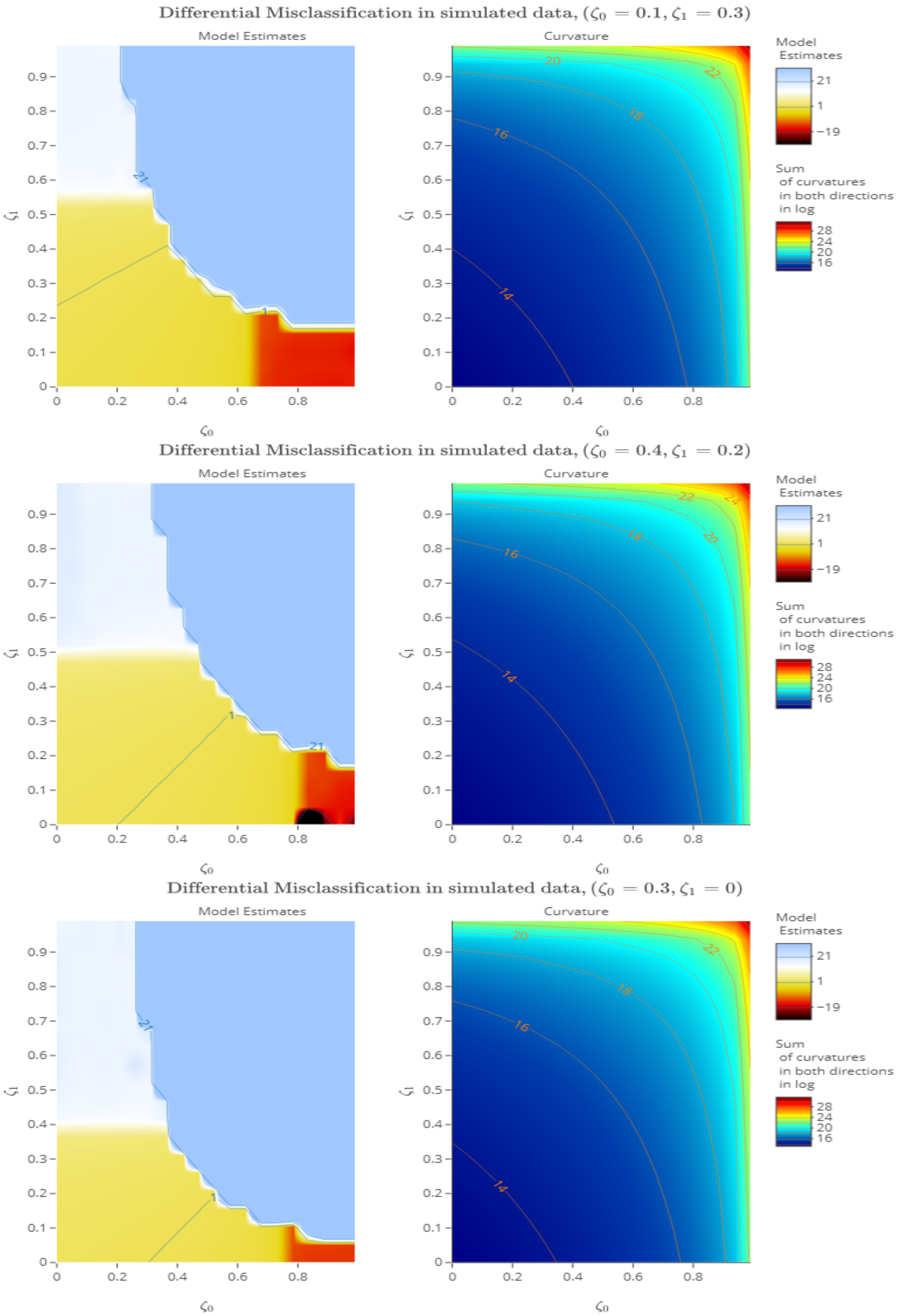
Table 5.17 shows an extract of model estimates and curvatures (in log transformation) for each of the direction ζ_0 and ζ_1 and their sums under a differential misclassification of $Y_i^* = 1$. Again, it can be seen that when a correct model is fitted, the estimate seems to be unbiased and significant.

The left graph of Figure 5.9 and 5.10 is the model estimates while the right graph is the sum of curvatures. For each dataset, we can identify models with estimated parameter 1 using contour line 1. These combinations of ζ_0, ζ_1 might be the plausible range of the error size present in each dataset. The right shows the sum of curvatures for each simulated data. Again, the assumed model has the lowest curvature. It is important to investigate how sensitive the conclusion and stability assessment is along the error values.

Figure 5.6 depicts the sum of normal curvatures in log transformation under differential misclassification of $Y_i^* = 1$ for the six datasets. It follows that the smallest curvature occurs at the origin $(\zeta_0 = 0, \zeta_1 = 0)$, while the highest curvature is near $(\zeta_0 = 1, \zeta_1 = 1)$. With no exception, all studied scenario individually shows that assume model is the most stable model compared to any other model. Hence, when there is no validation study and there is no prior knowledge of measurement error contained in a study, we can chose assume model.

Table 5.17: Extract of Model Estimates and Sum of Curvatures under a Differential Misclassification from simulation study IV. The six datasets selected are $(0.1, 0.3), (0.4, 0.2), (0.3, 0), (0.4, 0.6), (0.7, 0.8), (0.9, 0.6)$

SIMdata		Miscl rates			Estimates			Curvature	
ζ_0	ζ_1	ζ_0	ζ_1	$\hat{\theta}\|\zeta_0, \zeta_1$	95%CI	Pval	$\log(\kappa_{h_0})$	$\log(\kappa_{h\zeta_1})$	$\sum_{j=0}^{1}\log(\kappa_{h_{\zeta_j}})$
0.1	0.3	0	0	0.44	(0.18,0.69)	<0.001	6.106	6.662	12.768
		0.1	0.2	0.72	(0.43,1.02)	<0.001	6.316	7.109	13.425
		0.1	0.3	1.06	(0.72,1.39)	<0.001	6.316	7.376	13.692
		0.2	0.4	1.37	(0.93,1.81)	<0.001	6.552	7.718	14.269
0.4	0.2	0	0	1.3	(1.02,1.58)	<0.001	5.563	6.892	12.455
		0.2	0.3	1.9	(1.52,2.28)	<0.001	6.009	7.601	13.615
		0.4	0.2	1.09	(0.74,1.44)	<0.001	6.584	7.338	13.923
		0.4	0.3	1.49	(1.09,1.88)	<0.001	6.584	7.601	14.189
0.3	0	0	0	1.52	(1.25,1.79)	<0.001	5.771	7.447	13.218
		0	0.1	1.81	(1.52,2.11)	<0.001	5.771	7.658	13.429
		0.3	0	1.01	(0.71,1.3)	<0.001	6.485	7.447	13.932
		0.2	0.3	2.63	(2.06,3.21)	<0.001	6.218	8.160	14.378
0.4	0.6	0	0	0.19	(-0.11,0.48)	0.21	5.563	5.856	11.419
		0.5	0.1	-0.68	(-1.04,-0.32)	<0.001	6.949	6.067	13.016
		0.4	0.6	1.07	(0.6,1.55)	<0.001	6.584	7.689	14.273
		0.3	0.5	0.8	(0.41,1.19)	<0.001	6.276	7.243	13.519
0.7	0.8	0	0	0.13	(-0.24,0.51)	0.49	5.044	5.073	10.117
		0.5	0.4	-0.08	(-0.51,0.35)	0.71	6.431	6.095	12.525
		0.7	0.8	1.08	(0.32,1.84)	0.005	7.452	8.292	15.744
		0.1	0.4	0.63	(0.23,1.03)	0.002	5.255	6.095	11.350
0.9	0.6	0	0	2.19	(1.68,2.7)	<0.001	3.921	5.856	9.778
		0.2	0.1	2.1	(1.58,2.62)	<0.001	4.368	6.067	10.435
		0.9	0.6	1.08	(0.27,1.89)	0.009	8.526	7.689	16.215
		0.4	0.3	2.16	(1.62,2.7)	,0.001	4.943	6.570	11.513



Differential Misclassification in simulated data, $(\zeta_0 = 0.1, \zeta_1 = 0.3)$

Differential Misclassification in simulated data, $(\zeta_0 = 0.4, \zeta_1 = 0.2)$

Differential Misclassification in simulated data, $(\zeta_0 = 0.3, \zeta_1 = 0)$

Figure 5.9: Simulation IVa: Model Estimates and Sum of Curvatures of both directions of ζ_0 and ζ_1 from the last three datasets under a Differential Misclassification of $Y_i^* = 1$. First row; Model Estimates and Sum of Curvatures from SIMdata $(\zeta_0 = 0.1, \zeta_1 = 0.3)$, Middle row; Model Estimates and Sum of Curvatures from SIMdata $(\zeta_0 = 0.4, \zeta_1 = 0.2)$, and Last row; Model Estimates and Sum of Curvatures from SIMdata $(\zeta_0 = 0.3, \zeta_1 = 0)$.

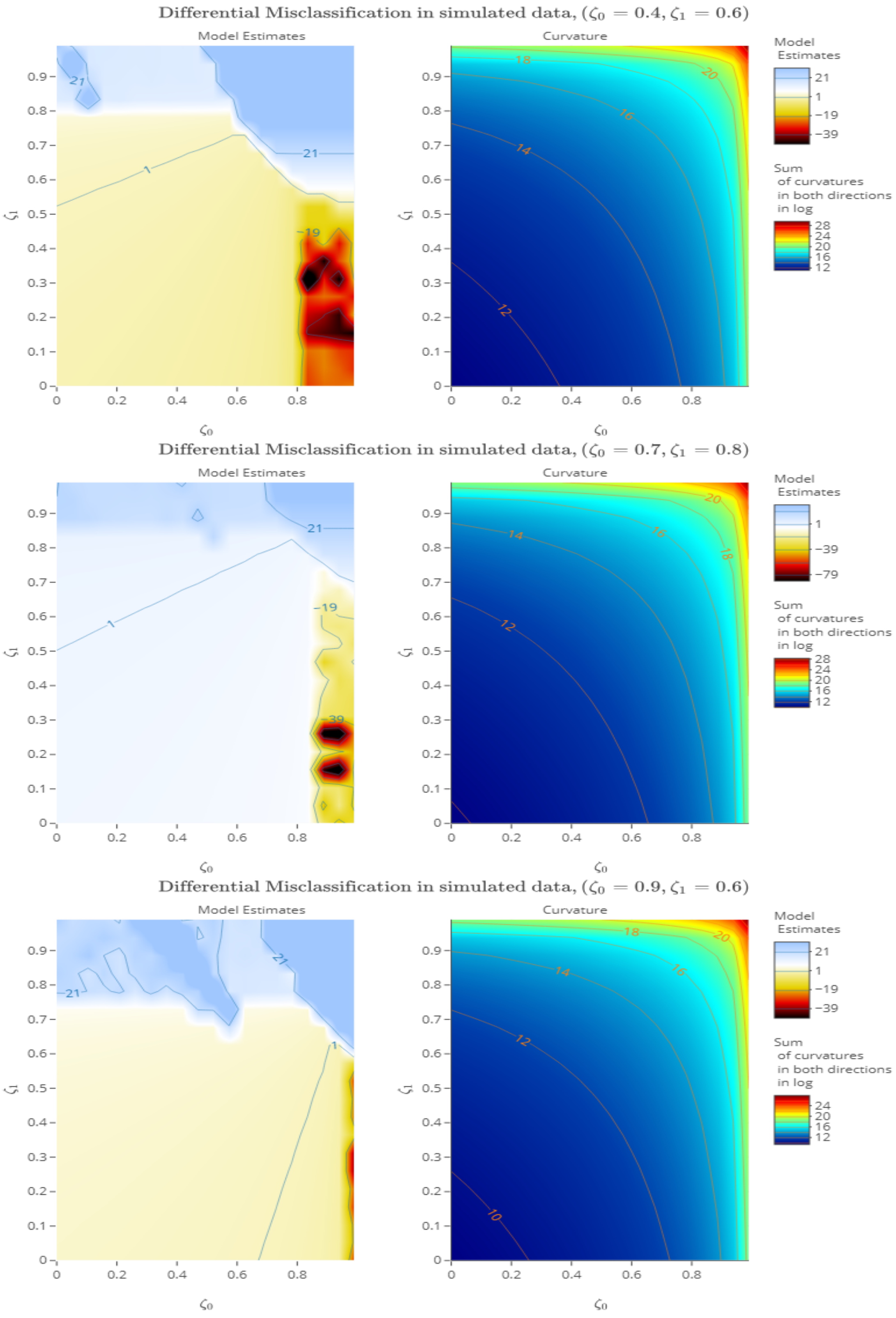


Figure 5.10: Simulation IVb: Model Estimates and Sum of Curvatures of both directions of ζ_0 and ζ_1 from the last three datasets under a Differential Misclassification of $Y_i^* = 1$. First row; Model Estimates and Sum of Curvatures from SIMdata $(\zeta_0 = 0.4, \zeta_1 = 0.6)$, Middle row; Model Estimates and Sum of Curvatures from SIMdata $(\zeta_0 = 0.7, \zeta_1 = 0.8)$, and Last row; Model Estimates and Sum of Curvatures from SIMdata $(\zeta_0 = 0.9, \zeta_1 = 0.6)$.

Chapter 6

Discussion and Conclusions

Statistical inference from an assumed model might be misleading if the impact of potential measurement error was not considered in statistical techniques. As a result, there exists a breadth of literature on different approaches used to account for measurement error. For instance, in the case of binary data, past research proposed methods to account for misclassification in outcome and/or covariates that might be present in an observed study. The existing methods in binary data framework, focus on the use of validation study to account for measurement error in the main study. A shortcoming of this approach is lack of validation data to inform the correction of measurement error in the main study. In addition, many of these methods used modified models (that is, extended logistic regression models), where the probability of measurement error is introduced as a modification weight in the assumed model. However, very little has be done to assess how stable an assumed model is under the influence of a minor modification of a binary model with outcome misclassification. It is therefore essential to investigate the performance of a naive logistic regression model fitted to main/observed study against a modified model.

The investigation is aimed at understanding how model estimates behave when a slight change is introduced in an assumed model, and model stability/sensitivity within a plausible range of error size. In practice, the true value of the error size is unknown. One possible solution is to find good information about a range of plausible values of the error size either from the literature or a validation data. However, it is sometimes difficult to get prior information about the likely range of error size. If there is no information about the error size, an alternative is to consider range of plausible values between 0 and 1. Although for the purpose of model development, the error size used in our study range from 0 to 1, note that in specific context the two extremes may not be plausible. Regardless of a validation data or a prior information about the error size, it is important to perform sensitivity analysis within plausible values, and report all the findings. If the assumed model estimates are unbiased and more stable than modified model under the impact of measurement error, the derived measure could be a useful tool in filling the gap between existing methods with only the main study, and those relying on combined main/validation substudies.

Therefore, the core idea of our research focuses on influence measure derived from curvature to study model stability under the impact of measurement error. It should be noted that using curvature as an influence measure leads to different interpretations. The key is to derive influence measure when modification is incorporated in the general likelihood function of an assumed binary model, where response variable is subject to misclassification. Each point on the influence graph or surface represents a distinct model (assumed or modified model). The curvature measures how curvy the surface is at a distinct model. Hence, when using curvature as an influence measure, it can be used to interpret how stable or sensitive a model is with respect to slight changes in the model assumption of measurement error.

A well known approach of influence measure is the use of eigenvalues as the basis and the corresponding eigenvectors as directions (Zhu and Zhang, 2004). This approach of influence method has been followed before by Steen et al., (2001), where they assessed individual specific modifications for missing ordinal responses in longitudinal models/settings. Another proposed method is the use of mean or gaussian curvature (Cook, 1984). Thus, any of the aforementioned influence methods could be used to detect the most sensitive observations when a slight change is introduced to a model, or outlying values (emerge) that seriously influence coefficient estimates. However, these methods do not directly address the aim and objectives of this research, where investigating the stability of assumed or modified models turns out to be of paramount importance of our study. In other words, the influence method is very extensive and can potentially be represented in numerous forms. Our method is based on assessing the stability of an assumed model under modifications in the case of measurement error for binary outcome. There are key differences between assessing influential individual measurements around missing at random model for continuous outcome (Steen et al., 2001) and the method presented here under error-prone binary response.

In this study, more specifically, we set out to derive influence measure within the problem of error-prone measurements of attributes of binary outcome. In both cases, we showed the theoretical framework of assessing model stability. When correcting for measurement error, the two most common misclassification patterns are nondifferential and differential misclassification. Hence, in our approach, we allowed for the modification of an assumed model based on nondifferential and differential misclassification pattern. In the case of nondifferential misclassification, which is modelled by introducing scalar modification weight into an assumed model, we proved that the assumed model is the most stable compared to any other modified models, when there is no information about the error size. This is demonstrated in the mathematical proofs in Section 3.4.1 of Chapter Three, and Section 4.2.1 of Chapter Four. Similarly, in the case of a differential misclassification pattern, where vector of modification weights are introduced, the assumed model emerges as the most stable when there is no information about the error size, as shown in mathematical proof in Section 3.4.2 of Chapter Three, and 4.2.2 of Chapter Four. For each case of $Y_i^* = 1$ and $Y_i^* = 0$, the empirical performances and numerical results under nondifferential and differential patterns are shown using the motivating example data and simulation studies in Chapter Five.

Although our theoretical framework showed that the standard logistic model is more stable than modified models under misclassification, model estimates appear to be biased when an incorrect model is applied(fitted). This can be seen in our simulation results in Section 5.2 to 5.5 of Chapter Five. In our study, there is no prior information about the error size. Hence, for model development on how to assess model stability under the impact of measurement error, the error size applied ranges from 0 to 1. However, in a specific context, extreme values may not be plausible. We showed how model estimates and stability behaves as the error size increases from 0 to 1. In real scenarios where the true value of error size is unknown, when measurement error is suspected, a set of plausible values could be found in literature (or from a validation data) which can be incorporated in a model. On the other hand, if there is no prior information about the error size, one possible solution is to consider plausible range of values. It is important to examine how model estimates and curvature behave within the plausible range of values. If the conclusion does not change much across a range of plausible values of error size, then it is robust. On the other hand, when a slight change in the error size changes the conclusion, and the curve of the curvature is so much such that the model gives you a very different result, then it is important to report all the findings. Therefore, we recommend that regardless of the information about the error size, model estimates and curvature within plausible range of values should be examined.

We illustrated how to incorporate a vector of modification weights in the likelihood function of an assumed model, and also shown how to derive influence measure. One of the advantages of our method is that it is easy to derive and computation is straightforward. The influence measures derived in our research are for special cases of attribute of binary outcome, but these are not the only possible misclassifications. Further, it is possible to consider a study where both attributes of a binary outcome are subject to measurement error at the same time. In such a scenario, the mechanism of measurement error is different (see Figure 1.1 from Section 1.2 in Chapter One). It has to be emphasized that the modification introduced in an assumed model is allowed to be a vector of modification weights. Therefore, our method can be extended to a study where both attributes of a binary outcome are misclassified. In this scenario, at least two modification weights are likely to be introduced into the assumed model depending on the assumptions made. The modification weights in this case share similarities with the misclassification rates considered in our study, in which $\zeta = P(Y_i^* = 0|Y_i = 1)$ and/or $\gamma = P(Y_i^* = 1|Y_i = 0)$ are misclassification rates. However, by assuming the occurrence of ζ and γ at the same time, in the case of nondifferential mechanism, the distribution of the observed outcome variable is conditional on ζ and γ, resulting in modified model $Y_i^*|X_i \sim Bin(1, (1-\zeta)\pi_i + \gamma(1-\pi_i))$. Obviously we can see that there is a difference between $Y_i^*|X_i \sim Bin(1, (1-\zeta)\pi_i + \gamma(1-\pi_i))$ and the two modified models (3.4.1) and (4.2.1) described in Chapter Three and Four respectively. Similarly, it follows that under the case of differential measurement error pattern, there is a difference between modified models. Since the assumed model is the most stable compared to modified models in our study, it is expected that by assuming ζ and γ simultaneously, the most stable model may likely be the assumed model. However, the latter comment is intuitive. Further investigation outlining change (if any) in the behaviour of the influence measures needs to be done.

Future work could be carried out to assess assumed or modified models under the impact of error-prone exposure. For instance, an example in epidemiology studies where demographic and/or multiple risk factors such as age, gender, ethnicity, biomarkers, preexisting diseases (to mention a few) are observed in error. Often in such studies, they tend to identify associations between primary or secondary outcome and exposure variables, where they adjust for some other variables, but a subset of these variables might be error-prone. When a subset of the covariates are measured with error, caution should be applied when specifying the modified model. This is because misspecifying the association between error -free and mismeasured covariates would introduce bias during parameter estimation. Also, selecting an inappropriate modification vector to a statistical model may lead to misleading inferences about the effect of the modification, since the perturbation vector is central to the development of the inference measure (Zhu et al., 2007). Hence, the number of modification weights to be incorporated in an assumed model needs careful consideration depending on the assumptions. Further research is required.

In addition, observe studies with more than one exposure variable merit further research. Possible measurement errors that may arise from such study could be that only the response or a subset of exposure may be subject to measurement error, or even both response and covariates are mismeasured. When considering error measurements of response and exposure variables, the association between error-free and mismeasured covariates must be correctly specified including the probabilities of measurement error of both variables. In this scenario, modified models under the impact of nondifferential, independent and dependent differential misclassification assumptions were provided in Section 3 of Chapter Two. Hence, following our approach, influence measure can be derived in this scenario. In practice, we adjust for variables during analysis. Adjusting for variables may change the results of our study. Thus, this is another area of research that needs further investigation.

Normal curvature along individual direction of modification weights may be complex in higher dimensions. For instance, an example is a special case of differential measurement error pattern under error prone $Y_i^* = 1$ and $Y_i^* = 0$. Here, the assumption is that the probability of measurement error is dependent on the error free measurement of binary exposure variable x_i. In this case, the modification weights are $(\zeta_0, \zeta_1, \gamma_0, \gamma_1)$. Hence, the distribution of the observed outcome variable conditional on ζ_0, ζ_1, γ_0 and γ_1 is $Y_i^*|X_i \sim Bin(1, (1-\zeta_0+\zeta_0 x_i-\zeta_1 x_i)\pi_i+(\gamma_0-\gamma_0 x_i+\gamma_1 x_i)(1-\pi_i))$. Deriving an influence measure based on normal curvature under this setting is a bit complex. Therefore, a good knowledge of differential geometry, theory of differential forms and Riemannian metrics is required to construct normal curvature in higher dimensional space. Furthermore, a computational, user- friendly statistical software or computer code in implementing this procedure would be a useful development.

The modification around binary outcome and our influence method is by no means limited to the logistic model. Another area of interest is improving or correcting outcome and/or covariates misclassification in a range of common models to include cox and ordinal. Although these models are outside the logistic regression model framework, the concept of our method can be applied to other model settings.

In summary, observed studies are prone to measurement error. Hence, in our research, we demonstrated how to investigate model stability and the behaviour of model estimates when one of the two attributes of a binary response variable is prone to nondifferential and differential misclassification. Although from our theoretical results, when there is no information about the error size, the assumed model appears to be the most stable model compared to the modified models, the assumed model estimates could be biased when measurement error is present. Therefore, it is important to find good information about the error size either from a validation study or the literature. In practice, it is sometimes difficult to get prior information about the likely range of error size. Hence, regardless of prior information about the error size, it is important to investigate model stability and how model estimates behave within a plausible range of error size, and report all the findings.

Bibliography

Anderson, T. (2003), *An Introduction to Multivariate Statistical Analysis. Wiley Series in Probability and Statistics. Wiley-Interscience; 2003.*

Barnard, G. (1980), 'Discussion of professor box's paper', *Journal of the Royal Statistical Society:Series A* **143**, 404–406.

Barron, B. (1977), 'The effects of misclassification on the estimation of relative risk', *Biometrics* **33**(2), 414–418.

Bedrick, E., Christensen, R. and Johnson, W. (1996), 'A new perspective on priors for generalized linear models', *Journal of the American Statistical Association* **91**(11), 1450––1460.

Bedrick, E., Christensen, R. and Johnson, W. (1997), 'Bayesian binomial regression: predicting survival at a trauma center', *American Statistician* **51**(8), 211—218.

Behnken, D. and Draper, N. (1972), 'Residuals and their variance patterns', *Technometrics* **14**(1), 101–111.

Berger, J. (1990), 'Robust bayesian analysis: sensitivity to the prior', *Journal of Statistical Planning and Inference* **25**(3), 303–328.

Berger, J. (1994), 'An overview of robust bayesian analysis', *TEST* **3**, 5–58.

Berger, J. (2000), *Bayesian robustness. In: Rios Insua, D.; Ruggeri, F., editors. Robust Bayesian analysis Insua, D.; Ruggeri, F., editors. Robust Bayesian analysis*, Springer-Verlag; New York:2000.

Box, G. and Draper, N. (1975), 'Robust design', *Biometrika* **62**(2), 347–352.

Breslow, N. (1990), 'Biostatistics and bayes', *Statistical Science* **5**(3), 269–298.

Bross, I. (1954), 'Misclassification in 2 x 2 tables', *Biometrics* **10**(4), 478–486.

Carroll, R., Ruppert, D. and Stefanski, L. (1995), *Measurement Error in Nonlinear Models*, London: Chapman Hall.

Carroll, R., Ruppert, D., Stefanski, L. and Crainiceanu, C. (2006), *Measurement Error in Nonlinear Models. A Modern Perspective, Second Edition*, Chapman and Hall; London.

Carroll, R., Spiegelman, C., Gordonlan, K., Bailey, K. and Abbott, R. (1984), 'On errors-in-variables for binary regression models', *Biometrika* **71**(1), 19—25.

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

Carroll, R., Wang, S. and Wang, C. (1995b), 'Asymptotic for prospective analysis of stratified logistics case control studies', *Journal of the American Statistical Association* **90**, 157—169.

Chen, F., Zhu, H., Song, X. and Lee, S. (2010), 'Perturbation selection and local influence analysis for generalized linear mixed models', *Journal of Computational and Graphical Statistics* **19**(4), 826–842.

Choo, H., Ibrahim, J., Sinha, D. and Zhu, H. (2009), 'Bayesian case influence diagnostics for survival models', *Biometrics* **65**(1), 116–124.

Cole, S., Chu, H. and Greenland, S. (2006), 'Multiple-imputation for measurement-error correction', *International Journal of Epidemiology* **35**(4), 1074–1081.

Cook, R. (1977), 'Detection of influential observation in linear regression', *Technometrics* **19**(1), 15–18.

Cook, R. (1986), 'Assessment of local influence', *Journal of the Royal Statistical Society: Series B* **48**(2), 133–169.

Copas, J. and Eguchi, S. (2001), 'Local sensitivity approximations for selectivity bias', *Journal of the Royal Statistical Society, Series B* **63**(4), 871–895.

Copas, J. and Eguchi, S. (2005), 'Local model uncertainty and incomplete data bias', *Journal of the Royal Statistical Society, Series B* **67**(4), 459–513.

Copas, J. and Eguchi, S. (2010), 'Likelihood for statistically equivalent models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(2), 193–217.

Copas, J. and Li, H. (1997), 'Inference for non-random samples', *Journal of the Royal Statistical Society: Series B* **59**(1), 55–95.

Copas, J. and Shi, J. (2000), 'Meta-analysis, funnel plots and sensitivity analysis', *Biostatistics* **3**(1), 247–262.

Critchley, F., Atkinson, R., Lu, G. and Biazi, E. (2001), 'Influence analysis based on the case sensitivity function', *Journal of the Royal Statistical Society:Series B* **63**(2), 307–323.

Dempster, A., Laird, N. and Rubin, D. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **39**(1), 1–22.

Diggle, P. and Kenward, M. (1994), 'Informative dropout in longitudinal data analysis', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **43**(1), 49–93.

Edwards, J., Cole, S., Troester, M. and Richardson, D. (2013), 'Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data', *American Journal of Epidemiology* **177**(9), 904–912.

Freedman, L., Fainberg, V., Kipnis, V., Midthune, D. and Carroll, R. (2004), 'A new method for dealing with measurement error in explanatory variables of regression models', *Biometrics* **60**(1), 172–181.

Freedman, L., Midthune, D., Carroll, R. and Kipnis, V. (2008), 'A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression', *Statistics in Medicine* **27**(25), 5195–5216.

Gerlach, R. and Stamey, J. (2007), 'Bayesian model selection for logistic regression with misclassified outcomes', *Statistical Modelling* **7**(3), 255–273.

Gleser, L. (1990), *Improvemens in the naive approach to estimation in nonlinear errors in variables regression model. In statistical analysis of measurement error models an Application. P J Brown and W J Fuller(eds). Providence: American Statistical Society.*

Greenland, S. (1980), 'The effect of misclassification in the presence of covariates', *American Journal of Epidemiology* **112**(4), 564–569.

Greenland, S. (1988), 'Variance estimation for epidemiologic effect estimates under misclassificatio', *Statistics in Medicine* **7**(7), 745–757.

Greenland, S. (1996), 'Basic methods for sensitivity analysis of biases', *International Journal of Epidemiology* **25**(6), 1107—1116.

Hogg, T., Petkau, J., Zhao, Y., Gustafson, P., Wijnands, J. and Tremlett, H. (2017), 'Bayesian analysis of pair-matched case-control studies subject to outcome misclassification', *Statistics in Medicine* **36**(18), 4196—4213.

Ibrahim, J., Chen, M., Lipsitz, S. and Herring, A. (2005), 'Missing-data methods for generalized linear model: A comparative review', *Journal of the American Statistical Association* **100**(469), 332–346.

Ioan, C. (2010), 'The space geometry of production functions', *[Online; accessed: 13th May 2021],* ***Url:*** *https://www.researchgate.net/profile/Catalin-Ioan* .

Karen, M. and N.Loki (2008), ‘Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment’, *Statistics in Medicine* **27**(30), 6332–6350.

Kass, R. and Raftery, A. (1995), ‘Bayes factors’, *Journal of the American Statistical Association* **90**(23), 773—795.

Lee, J. M. (1997), *Riemannan Manifolds: An Introduction to Curvature*, New York: Springer-Verlag.

Lin, N., Shi, J. and Henderson, R. (2012), ‘Doubly misspecified models’, *Biometrika* **99**(2), 285–298.

Little, R. and Rubin, D. (2002), *Statistical Analysis With Missing Data*, Wiley, New York, Second Edition.

Little, R. and Zhang, N. (2011), ‘Subsample ignorable likelihood for regression analysis with missing data’, *Journal of the Royal Statistical Society: Series C* **60**(4), 591–605.

Lyles, R. (2002), ‘A note on estimating crude odds ratios in case–control studies with differentially misclassified exposure’, *Biometrics* **58**(4), 1034–1036.

Lyles, R., L.Tang, H.Superak, C.King, Celentano, C., Lo, Y. and Sobel, J. (2011), ‘Validation data-based adjustments for outcome misclassification in logistic regression: An illustration’, *Epidemiology* **22**(4), 589–597.

Lyles, R., Zhang, F. and Drews-Botch, C. (2007), ‘Combining internal and external validation data to correct for exposure misclassification: a case study’, *Epidemiology* **18**(3), 321–328.

Magder, L. and Hughes, J. (1997), ‘Logistic regression when the outcome is measured with uncertainty’, *American Journal of Epidemiology* **146**(2), 195—203.

Marshall, R. (1990), ‘Validation study methods for estimating proportions and odds ratios with misclassified data’, *Journal of Clinical Epidemiology* **43**(9), 941–947.

Mcinturff, P., Johnson, W., Cowling, D. and Gardner, I. (2004), ‘Modelling risk when binary outcomes are subject to error’, *Statistical Modelling* **23**(7), 1095—1109.

Molenberghs, G., Benunckens, C., Sotto, C. and Kenward, M. (2008), ‘Every missingness not at random model has a missingness at random counterpart with equal fit’, *Journal of the Royal Statistical Society: Series B* **70**(2), 371–388.

Molenberghs, G., Verbeke, G., Thijs, H., Lesaffre, E. and Kenward, M. (2001), ‘Sensitivity analysis for nonrandom dropout: A local influence approach’, *Biostatistics* **57**(1), 7–14.

Morrissey, M. and Spiegelman, D. (1999), ‘Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons’, *Biometrics* **55**(2), 338–344.

Neuhaus, J. (1999), 'Bias and efficiency loss due to misclassified responses in binary regression', *Biometrika* **86**(4), 843–855.

Poon, W. and Y.Poon (1999), 'Conformal normal curvature and assessment of local influence', *Journal of the Royal Statistical Society:Series B* **61**(1), 51–61.

Pregibon, D. (1981), 'Logistics regression diagnostics', *The Annals of Statistics* **9**(4), 705–724.

Rao, C. (1973), *Linear statistical inference and its applications. 2nd ed. New York: John Wiley and Sons.*

Richardson, S. and Gilks, W. (1993), 'A bayesian approach to measurement error problems in epidemiology using conditional independence models', *American Journal of Epidemlotogy* **138**(6), 430–442.

Rosner, B., Spiegelman, D. and Willett, W. (1990), 'Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error', *American Journal of Epidemiology* **132**, 157—169.

Rosner, B., Spiegelman, D. and Willett, W. (1992), 'Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error', *American Journal of Epidemiology* **136**(11), 1400–1413.

Rubin, D. (1976), 'Inference and missing data', *Biometrika* **63**(3), 581—592.

Rubin, D. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Schafer, D. (1987), 'Covariate measurement error in generalized linear models', *Biometrika* **74**(2), 385—391.

Schafer, J. (1997), *Analysis of Incomplete Multivariate Data. Monographs on Statistics & Applied Probability. 72. Chapman & Hall/CRC;.*

Searle, S. (1982), *Matrix algebra useful for statistics. New York: John Wiley and Sons.*

Spiegelman, D., Carroll, R. and Kipnis, V. (2001), 'Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument', *Statistics in Medicine* **20**(1), 139–160.

Stamey, J., Young, D. and W, S. J. (2008), 'A bayesian approach to adjust for diagnostic misclassification between two mortality causes in poisson regression', *Statistics in Medicine* **27**(13), 2440–2452.

Steen, K., Molenberghs, G., G.Verbeke and H.Thijs (2001), 'A local influence approach to sensitivity analysis of incomplete longitudinal ordinal data', *Statistical Modelling* **1**, 125–142.

Tang, L., Lyles, R., King, C., Celentano, D. and Lo, Y. (2015), 'Binary regression with differentailly misclassified response and exposure variables', *Statistics in Medicine* **34**(9), 1605–1620.

Tang, L., Lyles, R., King, C., Hogan, J. and Lo, Y. (2015), 'Regression analysis for differentailly misclassified correlated binary outcomes', *Journal of the Royal Statistical Society: Series C* **64**(3), 433–449.

Thomas, D., Stram, D. and Dwyer, J. (1993), 'Exposure measurement error: influence on exposure-disease relationships and methods of correction', *Annual Review of Public Health* **14**, 69–93.

Thurston, S., Spiegelman, D. and Ruppert, D. (2003), 'Equivalence of regression calibration methods in main study/external validation study designs.', *Journal of Statistical Planning and Inference* **113**(2), 527—539.

Thurston, S., William, P., Hauser, R., Hu, H., Hernsndez-Avila, M. and Spiegelman, D. (2003), 'A comparison of regression calibration approaches for designs with internal validation data', *Journal of StatisticalPlanning and Inference* **131**(1), 175–190.

Wang, N. and Robin, J. (1998), 'Large sample theory for parametric multiple imputation procedures', *Biometrika* **85**(4), 935–948.

Zhu, H. and Ibrahim, J. (2011), 'Bayesian influence analysis: a geometric approach', *Biometrika* **98**(2), 307—323.

Zhu, H., Ibrahim, J., Cho, H. and N.Tang (2012), 'Bayesian case influence measures for statistical models with missing data', *Journal of Computational and Graphical Statistics* **21**(1), 253–271.

Zhu, H., Ibrahim, J., Lee, S. and H.Zhang (2007), 'Perturbation selection and influence measures in local influence analysis', *The Annals of Statistics* **35**(6), 2565–2588.

Zhu, H., J.Ibrahim, Cho, N. and Tang, N. (2010), *A general review of Bayesian influence analysis. In: Chen MH, Dey DK, Muller P, Sun D, Ye K, editors. Frontier of statistical decision making and Bayesian analysis*, New York: Springer-Verlag; 2010.

Zhu, H. and Lee, S. (2001), 'Local influence for incomplete data models', *Journal of the Royal Statistical Society, Series B* **63**(1), 111–126.

Zhu, H. and Zhang, H. (2004), 'A diagnostic procedure based on local influence', *Biometrika* **91**(3), 579–589.