Risk estimation using an exact method of fitting log-link models

by

Chao Zhu

MSc in Statistics

BSc (Hons) in Actuarial Mathematics

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy (Medical Study)

Menzies Institute for Medical Research

University of Tasmania

February 2021

Statements and Declarations

Declaration of Originality

This thesis contains no material which has been accepted for a degree or diploma by the University or any other institution, except by way of background information and duly acknowledged in the thesis, and to the best of my knowledge and belief no material previously published or written by another person except where due acknowledgement is made in the text of the thesis, nor does the thesis contain any material that infringes copyright.

Signed: …… Date: ……20/10/2020……

Statement of Authority of Access

This thesis is not to be made available for loan or copying for two years following the date this statement was signed. Following that time, the thesis may be made available for loan and limited copying and communication in accordance with the Copyright Act 1968.

Signed: …… Date: ……20/10/2020……

Statement Regarding Published Work

The publishers of the papers comprising Chapters 2, 3, 4 own the copyright in the content, and access to the material should be sought from the respective journals. The remaining non-published content of the thesis may be made available for loan and limited copying in accordance with the Copyright Act 1968.

Signed: …… Date: ……20/10/2020……

Statement of Co-Authorship

This thesis includes chapters for which Chao Zhu is not the sole author. He was the first author in the research of each chapter with assistance from the co-authors. The contributions and academic information for each author are acknowledged below.

Candidate: Chao Zhu, Menzies Institute for Medical Research, University of Tasmania.

Author 1: Professor Leigh Blizzard (primary supervisor), Menzies Institute for Medical Research, University of Tasmania.

Author 2: Professor David W Hosmer (co-supervisor), University of Vermont.

Author 3: Dr Jim Stankovich (co-supervisor), Monash University.

Author 4: Dr Karen Wills (co-supervisor), Menzies Institute for Medical Research, University of Tasmania.

Chapter 2

Chao Zhu as the primary and the first author contributed 60% to the topic exploration and decision, theorem summarisation and proof, example analysis, simulation designation and drafted chapter. Also, Zhu built and tested the relevant package for the chapter in R. Professor Leigh Blizzard as senior author and primary supervisor contributed 25% to the summarisation and proving of the theorem, example selection and analysis, simulation designation and revision of the manuscript. Also, Prof. Blizzard performed and tested the relevant package for the chapter in Stata. The other three authors Professor David W Hosmer, Dr Jim Stankovic and Dr Karen Wills contributed 5% each to the paper review and revision. All authors approved the final version of the manuscript.

Statements and Declarations

Chapter 3

Chao Zhu as the primary and the first author contributed 75% to the topic exploration and decision, theorem summarisation and proof, example analysis, simulation designation and drafted chapter. Also, Zhu built and tested the relevant package for the chapter in R. Professor Leigh Blizzard as senior author and primary supervisor contributed 10% to the summarisation and proving of the theorem, example selection and analysis, simulation designation and revision of the manuscript. He performed and tested the relevant package for the chapter in Stata. The three other authors Professor David W Hosmer, Dr Jim Stankovic and Dr Karen Wills each contributed 5% to the paper review and revision. All authors approved the final version of the manuscript.

Chapter 4

Chao Zhu as the primary and the first author contributed 75% to the topic exploration and decision, theorem summarisation and proving, example analysis, simulation designation and drafted chapter. He built and tested the relevant package for the chapter in R. Professor Leigh Blizzard as senior author and primary supervisor contribute 10% to the summarisation and proving of the theorem, example selection and analysis, simulation designation and revision of the manuscript. He performed and tested the relevant package for the chapter in Stata. The three other authors Professor David W Hosmer, Dr Jim Stankovic and Dr Karen Wills each contributed 5% to the paper review and comments. All authors approved the final version of the manuscript.

We, the undersigned, agree with the above stated “proportion of work undertaken” for each of the above chapters contributing to this thesis:

Statements and Declarations

Signed: ……

Chao Zhu

Candidate

Menzies Institute for Medical Research

University of Tasmania

Date: ……20/10/2020……

Signed: ……

Professor Leigh Blizzard

Primary Supervisor

Menzies Institute for Medical Research

University of Tasmania

Date: ……23 September 2020…

……

Professor Alison Venn

Chair of Menzies Institute for Medical Research

University of Tasmania

……13 October 2020……

Abstract

The relative risk has been widely reported as a ratio measure of association between covariates for study factors and a binary outcome of interest in medical research. It is possible to estimate relative risk through the log binomial model, a member of the family of generalised linear models with binomial errors and logarithmic link. However, since it was first proposed, this model has encountered numerical difficulties which restrict its use in studies using real-world data. The standard fitting algorithm of the log binomial model may fail to converge when the maximum likelihood (ML) solution is on the boundary of the allowable parameter space. If the ML solution lies on the boundary, special methods are needed because at least one vector of covariate values (referred to as boundary vector) has an estimated probability of unity when evaluated at the ML solution. For a model with a single covariate, Deddens et al. (2003) proposed an exact method based on re-parametrisation of the covariate. Petersen and Deddens (2010) proposed an extension of the exact method to general cases, but the method was incomplete, and the details to implement the method were missing. In this thesis, we provide details, including formulae (with proof) for estimating the covariances necessary to implement the method, explanation (with proof) of an inter-dependency between coefficient estimates, and proof that the method can be applied in general. The relevant R package for implementing the exact method is provided.

Another measure of the effect of a risk factor is the risk difference, which is recommended to be reported in clinical trials to assist clinicians in making evidence-based decisions about treatment allocation. It is possible to estimate the risk difference by fitting an identity-link binomial model. However, the standard fitting algorithm of the identity-link binomial model may fail to converge due to two sources of numerical difficulties. Use of an inadmissible starting value is sometimes responsible for failed convergence in the identity-link binomial

model as it can cause the fitted probability of some observations to be less than zero or greater than unity. The standard fitting algorithm, therefore, may not be able to iteratively correct the results if it starts the iteration from an inadmissible starting value. To solve this problem, we have introduced a well-designed starting value calibration for obtaining an admissible starting value of a standard fitting algorithm in an identity-link binomial model. Numerical difficulties may also be encountered if the ML solution is on the boundary of the allowable parameter space. The standard fitting algorithm in the identity-link binomial model will usually fail to converge when the ML solution lies on the boundary of parameter space. Given its similarity to the log binomial model, an extension of the exact method is introduced to overcome the difficulties in the identity-link binomial model. However, there are two boundaries, lower and upper, in the parameter space of identity-link binomial models, whereas the log binomial model only has an upper bound. We provide a strategy to compare and locate the ML solution. Eight theorems and two corollaries with proof are presented to obtain the estimates of coefficients and the relevant variance-covariance matrix. We demonstrate the application of the exact method in detail using example data. A real-world dataset and a designed simulation are provided to further discuss and compare the results of the exact method with other approaches. The relevant R package for implementing the exact method is provided.

The risk ratio/relative risk as a measure of effect is also used in the clustered/ longitudinal dataset. Fitting a marginal log binomial model estimated by generalised estimating equation (marginal LBM by GEE) provides a possible way to estimate the relative risk in correlated data. However, the algorithm may fail to converge even from admissible starting values. The previously published studies of the marginal LBM by GEE have focused on convergence rates and the selection of a working correlation structure. To date, there is no published work

accounting for the causes of non-convergence or proposing remedies for it. By investigating data with convergence issues, we found that formulating the marginal LBM by GEE as a population-averaged model might also fail to converge or converge to an inappropriate solution when there is a fitted probability that is extraordinarily close or equal to unity. It is a similar issue to the log binomial model for independent data. We extend the exact method to the marginal LBM by GEE and provide details for its implementation. The properties of the exact estimator are investigated by simulation, and the results are compared with those of a marginal modified Poisson with log-link function estimated by generalised estimating equation (marginal Poisson by GEE). The relevant R package for implementing the exact method is provided.

In this thesis, we studied the numerical difficulties in the log binomial model, the identity-link binomial model and the marginal LBM by GEE. Two algorithms were introduced to address the difficulties due to the inadmissible starting values for the log and identity-link binomial model. In the presence of boundary vectors, the exact method is effective in estimating the coefficients of covariates in those three models. It can eliminate the influence of boundary vector and improve the model fitting.

Acknowledgements

I would like to first thank my primary supervisor, Professor Leigh Blizzard, for his support. He was always willing to take time to work with me. His high standard of supervision and extensive academic experience guided my research in the right direction. He has not only guided my academic career but also helped me to understand Australian culture. He invited me to his family parties at Easter and Christmas. I feel honored and lucky to have him as my primary supervisor.

I would also like to thank my other supervisors. First, thanks to Dr Karen Wills for being an excellent facilitator in our group meetings when people disagree with each other. She provided useful comments on my works and kept my progress on track. I also would like to thank Dr Jim Stankovich from Monash University. He gave me a lot of good advice on the structure of the thesis and English editing. Lastly, I would like to thank Emeritus Professor David Hosmer, who is an expert in the field of biostatistics. He kindly offered his time and advice, and I very much appreciate his teaching and support.

I would also like to thank several fellow students and staff (in no particular order) at Menzies whose support both in-person and academically made the process so much easier: Ishanka, Petr, Fateme, Monique, Zhen, Yuan. Thank you for your advice, support, encouragement, and friendship.

I also wish to thank my Graduate Research Coordinators, Professor Wendy Oddy, for her invaluable advice on how to shape up my PhD journey. My sincere thank also goes to all academic and administrative staff at Menzies.

Also, I would like to thank the two anonymous examiners who read my thesis. I anticipate their thorough review and thoughtful comments will be helpful and encouraging.

Lastly, I would like to thank my family members: my parents, my wife and my daughter. Without their financial and emotional supports, I could not complete my PhD. In particular, I would like to thank my Father Shiwang Zhu and Mother Yueqin Pei, who always encourage me to pursue my dreams and who looked after my daughter throughout my entire PhD. I owe all of my achievements to you and the sacrifices you have made in my life. I also would like to thank my wife, who is doing her PhD at Macquarie University. We are spouses in life and friends in academia and support each other as we forge ahead.

Scientific Presentations Arising from the Thesis

July 2018

2018 Joint Statistical Meetings, Vancouver, Canada

“Log binomial regression when the maximum likelihood solution is on the boundary of the parameter space.” (Oral presentation)

July 2019

2019 Joint Statistical Meetings, Denver, US

“Relative risk estimation in clustered/longitudinal data using generalised estimating equations (GEE).” (Speed presentation)

These two presentations included some of the work presented in Chapters 2 and 4 respectively.

List of Tables

List of Tables

List of Figures

Table of Contents

Chapter 1 Introduction

1.1 Background

The generalised linear models (GLM) are a broad class of conventional linear regression models, first unified by Nelder and Wedderburn [1] and discussed in detail in the book of McCullagh and Nelder [2]. A GLM usually consists of three parts: the response variable *Y*, which is assumed to follow a distribution from the exponential family, such as the normal, binomial, Poisson and Gamma distributions; a linear combination $\mathbf{X}'\boldsymbol{\beta}$ of a set of explanatory variables *X* and the relevant coefficients $\boldsymbol{\beta}$; and a link-function *g* which establishes the connection between the response variable *Y* and the linear combination of $\mathbf{X}$ and $\boldsymbol{\beta}$. For the i^{th} observation, the general form of the probability density function in a GLM is

$$\mathrm{f}\left(y_i \mid \theta_i, \phi\right) = \exp\left(\frac{y_i\theta_i - b\left(\theta_i\right)}{a\left(\phi\right)} + c\left(y_i, \phi\right)\right), \tag{1.1}$$

and the joint probability density function with *n* independent observations is written as:

$$l\left(\theta, \phi\right) = \prod_{i=1}^{n} \exp\left(\frac{y_i\theta_i - b\left(\theta_i\right)}{a\left(\phi\right)} + c\left(y_i, \phi\right)\right), \tag{1.2}$$

which is also called the likelihood function generally. In practice, it is more convenient to work with the logarithm of the likelihood, the log-likelihood function,

$$L\left(\theta, \phi\right) = \sum_{i=1}^{n}\left(\frac{y_i\theta_i - b\left(\theta_i\right)}{a\left(\phi\right)} + c\left(y_i, \phi\right)\right). \tag{1.3}$$

In the above functions, ϕ is called the dispersion parameter, which is constant over observations, and θ_i is called the canonical parameter. The forms of the functions *a*, *b* and *c* vary depending on the distribution of *Y*. The detailed forms of these parameters and functions in each distribution are presented in the book of McCullagh and Nelder [2].

In medical research, a form of GLM with the outcome variable following a binomial distribution has been commonly used to study the relationship between a set of independent variables and a specific binary outcome. Since the probability function in the binomial distribution can be written in an exponential family form as:

$$\begin{aligned} \mathrm{f}\left(y_i \mid \mu_i\right) &= \binom{m_i}{y_i} \mu_i^{y_i} \left(1-\mu_i\right)^{m_i - y_i} \\ &= \exp\left[y_i \log\left(\frac{\mu_i}{1-\mu_i}\right) + m_i \log\left(1-\mu_i\right) + \log\binom{m_i}{y_i} \right], \end{aligned} \tag{1.4}$$

it is not difficult to identify each element in function (1.1) as:

$$\theta_i = \log\left(\frac{\mu_i}{1-\mu_i}\right),\ b\left(\theta_i\right) = m_i \log\left[1+\exp\left(\theta_i\right)\right],\ \text{and}\ c\left(y_i\right) = \log\binom{m_i}{y_i},$$

where μ_i is the expected value of y_i. The dispersion parameter ϕ is equal to 1. In practice, a more natural way in which the binomial distribution arises is that the outcome variable in an independent data follows a Bernoulli distribution which is a specific binomial distribution with $m_i = 1$ [2, 3]. The probability function in the Bernoulli distribution is:

$$\begin{aligned} \mathrm{f}\left(y_i \mid \mu_i\right) &= \mu_i^{y_i} \left(1-\mu_i\right)^{1-y_i} \\ &= \exp\left[y_i \log\left(\frac{\mu_i}{1-\mu_i}\right) + \log\left(1-\mu_i\right) \right]. \end{aligned} \tag{1.5}$$

The likelihood function can be written as:

$$\begin{aligned} l &= \prod_{i=1}^{n} \mu_i^{y_i} \left(1-\mu_i\right)^{1-y_i} \\ &= \prod_{i=1}^{n} \exp\left[y_i \log\left(\frac{\mu_i}{1-\mu_i}\right) + \log\left(1-\mu_i\right) \right], \end{aligned} \tag{1.6}$$

and the relevant log-likelihood function is written as:

$$L = \sum_{i=1}^{n} y_i \log \mu_i + \left(1-y_i\right)\log\left(1-\mu_i\right). \tag{1.7}$$

All GLMs with the outcome variable following a binomial distribution share the same likelihood (1.6) and log-likelihood function (1.7).

Two extensions of the GLM, known as the log binomial model and the identity-link binomial model, are viable approaches to summarise the risk of the outcome in a study group versus a reference group. The log binomial model allows the estimation of relative risk, which is a ratio of the probability of an outcome in the study group to the probability of the outcome in the reference group. The identity-link binomial model offers a way to estimate the risk difference, which is calculated by taking the difference in the risk of getting an event in the reference group from the risk of getting an event in the study group. Relative risk and risk difference are both intuitive summary metrics, and presenting them allows for a relatively clear interpretation of study results [4]. However, to estimate relative risk and risk difference by fitting a log binomial model and an identity-link binomial model often comes with challenges. That is, standard fitting algorithms may meet numerical difficulties and fail to converge in some cases. In this chapter, I provide a general background regarding the potential underlying reasons for numerical difficulties in the log binomial model and the identity-link binomial model and previous work that has been carried out to try to solve these problems.

1.2 Numerical instability of estimating the relative risk in the log binomial model

Relative risk is the ratio measure of choice for summarising the impact of exposure on the incidence proportion ("risk") in epidemiological studies [5]. Fitting a log binomial model with binomial errors and a logarithmic link to binary outcome data makes it possible to estimate risk and risk ratios in longitudinal studies, and prevalence ratios in cross-sectional studies.

Assuming n observations with a binary outcome variable Y and J covariates in the model, the likelihood function of the log binomial model is defined previously in function (1.6). For the convenience of the reader, I repeat it here:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \mu_i^{y_i} (1-\mu_i)^{1-y_i}$$

where μ_i is the probability of the outcome conditional on the i^{th} observation, which is equal to an exponential linear combination with parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots \beta_J)$ and covariates $\mathbf{x}_i = (1, x_{i1} x_{i2}, \ldots x_{iJ})$ with constant and denoted as:

$$\Pr(Y_i = 1 \mid \mathbf{x}_i) = \mu(\mathbf{x}_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta}),\ i = 1, 2, ..., n\,. \qquad (1.8)$$

This conditional probability function means that the fitting procedure of the log binomial model requires constraints on $\boldsymbol{\beta}$ to ensure that $\exp(\mathbf{x}_i'\boldsymbol{\beta})$ is within the interval [0, 1] [6]. By taking logarithms of both sides of (1.8), the function becomes:

$$\log(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_J x_{iJ},\ i = 1, 2, ...n\,, \qquad (1.9)$$

and the interval is mapped to a semi-open interval $(-\infty, 0]$. This leads to an inequality constraint $-\infty < \mathbf{x}_i'\boldsymbol{\beta} \le 0$ on $\boldsymbol{\beta}$ that was first asserted by Wedderburn [7]. With this inequality constraint, the allowable parameter space of $\boldsymbol{\beta}$ is defined [8, 9] as

$$\Theta = \{\boldsymbol{\beta} : \mathbf{x}_i'\boldsymbol{\beta} \le 0,\ \text{for all } i = 1, 2, ..., n\}\,. \qquad (1.10)$$

This restricted parameter space means that the standard fitting algorithm of the log binomial model may meet numerical difficulties, and either fail to converge, or converge to an inappropriate solution. Various types of difficulties are described below.

One is data separation, which occurs when all observations of a particular predictor have the same outcome. This is a common issue in models with a binary outcome variable and is not unique to the log binomial model. The second type of difficulty is issues with the fitting

algorithm. Williamson et al. [10] presented an example to illustrate a failure of convergence with the log binomial model. Marschner later restated this as a repulsion problem that can occur when using Fisher scoring to fit a log binomial model. In short, the Fisher scoring algorithm may have a repelling fixed point. This means that, even if the maximum likelihood (ML) solution is a stationary point inside the allowable parameter space, Fisher scoring is unable to converge and exhibits cyclical behaviour [8]. The third difficulty is the use of an inappropriate starting value with the standard fitting algorithm. The default starting value algorithm in statistical packages may provide a poor starting value, which results in the fitting algorithm beginning the iteration outside the boundary of the allowable parameter space. In this issue, the fitting algorithm may not be able to self-correct back to the boundary, which will lead to some fitted probabilities of observations exceeding unity. The last difficulty is that the standard fitting algorithm may meet numerical instability and fail to converge when the ML solution lies on the boundary of the allowable parameter space Θ. When this occurs, there is at least one covariate vector with the corresponding linear predictor equal to 0, which means the estimated probability is equal to 1. We refer to any such sets of covariate vectors as boundary vectors. As an example, consider the data presented in Table 1.1. The log binomial model $\Pr(Y_i = 1 | \mathrm{x}_i) = \exp(\beta_0 + \beta_1 x_i),\ i = 1, 2, ...40$ for these data has the ML solution of $\hat{\beta}_0 = -0.344616$ and $\hat{\beta}_1 = 0.344616$, which lies on the boundary of the allowable parameter space (the area covered by a blue grid shown in Figure 1.1. The boundary vectors are the five observations for which $x = 1$ and where the event occurred.

Table 1.1 An example dataset from Williamson et al. [10] where the ML solution lies on the boundary of the parameter space.

Exposure	Event	No event	Total
$x = -1$	10	8	18
$x = 0$	18	9	27
$x = 1$	5	0	5

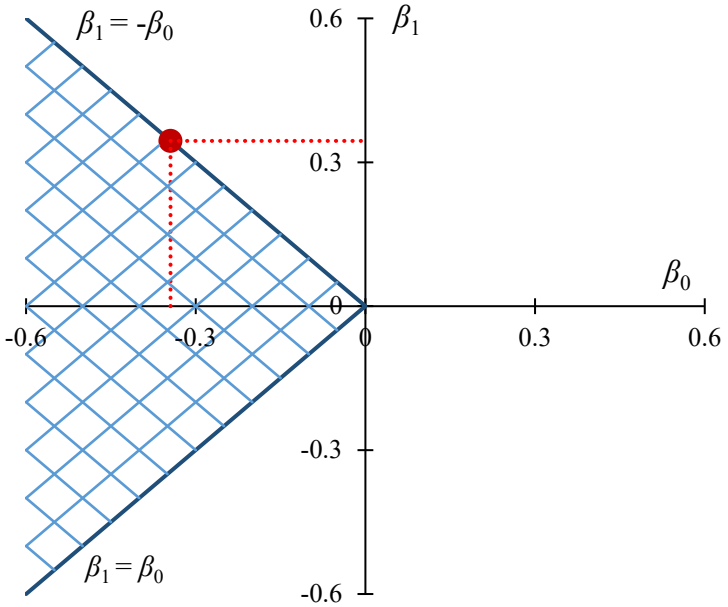

Figure 1.1. The area covered by the grid shows the allowable parameter space $\Theta = \{\boldsymbol{\beta} : \beta_0 + \beta_1 x \le 0, \text{ for all } i = 1, 2, ..., 50\}$: values of $\hat{\beta}_0$ and $\hat{\beta}_1$ in the inequality $\beta_0 + \beta_1 x \le 0$ ensure that the fitted probabilities are within the interval [0, 1] for all observations in the dataset of Table 1.1. The ML solution of $\hat{\beta}_0 = -0.344616$ and $\hat{\beta}_1 = 0.344616$ marked as a red point lies on the boundary of the parameter space.

The presence of a boundary vector makes the estimation of the ML solution problematic when using standard fitting algorithms. Even if the iteration begins from an appropriate starting value, it may still stop at a result outside the parameter space Θ, leading to a failure of convergence. To explain why this problem arises, I begin with the information matrix of the log binomial model. By taking the first partial derivative of (1.7), I obtain the score function of log binomial model as follows:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{x_{ij}(y_i - \mu_i)}{1 - \mu_i} \right] \tag{1.11}$$

The second partial derivative of (1.7) in matrix form is

$$\mathbf{H} = \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{j_1} \partial \beta_{j_2}} = \mathbf{X}' \left\{ \mathbf{diag} \left[-\frac{\mu_i (1 - y_i)}{(1 - \mu_i)^2} \right] \right\} \mathbf{X}, \tag{1.12}$$

which is also known as the Hessian matrix. The observed information matrix (OIM) is the negative of the Hessian, which is obtained by substituting the estimate of each fitted probability $\hat{\mu}_i$ evaluated at $\hat{\boldsymbol{\beta}}$, written as:

$$\widehat{\mathbf{OI}} = -\mathbf{H} = \mathbf{X}' \left\{ \mathbf{diag} \left[\frac{\hat{\mu}_i (1 - y_i)}{(1 - \hat{\mu}_i)^2} \right] \right\} \mathbf{X}. \tag{1.13}$$

The relevant expected information matrix (EIM) is:

$$\widehat{\mathbf{EI}} = -\mathrm{E}(\mathbf{H}) = \mathbf{X}' \left[\mathbf{diag} \left(\frac{\hat{\mu}_i}{1 - \hat{\mu}_i} \right) \right] \mathbf{X}. \tag{1.14}$$

The standard fitting algorithms used to estimate the ML solution in the log binomial model (such as Newton-Raphson and iteratively re-weighted least squares) mostly rely on the score function and either the OIM or the EIM. If a boundary vector is presented in data, it will result in a fitted probability $\hat{\mu}_i$ extremely close or equal to 1. This will lead to errors in the calculation of the denominator in the score function (1.11) and Hessian matrix (1.12), thereby affecting the convergence of the model. Some previous work has been done in an attempt to find feasible solutions to overcome the numerical difficulties in log binomial model. They are summarised below.

The majority of work to date on the log binomial model has been focused on finding approaches to solve the failure of convergence. Blizzard and Hosmer [11] suggested that some convergence problems could be avoided by a simple modification of the fitting algorithm, where the log-likelihood is only evaluated for observations with $y_i = 0$. Other researchers presented some “work-around” methods (methods that do not attempt to estimate the ML

solution) to approximate the risk ratio reasonably. Wacholder [12] proposed that any fitted values found to be outside the parameter space should be set to values known to be inside the space at each iteration. Stata's manual states that they have applied the idea into their command *binreg* [13]. Zhang and Yu [14] made use of a well-known method for converting odds ratios to relative risks using a baseline prevalence. They developed a method which uses logistic regression followed by the conversion of odds ratios to risk ratios using this conversion method. This method is cumbersome and not directly applicable to estimates adjusted for continuous covariates. Breslow [15] showed that by assuming a constant risk period, the conditional hazard ratio estimated by Cox's method for a closed cohort is equal to the cumulative incidence ratio (risk ratio). Lee and Chia [16] advocated that Cox regression be adapted to approximate the risk ratio by building a dataset in which each person has a pre-set and fixed follow-up time. Schouten et al. [17] proposed an approach to approximately estimate the parameters of the log binomial model. The approach estimates the model by fitting a logistic regression with the expanded data, which duplicated the observations of $y = 1$ and replaced with $y = 0$. Deddens et al. [18] advocated a method the called the COPY method. This solves the non-convergence issue in the log binomial model by adding 999 copies of the data with one copy having the outcome indicator reversed. McNutt et al. [19] recommended modified Poisson regression to estimate the coefficients from the log binomial model. Since the Poisson errors may overestimate the binomial errors when the outcome is not rare [19], Zou [20] proposed to use the information sandwich estimator to obtain variance estimates, which is robust to the error misspecification. Later on, Carter et al. [21] showed that estimates from modified Poisson regression consistently estimate the coefficients in the log binomial model. The work-around methods listed above can approximately estimate the coefficients of the log binomial model in some cases.

Marschner and Gillett [22] took advantage of the property of exponential linear combination in the log binomial model, treated it as the product of the components, and used the Expectation-Maximization (EM) algorithm to locate the ML solution. In the method, the outcomes are viewed as the product of the independent unobserved binary latent outcomes, thereby the corresponding event probability associated with the observed outcomes is the joint probability associated with the latent outcomes. Since the latent outcome model is defined over a restricted parameter space, the authors believed the issues of parameter space in the log binomial model could be simply avoided.

Moreover, some researchers thought the convergence problem in the log binomial model was due to the contradiction between the unconstrained standard fitting algorithm and the constrained allowable parameter space. Thereby, another reliable approach is to apply a linearly constrained optimisation method. Lange [23] described an approach called the adaptive barrier approach for estimating models in the GLM with a linear inequality constraint. A general implementation is available through the constrOptim function in R. Some researchers [24, 25] implemented this approach in their studies. All of these studies on constrained optimisation are summed up in a paper written by Andrade and Andrade [26].

However, none of them attempted to locate the ML solution in the log binomial model by directly eliminating the impact of the boundary vector. Petersen and Deddens [9] proposed an "exact" method to estimate the ML solution when it lies on the boundary of the parameter space. However, this method was incomplete. For example, the formulae to obtain the covariance between the estimate of the covariates and the estimate of the intercept and the necessary details to implement the exact method were missing. (It is discussed in more detail in Chapter 2 of this thesis.) In consequence, it has not gained traction among practitioners. To

remedy that, I provide the missing details with four theorems and two corollaries as well as the description of strategies required to implement the method. Mathematical details are reported in Appendix A.

1.3 Numerical instability of estimating risk difference in the identity-link binomial model

As a measure to evaluate the absolute effect in the risk of an outcome between the reference group and the study group, the absolute risk difference has been recommended by many reputable medical journals as a measure to inform clinical decisions regarding the benefit and harm of study treatment [4, 27-33]. A regression model makes it possible to estimate risk and risk difference, including adjustment for confounders and assessment of interactions. A linear probability model is used to perform this purpose. It has a binary outcome variable with the residuals following a normal distribution and is fitted as a linear regression model with the least squares method. A deficiency of the model is that the estimated coefficients can imply fitted probabilities outside the interval [0, 1], which are invalid as probabilities mathematically.

The GLM with a binomial error and identity link (identity-link binomial model) offers an opportunity to estimate the risk difference through its coefficients. Just the same as GLMs with a binomial error and other links, it uses the likelihood function as follows:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \mu_i^{y_i} \left(1-\mu_i\right)^{1-y_i},$$

where μ_i is the probability of outcome conditional on the i^{th} observation, which is equal to a linear combination with parameters $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \ldots \beta_J\right)$ and covariates $\mathbf{x}_i = \left(1, x_{i1}, x_{i2}, \ldots x_{iJ}\right)$ denoted as:

$$\Pr\left(Y_i = 1 \mid \mathbf{x}_i\right) = \mu\left(\mathbf{x}_i\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_J x_{iJ},\ i = 1, 2, ...n. \quad (1.15)$$

The conditional probability function implies an inequality constraint on the linear combination $0 \le \mathbf{x}_i'\boldsymbol{\beta} \le 1$. Therefore, the allowable parameter space of the identity-link binomial model is defined as:

$$\Theta = \left\{\boldsymbol{\beta} : 0 \le \mathbf{x}_i'\boldsymbol{\beta} \le 1, \text{ for all } i = 1, 2, ..., n\right\}. \quad (1.16)$$

The story of the parameter space is more complicated in an identity-link binomial model compared with the one in a log binomial model. The boundary of parameter space in an identity-link binomial model is not only fenced by an upper bound but also a lower bound. The lower or upper bounds will be reached respectively if the fitted probabilities are equal to 0 or 1. Because of those two boundaries, the standard fitting algorithm of identity-link binomial model is most likely to meet numerical difficulties and fails to converge. These difficulties can be attributed mainly to the following two reasons.

1. Inappropriate starting values. In the identity-link binomial model, an inadmissible starting value could lead to the fitting algorithm failing to converge or converging to a set of inappropriate estimates of coefficients resulting in fitted probabilities outside the interval [0, 1].

   An appropriate starting value can simply solve this issue. We introduced a starting value calibration into the identity-link binomial model. It applies the Min-Max Normalisation approach to correct the inappropriate values obtained from the default starting value algorithm, thereby making all fitted probabilities inside the interval [0, 1]. This approach is detailed in Chapter 3 of this thesis.

2. The ML solution lies on the boundaries of the parameter space.

> Under this situation, the standard fitting algorithm is likely to meet numerical difficulties and fail to converge because of the unconstrained fitting processes, even though the iteration begins from an appropriate starting value.
>
> A specialised method is required to overcome this difficulty. Because of the similarity between the log binomial model and the identity-link binomial model, I extend the exact method to an identity-link binomial model to overcome the convergence difficulty due to the boundary vectors. Eight theorems and two corollaries with proofs are introduced to support our method. The details are presented in Chapter 3 of this thesis.

Some alternative approaches have been proposed to estimate the risk difference. McNutt et al. [19] described the use of Poisson regression models to estimate relative risk when the standard fitting algorithm fails to converge, and Zou [20] proposed the use of the robust sandwich variance estimator [34, 35] to correct the estimated standard errors. Carter et al. [21] showed that the coefficients from the Poisson model consistently estimate the coefficients of the log binomial model. This model is referred to here as the modified Poisson model. Fitted with an identity link, this model could be used to approximately estimate the coefficients of the identity-link binomial model when the convergence issue was present [36]. However, because of the non-constrained standard fitting algorithm applied, estimates obtained by modified Poisson may result in a solution with the fitted probabilities exceeding unity [22, 37]. The same issue could arise when fitting the identity-link binomial model as well [38]. Moreover, Cheung [39] pointed out that there may be convergence issues in estimating the risk difference by using the modified Poisson approach with an identity link, and recommended a modified least-squares regression method with robust variance estimation, where the risk is

represented as the expected value of a binary outcome, and the ordinary least-squares is used to estimate the parameters. Although there are no convergence issues with modified least-squares regression, the unrestricted fitting algorithm may produce a solution with fitted probabilities outside the interval [0, 1] [38]. Bieler et al. [40] proposed a model-adjusted risk differences method that fits the model as a logistic regression model and converts back to risk difference through the average marginal prediction function defined by Graubard and Korn [41]. It obtains approximate estimates of marginal average risk difference and the corresponding variance-covariance matrix. However, as an approximation of the risk difference, an appropriate comparison with those provided by other alternative approaches was not available. This made it hard for users to make an informed decision on which approach to adopt.

Marschner [42] introduced the expectation-maximisation (EM) algorithm to overcome convergence issues with identity-link modified Poisson model and make it possible to approximately estimate risk difference. The method consists of performing *K* separate maximisations using the EM algorithm. These *K* maximisations are automatically cycled through, and if any of the constrained maximisers produce a stationary point, the process is terminated. If no stationary point is found, then the nonstationary constrained maximum corresponds to the MLE. Later on, by combining the multinomial-Poisson transformation [43] and the identity-link modified Poisson model, Donoghoe and Marschner [38] introduced a novel approach named as the additive binomial model in their paper to approximately estimate the coefficients of the identity-link binomial model (their approach is referred as to the additive Poisson model to distinguish it from the identity-link binomial model in this thesis). Their approach provides extra flexibility for including monotonic regression function to allow semi-parametric adjustment of risk differences.

Based on an adaptive barrier approach, Kovalchik et al. [44] developed a general linear-expit regression model (LEXPIT) that combines linear and nonlinear risk effects to approximately estimate absolute risk and risk variance. However, Donoghoe and Marschner [38] stated that LEXPIT did not produce a good performance in the simulations compared with their additive Poisson model. A nuisance model with doubly robust estimator has also been introduced to estimate risk ratio and risk difference approximately [45]. The approach produces the conditional log-odds product to build an unconstrained nuisance model.

None of the previous approaches directly overcame the numerical difficulties caused by an ML solution on the boundary of the parameter space in the identity-link binomial model. Application of the exact method makes it possible to eliminate the impact of boundary vectors. An appropriate comparison between the results of the exact method and the approximations are presented in Chapter 3 of this thesis.

1.4 Numerical instability of estimating the relative risk in the clustered/longitudinal data

To estimate the relative risk through a model in the independent data, the log binomial model is a viable option. However, in many real-world situations, observations may not be independent and could have correlated errors. In these situations, one possible way to analyse data and produce a better model fit than assuming that observations are independent is to assume an intraclass correlation and appropriately to assign an intraclass correlation structure in the fitting procedure. For example, failure to consider the correlation between a cluster of patients who have been taken care of by the same doctor may result in estimation bias. Therefore, an appropriate statistical model to estimate the relative risk in the

clustered/longitudinal data is needed. The generalised estimating equation (GEE) [46, 47] based on the quasi-likelihood provides a convenient alternative to maximum likelihood estimation method for correlated data. It estimates the averaging effects (population-averaged effect) over all clusters by including the within-cluster dependence. As an extension of GLM, GEE can be used to estimate several types of models with the correlated outcome by changing the link function and the distribution of the outcome variable. Analogous to the log binomial model, the marginal log binomial model estimated by GEE (referred to as marginal LBM by GEE) is recognised as a feasible approach to estimate relative risk, with adjustment for potential confounders and interaction effects. The marginal LBM by GEE as a semiparametric approach does not require a fully defined likelihood function to estimate the model (detailed in Chapter 4 of this thesis). Instead, it evaluates the model by introducing the first and second order components directly into an estimating equation.

Suppose that each cluster i, $i = 1, 2, \ldots, K$ includes n_i observations on related subjects or repeated observations on a single subject. We denote the response variable of the p^{th} observation in cluster i as $y_{ip}, p = 1, 2, \ldots, n_i$, and the response vector under i^{th} as $\mathbf{y}_i = \left(y_{i1}, y_{i2}, \ldots y_{in_i}\right)'$. The link function g builds a relationship between the mean of y_{ip} denoted as μ_{ip} and the vector product of covariates vector $\mathbf{x}_{ip} = \left(1, x_{ip1}, x_{ip2}, \ldots, x_{ipJ}\right)$ and regression parameters $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \beta_2, \ldots, \beta_J\right)$, which is written as:

$$g\left(\mu_{ip}\right) = \mathbf{x}_{ip}'\boldsymbol{\beta}. \qquad (1.17)$$

It is the natural logarithmic function in the marginal LBM by GEE. The variance of y_{ip} represented as $\text{Var}\left(y_{ip}\right) = \phi v\left(\mu_{ip}\right)$ depends on the distribution of y_{ip}, and generally consists of the mean μ_{ip} and scale parameter ϕ which is usually estimated by Pearson residuals [48]

(detailed in chapter 4). In a marginal LBM by GEE, the mean and variance are μ_{ip} and $\mathrm{Var}\left(y_{ip}\right)=\phi\mu_{ip}\left(1-\mu_{ip}\right)$. Within cluster *i*, the variance-covariance matrix is structured as

$$\mathbf{V}_i=\phi\mathbf{A}_i^{1/2}\mathbf{R}_i\left(\boldsymbol{\alpha}\right)\mathbf{A}_i^{1/2}, \tag{1.18}$$

where $\mathbf{A}_i=\mathbf{diag}\left[v\left(\mu_{i1}\right),v\left(\mu_{i2}\right),...v\left(\mu_{in_i}\right)\right]$ and $\mathbf{R}_i\left(\boldsymbol{\alpha}\right)$ is a $n_i\times n_i$ working correlation matrix which describes the correlation structure between observations in cluster *i*. The parameter vector $\boldsymbol{\alpha}$ summarises the correlation between observations [49]. An estimate of $\boldsymbol{\beta}$ is obtained by solving the estimating equations

$$U\left(\boldsymbol{\beta}\right)=\sum_{i=1}^{K}\mathbf{D}_i'\mathbf{V}_i^{-1}\left(\mathbf{y}_i-\boldsymbol{\mu}_i\right)=\mathbf{0}, \tag{1.19}$$

where $\boldsymbol{\mu}_i=\left(\mu_{i1},\mu_{i2},...\mu_{in_i}\right)$ and $\mathbf{D}_i=\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$. The estimation of the variance-covariance matrix $\hat{\boldsymbol{\beta}}$ is based on the sandwich estimator

$$\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}=\left(\sum_{i=1}^{K}\mathbf{D}_i'\hat{\mathbf{V}}_i^{-1}\mathbf{D}_i\right)^{-1}\mathbf{M}\left(\sum_{i=1}^{K}\mathbf{D}_i'\hat{\mathbf{V}}_i^{-1}\mathbf{D}_i\right)^{-1}, \tag{1.20}$$

where $\mathbf{M}=\sum_{i=1}^{K}\mathbf{D}_i'\hat{\mathbf{V}}_i^{-1}\left(\mathbf{y}_i-\hat{\mathbf{u}}_i\right)\left(\mathbf{y}_i-\hat{\mathbf{u}}_i\right)\hat{\mathbf{V}}_i^{-1}\mathbf{D}'$, and $\sum_{i=1}^{K}\mathbf{D}_i'\hat{\mathbf{V}}_i^{-1}\mathbf{D}_i$ is called naive variance-covariance matrix. The estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}$ are consistent even if the working correlation structures are misspecified [46, 50].

In correlated data with a binary outcome variable, if the solution $\hat{\boldsymbol{\beta}}$ obtained from the marginal LBM by GEE lies on the boundary of the parameter space, it will lead to a fitted probability (also known as estimated mean $\hat{\mu}_{ip}$) equal to 1. Such fitted probability $\hat{\mu}_{ip}=1$ will cause problems with the calculation of $\hat{\mathbf{A}}_i^{-1}$ and lead to a failure of convergence in the standard fitting algorithm. In such situations, since the failure of the convergence is also due

to the existence of the boundary vectors, the exact method may be possible to overcome the problem by eliminating the impact of boundary vector. We extend the exact method to the marginal LBM by GEE to deal with the problems of convergence issues due to boundary vectors. Five known criteria (RJ [51], QIC [52], CIC [53], SC [54] and GP [55]) were used to evaluate the improvement of the model fitting after eliminating the boundary vectors using the exact method in a real-world example and a simulation. The details of the procedure to obtain an approximate solution is presented in Chapter 4.

1.5 Research aims

1. To provide the mathematical details and practical guidance necessary to implement the exact method of fitting the log binomial model, to demonstrate and evaluate fits by the exact method to example and simulated data, and to compare the estimates to approximations made by alternative methods of fitting the log binomial model.

2. To extend the exact method to fitting the identity-link binomial model, to provide the mathematical details and practical guidance necessary for that purpose, to demonstrate and evaluate fits by the exact method to example and simulated data, and to compare the estimates to approximations made by alternative methods of fitting the identity-link binomial model.

3. To extend the exact method to fitting the marginal LBM by GEE, to provide the mathematical details and practical guidance necessary for that purpose, to demonstrate and evaluate fits by the exact method to example and simulated data, and to compare the estimates to approximations made by the marginal modified Poisson model estimated by GEE.

The problems that can arise in fitting probability models by standard statistical software in each of these data scenarios are demonstrated using a range of example and published datasets. Each dataset is able to be estimated by the exact method, and the numerical estimates that are reported in each case are able to be replicated, using R software packages prepared as part of this thesis. The packages can be downloaded from GitHub via the links provided in Appendixes.

Chapter 2 Resolution of numerical difficulties in fitting the log binomial model to estimate the relative risk

2.1 Introduction

The risk ratio (relative risk) is the ratio measure of choice for summarising the impact of exposure on the incidence proportion in epidemiologic studies [5]. The log binomial model makes it possible to estimate relative risk with adjustment for confounders. However, numerical difficulties often arise when fitting the log binomial model using statistical software. The difficulties are apparent when the iterative model-fitting algorithm fails to converge or, depending on the software, if convergence occurs but to a warning is issued that one or more of the fitted values exceed unity. Fitted values greater than unity are inadmissible for a model of probabilities. There are four causes for numerical difficulties: data separation, rare failures of the fitting algorithm, poor choice of starting values and the maximum likelihood (ML) solution on the boundary of the parameter space.

The first numerical difficulty is data separation, a common issue in the fitting procedure for models with a binary outcome variable, which is not unique to the log binomial model. The second numerical difficulty is a repelling point issue, which occurs specifically in the fitting algorithm built on the Fisher scoring algorithm. This numerical difficulty was initially identified in the paper of Williamson et al. [10], and was restated as an example of a repelling fixed point issue in the Fisher scoring algorithm by Marschner [8]. We briefly explain the problem and how to avoid it in section 2.3.6. For solving issues related to inappropriate starting values, the third numerical difficulty, we provide an approach in section 2.3.7 to adjust the value obtained from the default starting value algorithm. The main focus of this chapter is on the estimation of the ML solution when it lies on the boundary of the parameter

space, the fourth numerical difficulty. To do so, we make use of an "exact" method suggested by Deddens et al. [18] for estimating boundary solutions for models with a single covariate, and later extended to the general case by Petersen and Deddens [9].

The conditional sampling properties of this model for boundary solutions have not been studied. More generally, this is the case also for the alternative methods that have been proposed to deal with the intractability of the log binomial model [8]. Moreover, some authors [6, 56] have recommended the use of non-ML methods when boundary estimates or non-convergence are encountered with the ML estimator. To address this critical deficiency specifically in respect of the exact method, we undertake simulation studies to investigate the bias, efficiency and confidence interval coverage of the exact estimates of a boundary solution.

The chapter is organised as follows. In Section 2.2, I describe the log binomial model and briefly review the attempts of previous alternative approaches to overcome numerical difficulties. Petersen and Deddens [9] did not provide the details necessary to implement their exact method in general cases, so I provide the missing information in section 2.3 with four theorems and two corollaries. Mathematical details are provided in Appendix A. Sections 2.4 and 2.5 demonstrate the use of the method. Comparisons are made with the results from standard software and several alternative methods. Section 2.6 contains simulation results for data having a single boundary vector. Further results are reported in Appendix A, including a comparison of the performance of the exact method with that of a popular non-ML method. Section 2.7 reports our conclusions.

2.2 Alternative approaches for estimating the coefficients in the log binomial model

2.2.1 The log binomial model

Consider *n* independent observations of a binary outcome variable Y and J non-constant covariates $(X_1, X_2, ...X_J)$. Denote the observed data as $(y_i, \mathbf{x}_i)$ for $i = 1, 2, ...n$ where $\mathbf{x}_i' = (1, x_{i1}, x_{2i}, ...x_{iJ})$. Under the log binomial model, the conditional probability of the outcome given the covariates is:

$$\Pr(Y_i = 1 \mid \mathbf{x}_i) = \mu(\mathbf{x}_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta}),\ i = 1, 2, ...n, \qquad (2.1)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ...\beta_J)'$ is the parameter vector. Suppose there are n observations in the data. Simply rewrite the $\mu(\mathbf{x}_i)$ as μ_i. Then, the log-likelihood function of the log binomial model is:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log(\mu_i) + (1 - y_i)\log(1 - \mu_i). \qquad (2.2)$$

The relationship between the conditional probability of the outcome variable and the linear combination in the equation (2.1) is reformed by the log-link function as:

$$\log(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_J x_{iJ},\ i = 1, 2, ...n. \qquad (2.3)$$

μ_i as a probability is a number within a closed interval [0, 1] which is mapped onto a semi-open interval $(-\infty, 0]$ through a log-link function. This is generally pointed out and proved by Wedderburn [7], and summed up to an inequality constraint $-\infty < \mathbf{x}_i'\boldsymbol{\beta} \le 0$ on $\boldsymbol{\beta}$. With this inequality constraint, the allowable parameter space of $\boldsymbol{\beta}$ is defined [8, 9] as

$$\Theta = \{\boldsymbol{\beta} : \mathbf{x}_i'\boldsymbol{\beta} \le 0, \text{ for all } i = 1, 2, ..., 0\}. \qquad (2.4)$$

Maximum likelihood estimation (MLE) is a standard approach to obtain an optimised estimate of $\boldsymbol{\beta}$ that maximises the log-likelihood over the parameter space Θ, defined as:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta} \in \Theta} L(\boldsymbol{\beta}), \qquad (2.5)$$

which means that the MLE of the log binomial model cannot lie outside the parameter space Θ. This has also been previously emphasised by Deddens and Petersen [57], Lumley et al. [6] and Marschner [8]. With this constrained parameter space Θ, the unconstrained standard fitting algorithm may meet numerical difficulties when the ML solution is on the boundary of parameter space, leading to a fitted probability $\hat{\mu}_i$ of a covariate vector equal to 1, referred to as a boundary vector.

2.2.2 Previous approaches used for estimating the log binomial model

To estimate the log binomial model and overcome the numerical difficulties, previous studies introduced some MLE and non-MLE approaches which are outlined below.

2.2.2.1 MLE approaches

As an MLE approach, the exact method has the advantage of being able to successfully locate the ML solution when it is on the boundary of the parameter space Θ. It was proposed by Deddens et al. [18] for a model with a single covariate, and extended to general cases by Petersen and Deddens [9]. The exact method involves re-parameterisation of the covariates in the log binomial model to eliminate the impact of any boundary vectors in the fitting procedure. The exact method successfully located the ML solution for an example with two covariates and two boundary vectors in Petersen and Deddens' paper [9]. However, it does not provide the essential formula for estimating the covariance between coefficients, which is necessary to obtain the estimated standard errors for each coefficient in the exact method. Moreover, the details to implement the exact method and an appropriate simulation to demonstrate the performance of the method were not provided in the paper. Therefore, the method did not gain enough attention from the researchers. Marschner and Gillett [22] used the multiplicative probability structure of the log binomial model to construct a latent outcome

model for each term in the linear predictor and implemented the EM algorithm to estimate the coefficients on the constrained parameter space iteratively. The approach can successfully fit the log binomial model. However, when the ML solution lies on the boundary of the parameter space, it may converge to an approximation of the ML solution and fail to provide the estimate of the standard error of coefficients.

2.2.2.2 Use of the generalised linear model (GLM) function offered by statistical software to fit an approximate model

Iteratively re-weighted least squares (IRLS) and Newton-Raphson (NR) are two standard fitting algorithms commonly used in the GLM function of statistical software to fit the log binomial model. The *glm* function in R uses the IRLS algorithm to update the estimated parameter values iteratively. If an iteration obtains estimated parameter values outside the parameter space, R-*glm* will apply a non-standard approach, called step-halving, to help the estimation by repeatedly halving the step size of the regular IRLS update, until the updated parameter values are inside the parameter space [8]. However, the fitting algorithm may fall into an infinite loop and fail to converge. In this problem, the iterative process begins to move away from the ML solution after some iterations. The fitting algorithm then self-corrects, returns to a location close to the ML solution, and the iterative process resumes moving away from it again. Eventually, the iterative process reaches the set maximum number of iterations and fails to converge. The problem initially mentioned by Williamson et al. [10]. Marschner [8] further elaborated the problem as a repelling issue of IRLS to a fixed point, and introduced *glm*2 built on R-*glm* to fix the issue. R-*glm*2 solved the issue by merely adding a condition on the step-halving to make it only accept the solution as the log-likelihood increased [58]. Theoretically, both *glm* and *glm*2 can converge to a solution by repetitively invoking step-halving if the iteration begins from an appropriate starting value. However, if the ML solution

is on the boundary of parameter space Θ, the solution obtained from those fitting algorithms may not be the ML solution but an approximation due to the impact of the boundary vector. NR is generally considered to be less stable in the estimation of MLE compared with Fisher scoring. However, it has the advantage of avoiding the repelling fixed point issue when the observed information matrix (OIM) is used. SAS-*GENMOD* combined the benefits in both fitting algorithms. It begins the iteration with IRLS and switches to NR with OIM [8]. Nevertheless, SAS-*GENMOD* could fail to converge when the MLE is on the boundary of the parameter space. The COPY method introduced by Deddens et al. [18] is an approximate estimating approach which involves augmenting the data with 999 copies of the original data but with the outcome indicator reversed in one of the copies, then fitting the model with the augmented data using the standard fitting algorithm. Although the approach avoids issues with the boundary vector, it only provides an approximation of ML solution, which may be biased [6, 59]. Therefore, a MLE appropriate method to address the ML solution when it lies on the boundary of the parameter space remains absent.

2.2.2.3 Approaches based on solving quasi-likelihood estimating equations

Some researchers considered that the score of the log-likelihood function equals zero in the log binomial model is a special case of generalised estimating equations, which is based on the quasi-likelihood function [48, 60]. Since solving the quasi-likelihood estimating equations offer consistent estimation without imposing parameter constraints, several alternative approaches were introduced to approximately estimate the coefficients in the log binomial model [6, 17, 20, 21, 36, 61]. These approaches, including modified Poisson formally introduced by Zou [20] and quasi-MS first suggested by Schouten et al. [17], explained by Lumley et al. [6], and finally improved and summed up by Fitzmaurice et al. [61], were non-MLE methods and could only produce an approximation of the ML solution. Moreover, since these alternative

approaches' fitting procedure does not take into account the parameter space Θ of the log binomial model, the solution may result in a fitted probability outside the interval [0, 1] [8, 37].

2.2.2.4 Approaches based on constrained optimisation

A more direct approach is to introduce an appropriate linear inequality constraint in the traditional fitting algorithm to solve the numerical difficulties due to the ML solution being on the boundary of the parameter space Θ. This new approach, called the adaptive barrier approach, was introduced by Lange [23]. Some researchers [6, 24, 25] implemented this approach in their studies. These studies that relied on the constrained optimisation are summarised in the paper by Andrade and Andrade [26].

All of the previous methods [17, 18, 20, 22, 26, 61] were summarised and compared in a review paper written by Marschner [8]. We accepted his recommendation and compared our results from the exact method with the EM method (MLE approach), the modified Poisson method (non-MLE approach), and the most recent approach introduced by Andrade and Andrade [26] in a real-world example and a well-designed simulation study.

2.3 The exact method

In Section 2.3.1, the fundamentals of the exact method proposed by Petersen and Deddens [9] are presented. Because some essential details were missing in their paper, I explain what is missing in Sections 2.3.2 to 2.3.5 and provide the solutions with five theorems and two corollaries. Furthermore, I provide a well-designed strategy for identifying boundary vectors and fitting models in Section 2.3.8 so that the exact method can be applied in different statistical packages.

2.3.1 Fundamentals of the exact method

Suppose that the ML solution lies on a boundary of the allowable parameter space with maximum value $\mu(\mathbf{x}_i)=1$ for some $i \in 1,2,...n$, and that this maximum value is attained by $R \geq 1$ distinct sets of covariates (including the constant). We refer to these covariate vectors as boundary vectors. Denote the r^{th} boundary vector as $\mathbf{x}^{(r)}=\left(1, x_1^{(r)}, x_2^{(r)}, ...x_J^{(r)}\right)$. If the covariate values $\mathbf{x}^{(r)}$ of the r^{th} $(r=1,2,...R)$ boundary vector are shared by n_r observations, the method outlined by Petersen and Deddens [9] for estimating the model involves:

1. eliminating the constant by subtracting from the constant and each non-constant covariate its respective value in the boundary vector:

$$z_{ij}^{(1)} = x_{ij} - x_j^{(1)}, \quad j=0,1,2,...J \tag{2.6}$$

2. when there are multiple boundary vectors $(R>1)$, eliminating the first $R-1$ non-constant covariates by re-parametrising the covariates according to the scheme:

$$z_{ij}^{(r)} = z_{ij}^{(r-1)} - \left(\frac{t_j^{(r)}}{t_{r-1}^{(r)}}\right) z_{i,r-1}^{(r-1)} \quad \text{where } t_j^{(r)} = z_{ij}^{(r-1)}\Big|_{x_{ij}=x_j^{(r)}}, \quad r=2,3,...R \tag{2.7}$$

3. dropping the observations with covariate values $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ...\mathbf{x}^{(R)}$ respectively, which make no contribution to the log-likelihood, and fitting the model $\mu\left(\mathbf{z}_i^{(R)}\right) = \exp\left(\sum_{j=R}^{J} \beta_j z_{ij}^{(R)}\right)$ without a constant and with $J-R+1$ covariates to the remaining $n-n_1-n_2-...-n_R$ observations to obtain the estimates $\hat{\beta}_R, \hat{\beta}_{R+1}, ...\hat{\beta}_J$ of the coefficients of the non-eliminated non-constant covariates,

4. estimating the coefficients $\hat{\beta}_r$, $r=1,2,...R-1$ of the $R-1$ eliminated covariates as:

$$\hat{\beta}_r = -\frac{\sum_{j=r+1}^{J} \hat{\beta}_j t_j^{(r+1)}}{t_r^{(r+1)}} \tag{2.8}$$

5. estimating the standard errors of the estimated coefficients of eliminated covariates as:

$$\widehat{\mathrm{SE}}\left(\hat{\beta}_r\right)=\sqrt{\sum_{j=r+1}^{J}\left[\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\left(\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\right)^2\right]+\sum_{j_1=r+1}^{J}\sum_{\substack{j_2=r+1\\ j_2\neq j_1}}^{J}\left\{\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)\left[\frac{t_{j_1}^{(r+1)}t_{j_2}^{(r+1)}}{\left(t_r^{(r+1)}\right)^2}\right]\right\}} \qquad (2.9)$$

for $r=1,2,...R-1$ where $\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)$ denotes the estimated variance of the estimated coefficient $\hat{\beta}_j$, and $\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)$ denotes the estimated covariance between the estimated coefficients $\hat{\beta}_{j_1}$ and $\hat{\beta}_{j_2}$;

6. estimating the coefficient of the constant covariate from the boundary condition:

$$\hat{\beta}_0=-\sum_{j=1}^{J}\hat{\beta}_j x_j^{(1)} \qquad (2.10)$$

7. estimating the standard error of the estimated coefficient of the constant covariate as:

$$\widehat{\mathrm{SE}}\left(\hat{\beta}_0\right)=\sqrt{\sum_{j=1}^{J}\left[\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\left(x_j^{(1)}\right)^2\right]+\sum_{j_1=1}^{J}\sum_{\substack{j_2=1\\ j_2\neq j_1}}^{J}\left[\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)x_{j_1}^{(1)}x_{j_2}^{(1)}\right]} \qquad (2.11)$$

2.3.2 Obtaining estimates of the covariances between non-constant covariates

If $R=1$, all covariances between the estimated coefficients of non-constant covariates are provided in the regression output. When $R>1$, however fitting the model $\mu\left(\mathbf{z}_i^{(R)}\right)=\exp\left(\sum_{j=R}^{J}\beta_j z_{ij}^{(R)}\right)$ without a constant and with $J-R+1$ non-constant covariates does not provide estimates of the covariances between the estimated coefficients of the $R-1$ eliminated covariates and the estimated coefficients of the $J-R+1$ remaining non-constant covariates. Those covariances are required in (2.9) to calculate the estimated standard errors of the estimated coefficients of the $R-1$ eliminated covariates, and in (2.11) to calculate the estimated standard error of the estimated coefficient of the constant covariate. For example, if $R=2$, a model with $J=2$ covariates provides exact estimates of β_2 and $\widehat{\mathrm{SE}}\left(\hat{\beta}_2\right)$ but does

not provide an estimate of $\widehat{\mathrm{Cov}}\left(\hat{\beta}_1, \hat{\beta}_2\right)$ that is required to calculate $\widehat{\mathrm{SE}}\left(\hat{\beta}_1\right)$ and $\widehat{\mathrm{SE}}\left(\hat{\beta}_0\right)$. For this specific case, Petersen and Deddens [9] suggested reversing the order of the boundary vectors to produce two equations in the two unknowns $\widehat{\mathrm{SE}}\left(\hat{\beta}_0\right)$ and $\widehat{\mathrm{Cov}}\left(\hat{\beta}_1, \hat{\beta}_2\right)$. The solution provides the estimate of $\widehat{\mathrm{SE}}\left(\hat{\beta}_0\right)$, and the value of $\widehat{\mathrm{Cov}}\left(\hat{\beta}_1, \hat{\beta}_2\right)$ that is necessary to calculate $\widehat{\mathrm{SE}}\left(\hat{\beta}_1\right)$. In general, this approach requires the solution of a $R \times R$ system of simultaneous equations, and is feasible for all cases with $J = R > 1$. It is not applicable if $J > R > 1$.

To overcome this deficiency, we provide a solution in Theorem 2.1, whose proof is presented in Appendix A.

Theorem 2.1

For a log binomial model with $J \geq 1$ independent non-constant covariates fitted by the exact method to data having R $(1 \leq R \leq J)$ distinct sets of values of the covariates (including the constant) for which the estimated outcome probability evaluated at the ML solution is unity, the estimated covariances between exact estimates of the coefficients of the $R-1$ eliminated non-constant covariates and exact estimates of the coefficients of the $J-R+1$ remaining non-constant covariates are given by:

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_r, \hat{\beta}_s\right) = -\left\{ \frac{t_s^{(r+1)}}{t_r^{(r+1)}} \widehat{\mathrm{Var}}\left(\hat{\beta}_s\right) + \sum_{\substack{j=r+1 \\ j \neq s}}^{J} \left[\frac{t_j^{(r+1)}}{t_r^{(r+1)}} \widehat{\mathrm{Cov}}\left(\hat{\beta}_j, \hat{\beta}_s\right) \right] \right\} \qquad (2.12)$$

for $r = 1, 2, \ldots R-1$ and $s = r+1, \ldots J$.

2.3.3 Estimated covariance between the constant and each non-constant covariate

A method of estimating the covariances between the estimated coefficient of the constant covariate and the estimated coefficients of the non-constant covariates was not provided by Petersen and Deddens [9]. To overcome this deficiency, we provide a solution in Theorem 2.2, whose proof is presented in Appendix A.

Theorem 2.2

For a log binomial model with $J \geq 1$ *independent non-constant covariates fitted by the exact method to data having* R $(1 \leq R \leq J)$ *distinct sets of values of the covariates (including the constant) for which the estimated outcome probability evaluated at the ML solution is unity, the estimated covariances between the exact estimate of the coefficient of the constant covariate and exact estimates of the coefficients of the* J *non-constant covariates are given by:*

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_0, \hat{\beta}_j\right) = -\left[x_j^{(1)} \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=1 \\ j_1 \neq j}}^{J} x_{j_1}^{(1)} \widehat{\mathrm{Cov}}\left(\hat{\beta}_j, \hat{\beta}_{j_1}\right) \right] \qquad (2.13)$$

for $j = 1, 2, \ldots J$.

2.3.4 The special case of perfect linear correlation between estimates of model coefficients

We found that in the special case where the number of covariates is the same as the number of boundary vectors $(J = R)$, there is inter-dependency between the exact estimates of the coefficients of the non-constant covariates in the log binomial model. Petersen and Deddens [9] did not comment on this issue. For guidance, we provide Theorem 2.3 and Corollary 2.3.1 with proofs in Appendix A.

Theorem 2.3

For a log binomial model with $J \geq 1$ *independent non-constant covariates fitted by the exact method to data having* R $(R = J)$ *distinct sets of values of the covariates (including the constant) for which the estimated outcome probability evaluated at the ML solution is unity, there is a perfect correlation between each pair of exact estimates of the coefficients of the non-constant covariates:*

$$\left|\rho_{j_1,j_2}\right| = \frac{\left|\widehat{\mathrm{Cov}}(\hat{\beta}_{j_1}, \hat{\beta}_{j_2})\right|}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_{j_1})\widehat{\mathrm{Var}}(\hat{\beta}_{j_2})}} = \frac{\left|\widehat{\mathrm{Cov}}(\hat{\beta}_{j_1}, \hat{\beta}_{j_2})\right|}{\widehat{\mathrm{SE}}(\hat{\beta}_{j_1})\widehat{\mathrm{SE}}(\hat{\beta}_{j_2})} = 1 \qquad (2.14)$$

For $j_1, j_2 = 1, 2, \ldots J; j_1 \neq j_2$ *where* $\hat{\beta}_j$, $j = 1, 2, \ldots J$ *denotes the estimates of the coefficients of the non-constant covariates, and* $\widehat{\mathrm{Var}}(\hat{\beta}_j)$ *and* $\widehat{\mathrm{SE}}(\hat{\beta}_j)$ *denote the estimates of their estimated variance and standard error respectively*.

Corollary 2.3.1

If the log binomial model satisfies the conditions of Theorem 3, the standardised values (*Z statistics*) *of the exact estimates of the* $J + 1$ *model coefficients are equal in absolute size:*

$$\frac{\left|\hat{\beta}_0\right|}{\widehat{\mathrm{SE}}(\hat{\beta}_0)} = \frac{\left|\hat{\beta}_1\right|}{\widehat{\mathrm{SE}}(\hat{\beta}_1)} = \frac{\left|\hat{\beta}_2\right|}{\widehat{\mathrm{SE}}(\hat{\beta}_2)} = \ldots = \frac{\left|\hat{\beta}_J\right|}{\widehat{\mathrm{SE}}(\hat{\beta}_J)} \qquad (2.15)$$

where $\hat{\beta}_0$ *denotes the estimate of the coefficient of the constant covariate, and* $\widehat{\mathrm{SE}}(\hat{\beta}_0)$ *denotes the estimate of its estimated standard error.*

2.3.5 Covariate order when there are multiple distinct boundary vectors

When there are multiple boundary vectors $(R > 1)$, the exact method requires re-parameterisation of covariates according to the scheme $z_{ij}^{(1)} = x_{ij} - x_j^{(1)}$ for $r = 1$ and

$z_{ij}^{(r)} = z_{ij}^{(r-1)} - \left(t_j^{(r)} \Big/ t_{r-1}^{(r)} \right) z_{i,r-1}^{(r-1)}$ for $r = 2,3,...R$. The re-parameterisations for $r = 2,3,...R.$ cannot succeed if $t_{r-1}^{(r)} = 0$ at any step. As an example with $R = 2,$ suppose that the first covariate is a binary variable for sex, and that the first and second boundary vectors represent subjects of the same sex so that $x_1^{(1)} = x_1^{(2)}$. The first re-parameterisation eliminates constant covariate. For the first covariate, this produces $t_1^{(2)} = z_{i1}^{(1)}\Big|_{x_{i1}=x_1^{(2)}} = x_1^{(2)} - x_1^{(1)} = 0$ and results in division by zero in the second re-parameterisation. Petersen and Deddens [9] did not comment on this issue.

Our strategy is to exchange any covariate for which $t_{r-1}^{(r)} = 0$ at any step $r = 2,3,...R$ with any remaining covariate for which $t_j^{(2)} \neq 0.$ This requires re-ordering the covariates before proceeding. But that raises the question of whether it is possible always to find a covariate with nonzero $t_{r-1}^{(r)}$ to replace one with $t_{r-1}^{(r)} = 0.$

To answer the question, we provide Theorem 2.4 and Corollary 2.4.1. The proofs are presented in Appendix A.

Theorem 2.4

For a log binomial model with $J \geq 1$ non-constant covariates fitted by the exact method to data having R $(1 \leq R \leq J)$ distinct sets of values of the covariates (including the constant) for which the estimated outcome probability evaluated at the ML solution is unity, these R distinct sets of values of the covariates are always linearly independent.

Corollary 2.4.1

For a log binomial model with $J \geq 1$ *non-constant covariates fitted by the exact method to data having* R $(1 < R \leq J)$ *distinct sets of values of the covariates (including the constant) for which the estimated outcome probability evaluated at the ML solution is unity, there is at least one covariate for which* $t_{r-1}^{(r)} \neq 0,\ r = 2,3,...R$.

This corollary guarantees that if there are multiple $(R > 1)$ distinct boundary vectors, and $t_{r-1}^{(r)} = 0$ for any $r = 2,3,...R$, it is possible always to find and exchange it with another covariate for which $t_{r-1}^{(r)} \neq 0$.

2.3.6 Choice of the fitting algorithm

As mentioned previously in Section 2.2.3, the major options for fitting a log binomial model are the NR and IRLS algorithms. Each requires evaluation of terms involving $\left[1-\hat{\mu}\left(x_i\right)\right]^{-1}$, posing difficulties in the estimation of boundary solutions at which at least one fitted probability $\hat{\mu}\left(x_i\right)$ is equal to unity. NR additionally requires evaluation of terms involving $\left[1-\hat{\mu}\left(x_i\right)\right]^{-2}$, and is less tractable than IRLS in consequence. However, there is an advantage of NR in estimating the ML solution when the OIM is used. It shows a better stabilisation than IRLS when the iteration process gets close to the ML solution. This leads to a strategy which attempt fits of the log binomial model firstly by IRLS, and secondly by NR. Marschner [8] described the characteristics of NR and IRLS and presented the same strategy in his review paper. For estimation in R, for which IRLS is the fitting algorithm, we used a custom-built NR algorithm with unlimited step-halving.

This dual approach is necessary because neither algorithm is guaranteed to converge. Using IRLS, consider the example data in Table 2.1, which is taken from Williamson et al. [10]. The log binomial model $\Pr(Y_i = 1 | \mathrm{x}_i) = \exp(\beta_0 + \beta_1 x_i),\ i = 1, 2, ...40$ for these data has the ML solution of $\hat{\beta}_0 = -0.708$ and $\hat{\beta}_1 = -0.472$. It lies in the interior of the allowable parameter space, but estimation by IRLS is unsuccessful. Marschner [8] characterised the ML solution as a "repelling" point to which IRLS will not converge and introduced *glm*2 built on R-*glm* to fix the issue. R-*glm*2 solved the issue by merely adding a condition on the step-halving to make it only accept the solution as the log-likelihood increased [58].

For the data in Table 2.1, the IRLS fitting algorithm fails to converge to the ML solution. As Marschner [8] explains, Fisher scoring will not converge irrespective of whether, at any iterative step, the current set of estimates is very close to the ML solution. Marschner states that this is the case even if iterations commence from very good initial values. The starting values I used were $\hat{\beta}_{0,start} = -0.703$ and $\hat{\beta}_{1,start} = -0.597$. The detailed procedure for obtaining them is described in the next section.

Table 2.1. Example data of Williamson et al. [10]

Exposure	Event	No event	Total
$x = -1$	2	2	4
$x = 0$	14	3	17
$x = 1$	2	17	19

2.3.7 Starting values

We found that the standard fitting algorithm may encounter numerical difficulties and fail to self-correct the issue if the iteration begins from an inappropriate starting value, which will lead to some fitted probabilities exceeding unity. We tested the default starting value

algorithm under three statistical packages (SAS, R and Stata). R and Stata share the same algorithm for obtaining starting values, which did not provide a satisfying result mostly in the case that the ML solution lies close to or on the boundary of the parameter space. SAS uses a different algorithm to obtain starting values, but does not attain the ML solution when it is on a boundary although a close approximation is found in some instances.

The approach used is to deduct from the coefficient of the constant (if coefficients are to be supplied as initial values) or from the linear predictor (if fitted values are to be supplied as initial estimates) an amount just adequate to ensure that all initial fitted probabilities are less than unity. For example, the starting values for the data in Table 2.1 obtained from the default algorithm in R are $\hat{\beta}_{0,default} = -0.196$ and $\hat{\beta}_{1,default} = -0.597$. However, these values are inappropriate because they make the fitted probability corresponding to $x = -1$ greater than unity ($\hat{\mu}_{x=-1} = 1.493$). The fitting algorithm in R stops and asks the user to provide appropriate starting values. My starting value algorithm fixes the problem by subtracting the logarithm of the maximum fitted probability ($\hat{\mu}_{x=-1} = 1.493$) from the starting value for the constant β_0 and then adding the logarithm of a value in the range [0, 1] but close to unity. In my starting value algorithm, this value is $\log(0.9)$. The fixed constant of starting values $\hat{\beta}_{0,start}$ is obtained as follows:

$$\begin{aligned}\hat{\beta}_{0,start} &= \hat{\beta}_{0,default} - \log\left[\max\left(\hat{\mu}\right)\right] + \log(0.9) \\ &= -0.196 - 0.401 - 0.105 \\ &= -0.702\end{aligned} \quad .$$

If the standard fitting algorithm begins the iteration from an appropriate starting value, it can converge to a solution with enough step-halving. However, the solution is only an approximate one if the ML solution is on the boundary because the standard fitting algorithm

will have to stop the iteration somewhere before touching the boundary due to the impact of the boundary vector. Therefore, the exact method is still needed to eliminate the impact of the boundary vector and locate the ML solution in this case.

2.3.8 Identifying the boundary vector(s) in advance to apply the exact method

From the outset, Deddens et al. [18] acknowledged a limitation of the exact method. Before it can be used, the covariate vectors with fitted probabilities equal to unity need to be identified. That is unknown in advance of estimation. There may be multiple covariate vectors with approximate fitted probabilities close to unity, and no way of determining with certainty which of them will prove to be the boundary vector(s).

Our strategy to address the issue is as follows:

1. Approximate the ML solution as accurately as possible. (the details are in the next paragraph);
2. Identify all (distinct) covariate vectors with an approximate fitted probability close to unity;
3. Test each candidate boundary vector and boundary vector combination until the exact solution is found. The approximate fitted probabilities (largest first) are used to prioritise the order of selection.

Our strategy for step 1 is software-specific. The implementation of IRLS in R-*glm*, once admissable starting values are supplied, provides estimates that are generally adequate as an approximation of the ML solution. For Stata users, a custom-made Newton-Raphson algorithm with Fisher scoring (asymptotically equivalent to IRLS) and unlimited step-halving is used to provide a reasonable first approximation of the ML solution.

An R package *lbm* based on the exact method for fitting a log binomial model has been released on Github. The related R documents are set out in Section 6.1 of Chapter 6.

2.4 Illustration of our method with an example dataset

2.4.1 The example data

The data includes 11 observations shown in Table 2.2. The response variable Y for each analysis is a binary (0/1) variable. The covariates X_1, X_2 and X_3 are three continuous variables. To estimate the probability of $Y=1$, a log binomial model was fitted, but failed to converge. In what follows, it will be discovered that there are two boundary vectors in this dataset: observations 10 and 11 have fitted probabilities of unity when evaluated at the ML solution.

Table 2.2. Example data

obs	y	x_1	x_2	x_3
1	0	14	3.90	14.500
2	0	22	3.18	4.504
3	0	12	4.72	13.594
4	0	14	4.13	6.303
5	0	18	3.69	4.890
6	1	14	3.42	12.990
7	0	34	1.80	4.425
8	0	18	3.47	4.934
9	1	35	2.05	3.798
10	1	26	1.83	3.895
11	1	17	2.83	9.690

2.4.2 Application of the exact method

When R-*glm* was used to initially fit the data, the fitting algorithm stopped and asked the user to provide appropriate starting values. The starting values used by R-*glm* – $\hat{\beta}_{0,default}=2.299$, $\hat{\beta}_{1,default}=-0.008$, $\hat{\beta}_{2,default}=-1.550$, and $\hat{\beta}_{3,default}=0.225$ – are inappropriate because the fitted probabilities of observations 7 and 10 evaluated at those

initial values exceed unity: $\hat{\mu}_{7,default} = 1.270$ and $\hat{\mu}_{10,default} = 1.145$. To correct this problem and obtain appropriate starting values, the algorithm described in Section 2.3.7 is applied by subtracting the logarithm of $\hat{\mu}_{7,default} = 1.270$ from $\hat{\beta}_{0,default}$ and adding a logarithm of 0.9. The details are as follows:

$$\begin{aligned}\hat{\beta}_{0,start} &= \hat{\beta}_{0,default} - \log\left(\hat{\mu}_{7,default}\right) + \log(0.9) \\ &= 2.299 - \log(1.270) + \log(0.9) \\ &= 1.955\end{aligned} \quad .$$

Thus the corrected starting values were $\hat{\beta}_{0,start} = 1.955$, $\hat{\beta}_{1,start} = -0.008$, $\hat{\beta}_{2,start} = -1.550$, and $\hat{\beta}_{3,start} = 0.225$. Commencing from the corrected starting values, the fitting algorithm of R-*glm* reaches an approximation of the ML solution with estimated coefficients, $\hat{\beta}_{0,approxi} = 6.5206670$, $\hat{\beta}_{1,approxi} = -0.1098078$, $\hat{\beta}_{2,approxi} = -2.5921914$, and $\hat{\beta}_{3,approxi} = 0.2767768$. (Note that the decimal number is increased to distinguish the approximation of the ML solution from the exact solution. In this small example dataset, the two sets of estimates are extremely close but, in other and larger datasets, the difference between them may be more substantial).

The fitted values evaluated at the approximate solution revealed that the covariate vectors $\left(x_0^{(1)}, x_1^{(1)}, x_2^{(1)}, x_3^{(1)}\right) = (1, 17, 2.83, 9.690)$ and $\left(x_0^{(2)}, x_1^{(2)}, x_2^{(2)}, x_3^{(2)}\right) = (1, 26, 1.83, 3.895)$ of observations 11 and 10 respectively (in decreasing order of initial fitted probability) were each a strong candidate to constitute a boundary vector. The data were re-parameterised firstly as:

$$z_{ij}^{(1)} = x_{ij} - x_j^{(1)}, \; i = 1, 2, \ldots 11; \; j = 0, 1, 2, 3$$

to eliminate the constant and observation 11, and secondly as:

$$z_{ij}^{(2)} = z_{ij}^{(1)} - \left(\frac{x_j^{(2)} - x_j^{(1)}}{x_1^{(2)} - x_1^{(1)}} \right) z_{i1}^{(1)}, \;\; i = 1,2,...10; j = 1,2,3$$

to eliminate the first covariate and observation 10. The first covariate is eliminated because:

$$z_{i1}^{(2)} = z_{i1}^{(1)} - \left(\frac{x_1^{(2)} - x_1^{(1)}}{x_1^{(2)} - x_1^{(1)}} \right) z_{i1}^{(1)} = 0, \;\; i = 1,2,...10;$$

The reduced model $\mu\left(\mathbf{z}_i^{(2)}\right) = \exp\left(\sum_{j=2}^{3} \beta_j z_{ij}^{(2)}\right)$ with two covariates (the re-parameterised values of x_2 and x_3) and without a constant was then fitted to the remaining $i = 1,2,...,9$ observations. In confirmation that the two covariate vectors were boundary vectors, this estimation was successful. It resulted in the coefficient estimates $\hat{\beta}_2 = -2.5921916$ and $\hat{\beta}_3 = 0.2767768$, the standard error estimates $\widehat{\mathrm{SE}}\left(\hat{\beta}_2\right) = 1.5029337$ and $\widehat{\mathrm{SE}}\left(\hat{\beta}_3\right) = 0.2508485$, and the covariance estimate $\widehat{\mathrm{Cov}}\left(\hat{\beta}_2, \hat{\beta}_3\right) = -0.3380011$. The maximised value of the log-likelihood was –3.191690. Using $t_1^{(2)} = x_1^{(2)} - x_1^{(1)} = 26 - 17$, $t_2^{(2)} = x_2^{(2)} - x_2^{(1)} = 1.83 - 2.83$ and $t_3^{(2)} = x_3^{(2)} - x_3^{(1)} = 3.895 - 9.690$, the estimate of $\hat{\beta}_1$ was calculated from (2.8) as $\hat{\beta}_1 = -0.1098078$, its standard error was calculated from (2.9) as $\widehat{\mathrm{SE}}\left(\hat{\beta}_1\right) = 0.0749095$, and the covariances with the estimated coefficients of non-eliminated covariates were calculated from (2.12) as (2.16) $\widehat{\mathrm{Cov}}\left(\hat{\beta}_1, \hat{\beta}_2\right) = 0.0333437$ and $\widehat{\mathrm{Cov}}\left(\hat{\beta}_1, \hat{\beta}_3\right) = 0.0029610$. Estimates of the constant and its estimated standard error are derived from (2.10) and (2.11) as $\hat{\beta}_0 = 6.5206677$ and $\widehat{\mathrm{SE}}\left(\hat{\beta}_0\right) = 3.3566005$ respectively. Finally, the estimated covariances of the coefficient of the constant covariate with the estimated coefficients of the non-constant covariates were calculated from (2.13) as $\widehat{\mathrm{Cov}}\left(\hat{\beta}_0, \hat{\beta}_1\right) = -0.2184495$, $\widehat{\mathrm{Cov}}\left(\hat{\beta}_0, \hat{\beta}_2\right) = -3.6840449$ and $\widehat{\mathrm{Cov}}\left(\hat{\beta}_0, \hat{\beta}_3\right) = 0.2964624$.

These estimates of standard error and covariance were evaluated using the OIM. The fitted probabilities for each observation evaluated at the exact estimates are provided in Table A1 of Appendix A.

2.4.3 Results obtained from standard statistical software packages

The standard software packages that have been tested are R-*glm* (version 3.5.0), SAS-*GENMOD* (version 9.4), SPSS-*genlin* (version 24) and Stata-*glm* (version 15.0). The R-*glm* algorithm is an implementation of IRLS. SAS-*GENMOD* uses the NR algorithm but with an option to use Fisher scoring (NR–Fisher). SPSS-*genlin* allows the user to choose between NR, NR–Fisher and a hybrid method. Stata provides Stata-*glm* (NR), Stata-*glm* (NR–Fisher), Stata-*glm* (IRLS), and Stata-*binreg* (IRLS). Stata-*glm* (NR) and Stata-*glm* (NR–Fisher) provide options for three additional fitting algorithms to be used interchangeably with or in place of NR. SAS and Stata allow for the estimates of variance to be based on the expected information matrix (EIM) rather than OIM. All four software packages allow the user to over-ride the default starting values. The current version of R-*glm* uses as default the same starting values used by Stata but, if they are inadmissible, halts execution until admissible starting values are supplied by the user. Stata-*glm* (NR) and Stata-*glm* (NR–Fisher) provide an option that initiates a search for improved starting values.

In order to compare the performance of the approaches from four statistical packages with the exact method, the model is estimated under their default starting value algorithms. For the data in Table 2.2:

1. SAS-*GENMOD* (NR) produced an inadmissible solution with coefficient errors (calculated as percentage differences relative to the ML values of the coefficient) of 8–124%;

2. SPSS-*genlin* excluded observation 10 and 11 from estimation and reached inadmissible solutions with coefficient errors of 36–1015% (NR) or 15-56% (NR–Fisher);
3. Stata-*glm* (NR) and Stata-*glm* (NR–Fisher) failed to converge even when admissible starting values were supplied by the user, or the search option was specified;
4. Stata-*binreg* iterated to an admissible solution but with coefficient errors of 3–211%.
5. Stata-*glm* (IRLS) converged to an inadmissible solution with coefficient errors of 1–12%.

Table 2.3 gives details of the estimated coefficients obtained from each approach. Because the results of R-*glm* (IRLS) were used as an approximate solution to identify the boundary vectors shown in Section 2.2.4, they are not included in Table 2.3 for comparison.

Table 2.3: The estimated coefficients of the model are obtained from the five approaches provided by the three statistical packages.

	Exact method	SPSS (NR-Fisher)	SPSS (NR)	Stata-*binreg*	Stata-*glm* (IRLS)	SAS-*GENMOD*
β_0	6.5207	6.3097	22.9980	20.3244	7.1509	0.5696
β_1	–0.1098	–0.1191	–0.4489	–0.3167	–0.1217	0.0267
β_2	–2.5922	–1.1815	–4.3301	–5.4502	–2.6244	–1.2759
β_3	0.2768	0.0256	0.0121	0.2858	0.2441	0.2535

2.4.4 Approximation of the solution using alternative methods

The modified Poisson approach produces an inadmissible solution with coefficient errors of 3–30%. The EM method implemented through R-*logbin* reaches an admissible set of estimates with coefficient errors of 76–99%, and warns that the MLE is on the boundary of parameter space and cannot report the estimated standard errors. The R-*lbreg*, which is a package based on constrained optimisation released by Andrade and Andrade [26], reaches an admissible solution with coefficient errors of 0.01– 0.02%.

Table 2.4: The estimated coefficients of the model are obtained from three alternative approaches.

	Exact method	Poisson	R-*lbreg*	R-*logbin*
β_0	6.5207	5.6902	6.5211	1.5609
β_1	–0.1098	–0.0766	–0.1098	–0.0207
β_2	–2.5922	–2.6656	–2.5925	–0.7499
β_3	0.2768	0.3127	0.2768	0.0026

2.5 A real world example

In their demonstration of model-building techniques for the logistic regression model, Hosmer et al. [62] used information collected at baseline in the Global Longitudinal Study of Osteoporosis in Women (GLOW) on a sub-sample of $n = 500$ women enrolled at six sites in the United States. The GLOW500 data are available from the following website, John Wiley & Sons, Inc:

http://wiley.mpstechnologies.com/wiley/BOBContent/searchLPBobContent.do

The binary (0/1) outcome variable (*FRACTURE*) is an indicator of fracture in the first year of follow-up. The study factors selected for the demonstration were the subjects' age in years (*AGE*), weight in kgs (*WEIGHT*), height in cms (*HEIGHT*), body mass index in kgs/m² (*BMI*), self-reported risk of fracture (*RATERISK*), and binary (0/1) indicators for any prior fracture since age 45 (*PRIORFRAC*), whether mother had had a fracture (*MOMFRAC*), menopause status at age 44 (*PREMENO*), whether arms are needed to stand from sitting in a chair (*ARMASSIST*), and ever-smoker status (*SMOKE*). In drawing the sub-sample, fractures were over-sampled to produce a fracture proportion of 25% (125/500) compared to around 4% in the full dataset. As a result, associations found in modelling these data may not apply to the entire cohort.

To estimate the risk of fracture given the values of plausible covariates, a log binomial model was estimated using a model-fitting procedure similar to that adopted by Hosmer et al. [62].

Similar conclusions were reached in respect to the scaling of *AGE* and *HEIGHT*, and the details are omitted for brevity. Consistent with the logistic regression analysis of this data subset, the coefficient of *HEIGHT* is negative. In a departure from the logistic analysis, *WEIGHT* was a significant predictor of the risk of *FRACTURE* when adjusted for *HEIGHT*. To capture its quadratic relationship with risk of fracture, the square of *WEIGHT* was included as a covariate. The influence of *WEIGHT* was most pronounced for those whose mothers had a fracture, and the final model includes two interaction terms $WEIGHT \times MOMFRAC$ and $WEIGHT^2 \times MOMFRAC$. An interaction of *WEIGHT* with *MOMFRAC* was not considered for inclusion in the logistic model by Hosmer et al. [62] because *WEIGHT* was not a significant predictor on the logit scale in univariable analysis. The logistic regression model included the product terms $AGE \times PRIORFRAC$ and $MOMFRAC \times ARMASSIST$, but only the former was retained in the log binomial model when the interactions with *WEIGHT* were included. The composite variable *BMI* was not a statistically significant predictor of the risk of *FRACTURE*, and was not included in the model. Nor were the binary covariates *PREMENO* and *SMOKE*.

Table 2.3 shows the log binomial model fitted using the exact method. Of relevance when interpreting the coefficient of the constant, the values of the continuous covariates are mean-centred. The ML solution is –240.1546083, located on the boundary of the parameter space. In order to obtain the approximate solution for identifying the boundary vectors, a user-supplied starting value is attained by using the approach introduced in Section 2.3.7, $\hat{\beta}_{constant} = -1.5932$, $\hat{\beta}_{AGE} = 0.0290$, $\hat{\beta}_{WEIGHT} = -0.0012$, $\hat{\beta}_{WEIGHT^2} = -0.0000$, $\hat{\beta}_{HEIGHT} = -0.0178$, $\hat{\beta}_{PRIORFRAC} = 0.3704$, $\hat{\beta}_{MOMFRAC} = 0.6341$, $\hat{\beta}_{ARMASSIST} = 0.1838$, $\hat{\beta}_{RATERISK} = 0.1694$, $\hat{\beta}_{AGE \times PRIORFRAC} = -0.0300$, $\hat{\beta}_{WEIGHT \times MOMFRAC} = 0.0166$, and

$\hat{\beta}_{WEIGHT^2 \times MOMFRAC} = -0.0027$. There are four boundary vectors. One represents a woman of very short stature who needed to use her arms to stand from sitting in a chair. The other three had suffered a fracture prior to age 45, their mothers had had a fracture, and they each considered their own risk of fracture to be greater than others of the same age. Two of the three needed to use their arms to get out of a chair.

Table 2.5 Results of fitting a log binomial model to the GLOW500 data by the exact method

	Coef.	Std. Err.	*z*	*p*	95% CI
AGE	0.0438	0.0115	3.82	<0.001	(0.0213, 0.0662)
WEIGHT	0.0093	0.0056	1.67	0.095	(–0.0016, 0.0202)
$WEIGHT^2$	–0.0001	0.0002	–0.25	0.802	(–0.0005, 0.0004)
HEIGHT	–0.0428	0.0052	–8.17	<0.001	(–0.0530, –0.0325)
PRIORFRAC	0.6331	0.1459	4.34	<0.001	(0.3472, 0.9190)
MOMFRAC	1.0121	0.1758	5.76	<0.001	(0.6675, 1.3567)
ARMASSIST	0.2524	0.1384	1.82	0.068	(–0.0189, 0.5236)
RATERISK	0.2635	0.1040	2.53	0.011	(0.0595, 0.4674)
AGE×PRIORFRAC	–0.0466	0.0127	–3.65	<0.001	(–0.0716, –0.0216)
WEIGHT×MOMFRAC	0.0200	0.0114	1.76	0.079	(–0.0023, 0.0424)
$WEIGHT^2 \times MOMFRAC$	–0.0036	0.0012	–3.00	0.003	(–0.0060, –0.0013)
constant	–2.3465	0.2324	–10.10	<0.001	(–2.8021, –1.8909)

In comparison, R-*glm* (IRLS) with user-supplied admissible starting values reached an admissible solution with log-likelihood of 240.1546415 and coefficient errors of 0.1–4%. SAS-*genmod* (NR–Fisher) reached an admissible solution with coefficient errors of 0.2–4%, but a "Negative of Hessian not positive definite" warning was issued, and some standard error estimates were not produced. Stata-*glm* (IRLS) failed to converge (default starting values) or converged to an inadmissible solution with coefficient errors of 0.2–1% (user-supplied admissible starting values).

The modified Poisson approach produced an inadmissible solution with coefficient errors of 0.4–63%. After 5088 iterations, R-*logbin* reached an admissible set of estimates with the log-likelihood of –254.9663114 and coefficient errors of 5–100%, but failed to provide the standard error and issued a warning message, "MLE on the boundary of parameter space,

cannot use asymptotic covariance matrix". The R-*lbreg* reached an admissible solution with the log-likelihood of –240.1546645 and coefficient errors of 0.02–0.47%. The log-likelihood was not provided as a reference in the approaches that produced an inadmissible solution.

2.6 Simulations

The simulations were undertaken to confirm that the exact estimates had satisfactory properties in small samples. Each set of simulations involved 10000 replications of a dataset of 500 observations of a binary outcome indicator Y, a dichotomous covariate X_1 and a continuous covariate X_2. The observations are denoted as (y_i, x_{i1}, x_{i2}), $i = 1, 2, ...500$. The design values of the parameters β_0, β_1 and β_2 of the linear predictor $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ used in each set of simulations are shown in Table 2.4. The conclusions are not influenced by the specific values of β_0, β_1 and β_2 chosen. The values x_{i1} of X_1 were drawn at random from a Bernoulli distribution with success probability $p_{x_1} = 0.5$. The values x_{i2} of X_2 were drawn at random from a uniform distribution with limiting values a and b that are respectively the minimum and maximum of $[\ln(0.05) - \beta_0 - \beta_1 \times 0]/\beta_2$ and $[\ln(1) - \beta_0 - \beta_1 \times 1]/\beta_2$. The values y_i of Y were drawn at random from a Bernoulli distribution with success probability $p_y = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$. A realisation with $y_i = 1$ could produce a boundary vector if the largest value $(x_{i1} = 1)$ of X_1 and a large $(x_{i2} \approx b)$ value of X_2 were drawn in settings with $\beta_2 = 0.2$, or if the smallest value $(x_{i1} = 0)$ of X_1 and a small $(x_{i2} \approx a)$ value of X_2 were drawn in settings with $\beta_2 = -0.2$.

Table 2.6: Design of the simulations

Setting	β_0	β_1	β_2	a^*	b^*
1	$\ln(0.1)$	0.6	0.2	8.047190	20.025851
2	$\ln(0.1)$	0.6	–0.2	–2.002585	–0.804719
3	$\ln(0.2)$	0.6	0.2	–3.465736	8.512925
4	$\ln(0.2)$	0.6	–0.2	–0.851293	0.346574
5	$\ln(0.3)$	0.6	0.2	–6.931472	5.047190
6	$\ln(0.3)$	0.6	–0.2	–0.504719	0.693147
7	$\ln(0.4)$	0.6	0.2	–8.958797	3.019864
8	$\ln(0.4)$	0.6	–0.2	–0.301986	0.895880

* Lower and upper limits of the continuous covariate X_2 taking values generated at random from the uniform distribution $U(a,b)$.

2.6.1 Results for coefficient estimates

Table 2.5 shows the simulation results for the slope estimates $\hat{\beta}_1$ and $\hat{\beta}_2$. The results for the intercept $\hat{\beta}_0$ have been omitted for brevity, but they can be found in Appendix A Table A1. Results are shown separately for fits that did not produce a boundary vector and for fits that did. The estimates for fits without a boundary vector are from a standard log binomial model. The estimates for fits with a boundary vector were made using the exact method.

Table 2.7: Simulation results

		Simulations without a boundary vector				Simulations with a boundary vector			
		n	Bias*	MSE†	Coverage‡	*n*	Bias*	MSE†	Coverage‡
1	β_1	4408	-2.287	3.489	94.2	5592	1.065	3.209	94.4
	β_2		-3.579	0.079	95.2		1.117	0.055	94.8
2	β_1	4507	-2.770	3.478	94.5	5493	0.951	3.150	94.9
	β_2		-3.561	0.079	95.4		1.415	0.056	95.2
3	β_1	4536	-2.469	3.383	95.2	5464	1.090	3.222	94.6
	β_2		-3.683	0.078	95.4		1.429	0.057	94.7
4	β_1	4645	-2.772	3.495	94.3	5355	1.155	3.198	95.1
	β_2		-3.589	0.080	95.0		1.197	0.055	95.2
5	β_1	4637	-2.822	3.434	94.5	5363	1.481	3.212	94.6
	β_2		-3.609	0.080	94.8		1.082	0.055	94.9
6	β_1	4575	-2.452	3.412	95.2	5425	1.760	3.193	95.6
	β_2		-3.614	0.078	95.7		1.130	0.054	95.3
7	β_1	4647	-2.215	3.370	94.9	5353	1.648	3.267	94.5
	β_2		-3.530	0.080	95.3		1.285	0.055	95.3
8	β_1	4519	-2.613	3.399	94.7	5481	2.100	3.220	95.0
	β_2		-3.726	0.079	95.5		1.084	0.055	94.9

* Average percent relative bias: $(100/n)\sum_{k=1}^{n}\left[\left(\hat{\beta}_{jk}-\beta_j\right)/\beta_j\right]$, $j=0,1,2; k=1,2,...n$

† 100 times the average mean squared error: $(100/n)\sum_{k=1}^{n}\left[\left(\hat{\beta}_{jk}-\beta_j\right)^2+\widehat{\mathrm{Var}}\left(\hat{\beta}_{jk}\right)\right]$,

$j=0,1,2; k=1,2,...n$.

‡ Percentage of 95 percent confidence intervals covering the design value of the coefficient.

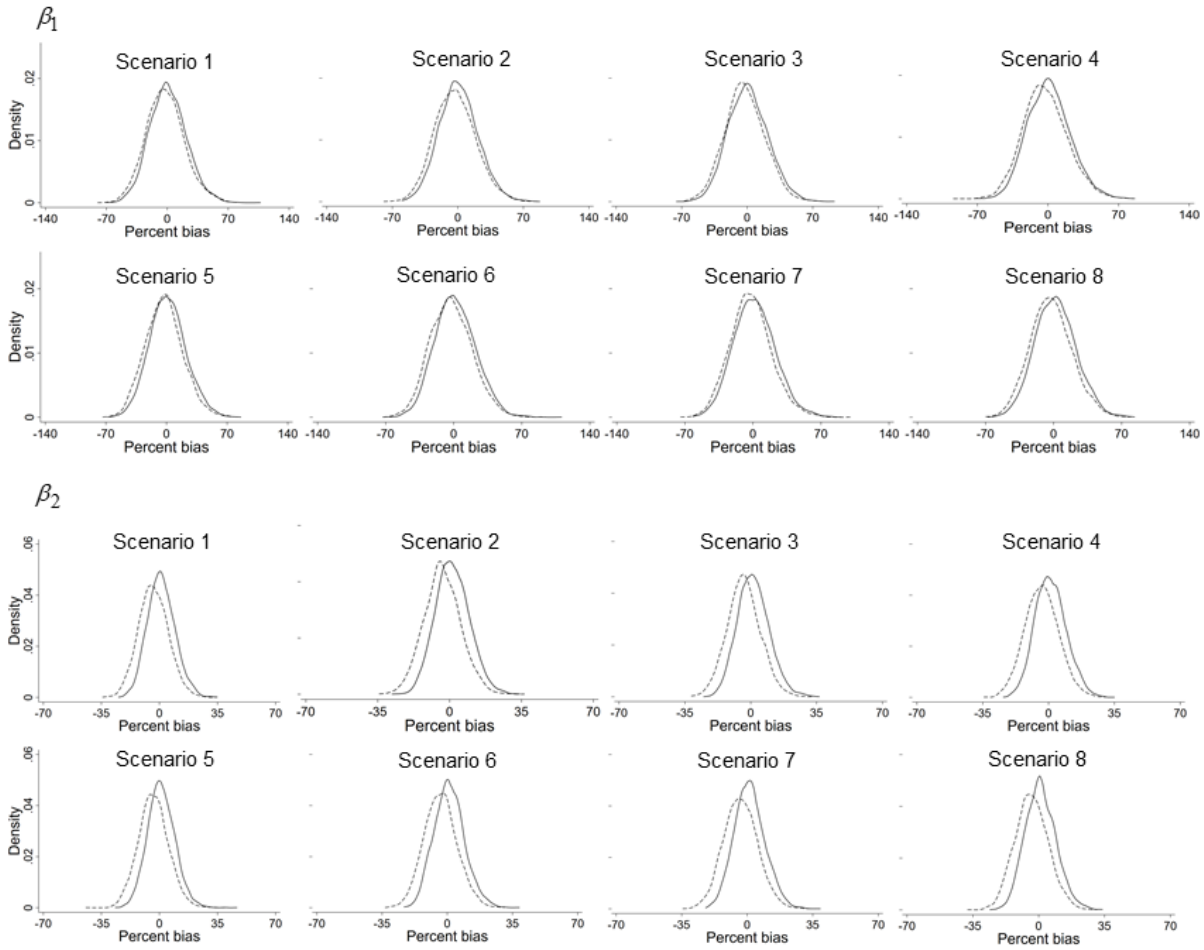

Figure 2.1. Density plots of the sampling distributions (n=500, 10000 replications, 8 settings) of the binary covariate (top panel) and of the continuous covariate (bottom panel). The solid lines represent replications with a boundary vector, the occurrence of which requires both coefficients to be fully estimated. The dash lines represent replications without a boundary vector, which will be the case if at least one of the coefficients is underestimated (negative bias).

Firstly, the results are shown for approximately 44% of repetitions that did not produce a boundary vector. Average percent relative bias was negative for these coefficient estimates, indicating that most were closer to the null than their design values. That is to be expected because those estimates did not produce fitted probabilities large enough to constitute a boundary vector. Secondly, the exact estimates for approximately 55% of repetitions that produced a boundary vector are shown. The average percent bias was small (around 1% in each case) and mildly positive, indicating that they were typically a little further from the null than their design values. Again, that is to be expected, because these coefficients need to be fully estimated in order to produce a maximum fitted probability of unity. This distributional

shift for replications without a boundary vector is displayed in Figure 2.1, which provides density plots of percent bias for the binary and continuous covariates.

Table 2.6 also provides the estimation results by the modified Poisson method and R-*lbreg* for the replications that produced a boundary vector. R-*logbin* was not included in the simulation because it has difficulty producing the estimate of the standard error in the replications when boundary vector is present. The average percent bias in the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ of the slope coefficients are similar in the exact method and R-*lbreg*, but were about 2 to 4 times higher in the modified Poisson method.

Table 2.8: Simulation results.

			Exact method			Modified Poisson			R-*lbreg*		
		n	Bias*	MSE†	Cov‡	Bias*	MSE†	Cov‡	Bias*	MSE†	Cov‡
1	β_1	5592	1.065	3.209	94.4	3.212	3.658	95.4	1.076	3.207	94.5
	β_2		1.117	0.055	94.8	4.100	0.100	95.5	1.138	0.055	94.0
2	β_1	5493	0.951	3.150	94.9	3.129	3.614	95.8	0.962	3.149	94.9
	β_2		1.415	0.056	95.2	4.395	0.100	95.3	1.435	0.055	94.4
3	β_1	5464	1.090	3.222	94.6	3.208	3.681	95.9	1.107	3.223	94.6
	β_2		1.429	0.057	94.7	4.336	0.100	95.4	1.450	0.057	93.9
4	β_1	5355	1.155	3.198	95.1	3.349	3.674	95.4	1.171	3.197	95.1
	β_2		1.197	0.055	95.2	4.104	0.098	95.5	1.224	0.055	94.2
5	β_1	5363	1.481	3.212	94.6	3.582	3.651	95.4	1.496	3.210	94.6
	β_2		1.082	0.055	94.9	4.074	0.099	94.9	1.103	0.054	93.8
6	β_1	5425	1.760	3.193	95.6	3.979	3.632	96.2	1.771	3.191	95.6
	β_2		1.130	0.054	95.3	4.242	0.098	95.7	1.151	0.054	94.2
7	β_1	5353	1.648	3.267	94.5	3.912	3.725	95.1	1.664	3.265	94.5
	β_2		1.284	0.055	95.3	4.352	0.100	95.2	1.307	0.055	94.4
8	β_1	5481	2.100	3.220	95.0	4.309	3.732	95.4	2.119	3.218	94.9
	β_2		1.084	0.055	94.9	4.018	0.099	95.3	1.106	0.054	93.8

* Average percent relative bias: $(100/n)\sum_{k=1}^{n}\left[\left(\hat{\beta}_{jk}-\beta_j\right)/\beta_j\right]$, $j=0,1,2; k=1,2,...n$

† 100 times the average mean squared error: $(100/n)\sum_{k=1}^{n}\left[\left(\hat{\beta}_{jk}-\beta_j\right)^2+\widehat{\mathrm{Var}}\left(\hat{\beta}_{jk}\right)\right]$, $j=0,1,2; k=1,2,...n$.

‡ Percentage of 95 percent confidence intervals covering the design value of the coefficient.

Figure 2.2 provides the density curve of average percent bias in the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ of the slope coefficients for the 10000 replications. The curves coincided in the exact method and R-*lbreg*, but spread out more to both tails in modified Poisson compared with the prior two methods.

We compared the log-likelihood in the exact method and R-*lbreg* for 10000 replications as well. As expected, the log-likelihoods were almost identical in both methods for the replications that did not present a boundary vector. However, they became larger in the exact method for the replications in which a boundary vector was present since the coefficients were fully estimated after eliminating the impact of the boundary vector.

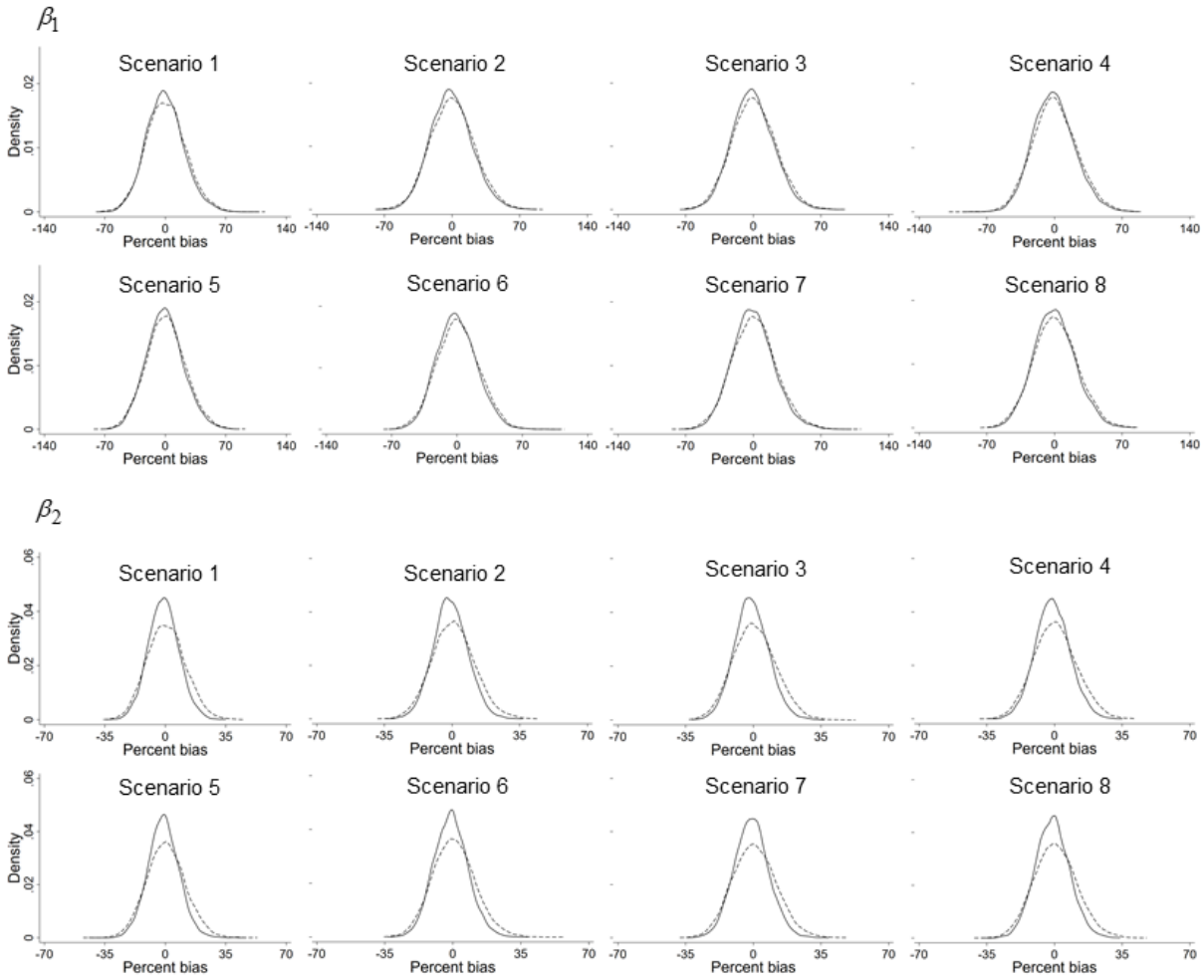

Figure 2.2. Density plots of the sampling distributions (n = 500, 10000 replications, 8 settings) of the binary covariate (top panel) and of the continuous covariate (bottom panel). The solid lines represent the density curve of the average percent bias in the exact method. The dotted lines represent the density curve of average percent bias in R-*lbreg*, which are coincident with the curve of the exact method. The dash lines represent the density curve of the average percent bias in modified Poisson.

2.6.2 Results for mean squared error and confidence interval coverage

Reflecting the lesser contribution of small-sample bias, 100 times the average mean squared errors of exact estimates that produce a boundary vector are lower than those of the fits that do not (Table 2.5). They are consistent with the 100 times the average mean squared errors of estimates in R-*lbreg*, and both are lower than the average mean squared errors of the corresponding modified Poisson estimates (Table 2.6). The 95% confidence interval coverage by the exact method is close to the target (Table 2.5) whereas, for the fits without boundary vectors, the coverage is close to the target but slightly on the conservative side (Table 2.5). These good coverage properties are consistent with the approximate normality of the

sampling distributions displayed in the density plots (see Figure 2.1). For fits with boundary vectors, the coverage is on the conservative side in the modified Poison method, and is slightly lower than the target in the R-*lbreg* (Table 2.6).

2.6.3 Choice of the variance estimator

The estimates of mean squared error and confidence interval coverage in Table 2.6 for the exact method are calculated from variance estimates based on the observed information matrix (OIM). There are multiple other candidates for the choice of variance estimator, including the use of bootstrap or jackknife methods.

Table A3 in Appendix A allows a comparison of 95% confidence interval coverage in each of the 8 settings for variance estimates based on the OIM, the expected information matrix (EIM), and the robust or sandwich estimator of variance [34, 35]. There are only minor differences in coverage, but, in this set of simulations, the OIM estimates had a slight advantage. In some other simulations not reported, we have found the robust estimates to have a slight advantage.

2.7 Conclusion

Fitting a log binomial model using standard software can result in numerical difficulties. If this is due solely to commencing iteration from inadmissible starting values that produce at least one fitted probability in excess of unity, this can be rectified by supplying admissible starting values. If the ML solution lies on the boundary, however, special methods are needed because one or more of the fitted probabilities has a value of unity. To resolve that issue for a model with a single covariate, Deddens et al. [18] proposed an exact method based on re-

parametrisation of the covariate. Subsequently, Petersen and Deddens [9] outlined an extension to general cases, but without providing the details necessary to implement it [26].

The missing details are provided in this chapter. It contains novel contributions:

1. A theorem (with proof) for estimating the covariance of the estimated coefficient of any non-constant covariate excluded from the regression model with the estimated coefficient of each other non-constant covariate;
2. A theorem (with proof) for estimating the covariances between the estimated coefficients of the non-constant covariates and the estimated coefficient of the constant;
3. A theorem (with proof) to verify the perfect correlation between each pair of exact estimates of the coefficients of the covariates when their number is equal to the number of boundary vectors. We also prove the corollary that, in those circumstances, the standardised values of the ML estimates of the coefficients (including that of the constant covariate) are equal in absolute value;
4. A theorem and corollary (with proofs) to establish that the re-parameterisation of covariates required to implement the exact method can be successfully undertaken if the ML solution exists.
5. A strategy for re-ordering the covariates when there are multiple (R) boundary vectors and the first R-1 covariates in the order entered by the user have shared values;
6. A strategy for addressing the issue that the NR and IRLS fitting algorithms are not guaranteed to converge;
7. A strategy for testing whether the initial values are admissible for a probability model, and correcting them if they are not;
8. A strategy for identifying the boundary vectors in advance of estimating the ML solution.

An example dataset and a real-world dataset were used to demonstrate our implementation of the exact method. For each, the ML solution was on the boundary of the parameter space with two distinct boundary vectors in the first example, and four distinct boundary vectors in the real-world second case. The exact method produced the ML solution. The fitting algorithms provided in the four statistical software packages we tested, and either were not successful in estimating any solution or, if convergence was attained, provided approximations that in some cases were poor. The modified Poisson approach [19-21] produced poor approximations. The R-*logbin* method using the EM algorithm [22] successfully reached an approximate solution in these data examples, but failed to provide the estimates of the standard error of coefficients with multiple boundary vectors. The R-*lbreg* method using constrained optimisation [26] reached an approximate solution with the smallest coefficient errors compared with other alternative methods.

The simulations revealed a minimal bias of fits by the exact method to data with a single boundary vector. In our simulations, any "bias" was an artefact of the small sample size and the success or failure of the attempt to produce a boundary vector. The 95% confidence interval coverage was close to the target, with a small advantage to the OIM estimator of variance. In the comparison of regression fits by the modified Poisson to the same datasets, the average percent relative bias of the modified Poisson estimates was up to 4 times higher than the "bias" of the exact method. In addition, the functional form of the Poisson regression model allows the predicted probabilities to exceed unity, which technically is not allowable for a model of probabilities. As previously [37], we recommend against using the modified Poisson approach to estimate risk and risk ratios. Marschner and Gillett [22] reached the same conclusion particularly in scenarios when the average risk is high. Our results suggest that large individual fitted probabilities are the source of the problem. The R-*lbreg* method using

constrained optimisation provides a good approximation for estimating the ML solution in log binomial model. However, since the impact of the boundary vectors is not eliminated, the coefficients of covariates are not fully estimated, which results in a larger log-likelihood compared with the exact method in all replications when a boundary vector was presented.

In conclusion, the exact method provides an elegant and effective way of estimating the log binomial model when the ML solution lies on the boundary of the parameter space. Our software in R and Stata for fitting the model ensures that the fitting algorithm starts from admissible initial values, but defaults to the standard log binomial model when there is no boundary vector. That makes the software seamless to use by a practitioner. It also overcomes an objection to the use of log binomial model raised by Marschner and Gillett [22]. They note that prospective studies – particularly clinical trials – often require pre-specified analysis plans, making it problematic to commit to the log binomial model before one knows it will be successful. No longer is that the case. Our implementation of the exact method should provide complete confidence in using relative risk regression as a primary method of analysis for prospective studies.

Chapter 3 Overcoming numerical difficulties in estimating absolute risk difference by fitting identity-link binomial model

3.1 Introduction

Risk difference in epidemiology is the observed difference in risk of a binary outcome between reference and study groups. It is a measure of absolute effect and is recommended for clinical trials, especially randomised controlled trials, to inform decisions regarding benefit or harm of treatment [4, 27-33]. The risk difference can be estimated using an identity-link binomial model, which is a generalised linear model with a binomial error and an identity-link function. However, two numerical difficulties may affect convergence in the identity-link binomial model. First, the standard fitting algorithm may commence from an inappropriate starting value, which could lead to fitted probabilities go outside the interval [0, 1]. In most conditions, the fitting algorithm will not self-correct the error once the iteration commences from an inappropriate starting value, and will fail to converge or converge to a solution outside the allowable parameter space. This issue can be simply resolved using a starting point calibration method, described in Section 3.2.2. The second numerical difficulty is encountered when the maximum likelihood (ML) solution is on the boundary of parameter space. In this situation, there is at least one observation with the fitted probability that is equal to 0 or 1 in the model. This introduces difficulties in the calculation of the information matrix and the relevant variance-covariance matrix (detailed in Section 3.2.1). To solve this difficulty, a specialised approach is needed.

Several approaches have been introduced for estimating the risk difference. One approach uses an identity-link modified Poisson model with robust sandwich variance estimates to approximately estimate the coefficients of the identity-link binomial model [36]. It is an

extension of a log-link modified Poisson model with robust sandwich variance estimates [19, 20, 34, 35], which is commonly used for estimating the relative risk. However, the estimates obtained by the modified Poisson approach may result in an inappropriate solution with fitted probabilities outside the interval [0, 1] when the ML solution is on the boundary of the allowable parameter space [22, 37]. This issue can take place in estimating identity-link binomial model through the modified Poisson approach also. Moreover, Cheung [39] stated that convergence issues could be encountered in the risk difference estimation using the modified Poisson approach with an identity link as well. Instead, he recommended a modified least-squares regression approach with a robust variance estimate to estimate risk difference. Although there are no convergence issues with modified least-squares regression, the unrestricted fitting algorithm could produce a solution with fitted probabilities beyond the boundary of the interval [0, 1]. Bieler et al. [40] introduced the model-adjusted risk differences method based on logistic regression models using the function of the average marginal prediction defined by Graubard and Korn [41] to estimate risk difference. Although an approximate marginal average risk difference and variance-covariance matrix are obtained, the authors did not compare their approach with other methods, so it is not known whether this approach provides any incremental improvement on previous methods. Marschner [42] stated that the expectation-maximisation (EM) algorithm can somewhat avoid these two issues – the unconstrained parameter space and the failure of convergence – in estimating identity-link binomial models by the identity-link Poisson model. Donoghoe and Marschner [38] subsequently took a further step to resolve the convergence issue in the identity-link binomial model. They used a multinomial-Poisson transformation [43] to convert the problem into an equivalent additive Poisson model, and fit it as an additive binomial model using the EM algorithm to estimate the identity-link binomial model and risk difference. This is referred to as additive Poisson model in the remainder of this chapter. Another two

approaches proposed to approximately estimate RD are the linear-expit model (LEXPIT) combined with adaptive barrier approach [44] and the nuisance model with doubly robust estimator [45]. None of these approaches estimates the RD directly in the identity-link binomial model, and the convergence issue remains unresolved.

As discussed in Chapter 2, Petersen and Deddens [9] introduced an exact method that has addressed the convergence issue in the log binomial model when the ML solution lies on the boundary of the parameter space. The exact method can also provide a solution to convergence issues in the identity-link binomial model due to the similarity in the likelihood function. Notably, the convergence issue is more complicated in the identity-link binomial model, which has both lower and upper boundaries of the parameter space as compared with only an upper boundary in the log binomial model.

To this end, an approach has been developed to overcome the numerical difficulties in the identity-link binomial model that consists of a starting value calibration and a method based on the rationale of the exact method. The following sections in this chapter include: a brief description of the identity-link binomial model and a well-designed starting value calibration method in section 3.2.2; application of the exact method to overcome the numerical difficulties due to the ML solution lying on the boundary of the parameter space in section 3.3; application of the exact method to example data in section 3.4; application of the exact method to real-world data in section 3.5; a comparison of the exact method with alternative approaches using simulation studies in section 3.6. The proofs of the theorems and corollaries in this chapter are reported in Appendix B. An R package *bm* based on the exact method for fitting the identity-link binomial model has been released on Github. The related R documents are set out in Section 6.2 of Chapter 6.

3.2 Overview of identity-link binomial model

3.2.1 The identity-link binomial model

Assume that there are n independent observations with a binary outcome variable Y and J covariates $(1, X_1, X_2, ...X_J)$ with constant. Then, the likelihood function of the identity-link binomial model is defined as:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \mu_i^{y_i} (1-\mu_i)^{(1-y_i)} \tag{3.1}$$

where μ_i is the probability of the outcome conditional on the i^{th} observation which is equal to a linear combination of parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots \beta_J)$ and covariates $\mathbf{x}_{ij} = (1, x_{i1}, x_{i2}, \ldots x_{iJ})$ with constant and denoted as

$$\Pr(Y_i = 1 \mid \mathbf{x}_i) = \mu(\mathbf{x}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_J x_{iJ},\ i = 1, 2, ...n\,. \tag{3.2}$$

To ensure the conditional probability remain inside [0, 1], the allowable parameter space of the identity-link binomial model is defined as

$$\Theta = \{\boldsymbol{\beta} : 0 \le \mathbf{x}_i'\boldsymbol{\beta} \le 1,\ \text{for all } i = 1, 2, ..., n\}\,. \tag{3.3}$$

To simplify the calculation, the likelihood function (3.1) is usually transformed into a log-likelihood function,

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \log(\mu_i) + (1-y_i) \log(1-\mu_i) \right]. \tag{3.4}$$

By taking the first partial derivative of (3.4), we obtain the score function as:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{x_{ij}(y_i - \mu_i)}{\mu_i(1-\mu_i)} \right]. \tag{3.5}$$

The second partial derivative of (3.4) in matrix form known as the Hessian matrix is:

$$\mathbf{H} = \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{j_1} \partial \beta_{j_2}} = \mathbf{X}' \left(\mathbf{diag} \left\{ - \frac{\hat{\mu}_i{}^2 - 2\hat{\mu}_i y_i + y_i}{\left[\mu_i (1 - \mu_i) \right]^2} \right\} \right) \mathbf{X}. \qquad (3.6)$$

The observed information matrix is the negative of the Hessian matrix. It is obtained by substituting the estimate of each fitted probability p_i evaluated at $\hat{\boldsymbol{\beta}}$, written as:

$$\widehat{\mathbf{OI}} = -\mathbf{H} = \mathbf{X}' \left(\mathbf{diag} \left\{ \frac{\hat{\mu}_i{}^2 - 2\hat{\mu}_i y_i + y_i}{\left[\hat{\mu}_i (1 - \hat{\mu}_i) \right]^2} \right\} \right) \mathbf{X}. \qquad (3.7)$$

The relevant expected information matrix of (3.7) is:

$$\widehat{\mathbf{EI}} = -\mathrm{E}(\mathbf{H}) = \mathbf{X}' \left\{ \mathbf{diag} \left[\frac{1}{\hat{\mu}_i (1 - \hat{\mu}_i)} \right] \right\} \mathbf{X}. \qquad (3.8)$$

There are two standard fitting algorithms commonly applied to locate the maximum likelihood estimate (MLE) in the identity-link binomial model: the Newton-Raphson (NR) algorithm which iteratively updates estimates by the product of the score function and the inverse of the information matrix; and iteratively reweighted least squares (IRLS) which is an implementation of the weighted least square algorithm, which is equal to NR algorithm with an expected information matrix specified. Since the iteration of the standard fitting algorithm may sometimes take updated estimates outside the parameter space, truncating fitted values or step-halving or both are generally implemented in statistical packages to resolve the problem. Truncating fitted values truncates the probability to a value just inside the boundary of the parameter space if a fitted probability evaluated on the updated estimates is outside the parameter space Θ [12]. Step-halving repeatedly halves the step size of the regular update of the standard fitting algorithm until the updated estimates are inside the parameter space [6].

Even with the use of these techniques, however, the standard fitting algorithm of an identity-link binomial model may encounter numerical difficulties and fail to converge when the ML

solution $\hat{\boldsymbol{\beta}}$ lies on or near the boundary of parameter space Θ. In these particular cases, at least one fitted probability $\hat{\mu}_i$ evaluated at $\hat{\boldsymbol{\beta}}$ is extremely close or equal to 0 or 1. The product of $\hat{\mu}_i$ and $1-\hat{\mu}_i$ in the denominator of the formulas (3.7) could thus be 0 or a very small number, and the score function and the information matrix cannot be calculated, resulting in a failure of the fitting algorithm.

3.2.2 Starting value calibration (Min-Max normalisation)

As mentioned in the introduction to this chapter, one numerical difficulty is due to the use of an inappropriate starting value in the standard fitting algorithm. Statistical packages such as R, SPSS, Stata and SAS have their own starting value algorithms for the identity-link binomial model. However, they frequently fail to identify an appropriate point to start the iteration. Once the fitting algorithm begins the iteration from an inappropriate starting value, it is usually unable to self-correct the fitting procedure. Resulting estimates may be outside the allowable parameter space Θ, leading to a fitted probability of observation which is outside [0, 1].

To correct a misspecified starting value, I introduce min-max normalisation (also called unity-based normalisation) that is usually applied in machine learning as a calibration algorithm to normalise a set of values into the range [0, 1] [63-65]. For the identity-link binomial model, the fitting procedure can start from either a set of initial probabilities corresponding to each observation, or from a user-provided starting value of the model coefficients. If the fitting procedure starts from a set of initial probabilities, the starting values of the coefficients can be obtained by using ordinary least squares. (Note that there are no special rules for selecting the initial probabilities. Data analysts may choose to devise their own initial probabilities based on their understanding of the data. The software packages SAS, Stata, R

and SPSS use the first step of their default fitting algorithm applied to values of the binary outcome re-scaled to lie within the range 0 to 1. SAS replaces $y = 0$ by $y = 0.1$ and $y = 1$ by $y = 0.9$ for this purpose, whilst R and Stata use $y = 0.25$ and $y = 0.75$ respectively.) If the starting values of the coefficients obtained from the prior steps are outside the parameter space Θ, they can be corrected by the following functions

$$\hat{\beta}'_{0,start} = \frac{\hat{\beta}_{0,start} - \min(\hat{\boldsymbol{\mu}})}{\max(\hat{\boldsymbol{\mu}}) - \min(\hat{\boldsymbol{\mu}})} \text{ and } \hat{\beta}'_{j,start} = \frac{\hat{\beta}_{j,start}}{\max(\hat{\boldsymbol{\mu}}) - \min(\hat{\boldsymbol{\mu}})} \text{ for } j = 1, 2, ...J, \quad (3.9)$$

where $\hat{\beta}_{0,start}$ and $\hat{\beta}_{j,start}$ are the starting values initially obtained from the default starting value algorithm under statistical packages, $\hat{\beta}'_{0,start}$ and $\hat{\beta}'_{j,start}$ are the relevant normalised estimates, and $\hat{\boldsymbol{\mu}}$ is the fitted probabilities vector. This calibration ensures that the fitting algorithm always begins the iteration inside the parameter space Θ. With an appropriate starting value, the IRLS with step-halving can converge to a solution. However, if the ML solution lies on the boundary of the parameter space, the IRLS with step-halving can only converge to an approximation of the ML solution.

In the following section, I will discuss the second numerical difficulty, which is the MLE on the boundary of the parameter space. The exact method will be applied and discussed to improve model fitting.

3.3 Maximum likelihood solution is on the boundary of the parameter space

Previously, an inequality constraint $0 \leq \mathbf{x}'\boldsymbol{\beta} \leq 1$ is used to define the allowable parameter space Θ in the identity-link binomial model. It intuitively shows that there are three situations in which the ML solution could be on the boundary of the parameter space: lying on the upper bound, the lower bound, or on both bounds of the Θ. If the ML solution is on

the lower or upper bounds, there is at least one fitted probability equal to either 0 or 1. Similarly, if the ML solution is on both the lower and upper bounds, there is at least one fitted probability equal to 0 and at least one equal to 1. We refer to any such set of covariate values as a boundary vector. Next, the exact method will be applied to improve the model fitting when the ML solution is on the boundary of the parameter space.

To apply the exact method, the likelihood function is rewritten (3.1) as:

$$l(\boldsymbol{\beta}) = \prod_{Y_i=1} \mu_i \prod_{Y_i=0} 1-\mu_i \tag{3.10}$$

where $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$, for $\mathbf{x}_{ij} = (1, x_{i1} x_{i2}, \ldots x_{iJ})$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots \beta_J)$. We denote the ML solution of the model as $\hat{\boldsymbol{\beta}}$. The fitted probability of observations evaluated at $\hat{\boldsymbol{\beta}}$ is denoted as $\hat{\mu}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$. Then, the likelihood function evaluated at $\hat{\boldsymbol{\beta}}$ becomes

$$\begin{aligned} l(\hat{\boldsymbol{\beta}}) &= \prod_{Y_i=1} \hat{\mu}_i \prod_{Y_i=0} 1-\hat{\mu}_i \\ &= \prod_{Y_i=1} \left(\hat{\beta}_0 + \sum_{j=1}^{J} x_{ij}\hat{\beta}_j \right) \prod_{Y_i=0} \left[1 - \left(\hat{\beta}_0 + \sum_{j=1}^{J} x_{ij}\hat{\beta}_j \right) \right] \end{aligned} \tag{3.11}$$

Petersen and Deddens [9] effectively addressed the problem of the ML solution lying on the boundary of the parameter space in the log binomial model by using the boundary vectors to re-parameterise the covariates so that the adverse effect of the boundary vectors on the fitting procedure is eliminated. This method can be applied to the first two cases: where the ML solution is on the upper bound of the parameter space (Section 3.3.1), and where the ML solution is on the lower bound of the parameter space (Section 3.3.2).

3.3.1 The ML solution is on the upper bound of the parameter space

Theorem 3.1

Suppose that the ML solution is on the upper bound of parameter space, which means the maximum fitted probability is equal to unity, $\max(\hat{\mu}_i)=1$ *for* $i=1,2,...n$ *. Assume that there are* $R\geq 1$ *distinct sets of covariate vectors with the fitted probability attained unity referred to as boundary vectors. Denote the* r^{th} $(r=1,2,...R)$ *boundary vector as* $\mathbf{x}^{(r)}=\left(1,x_1^{(r)}.x_2^{(r)},...x_J^{(r)}\right)$ *which shares covariate values with* n_r *observations. Then, the constant and the first* $R-1$ *covariates in the likelihood function* (3.11) *of the identity-link binomial model evaluated at the maximum likelihood solution* $\hat{\boldsymbol{\beta}}$ *can be re-parametrised as follows:*

$$l\left(\hat{\boldsymbol{\beta}}\right)=\prod_{\substack{Y_i=1,x_{ij}\neq x_j^{(1)},\\ x_{ij}\neq x_j^{(2)},...,\\ x_{ij}\neq x_j^{(R)}}}\hat{\mu}_i^*\prod_{Y_i=0}1-\hat{\mu}_i^* \tag{3.12}$$

where $\hat{\mu}_i^*=1+\sum_{j=R}^{J}z_{ij}^{(R)}\beta_j$ *. The function of each* $z_{ij}^{(r)}$ *is:*

$$z_{ij}^{(1)}=x_{ij}-x_j^{(1)},\ i=1,2,...n \text{ and } j=0,1,2,...J. \tag{3.13}$$

$$z_{ij}^{(r)}=z_{ij}^{(r-1)}-\left(\frac{t_j^{(r)}}{t_{r-1}^{(r)}}\right)z_{i,r-1}^{(r-1)} \text{ where } t_j^{(r)}=z_{ij}^{(r-1)}\Big|_{x_{ij}=x_j^{(r)}},\ r=2,3,...R. \tag{3.14}$$

After re-parameterisation, dropping the observations with covariate values $\mathbf{x}^{(1)},\mathbf{x}^{(2)},...,\mathbf{x}^{(R)}$ and the relevant $n_1,n_2,...n_R$ observations who share the same values, which make no contribution to the log-likelihood since they became zero vectors. Then, fitting the model $\hat{\mu}_i^*=1+\sum_{j=R}^{J}z_{ij}^{(R)}\beta_j$ – with an offset 1, without a constant, and with $J-R+1$ covariates – to the remaining $n-n_1-n_2-...-n_R$ observations. This enables us to obtain the estimates

$\hat{\beta}_R, \hat{\beta}_{R+1}, \ldots \hat{\beta}_J$ for the non-eliminated non-constant covariates. To obtain the estimates for the coefficients of $R-1$ re-parameterised non-constant covariates and corresponding standard errors and the covariances, Theorem 3.2 is introduced.

Theorem 3.2

For an identity-link binomial model with $J \geq 1$ *independent non-constant covariates fitted by the exact method to data having* R $(1 \leq R \leq J)$ *distinct sets of values of the covariates (including the constant) for which the estimated outcome probability evaluated at the ML solution is unity, the estimated coefficients and the corresponding standard errors of the* $R-1$ *eliminated non-constant covariates are given by*:

$$\hat{\beta}_r = -\sum_{j=r+1}^{J} \frac{t_j^{(r+1)}}{t_r^{(r+1)}} \hat{\beta}_j, \text{ for } r = 1, 2, \ldots R-1 \qquad (3.15)$$

and

$$\widehat{\mathrm{SE}}\left(\hat{\beta}_r\right) = \sqrt{\sum_{j=r+1}^{J} \left(\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=r+1 \\ j_1 \neq j_2}}^{J} \sum_{j_2=r+1}^{J} \left\{\frac{t_{j_1}^{(r+1)} t_{j_2}^{(r+1)}}{\left(t_r^{(r+1)}\right)^2}\right\} \widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1}, \hat{\beta}_{j_2}\right)}, \qquad (3.16)$$

and the estimated covariances between exact estimates of the coefficients of the $R-1$ *eliminated non-constant covariates and exact estimates of the coefficients of the* $J-R+1$ *remaining non-constant covariates are given by*:

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_r, \hat{\beta}_s\right) = -\left\{\frac{t_s^{(r+1)}}{t_r^{(r+1)}} \widehat{\mathrm{Var}}\left(\hat{\beta}_s\right) + \sum_{\substack{j=r+1 \\ j \neq s}}^{J} \left[\frac{t_j^{(r+1)}}{t_r^{(r+1)}} \widehat{\mathrm{Cov}}\left(\hat{\beta}_j, \hat{\beta}_s\right)\right]\right\}, \qquad (3.17)$$

for $r = 1, 2, \ldots R-1$ *and* $s = r+1, \ldots J$.

To obtain estimates for the constant, its standard error, and covariances between the constant and non-constant coefficients, Theorem 3.3 is introduced.

Theorem 3.3

For an identity-link binomial model with $J \geq 1$ *independent non-constant covariates fitted by the exact method to data having* R $(1 \leq R \leq J)$ *distinct sets of values of the covariates (including the constant) for which the estimated outcome probability evaluated at the ML solution is unity, the estimated coefficient and the corresponding standard error of the constant covariate are given by:*

$$\hat{\beta}_0 = 1 - \sum_{j=1}^{J} x_j^{(1)} \hat{\beta}_j \tag{3.18}$$

and

$$\widehat{\mathrm{SE}}\left(\hat{\beta}_0\right) = \sqrt{\sum_{j=1}^{J}\left(x_j^{(1)}\right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=1 \\ j_1 \neq j_2}}^{J} \sum_{j_2=1}^{J} x_{j_1}^{(1)} x_{j_2}^{(1)} \widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1}, \hat{\beta}_{j_2}\right)}, \tag{3.19}$$

and the estimated covariances between the exact estimate of the coefficient of the constant covariate and exact estimates of the coefficients of the J *non-constant covariates are given by*:

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_0, \hat{\beta}_j\right) = -\left[x_j^{(1)} \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=1 \\ j_1 \neq j}}^{J} x_{j_1}^{(1)} \widehat{\mathrm{Cov}}\left(\hat{\beta}_j, \hat{\beta}_{j_1}\right) \right] \tag{3.20}$$

for $j = 1, 2, \ldots J$.

The detailed proofs of Theorem 3.1, 3.2, and 3.3 are provided in Appendix B.

3.3.2 The ML solution is on the lower bound of the parameter space

Theorem 3.4

Suppose that the ML solution is on the lower bound of parameter space, which means the minimum fitted probability is equal to zero, $\min\left(\hat{\mu}_i\right) = 0$ *for* $i = 1, 2, \ldots n$. *Assume that there are* $R \geq 1$ *distinct sets of covariate vectors with the fitted probability attained zero referred to*

as boundary vectors. Denote the r^{th} $(r = 1, 2, ...R)$ *boundary vector as* $\mathbf{x}^{(r)} = \left(1, x_1^{(r)}.x_2^{(r)}, ...x_J^{(r)}\right)$ *which shares covariate values with* n_r *observations. Then, the constant and the first* $R-1$ *covariates in the likelihood function* (3.11) *of the identity-link binomial model evaluated at the maximum likelihood solution* $\hat{\boldsymbol{\beta}}$ *can be re-parametrised as follows*:

$$l\left(\hat{\boldsymbol{\beta}}\right) = \prod_{Y_i=1} \hat{\mu}_i^* \prod_{\substack{Y_i=0, x_{ij} \neq x_j^{(1)}, \\ x_{ij} \neq x_j^{(2)}, \cdots \\ x_{ij} \neq x_j^{(R)}}} 1 - \hat{\mu}_i^* \tag{3.21}$$

where $\hat{\mu}_i^* = \sum_{j=R}^{J} z_{ij}^{(R)} \beta_j$ *. The function of each* $z_{ij}^{(r)}$ *is*:

$$z_{ij}^{(1)} = x_{ij} - x_j^{(1)},\ i = 1, 2, ...n \text{ and } j = 0, 1, 2, ...J\,. \tag{3.22}$$

$$z_{ij}^{(r)} = z_{ij}^{(r-1)} - \left(\frac{t_j^{(r)}}{t_{r-1}^{(r)}}\right) z_{i,r-1}^{(r-1)} \text{ where } t_j^{(r)} = z_{ij}^{(r-1)}\Big|_{x_{ij}=x_j^{(r)}},\ r = 2, 3, ...R \tag{3.23}$$

After re-parameterisation, dropping the observations with covariate values $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ...\mathbf{x}^{(R)}$ and the relevant $n_1, n_2, ...n_R$ observations who share the same values, which make no contribution to the log-likelihood since they became zero vectors. Then, fitting the model $\hat{\mu}_i^* = \sum_{j=R}^{J} z_{ij}^{(R)} \beta_j$ – without a constant and with $J - R + 1$ covariates – to the remaining $n - n_1 - n_2 - ... - n_R$ observations. This enables us to obtain the estimates $\hat{\beta}_R, \hat{\beta}_{R+1}, ...\hat{\beta}_J$ for the non-eliminated non-constant covariates. To obtain the estimates for the coefficients of $R-1$ re-parameterised non-constant covariates and the relevant estimates of standard errors and the covariances, Theorem 3.5 is introduced.

Theorem 3.5

For an identity-link binomial model with $J \geq 1$ *independent non-constant covariates fitted by the exact method to data having* R $(1 \leq R \leq J)$ *distinct sets of values of the covariates (including the constant) for which the estimated outcome probability evaluated at the ML solution is zero, the estimated coefficients and the corresponding standard errors of the* $R-1$ *eliminated non-constant are given by:*

$$\hat{\beta}_r = -\sum_{j=r+1}^{J} \frac{t_j^{(r+1)}}{t_r^{(r+1)}} \hat{\beta}_j, \text{ for } r = 1, 2, \ldots R-1 \qquad (3.24)$$

and

$$\widehat{\mathrm{SE}}\left(\hat{\beta}_r\right) = \sqrt{\sum_{j=r+1}^{J} \left[\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\right]^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=r+1 \\ j_1 \neq j_2}}^{J} \sum_{j_2=r+1}^{J} \left\{\frac{t_{j_1}^{(r+1)} t_{j_2}^{(r+1)}}{\left[t_r^{(r+1)}\right]^2}\right\} \widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1}, \hat{\beta}_{j_2}\right)}, \quad (3.25)$$

and the estimated covariances between the exact estimates of the coefficients of the $R-1$ *eliminated non-constant covariates and the exact estimates of the coefficients of the* $J-R+1$ *remaining non-constant covariates are given by*:

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_r, \hat{\beta}_s\right) = -\left\{\frac{t_s^{(r+1)}}{t_r^{(r+1)}} \widehat{\mathrm{Var}}\left(\hat{\beta}_s\right) + \sum_{\substack{j=r+1 \\ j \neq s}}^{J} \left[\frac{t_j^{(r+1)}}{t_r^{(r+1)}} \widehat{\mathrm{Cov}}\left(\hat{\beta}_j, \hat{\beta}_s\right)\right]\right\}, \qquad (3.26)$$

for $r = 1, 2, \ldots R-1$ *and* $s = r+1, \ldots J$.

To obtain estimates for the constant, its standard error, and covariances between the constant and non-constant coefficients, Theorem 3.6 is introduced.

Theorem 3.6

For an identity-link binomial model with $J \geq 1$ *independent non-constant covariates fitted by the exact method to data having* R $(1 \leq R \leq J)$ *distinct sets of values of the covariates (including the constant) for which the estimated outcome probability evaluated at the ML solution is zero, the estimated coefficient and the corresponding standard error of the constant covariate are given by:*

$$\hat{\beta}_0 = -\sum_{j=1}^{J} x_j^{(1)} \hat{\beta}_j \tag{3.27}$$

and

$$\widehat{\mathrm{SE}}\left(\hat{\beta}_0\right) = \sqrt{\sum_{j=1}^{J}\left(x_j^{(1)}\right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=1 \\ j_1 \neq j_2}}^{J} \sum_{j_2=1}^{J} x_{j_1}^{(1)} x_{j_2}^{(1)} \widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1}, \hat{\beta}_{j_2}\right)} \tag{3.28}$$

and the estimated covariances between the exact estimate of the coefficient of the constant covariate and the exact estimates of the coefficients of the J *non-constant covariates are given by*:

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_0, \hat{\beta}_j\right) = -\left[x_j^{(1)} \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=1 \\ j_1 \neq j}}^{J} x_{j_1}^{(1)} \widehat{\mathrm{Cov}}\left(\hat{\beta}_j, \hat{\beta}_{j_1}\right) \right] \tag{3.29}$$

for $j = 1, 2, \ldots J$.

The detailed proof of Theorem 3.4, 3.5 and 3.6 are provided in Appendix B.

3.3.3 The special case of perfect linear correlation between estimates of model coefficients

When the ML solution is on the boundary, there is an inter-dependency between the estimates of the coefficients of the non-constant covariates in the identity-link binomial model if the number of covariates is equal to the number of boundary vectors $(J = R)$. For guidance, Theorem 3.7 and Corollary 3.7.1 are provided. The details of the proof are in Appendix B.

Theorem 3.7

For an identity-link binomial model with $J \geq 1$ *independent covariates fitted by the exact method to data having* R $(R = J)$ *distinct sets of values of the covariates for which the estimated outcome probability is unity when the ML solution is on the upper boundary of parameter space, or is zero when the ML solution is on the lower boundary, there is a perfect correlation between each pair of exact estimates of the non-constant coefficients:*

$$\left|\rho_{j_1,j_2}\right| = \frac{\left|\widehat{\mathrm{Cov}}(\hat{\beta}_{j_1}, \hat{\beta}_{j_2})\right|}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_{j_1})\widehat{\mathrm{Var}}(\hat{\beta}_{j_2})}} = \frac{\left|\widehat{\mathrm{Cov}}(\hat{\beta}_{j_1}, \hat{\beta}_{j_2})\right|}{\widehat{\mathrm{SE}}(\hat{\beta}_{j_1})\widehat{\mathrm{SE}}(\hat{\beta}_{j_2})} = 1 \qquad (3.30)$$

for $j_1, j_2 = 1, 2, \ldots J; j_1 \neq j_2$, where $\hat{\beta}_j,\ j = 1, 2, \ldots J$ denotes estimates of the non-constant coefficients. $\widehat{\mathrm{Var}}(\hat{\beta}_j)$ and $\widehat{\mathrm{SE}}(\hat{\beta}_j)$ denote the estimates of their estimated variance and standard error respectively.

Corollary 3.7.1

If the identity-link binomial model satisfies the conditions of Theorem 3.7, the standardised values of the exact estimates of the J *model coefficients are equal in absolute size:*

$$\frac{\left|\hat{\beta}_1\right|}{\widehat{\mathrm{SE}}(\hat{\beta}_1)} = \frac{\left|\hat{\beta}_2\right|}{\widehat{\mathrm{SE}}(\hat{\beta}_2)} = \ldots = \frac{\left|\hat{\beta}_J\right|}{\widehat{\mathrm{SE}}(\hat{\beta}_J)} \qquad (3.31)$$

where $\hat{\beta}_j,\ j = 1, 2, \ldots J$ *denotes estimates of the coefficients of the non-constant covariates, and* $\widehat{\mathrm{SE}}(\hat{\beta}_j)$ *denotes the estimate of their corresponding standard errors.*

3.3.4 Covariate order when there are multiple distinct boundary vectors

Assume there are $R > 1$ boundary vectors in a dataset. The exact method requires re-parameterisation of covariates as $z_{ij}^{(1)} = x_{ij} - x_j^{(1)}$ for $r = 1$ and $z_{ij}^{(r)} = z_{ij}^{(r-1)} - \left[t_j^{(r)} \Big/ t_{r-1}^{(r)} \right] z_{i,r-1}^{(r-1)}$ for $r = 2,3,...R$. If $t_{r-1}^{(r)} = 0$ happens at any step, the re-parameterisation fails and cannot continue. For an example with $R = 2$, suppose that sex x_1 is a binary covariate in the model. The relevant observations of the first and second boundary vectors share the same gender so that $x_1^{(1)} = x_1^{(2)}$. The first re-parameterisation eliminates the constant covariate in the model by one of the boundary vectors. It also produces $t_1^{(2)} = z_{i1}^{(1)}\Big|_{x_{i1}=x_1^{(2)}} = x_1^{(2)} - x_1^{(1)} = 0$ in another boundary vector and results in division by zero in the second re-parameterisation. To solve this issue, our strategy is to exchange the value of any covariate for which $t_{r-1}^{(r)} = 0$ at any step $r = 2,3,...R$ with the value of any remaining non-zero covariate. This requires re-ordering the covariates before proceeding. But that raises the question of whether it is possible always to find a covariate with non-zero $t_{r-1}^{(r)}$ to replace the one with $t_{r-1}^{(r)} = 0$. For reassurance, Theorem 3.8 and Corollary 3.8.1 are provided. The proofs are presented in Appendix B.

Theorem 3.8

For an identity-link binomial model with $J \geq 1$ non-constant covariates fitted by the exact method to data having R $(1 \leq R \leq J)$ distinct sets of values of the covariates (including the constant) for which the estimated outcome probability is unity when ML solution is on the upper boundary of parameter space, or is zero when ML solution is on the lower boundary, these R distinct sets of values of the covariates are always linearly independent.

Corollary 3.8.1

For an identity-link binomial model with $J \geq 1$ *non-constant covariates fitted by the exact method to data having* R $(1 < R \leq J)$ *distinct sets of values of the covariates (including the constant) for which the estimated outcome probability is unity when ML solution is on the upper boundary of parameter space, or is zero when ML solution is on the lower boundary, there is at least one covariate for which* $t_{r-1}^{(r)} \neq 0,\ r = 2,3,...R$.

This corollary guarantees that if there are multiple $(R > 1)$ distinct boundary vectors, and $t_{r-1}^{(r)} = 0$ for any $r = 2,3,...R$, it is possible always to find and exchange another covariate for which $t_{r-1}^{(r)} \neq 0.$ The detailed proof of the corollary is in Appendix B.

3.3.5 Identifying the boundary vector(s) in advance of applying the exact method

For a given set of data, it is usually unknown ahead of estimation whether the ML solution is on the boundary of the parameter space Θ in the identity-link binomial model. Also, a reliable strategy is required to identify boundary vectors from approximate fitted probabilities close to 0 or 1. As described at the beginning of Section 3.3, the ML solution could lie on the upper bound, the lower bound, or both. Accordingly, my strategy to address the problem consists of the following two parts, dealing with different situations of boundary vectors.

The first part of the strategy is regarding the case in which the ML solution only lies on the lower bound or upper bound, where the fitted probabilities of the boundary vectors are equal to 0 or equal to 1 respectively.

(a) Obtain an appropriate starting value and fit the model through IRLS with step-halving to obtain a solution. However, the solution may just be an approximate solution rather than the ML solution if a fitted probability is extremely close to either 0 or 1.

(b) Identify all (distinct) covariate vectors with approximate fitted probabilities from step (a) close to 0 or 1. These are candidate boundary vectors.

(c) Separate all candidate boundary vectors into lower and upper bound groups according to whether the approximate fitted probability is close to 0 or 1. Rank the candidates in the order closest to 0 or 1 in each group. Test each as a candidate boundary vector, and combinations of them as a combination of boundary vectors, until the ML solution is found. The approximate fitted probabilities (closest to 0 or 1) are used to prioritise the order of selection.

With the first part of the strategy, the exact method can effectively locate the ML solution when it is either on the lower or upper bounds of the parameter space.

The second part of the strategy is regarding the case in which the ML solution lies on both lower and upper bounds, where some of the boundary vectors with fitted probabilities equal to 0 and the other boundary vectors with fitted probabilities equal to 1.

When the ML solution lies on both lower and upper bounds, the exact method cannot fully estimate the model. The method is not able to simultaneously deal with the situation where one fitted probability is equal to 0 and another is equal to 1. This is because the likelihood function is not able to be simplified into an appropriate form to be fitted in the fitting algorithm by the re-parameterisation of the covariates. When the ML solution lies on both lower and upper bounds, the exact method can only select the boundary vector from one side of the boundary to re-parameterise the covariates, but cannot choose both for re-

parameterisation at the same time. After the exact method eliminates all the boundary vectors on the selected bound in the data, the IRLS with step-halving would be able to converge to an approximate solution. This approximate solution can be a candidate solution. Once both candidate solutions have been obtained by the upper or lower bound hypothesis, the final solution will be the one that maximises the log-likelihood.

The solution obtained by the exact method is an approximate solution in the case with the presence of both lower and upper boundary vectors. (It may not be appropriate to name this method an exact method since it can only locate an approximate solution when the ML solution lies on both bounds. In order to maintain the consistency of the whole thesis, I prefer to keep the exact method as the name of the method.) This is because the solution is obtained by only eliminating the boundary vectors on one side of lower or upper bounds, the other side of the boundary vectors are still affecting the estimation of the model. Therefore, the coefficients are not fully estimated. Although it is an approximate solution, compared with other approximate methods, its performance in model fit is worth exploring because it is obtained by eliminating one side of the boundary vector to reduce the influence on the model estimation. Further discussions are presented in sections 3.5 and 3.6 with a real-world example and a simulation study.

3.3.6 Selection of information matrix

In some cases, using the expected information matrix to estimate the variance-covariance matrix may cause it to be underestimated. For example, in data, the ML solution lies on both lower and upper bounds of parameter space. Assume that the IRLS with step-halving locates an approximate solution by eliminating the boundary vectors on the upper bound using the exact method. Since the boundary vectors on the lower bound are still involved in the data,

the relevant approximate fitted probabilities may have minimal values which are extremely close to 0. These fitted probabilities with extremely small values will result in a set extremely large numbers in the Hessian matrix $\mathbf{X}'\left(\mathbf{diag}\left\{-\left(\mu_i^2-2\mu_i y_i+y_i\right)\Big/\left[\mu_i\left(1-\mu_i\right)\right]^2\right\}\right)\mathbf{X}$. Under this circumstance, those very large numbers will dominate the weights in the calculation of the expected information matrix (EIM) $\mathbf{X}'\left(\mathbf{diag}\left\{1\Big/\left[\hat{\mu}_i\left(1-\hat{\mu}_i\right)\right]\right\}\right)\mathbf{X}$.

The same problem will happen if an approximate solution is obtained by eliminating the boundary vectors on the lower bound using the exact method. In this case, since the boundary vectors on the upper bound are still involved in the data, the relevant approximate fitted probabilities may have values which are extremely close to 1. Those fitted probabilities will also make $1\Big/\left[\hat{\mu}_i\left(1-\hat{\mu}_i\right)\right]$ become an extremely large value so that it dominates the weights in the calculation of EIM, influencing the estimation of the variance-covariance matrix.

In contrast, the observed information matrix (OIM) does not have this problem. The diagonal matrix $\mathbf{X}'\left(\mathbf{diag}\left\{\left(\hat{\mu}_i^2-2\hat{\mu}_i y_i+y_i\right)\Big/\left[\hat{\mu}_i\left(1-\hat{\mu}_i\right)\right]^2\right\}\right)\mathbf{X}$ in the formula of OIM will simplify to $\mathbf{diag}\left[1\Big/\left(1-\hat{\mu}_i\right)^2\right]$ when $y=0$ and $\mathbf{diag}\left[1\Big/\hat{\mu}_i^2\right]$ when $y=1$. (Note that the observation with an extremely small fitted probability is almost impossible to have $y=1$. Similarly, the observation with a fitted probability extremely close to 1 is almost impossible to have $y=0$.) Returning to the example at the beginning of this subsection, the observation with an extremely small fitted probability close to 0 most likely will have $y=0$. Therefore, the formula of OIM will be simplified as $\mathbf{diag}\left[1\Big/\left(1-\hat{\mu}_i\right)^2\right]$ so that a fitted probability that is close to 0 can only make $1\Big/\left(1-\hat{\mu}_i\right)^2$ result in a number that is quite close to 1, which will not

interfere with the calculation of the OIM and influence the estimation of the relevant standard error.

The influence of those inaccurate EIM on the estimation of the standard error is not clear and requires further study, but it will not be discussed in this thesis. However, in my study, some cases have been met that the standard error obtained by EIM will be very small, and should be questioned. Therefore, I recommend using the observed information matrix to estimate the standard error of the coefficient in the identity-link binomial model.

3.4 Demonstration of the identity-link binomial model using the exact method with an example dataset

3.4.1 The example data

For ease of demonstration, the example data consists of the nine observations shown in Table 3.1. This small dataset suffices to reveal the problem that can arise in the larger datasets found in real-world applications (see Section 5). It is an example in which the ML solution is on the lower bound of the parameter space with two boundary vectors. The response variable Y is a binary (0/1) variable. The covariates X_1, X_2, and X_3 are three continuous variables. To estimate the probability of $Y=1$, an identity-link binomial model with the exact method was fitted. In what follows, it will be discovered that there are two boundary vectors in this data: observations 3 and 4 have fitted probabilities of 0 when evaluated at the ML solution. The results of the exact method will be compared with four statistical packages and two alternative methods.

Table 3.1. Example data

obs	y	x_1	x_2	x_3
1	0	14	3.90	14.500
2	0	22	3.18	4.504
3	0	12	4.72	13.594
4	0	18	3.69	4.890
5	1	14	3.42	12.990
6	0	34	1.80	4.425
7	0	18	3.47	4.934
8	1	35	2.05	3.798
9	1	26	1.83	3.895

3.4.2 Application of the exact method

The standard fitting algorithm in R-*glm*, which is based on IRLS with step-halving, is used to obtain an approximate solution. For these data, the fitting algorithm stops on the first iteration and asks the user to provide appropriate starting values. The problem is that the starting values used by R-*glm* – $\hat{\beta}_{0,default} = 2.839$, $\hat{\beta}_{1,default} = -0.036$, $\hat{\beta}_{2,default} = -0.661$, and $\hat{\beta}_{3,default} = 0.043$ – are inappropriate because the fitted probabilities of observations 3 and 4 evaluated at those initial values are less than zero: $\hat{\mu}_{3,default} = -0.129$ and $\hat{\mu}_{4,default} = -0.033$. To correct this and obtain appropriate starting values, the algorithm mentioned in Section 3.2.2 is applied as follows:

$$\hat{\beta}_{0,start} = \frac{\left(\hat{\beta}_{0,default} - \hat{\mu}_{3,default}\right)\times 0.9}{\max(\hat{\boldsymbol{\mu}}) - \hat{\mu}_{3,default}} + 0.01 = \frac{(2.839 + 0.129)\times 0.9}{0.868 + 0.129} + 0.01 = 2.689,$$

$$\hat{\beta}_{1,start} = \frac{\hat{\beta}_{1,default}\times 0.9}{\max(\hat{\boldsymbol{\mu}}) - \hat{\mu}_{3,default}} = \frac{-0.036\times 0.9}{0.868 + 0.129} = -0.032,$$

$$\hat{\beta}_{2,start} = \frac{\hat{\beta}_{2,default}\times 0.9}{\max(\hat{\boldsymbol{\mu}}) - \hat{\mu}_{3,default}} = \frac{-0.661\times 0.9}{0.868 + 0.129} = -0.597,$$

and

$$\hat{\beta}_{3,start} = \frac{\hat{\beta}_{3,default}\times 0.9}{\max(\hat{\boldsymbol{\mu}}) - \hat{\mu}_{3,default}} = \frac{0.043\times 0.9}{0.868 + 0.129} = 0.039,$$

where $\max(\hat{\boldsymbol{\mu}})$ is the approximate maximum fitted probability (this is 0.868, the approximate fitted probability of observation 9). Each starting value is multiplied by 0.9 and the $\hat{\beta}_{0,start}$ is added by 0.01 to ensure that the maximum fitted probability is less than unity and the minimum is greater than zero. By using the corrected starting values, $\hat{\beta}_{0,start} = 2.689$, $\hat{\beta}_{1,start} = -0.032$, $\hat{\beta}_{2,start} = -0.597$, and $\hat{\beta}_{3,start} = 0.039$, the fitting algorithm of R-*glm* can converge to an approximate solution with estimated coefficients, $\hat{\beta}_{0,approxi} = 2.688$, $\hat{\beta}_{1,approxi} = -0.034$, $\hat{\beta}_{2,approxi} = -0.631$, and $\hat{\beta}_{3,approxi} = 0.051$. The approximate fitted probabilities of observations 3 and 4 are extremely close to 0, which suggests the ML solution could lie on the lower bounds of the parameter space.

By assuming that the ML solution lies on the lower bound of parameter space, the estimation begins the re-parameterisation from observation 3 by the equation (3.22) as:

$$z_{ij}^{(1)} = x_{ij} - x_j^{(1)},\ i = 1,2,...9 \text{ and } j = 0,1,2,3.$$

Since observation 3 is the boundary vector $x_j^{(1)}$, it became a zero vector because of:

$$z_{3j}^{(1)} = x_{3j} - x_j^{(1)} = 0,\ \ j = 0,1,2,3.$$

The first re-parameterisation eliminated the constant β_0 and observation 3 from the model. It continued with observation 4 by the following equation:

$$z_{ij}^{(2)} = z_{ij}^{(1)} - \frac{t_j^{(2)}}{t_1^{(2)}} z_{i1}^{(1)} = z_{ij}^{(1)} - \frac{x_j^{(2)} - x_j^{(1)}}{x_1^{(2)} - x_1^{(1)}} z_{i1}^{(1)},\ i = 1,2,...9, i \neq 3, \text{ and } j = 1,2,3.$$

Observation 4 became a zero vector after the re-parameterisation since:

$$
\begin{aligned}
z_{4j}^{(2)} &= z_{4j}^{(1)} - \frac{x_{4j} - x_j^{(1)}}{x_{41} - x_1^{(1)}} z_{i1}^{(1)} \\
&= \left(x_{4j} - x_j^{(1)}\right) - \left(\frac{x_{4j} - x_j^{(1)}}{x_{41} - x_1^{(1)}}\right)\left(x_{41} - x_1^{(1)}\right), \\
&= 0
\end{aligned}
$$

for $i = 1, 2, ...9,\ i \neq 3$ and $j = 1, 2, 3$.

Moreover, the covariate X_1, the relevant coefficient β_1 and observation 4 were also eliminated from the data.

The final model after the re-parameterisation is $\hat{\mu}_i^* = \sum_{j=2}^{3} z_{ij}^{(2)} \beta_j$ with only two covariates $z_2^{(2)}$ and $z_3^{(2)}$ (the re-parameterised values of x_2 and x_3), and two relevant coefficients β_2 and β_3. The model converged effectively to the estimates $\hat{\beta}_2 = -0.630$, and $\hat{\beta}_3 = 0.051$. By the function (3.24) and (3.27), we got the estimate $\hat{\beta}_1 = -0.034$ and $\hat{\beta}_0 = 2.683$. The log-likelihood obtained from the identity-link binomial model with the exact method is –3.4010849, which is improved compared with the approximate solution. The improvement could become significant in data with a larger sample size and more boundary vectors involved.

3.4.3 The results from four statistical packages and two alternative methods

We fitted the data using the corresponding generalised linear model function in four statistical packages (R-3.6.3, Stata-16, SPSS-24 and SAS-9.4) with a binomial family and log link function. SAS-*genmod* failed to converge and reported an error message that the mean parameter is either invalid or at a limit of its range for some observations. Stata-*glm* (NR) failed to converge after 50 iterations. Stata-*glm* (IRLS) converged to a set of inadmissible

estimates and warned that parameter estimates produce inadmissible mean estimates in one or more observations. We found that the fitted probabilities of observations 3 and 4 became negative. SPSS failed to converge and reported an error message. R-*glm* showed an error message and stopped the fitting procedure to request an appropriate user-provided starting value. With a user-provided starting value, it converged to an approximate solution as the one used to identify the boundary vectors in Section 3.4.2.

We used the R-*glm* and Stata-*glm* with a Poisson family and log link function to implement the modified Poisson approach. The modified Poisson in Stata-*glm* failed to converge. The R-*glm* showed an error message and halted the fitting procedure to request an appropriate user-provided starting value. After providing an appropriate starting value, modified Poisson in R-*glm* can converge to an approximate solution with the log-likelihood of –3.4985379. However, as the IRLS with step-halving can provide an approximate solution for the identity-link binomial model when an appropriate starting value is present, the value of using modified Poisson is limited. The additive Poisson approach converged to a result with the log-likelihood of –4.4420667. The detailed estimates of three approaches with the relevant 95% confidence interval are listed in Table 3.2.

Table 3.2 The coefficients and the 95% confidence interval obtained from four approaches.

	Exact method	IRLS with step-halving	Additive Poisson	Modified Poisson
$\hat{\beta}_0$	2.683 (–1.918, 7.284)	2.688 (–0.568, 5.943)	0.961 (–3.003, 4.925)	1.672 (–3.498, 6.841)
$\hat{\beta}_1$	–0.034 (–0.132, 0.064)	–0.034 (–0.114, 0.046)	0.000 (–0.092, 0.092)	–0.012 (–0.134, 0.109)
$\hat{\beta}_2$	–0.630 (–1.458, 0.199)	–0.631 (–1.163, –0.098)	–0.207 (–0.866, 0.451)	–0.452 (–1.333, 0.429)
$\hat{\beta}_3$	**0.051** **(0.010, 0.093)**	**0.051** **(0.021, 0.081)**	0.005 (–0.091, 0.101)	0.045 (–0.006, 0.096)

It is worth noting that the results in the modified Poisson and the additive Poisson approach result in poor model fitting. These poor estimates may affect the statistical significance of the results, as the covariate X_3 in this example.

3.5 A real world example

We used a real world dataset from the book by Hosmer et al. [62]. They obtained the data by sampling a sub-sample of $n = 1000$ from the April 2008 release (Version 4.0) of the National Burn Repository research dataset (BURN) that includes 306,304 patients who were hospitalised due to burn injury between 1973 and 2007 in the United States. The mortality in hospital achieves 15 percent in the sample data. The binary (0/1) outcome variable is an indicator of death (*DEATH*) for the hospital discharge status of the patient. The study factors include age at admission (*AGE*), gender (*GENDER,* Male = 1 and Female = 0), race (*RACE,* Caucasian = 1 and Not Caucasian = 0), burn involved inhalation injury (*INH_INJ,* Yes = 1 and No = 0), the flame involved in burn injury (*FLAME,* Yes = 1 and No = 0), and the total percentage of burn surface area (*TBSA*, 0-100%).

To estimate mortality rate and quantify the effect of individual risk factors, an identity-link binomial model was fitted using a model-fitting procedure similar to that introduced by Hosmer, Lemeshow and Sturdivant [62]. Variables were compiled as per the criteria in their book, which include those with statistical significance and/or having clinical relevance with mortality risk. The details are omitted for brevity, but the findings and the scaling of covariates are summarised to briefly understand the data and model. *AGE* was divided into four groups (54 years or under, 55-64 years, 65-74 years, over 75 years), and four binary variables *AGE1, AGE2, AGE3* and *AGE4* were generated. *AGE1* was set as the reference group in the model.

Table 3.3 The coefficients and the 95% confidence interval obtained from four approaches.

	Exact method	IRLS with step-halving	Modified Poisson
Constant	–0.0013 (–0.0017, –0.0009)	0.0000 (–0.0158, 0.0157)	–0.0011 (–0.0014, –0.0007)
TBSA	0.0065 (0.0047, 0.0083)	0.0072 (0.0059, 0.0085)	0.0053 (0.0037, 0.0069)
INH_INJ	0.2599 (0.1487, 0.3710)	0.1516 (0.0577, 0.2457)	0.2472 (0.1131, 0.3813)
RACE	0.0007 (0.0005, 0.0008)	–0.0007 (–0.0164, 0.0150)	0.0005 (0.0004, 0.0007)
AGE2	0.1075 (0.0169, 0.1981)	0.1189 (0.0421, 0.1956)	0.1099 (0.0263, 0.1935)
AGE3	0.1519 (0.0287, 0.2750)	0.1396 (0.0249, 0.2543)	0.1637 (0.0244, 0.3030)
AGE4	0.4588 (0.3315, 0.5861)	0.3897 (0.2810, 0.4983)	0.4475 (0.2674, 0.6275)

The model was fitted without an issue through our package *bm* in R-3.6.1. Hosmer et al. [62] excluded *GENDER* because of no evidence of confounding. They removed *FLAME* from their model also because using simple yes or no coding is not precise enough to be helpful after the consultation from an experienced burn surgeon. We accepted their decision and excluded those two covariates from our model for the same reasons. *INH_INJ*, *RACE* and *TBSA* were involved in the final model due to their significant association with mortality. Each possible interaction of covariates had been examined. There was no significant association with mortality for any of the interaction terms.

With an appropriate starting value, the IRLS with step-halving converged to an approximate solution with the fitted probabilities that were close to the boundaries of the parameter space. To further improve the model fitting, the identity-link binomial model using the exact method is respectively fitted by assuming the ML solution lied either on the lower or upper bound of the parameter space. The log-likelihood obtained by assuming the ML solution on the lower bound was –205.5336, which was greater than the log-likelihood of –206.1732 obtained by

assuming the ML solution on the upper bound. Therefore, the solution with greater log-likelihood was accepted as the final solution for the model with two observations were confirmed as boundary vectors. Both observations (one was Caucasian, and another one was not) had the fitted probabilities equal to 0, and were from the age group 1 with a very small percentage of *TBSA* (0.1 and 0.2 percent) and no inhalation injury involved. These characters made them have an extremely low risk of mortality in the hospital after a burn injury. The results of the exact method were compared with two approximate solutions from alternative approaches in Table 3.3.

In comparison, Stata-*glm* (NR) and Stata-*glm* (IRLS) failed to converge after 50 iterations and warned that "parameter estimates produce inadmissible mean estimates in one or more observations". SAS-*genmod* failed to converge and warned "The mean parameter is either invalid or at a limit of its range for some observations". SPSS could not converge and warned that "There is at least one invalid case in the last iteration. A case is invalid if there are errors in computing the inverse identity link function, the log-likelihood, the gradient, or the Hessian matrix in the iterative process. Only the iteration history is displayed". R-*glm* showed an error message and stopped the fitting procedure to request an appropriate user-provided starting value. With a user-provided starting value, it converged to an approximate solution with the log-likelihood of –208.9734, like the one obtained from the IRLS with step-halving.

We used R-*glm* and Stata-*glm* with Poisson family and a log link to fulfil the modified Poisson approach. Stata-*glm* failed to converge. R-*glm* failed to begin the iteration with the default starting values. With user-provided starting values, it converged to an approximate solution with the log-likelihood of –208.7647 after the 662 iterations. After 380,974

iterations, the additive Poisson approach converged to a solution with the log-likelihood of –216.0718, and failed to report the estimates of standard error and warned that "MLE is on the boundary of parameter space, cannot use asymptotic covariance matrix". Since the estimates of the standard error and the relevant 95 percent confidence interval were missing, their results are not listed in Table 3.3.

Since there were two fitted probabilities of observations extremely close to 1, the ML solution of the real-world example may lie on both lower and upper bounds of the parameter space. This leads to a special case described in Section 3.3.5 in which, when the ML solution lies on both upper and lower bound, the exact method only can attain an approximate solution by comparing the log-likelihood obtained by the upper or lower bound hypothesis. The approximate solution will be the one maximised the log-likelihood. Although the approximate solution is not the ML solution of the model, it obtained by eliminating one side of the boundary vectors. Thus, the exact method solution remains to be the closest to the ML solution of the example among all other alternative approaches.

3.6 Simulations

The simulations were designed to demonstrate the performance of the exact method in the estimation of the identity-link binomial model. To perform this objective, 20000 sample datasets were randomly generated based on eight scenarios. Each sample data includes 1000 observations with a binary outcome variable Y, a dichotomous covariate X_1 and a continuous covariate X_2. Table 3.4 lists the designed values for the relevant parameters β_0, β_1 and β_2 in each scenario. The conclusions are not influenced by the specific values of β_0, β_1 and β_2 chosen. The value of outcome variable Y was drawn at random from a Bernoulli

distribution with success probability $\mu_y = \beta_0 + x_1\beta_1 + x_2\beta_1$. The dichotomous covariate X_1 was drawn at random from a Bernoulli distribution with probability $\mu = 0.5$. The X_2 as a uniformly distributed continuous covariate was randomly generated with limiting values a and b that are respectively the minimum and maximum of $\left(0 - \beta_0 - \beta_1 \times 0\right)/\beta_2$ and $\left(1 - \beta_0 - \beta_1 \times 1\right)/\beta_2$ when the value of β_2 is positive, and invert their order when the value of β_2 is negative. We define the covariates vector with the fitted probability is equal to 0 or 1 as a boundary vector on the lower or upper bound of the parameter space respectively. A realisation with $y = 1$ could produce a boundary vector on the upper bound if the largest value $\left(x_1 = 1\right)$ of X_1 and a large $\left(x_2 \approx b\right)$ value of X_2 were drawn in settings with $\beta_2 = 0.1$, or a realisation with $y = 0$ could produce a boundary vector on the lower bound if the smallest value $\left(x_1 = 0\right)$ of X_1 and a small $\left(x_2 \approx a\right)$ value of X_2 were drawn in settings with the same value of β_2. (The boundary vector could be produced similarly when the value of β_2 is negative.)

Table 3.4: Design of the simulations

Setting	β_0	β_1	β_2	a^*	b^*
1	1.0	0.06	0.1	–10.0	–0.60
2	1.0	0.06	–0.1	0.60	10.0
3	1.5	0.06	0.1	–15.0	–5.60
4	1.5	0.06	–0.1	5.60	15.0
5	2.0	0.06	0.1	–20.0	–10.6
6	2.0	0.06	–0.1	10.6	20.0
7	2.5	0.06	0.1	–25.0	–15.6
8	2.5	0.06	–0.1	15.6	25.0

‘ Lower and upper limits of the continuous covariate X_2 taking values generated at random from the uniform distribution $U\left(a,b\right)$.

3.6.1 Results for coefficient estimates

Table 3.5 shows the simulation results for the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ between the replications with and without boundary vector. The results for the $\hat{\beta}_0$ have been omitted for brevity, but they can be found in Appendix B. Firstly, the results showed that 41.3 to 48.7 percent of repetitions did not produce a boundary vector. The average percent bias in the group without a boundary vector was negative for the coefficient estimates, indicating that most were less than the design values, and the mean of the bias was at the left side of the null. That is to be expected because those estimates did not produce fitted probabilities large enough to constitute a boundary vector. Moreover, because the coefficients were not fully estimated, they need a wider range of confidence intervals to cover the design values. A slightly larger value of average mean squared error in the simulations without a boundary vector also described this partially. In contrast, the average percent bias in the group with a boundary vector was small (approximately 0.42 to 3.2% in each scenario) and mildly negative, indicating that most were close to the design values, and the mean of the bias was minor to the left of the null. That is to be expected because these coefficients needed to be fully estimated to produce a boundary vector. Reflecting the lesser contribution of the average percent bias, the average mean squared errors of exact estimates that produced a boundary vector were lower than those of the fits that did not (Table 3.5). The results for the repetitions with a boundary vector on the lower bound have been compared with those from the upper, and no notable difference was shown. The results can be found in Appendix B.

We noticed that the average percent bias in the group without a boundary vector was approximately 3.8 to 8.7 times more than the group with a boundary vector. That leads to a reasonable conjecture that the sample with a boundary vector contains the information from not only the population parameters, but also the extra information about the boundary of

parameters space. This extra information enabled the fitting procedure to converge to an estimate that is closer to the design value. However, the sample without a boundary vector contains only the information from the parameters. Therefore, it could not produce an estimate with less "bias". Further study is needed to prove the correctness of the conjecture, which will not be covered in this thesis.

Table 3.5: Simulation results

		Simulations without a boundary vector				Simulations with a boundary vector			
Setting		*n*	Bias*	MSE†	Coverage‡	*n*	Bias*	MSE†	Coverage‡
1	β_1	9,317	–12.258	0.117	95.5	10,683	–2.334	0.101	94.7
	β_2	9,317	–2.117	0.003	97.4	10,683	–0.490	0.001	96.6
2	β_1	9,338	–12.637	0.118	95.6	10,662	–2.224	0.098	95.2
	β_2	9,338	–2.112	0.003	97.4	10,662	–0.485	0.001	97.0
3	β_1	9,261	–13.469	0.118	95.3	10,739	–1.546	0.099	94.9
	β_2	9,261	–2.062	0.003	97.0	10,739	–0.547	0.001	96.5
4	β_1	9,421	–14.612	0.117	95.5	10,579	–2.087	0.101	94.3
	β_2	9,421	–1.948	0.003	97.5	10,579	–0.509	0.001	96.5
5	β_1	9,738	–13.035	0.116	95.6	10,262	–3.210	0.101	94.7
	β_2	9,738	–1.971	0.003	97.7	10,262	–0.405	0.001	96.4
6	β_1	9,715	–12.683	0.116	95.8	10,285	–2.889	0.100	94.7
	β_2	9,715	–1.985	0.003	97.5	10,285	–0.422	0.001	96.5
7	β_1	8,256	-14.092	0.119	94.8	11,744	-2.717	0.100	94.8
	β_2	8,256	-2.310	0.003	97.2	11,744	-0.480	0.001	96.7
8	β_1	8,433	-14.371	0.119	94.9	11,567	-2.759	0.100	94.9
	β_2	8,433	-2.238	0.002	97.6	11,567	-0.482	0.001	96.7

* Average value of percent bias: $(100/n)\sum_{k=1}^{n}\left[\left(\hat{\beta}_{jk}-\beta_j\right)/\beta_j\right], \quad j=0,1,2; k=1,2,...n$

† Average value of mean squared error: $(100/n)\sum_{k=1}^{n}\left[\left(\hat{\beta}_{jk}-\beta_j\right)^2+\widehat{Var}\left(\hat{\beta}_{jk}\right)\right]$ for $j=0,1,2$ and $k=1,2,\ldots n$.

‡ Percentage of 95 percent confidence intervals covering the design value of the coefficient.

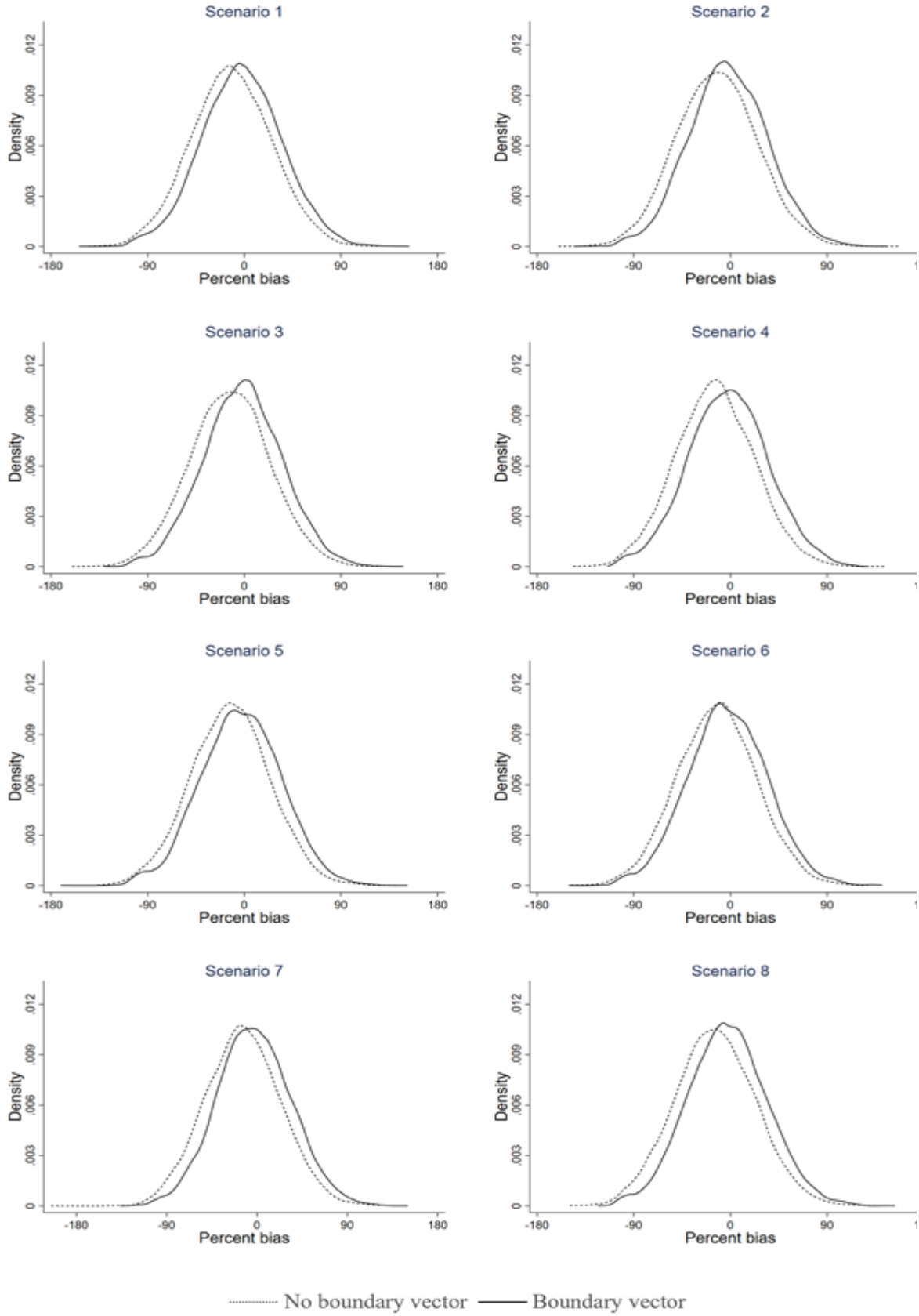

Figure 3.1: Density plots of the sampling distributions (n=1000, 20000 replications, eight settings) of the binary covariate β_1. The solid lines represent replications with a boundary vector. The dotted lines represent replications without a boundary vector.

Figure 3.1 provides density plots of the percent bias for the estimate $\hat{\beta}_1$ between the replications with and without boundary vector, which visually presents the distributional shift (Figures for the $\hat{\beta}_0$ and $\hat{\beta}_2$ see Appendix B). The peak of the kernel density curves in the replications without a boundary vector were shifted from the null toward left in all scenarios compared with those with a boundary vector.

3.6.2 The comparison with the alternative approach

For comparison, I attempted to fit the simulated samples using the modified Poisson with default starting value algorithm and additive Poisson approach. For the modified Poisson approach, the fitting procedure failed to begin the iteration without the starting values provided. This approach is excluded from the comparison rather than self-providing these values. As mentioned previously, when an appropriate starting value is present, the IRLS with step-halving can provide an approximate solution for the identity-link binomial model therefore, the value of using the modified Poisson is limited. In terms of the additive Poisson approach, whilst it could converge to a solution, the solution is biased. Moreover, the approach could not obtain an estimate of standard error for about 12-61% of repetitions with a boundary vector. Table 3.6 provides the results from the additive Poisson approach for the repetitions with a boundary vector (the results only included the repetitions in which the estimates of the standard error were not missing). The average percent bias is about 1.1 to 1.6 times larger than in the exact method.

The average mean squared errors of the additive Poisson approach were comparable to the corresponding exact method estimates (Table 3.6) in the samples with no missing estimate of standard error. However, the 95% confidence interval coverage by this approach did not reach (92.8-96.9%) the targets. This is likely to be caused by missing standard errors which

result in some results of the replications being excluded from the calculation of 95% confidence intervals in the additive Poisson approach. In contrast, the exact method has a much better performance in reaching the target coverage in the entire dataset (n =20000) (Table 3.6). It is also on average closer to the target in the repetitions with a boundary vector.

Table 3.6: Simulation results (with results in which the standard error is not missing for the additive Poisson method shown for comparison)

	All simulations (n = 20000)			Simulations with a boundary vector							
	Exact method			Exact method				Additive Poisson[§]			
Setting	Bias[*]	MSE[†]	Cover[‡]	n	Bias[*]	MSE[†]	Cover[‡]	n	Bias[*]	MSE[†]	Cover[‡]
1 β_0	–0.482	0.136	95.5	4636	0.433	0.113	93.1	4636	0.475	0.115	92.8
β_1	–6.957	0.108	95.1	4636	–7.020	0.101	93.5	4636	–7.664	0.101	92.9
β_2	–1.248	0.002	97.0	4636	–1.024	0.001	95.9	4636	–1.126	0.001	96.3
2 β_0	–0.465	0.135	95.6	9428	0.131	0.120	95.3	9428	0.142	0.115	93.9
β_1	–7.086	0.107	95.4	9428	–2.827	0.098	95.2	9428	–4.482	0.096	94.1
β_2	–1.245	0.002	97.2	9428	–0.536	0.001	96.9	9428	–0.597	0.001	96.1
3 β_0	–0.720	0.321	95.7	4773	–0.015	0.235	94.9	4773	–0.084	0.246	95.2
β_1	–7.067	0.108	95.0	4773	–6.597	0.099	94.5	4773	–7.231	0.098	93.5
β_2	–1.248	0.002	96.8	4773	–1.053	0.001	96.3	4773	–1.149	0.002	96.9
4 β_0	–0.665	0.317	95.6	9349	–0.109	0.268	94.7	9349	–0.121	0.262	93.9
β_1	–7.987	0.108	94.9	9349	–2.598	0.101	94.2	9349	–4.280	0.099	93.0
β_2	–1.187	0.002	97.0	9349	–0.571	0.001	96.5	9349	–0.630	0.001	95.8
5 β_0	–0.774	0.598	96.0	4738	–0.178	0.440	94.5	4738	–0.257	0.464	95.2
β_1	–7.994	0.108	95.1	4738	–8.162	0.101	94.1	4738	–8.782	0.101	93.5
β_2	–1.167	0.002	97.0	4738	–0.955	0.001	95.9	4738	–1.056	0.002	96.5
6 β_0	–0.798	0.599	96.3	9052	–0.152	0.471	95.6	9052	–0.168	0.464	94.7
β_1	–7.647	0.108	95.3	9052	–3.093	0.100	94.8	9052	–4.914	0.097	93.3
β_2	–1.181	0.002	97.0	9052	–0.502	0.001	96.3	9052	–0.553	0.001	95.7
7 β_0	–0.923	0.983	96.2	4630	–0.374	0.708	94.8	4630	–0.462	0.755	95.7
β_1	–7.413	0.108	94.8	4630	–7.685	0.100	94.0	4630	–8.359	0.100	93.3
β_2	–1.236	0.002	96.9	4630	–1.008	0.001	95.9	4630	–1.114	0.002	96.5
8 β_0	–0.910	0.977	96.3	10056	–0.318	0.770	95.9	10056	–0.361	0.767	95.4
β_1	–7.655	0.108	94.9	10056	–3.021	0.099	95.0	10056	–4.592	0.097	93.9
β_2	–1.222	0.002	97.1	10056	–0.569	0.001	96.8	10056	–0.641	0.001	96.1

[*] Average value of percent bias: $(100/n)\sum_{k=1}^{n}\left[\left(\hat{\beta}_{jk}-\beta_j\right)/\beta_j\right], \quad j=0,1,2; k=1,2,...n$

[†] Average value of mean squared error: $(100/n)\sum_{k=1}^{n}\left[\left(\hat{\beta}_{jk}-\beta_j\right)^2+\widehat{Var}\left(\hat{\beta}_{jk}\right)\right], \quad j=0,1,2; k=1,2,...n.$

[‡] Percentage of 95 percent confidence intervals covering the design value of the coefficient.
[§] The simulation results with a boundary vector in the additive Poisson only include the repetitions that the estimated standard errors are reported.

3.7 Conclusion

In clinical trials for which risk of occurrence of an event or non-event is an outcome measure, it is recommended that risk differences be reported to assist clinicians in decision making.

The identity-link binomial model theoretically provides a way to estimate the risk difference. However, the standard fitting algorithm in statistical packages based on either Newton-Raphson or IRLS algorithm is often troubled by numerical difficulties due to inappropriate starting values and failure of convergence when the ML solution is close to or on the boundary of the parameter space. We found that the IRLS with step-halving can converge to the ML solution if an appropriate starting value is provided. However, in some cases, when the ML solution lies on the boundary of the parameter space, the solution obtained from the IRLS with step-halving can only be presented as an approximate one due to the impact of the boundary vector on the fitting algorithm. The difference between the approximate and the ML solution could be significant in some cases. Petersen and Deddens [9] firstly proposed the exact method to solve the convergence difficulty due to the ML solution is on the boundary of parameter space in the log binomial model. Because of the similarity of log binomial and identity-link binomial models, the exact method is extended into the identity-link binomial model to overcome the issue caused by the boundary vector and further improve the model fitting. The details and the specifics to apply the exact method in the identity-link binomial model are listed in this chapter. It is summarised in the following seven aspects:

(a) A starting value calibration offers an appropriate starting value for the standard fitting algorithm;
(b) Three theorems (with proof) for estimating the ML solution when it is on the upper bound of parameter space in the identity-link binomial model;
(c) Three theorems (with proof) for estimating the ML solution when it is on the lower bound of parameter space in the identity-link binomial model;
(d) A theorem (with proof) to demonstrate a perfect linear correlation between each pair of estimates of coefficients of non-constant covariates in the identity-link binomial

model when the ML solution is on the boundary, and the number of coefficients of non-constant covariates is equal to the number of boundary vectors.

(e) A proved corollary to show that the standardised values of the ML estimates of coefficients of non-constant covariates are equal in absolute value;

(f) A theorem and corollary to establish that the re-parameterisation of covariates required to implement the exact method can be successfully undertaken if the ML solution exists.

(g) An integrated scheme to overcome the numerical difficulties and locate the ML solution in the identity-link binomial model.

To demonstrate the implementation of the exact method in detail, an example dataset and a real-world dataset have been provided. The exact method converged to the ML solution with two boundary vectors on the lower bound of the parameter space in the example dataset. In the real-world dataset, it successfully found two boundary vectors on the lower bound and improved the approximate solution obtained from the IRLS with step-halving by eliminating the impact of the boundary vectors. (Since the R-*glm* is coded based on the IRLS with step-halving [58], this function is used to implement the IRLS with step-halving in this thesis) However, the real-world dataset may be a case that the ML solution lies on both lower and upper bound. Eliminating the impact of the boundary vector on the lower bound through the exact method can only improve the solution, but the model may not be fully estimated due to the uneliminated impact of the boundary vector on the upper bound. As mentioned in Section 3.3.5, the exact method can only eliminate the boundary vector on one side by the re-parameterising the covariates if the ML solution lies on both bounds. Therefore, the solution obtained by eliminating the boundary vector on one side is still an approximate one. The standard fitting algorithm of the identity-link binomial model in three statistical packages

(SAS, Stata and SPSS) failed to converge in both examples. The fitting algorithm in R-*glm* failed to begin the iteration and asked for an appropriate starting value. After an appropriate starting value was provided, it converged to an identical solution as IRLS with step-halving since the fitting algorithm of R-*glm* is coded based on the IRLS with step-halving [58]. The modified Poisson failed to begin the iteration with the default starting values. After an appropriate starting value was provided, it converged to an approximate solution which did not represent an improved performance than the approximate solution obtained from the IRLS with step-halving. Moreover, as mentioned in Sections 3.4.3 and 3.6.2, when an appropriate starting value is present, the IRLS with step-halving can provide an approximate solution for the identity-link binomial model therefore, the value of using the modified Poisson is limited. The additive Poisson approach produced an inadmissible solution in both datasets, and did not report the standard errors in real world dataset.

We performed a simulation to evaluate the performance of the exact method in the ML estimation of the identity-link binomial model. The simulation results obtained by the exact method revealed a relatively lower average value of percent bias in the repetitions with a boundary vector compared to those without. As expected, the simulation results illustrate that the coefficients in the repetitions with a boundary vector were fully estimated to produce a boundary vector compared to those without. Thereby, the exact method improved the model fitting by properly dealing with the boundary vector.

The R-*glm* with Poisson family and log link was used to implement the modified Poisson method in simulation. The method entirely failed to begin the iteration with a default starting value. Thus, it was excluded from the comparison. In the repetitions with a boundary vector, the additive Poisson approach converged to an inadmissible solution with a larger average

percent bias compared with the exact method and failed to report the estimate of the standard error in 12-61% repetitions. Moreover, the 95% confidence interval coverage was around the target in the exact method and was about 2 to 3 percentage points below the 95% in the additive Poisson model in contrast. The simulation results suggest that the exact method displays good performance in the data with a boundary vector.

In conclusion, the exact method showed an advantage in the estimation of the ML solution when it is on the boundary of the parameter space in the identity-link binomial model. We recommend it to fit the model and estimate the risk difference. A relevant package *bm* in R to estimate the ML solution in identity-link binomial model based on all of the contributions in this chapter has been released.

Chapter 4 Relative risk estimation of clustered/longitudinal data using generalised estimating equations

4.1 Introduction

The risk ratio/relative risk (RR) is the ratio measure of choice for summarising the impact of exposure on the incidence proportion [5]. It has application to the analysis of randomised controlled trials with discrete event outcomes that are not a function of time (for which failure-time analysis is used). The statistical methods employed to estimate RR and adjust for confounders in a regression model differ by data type. In independent data (meaning that the different subjects do not depend on each other), a possible way is to fit a log binomial model (LBM) as discussed in Chapter 2 of this thesis. Nevertheless, LBM is not suitable for the data with correlated outcomes, such as clustered/longitudinal data. In the analysis of correlated data, the approaches used for addressing within-cluster correlations can be grouped into two classes described by [50], subject-specific and population-averaged.

In subject-specific approaches, the heterogeneity across subjects is explicitly modelled by adopting unobservable random effects [50]. The subject-specific model fully parameterises the distribution of the population, and the coefficients of the model have an interpretation for individuals [49]. Mixed model [66, 67] is an example of the subject-specific model, which estimates the random effect by assuming it follows a parametric distribution across the population. Conversely, in the population-averaged approaches, heterogeneity between subjects is not explicitly taken into account [3, 49]. Instead, the population-averaged response is modelled as a function of covariates. The population-averaged model parameterises only the marginal distribution of the population, and the coefficients of the model have an interpretation for the population rather than for any individual [50]. Generalised estimating

equations (GEE) introduced by Liang and Zeger [46] is a population-averaged approach. It is an extension of generalised linear model (GLM) and is used for estimating the parameters of a marginal GLM with a user-defined correlation structure between outcomes [46]. In this chapter of the thesis, I study the application of the exact method in a marginal log binomial model estimated by GEE (marginal LBM by GEE).

Naturally, a marginal LBM by GEE becomes a feasible approach to estimate RR with adjustment for confounders in a model with a correlated response on the population-averaged level. However, the standard fitting algorithm of marginal LBM by GEE in the statistical software may meet numerical difficulties, leading to a failure of convergence [68, 69]. When a convergence issue occurs, one possible option is to alter the RR estimation toward approximately estimating the odds ratio using the marginal logistic model estimated by GEE. As an estimate of the risk ratio, the approximation is adequate when the outcome is rare in all exposure and confounder categories [5], but the odds ratio increasingly overstates its target as the outcome becomes more common [70]. Another possible choice is to use a marginal modified Poisson with log-link function estimated by GEE (marginal Poisson by GEE) to approximately estimate the coefficients in marginal LBM by GEE. This is a common method to approximate RR when a convergence issue occurs in LBM [19, 20]. However, the solution obtained by marginal Poisson by GEE may result in a fitted probability (estimated mean of observation) exceeding the interval [0, 1] [69].

There is no published account of the causes of non-convergence or remedies for the convergence issues in the marginal LBM by GEE. In the LBM, one of the reasons for convergence issues is that the maximum likelihood solution lies on the boundary of the parameter space, leading to the fitted probability equal to 1 [8, 9] (detailed in Chapter 2 of this

thesis). Because of the similarity of marginal LBM by GEE and LBM in the mathematical form, it is reasonable to presume that one interpretation of numerical difficulty is the presence of boundary vector, which is defined as the covariates vector with the corresponding fitted probability equal to 1. If a boundary vector is present, the standard fitting algorithm may meet numerical difficulties and fail to converge. A feasible approach is needed to overcome this difficulty.

Petersen and Deddens [9] introduced the exact method to resolve the convergence issue caused by the boundary vector in LBM. The exact method involves transforming the data to eliminate the boundary vectors, resulting in a re-parameterisation of the likelihood function. Their scheme can solve the convergence issue, but it is incomplete and without details to implement (detailed in Section 4.3). Because of the similarity of marginal LBM by GEE and LBM in the mathematical form, an extension of the exact method in marginal LBM by GEE may overcome the numerical difficulty caused by boundary vector.

In this chapter, I propose an extension of the exact method in marginal LBM by GEE. Two theorems with proof and details to implement the method are provided to overcome the convergence issue due to the numerical difficulty attributed to boundary vector. The following sections in this chapter include: a brief introduction of GEE and the connection with quasi-likelihood in section 4.2; application of the exact method to overcome the numerical difficulties due to the ML solution lying on the boundary of the parameter space in section 4.3; a comparison of the exact method with alternative approaches using simulation studies in section 4.4; application of the exact method to real-world data in section 4.5. The proofs of the theorems and corollaries in this chapter are reported in Appendix C. An R

package *lb.gee* based on the exact method for fitting the identity-link binomial model has been released on Github. The related R documents are set out in Section 6.3 of Chapter 6.

4.2 The background of GEE

Proposed by Liang and Zeger [46], GEE is a generalisation approach based on quasi-likelihood [60]. It offers a way to estimate the regression parameters in marginal models, with no assumptions being made on the joint distribution of the responses. To understand the association between GEE and quasi-likelihood estimation, the quasi-likelihood approach is described as follows before discussing its extension to marginal models for clustered responses.

4.2.1 Quasi-likelihood

As an extension of the likelihood function, the quasi-likelihood is initiated by Wedderburn [60], and its properties are further discussed by McCullagh [48]. It is used for estimating the regression coefficients without fully specifying a probability distribution for the data. The quasi-likelihood allows for overdispersion and makes assumptions only on the mean and the variance of the data distribution [60].

Let $y_i,\ i=1,2,...n$ be n independent responses with mean $\mathrm{E}(y_i)=\mu_i$ and suppose that each μ_i is a known function of parameters $\boldsymbol{\beta}=(\beta_1,\beta_2,...\beta_J)$. The response variance is $\mathrm{Var}(y_i)=\phi v(\mu_i)$, where ϕ is a scalar (dispersion) parameter, and v is a known variance function [60]. Then the quasi-likelihood for each observation is defined as follows:

$$\frac{\partial \mathrm{Q}(y_i,\mu_i)}{\partial \mu_i}=\frac{y_i-\mu_i}{\phi v(\mu_i)}$$

or equivalently

$$\mathrm{Q}(y_i,\mu_i)=\int_{y_i}^{\mu_i}\frac{y_i-t_i}{\phi v(t_i)}dt_i\,.$$

The quasi-likelihood for the complete data is given by the sum of the quasi-likelihood for each observation,

$$\mathrm{Q}(y,\mu)=\sum_{i=1}^{n}\mathrm{Q}(y_i,\mu_i)\,.$$

The quasi-score function written in matrix form is obtained by differentiating the function $\mathrm{Q}(\mathbf{y},\boldsymbol{\mu})$ with respect to the parameters $\boldsymbol{\beta}$ as follows [48]:

$$\mathrm{S}(\boldsymbol{\beta})=\frac{\partial \mathrm{Q}(\mathbf{y},\boldsymbol{\mu})}{\partial \boldsymbol{\beta}}=\mathbf{D}'\mathbf{V}^{-1}(\mathbf{y}-\boldsymbol{\mu}),$$

where $\mathbf{D}$ is a $n\times J$ matrix with elements $\partial\mu_i/\partial\beta_j$, and $\mathbf{V}=\mathrm{diag}\left[\phi v(\mu_i)\right]$. The estimates of the regression parameters $\boldsymbol{\beta}$ are obtained by solving a set of quasi-score functions equal to 0 as follows:

$$\mathbf{D}'\mathbf{V}^{-1}(\mathbf{y}-\boldsymbol{\mu})=0\,.$$

It has been proved that if the response distribution is a member of the exponential family, the log-likelihood function is identical to the quasi-likelihood [60].

4.2.2 Generalised estimating equations

GEE has been recognised as a reliable estimation technique for fitting marginal models that take account of inter-dependencies in response [46]. It is a semiparametric approach used for estimating the regression coefficients of a marginal GLM with correlated responses without requiring a full specification of the joint distribution. The approach does not depend on a likelihood function, but on quasi-likelihood in which only the mean and variance of the response (first two moments) is specified for estimating the regression parameters [46].

Suppose that there are K clusters/subjects in the data, and n_i observations included in the cluster i, $i = 1, 2, ...K$, which could be related subjects or observations on a single subject. Denote response variable of p^{th} observation in the cluster i as y_{ip}, $p = 1, 2, ...n_i$. The relationship between the mean μ_{ip} of y_{ip} and the linear combination of covariates vector $\mathbf{x}_{ip} = \left(x_{ip1}, x_{ip2}, ...x_{ipJ}\right)$ and the marginal model regression parameters $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \beta_2, ...\beta_J\right)$ is written as:

$$\mathrm{g}\left(\mu_{ip}\right) = \mathbf{x}'_{ip}\boldsymbol{\beta} = \beta_0 + \beta_1 x_{ip1} + \beta_2 x_{ip2} + ... + \beta_J x_{ipJ},$$

where g is a link function (a list of the available link function with different distributions in the exponential family is in the book of Hardin and Hilbe [49]). The variance of y_{ip} is represented as $\mathrm{Var}\left(y_{ip}\right) = \phi v\left(\mu_{ip}\right)$, where ϕ is a scalar (dispersion) parameter, and v is a known variance function. When the outcome, y_{ip}, is following a binomial distribution and the link function g is the natural logarithm, it is called marginal LBM by GEE with mean $\mu_{ip} = \exp\left(\mathbf{x}'_{ip}\boldsymbol{\beta}\right)$ and variance $\mathrm{Var}\left(y_{ip}\right) = \phi\mu_{ip}\left(1 - \mu_{ip}\right)$. For the i^{th} cluster, the variance-covariance matrix is structured as

$$\mathbf{V}_i = \phi\mathbf{A}_i^{1/2}\mathbf{R}_i\left(\boldsymbol{\alpha}\right)\mathbf{A}_i^{1/2},$$

where $\mathbf{A}_i = \mathbf{diag}\left[v\left(\mu_{i1}\right), v\left(\mu_{i2}\right), ...v\left(\mu_{in_i}\right)\right]$ and $\mathbf{R}_i\left(\boldsymbol{\alpha}\right)$ is a $n_i \times n_i$ "working" correlation matrix which specifies the correlation structure between observations in cluster *i*. The parameter vector $\boldsymbol{\alpha}$ summarises the correlation between observations [46]. The estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, is obtained by solving the following estimating equation:

$$U\left(\boldsymbol{\beta}\right) = \sum_{i=1}^{K} \mathbf{D}'_i\mathbf{V}_i^{-1}\left(\mathbf{y}_i - \boldsymbol{\mu}_i\right) = \mathbf{0},$$

where $\boldsymbol{\mu}_i = \left(\mu_{i1}, \mu_{i2}, ...\mu_{in_i}\right)$ and $\mathbf{D}_i = \partial\boldsymbol{\mu}_i / \partial\boldsymbol{\beta}$. It can be seen that if $\mathbf{R}_i(\boldsymbol{\alpha})$ is an identity matrix, then the estimating equations in GEE are the same as the one in quasi-likelihood.

The word "working" in the name of the "working" correlation matrix $\mathbf{R}_i(\boldsymbol{\alpha})$ means that the correlation matrix does not need to be correctly specified. The consistency of the estimates $\hat{\boldsymbol{\beta}}$ depends only on the correct specification of the mean, not on the correct choice of $\mathbf{R}_i(\boldsymbol{\alpha})$ [46]. Under mild regularity conditions (the function $\sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}^{-1}\left(\mathbf{y}_i - \boldsymbol{\mu}_i\right)$ is continuously differentiable), Liang and Zeger [46] showed that $\sqrt{K}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ is asymptotically normal with mean zero and covariance matrix $\mathbf{V}_{\boldsymbol{\beta}}$ given by

$$\mathbf{V}_{\boldsymbol{\beta}} = \lim_{K\to\infty} K\left(\sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i\right)^{-1} \left[\sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathrm{Cov}\left(\mathbf{y}_i\right) \mathbf{V}_i^{-1} \mathbf{D}'\right] \left(\sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i\right)^{-1}.$$

The estimate of $\mathbf{V}_{\boldsymbol{\beta}}$ can be obtained consistently by replacing $\mathrm{Cov}\left(\mathbf{y}_i\right)$ by $\left(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i\right)\left(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i\right)'$ and $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and ϕ by their estimates as follows [46]:

$$\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}} = \left(\sum_{i=1}^{K} \mathbf{D}_i' \hat{\mathbf{V}}_i^{-1} \mathbf{D}_i\right)^{-1} \mathbf{M} \left(\sum_{i=1}^{K} \mathbf{D}_i' \hat{\mathbf{V}}_i^{-1} \mathbf{D}_i\right)^{-1}$$

where $\mathbf{M} = \sum_{i=1}^{K} \mathbf{D}_i' \hat{\mathbf{V}}_i^{-1}\left(\mathbf{y}_i - \hat{\mathbf{u}}_i\right)\left(\mathbf{y}_i - \hat{\mathbf{u}}_i\right) \hat{\mathbf{V}}_i^{-1} \mathbf{D}'$, and $\sum_{i=1}^{K} \mathbf{D}_i' \hat{\mathbf{V}}_i^{-1} \mathbf{D}_i$ is model-based (also called naive) variance-covariance matrix. The matrix $\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}$ is called the robust (or empirical) variance-covariance matrix. The estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}$ are consistent even if the working correlation structure is misspecified [46].

Given an estimate of $\mathbf{R}_i(\boldsymbol{\alpha})$ and ϕ, an updated estimate of $\boldsymbol{\beta}$ can be obtained by the iteratively reweighted least squares described by McCullagh [48]. At a given iteration, the current Pearson residuals are defined as follows:

$$\hat{r}_{ip} = \frac{y_{ip} - \hat{\mu}_{ip}}{\sqrt{v(\hat{\mu}_{ip})}},$$

which is used for estimating ϕ by [46, 48, 60]

$$\hat{\phi} = \sum_{i=1}^{K} \sum_{p=1}^{n_i} \frac{\hat{r}_{ip}^2}{N-J},$$

where $N = \sum n_i$. The general approach to estimate $\boldsymbol{\alpha}$ is given by Liang and Zeger [46] as follows:

$$\hat{\alpha}_{uv} = \sum_{i=1}^{K} \frac{\hat{r}_{iu}\hat{r}_{iv}}{(N-J)}.$$

The detailed information of the correlation structure and the corresponding specific function for estimating correlations are in the book of Hardin and Hilbe [49].

In some cases, the fitting algorithm of marginal LBM by GEE may encounter numerical difficulty and fail to converge when there is a mean of y_{ip}, $\hat{\mu}_{ip}$, evaluated at $\hat{\boldsymbol{\beta}}$ that is equal to 1 (the corresponding vector of covariates is referred as boundary vector). Such a value could be problematic as it could lead to the inefficient calculation of $\hat{\mathbf{V}}_i^{-1}$ as $\hat{\mu}_{ip} = 1$ which results in the $\widehat{\mathrm{Var}}(y_{ip}) = 0$, causing failure of the fitting algorithm.

4.3 Marginal LBM by GEE using the exact method

As discussed in Chapter 2, the exact method is initiated by Deddens et al. [18] and generalised by Petersen and Deddens [9] to resolve the convergence issue due to boundary vectors in the LBM. They introduced two theorems with formulae to estimate the coefficients and the relevant standard errors in the model. However, the method was incomplete because formulae to estimate covariances were not provided, which means that the exact method could not be implemented in general cases. In addition, the details to implement this method were also missing, and subsequently this method was not implemented in practice.

In the following part of this section, two theorems with formulae are provided to estimate the covariances between the estimated coefficients of the eliminated covariates (including the constant) with each other and with the estimated coefficients of the non-eliminated covariates.

4.3.1 Implementation of the exact method

The exact method for the LBM can be extended to marginal LBM by GEE based on the similarities between the two models and the characteristics of the boundary vector, which include:

1. both models fail to converge due to a similar reason, which is the presence of the boundary vector,
2. they share the functions of the same form in the mean $\mu_{ip} = \exp\left(\mathbf{X}'_{ip}\boldsymbol{\beta}\right)$ and variance $Var\left(Y_{ip}\right) = \mu_{ip}\left(1-\mu_{ip}\right)$ of the outcome variable,
3. eliminating the boundary vector will not influence the estimation of the working correlation matrix and the dispersion parameter. The estimated mean corresponding to

the boundary vector is equal to 1, which leads to a Pearson residual equal to 0 and no contribution to the estimation of the correlation and the dispersion parameter.

Some formulae for implementing the exact method for the LBM can be found in the paper by Petersen and Deddens [9]. The remaining formulae necessary to implement the method are provided in chapter 2 of this thesis. Here I introduce the implementation of the method in marginal LBM by GEE.

Assume that there are R $\left(0 < R \le J\right)$ distinct covariate vectors of observations in the data with the estimated mean evaluated at the estimates $\hat{\boldsymbol{\beta}}$ equal to 1, which are defined as the boundary vectors. Denote the r^{th} $\left(r = 1, 2, ...R\right)$ boundary vector as $\mathbf{x}^{(r)}$, which shares the covariate values with n_r observations. The fitting procedure of the exact method can be summarised in seven steps:

1. Eliminate the constant by subtracting from the constant and each non-constant covariate its respective value in the boundary vector:

$$z_{ijp}^{(1)} = x_{ijp} - x_j^{(1)}, \quad j = 0, 1, 2, ...J$$

where i denotes cluster, j denotes the position of a covariate, and p denotes the position in the cluster.

2. When there are multiple boundary vectors $(R > 1)$, eliminate the first $R-1$ non-constant covariates by re-parametrising the covariates according to the scheme:

$$z_{ijp}^{(r)} = z_{ijp}^{(r-1)} - \left(\frac{t_j^{(r)}}{t_{r-1}^{(r)}}\right) z_{ip,r-1}^{(r-1)} \quad \text{where } t_j^{(r)} = z_{ijp}^{(r-1)}\Big|_{x_{ijp}=x_j^{(r)}}, \quad r = 2, 3, ...R$$

3. Drop the observations with covariate vectors $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ...\mathbf{x}^{(R)}$ respectively, which make no contribution to the estimation, and fit the model $\mu_{ip} = \exp\left(\sum_{j=R}^{J} \beta_j z_{ijp}^{(R)}\right)$ without a constant and with $J - R + 1$ covariates to the remaining $n - n_1 - n_2 - ... - n_R$ observations to obtain the estimates $\hat{\beta}_R, \hat{\beta}_{R+1}, ...\hat{\beta}_J$ of the coefficients of the non-eliminated non-constant covariates and their estimated variances $\widehat{\text{Var}}\left(\hat{\beta}_j\right)$, $j = R, R+1, ...J$ and covariances $\widehat{\text{Cov}}\left(\hat{\beta}_{j_1}, \hat{\beta}_{j_2}\right)$, $j_1 \neq j_2$, j_1 and $j_2 = R, R+1, ...J$.

4. Estimate the coefficients $\hat{\beta}_r$, $r = 1, 2, ...R-1$ of the $R-1$ eliminated covariates as:

$$\hat{\beta}_r = -\frac{\sum_{j=r+1}^{J} \hat{\beta}_j t_j^{(r+1)}}{t_r^{(r+1)}}$$

5. Estimate the standard error of the estimates of the $R-1$ eliminated covariates as:

$$\widehat{\text{SE}}\left(\hat{\beta}_r\right) = \sqrt{\sum_{j=r+1}^{J} \left\{\left[\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\right]^2 \widehat{\text{Var}}\left(\hat{\beta}_j\right)\right\} + \sum_{j_1=r+1}^{J} \sum_{\substack{j_2=r+1 \\ j_2 \neq j_1}}^{J} \left(\left\{\frac{t_{j_1}^{(r+1)} t_{j_2}^{(r+1)}}{\left[t_r^{(r+1)}\right]^2}\right\} \widehat{\text{Cov}}\left(\hat{\beta}_{j_1}, \hat{\beta}_{j_2}\right)\right)},$$

for $r = 1, 2, ...R-1$, where $\widehat{\text{Var}}\left(\hat{\beta}_j\right)$ denotes the estimated variance of $\hat{\beta}_j$, and $\widehat{\text{Cov}}\left(\hat{\beta}_{j_1}, \hat{\beta}_{j_2}\right)$ denotes the estimated covariance between the estimated coefficients $\hat{\beta}_{j_1}$ and $\hat{\beta}_{j_2}$.

6. Estimate the coefficient of the constant covariate from the boundary condition as:

$$\hat{\beta}_0 = -\sum_{j=1}^{J} \hat{\beta}_j x_j^{(1)}$$

7. Estimate the standard error of the constant as:

$$\widehat{\text{SE}}\left(\hat{\beta}_0\right) = \sqrt{\sum_{j=1}^{J} \left\{\left[x_j^{(1)}\right]^2 \widehat{\text{Var}}\left(\hat{\beta}_j\right)\right\} + \sum_{j_1=1}^{J} \sum_{\substack{j_2=1 \\ j_2 \neq j_1}}^{J} \left[x_{j_1}^{(1)} x_{j_2}^{(1)} \widehat{\text{Cov}}\left(\hat{\beta}_{j_1}, \hat{\beta}_{j_2}\right)\right]}.$$

To complete the implementation of the exact method for the LBM, it is necessary to provide formulae for the covariances of the estimated coefficients of the eliminated covariates (including the constant) with each other and with the estimated coefficients of the non-eliminated covariates. Two theorems with formulae are provided in Chapter 2 of this thesis. They extend in a straightforward way to the marginal LBM by GEE as outlined below. The proofs of two theorems are provided in Appendix C.

Theorem 4.1

For a marginal log binomial model estimated by generalised estimating equations with $J \geq 1$ *independent non-constant covariates fitted by the exact method to data having* R $(1 \leq R \leq J)$ *distinct sets of values of the covariates (including the constant) for which the estimated mean of outcome evaluated at the solution is unity, the estimated covariances between the estimates of the coefficients of the* $R-1$ *eliminated non-constant covariates and the estimates of the coefficients of the other non-constant covariates are given by:*

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_r, \hat{\beta}_s\right) = -\left\{ \frac{t_s^{(r+1)}}{t_r^{(r+1)}} \widehat{\mathrm{Var}}\left(\hat{\beta}_s\right) + \sum_{\substack{j=r+1 \\ j \neq s}}^{J} \left[\frac{t_j^{(r+1)}}{t_r^{(r+1)}} \widehat{\mathrm{Cov}}\left(\hat{\beta}_j, \hat{\beta}_s\right) \right] \right\},$$

for $r = 1, 2, \ldots R-1$ *and* $s = r+1, \ldots J$.

To attain the covariances between the estimated coefficient of the constant covariate and the estimated coefficients of the non-constant covariates, Theorem 4.2 is provided.

Theorem 4.2

For a marginal log binomial model estimated by generalised estimating equations with $J \geq 1$ *independent non-constant covariates fitted by the exact method to data having* R $(1 \leq R \leq J)$ *distinct sets of values of the covariates (including the constant) for which the estimated mean*

of outcome evaluated at the solution is unity, the estimated covariances between the exact estimate of the coefficient of the constant covariate and exact estimates of the coefficients of the J non-constant covariates are given by:

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_0, \hat{\beta}_j\right) = -\left[x_j^{(1)} \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=1 \\ j_1 \neq j}}^{J} x_{j_1}^{(1)} \widehat{\mathrm{Cov}}\left(\hat{\beta}_j, \hat{\beta}_{j_1}\right) \right], \; j = 1, 2, \ldots J .$$

4.3.2 Identifying the boundary vector(s)

Prior to the estimation of the marginal LBM by GEE, it is not known whether or not a covariate vector will constitute a boundary vector. Moreover, a reliable scheme to correctly address the boundary vectors in data is also needed. Therefore, a strategy is provided to overcome the issue as follows:

(a) Fit the standard fitting algorithm of marginal LBM by GEE to obtain an approximate solution, which may not fully estimate the model and the correlation due to the impact of boundary vectors.

(b) Identify all distinct covariate vectors for observations with estimated means close to 1 evaluated at the approximate solution, and treat them as candidate boundary vectors.

(c) Test each candidate boundary vector, and combinations of them as a combination of the candidate boundary vectors, until the exact solution is found. The approximate fitted probabilities (largest first) are used to prioritise the order of selection.

4.3.3 Model selection criteria for the marginal LBM by GEE

Model selection is aimed at finding the model that is nearest to the true model from a set of potential candidate models. In the model relying on a likelihood function, the solution that maximises the likelihood is acknowledged as the best fit to the data. Marginal LBM by GEE,

however, is based on a quasi-likelihood function which does not fully specify the joint distribution, but only depends on the first two moments (detailed in Section 4.2 of this chapter). Therefore, the likelihood-based criteria (such as AIC and BIC) can not be used for the model selection in GEE. Some criteria have previously been introduced in GEE for model selection. Each of them has its own advantages and disadvantages for informing the selection of covariates and correlation structure for the model [52-55, 71].

Subsets of covariates or the correlation structure are the two main determinants in the model selection of GEE. Changes in either of those two components will lead to changes in the criteria [52-55, 71]. The way to evaluate the performance of criteria in the model selection is to compare the variations of criteria attributed to the changes in one component by holding another constant. This was documented in the design of the simulation study in previous literature about the comparison of the criteria in GEE [53, 55, 71, 72]. In the exact method, the subsets of covariates and the correlation structure remain constant, and the model will be fitted twice before eliminating the boundary vectors and after eliminating the boundary vectors by the exact method. By doing so, the only factor that could lead to a change in the criteria is the elimination of the boundary vectors. In Chapter 2 and 3 of this thesis, it has been seen that the model fit is improved after eliminating the boundary vector. Similarly, after eliminating the boundary vector, an improvement in the criterion is expected to be observed in the marginal LBM by GEE. The five criteria in GEE are explored in the next part of this section. A simulation study in Section 4.4 will assess their performances in the exact method.

1. *Rotnitzky – Jewell* (RJ)

Rotnitzky and Jewell [51] provided and proved the asymptotic properties of the working Wald statistic for regression coefficients in the generalised estimating equations. Hin et al. [71] summarised their idea into a criterion to select an intra-cluster correlation structure in GEE named as Rotnitzky – Jellew's criterion, $\mathrm{RJ}(R)$. The criterion formula is written as:

$$\begin{aligned}\mathrm{RJ}(R) &= \sqrt{\left[1-RJ1\right]^2+\left[1-RJ2\right]^2} \\ &= \sqrt{\left[1-\frac{\mathrm{trace}(\boldsymbol{\Gamma})}{p}\right]^2+\left[1-\frac{\mathrm{trace}(\boldsymbol{\Gamma}^2)}{p}\right]^2}\end{aligned}$$

where $\boldsymbol{\Gamma}=\left(\sum_{i=1}^{K}\mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{D}_i\right)^{-1}\mathbf{M}$, and p is the number of covariates. In the situation where a working correlation structure is correctly assigned, the matrix $\boldsymbol{\Gamma}$ is approximately to be an identity matrix. Therefore, $\mathrm{trace}(\boldsymbol{\Gamma})/p$ and $\mathrm{trace}(\boldsymbol{\Gamma}^2)/p$ are a number very close to 1, thus $\mathrm{RJ}(R)$ approaches 0.

2. *Shults–Chaganty criterion* (SC)

Shults and Chaganty [54] argued that an appropriate working correlation structure should minimise the generalised error sum of squares written as:

$$\begin{aligned}\mathrm{GESS}(\boldsymbol{\alpha},\boldsymbol{\beta}) &= \sum_{i=1}^{K}(\mathbf{y}_i-\mathbf{u}_i)'\mathbf{V}_i^{-1}(\mathbf{y}_i-\mathbf{u}_i) \\ &= \sum_{i=1}^{K}\mathbf{Z}_i'(\boldsymbol{\beta})\mathbf{R}^{-1}(\boldsymbol{\alpha})\mathbf{Z}_i(\boldsymbol{\beta})\end{aligned},$$

where $\mathbf{Z}_i(\boldsymbol{\beta})=\mathbf{A}_i^{-1/2}\left(\mathbf{y}_i-\mathbf{u}_i\right)$. A relevant criterion for model selection in GEE was defined as:

$$\mathrm{SC}(R)=\frac{\mathrm{GESS}(\boldsymbol{\alpha},\boldsymbol{\beta})}{(N-p-q)},$$

where $N=\sum_{i=1}^{K}n_i$ is the total number of observations, p is the number of covariates, and q is the number of correlation coefficients in the working correlation structure.

3. *Gaussian pseudolikelihood* (GP)

Carey and Wang [55] proposed a multivariate Gaussian criterion for choosing the correlation structure as an extension of the generalised error sum of squares. The criterion is written as:

$$\mathrm{GP}(R) = -0.5\sum_{i=1}^{K}\left[(\mathbf{y}_i - \mathbf{u}_i)'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{u}_i) + \ln(|\mathbf{V}_i|)\right].$$

A larger $\mathrm{GP}(R)$ indicates a better correlation structure.

4. *Quasi-likelihood under the independence model criterion* (QIC)

Pan [52] produced a criterion for GEE by modifying the Akaike information criterion (AIC). The criterion consists of two terms: the value of the quasi-likelihood under the independence assumption evaluated at $\hat{\boldsymbol{\beta}}_R$, and a penalty term. It can be written as:

$$\mathrm{QIC}(R) = -2\,\mathrm{Q}\left[\hat{\boldsymbol{\beta}}_R, \hat{\phi}\right] + 2\,\mathrm{trace}\left(\hat{\boldsymbol{\Omega}}_I \hat{\mathbf{V}}_R\right),$$

where:

- $\mathrm{Q}\left[\hat{\boldsymbol{\beta}}_R, \hat{\phi}\right] = \sum_{i=1}^{K}\sum_{j=1}^{n_i}\mathrm{Q}\left[\hat{\boldsymbol{\beta}}_R, \hat{\phi}; \left(y_{ij}, x_{ij}\right)\right]$ is the quasi-likelihood with $\mathrm{Q}(u, \hat{\phi}; y) = \int_y^u \frac{y-t}{\hat{\phi}\,\mathrm{V}(t)}dt$. $\hat{\boldsymbol{\beta}}_R$ and $\hat{\phi}$ are obtained under the hypothesised correlation structure.
- $\hat{\boldsymbol{\Omega}}_I = \sum_{i=1}^{K}\mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{D}_i\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_R, \mathbf{R}=\mathbf{I}}$ is the model-based covariance estimate under the independent correlation assumption, evaluated at $\hat{\boldsymbol{\beta}}_R$.
- $\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_R}$ is the robust (sandwich) variance-covariance estimator.

A lower $\mathrm{QIC}(R)$ indicates a better model fitting.

5. *Correlation information criterion* (CIC)

Hin and Wang [53] found that the performance of $\mathrm{QIC}(R)$ is affected by the quasi-likelihood term that is theoretically independent of the correlation structure. In addition, they stated that the penalty term of $\mathrm{QIC}(R)$ had a better performance in the selection of correlation structure and proposed a correlation information criterion (CIC) by only including the penalty term in $\mathrm{QIC}(R)$ as:

$$\mathrm{CIC}(R) = \mathrm{trace}\left(\hat{\boldsymbol{\Omega}}_I \hat{\mathbf{V}}_R\right).$$

A better correlation structure should minimise the $\mathrm{CIC}(R)$.

There are also some other criteria introduced by researchers previously for the model selection in GEE. However, they are not covered in this thesis due to their unsatisfactory performance in the model selection when the outcome variable is binary [73] and have been used rarely in practice.

4.3.4 The presence of a boundary vector may lead to the underestimation of the robust standard error and make model selection criteria fail

The presence of the boundary vector does not necessarily make the standard fitting algorithm of marginal LBM by GEE fail to converge. When an appropriate starting value is provided, the standard fitting algorithm can converge to an approximate solution in which the estimated mean of a boundary vector is extremely close to 1. This solution could be problematic as it may distort the calculation of $\hat{\mathbf{V}}_i^{-1}$, leading to an underestimation of the robust variance-covariance matrix and the corresponding standard error. For example, with an estimated mean extremely close to 1, the value of $\widehat{\mathrm{Var}}\left(y_{ip}\right) = \mu_{ip}\left(1-\mu_{ip}\right)$ will be enormously small, and its reciprocal becomes tremendously large. This huge value will dominate the weights in the calculation of $\hat{\mathbf{V}}_i^{-1}$, and consequently make $\hat{\mathbf{V}}_i^{-1}$ fail to represent the entire picture of

variation of the estimated means, then affect all of the estimations based on $\hat{\mathbf{V}}_i^{-1}$. Since the estimation of the robust standard error and the model selection criteria depend on $\hat{\mathbf{V}}_i^{-1}$, they may lose their power in this situation.

In contrast, because the boundary vectors are eliminated in the exact method by a re-parameterisation, and will not be mathematically involved in the estimation of the model, they will not affect the calculation of $\hat{\mathbf{V}}_i^{-1}$. More studies are needed to evaluate how this issue impacts the estimation of robust variance-covariance matrix and the criteria. The five criteria mentioned in this section are offered as options in the package *lb.gee* written in R-3.6.3.

4.4 Simulations

The simulation study is designed with two objectives. The first one is to evaluate the performance of the exact method in estimating marginal LBM by GEE. The aforementioned criteria are used for identifying the boundary vectors. The influence of the presence of boundary vectors on the performance of each criterion is also assessed by the number of improved replications (improved replication means that the value of criterion is improved in replication after eliminating the boundary vector). The second objective is to evaluate the performance in the model fit of the marginal LBM by GEE by comparing its estimates with those of the marginal Poisson model. As described in Section 4.1, the marginal Poisson by GEE is commonly used to approximate the coefficients of the marginal LBM by GEE model when convergence problems are encountered in fitting the latter. Therefore, it is worthwhile to compare the two approaches.

Two statistics were used in evaluating the model fit: the average percent relative bias of the estimate from the designed value $(100/N)\sum_{n=1}^{N}\left(\hat{\beta}_{pn}-\beta_p\right)/\beta_p$, and the percentage of 95 percent confidence intervals covering the designed value of the coefficient. Since the selection of intraclass correlation structure is not the aim of our study, the data is only generated following an exchangeable correlation structure with a correlation coefficient $\rho = 0.3$ or 0.6 (the exchangeable correlation structure is the most common correlation structure applied in practice).

All calculations were performed using R version 3.6.3. The correlated binary outcome variable was generated through the R function `rmvbin` in the package `bindata`, which was built using the method by Leisch et al. [74]. The function for the marginal means is as follows:

$$\log(\mu_{ip}) = \beta_0 + \beta_1 x_{1ip} + \beta_2 x_{2ip},\ i = 1,2,...,K \text{ and } p = 1,2,...,n_i,$$

where β_0, β_1 and β_2 are three coefficients with design values $\log(0.1)$, $\log(1.25)$ and $\log(2)$. K is the total number of clusters in the sample. Let n_i denote the number of observations under the i^{th} cluster. The x_{1ij} is a binary covariate randomly generated from a Bernoulli distribution with mean 0.5, and x_{2ij} is a continuous covariate randomly generated from a uniform distribution with the upper bound attained through function

$$x_{2upper} = \frac{\log(1) - \log(0.1) - \log(1.25)}{\log(2)}$$

and the lower bound obtained by

$$x_{2lower} = \frac{\log(0.15) - \log(0.1)}{\log(2)}.$$

Table 4.1 shows the settings for eight scenarios. Each scenario involved $N = 10000$ replications of a dataset of K observations of a binary outcome.

Table 4.1: The settings in the simulation.

Setting	K^*	$\overline{n}_i$ §	Cluster size	ρ †	β_0	β_1	β_2
1	500	3	Fixed	0.3	$\ln(0.1)$	$\ln(1.25)$	$\ln(2)$
2				0.6	$\ln(0.1)$	$\ln(1.25)$	$\ln(2)$
3	500	3	Floating	0.3	$\ln(0.1)$	$\ln(1.25)$	$\ln(2)$
4				0.6	$\ln(0.1)$	$\ln(1.25)$	$\ln(2)$
5	50	10	Fixed	0.3	$\ln(0.1)$	$\ln(1.25)$	$\ln(2)$
6				0.6	$\ln(0.1)$	$\ln(1.25)$	$\ln(2)$
7	50	10	Floating	0.3	$\ln(0.1)$	$\ln(1.25)$	$\ln(2)$
8				0.6	$\ln(0.1)$	$\ln(1.25)$	$\ln(2)$

K^* is the total number of clusters in each generated dataset. $\overline{n}_i$ § is the average number of observations in each cluster with $i = 1,2,3,4$ denoting settings (1,2), (3,4), (5,6) and (7,8) respectively. ρ † is the correlation coefficient in the exchangeable correlation structure. In order to study the performance of the exact method in data with balanced and unbalanced cluster, samples with fixed and floating cluster size were generated respectively. In the floating clusters, the cluster size was generated randomly from a Poisson distribution with mean equal to 3 or 10. In contrast, in the fixed clusters, each cluster has a fixed size of $\overline{n}_i$ §.

4.4.1 Results for the comparison of the model selection criteria

All replications were fitted by marginal LBM by GEE using the exact method. Table 4.2 shows the results of the marginal LBM by GEE using the exact method from eight scenarios under five criteria. As expected, by appropriately eliminating boundary vectors using the exact method, the overall model fit evaluated under the respective criterion is improved to a different extent in each scenario. About 82-87 percent of replications in the first four scenarios and 42-59 percent of replications in the last four scenarios produce an improved result in RJ and CIC when applying the exact method. This percentage reduced to about 72-75 percent in the first four and 34-55 percent in the last four in QIC, 48-62 and 14-26 percent in SC, and 15 to 20 and 7-8 percent in GP. As a result, the number of improved repetitions in each criterion is reduced by half in the last four scenarios compared to the first four. It is due to the reduction in the number of clusters and the increment of the cluster sizes. In previous

studies, the changes in those two sizes have been shown to be critical factors affecting the model selection and the model fit in GEE through the variance-covariance function $\mathbf{V}_i$ [69, 75-77]. The results from the simulation show that the presence of the boundary vector in the replications may be affected by the number of clusters and the cluster sizes. However, the relationship between the presence of boundary vectors and the variation of the total number of clusters and the cluster size is not clear and needs further study, but it will not be discussed in this thesis.

After eliminating the boundary vector by the exact method, the total number of repetitions with improved results are different in each of the five criteria. There is a wide gap in the number of improved replications between the five criteria. The likely reason is that the boundary vector has a different effect on the performance of each criterion. As previously mentioned in Section 4.3.4, the presence of boundary vector in the fitting procedure may affect the estimation of $\hat{\mathbf{V}}_i^{-1}$, and the relevant criterion obtained based on it. The results have shown a significant improvement in RJ, QIC and CIC while SC and GP did not after the boundary vectors are eliminated by the exact method. It can be inferred that RJ, QIC and CIC are more sensitive to the presence of boundary vectors as compared with the other two, indicating that RJ, QIC and CIC are more efficient at identifying boundary vectors in an exact method.

The coverage rate for the estimates β_1 and β_2 in the five criteria are about 92-95 percent which is slightly lower than the target 95 percent, likely due to an underestimate of the robust standard error [75, 76, 78, 79]. Some robust standard error corrections and alternative covariance estimators in GEE have been previously proposed to correct the problem [80-88]. They have their advantages and disadvantages in terms of variance estimation and covariance

estimation. I do not discuss these methods in this thesis since the performance comparison between those methods is beyond the scope.

Table 4.2: The simulation results from marginal LBM by GEE in the exact method under five criteria.

		RJ			QIC			CIC			SC			GP	
		n*	Bias§	Cov†	n*	Bias§	Cov†	n*	Bias§	Cov†	n*	Bias§	Cov†	n*	Bias§	Cov†
1	β_1	8246	–0.95	94.9	7539	–0.54	94.7	8339	–0.05	94.8	5753	–1.48	94.7	1772	–0.92	94.7
	β_2		–0.37	93.8		–0.26	93.9		–0.07	94.0		–0.58	93.8		–0.47	93.9
2	β_1	8706	0.23	94.1	7419	0.03	93.3	8379	0.38	93.7	6238	–0.68	94.0	1489	–0.24	93.0
	β_2		–0.20	94.3		–0.26	93.9		–0.15	94.0		–0.46	94.1		–0.38	93.6
3	β_1	8044	–1.17	94.3	7195	–1.09	94.2	8576	–0.38	94.4	5368	–1.71	94.2	2010	–1.48	94.1
	β_2		–0.75	93.6		–0.72	93.5		–0.47	93.7		–0.95	93.6		–0.89	93.5
4	β_1	8182	0.35	94.2	7191	–0.11	93.9	8651	0.34	94.4	4767	–0.49	93.8	1678	–0.32	93.3
	β_2		–0.61	93.7		–0.73	93.6		–0.62	93.7		–0.79	93.4		–0.77	93.4
5	β_1	4528	–5.44	91.8	5539	–4.73	91.7	5859	–4.69	91.9	1433	–5.17	91.5	789	–5.18	91.6
	β_2		–2.58	92.2		–2.44	92.0		–2.48	92.0		–2.50	92.1		–2.50	92.1
6	β_1	4783	–0.64	92.6	3446	–0.83	92.5	4377	–1.00	92.7	1585	–0.61	92.3	719	–0.66	92.2
	β_2		–1.07	93.9		–0.79	93.9		–0.91	93.7		–0.96	94.0		–0.94	94.0
7	β_1	4521	–5.12	92.5	5164	–4.42	92.4	5595	–4.47	92.5	1985	–4.62	92.4	732	–4.72	92.5
	β_2		–2.68	92.0		–2.56	91.8		–2.59	91.9		–2.59	91.8		–2.61	91.9
8	β_1	4720	–0.86	92.4	3431	–1.06	92.6	4292	–1.25	92.9	2557	–0.87	92.3	753	–0.81	92.1
	β_2		–1.10	93.8		–0.79	93.7		–0.89	93.6		–0.95	93.5		–0.95	93.6

*n is the number of repetitions with improved results using the exact method under each criterion.

§Average percent relative bias of the estimate from the design value: $(100/N)\sum_{n=1}^{N}\left[\left(\hat{\beta}_{jn}-\beta_j\right)/\beta_j\right]$, $j=0,1,2$ and $n=1,2,...N$.

† Cov is the percentage of 95 percent confidence intervals covering the design value of the coefficient.

4.4.2 Comparison with the alternative approaches

This section further explores the performance of marginal LBM by GEE using the exact method. The results of the marginal LBM by GEE using the exact method are compared with the marginal Poisson by GEE, which is an approximate alternative approach used when there are numerical difficulties in the standard fitting algorithm for the marginal LBM by GEE. The comparison only includes the results of the marginal LBM by GEE using the exact method from the criteria RJ, QIC and CIC. The results from the criteria SC and GP are not included because of their unsatisfied performance in identifying boundary vectors.

Table 4.3: Simulation results (with results from the marginal Poisson by GEE shown for comparison)

		Exact method						Poisson	
		RJ		QIC		CIC			
		Bias*	Cov§	Bias*	Cov§	Bias*	Cov§	Bias*	Cov§
1	β_1	–0.95	94.9	–0.54	94.7	–0.05	94.8	–0.63	95.0
	β_2	–0.37	93.8	–0.26	93.9	–0.07	94.0	–0.28	94.4
2	β_1	0.23	94.1	0.03	93.3	0.38	93.7	–0.45	94.7
	β_2	–0.20	94.3	–0.26	93.9	–0.15	94.0	–0.31	94.5
3	β_1	–1.17	94.3	–1.09	94.2	–0.38	94.4	–1.27	94.3
	β_2	–0.75	93.6	–0.72	93.5	–0.47	93.7	–0.79	94.3
4	β_1	0.35	94.2	–0.11	93.9	0.34	94.4	–0.73	95.0
	β_2	–0.61	93.7	–0.73	93.6	–0.62	93.7	–0.86	94.3
5	β_1	–5.44	91.8	–4.73	91.7	–4.69	91.9	–3.81	93.4
	β_2	–2.58	92.2	–2.44	92.0	–2.48	92.0	–1.96	92.9
6	β_1	–0.64	92.6	–0.83	92.5	–1.00	92.7	–0.52	94.1
	β_2	–1.07	93.9	–0.79	93.9	–0.91	93.7	–0.86	93.3
7	β_1	–5.12	92.5	–4.42	92.4	–4.47	92.5	–3.47	93.9
	β_2	–2.68	92.0	–2.56	91.8	–2.59	91.9	–1.93	92.6
8	β_1	–0.86	92.4	–1.06	92.6	–1.25	92.9	–0.36	94.0
	β_2	–1.10	93.8	–0.79	93.7	–0.89	93.6	–0.80	93.1

§Average percent relative bias of the estimate from the design value: $(100/N)\sum_{n=1}^{N}\left[\left(\hat{\beta}_{jn}-\beta_j\right)/\beta_j\right]$, $j=0,1,2$ and $n=1,2,...N$.

§Cov is the percentage of 95 percent confidence intervals covering the design value of the coefficient.

Table 4.3 presents the results for the average percent bias and the 95 percent coverage rate of the marginal LBM by GEE using the exact method from three criteria and the marginal Poisson by GEE. The results of the marginal LBM by GEE using the exact method are not significantly improved compared with the results of the marginal Poisson by GEE in terms of average percent bias. However, the improvements are more evident in the density plots of the bias. Figure 4.1 shows the density plots of the average percent bias for the estimate $\hat{\beta}_1$ between the results from marginal LBM by GEE using the exact method under three criteria and marginal Poisson by GEE (Figures for the $\hat{\beta}_0$ and $\hat{\beta}_2$ see Appendix C). The density plots are drawn in ascending order of the average percent bias of all replications in each scenario. From figure 4.1, the Normal distribution curves of the marginal LBM by GEE using

the exact method for three criteria in each scenario are close together and does not spread out as much as the curve of the marginal Poisson by GEE. This suggests that compared with the marginal Poisson by GEE, the biases in the marginal LBM by GEE using the exact method are more approaching null.

Marschner and Gillett [22] pointed out that using the unconstrained Poisson method to approximately estimate a log binomial model may result in some risks substantially exceeding 1. Zhu et al. [37] confirmed this issue later in their study and recommended not to use it. This issue also exists in the marginal Poisson by GEE [69]. We found that 40-44% of repetitions in the marginal Poisson by GEE produced an estimated mean exceeding 1. Therefore, the validity of estimates in those repetitions cannot be ensured.

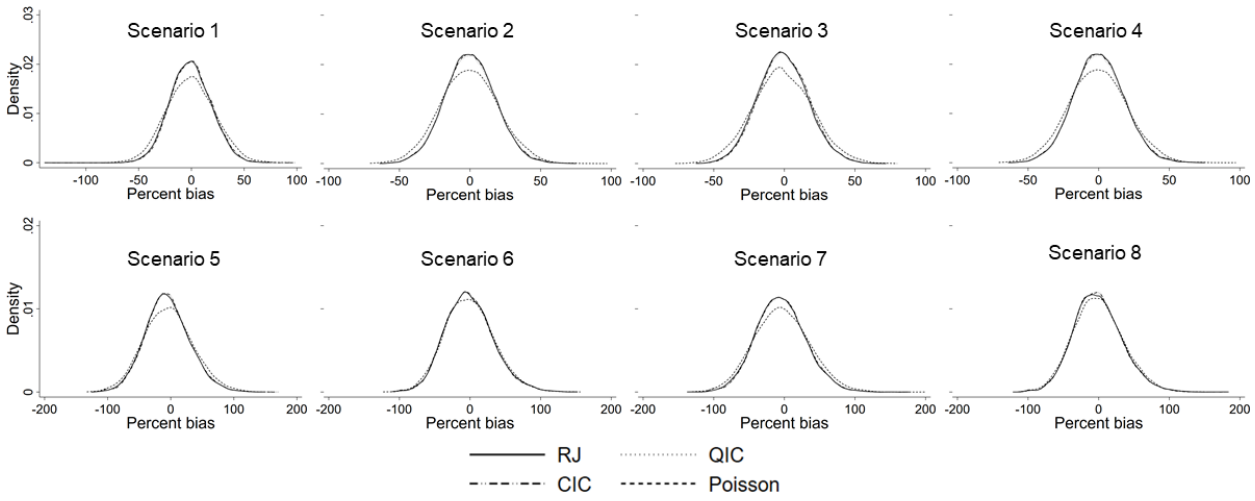

Figure 4.1: Density plots of the sampling distributions (10000 replications, eight scenarios) of the binary covariate β_1. The density plots are drawn in ascending order of the average percent bias of all replications in each scenario. The density curves of the exact method under three criteria in each scenario are close together and do not spread out as much as the curve of Poisson.

4.5 A real world example

We further evaluated the performance of marginal LBM by GEE using the exact method by a dataset of polypharmacy from Hosmer et al. [62]. The example using a sample of $N = 500$

patients who were derived from a cohort of subjects having data in each of the seven years. The outcome is polypharmacy, which is a binary variable describing whether the patient is taking three or more drug classes. The study factors included in the final model were the subjects' age in years and months rounded into two decimal places (*AGE*, from 1.17 to 18.92), number of outpatient mental health visits (*MHV*4, 0 = None, 1 = 1 to 5, 2 = 6 to 14 and 3 = Greater than 14), number of inpatient mental health visits (*INPTMHV*3, 0 = None, 1 = One and 2 = More than one), race (*RACE*, 0 = White, 1 = Black and Other), and gender (*GENDER*, 0 = Female and 1= Male). The reported age range is that of the range of recorded values in this data subset.

To investigate the association between the factors and polypharmacy, data were fitted using the marginal LBM by GEE. The subject ID (*ID*) of the patient was chosen as a cluster identification variable. We used a similar analytical procedure to the one used by Hosmer et al. [62], and only the factors which showed a significant association with outcomes were included in the model. Hosmer et al. [62] created a new dichotomous covariate by combining last two levels in *INPTMHV*3, replacing *AGE* by a natural log-transformed age, and combining Black and Other in *RACE* to make it become a dichotomous covariate. In this example, *INPTMHV*3 and *RACE* will remain unchanged, and the categories in *INPTMHV*3 and *RACE* will be retained. For the continuous covariate *AGE*, I also found that the natural log age could improve the fit of the model instead of the linear form by looking at the scale of age. However, if the *AGE* is rescaled in the natural log form, the boundary vector will no longer exist. Therefore, I decide to leave the *AGE* as linear to keep a boundary vector in the example. All pairs of covariates were tested for interaction, and none of the interactions was significant.

We assumed an identical correlation between observations within a subject and fitted the marginal LBM by GEE using the exact method with an exchangeable correlation structure. This model was fitted successfully, with RJ = 0.0843, QIC = 3501.4838 and CIC = 16.1375. After switching to an autoregressive correlation structure, the RJ = 0.7756, QIC = 3585.0087 and CIC = 16.3996. These results suggest that by assuming an exchangeable correlation structure within clusters, a better model fit can be obtained than using autoregressive correlation structure. Applying the other two criteria listed in Section 4.3 yielded the same conclusion (corresponding results are omitted for brevity). In addition, the solution shows that there was one boundary vector involved in the model, and only RJ, QIC and CIC successfully addressed it. The patient corresponding to the boundary vector represents the oldest white man in the sample data who made more than 14 outpatient mental health visits and one inpatient mental health visit in 2008.

Table 4.4: Results of the final model to the POLYPHARMACY data obtained by three approaches. The 95% confidence intervals associated with the estimates are shown in parentheses.

	Binomial (Exact) ($\rho^* = 0.4167$)	Poisson ($\rho^* = 0.4192$)	SAS-GEE ($\rho^* = 0.4165$)
*MHV*41	0.1097 (-0.1113, 0.3307)	0.1181 (-0.1037, 0.3398)	0.1100 (-0.1107, 0.3308)
*MHV*42	0.4615 (0.2259, 0.6971)	0.4849 (0.2464, 0.7233)	0.4617 (0.2259, 0.6975)
*MHV*43	0.7033 (0.4622, 0.9443)	0.7056 (0.4574, 0.9537)	0.6992 (0.4567, 0.9416)
*INPTMHV*31	0.4477 (0.2822, 0.6132)	0.3329 (0.0743, 0.5915)	0.4225 (0.2162, 0.6288)
*INPTMHV*32	0.2447 (-0.0882, 0.5775)	0.2136 (-0.1721, 0.5993)	0.2415 (-0.0966, 0.5797)
*RACE*1	-0.3143 (-0.5851, -0.0434)	-0.2910 (-0.5785, -0.0035)	-0.2977 (-0.5718, -0.0236)
*RACE*2	-0.5244 (-1.6858, 0.6369)	-0.4780 (-1.6934, 0.7374)	-0.5001 (-1.6613 0.6610)
GENDER	0.3460 (0.0555, 0.6366)	0.3125 (0.0180, 0.6070)	0.3336 (0.0426, 0.6246)
AGE	0.0924 (0.0732, 0.1115)	0.0891 (0.0649, 0.1134)	0.0899 (0.0675, 0.1124)
Constant	-3.2445 (-3.6691, -2.8199)	-3.1979 (-3.6794, -2.7165)	-3.2076 (-3.6687, -2.7464)

The outcome is polypharmacy, which is a binary variable describing whether the patient is taking three or more drug classes. The selected study factors were the subjects' age in years and months rounded into two decimal places (*AGE*, from 1.17 to 18.92), number of outpatient mental health visits (*MHV*4, 0 = None, 1 = 1 to 5, 2 = 6 to 14 and 3 = Greater than 14), number of inpatient mental health visits (*INPTMHV*3, 0 = None, 1 = One and 2 = More than one), race (*RACE*, 0 = White, 1 = Black and Other), and gender (*GENDER*, 0 = Female and 1= Male). ρ^* is the correlation coefficients under the exchangeable correlation structure. To evaluate the solution under the same condition and maintain the consistency of statistics, the convergence tolerance is manually set equal to 1E-06 among three approaches.

Table 4.5: The relevant criterion values of three approaches.

	Exact method	Poisson	SAS-*GEE*
RJ	0.0843	0.1857	0.1442
CIC	16.1375	24.1955	26.2650
QIC	3501.4838	3527.6126	3503.6123

In Table 4.4, the results obtained from the marginal LBM by GEE using the exact method are compared with the results of the other two alternative methods, the marginal Poisson by GEE

and the SAS-GEE. To evaluate the solution under the same condition and maintain the consistency of statistics, the convergence tolerance is manually set equal to 1E-06 among three approaches. The marginal Poisson by GEE and SAS-GEE (function under SAS-9.4) are chosen as competitors as they are the only two approaches that can address the convergence issue in this data. Compared with marginal LBM by GEE using the exact method, SAS-*GEE* converges to an inappropriate solution with higher criteria (Table 4.5). The higher criterion value indicates that the model does not fit the data well compared with the lower criterion. In the results obtained by SAS-*GEE*, the estimated mean of the boundary vector is equal to 0.9499. This indicates that SAS-*GEE* stops the fitting algorithm at a distance from the boundary to avoid the generation of boundary vector. Although this may avoid the convergence problem due to the presence of the boundary vectors, it may cause the model to be unable to be fully estimated. The marginal Poisson by GEE converges to an unsatisfactory solution with the highest criteria (Table 4.5). The estimated mean of the boundary vector is equal to 0.8518.

The other functions that fit the marginal LBM by GEE in R, Stata, and SAS are also tested, but they are excluded for the above comparisons, with reasons detailed below. R-*geese* in the package R-*geepack*-1.3-1 fails to begin the iteration with a default starting value and asked for an admissible user-supplied starting values. After being provided with an appropriate starting value, it reaches an inappropriate solution with RJ, QIC and CIC larger than the ones in the marginal LBM by GEE using the exact method. Stata-*xtgee* with the default starting values in Stata-16 fails to converge and reports that the estimates were diverging. The issue remains after providing an admissible starting value. Likewise, SAS-*GENMOD*, another function to estimate marginal LBM by GEE under SAS-9.4 was excluded due to its failure of convergence.

4.6 Discussion

Failure of convergence in the standard fitting algorithm in marginal LBM by GEE remains an unresolved issue. No study has been done yet to address it. The issue may occur due to the presence of boundary vectors. To overcome the issue, the exact method is extended into marginal LBM by GEE. In the exact method, the boundary vectors are eliminated by re-parameterizing the model. Two theorems with proofs for the missing covariance formula in the exact method were provided. The performance of the method is explored by performing a simulation study with five different criteria (RJ, QIC, CIC, SC and GP) used for model selection in GEE. An example is used to demonstrated how to implement the marginal LBM by GEE using the exact method in practice.

The present results in the simulation demonstrate that the exact method successfully resolves the convergence issue and improves the model fitting to a different extent using diverse criteria. The greatest improvement is observed in RJ, QIC and CIC, following by SC and GP. These results suggest that the RJ, CIC and QIC were sensitive to the exact method while the SC and GP were not. The average percent bias and 95 percent coverage were comparable between these criteria. In addition, there was a trend toward a reduced bias and improved coverage due to the increased number of improved replications in the model fitting.

The simulation results of marginal LBM by GEE using the exact method are compared with the results in marginal Poisson by GEE. The results showed no evident difference in the mean of the density of the average percent bias between the two approaches. However, the peak of the density curve was higher and the shape of the curve less spread out in the marginal LBM by GEE using the exact method. In addition, there are about 40-44% of repetitions in the

marginal Poisson by GEE produced an estimated mean outside the interval [0, 1]. Although marginal Poisson by GEE has been widely used as an alternative of marginal LBM by GEE in which there is a convergence issue, the results from this chapter suggest that using marginal Poisson by GEE could be problematic due to its larger bias in estimations and potentially produce a solution with an estimated mean outside the interval [0, 1].

In an example dataset, one boundary vector is identified by the criteria RJ, QIC and CIC in the marginal LBM by GEE using the exact method. The standard fitting algorithms are tested for fitting a GEE model from three statistical software packages. They either are not successful in estimating a solution or, if convergence is attained, provides approximations that in some cases were poor. The marginal Poisson by GEE also produced suboptimal approximations.

In conclusion, the exact method shows a good performance in the estimations of marginal LBM by GEE when a boundary vector is present. It successfully improved the model fit in evaluations using different criteria for model selecting. Marginal Poisson by GEE could produce an estimate with a larger bias and inappropriate estimated mean, thus should not be used as an alternative of marginal LBM by GEE.

Chapter 5 Discussion

5.1 Background

Greenland [5] argues that for summarising the impact of exposure on a binary outcome, risk differences and ratios are fundamental measures of effect and the effect measures of choice. The risk ratio/relative risk is the ratio measure of choice for summarizing the impact of exposure on the incidence proportion ("risk") in epidemiologic studies [5]. It can be estimated by fitting a log binomial model in independent data (meaning that the different observations do not depend on each other) and a marginal log binomial model estimated by GEE (marginal LBM by GEE) in clustered/longitudinal data. Risk difference is the observed difference in risk of a binary outcome between reference and study groups. It can be estimated through an identity-link binomial model. Arguably, the risk ratio and risk difference are axiomatic as measures of effect for a closed cohort, including trials with binary outcomes. However, the standard fitting algorithm of these three models often experiences numerical difficulties and fails to converge. In this thesis, I discuss the numerical difficulties due to an inappropriate starting value in the log binomial model and the identity-link binomial model, and the difficulties caused by the ML solution lying on the boundary of the parameter space in all three models.

The numerical difficulties due to a poor starting value can be solved by providing an appropriate starting value. A model-based starting value selecting approach was provided in terms of the log binomial model and the identity-link binomial model in Chapters 2 and 3, respectively. For the numerical difficulties due to the maximum likelihood solution lying on the boundary of the allowable parameter space, the issue is more complicated due to the specialized characteristics of each of the three models.

As demonstrated in Chapter 2, the fitting procedure of the log binomial model requires constraints on $\boldsymbol{\beta}$ to ensure that the fitted probabilities remain inside the interval [0, 1] [6]. The interval is mapped to $(-\infty, 0]$ by the log link function [7], which means the parameter space of the log binomial model has an upper bound at 0. If the maximum likelihood solution lies on the boundary of the parameter space, it means that at least one fitted probability is equal to 1 $(\mu = 1)$, which leads to the corresponding estimated variance $\text{Var}(\mu) = \mu(1-\mu)$ equal to zero. This will cause disruption to the calculation of the inverse of the Fisher information matrix, and the standard fitting algorithm will encounter numerical difficulties and fail to converge. Some alternative methods [17-20, 22, 26, 61] have been introduced previously to overcome this issue (details are in Chapter 2). Some of them have been shown to produce the ML solution for simple models, such as a model with only one boundary vector presents [18, 22]. However, none of them has entirely succeeded in estimating the ML solution in a complicated model with multiple boundary vectors involved.

In the identity-link binomial model, the numerical difficulties due to the ML solution lying on the boundary of the parameter space is even more complicated. This is because the identity-link binomial model can have two boundaries, the lower and upper bounds. When the ML solution lies on the boundaries of the parameter space, it can be one of the following three situations: (1) the ML solution lies on the lower bound in which at least one fitted probability is equal to 0; (2) the ML solution lies on the upper bound in which at least one fitted probability is equal to 1; or (3) the ML solution lies on both the lower and upper bounds in which one fitted probability is equal to 0 and at least one fitted probability is equal to 1, respectively. When the ML solution lies on the boundaries, the calculation of the inverse of the Fisher information matrix will encounter problems, and the standard fitting algorithm will

meet the numerical difficulties and fail to converge. Some workaround methods were designed to estimate the risk difference previously (details are in Chapter 3). However, none of them was entirely successful and did not directly work on the numerical difficulties caused by the ML solution on the boundary in identity-link binomial models. The convergence issue restricted a broad implementation of the identity-link binomial model.

The GEE is a semiparametric approach for estimating coefficients of the model in correlated data and does not rely on a likelihood function to fit the model. Instead, it obtains the first and second moments of the outcome variable based on the quasi-likelihood function and combines with the correlation matrix to build the estimating equations. Eventually, the population parameters are iteratively estimated through a modified Fisher's scoring algorithm established by estimating equation and its derivation [46]. To estimate the relative risk in the clustered/longitudinal data, the marginal LBM by GEE provides a viable choice. However, if there is a covariate vector with an estimated mean μ equal to 1, the standard fitting algorithm will meet numerical difficulties, and fail to converge or converge to an inappropriate solution.

Petersen and Deddens [9] introduced the exact method to overcome the numerical difficulties in the log binomial model caused by the ML solution lying on the boundary of the parameter space. The exact method eliminates the boundary vectors by re-parameterizing the covariates and makes the model fit without the impact of the boundary vectors. Petersen and Deddens [9] provided the formulas to estimate the coefficients and the corresponding standard errors of covariates, and used a simple example with two boundary vectors to explain how to implement their method. In the example, the method successfully reached the ML solution, which was on the boundary of the parameter space. However, there are some deficiencies

which restrict the application of the method. The author pointed out a deficiency in that two applicable requirements must be met before the exact method can be implemented. The first requirment is that the model must be identified as a case in which the ML solution lies on a boundary of the parameter space. The second is that the boundary vectors should be determined before fitting the model. In practice, it is usually impossible to know the location of the ML solution and the boundary vectors before the model is fitted. Another deficiency is that the covariances between the coefficients of the covariates are key factors to estimate the standard errors of the relevant coefficients. However, a formula to estimate the covariances was missing. Petersen and Deddens [9] provided an informal method to obtain the covariances in a simple example with two covariates and two boundary vectors by exchanging the order in which the boundary vectors were eliminated in the re-parameterization procedure. However, since the details to implement the method were missing, it is not clear that this informal method to obtain the covariances can be implemented in general cases. On account of the incompleteness and the lack of implementation details, the method did not receive enough attention in the world.

This thesis aims to supplement the necessary contents for the generalization of the exact method and extend it to overcome the numerical instability in the identity-link binomial model and improve the model fitting in the marginal LBM by GEE. The respective aim of the three studies included in this thesis was summarized as follows:

1. To provide the mathematical details and practical guidance necessary to implement the exact method of fitting the log binomial model, to demonstrate and evaluate fits by the exact method to example and simulated data, and to compare the estimates to

approximations made by alternative methods of fitting the log binomial model. (Chapter 2)

2. To extend the exact method to fitting the identity-link binomial model, to provide the mathematical details and practical guidance necessary for that purpose, to demonstrate and evaluate fits by the exact method to example and simulated data, and to compare the estimates to approximations made by alternative methods of fitting the identity-link binomial model. (Chapter 3)

3. To extend the exact method to fitting the marginal LBM by GEE, to provide the mathematical details and practical guidance necessary for that purpose, to demonstrate and evaluate fits by the exact method to example and simulated data, and to compare the estimates to approximations made by the marginal Poisson by GEE. (Chapter 4)

5.2 Overview of the results

Petersen and Deddens [9] introduced the exact method to solve the convergence issue in the log binomial model due to the ML solution on the boundary of the parameter space. Yet, the method has not received enough attention due to the lack of implementation details and two unachievable prerequisites. Chapter 2 re-summarizes the numerical difficulties in the log binomial model and provides the solutions for two types of difficulties, the poor starting value and the ML solution on the boundary of the parameter space. A designed algorithm is provided to correct the inappropriate starting value obtained from the default starting value algorithm. Two theorems with proofs were introduced to estimate the covariances between covariates and fill in the missing part of the exact method in the estimation of the ML

solution when it lies on the boundary of the parameter space. A theorem with proof is introduced to explain the perfect linear correlation between the exact estimates of the model when the number of non-constant covariates is equal to the number of boundary vectors. With the prior theorem satisfied, a corollary with proof is provided to explain an equivalence among the absolute of the standardized estimates of the coefficients (including the constant). Finally, a theorem and corollary (with proofs) to establish that the re-parameterization of covariates required to implement the exact method can be successfully undertaken if the ML solution exists. The application of the method is explained in detail through an example data and real-world data. A simulation study is used to compare the results from using the exact method with the results from alternatives.

In Chapter 3, I have described two numerical difficulties in the identity-link binomial model, the inappropriate starting values and ML solution on the boundary of the parameter space. A starting value calibration algorithm was produced to correct inappropriate values obtained from the default starting value algorithm. To overcome the numerical difficulties due to the ML solution on the boundary of the parameter space, the exact method was extended to the identity-link binomial model. Since the identity-link binomial model has both lower and upper bounds of the parameter space, the numerical difficulties due to boundary vectors are more complicated in the identity-link binomial model as compared with the log binomial model. I produce eight theorems and two corollaries with proofs to provide the theoretical justification for the application of the exact method. The application of the method is demonstrated through an example and real-world data. The results from using the exact method are compared with the results from other approaches in a simulation study.

The numerical difficulties in the marginal LBM by GEE caused by the boundary vectors in clustered/longitudinal data are similar to the difficulties due to the boundary vectors in the log binomial model fitting the independent data. In Chapter 4, I have demonstrated that the exact method can overcome the numerical difficulties in the marginal LBM by GEE due to the boundary vector. A theorem is provided to estimate the covariances between the estimated coefficients of the covariates in the marginal LBM by GEE, which is similar to the one in the log binomial model. Since the marginal LBM by GEE does not require the full specification of the joint distribution of the responses variable but rather only the first two moments [46, 47, 89], it does not have a likelihood function. In addition, the likelihood-based approaches (such as the maximum likelihood estimation and Akaike Information Criterion) are not available for model selection. To evaluate the model and improve the fitting in GEE, some model selection criteria were released previously [51-55]. We assessed the results from using the exact method under five criteria (RJ [51], QIC [52], CIC [53], SC [54] and GP [55]) in the model selection of the marginal LBM by GEE in a simulation. The model fitting under the RJ, QIC and CIC criteria is improved after the exact method is used to eliminate the impact of boundary vectors. A real-world example is used to explain the fitting process of the marginal LBM by GEE using the exact method. Results from the exact method were compared with those from alternative approaches under three criteria, RJ, QIC and CIC. The results suggest that the exact method outperformed alternative approaches in terms of model fitting by appropriately eliminating the impact of boundary vectors on the estimation of marginal LBM by GEE.

5.3 Contribution and significance of the work

The role of risk differences and ratios as fundamental measures of effect in summarizing the impact of exposure on a binary outcome has been evident in epidemiological studies since Greenland [5] pointed out in 1987. Because of numerical difficulties in estimation of risk,

epidemiologists and others lack the basic tools necessary to estimate this preferred measure of effect and have to resort to logistic regression estimating odds ratios that lack interpretability in terms of incidence proportions, or modified Poisson that provides biased estimates.

In this thesis, I have outlined the numerical difficulties that can arise in fitting a log binomial model and a marginal LBM by GEE to estimate relative risk and in fitting the identity-link binomial model to estimate the risk difference. The numerical difficulties arise due to inappropriate starting values in log binomial model and identity-link binomial model, and due to the ML solution lying on the boundary of the parameter space in the case of any of the three models. The difficulty due to inappropriate starting values can be resolved by replacing them with values that are appropriate (admissable) for a probability model, and I propose a starting value algorithm to do this for the log binomial model and the identity-link binomial model. To resolve numerical difficulties due to the ML solution lying the boundary of the parameter space, the exact method was proposed by Deddens et al. [18] for a log binomial model with a single covariate and extended to more general cases by Petersen and Deddens [9]. However, the method was incomplete, and the details necessary for its implementation were missing. In this thesis, I have provided the missing methodology and implementation details for the log binomial model. In addition, I have adapted it for use with the identity-link binomial model and the marginal LBM by GEE to overcome model-specific issues in those cases as well. The numerical difficulties due to the ML solution on the boundary of the parameter space are overcome by the exact method successfully.

Chapters 2, 3 and 4 of this thesis describe the details of implementing the exact method in these three models. The success of the implementation for each model is verified using

example datasets and by comparison of its performance against that of alternative methods in simulated data, and the practical utility of each model is demonstrated using real-world data. These contributions are reported in three manuscripts prepared for publication in peer-review journals. For the convenience of users wishing to implement the exact method in the fitting of the three models, I have released three packages in R on GitHub for peer evaluation and testing.

The comparison results in our real-world data analysis and simulation further demonstrate that, compared with alternative approaches, the implementation of the exact method not only successfully overcomes the numerical difficulties due to the boundary vector in the three models but also produces better estimates with the small sample biases.

Based on all contributions in this thesis led by Zhu, researchers are able to estimate the relative risk and the risk difference, with adjustment of potential confounders and interactions.

5.4 Limitations of the exact method

The exact method has two limitations. The first limitation is the time-consuming issue of the exact method. Petersen and Deddens [9] stated that identifying the boundary vectors in a model was an essential step before applying the exact method. However, it is impossible to meet this requirement. I solve this issue by obtaining a set of boundary vector candidates based on an approximate solution and by designing an ad-hoc strategy as outlined in Chapter 2, 3 and 4 of this thesis. This strategy is feasible but time-consuming because the fitting algorithm in the exact method needs to go through each boundary vector candidate and their combinations to obtain an optimum solution for model fitting. Table 5.1 shows the total number of times that

need to be checked to locate the final solution in a log binomial model with *n* boundary vector candidates by the fitting algorithm in the exact method. The table only shows the total number combinations of *n* less than seven boundary vector candidates, but *n* can be any non-zero positive integer. The processing time will become a serious issue when there are more than six boundary vector candidates. This may take longer in the identity-link binomial model because the parameter space has both lower and upper boundaries. The fitting algorithm needs to go through each boundary vector candidate and their combinations from either lower or upper bound and end by another.

Table 5.1 The *k*-combinations of *n* boundary vector candidates $\binom{n}{k}$.

	$n^{\dagger}$						
$k^{\ddagger}$	1	2	3	4	5	6	7
1	1	2	3	4	5	6	7
2		1	3	6	10	15	21
3			1	4	10	20	35
4				1	5	15	35
5					1	6	21
6						1	7
7							1
Total[§]	1	3	7	15	31	63	127

† *n* is the number of boundary vector candidates in the model.
‡ Subset *k* is chosen from n distinct boundary vector candidates.
§ The total number of times that need to be checked to locate the final solution in a log binomial model with *n* boundary vector candidates by the fitting algorithm in the exact method.
The table only shows the total number combinations of *n* less than seven boundary vector candidates, but *n* can be any non-zero positive integer.

The second limitation of the exact method only exists in the identity-link binomial model. The exact method can help the fitting algorithm to address the ML solution when it lies on either lower or upper bound through a re-parameterization process. However, the method is only able to obtain an approximation instead of the ML solution when it lies on both boundaries. This is because the re-parameterization process in the exact method is not able to

switch to another boundary once it begins the process from one boundary. It is possible to be resolved in future work.

5.5 Future topics

To better understand the exact method, two topics may merit further investigation.

1. The goodness of fit test under the exact method. The model fitting is improved by overcoming the numerical difficulties due to the ML solution on the boundary of the parameter space through the exact method. However, it is still critical to assess whether a fitted model adequately represents the data. Goodness-of-fit measures have been developed for the log binomial model [11]. It remains unknown whether and in what circumstances the goodness-of-fit test can be applied to the log binomial model and the identity-link binomial model with the exact method. There are three possible issues need to be addressed under this topic:

    (1) whether the performance of these tests (in respect of rejection rates, power to reject the null hypothesis when an incorrect model is specified, and the applicable degrees of freedom for tests that are extensions of the Hosmer-Lemeshow test) is altered when the ML solution lies on a boundary of the parameter space;

    (2) whether these tests can be applied to the identity-link model, and whether their performance is altered when the ML solution lies on a boundary of the parameter space;

(3) whether goodness-of-fit tests proposed for use with marginal logistic regression models of correlated binary responses can be extended to marginal log-link models of correlated binary responses, and whether their performance is altered when the ML solution lies on a boundary of the parameter space.

2. To use a regression model to estimate the risk difference in cluster/longitudinal data, a marginal identity-link binomial model estimated by GEE (marginal identity-link binomial GEE) is a viable option. However, the fitting algorithm may converge to an inappropriate solution, resulting in an estimated mean less than zero or greater than unity. The exact method would be expected to overcome the issue. The subjects of the study could be similar to Chapter 4 of the thesis. In addition, the performance of five model selection criteria (RJ, QIC, CIC, SC, and GP) in the marginal LBM by GEE with the exact method have been compared in Chapter 4. RJ, QIC and CIC express more significant improvements in model fitting after eliminating the boundary vectors by the exact method. It is also worth investigating whether these three criteria exhibit the same performance in a marginal identity-link binomial GEE.

Chapter 6 The documents of R packages

This chapter shows the R documents arising by this thesis. The format is consistent with the official manual of R. The packages corresponding to section 6.1, 6.2 and 6.3 can be downloaded from Github via the link under each section heading.

6.1 R-*lbm* package

Github link: https://github.com/zhuchao0228/lbm

Description

When the maximum likelihood (ML) solution lies on the boundary of the parameter space in the log binomial model, a special method is needed since the standard fitting algorithm may meet numerical difficulties. Exact method can overcome the difficulties and address the ML solution when it lies on the boundary of the parameter space. `lbm` implemented the exact method to address the ML solution in the log binomial model.

Usage

```
lbm(formula, data,contrasts = NULL,subset,na.action,lfv=0.95,
    vce = "oim",rescode=NULL,control=lbm.control(),...)
```

Arguments

`formula`	an object of class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted. The details of model specification are given under ‘Details’.
`data`	an optional data frame, list or environment (or object coercible by `as.data.frame` to a data frame) containing the variables in the model. If not found in `data`, the variables are taken from the `environment(formula)`, typically the environment from which `lbm` is called.
`contrasts`	an optional list. See the `contrasts.arg` of `model.matrix.default`.
`subset`	an optional vector specifying a subset of observations to be used in the fitting process.
`na.action`	a function which indicates what should happen when the data contain `NA`s. The default is set by the `na.action` setting of `na.action`, and is `na.fail` if that is unset. The default is `na.omit`. Another possible value is `NULL`, no action. Value `na.exclude` can be useful.
`lfv`	a testing range option which decides the range of boundary vector candidates included for testing. The default value is 0.95, which means the covariate vectors with the fitted probability greater than 0.95 will be incorporated in the boundary testing system as boundary vector candidates.

vce	the type of the information matrix used to attain the variance-covariance matrix. Two options could be selected, observed information matrix (OIM) and expected information matrix (EIM). The default vce is "OIM". This argument only works in the data with the boundary vector. If there is no boundary vector included in the data, the results are from glm. In the glm, the variance-covariance matrix is calculated by the expected information matrix.
rescode	is an option to code the response variable if it is a factor.
control	The control argument of lbm is by default passed to the arguments of lbm.control.

Details

A typical predictor has the form response ~ terms where the response is the (numeric) response vector, and terms is a series of terms which specifies a linear predictor for the response. A terms specification of the form first + second indicates all the terms in first together with all the terms in second with any duplicates removed. A specification of the form first:second indicates the set of terms obtained by taking the interactions of all terms in first with all terms in second. The specification first*second indicates the cross of first and second. This is the same as first + second + first:second. The terms in the formula will be re-ordered so that main effects come first, followed by the interactions, all second-order, all third-order and so on: to avoid this pass a terms object as the formula.

Value

lbm returns an object of class inheriting from "lbm" which inherits from the class "lbm". The function summary (i.e., summary.lbm) can be used to obtain or print a summary of the results table. The argument CF.lvl in summary represents the level of confidence interval claimed in the model. The default value is CF.lvl=0.95. Optionally, Risk ratio estimates and their related confidence interval are offered as an argument RR in summary. The default RR=FALSE is not to display them.

An object of class "lbm" is a list containing at least the following components:

coefficients	a named vector of coefficients
residuals	the working residuals, which are the residuals in the final iteration.
fitted.values	the fitted mean values, obtained by transforming the linear predictors by the inverse of the log link function.
linear.predictors	the linear fit on the log scale.
deviance	twice the absolute value of maximized log-likelihood.
aic	a version of Akaike Information Criterion (minus twice the maximized log-likelihood plus twice the number of parameters) computed by the aic component of the family. For the binomial model, the dispersion is fixed at one, and the number of parameters is the number of coefficients.
null.deviance	the deviance for the null model. The null model will only include an intercept if there is one in the model.
df.residual	the residual degrees of freedom.
df.null	the residual degrees of freedom for the null model.
response	the response vector used in the mode.l

`vcov`	the unscaled (`dispersion = 1`) estimated covariance matrix of the estimated coefficients.
`vce`	the type of information matrix applied.
`call`	the matched call.
`na.action`	(where relevant) information returned by `stats::model.frame` on the special handling of `NA`.
`contrasts`	(where relevant) the contrasts used.
`formula`	the formula supplied.
`factor`	the order of factors used in the response variable.
`bvector`	the matrix of boundary vectors.
`bv`	logical. Determines whether the model has boundary vectors.

References

Petersen, M. R. & Deddens, J. A. (2010). Maximum likelihood estimation of the log-binomial model. *Communications in Statistics - Theory and Methods*, 39: 5, 874 - 883.

See Also

`glm`, `lm`.

Examples

```
## Two examples are from Petersen, M. R. & Deddens, J. A. (2010).

## Example 1.
x<-c(1:10)
y<-c(0,0,0,0,1,0,1,1,1,1)
data<-data.frame(x,y)
a<-lbm(formula=y~x,data=data,vce="eim")

## Example 2.
x1<-c(1:11)
x2<-x1^2
y<-c(10,6,4,3,3,2,3,3,4,6,10)
dat<-cbind(x1,x2,y)
dat1<-apply(dat, 1, function(t) {
  temp<-data.frame(x1=rep(t[1],10),x2=rep(t[2],10),y=0)
  temp$y[1:t[3]]<-1
  return(temp)
})
data<-do.call(rbind, dat1)
a<-lbm(formula=y~x1+x2, data=data)
summary(a)
```

6.2 R-*bm* package

Github link: https://github.com/zhuchao0228/bm

Description

Fit a generalized linear model with binomial error and a link function by the exact method. The boundary of parameter space in log and identity link binomial model may meet numerical difficulties and fail to converge when the maximum likelihood solution lies either close to or on the boundary of parameter space. The exact method eliminates the boundary vector by the re-parameterization of the covariate and fits the model without the impact of the boundary vector.

Usage

```
bm(formula, data, contrasts = NULL, subset,na.action,
   link="identity", lfv.u=0.95, lfv.l=0.01, vce = "oim",
   rescode=NULL, control=bm.control(),...)
```

Arguments

formula	an object of class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted. The details of model specification are given under 'Details'.
data	an optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model. If not found in data, the variables are taken from environment(formula), typically the environment from which bm is called.
contrasts	an optional list. See the contrasts.arg of model.matrix.default.
subset	an optional vector specifying a subset of observations to be used in the fitting process.
na.action	a function which indicates what should happen when the data contain NAs. The default is set by the na.action setting of na.action, and is na.fail if that is unset. The 'factory-fresh' default is na.omit. Another possible value is NULL, no action. Value na.exclude can be useful.
link	a specification for the link function of the binomial model. The link function can be identity and log. The default link is identity.
lfv.u	a testing range of probability option which decides the range of boundary vector candidates included for testing. The default value is 0.95, which means the covariate vectors with the fitted probability greater than 0.95 will be included in the boundary testing system as boundary vector candidates.
lfv.l	a testing range of probability option which decides the range of boundary vector candidates included for testing. The default value is 0.01, which means the covariate vectors with the fitted probability less than 0.01 but greater than 0 will be included in the boundary testing system as boundary vector candidates. (This argument only works in the identity-link binomial model because the log binomial model does not have a lower bound of parameter space.)
vce	the type of the information matrix used to attain the variance-covariance matrix. Two options could be selected, observed information matrix (OIM) and expected information matrix (EIM). The default vce is "OIM". This argument only works in the data with a boundary vector. If there is no boundary vector included in the data, the results are directly from glm. In the glm, the variance-covariance matrix is calculated based on the expected information matrix.
rescode	is an option to code the response variable if it is a factor.
control	The control argument of bm is by default passed to the arguments of bm.control.

Details

A typical predictor has the form response ~ terms where the response is the (numeric) response vector, and terms is a series of terms which specifies a linear predictor for the response. A terms specification of the form first + second indicates all the terms in first together with all the terms in second with any duplicates removed. A specification of the form first:second indicates the set of

terms obtained by taking the interactions of all terms in first with all terms in second. The specification `first*second` indicates the cross of first and second. This is the same as `first + second + first:second`. The terms in the formula will be re-ordered so that main effects come first, followed by the interactions, all second-order, all third-order and so on: to avoid this pass a terms object as the formula.

Value

`bm` returns an object of `class` inheriting from `"bm"` which inherits from the class `"bm"`. The function `summary` (i.e.,`summary.bm`) can be used to obtain or print a summary of the results table. The argument `CF.lvl` in `summary` represents the level of confidence interval claimed in the model. The default value is `CF.lvl=0.95`. Optionally, Risk ratio estimates and their related confidence interval are offered as an argument `RR` in `summary` for log-link function only. The default `RR=FALSE` is not to display them.

An object of class `"bm"` is a list containing at least the following components:

`coefficients`	a named vector of coefficients
`residuals`	the working residuals, which are the residuals in the final iteration.
`fitted.values`	the fitted mean values, obtained by transforming the linear predictors by the inverse of the log link function.
`linear.predictors`	the linear fit on the log scale.
`deviance`	twice the absolute value of maximized log-likelihood.
`aic`	a version of Akaike Information Criterion (minus twice the maximized log-likelihood plus twice the number of parameters) computed by the `aic` component of the family. For the binomial model, the dispersion is fixed at one, and the number of parameters is the number of coefficients.
`null.deviance`	the deviance for the null model. The null model will only include an intercept if there is one in the model.
`df.residual`	the residual degrees of freedom.
`df.null`	the residual degrees of freedom for the null model.
`response`	the response vector used in the mode.l
`vcov`	the unscaled (`dispersion = 1`) estimated covariance matrix of the estimated coefficients.
`vce`	the type of standard error estimated based on the information matrix (observed or expected) applied.
`call`	the matched call.
`na.action`	(where relevant) information returned by `stats::model.frame` on the special handling of `NA`.
`contrasts`	(where relevant) the contrasts used.
`formula`	the formula supplied.
`factor`	the order of factors used in the response variable.
`bvector`	the matrix of boundary vectors.
`bv`	logical. Determines whether the model has boundary vectors.
`link`	link function applied in the model

`bound`	an indicator to describe the location of MLE on the boundary of parameter space in the identity-link binomial model only.

References

Petersen, M. R. & Deddens, J. A. (2010). Maximum likelihood estimation of the log-binomial model. *Communications in Statistics - Theory and Methods*, 39: 5, 874 - 883.

See Also

`glm`, `lm`, `lbm`.

Examples

```
## Two examples are from Petersen, M. R. & Deddens, J. A. (2010).

## Example 1.
x<-c(1:10)
y<-c(0,0,0,0,1,0,1,1,1,1)
data<-data.frame(x,y)
a<-bm(formula=y~x,data=data,link=log,vce=eim)

## Example 2.
x1<-c(1:11)
x2<-x1^2
y<-c(10,6,4,3,3,2,3,3,4,6,10)
dat<-cbind(x1,x2,y)
dat1<-apply(dat, 1, function(t) {
  temp<-data.frame(x1=rep(t[1],10),x2=rep(t[2],10),y=0)
  temp$y[1:t[3]]<-1
  return(temp)
})
data<-do.call(rbind, dat1)
a<-bm(formula=y~x1+x2,data=data)
summary(a)
a<-bm(formula=y~x1+x2,data=data,link=identity)
summary(a)
```

6.3 R-*lb.gee* package

Github link: https://github.com/zhuchao0228/lb.gee

Description

Fit a generalized estimating equation (GEE) with binomial error and a log link function by the exact method. When the fitted probabilities are either close or equal to 1 (defined as a boundary vector), the fitting algorithm of GEE may meet numerical instabilities and fail to converge or converge to an inappropriate solution. The exact method eliminates the boundary vector by the re-parameterization of the covariate and fits the model without the impact of the boundary vector.

Usage

```
lb.gee(formula = formula(data),contrasts = NULL,subset,na.action,
       data = parent.frame(),rescode=NULL,corstr,lfv=0.95,id,
       control=lb.gee.control(),criter="cic",ext=TRUE,ext.lvl=(1-1e-6),...)
```

Arguments

formula	an object of class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted. The details of model specification are given under ‘Details’.
contrasts	an optional list. See the contrasts.arg of model.matrix.default.
subset	an optional vector specifying a subset of observations to be used in the fitting process.
na.action	a function which indicates what should happen when the data contain NAs. The default is set by the na.action setting of na.action, and is na.fail if that is unset. The 'factory-fresh' default is na.omit. Another possible value is NULL, no action. Value na.exclude can be useful.
data	an optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model. If not found in data, the variables are taken from environment(formula), typically the environment from which bm is called.
rescode	is an option to code the response variable if it is a factor.
corstr	a character string specifying the correlation structure. The following correlation structure are permitted: "independence", "exchangeable", "ar1", and "unstructured".
lfv	a testing range of mean option which decides the range of boundary vector candidates included for testing. The default value is 0.95, which means the covariate vectors with the estimated mean greater than 0.95 will be incorporated in the boundary testing system as boundary vector candidates.
id	a vector which identifies the clusters. The length of id should be the same as the number of observations.
control	The control argument of lb.gee is by default passed to the arguments of lb.gee.control.
criter	a character string specifying the criterion used for the model selection. The following criteria are permitted: "RJ", "QIC", "CIC", "SC", and "GP".

Details

A typical predictor has the form response ~ terms where the response is the (numeric) response vector, and terms is a series of terms which specifies a linear predictor for the response. A terms specification of the form first + second indicates all the terms in first together with all the terms in second with any duplicates removed. A specification of the form first:second indicates the set of terms obtained by taking the interactions of all terms in first with all terms in second. The specification first*second indicates the cross of first and second. This is the same as first + second + first:second. The terms in the formula will be re-ordered so that main effects come first, followed by the interactions, all second-order, all third-order and so on: to avoid this pass a terms object as the formula.

Value

lb.gee returns an object of class which inherits from the class "lb.gee". The function summary (i.e., summary.lb.gee) can be used to obtain or print a summary of the results table. The argument CF.lvl in summary represents the level of confidence interval claimed in the model. The default value is CF.lvl=0.95. The estimates of relative risk and the corresponding confidence interval are offered as an argument RR in summary. The default RR=FALSE is not to display them.

An object of class "lb.gee" is a list containing at least the following components:

`beta`	a named vector of coefficients.
`alpha`	the estimated correlation parameters.
`gamma`	the scale parameter (dispersion parameter) of the distribution's variance.
`vbeta`	the robust variance-covariance matrix for the estimates of coefficients.
`vbeta.naiv`	the model-based variance-covariance matrix for the estimates of coefficients.
`criterion`	a return value of user-chosen criterion.
`response`	the response vector used in the model.
`call`	the matched call.
`na.action`	(where relevant) information returned by `stats::model.frame` on the special handling of `NA`.
`contrasts`	(where relevant) the contrasts used.
`formula`	the formula supplied.
`factor`	the order of factors used in the response variable.
`bvector`	the `data.frame` of boundary vectors.
`bv`	logical. Determines whether the model has boundary vectors.

References

Petersen, M. R. & Deddens, J. A. (2010). Maximum likelihood estimation of the log-binomial model. *Communications in Statistics - Theory and Methods*, 39: 5, 874 - 883.

Yan, J. & J.P. Fine (2004). Estimating Equations for Association Structures. *Statistics in Medicine*, 23, 859 - 880.

See Also

`geese`.

Bibliography

1. Nelder JA, Wedderburn RWM. Generalized linear models. *Journal of the Royal Statistical Society Series A (General)* 1972; 135: 370-384. DOI: 10.2307/2344614.
2. McCullagh P, Nelder JA. *Generalized linear models*. London : Chapman and Hall, 1989. 2nd ed., 1989.
3. Hardin JW, Hilbe J. *Generalized linear models and extensions*. Report no. 1597181056, 2012. StataCorp LP.
4. Noordzij M, van Diepen M, Caskey FC, Jager KJ. Relative risk versus absolute risk: one cannot be interpreted without the other. *Nephrology Dialysis Transplantation* 2017; 32: ii13-ii18. DOI: 10.1093/ndt/gfw465.
5. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology* 1987; 125: 761-768.
6. Lumley T, Kronmal R, Ma S. Relative risk regression in medical research: models, contrasts, estimators, and algorithms. *UW Biostatistics Working Paper Series* 2006; Working Paper 293.
7. Wedderburn RW. On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* 1976; 63: 27-32.
8. Marschner IC. Relative risk regression for binary outcomes: methods and recommendations. *Australian & New Zealand Journal of Statistics* 2015; 57: 437-462.
9. Petersen MR, Deddens JA. Maximum likelihood estimation of the log-binomial model. *Communications in Statistics-Theory and Methods* 2010; 39: 874-883.
10. Williamson T, Eliasziw M, Fick GH. Log-binomial models: exploring failed convergence. *Emerging Themes in Epidemiology* 2013; 10: 14.
11. Blizzard CL, Hosmer DW. Parameter estimation and goodness-of-fit in log binomial regression. *Biometrical Journal* 2006; 48: 5-22.
12. Wacholder S. Binomial regression in GLIM: Estimating risk ratios and risk differences. *American Journal of Epidemiology* 1986; 123: 174-184. DOI: 10.1093/oxfordjournals.aje.a114212.
13. StataCorp L. Stata 13 Base Reference Manual. *College Station, TX: Stata Press* 2013.
14. Zhang J, Yu KF. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA* 1998; 280: 1690-1691. DOI: 10.1001/jama.280.19.1690.
15. Breslow N. Covariance analysis of censored survival data. *Biometrics* 1974: 89-99.
16. Lee J, Chia K. Estimation of prevalence rate ratios for cross sectional data: an example in occupational epidemiology. *British Journal of Industrial Medicine* 1993; 50: 861.
17. Schouten EG, Dekker JM, Kok FJ, Cessie SL, Van Houwelingen HC, Pool J, Vandnbroucke JP. Risk ratio and rate ratio estimation in case‐cohort designs: Hypertension and cardiovascular mortality. *Statistics in Medicine* 1993; 12: 1733-1745.
18. Deddens JA, Petersen MR, Lei X. Estimation of prevalence ratios when PROC GENMOD does not converge. *Proceedings of the 28th annual SAS users group international conference*. 2003, p. 270-228.
19. McNutt LA, Wu C, Xue X, Hafner JP. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *American Journal of Epidemiology* 2003; 157: 940-943.
20. Zou G. A modified Poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology* 2004; 159: 702-706.
21. Carter RE, Lipsitz SR, Tilley BC. Quasi-likelihood estimation for relative risk regression models. *Biostatistics* 2005; 6: 39-44.

22. Marschner IC, Gillett AC. Relative risk regression: reliable and flexible methods for log-binomial models. *Biostatistics* 2012; 13: 179-192.
23. Lange K. *Numerical analysis for statisticians*. Springer Science & Business Media, 2010.
24. Andrade BB, Carabin H. On the estimation of relative risks via log binomial regression. *Revista Brasileira de Biometria* 2011; 29: 1-15.
25. Luo J, Zhang JJ, Sun H. Estimation of relative risk using a log-binomial model with constraints. *Computation Stat* 2014; 29: 981-1003. DOI: 10.1007/s00180-013-0476-8.
26. Andrade BB, Andrade JMdL. Some results for maximum likelihood estimation of adjusted relative risks. *Communications in Statistics - Theory and Methods* 2018; 47: 5750-5769. DOI: 10.1080/03610926.2017.1402045.
27. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine* 1988; 318: 1728-1733.
28. Bucher HC, Weinbacher M, Gyr K. Influence of method of reporting study results on decision of physicians to prescribe drugs to lower cholesterol concentration. *BMJ* 1994; 309: 761-764. DOI: 10.1136/bmj.309.6957.761.
29. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *BMJ* 1995; 310: 452-454. DOI: 10.1136/bmj.310.6977.452.
30. Sackett DL. Evidence-based medicine. *Seminars in Perinatology* 1997; 21: 3-5. DOI: 10.1016/S0146-0005(97)80013-4.
31. Furukawa TA. From effect size into number needed to treat. *The Lancet* 1999; 353: 1680.
32. Nuovo J, Melnikow J, Chang D. Reporting number needed to treat and absolute risk reduction in randomized controlled trials. *JAMA* 2002; 287: 2813-2814.
33. Akobeng A. Understanding measures of treatment effect in clinical trials. *Archives of Disease in Childhood* 2005; 90: 54-56.
34. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 1980; 48: 817-838.
35. Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* 1967, pp.221-233. University of California Press.
36. Spiegelman D, Hertzmark E. Easy SAS calculations for risk or prevalence ratios and differences. *American Journal of Epidemiology* 2005; 162: 199-200.
37. Zhu C, Blizzard L, Stankovich J, Wills K, Hosmer DW. Be wary of using Poisson regression to estimate risk and relative risk. *Biostatistics and Biometrics Open Access Journal* 2018; 4: 555649. DOI: 10.19080/BBOAJ.2018.04.555649.
38. Donoghoe MW, Marschner IC. Stable computational methods for additive binomial models with application to adjusted risk differences. *Computational Statistics & Data Analysis* 2014; 80: 184-196.
39. Cheung YB. A modified least-squares regression approach to the estimation of risk difference. *American Journal of Epidemiology* 2007; 166: 1337-1344. DOI: 10.1093/aje/kwm223.
40. Bieler G, Brown G, Williams R, Brogan D. Estimating Model-Adjusted Risks, Risk Differences, and Risk Ratios From Complex Survey Data. *American Journal of Epidemiology* 2010; 171: 618-623.
41. Graubard B, Korn E. Predictive margins with survey data. *Biometrics* 1999; 55: 652.
42. Marschner IC. Stable computation of maximum likelihood estimates in identity link Poisson regression. *Journal of Computational and Graphical Statistics* 2010; 19: 666-683.
43. Baker SG. The multinomial-Poisson transformation. *Journal of the Royal Statistical Society Series D (The Statistician)* 1994; 43: 495-504. DOI: 10.2307/2348134.

44. Kovalchik SA, Varadhan R, Fetterman B, Poitras NE, Wacholder S, Katki HA. A general binomial regression model to estimate standardized risk differences from binary response data. *Statistics in Medicine* 2013; 32: 808-821.
45. Richardson TS, Robins JM, Wang L. On modeling and estimation for the relative risk and risk difference. *Journal of the American Statistical Association* 2017; 112: 1121-1130.
46. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; 73: 13-22.
47. Zeger SL, Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986: 121-130.
48. McCullagh P. Quasi-likelihood functions. *The Annals of Statistics* 1983: 59-67.
49. Hardin JW, Hilbe JM. *Generalized Estimating Equations*. CRC Press, 2012.
50. Zeger SL, Liang K-Y, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988: 1049-1060.
51. Rotnitzky A, Jewell NP. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* 1990; 77: 485-485.
52. Pan W. Akaike's information criterion in generalized estimating equations. *Biometrics* 2001; 57: 120-125.
53. Hin L-Y, Wang Y-G. Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine* 2009; 28: 642-658.
54. Shults J, Chaganty NR. Analysis of serially correlated data using quasi-least squares. *Biometrics* 1998: 1622-1630.
55. Carey VJ, Wang Y-G. Working covariance model selection for generalized estimating equations. *Statistics in Medicine* 2011; 30: 3117-3124.
56. Yelland LN, Salter AB, Ryan P. Performance of the modified poisson regression approach for estimating relative risks from clustered prospective data. *American Journal of Epidemiology* 2011; 174: 984-992. DOI: 10.1093/aje/kwr183.
57. Deddens JA, Petersen MR. Re:"Estimating the relative risk in cohort studies and clinical trials of common outcomes". *American journal of epidemiology* 2004; 159: 213-214.
58. Marschner IC. glm2: Fitting Generalized Linear Models with Convergence Problems. *R J* 2011; 3: 12-15. DOI: 10.32614/RJ-2011-012.
59. Lipsitz SR, Fitzmaurice GM, Arriaga A, Sinha D, Gawande AA. Using the jackknife for estimation in log link Bernoulli regression models. *Statistics in medicine* 2015; 34: 444-453. 2014/11/12. DOI: 10.1002/sim.6348.
60. Wedderburn RW. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* 1974; 61: 439-447.
61. Fitzmaurice GM, Lipsitz SR, Arriaga A, Sinha D, Greenberg C, Gawande AA. Almost efficient estimation of relative risk regression. *Biostatistics* 2014; 15: 745-756. DOI: 10.1093/biostatistics/kxu012.
62. Hosmer DW, Lemeshow S, Sturdivant R. *Applied Logistic Regression*. 3 ed. Hoboken, NJ: John Wiley & Sons, 2013.
63. Aksoy S, Haralick RM. Feature normalization and likelihood-based similarity measures for image retrieval. *Pattern Recognition Letters* 2001; 22: 563-582. DOI: https://doi.org/10.1016/S0167-8655(00)00112-4.
64. Suarez-Alvarez MM, Pham D-T, Prostov MY, Prostov YI. Statistical approach to normalization of feature vectors and clustering of mixed datasets. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 2012; 468: 2630-2651.
65. Basheer IA, Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application. *Journal of microbiological methods* 2000; 43: 3-31.

66. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982: 963-974.
67. Stiratelli R, Laird N, Ware JH. Random-effects models for serial observations with binary response. *Biometrics* 1984: 961-971.
68. Yelland LN, Salter AB, Ryan P. Relative risk estimation in cluster randomized trials: a comparison of generalized estimating equation methods. *The International Journal of Biostatistics* 2011; 7: 1-26.
69. Pedroza C. Estimating relative risks in multicenter studies with a small number of centers—which methods to use? A simulation study. *Trials* 2017; 18: 512.
70. Davies HTO, Crombie IK, Tavakoli M. When can odds ratios mislead? *BMJ* 1998; 316: 989-991. DOI: 10.1136/bmj.316.7136.989.
71. Hin L-Y, Carey VJ, Wang Y-G. Criteria for working-correlation-structure selection in GEE: Assessment via simulation. *The American Statistician* 2007; 61: 360-364.
72. Shults J, Sun W, Tu X, Kim H, Amsterdam J, Hilbe JM, Ten-Have T. A comparison of several approaches for choosing between working correlation structures in generalized estimating equation analysis of longitudinal binary data. *Statistics in Medicine* 2009; 28: 2338-2355. DOI: 10.1002/sim.3622.
73. Wang M. Generalized estimating equations in longitudinal data analysis: a review and recent developments. *Advances in Statistics* 2014; 2014.
74. Leisch F, Weingessel A, Hornik K. On the generation of correlated articial binary data. *Working Paper Series, SFB "Adaptive Information Systems and Modelling in Economics and Management Science"*. Vienna University of Economics 1998.
75. Gunsolley J, Getchell C, Chinchilli V. Small sample characteristics of generalized estimating equations. *Communications in Statistics-simulation and Computation* 1995; 24: 869-878.
76. Wang M, Kong L, Li Z, Zhang L. Covariance estimators for generalized estimating equations (GEE) in longitudinal analysis with small samples. *Statistics in medicine* 2016; 35: 1706-1721.
77. McNeish DM, Harring JR. Clustered data with small sample sizes: Comparing the performance of model-based and design-based approaches. *Communications in Statistics - Simulation and Computation* 2017; 46: 855-869. DOI: 10.1080/03610918.2014.983648.
78. Cox DR, Hinkley DV. *Theoretical statistics*. CRC Press, 1979.
79. Feng Z, McLerran D, Grizzle J. A comparison of statistical methods for clustered data analysis with Gaussian error. *Statistics in medicine* 1996; 15: 1793-1806.
80. Westgate PM. A covariance correction that accounts for correlation estimation to improve finite-sample inference with generalized estimating equations: a study on its applicability with structured correlation matrices. *Journal of statistical computation and simulation* 2016; 86: 1891-1900.
81. Fan C, Zhang D, Zhang C-H. A comparison of bias-corrected covariance estimators for generalized estimating equations. *Journal of biopharmaceutical statistics* 2013; 23: 1172-1187.
82. Wang M, Long Q. Modified robust variance estimator for generalized estimating equations with improved small-sample performance. *Statistics in medicine* 2011; 30: 1278-1291.
83. Pan W. On the robust variance estimator in generalised estimating equations. *Biometrika* 2001; 88: 901-906.
84. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics* 2001; 57: 126-134.
85. Kauermann G, Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* 2001; 96: 1387-1396.

86. Gosho M, Sato Y, Takeuchi H. Robust covariance estimator for small-sample adjustment in the generalized estimating equations: A simulation study. *Science Journal of Applied Mathematics and Statistics* 2014; 2: 20-25.
87. Fay MP, Graubard BI. Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics* 2001; 57: 1198-1206.
88. Morel JG, Bokossa M, Neerchal NK. Small sample correction for the variance of GEE estimators. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 2003; 45: 395-409.
89. Ghisletta P, Spini D. An Introduction to Generalized Estimating Equations and an Application to Assess Selectivity Effects in a Longitudinal Study on Very Old Individuals. *Journal of Educational and Behavioral Statistics* 2004; 29: 421-437. DOI: 10.3102/10769986029004421.
90. Mackiw G. A note on the equality of the columr, and row rank of a matrix. *Mathematics Magazine* 1995; 68: 285.
91. Wardlaw WP. Row rank equals column rank. *Mathematics Magazine* 2005; 78: 316-318.

Appendix A

Supplementary materials of the log binomial model in the exact method

The exact method

The log binomial model

Consider *n* independent observations of a binary (0/1) outcome variable Y and J non-constant covariates $(X_1, X_2, ...X_J)$. The covariates can be any mix of continuous and categorical variables, and nonlinear combinations or transformations of them. Denote the observed data as $(y_i, \mathbf{x}_i), i = 1, 2, ...n$ where $\mathbf{x}_i' = (x_0, x_{i1}, x_{2i}, ...x_{iJ})$ with $x_0 = 1$. Under the log binomial model, the conditional probability of the outcome given the covariates is:

$$\Pr(Y_i = 1 \mid \mathbf{x}_i) = \mu(\mathbf{x}_i) = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ...\beta_J x_{iJ}), \ i = 1, 2, ...n \qquad \text{(A1)}$$

Suppose the sample data contains n_0 observations with $y = 0$ and $n - n_0$ observations with $y = 1$, and that the observations with $y = 0$ are the first n_0 observations and the observations with $y = 1$ are the last $n - n_0$ observations. The likelihood of the parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ...\beta_J)$ is:

$$L(\boldsymbol{\beta} \mid y_i, \mathbf{x}_i) = \prod_{i=1}^{n_0} [1 - \mu(\mathbf{x}_i)] \times \prod_{i=n_o+1}^{n} \mu(\mathbf{x}_i)$$

Substituting $\mu(\mathbf{x}_i) = \exp\left(\beta_0 + \sum_{j=1}^{J} \beta_j x_{ij}\right)$:

$$L(\boldsymbol{\beta} \mid y_i, \mathbf{x}_i) = \prod_{i=1}^{n_0} \left[1 - \exp\left(\beta_0 + \sum_{j=1}^{J} \beta_j x_{ij}\right)\right] \times \prod_{i=n_0+1}^{n} \left[\exp\left(\beta_0 + \sum_{j=1}^{J} \beta_j x_{ij}\right)\right]$$

Application of the exact method

Suppose that the ML solution lies on a boundary of the allowable parameter space with maximum value $\mu(\mathbf{x}_i) = 1$ for some $i \in 1, 2, ...n$, and that this maximum value is attained by $R \geq 1$ distinct sets of covariates (including the constant). We refer to these covariate vectors as boundary vectors. Denote the r^{th} boundary vector as $\mathbf{x}^{(r)} = \left(1, x_1^{(r)}, x_2^{(r)}, ...x_J^{(r)}\right)$. The boundary condition satisfied by this vector is:

$$\beta_0 + \sum_{j=1}^{J} \beta_j x_{ij}^{(r)} = 0, \ r = 1, 2, ...R \qquad \text{(A2)}$$

If the covariate values $\mathbf{x}^{(r)}$ of the r^{th} $(r = 1, 2, ...R)$ boundary vector are shared by n_r observations, the method proposed by Petersen and Deddens [9] for estimating the model is:

1. eliminating the constant by subtracting from the constant and each non-constant covariate its respective value in the boundary vector:
$$z_{ij}^{(1)} = x_{ij} - x_j^{(1)}, \ j = 0, 1, 2, ...J \qquad \text{(A3)}$$
2. when there are multiple boundary vectors $(R > 1)$, eliminating the first $R - 1$ non-constant covariates by re-parametrizing the covariates according to the scheme:
$$z_{ij}^{(r)} = z_{ij}^{(r-1)} - \left[\frac{t_j^{(r)}}{t_{r-1}^{(r)}}\right] z_{i,r-1}^{(r-1)} \ \text{ where } t_j^{(r)} = z_{ij}^{(r-1)}\Big|_{x_{ij} = x_j^{(r)}}, \quad r = 2, 3, ...R \qquad \text{(A4)}$$

3. dropping the observations with covariate values $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ...\mathbf{x}^{(R)}$ respectively, which make no contribution to the loglikelihood, and fitting the model $\mu\left(\mathbf{z}_i^{(R)}\right) = \exp\left(\sum_{j=R}^{J} \beta_j z_{ij}^{(R)}\right)$ without a constant and with $J-R+1$ covariates to the remaining $n-n_1-n_2-...-n_R$ observations to obtain the estimates $\hat{\beta}_R, \hat{\beta}_{R+1},,...\hat{\beta}_J$ of the coefficients of the non-eliminated non-constant covariates,
4. estimating the coefficients $\hat{\beta}_r$, $r=1,2,...R-1$ of the $R-1$ eliminated covariates from the boundary condition:

$$\hat{\beta}_r = -\frac{\sum_{j=r+1}^{J} \hat{\beta}_j t_j^{(r+1)}}{t_r^{(r+1)}} \qquad \text{(A5)}$$

5. estimating the standard errors of the estimated coefficients of the eliminated covariates as:

$$\widehat{\mathrm{SE}}\left(\hat{\beta}_r\right) = \sqrt{\sum_{j=r+1}^{J}\left[\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\left(\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\right)^2\right] + \sum_{j_1=r+1}^{J}\sum_{\substack{j_2=r+1 \\ j_2\neq j_1}}^{J}\left\{\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)\left[\frac{t_{j_1}^{(r+1)} t_{j_2}^{(r+1)}}{\left(t_r^{(r+1)}\right)^2}\right]\right\}} \qquad \text{(A6)}$$

for $r=1,2,...R-1$ where $\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)$ denotes the estimated variance of the estimated coefficient $\hat{\beta}_j$, and $\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)$ denotes the estimated covariance between the estimated coefficients $\hat{\beta}_{j_1}$ and $\hat{\beta}_{j_2}$;

6. estimating the coefficient of the constant covariate from the boundary condition (A2) as:

$$\hat{\beta}_0 = -\sum_{j=1}^{J} \hat{\beta}_j x_j^{(1)} \qquad \text{(A7)}$$

7. estimating the standard error of the estimated coefficient of the constant covariate as:

$$\widehat{\mathrm{SE}}\left(\hat{\beta}_0\right) = \sqrt{\sum_{j=1}^{J}\left[\left(x_j^{(1)}\right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\right] + \sum_{j_1=1}^{J}\sum_{\substack{j_2=1 \\ j_2\neq j_1}}^{J}\left[\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right) x_{j_1}^{(1)} x_{j_2}^{(1)}\right]} \qquad \text{(A8)}$$

Covariance estimates not provided by the procedure outlined by Petersen and Deddens [9]

If $R=1$, all covariances between the estimated coefficients of non-constant covariates are estimated in the regression. When $R>1$, however, fitting the model $\mu\left(\mathbf{z}_i^{(R)}\right) = \exp\left(\sum_{j=R}^{J} \beta_j z_{ij}^{(R)}\right)$ without a constant and with $J-R+1$ non-constant covariates does not provide estimates of the covariances between the estimated coefficients of the $R-1$ eliminated covariates and the estimated coefficients of the $J-R+1$ remaining non-constant covariates. Those covariances are required in (A6) and (A8). To overcome this deficiency, we provide Theorem A1.

Even when $R=1$, the procedure outlined by Petersen and Deddens [9] does not provide estimates of the covariances between the estimated coefficients of the constant covariate and the estimated coefficients of the non-constant covariates. To overcome that deficiency, we provide Theorem A2.

Theorem A1

For a log binomial model with $J \geq 1$ non-constant covariates fitted by the exact method to data having R $(1 \leq R \leq J)$ distinct sets of values of the covariates (including the constant) for which the estimated outcome probability evaluated at the ML solution is unity, the estimated covariances between exact estimates of the coefficients of the $R-1$ eliminated non-constant covariates and exact estimates of the coefficients of the $J-R+1$ remaining non-constant covariates are given by:

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_r,\hat{\beta}_s\right) = -\left\{\frac{t_s^{(r+1)}}{t_r^{(r+1)}}\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right) + \sum_{\substack{j=r+1 \\ j\neq s}}^{J}\left[\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_s\right)\right]\right\} \qquad \text{(A9)}$$

for $r = 1, 2, ...R-1; s = r+1, ...J$.

Proof of Theorem A1

In successive re-parameterizations, the boundary conditions require $\sum_{j=0}^{J} \hat{\beta}_j z_{ij}^{(1)} = 0,\ \sum_{j=1}^{J} \hat{\beta}_j t_j^{(2)} = 0,$ $\sum_{j=2}^{J} \hat{\beta}_j t_j^{(3)} = 0$ and, in general for $R > 1$:

$$\sum_{j=r}^{J} \hat{\beta}_j t_j^{(r+1)} = 0,\ r = 1, 2, ...R-1 \tag{A10}$$

for all observations that share those covariate values. Hence:

$$t_r^{(r+1)} \hat{\beta}_r = -\sum_{j=r+1}^{J} \hat{\beta}_j t_j^{(r+1)},\ r = 1, 2, ...R-1 \tag{A11}$$

Adding to each side of (A11) the estimated coefficient of one of the $J - R + 1$ covariates remaining in the model:

$$t_r^{(r+1)} \hat{\beta}_r + \hat{\beta}_s = -\left[\left(t_s^{(r+1)} - 1 \right) \hat{\beta}_s + \sum_{\substack{j=r+1 \\ j \neq s}}^{J} \hat{\beta}_j t_j^{(r+1)} \right],\ r = 1, 2, ..., R-1;\ s = r+1, ...J$$

Thus:

$$\begin{aligned} &\widehat{\mathrm{Var}}\left(t_r^{(r+1)} \hat{\beta}_r + \hat{\beta}_s \right) \\ &\quad = \widehat{\mathrm{Var}} \left\{ -\left[\left(t_s^{(r+1)} - 1 \right) \hat{\beta}_s + \sum_{\substack{j=r+1 \\ j \neq s}}^{J} \hat{\beta}_j t_j^{(r+1)} \right] \right\},\ r = 1, 2, ..., R-1;\ s = r+1, ...J \end{aligned} \tag{A12}$$

Expanding the left-hand side of (A12):

$$\widehat{\mathrm{Var}}\left(t_r^{(r+1)} \hat{\beta}_r + \hat{\beta}_s \right) = \left(t_r^{(r+1)} \right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_r \right) + 2 t_r^{(r+1)} \widehat{\mathrm{Cov}}\left(\hat{\beta}_r, \hat{\beta}_s \right) + \widehat{\mathrm{Var}}\left(\hat{\beta}_s \right) \tag{A13}$$

Expanding the right-hand side of (A12):

$$\begin{aligned}
&\widehat{\mathrm{Var}}\left\{-\left[\left(t_s^{(r+1)}-1\right)\hat{\beta}_s+\sum_{\substack{j=r+1\\ j\neq s}}^{J}\hat{\beta}_j t_j^{(r+1)}\right]\right\}\\
&=\left(t_s^{(r+1)}-1\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right)+2\sum_{\substack{j=r+1\\ j\neq s}}^{J}\left[t_j^{(r+1)}\left(t_s^{(r+1)}-1\right)\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_s\right)\right]+\widehat{\mathrm{Var}}\left[\sum_{\substack{j=r+1\\ j\neq s}}^{J}\hat{\beta}_j t_j^{(r+1)}\right]\\
&=\left[\left(t_s^{(r+1)}\right)^2-2t_s^{(r+1)}+1\right]\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right)+2\sum_{\substack{j=r+1\\ j\neq s}}^{J}\left[\left(t_j^{(r+1)}t_s^{(r+1)}-t_j^{(r+1)}\right)\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_s\right)\right]\\
&+\sum_{\substack{j=r+1\\ j\neq s}}^{J}\left[\left(t_j^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\right]+\sum_{j_1=r+1}^{J}\sum_{\substack{j_2=r+1\\ j_1,j_2\neq s\\ j_2\neq j_1}}^{J}\left[\left(t_{j_1}^{(r+1)}t_{j_2}^{(r+1)}\right)\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)\right]\\
&=\left(1-2t_s^{(r+1)}\right)\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right)-2\sum_{\substack{j=r+1\\ j\neq s}}^{J}\left[t_j^{(r+1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_s\right)\right]\\
&+\sum_{j=r+1}^{J}\left[\left(t_j^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\right]+\sum_{j_1=r+1}^{J}\sum_{\substack{j_2=r+1\\ j_2\neq j_1}}^{J}\left[\left(t_{j_1}^{(r+1)}t_{j_2}^{(r+1)}\right)\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)\right]
\end{aligned}\qquad \text{(A14)}$$

From (A11), $\widehat{\mathrm{Var}}\left[\hat{\beta}_r t_r^{(r+1)}\right]=\widehat{\mathrm{Var}}\left[-\sum_{j=r+1}^{J}\hat{\beta}_j t_j^{(r+1)}\right]$, $r=1,2,\ldots R-1$ and hence:

$$\left(t_r^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_r\right)=\sum_{j=r+1}^{J}\left[\left(t_j^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\right]+\sum_{j_1=r+1}^{J}\sum_{\substack{j_2=r+1\\ j_2\neq j_1}}^{J}\left[\left(t_{j_1}^{(r+1)}t_{j_2}^{(r+1)}\right)\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)\right]\qquad \text{(A15)}$$

Substituting for $\left(t_r^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_r\right)$ in (A14):

$$\begin{aligned}
&\widehat{\mathrm{Var}}\left\{-\left[\left(t_s^{(r+1)}-1\right)\hat{\beta}_s+\sum_{\substack{j=r+1\\ j\neq s}}^{J}\hat{\beta}_j t_j^{(r+1)}\right]\right\}\\
&\quad=\left(1-2t_s^{(r+1)}\right)\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right)+\left(t_r^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_r\right)-2\sum_{\substack{j=r+1\\ j\neq s}}^{J}t_j^{(r+1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_s\right)
\end{aligned}\qquad \text{(A16)}$$

Equating (A13) and (A16), the two sides of (A12), and simplifying:

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_r,\hat{\beta}_s\right)=-\left[\frac{t_s^{(r+1)}}{t_r^{(r+1)}}\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right)+\sum_{\substack{j=r+1\\ j\neq s}}^{J}\left\{\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_s\right)\right\}\right],\ r=1,2,\ldots R-1;\ s=r+1,\ldots J$$

<u>*Theorem* A2</u>

For a log binomial model with $J\geq 1$ non-constant covariates fitted by the exact method to data having R $(1\leq R\leq J)$ distinct sets of values of the covariates (including the constant) for which the estimated outcome probability evaluated at the ML solution is unity, the estimated covariances between the exact estimate of the coefficient of the constant covariate and exact estimates of the coefficients of the J non-constant covariates are given by:

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_0,\hat{\beta}_j\right)=-\left[x_j^{(1)}\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}x_{j_1}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_{j_1}\right)\right],\ j=1,2,\ldots J\qquad \text{(A17)}$$

Proof of Theorem A2

Adding the estimated coefficient $\hat{\beta}_j,\ j=1,2,\ldots J$ of any of the non-constant covariates to each side of the boundary condition (A7) yields:

$$\hat{\beta}_0+\hat{\beta}_j=-\left[\left(x_j^{(1)}-1\right)\hat{\beta}_j+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}\hat{\beta}_{j_1}x_{j_1}^{(1)}\right],\ \ j=1,2,...J$$

It follows that:

$$\widehat{\mathrm{Var}}\left(\hat{\beta}_0+\hat{\beta}_j\right)=\widehat{\mathrm{Var}}\left\{-\left[\left(x_j^{(1)}-1\right)\hat{\beta}_j+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}\hat{\beta}_{j_1}x_{j_1}^{(1)}\right]\right\} \tag{A18}$$

Expanding the left-hand side of (A18):

$$\widehat{\mathrm{Var}}\left(\hat{\beta}_0+\hat{\beta}_j\right)=\widehat{\mathrm{Var}}\left(\hat{\beta}_0\right)+2\widehat{\mathrm{Cov}}\left(\hat{\beta}_0,\hat{\beta}_j\right)+\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) \tag{A19}$$

Expanding the right-hand side of (A18):

$$\begin{aligned}&\widehat{\mathrm{Var}}\left\{-\left[\left(x_j^{(1)}-1\right)\hat{\beta}_j+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}\hat{\beta}_{j_1}x_{j_1}^{(1)}\right]\right\}\\ &\quad=\left(x_j^{(1)}-1\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)+2\left(x_j^{(1)}-1\right)\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}x_{j_1}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_{j_1}\right)+\widehat{\mathrm{Var}}\left(\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}\hat{\beta}_{j_1}x_{j_1}^{(1)}\right)\end{aligned}$$

and gathering like terms:

$$\begin{aligned}&\widehat{\mathrm{Var}}\left\{-\left[\left(x_j^{(1)}-1\right)\hat{\beta}_j+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}\hat{\beta}_{j_1}x_{j_1}^{(1)}\right]\right\}\\ &\quad=\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)-2\left[x_j^{(1)}\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}x_{j_1}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_{j_1}\right)\right]+\left(x_j^{(1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\\ &\quad+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}\left[\left(x_{j_1}^{(1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_{j_1}\right)\right]+2\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}x_j^{(1)}x_{j_1}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_{j_1}\right)\\ &\quad+\sum_{\substack{j_2=1\\ j_2\neq j}}^{J}\sum_{\substack{j_3=1\\ j_3\neq j\\ j_3\neq j_2}}^{J}\left[x_{j_2}^{(1)}x_{j_3}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_2},\hat{\beta}_{j_3}\right)\right]\end{aligned} \tag{A20}$$

From (A7), $\hat{\beta}_0=-\sum_{j=1}^{J}\hat{\beta}_j x_j^{(1)}$ and it follows that $\widehat{\mathrm{Var}}\left(\hat{\beta}_0\right)=\widehat{\mathrm{Var}}\left[-\sum_{j=1}^{J}\hat{\beta}_j x_j^{(1)}\right]$. Thus:

$$\widehat{\mathrm{Var}}\left(\hat{\beta}_0\right)=\sum_{j=1}^{J}\left[\left(x_j^{(1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\right]+\sum_{j_1=1}^{J}\sum_{\substack{j_2=1\\ j_1\neq j_2}}^{J}\left[x_{j_1}^{(1)}x_{j_2}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)\right]$$

Substituting for $\widehat{\mathrm{Var}}\left(\hat{\beta}_0\right)$ in (A20):

$$\begin{aligned}&\widehat{\mathrm{Var}}\left\{-\left[\left(x_j^{(1)}-1\right)\hat{\beta}_j+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}\hat{\beta}_{j_1}x_{j_1}^{(1)}\right]\right\}\\ &\quad=\widehat{\mathrm{Var}}\left(\hat{\beta}_0\right)+\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)-2\left[x_j^{(1)}\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}x_{j_1}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_{j_1}\right)\right]\end{aligned} \tag{A21}$$

Substituting (A19) and (A21) into (A18), and simplifying:

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_0,\hat{\beta}_j\right)=-\left[x_j^{(1)}\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}x_{j_1}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_{j_1}\right)\right],\quad j=1,2,\ldots J\ .$$

Theorem A3

For a log binomial model with $J\geq 1$ non-constant covariates fitted by the exact method to data having R $(R=J)$ distinct sets of values of the covariates (including the constant) for which the estimated outcome probability evaluated at the ML solution is unity, there is a perfect correlation between each pair of exact estimates of the coefficients of the non-constant covariates:

$$\left|\rho_{j_1,j_2}\right|=\frac{\left|\widehat{\mathrm{Cov}}(\hat{\beta}_{j_1},\hat{\beta}_{j_2})\right|}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_{j_1})\widehat{\mathrm{Var}}(\hat{\beta}_{j_2})}}=\frac{\left|\widehat{\mathrm{Cov}}(\hat{\beta}_{j_1},\hat{\beta}_{j_2})\right|}{\widehat{\mathrm{SE}}(\hat{\beta}_{j_1})\widehat{\mathrm{SE}}(\hat{\beta}_{j_2})}=1\ ,\quad j_1,j_2=1,2,\ldots J;\ j_1\neq j_2\ .$$

Proof of Theorem A3

From the boundary condition (A11) for boundary vectors evaluated for $r=R-1$:

$$\hat{\beta}_{R-1}t_{R-1}^{(R)}=-\hat{\beta}_R t_R^{(R)}$$

Because $J=R$, that expression can be written as:

$$\hat{\beta}_{J-1}t_{J-1}^{(R)}=-\hat{\beta}_J t_J^{(R)}$$

Hence the exact estimate of the coefficient of the $(J-1)^{st}$ covariate can be obtained as:

$$\hat{\beta}_{J-1}=-\frac{t_J^{(R)}}{t_{J-1}^{(R)}}\hat{\beta}_J \qquad \text{(A22)}$$

With the exact estimates of the estimated standard errors denoted for brevity as $\hat{s}_j=\widehat{\mathrm{SE}}\left(\hat{\beta}_j\right),\ j=0,1,2\ldots J$, equation (A15) can be expressed as:

$$\hat{s}_r=\sqrt{\sum_{j=r+1}^{J}\left[\left(\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\right)^2\left(\hat{s}_j\right)^2\right]+\sum_{j_1=r+1}^{J}\sum_{\substack{j_2=r+1\\ j_2\neq j_1}}^{J}\left[\frac{t_{j_1}^{(r+1)}t_{j_2}^{(r+1)}}{\left(t_r^{(r+1)}\right)^2}\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)\right]},\ r=1,2,\ldots R-1 \qquad \text{(A23)}$$

For models with $J=R$, only one covariate is estimated and there are no covariances with higher-numbered covariates. Hence (A23) evaluated for the $(J-1)^{st}$ covariate yields:

$$\hat{s}_{J-1}=\sqrt{\left(\frac{t_J^{(R)}}{t_{J-1}^{(R)}}\right)^2\left(\hat{s}_J\right)^2}=\left|\frac{t_J^{(R)}}{t_{J-1}^{(R)}}\right|\hat{s}_J \qquad \text{(A24)}$$

and (A9) evaluated for the $(J-1)^{st}$ covariate yields (Note that *J* is the last covariate. Therefore, the second part of (A9) eliminated):

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_{J-1},\hat{\beta}_J\right)=-\frac{t_J^{(R)}}{t_{J-1}^{(R)}}\widehat{\mathrm{Var}}\left(\hat{\beta}_J\right)=-\frac{t_J^{(R)}}{t_{J-1}^{(R)}}\left(\hat{s}_J\right)^2 \qquad \text{(A25)}$$

By changing the order of the covariates, and thereby varying which covariate is estimated, estimates can be obtained for all $J-1$ excluded covariates as follows:

$$\hat{\beta}_j = -\frac{t_J^{(R)}}{t_j^{(R)}}\hat{\beta}_J,\ \hat{s}_j = \left|\frac{t_J^{(R)}}{t_j^{(R)}}\right|\hat{s}_J \text{ and } \widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_J\right) = -\frac{t_J^{(R)}}{t_j^{(R)}}\left(\hat{s}_J\right)^2,\ \ j=1,2,...J-1 \qquad \text{(A26)}$$

From (A25):

$$\left|\widehat{\mathrm{Cov}}\left(\hat{\beta}_{J-1},\hat{\beta}_J\right)\right| = \left|\frac{t_J^{(R)}}{t_{J-1}^{(R)}}\right|\hat{s}_J\hat{s}_J \qquad \text{(A27)}$$

Substituting for $\hat{s}_{J-1} = \left|\frac{t_J^{(R)}}{t_{J-1}^{(R)}}\right|\hat{s}_J$ from (A27):

$$\left|\widehat{\mathrm{Cov}}(\hat{\beta}_{J-1},\hat{\beta}_J)\right| = \hat{s}_{J-1}\hat{s}_J$$

Hence:

$$\left|\rho_{J-1,J}\right| = \frac{\left|\widehat{\mathrm{Cov}}(\hat{\beta}_{J-1},\hat{\beta}_J)\right|}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_{J-1})\widehat{\mathrm{Var}}(\hat{\beta}_J)}} = \frac{\hat{s}_{J-1}\hat{s}_J}{\sqrt{\left(\hat{s}_{J-1}\right)^2\left(\hat{s}_J\right)^2}} = 1\,.$$

This relationship holds for any pair of coefficients. In general:

$$\left|\widehat{\mathrm{Cov}}(\hat{\beta}_{j_1},\hat{\beta}_{j_2})\right| = \hat{s}_{j_1}\hat{s}_{j_2},\quad j_1,j_2=1,2,...J;\ j_1\neq j_2 \qquad \text{(A28)}$$

and:

$$\left|\rho_{j_1,j_2}\right| = \frac{\left|\widehat{\mathrm{Cov}}(\hat{\beta}_{j_1},\hat{\beta}_{j_2})\right|}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_{j_1})\widehat{\mathrm{Var}}(\hat{\beta}_{j_2})}} = \frac{\hat{s}_{j_1}\hat{s}_{j_2}}{\sqrt{\left(\hat{s}_{j_1}\right)^2\left(\hat{s}_{j_2}\right)^2}} = 1,\quad j_1,j_2=1,2,...J;\ j_1\neq j_2\,.$$

<u>*Corollary* A3.1</u>

If the log binomial model satisfies the conditions of Theorem 3, the standardized values of the exact estimates of the $J+1$ model coefficients are equal in absolute size:

$$\frac{\left|\hat{\beta}_0\right|}{\widehat{\mathrm{SE}}(\hat{\beta}_0)} = \frac{\left|\hat{\beta}_1\right|}{\widehat{\mathrm{SE}}(\hat{\beta}_1)} = \frac{\left|\hat{\beta}_2\right|}{\widehat{\mathrm{SE}}(\hat{\beta}_2)} = ... = \frac{\left|\hat{\beta}_J\right|}{\widehat{\mathrm{SE}}(\hat{\beta}_J)} \qquad \text{(A29)}$$

<u>Proof of Corollary A3.1</u>

From (A26), $\hat{\beta}_j = -\frac{t_J^{(R)}}{t_j^{(R)}}\hat{\beta}_J$ and $\hat{s}_j = \left|\frac{t_J^{(R)}}{t_j^{(R)}}\right|\hat{s}_J,\ \ j=1,2,...J-1$ yielding:

$$\frac{\left|\hat{\beta}_j\right|}{\hat{s}_j} = \frac{\left|\frac{t_J^{(R)}}{t_j^{(R)}}\hat{\beta}_J\right|}{\left|\frac{t_J^{(R)}}{t_j^{(R)}}\right|\hat{s}_J} = \frac{\left|\frac{t_J^{(R)}}{t_j^{(R)}}\right|\left|\hat{\beta}_J\right|}{\left|\frac{t_J^{(R)}}{t_j^{(R)}}\right|\hat{s}_J} = \frac{\left|\hat{\beta}_J\right|}{\hat{s}_J},\quad j=1,2,...J-1 \qquad \text{(A30)}$$

Substituting $\hat{\beta}_j = -\frac{t_J^{(R)}}{t_j^{(R)}}\hat{\beta}_J$ from (A11) in the boundary condition (A7) for the first boundary vector gives:

$$\begin{aligned}\hat{\beta}_0 &= -\left(\hat{\beta}_1 x_1^{(1)} + \hat{\beta}_2 x_2^{(1)} + \cdots + \hat{\beta}_J x_J^{(1)}\right) \\ &= -\left(\frac{t_J^{(R)}}{t_1^{(R)}}\hat{\beta}_J x_1^{(1)} + \frac{t_J^{(R)}}{t_2^{(R)}}\hat{\beta}_J x_2^{(1)} + \cdots + \frac{t_J^{(R)}}{t_{J-1}^{(R)}}\hat{\beta}_J x_{J-1}^{(1)} + \frac{t_J^{(R)}}{t_J^{(R)}}\hat{\beta}_J x_J^{(1)}\right) \\ &= -\hat{\beta}_J\left(\frac{t_J^{(R)}}{t_1^{(R)}} x_1^{(1)} + \frac{t_J^{(R)}}{t_2^{(R)}} x_2^{(1)} + \cdots + \frac{t_J^{(R)}}{t_{J-1}^{(R)}} x_{J-1}^{(1)} + x_J^{(1)}\right)\end{aligned}$$

Equation (A8) for the standard error of the constant covariate can be factorised as follows:

$$\begin{aligned}\hat{s}_0 &= \sqrt{\sum_{j=1}^{J}\left[\left(x_j^{(1)}\right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\right] + \sum_{j_1=1}^{J}\sum_{\substack{j_2=1 \\ j_2\neq j_1}}^{J}\left[\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)x_{j_1}^{(1)}x_{j_2}^{(1)}\right]} \\ &= \sqrt{\widehat{\mathrm{Var}}\left(\hat{\beta}_1 x_1^{(1)} + \hat{\beta}_2 x_2^{(1)} + \cdots + \hat{\beta}_J x_J^{(1)}\right)}\end{aligned}$$

Substituting for $\hat{\beta}_j = -\frac{t_J^{(R)}}{t_j^{(R)}}\hat{\beta}_J$ from (A26):

$$\begin{aligned}\hat{s}_0 &= \sqrt{\widehat{\mathrm{Var}}\left(\frac{t_J^{(R)}}{t_1^{(R)}}\hat{\beta}_J x_1^{(1)} + \frac{t_J^{(R)}}{t_2^{(R)}}\hat{\beta}_J x_2^{(1)} + \cdots + \frac{t_J^{(R)}}{t_{J-1}^{(R)}}\hat{\beta}_J x_{J-1}^{(1)} + \frac{t_J^{(R)}}{t_J^{(R)}}\hat{\beta}_J x_J^{(1)}\right)} \\ &= \sqrt{\widehat{\mathrm{Var}}\left[\left(\frac{t_J^{(R)}}{t_1^{(R)}} x_1^{(1)} + \frac{t_J^{(R)}}{t_2^{(R)}} x_2^{(1)} + \cdots + \frac{t_J^{(R)}}{t_{J-1}^{(R)}} x_{J-1}^{(1)} + x_J^{(1)}\right)\hat{\beta}_J\right]} \\ &= \sqrt{\left(\frac{t_J^{(R)}}{t_1^{(R)}} x_1^{(1)} + \frac{t_J^{(R)}}{t_2^{(R)}} x_2^{(1)} + \cdots + \frac{t_J^{(R)}}{t_{J-1}^{(R)}} x_{J-1}^{(1)} + x_J^{(1)}\right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_J\right)} \\ &= \left|\frac{t_J^{(R)}}{t_1^{(R)}} x_1^{(1)} + \frac{t_J^{(R)}}{t_2^{(R)}} x_2^{(1)} + \cdots + \frac{t_J^{(R)}}{t_{J-1}^{(R)}} x_{J-1}^{(1)} + x_J^{(1)}\right| \hat{s}_J\end{aligned}$$

Therefore:

$$\frac{\left|\hat{\beta}_0\right|}{\hat{s}_0} = \frac{\left|-\hat{\beta}_J\left(\frac{t_J^{(R)}}{t_1^{(R)}} x_1^{(1)} + \frac{t_J^{(R)}}{t_2^{(R)}} x_2^{(1)} + \cdots + \frac{t_J^{(R)}}{t_{J-1}^{(R)}} x_{J-1}^{(1)} + x_J^{(1)}\right)\right|}{\hat{s}_J\left|\frac{t_J^{(R)}}{t_1^{(R)}} x_1^{(1)} + \frac{t_J^{(R)}}{t_2^{(R)}} x_2^{(1)} + \cdots + \frac{t_J^{(R)}}{t_{J-1}^{(R)}} x_{J-1}^{(1)} + x_J^{(1)}\right|} = \frac{\left|\hat{\beta}_J\right|}{\hat{s}_J} \qquad \text{(A31)}$$

Combining (A30) and (A31), we obtain:

$$\frac{\left|\hat{\beta}_0\right|}{\widehat{\mathrm{SE}}(\hat{\beta}_0)} = \frac{\left|\hat{\beta}_1\right|}{\widehat{\mathrm{SE}}(\hat{\beta}_1)} = \frac{\left|\hat{\beta}_2\right|}{\widehat{\mathrm{SE}}(\hat{\beta}_2)} = \ldots = \frac{\left|\hat{\beta}_J\right|}{\widehat{\mathrm{SE}}(\hat{\beta}_J)}.$$

Theorem A4

For a log binomial model with $J \geq 1$ non-constant covariates fitted by the exact method to data having R $(1 \leq R \leq J)$ distinct sets of values of the covariates (including the constant) for which the estimated outcome

probability evaluated at the ML solution is unity, these R distinct sets of values of the covariates are always linearly independent.

Proof of Theorem A4

Denote by $\mathbf{B}$ the $R\times(J+1)$ matrix with R rows and $J+1$ columns formed from the R distinct sets of covariate values (including the constant) for which the estimated outcome probability evaluated ML solution is unity, and by $\mathbf{x}^{(r)}=\left(1,x_1^{(r)},x_2^{(r)},....,x_{J-1}^{(r)},x_J^{(r)}\right)$ the set of covariate values (including the constant) in the r^{th} $(r=1,2,...R)$ such set. Denote the ML solution evaluated by the exact method if it exists as $\hat{\boldsymbol{\beta}}=\left(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2....\hat{\beta}_{J-1},\hat{\beta}_J\right)'$ with all elements non-zero. We refer to $\mathbf{x}^{(r)}$ as a boundary vector, and to $\mathbf{B}=\left(\mathbf{x}^{(1)},\mathbf{x}^{(2)},...,\mathbf{x}^{(R-1)},\mathbf{x}^{(R)}\right)'$ as the boundary matrix. The boundary matrix is:

$$\mathbf{B}=\begin{bmatrix}1 & x_1^{(1)} & x_2^{(1)} & . & . & x_{J-1}^{(1)} & x_J^{(1)}\\ 1 & x_1^{(2)} & x_2^{(2)} & . & . & x_{J-1}^{(2)} & x_J^{(2)}\\ . & . & . & . & . & . & .\\ . & . & . & . & . & . & .\\ 1 & x_1^{(R-1)} & x_2^{(R-1)} & . & . & x_{J-1}^{(R-1)} & x_J^{(R-1)}\\ 1 & x_1^{(R)} & x_2^{(R)} & . & . & x_{J-1}^{(R)} & x_J^{(R)}\end{bmatrix} \tag{A32}$$

Note that at the ML solution, if it exists, $\mathbf{B}\hat{\boldsymbol{\beta}}=\mathbf{0}$ where $\mathbf{0}=(0,0,0,...0,0)'$. That system of equations can be written as:

$$\begin{aligned}\hat{\beta}_0+\hat{\beta}_1x_1^{(1)}+\hat{\beta}_2x_2^{(1)}+...+\hat{\beta}_{(J-1)}x_{J-1}^{(1)}+\hat{\beta}_Jx_J^{(1)}&=0\\ \hat{\beta}_0+\hat{\beta}_1x_1^{(2)}+\hat{\beta}_2x_2^{(2)}+...+\hat{\beta}_{(J-1)}x_{J-1}^{(2)}+\hat{\beta}_Jx_J^{(2)}&=0\\ &.\\ &.\\ \hat{\beta}_0+\hat{\beta}_1x_1^{(R-1)}+\hat{\beta}_2x_2^{(R-1)}+...+\hat{\beta}_{(J-1)}x_{J-1}^{(R-1)}+\hat{\beta}_Jx_J^{(R-1)}&=0\\ \hat{\beta}_0+\hat{\beta}_1x_1^{(R)}+\hat{\beta}_2x_2^{(R)}+...+\hat{\beta}_{(J-1)}x_{J-1}^{(R)}+\hat{\beta}_Jx_J^{(R)}&=0\end{aligned} \tag{A33}$$

Assume that the rank of the boundary matrix is less than the number of its rows. For example, assume that $rank(\mathbf{B})=R-1$. By the fundamental theorem of linear algebra, only $R-1$ columns of $\mathbf{B}$ are linearly independent. Assume that the linearly independent columns are the first $R-1$ columns of $\mathbf{B}$. Using elementary column operations, the matrix $\mathbf{B}$ can be reduced to the simplified form:

$$\begin{bmatrix}1 & x_1^{(1)} & x_2^{(1)} & . & . & x_{R-1}^{(1)} & 0 & . & . & 0\\ 1 & x_1^{(2)} & x_2^{(2)} & . & . & x_{R-1}^{(2)} & 0 & . & . & 0\\ . & . & . & . & . & . & . & . & . & .\\ . & . & . & . & . & . & . & . & . & .\\ 1 & x_1^{(R-1)} & x_2^{(R-1)} & . & . & x_{R-1}^{(R-1)} & 0 & . & . & 0\\ 1 & x_1^{(R)} & x_2^{(R)} & . & . & x_{R-1}^{(R)} & 0 & . & . & 0\end{bmatrix} \tag{A34}$$

Evaluated at the ML solution $\hat{\beta}_j, j=0,1,2,...J$, if it exists, this system of equations can be written as:

$$\begin{aligned}
\hat{\beta}_0 + \hat{\beta}_1 x_1^{(1)} + \hat{\beta}_2 x_2^{(1)} + ... + \hat{\beta}_{(R-1)} x_{R-1}^{(1)} + \hat{\beta}_{(R)}(0) + ... + \hat{\beta}_J(0) = 0 \\
\hat{\beta}_0 + \hat{\beta}_1 x_1^{(2)} + \hat{\beta}_2 x_2^{(2)} + ... + \hat{\beta}_{(R-1)} x_{R-1}^{(2)} + \hat{\beta}_{(R)}(0) + ... + \hat{\beta}_J(0) = 0 \\
\vdots \\
\hat{\beta}_0 + \hat{\beta}_1 x_1^{(R-1)} + \hat{\beta}_2 x_2^{(R-1)} + ... + \hat{\beta}_{(R-1)} x_{R-1}^{(R-1)} + \hat{\beta}_{(R)}(0) + ... + \hat{\beta}_J(0) = 0 \\
\hat{\beta}_0 + \hat{\beta}_1 x_1^{(R)} + \hat{\beta}_2 x_2^{(R)} + ... + \hat{\beta}_{(R-1)} x_{R-1}^{(R)} + \hat{\beta}_{(R)}(0) + ... + \hat{\beta}_J(0) = 0
\end{aligned} \qquad \text{(A35)}$$

By subtracting R from previous $R-1$ equations, (A35) can be simplified to:

$$\begin{aligned}
\hat{\beta}_1\left(x_1^{(1)} - x_1^{(R)}\right) + \hat{\beta}_2\left(x_2^{(1)} - x_2^{(R)}\right) + ... + \hat{\beta}_{(R-1)}\left(x_{R-1}^{(1)} - x_{R-1}^{(R)}\right) = 0 \\
\hat{\beta}_1\left(x_1^{(2)} - x_1^{(R)}\right) + \hat{\beta}_2\left(x_2^{(2)} - x_2^{(R)}\right) + ... + \hat{\beta}_{(R-1)}\left(x_{R-1}^{(2)} - x_{R-1}^{(R)}\right) = 0 \\
\vdots \\
\hat{\beta}_1\left(x_1^{(R-1)} - x_1^{(R)}\right) + \hat{\beta}_2\left(x_2^{(R-1)} - x_2^{(R)}\right) + ... + \hat{\beta}_{(R-1)}\left(x_{R-1}^{(R-1)} - x_{R-1}^{(R)}\right) = 0
\end{aligned} \qquad \text{(A36)}$$

for which the system matrix is:

$$\mathbf{B}^* = \begin{bmatrix}
\left(x_1^{(1)} - x_1^{(R)}\right) & \left(x_2^{(1)} - x_2^{(R)}\right) & . & . & \left(x_{R-1}^{(1)} - x_{R-1}^{(R)}\right) \\
\left(x_1^{(2)} - x_1^{(R)}\right) & \left(x_2^{(2)} - x_2^{(R)}\right) & . & . & \left(x_{R-1}^{(2)} - x_{R-1}^{(R)}\right) \\
. & . & . & . & . \\
. & . & . & . & . \\
\left(x_1^{(R-1)} - x_1^{(R)}\right) & \left(x_2^{(R-1)} - x_2^{(R)}\right) & . & . & \left(x_{R-1}^{(R-1)} - x_{R-1}^{(R)}\right)
\end{bmatrix} \qquad \text{(A37)}$$

$\mathbf{B}^*$ is a $(R-1)\times(R-1)$ full rank square matrix, and hence $det\left(\mathbf{B}^*\right) \neq 0$. Denote $\hat{\boldsymbol{\beta}}^* = \left(\hat{\beta}_1, \hat{\beta}_2, ...\hat{\beta}_{R-1}\right)'$, then $\mathbf{B}^*\hat{\boldsymbol{\beta}}^* = 0$ has a single unique solution, but $\mathbf{B}^*\hat{\boldsymbol{\beta}}^* = 0$ comprises a homogeneous system, which has the zero solution at least. Hence the single unique solution is the zero solution. But this is a contradiction because none of the elements of $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ...\hat{\beta}_{J-1}, \hat{\beta}_J\right)'$ are zero. Hence the assumption that the rank of the boundary matrix is less than the number of its rows must be incorrect.

The argument preceding was specific to the case that $rank\left(\mathbf{B}\right) = R-1$, but it extends in straightforward fashion to the general case that $rank\left(\mathbf{B}\right) = R-m$ because equation system (A33) can be transformed into a homogeneous system of the form (A36) for any value of m such that $0 < m < R$. Hence we conclude that the rank of the boundary matrix cannot be less than the number of its rows, which is $rank\left(\mathbf{B}\right) = R$. Therefore, all rows are linearly independent.

<u>*Corollary* A4.1</u>

For a log binomial model with $J \geq 1$ non-constant covariates fitted by the exact method to data having R $(1 < R \leq J)$ distinct sets of values of the covariates (including the constant) for which the estimated outcome probability evaluated at the ML solution is unity, there is at least one covariate for which $t_{r-1}^{(r)} \neq 0,\ r = 2,3,...R$.

<u>Proof of corollary A4.1</u>

Appendix

Consider again matrix $\mathbf{B}$, the $R \times (J+1)$ matrix with R rows and $J+1$ columns formed from the R distinct sets of covariate values (including the constant) for which the estimated outcome probability evaluated ML solution is unity. From Theorem 4, the rows of $\mathbf{B}$ are linearly independent and hence $row\ rank(\mathbf{B}) = R$. By the fundamental theorem of linear algebra, the column rank of the boundary matrix must be R also. Proofs that $row\ rank(\mathbf{B}) = col\ rank(\mathbf{B})$ can be credited to Mackiw [90] and Wardlaw [91].

Therefore, if $t_{r-1}^{(r)} = 0$ at any step $r = 2,3,...R$ of the re-parameterization, it is always possible to find another covariate for which $t_{r-1}^{(r)} \neq 0$. In that case, the re-parameterization can proceed after re-ordering the covariates.

Table A1: The example data and the corresponding fitted probabilities for each observation evaluated at the exact estimates.

obs	y	x_1	x_2	x_3	$\hat{\mu}^*$
1	0	14	3.90	14.500	0.329
2	0	22	3.18	4.504	0.055
3	0	12	4.72	13.594	0.038
4	0	14	4.13	6.303	0.019
5	0	18	3.69	4.890	0.026
6	1	14	3.42	12.990	0.751
7	0	34	1.80	4.425	0.520
8	0	18	3.47	4.934	0.046
9	1	35	2.05	3.798	0.205
10	1	26	1.83	3.895	1.000
11	1	17	2.83	9.690	1.000

* $\hat{\mu}$ is the fitted probabilities evaluated at the exact estimates, $\hat{\beta}_0 = 6.5206677$, $\hat{\beta}_1 = -0.1098078$, $\hat{\beta}_2 = -2.5921916$, and $\hat{\beta}_3 = 0.2767768$.

Table A2: Simulation results

Setting		Simulations without a boundary vector				Simulations with a boundary vector			
		n	Bias*	MSE†	Coverage‡	*n*	Bias*	MSE†	Coverage‡
1	β_0	4408	–1.169	4.703	95.6	5592	0.724	4.231	95.1
	β_1		–2.287	3.489	94.2		1.065	3.209	94.4
	β_2		–3.579	0.079	95.2		1.117	0.055	94.8
2	β_0	4507	–1.309	4.732	95.0	5493	0.924	4.246	95.2
	β_1		–2.770	3.478	94.5		0.951	3.150	94.9
	β_2		–3.561	0.079	95.4		1.415	0.056	95.2
3	β_0	4536	–0.318	2.822	94.9	5464	0.764	2.878	94.7
	β_1		–2.469	3.383	95.2		1.090	3.222	94.6
	β_2		–3.683	0.078	95.4		1.429	0.057	94.7
4	β_0	4645	–0.398	2.913	94.9	5355	0.638	2.832	95.1
	β_1		–2.772	3.495	94.3		1.155	3.198	95.1
	β_2		–3.589	0.080	95.0		1.197	0.055	95.2
5	β_0	4637	0.640	2.650	94.9	5363	0.570	2.663	95.0
	β_1		–2.822	3.434	94.5		1.481	3.212	94.6
	β_2		–3.609	0.080	94.8		1.082	0.055	94.9
6	β_0	4575	0.752	2.659	95.3	5425	0.734	2.644	95.3
	β_1		–2.452	3.412	95.2		1.760	3.193	95.6
	β_2		–3.614	0.078	95.7		1.130	0.054	95.3
7	β_0	4647	2.418	2.881	95.7	5353	0.598	2.838	94.5
	β_1		–2.215	3.370	94.9		1.648	3.267	94.5
	β_2		–3.530	0.080	95.3		1.285	0.055	95.3
8	β_0	4519	2.187	2.919	95.6	5481	0.795	2.809	95.0
	β_1		–2.613	3.399	94.7		2.100	3.220	95.0
	β_2		–3.726	0.079	95.5		1.084	0.055	94.9

* Average percent relative bias: $(100/n)\sum_{k=1}^{n}\left[\left(\hat{\beta}_{jk}-\beta_j\right)\Big/\beta_j\right],\ j=0,1,2;k=1,2,...n$

† 100 times the average mean squared error: $(100/n)\sum_{k=1}^{n}\left[\left(\hat{\beta}_{jk}-\beta_j\right)^2+\widehat{\mathrm{Var}}\left(\hat{\beta}_{jk}\right)\right]$,

$j=0,1,2;k=1,2,...n$.

‡ Percentage of 95 percent confidence intervals covering the design value of the coefficient.

Table A3: Simulation results with n = 10000 replications.

		Exact method			Modified Poisson			R-*lbreg*		
		Bias*	MSE†	Cov‡	Bias*	MSE†	Cov‡	Bias*	MSE†	Cov‡
1	β_0	–0.110	4.439	95.3	0.655	5.527	95.1	–0.095	4.437	95.3
	β_1	–0.413	3.332	94.5	0.742	3.760	94.8	–0.403	3.332	94.4
	β_2	–0.953	0.066	95.0	0.639	0.098	95.2	–0.935	0.066	94.6
2	β_0	–0.087	4.465	95.1	0.656	5.538	95.0	–0.072	4.462	95.1
	β_1	–0.726	3.298	94.7	0.412	3.725	95.3	–0.716	3.297	94.7
	β_2	–0.827	0.066	95.3	0.715	0.098	94.8	–0.811	0.066	94.9
3	β_0	0.273	2.853	94.8	0.672	3.062	94.9	0.289	2.855	94.8
	β_1	–0.524	3.295	94.9	0.596	3.721	95.5	–0.510	3.296	94.9
	β_2	–0.889	0.067	95.0	0.680	0.099	95.2	–0.873	0.066	94.6
4	β_0	0.157	2.869	95.0	0.539	3.084	95.1	0.175	2.871	95.0
	β_1	–0.669	3.336	94.7	0.397	3.772	95.0	–0.655	3.336	94.7
	β_2	–1.026	0.066	95.1	0.420	0.098	94.9	–1.006	0.066	94.6
5	β_0	0.603	2.657	94.9	0.602	2.719	95.2	0.616	2.658	94.9
	β_1	–0.515	3.315	94.5	0.537	3.718	95.0	–0.499	3.314	94.5
	β_2	–1.093	0.066	94.9	0.432	0.099	94.6	–1.083	0.066	94.2
6	β_0	0.742	2.651	95.3	0.727	2.708	95.4	0.755	2.652	95.3
	β_1	–0.167	3.293	95.4	0.933	3.716	95.6	–0.153	3.292	95.3
	β_2	–1.041	0.065	95.5	0.572	0.097	95.2	–1.029	0.065	94.9
7	β_0	1.444	2.858	95.0	0.984	2.978	94.7	1.459	2.859	95.0
	β_1	–0.147	3.315	94.7	1.039	3.743	95.0	–0.131	3.314	94.7
	β_2	–0.953	0.067	95.3	0.625	0.100	94.7	–0.944	0.066	94.8
8	β_0	1.424	2.859	95.3	0.913	2.987	94.9	1.440	2.860	95.3
	β_1	–0.030	3.301	94.9	1.099	3.753	95.1	–0.012	3.300	94.8
	β_2	–1.090	0.066	95.2	0.528	0.099	94.8	–1.081	0.066	94.6

* Average percent relative bias: $(100/n)\sum_{k=1}^{n}\left[\left(\hat{\beta}_{jk}-\beta_j\right)/\beta_j\right],\ j=0,1,2;k=1,2,...n$

† 100 times the average mean squared error: $(100/n)\sum_{k=1}^{n}\left[\left(\hat{\beta}_{jk}-\beta_j\right)^2+\widehat{\mathrm{Var}}\left(\hat{\beta}_{jk}\right)\right]$,

$j=0,1,2;k=1,2,...n$.

‡ Percentage of 95 percent confidence intervals covering the design value of the coefficient.

Table A4: 95% confidence interval coverage with three estimators of estimated variance.

Setting		Replications without a boundary vector				Replications with a boundary vector			
			Variance estimator				Variance estimator		
		n	OIM*	EIM†	Robust‡	n	OIM*	EIM†	Robust‡
1	β_0	4408	95.6	95.1	95.0	5592	95.1	95.0	95.1
	β_1	4408	94.2	94.1	94.6	5592	94.4	94.1	94.9
	β_2	4408	95.2	95.0	95.1	5592	94.8	94.6	95.2
2	β_0	4507	95.0	94.6	94.9	5493	95.2	95.0	95.0
	β_1	4507	94.5	94.3	94.6	5493	94.9	94.3	95.4
	β_2	4507	95.4	94.9	94.9	5493	95.2	94.7	94.8
3	β_0	4536	94.9	94.7	94.7	5464	94.7	94.4	95.0
	β_1	4536	95.2	94.8	95.3	5464	94.6	94.3	95.6
	β_2	4536	95.4	94.8	94.6	5464	94.7	94.5	95.4
4	β_0	4645	94.9	94.6	94.8	5355	95.1	94.8	95.2
	β_1	4645	94.3	94.5	94.8	5355	95.1	94.9	95.1
	β_2	4645	95.0	94.6	94.7	5355	95.2	94.7	94.9
5	β_0	4637	94.9	94.7	95.0	5363	95.0	95.1	95.3
	β_1	4637	94.5	94.4	94.9	5363	94.6	94.5	95.0
	β_2	4637	94.8	94.5	94.5	5363	94.9	94.6	94.6
6	β_0	4575	95.3	94.7	95.2	5425	95.3	94.7	95.6
	β_1	4575	95.2	94.8	95.2	5425	95.6	95.2	95.7
	β_2	4575	95.7	94.9	95.1	5425	95.3	94.7	95.3
7	β_0	4647	95.7	95.0	94.5	5353	94.5	94.5	94.9
	β_1	4647	94.9	94.5	94.6	5353	94.5	94.4	95.3
	β_2	4647	95.3	95.0	94.5	5353	95.3	94.9	94.9
8	β_0	4519	95.6	94.9	94.6	5481	95.0	94.7	94.9
	β_1	4519	94.7	94.5	95.0	5481	95.0	94.8	95.2
	β_2	4519	95.5	94.9	94.5	5481	94.9	94.8	94.9

* Percentage of 95 percent confidence intervals covering the design value of the coefficient with variance estimated using the observed information matrix (OIM).

† Percentage of 95 percent confidence intervals covering the design value of the coefficient with variance estimated using the expected information matrix (EIM).

‡ Percentage of 95 percent confidence intervals covering the design value of the coefficient with variance estimated using the robust (sandwich) estimator.

Appendix B

Supplementary materials of the identity-link binomial model in the exact method

The identity-link binomial model

Assume that there are n independent observations with a binary outcome variable Y and J covariates $(1, X_1, X_2, ...X_J)$ with constant. The likelihood function of identity-link binomial model is defended as:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \mu_i^{y_i} (1-\mu_i)^{(1-y_i)} \tag{B1}$$

where p_i is the probability of outcome conditional on the i^{th} observation which is equal to a linear combination of parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots \beta_J)$ and covariates $\mathbf{x}_{ij} = (1, x_{i1} x_{i2}, \ldots x_{iJ})$ with constant denoted as

$$\Pr(Y_i = 1 \mid \mathbf{x}_i) = \pi(\mathbf{x}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_J x_{iJ}, \; i = 1, 2, ...n \,. \tag{B2}$$

To apply exact method, the likelihood function (3.1) need to be reformed as:

$$l(\boldsymbol{\beta}) = \prod_{Y_i=1} \mu_i \prod_{Y_i=0} 1-\mu_i \tag{B3}$$

where $p_i = \mathbf{x}_i' \boldsymbol{\beta}$, for $\mathbf{x}_{ij} = (1, x_{i1}, x_{i2}, \ldots x_{iJ})$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots \beta_J)$. Assume that there are R boundary vectors in the model. The model obtains maximum likelihood solution at $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_J)$. Then, the likelihood function evaluated at $\hat{\boldsymbol{\beta}}$ becomes:

$$\begin{aligned} l(\hat{\beta}) &= \prod_{Y_i=1} \hat{\mu}_i \prod_{Y_i=0} 1-\hat{\mu}_i \\ &= \prod_{Y_i=1} \left(\hat{\beta}_0 + \sum_{j=1}^{J} x_{ij} \hat{\beta}_j \right) \prod_{Y_i=0} \left[1 - \left(\hat{\beta}_0 + \sum_{j=1}^{J} x_{ij} \hat{\beta}_j \right) \right] \end{aligned} \tag{B4}$$

From here, the estimation is splitting into two situations which are ML solution is on the upper bound of parameter space or on the lower bound of parameter space.

The ML solution is on the upper bond of parameter space in the identity-link binomial model in the exact method.

Theorem B1

Suppose that the ML solution is on the upper bound of parameter space, which means the maximum fitted probability is equal to unity, $\max(\hat{\mu}_i) = 1$ *for* $i = 1, 2, ...n$. *Assume that there are* $R \geq 1$ *distinct sets of covariate vectors with the fitted probability attained unity referred as boundary vectors. Denote the* r^{th} $(r = 1, 2, ...R)$ *boundary vector as* $\mathbf{x}^{(r)} = \left(1, x_1^{(r)}.x_2^{(r)}, ...x_J^{(r)}\right)$ *which shares covariate values with* n_r *observations. Then, the constant and first* $R-1$ *covariates in the likelihood function* (B4) *of identity-link binomial model evaluated at the maximum likelihood solution* $\hat{\boldsymbol{\beta}}$ *can be re-parametrized as following:*

$$l\left(\hat{\boldsymbol{\beta}}\right)=\prod_{\substack{Y_i=1,x_{ij}\neq x_j^{(1)},\\ x_{ij}\neq x_j^{(2)},...,\\ x_{ij}\neq x_j^{(R)}}}\hat{\mu}_i^{*}\prod_{Y_i=0}1-\hat{\mu}_i^{*} \tag{B5}$$

where $\hat{\mu}_i^{*}=1+\sum_{j=R}^{J}z_{ij}^{(R)}\beta_j$. The function for each $z_{ij}^{(r)}$ is:

$$z_{ij}^{(1)}=x_{ij}-x_j^{(1)},\ i=1,2,...n \text{ and } j=0,1,2,...J\ . \tag{B6}$$

$$z_{ij}^{(r)}=z_{ij}^{(r-1)}-\left(\frac{t_j^{(r)}}{t_{r-1}^{(r)}}\right)z_{i,r-1}^{(r-1)} \text{ where } t_j^{(r)}=z_{ij}^{(r-1)}\Big|_{x_{ij}=x_j^{(r)}},\ r=2,3,...R\ . \tag{B7}$$

Proof of Theorem B1

Let $\mathbf{x}^{(1)}$ is the first boundary vector in the model, which makes the probability of observation is equal to unity under ML solution. Then, the likelihood function of identity-link binomial model is rewritten as:

$$\begin{aligned}l\left(\hat{\boldsymbol{\beta}}\right)&=\prod_{Y_i=1}\left(\hat{\beta}_0+\sum_{j=1}^{J}x_{ij}\hat{\beta}_j\right)\prod_{Y_i=0}\left[1-\left(\hat{\beta}_0+\sum_{j=1}^{J}x_{ij}\hat{\beta}_j\right)\right]\\&=\left\{\prod_{\substack{Y_i=1,\\ x_{ij}\neq x_j^{(1)}}}\left(\hat{\beta}_0+\sum_{j=1}^{J}x_{ij}\hat{\beta}_j\right)\prod_{Y_i=0}\left[1-\left(\hat{\beta}_0+\sum_{j=1}^{J}x_{ij}\hat{\beta}_j\right)\right]\right\}\left(\hat{\beta}_0+\sum_{j=1}^{J}x_j{}^{(1)}\hat{\beta}_j\right)\end{aligned} \tag{B8}$$

Since $\hat{\beta}_0+\sum_{j=1}^{J}x_j^{(1)}\hat{\beta}_j=1$ and

$$\hat{\beta}_0=1-\sum_{j=1}^{J}x_j^{(1)}\hat{\beta}_j\ , \tag{B9}$$

the (B8) can be rewritten as:

$$\begin{aligned}l\left(\hat{\boldsymbol{\beta}}\right)&=\prod_{\substack{Y_i=1,\\ x_{ij}\neq x_j^{(1)}}}\left(\hat{\beta}_0+\sum_{j=1}^{J}x_{ij}\hat{\beta}_j\right)\prod_{Y_i=0}\left[1-\left(\hat{\beta}_0+\sum_{j=1}^{J}x_{ij}\hat{\beta}_j\right)\right]\\&=\prod_{\substack{Y_i=1,\\ x_{ij}\neq x_j^{(1)}}}\left(1-\sum_{j=1}^{J}x_j^{(1)}\hat{\beta}_j+\sum_{j=1}^{J}x_{ij}\hat{\beta}_j\right)\prod_{Y_i=0}\left[1-\left(1-\sum_{j=1}^{J}x_j^{(1)}\hat{\beta}_j+\sum_{j=1}^{J}x_{ij}\hat{\beta}_j\right)\right]\\&=\prod_{\substack{Y_i=1,\\ x_{ij}\neq x_j^{(1)}}}\left[1+\sum_{j=1}^{J}\left(x_{ij}-x_j^{(1)}\right)\hat{\beta}_j\right]\prod_{Y_i=0}\left\{1-\left[1+\sum_{j=1}^{J}\left(x_{ij}-x_j^{(1)}\right)\hat{\beta}_j\right]\right\}\end{aligned} \tag{B10}$$

Let $z_{ij}^{(1)}=x_{ij}-x_j^{(1)}$. Then, rewrite (B10) to:

$$l\left(\hat{\boldsymbol{\beta}}\right)=\prod_{\substack{Y_i=1,\\ x_{ij}\neq x_j^{(1)}}}\left(1+\sum_{j=1}^{J}z_{ij}^{(1)}\hat{\beta}_j\right)\prod_{Y_i=0}\left[1-\left(1+\sum_{j=1}^{J}z_{ij}^{(1)}\hat{\beta}_j\right)\right]. \tag{B11}$$

By the second boundary vector $\mathbf{t}^{(2)}=\mathbf{x}^{(2)}-\mathbf{x}^{(1)}$, the likelihood function (B11) is rewritten as:

Appendix

$$l\left(\hat{\boldsymbol{\beta}}\right)=\left\{\prod_{\substack{Y_i=1, x_{ij}\neq x_j^{(1)},\\ x_{ij}\neq x_j^{(2)}}}\left(1+\sum_{j=1}^{J} z_{ij}^{(1)}\hat{\beta}_j\right)\prod_{Y_i=0}\left[1-\left(1+\sum_{j=1}^{J} z_{ij}^{(1)}\hat{\beta}_j\right)\right]\right\}\left(1+\sum_{j=1}^{J} t_j^{(2)}\hat{\beta}_j\right) \tag{B12}$$

Since $1+\sum_{j=1}^{J} t_j^{(2)}\hat{\beta}_j=1$ and $1+t_1^{(2)}\hat{\beta}_1+\sum_{j=2}^{J} t_j^{(2)}\hat{\beta}_j=1$, rearrange it as:

$$\hat{\beta}_1=-\sum_{j=2}^{J}\frac{t_j^{(2)}}{t_1^{(2)}}\hat{\beta}_j \tag{B13}$$

Substitute (B13) into (B12). It is reformed as:

$$l\left(\hat{\boldsymbol{\beta}}\right)=\prod_{\substack{Y_i=1,\\ x_{ij}\neq x_j^{(1)},\\ x_{ij}\neq x_j^{(2)}}}\left[1+\sum_{j=2}^{J}\left(z_{ij}^{(1)}-\frac{t_j^{(2)}}{t_1^{(2)}}z_{i1}^{(1)}\right)\hat{\beta}_j\right]\prod_{Y_i=0}\left\{1-\left[1+\sum_{j=2}^{J}\left(z_{ij}^{(1)}-\frac{t_j^{(2)}}{t_1^{(2)}}z_{i1}^{(1)}\right)\hat{\beta}_j\right]\right\} \tag{B14}$$

Let $z_{ij}^{(2)}=z_{ij}^{(1)}-\frac{t_j^{(2)}}{t_1^{(2)}}z_{i1}^{(1)}$. Then, rewrite (B14) as:

$$l\left(\hat{\boldsymbol{\beta}}\right)=\prod_{\substack{Y_i=1, x_{ij}\neq x_j^{(1)},\\ x_{ij}\neq x_j^{(2)}}}\left(1+\sum_{j=2}^{J} z_{ij}^{(2)}\hat{\beta}_j\right)\prod_{Y_i=0}\left[1-\left(1+\sum_{j=2}^{J} z_{ij}^{(2)}\hat{\beta}_j\right)\right] \tag{B15}$$

By repeating the previous step on the rest of each $R-2$ boundary vectors, the likelihood function is restructured by

$$\hat{\beta}_r=-\sum_{j=r+1}^{J}\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\hat{\beta}_j,\ \text{for } r=1,2,\ldots R-1 \tag{B16}$$

to:

$$l\left(\hat{\boldsymbol{\beta}}\right)=\prod_{\substack{Y_i=1, x_{ij}\neq x_j^{(1)},\\ x_{ij}\neq x_j^{(2)},\cdots\\ x_{ij}\neq x_j^{(R)}}}\left(1+\sum_{j=R}^{J} z_{ij}^{(R)}\hat{\beta}_j\right)\prod_{Y_i=0}\left[1-\left(1+\sum_{j=R}^{J} z_{ij}^{(R)}\hat{\beta}_j\right)\right]. \tag{B17}$$

After the reparameterization procedure, the reformed likelihood function is:

$$l\left(\hat{\boldsymbol{\beta}}\right)=\prod_{\substack{Y_i=1, x_{ij}\neq x_j^{(1)},\\ x_{ij}\neq x_j^{(2)},\cdots\\ x_{ij}\neq x_j^{(R)}}}\mu_i^{*}\prod_{Y_i=0}1-\mu_i^{*} \tag{B18}$$

Where $\mu_i^{*}=1+\sum_{j=R}^{J} z_{ij}^{(R)}\hat{\beta}_j$.

Theorem B2

For an identity-link binomial model with $J \geq 1$ independent non-constant covariates fitted by the exact method to data having R $(1 \leq R \leq J)$ distinct sets of values of the covariates (including the constant) for which the estimated outcome probability evaluated at the ML solution is unity, the estimated coefficients and the corresponding standard errors of the $R-1$ eliminated non-constant covariates are given by:

$$\hat{\beta}_r = -\sum_{j=r+1}^{J} \frac{t_j^{(r+1)}}{t_r^{(r+1)}} \hat{\beta}_j, \text{ for } r = 1, 2, \ldots R-1. \tag{B19}$$

and

$$\widehat{\mathrm{SE}}\left(\hat{\beta}_r\right) = \sqrt{\sum_{j=r+1}^{J} \left(\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=r+1 \\ j_1 \neq j_2}}^{J} \sum_{j_2=r+1}^{J} \left\{\frac{t_{j_1}^{(r+1)} t_{j_2}^{(r+1)}}{\left(t_r^{(r+1)}\right)^2}\right\} \widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1}, \hat{\beta}_{j_2}\right)} \tag{B20}$$

and the estimated covariances between exact estimates of the coefficients of the $R-1$ eliminated non-constant covariates and exact estimates of the coefficients of the $J-R+1$ remaining non-constant covariates are given by:

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_r, \hat{\beta}_s\right) = -\left\{\frac{t_s^{(r+1)}}{t_r^{(r+1)}} \widehat{\mathrm{Var}}\left(\hat{\beta}_s\right) + \sum_{\substack{j=r+1 \\ j \neq s}}^{J} \left[\frac{t_j^{(r+1)}}{t_r^{(r+1)}} \widehat{\mathrm{Cov}}\left(\hat{\beta}_j, \hat{\beta}_s\right)\right]\right\}, \tag{B21}$$

for $r = 1, 2, \ldots R-1$ and $s = r+1, \ldots J$.

Proof of Theorem B2

The formula (B19) has been proved through the proof of Theorem 1. The proof is focusing on the (B20) and (B21).

Since $\hat{\beta}_r = -\sum_{j=r+1}^{J} \frac{t_j^{(r+1)}}{t_r^{(r+1)}} \hat{\beta}_j$, we get the variance function of $\hat{\beta}_r$ as:

$$\widehat{\mathrm{Var}}\left(\hat{\beta}_r\right) = \widehat{\mathrm{Var}}\left[-\sum_{j=r+1}^{J} \frac{t_j^{(r+1)}}{t_r^{(r+1)}} \hat{\beta}_j\right] \tag{B22}$$

Expanding the (B22):

$$\widehat{\mathrm{Var}}\left(\hat{\beta}_r\right) = \sum_{j=r+1}^{J} \left(\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=r+1 \\ j_1 \neq j_2}}^{J} \sum_{j_2=r+1}^{J} \left[\frac{t_{j_1}^{(r+1)} t_{j_2}^{(r+1)}}{\left(t_r^{(r+1)}\right)^2}\right] \widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1}, \hat{\beta}_{j_2}\right) \tag{B23}$$

The formula of standard error for the first $R-1$ estimate of non-constant eliminating coefficients that did not achieve from the fitting procedure of model is obtained by square rooting both sides of (B23):

$$\widehat{\mathrm{SE}}\left(\hat{\beta}_r\right) = \sqrt{\sum_{j=r+1}^{J} \left[\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\right]^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=r+1 \\ j_1 \neq j_2}}^{J} \sum_{j_2=r+1}^{J} \left[\frac{t_{j_1}^{(r+1)} t_{j_2}^{(r+1)}}{\left(t_r^{(r+1)}\right)^2}\right] \widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1}, \hat{\beta}_{j_2}\right)}. \tag{B24}$$

To obtain the formula of covariance between the first $R-1$ estimate of non-constant eliminating coefficients that did not achieve from the fitting procedure of model and the rest of the coefficients in the model, we begin with restructuring (B19):

$$t_r^{(r+1)}\hat{\beta}_r = -\sum_{j=r+1}^{J}\hat{\beta}_j t_j^{(r+1)},\ r=1,2,...R-1 \tag{B25}$$

and adding each side of the equation the estimated coefficient of one of the $J-R+1$ covariates remaining in the model:

$$t_r^{(r+1)}\hat{\beta}_r + \hat{\beta}_s = -\left[\left(t_s^{(r+1)}-1\right)\hat{\beta}_s + \sum_{\substack{j=r+1\\ j\neq s}}^{J}\hat{\beta}_j t_j^{(r+1)}\right],\ r=1,2,...,R-1;\ s=r+1,\ldots J$$

Thus:

$$\begin{aligned}&\widehat{\mathrm{Var}}\left(t_r^{(r+1)}\hat{\beta}_r+\hat{\beta}_s\right)\\ &\quad=\widehat{\mathrm{Var}}\left\{-\left[\left(t_s^{(r+1)}-1\right)\hat{\beta}_s+\sum_{\substack{j=r+1\\ j\neq s}}^{J}\hat{\beta}_j t_j^{(r+1)}\right]\right\},\ r=1,2,...,R-1;\ s=r+1,\ldots J\end{aligned} \tag{B26}$$

Expanding the left-hand side of (B26):

$$\widehat{\mathrm{Var}}\left(t_r^{(r+1)}\hat{\beta}_r+\hat{\beta}_s\right)=\left(t_r^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_r\right)+2t_r^{(r+1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_r,\hat{\beta}_s\right)+\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right) \tag{B27}$$

Expanding the right-hand side of (B26):

$$\begin{aligned}&\widehat{\mathrm{Var}}\left\{-\left[\left(t_s^{(r+1)}-1\right)\hat{\beta}_s+\sum_{\substack{j=r+1\\ j\neq s}}^{J}\hat{\beta}_j t_j^{(r+1)}\right]\right\}\\ &=\left(t_s^{(r+1)}-1\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right)+2\sum_{\substack{j=r+1\\ j\neq s}}^{J}\left[t_j^{(r+1)}\left(t_s^{(r+1)}-1\right)\widehat{Cov}\left(\hat{\beta}_j,\hat{\beta}_s\right)\right]+\widehat{Var}\left[\sum_{\substack{j=r+1\\ j\neq s}}^{J}\hat{\beta}_j t_j^{(r+1)}\right]\\ &=\left[\left(t_s^{(r+1)}\right)^2-2t_s^{(r+1)}+1\right]\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right)+2\sum_{\substack{j=r+1\\ j\neq s}}^{J}\left[\left(t_j^{(r+1)}t_s^{(r+1)}-t_j^{(r+1)}\right)\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_s\right)\right]\\ &+\sum_{\substack{j=r+1\\ j\neq s}}^{J}\left[\left(t_j^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\right]+\sum_{j_1=r+1}^{J}\sum_{\substack{j_2=r+1\\ j_1,j_2\neq s\\ j_2\neq j_1}}^{J}\left[\left(t_{j_1}^{(r+1)}t_{j_2}^{(r+1)}\right)\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)\right]\\ &=\left(1-2t_s^{(r+1)}\right)\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right)-2\sum_{\substack{j=r+1\\ j\neq s}}^{J}\left[t_j^{(r+1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_s\right)\right]\\ &+\sum_{j=r+1}^{J}\left[\left(t_j^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\right]+\sum_{j_1=r+1}^{J}\sum_{\substack{j_2=r+1\\ j_2\neq j_1}}^{J}\left[\left(t_{j_1}^{(r+1)}t_{j_2}^{(r+1)}\right)\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)\right]\end{aligned} \tag{B28}$$

By expanding $\widehat{\mathrm{Var}}\left[\hat{\beta}_r t_r^{(r+1)}\right]=\widehat{\mathrm{Var}}\left[-\sum_{j=r+1}^{J}\hat{\beta}_j t_j^{(r+1)}\right],\ r=1,2,...R-1$, we have:

$$\left(t_r^{(r+1)}\right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_r\right) = \sum_{j=r+1}^{J}\left[\left(t_j^{(r+1)}\right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\right] + \sum_{j_1=r+1}^{J}\sum_{\substack{j_2=r+1 \\ j_2\neq j_1}}^{J}\left[\left(t_{j_1}^{(r+1)}t_{j_2}^{(r+1)}\right)\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)\right] \quad \text{(B29)}$$

Substituting for $\left(t_r^{(r+1)}\right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_r\right)$ in (B28):

$$\begin{aligned}&\widehat{\mathrm{Var}}\left\{-\left[\left(t_s^{(r+1)}-1\right)\hat{\beta}_s + \sum_{\substack{j=r+1 \\ j\neq s}}^{J}\hat{\beta}_j t_j^{(r+1)}\right]\right\} \\ &\quad = \left(1-2t_s^{(r+1)}\right)\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right) + \left(t_r^{(r+1)}\right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_r\right) - 2\sum_{\substack{j=r+1 \\ j\neq s}}^{J} t_j^{(r+1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_s\right)\end{aligned} \quad \text{(B30)}$$

Equating (B27) and (B30), the two sides of (B26), and simplifying:

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_r,\hat{\beta}_s\right) = -\left\{\frac{t_s^{(r+1)}}{t_r^{(r+1)}}\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right) + \sum_{\substack{j=r+1 \\ j\neq s}}^{J}\left[\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_s\right)\right]\right\},\ r=1,2,\ldots R-1;\ s=r+1,\ldots J$$

Theorem B3

For an identity-link binomial model with $J \geq 1$ independent non-constant covariates fitted by the exact method to data having R $(1 \leq R \leq J)$ distinct sets of values of the covariates (including the constant) for which the estimated outcome probability evaluated at the ML solution is unity, the estimated coefficient and the corresponding standard error of the constant covariate are given by:

$$\hat{\beta}_0 = 1 - \sum_{j=1}^{J} x_j^{(1)}\hat{\beta}_j \quad \text{(B31)}$$

and

$$\widehat{\mathrm{SE}}\left(\hat{\beta}_0\right) = \sqrt{\sum_{j=1}^{J}\left(x_j^{(1)}\right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=1 \\ j_1\neq j_2}}^{J}\sum_{j_2=1}^{J} x_{j_1}^{(1)}x_{j_2}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)} \quad \text{(B32)}$$

and the estimated covariances between the exact estimate of the coefficient of the constant covariate and exact estimates of the coefficients of the J non-constant covariates are given by:

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_0,\hat{\beta}_j\right) = -\left[x_j^{(1)}\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=1 \\ j_1\neq j}}^{J} x_{j_1}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_{j_1}\right)\right] \quad \text{(B33)}$$

for $j = 1,2,\ldots J$.

Proof of Theorem B3

The formula (B31) has been proved through the proof of Theorem 1. The proof is focusing on the (B32) and (B33).

Since $\hat{\beta}_0 = 1 - \sum_{j=1}^{J} x_j^{(1)}\hat{\beta}_j$, we get the variance function of $\hat{\beta}_0$ as:

$$\begin{aligned}\widehat{\mathrm{Var}}\left(\hat{\beta}_0\right) &= \widehat{\mathrm{Var}}\left(1-\sum_{j=1}^{J} x_j^{(1)}\hat{\beta}_j\right) \\ &= \widehat{\mathrm{Var}}\left(\sum_{j=1}^{J} x_j^{(1)}\hat{\beta}_j\right)\end{aligned} \tag{B34}$$

Expanding the (B34):

$$\widehat{\mathrm{Var}}\left(\hat{\beta}_0\right) = \sum_{j=1}^{J}\left(x_j^{(1)}\right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=1 \\ j_1\neq j_2}}^{J}\sum_{j_2=1}^{J} x_{j_1}^{(1)} x_{j_2}^{(1)} \widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1}, \hat{\beta}_{j_2}\right) \tag{B35}$$

The formula of standard error for the estimate of constant is obtained by square rooting both sides of (B23):

$$\widehat{\mathrm{SE}}\left(\hat{\beta}_0\right) = \sqrt{\sum_{j=1}^{J}\left(x_j^{(1)}\right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=1 \\ j_1\neq j_2}}^{J}\sum_{j_2=1}^{J} x_{j_1}^{(1)} x_{j_2}^{(1)} \widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1}, \hat{\beta}_{j_2}\right)} \tag{B36}$$

To obtain the formula of covariance between the estimate of constant and the estimate of coefficients in the model, we begin with adding a $\hat{\beta}_j$ to both sides of $\hat{\beta}_0 = 1-\sum_{j=1}^{J} x_j^{(1)}\hat{\beta}_j$:

$$\hat{\beta}_0 + \hat{\beta}_j = 1 - \left[\left(x_j^{(1)} - 1\right)\hat{\beta}_j + \sum_{\substack{j_1=1 \\ j_1\neq j}}^{J} \hat{\beta}_{j_1} x_{j_1}^{(1)}\right], \quad j = 1, 2, \ldots J \tag{B37}$$

Thus:

$$\begin{aligned}\widehat{\mathrm{Var}}\left(\hat{\beta}_0 + \hat{\beta}_j\right) &= \widehat{\mathrm{Var}}\left\{1 - \left[\left(x_j^{(1)} - 1\right)\hat{\beta}_j + \sum_{\substack{j_1=1 \\ j_1\neq j}}^{J} \hat{\beta}_{j_1} x_{j_1}^{(1)}\right]\right\} \\ &= \widehat{\mathrm{Var}}\left[\left(x_j^{(1)} - 1\right)\hat{\beta}_j + \sum_{\substack{j_1=1 \\ j_1\neq j}}^{J} \hat{\beta}_{j_1} x_{j_1}^{(1)}\right]\end{aligned} \tag{B38}$$

Expanding the left-hand side of (B38):

$$\widehat{\mathrm{Var}}\left(\hat{\beta}_0 + \hat{\beta}_j\right) = \widehat{\mathrm{Var}}\left(\hat{\beta}_0\right) + 2\widehat{\mathrm{Cov}}\left(\hat{\beta}_0, \hat{\beta}_j\right) + \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) \tag{B39}$$

Expanding the right-hand side of (B38):

$$\begin{aligned}&\widehat{\mathrm{Var}}\left[\left(x_j^{(1)} - 1\right)\hat{\beta}_j + \sum_{\substack{j_1=1 \\ j_1\neq j}}^{J} \hat{\beta}_{j_1} x_{j_1}^{(1)}\right] \\ &= \left(x_j^{(1)} - 1\right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + 2\left(x_j^{(1)} - 1\right)\sum_{\substack{j_1=1 \\ j_1\neq j}}^{J} x_{j_1}^{(1)} \widehat{\mathrm{Cov}}\left(\hat{\beta}_j, \hat{\beta}_{j_1}\right) + \widehat{\mathrm{Var}}\left(\sum_{\substack{j_1=1 \\ j_1\neq j}}^{J} \hat{\beta}_{j_1} x_{j_1}^{(1)}\right)\end{aligned} \tag{B40}$$

Since:

$$\widehat{\mathrm{Var}}\left(\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}\hat{\beta}_{j_1}x_{j_1}^{(1)}\right)=\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}\left[\left(x_{j_1}^{(1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_{j_1}\right)\right]+\sum_{\substack{j_2=1\\ j_2\neq j}}^{J}\sum_{\substack{j_3=1\\ j_3\neq j\\ j_3\neq j_2}}^{J}\left[x_{j_2}^{(1)}x_{j_3}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_2},\hat{\beta}_{j_3}\right)\right], \quad \text{(B41)}$$

expand (B40) by (B41) as:

$$\begin{aligned}
&\widehat{\mathrm{Var}}\left[\left(x_j^{(1)}-1\right)\hat{\beta}_j+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}\hat{\beta}_{j_1}x_{j_1}^{(1)}\right]\\
&\quad=\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)-2\left[x_j^{(1)}\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}x_{j_1}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_{j_1}\right)\right]+\left(x_j^{(1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\\
&\quad+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}\left[\left(x_{j_1}^{(1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_{j_1}\right)\right]+2\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}x_j^{(1)}x_{j_1}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_{j_1}\right)\\
&\quad+\sum_{\substack{j_2=1\\ j_2\neq j}}^{J}\sum_{\substack{j_3=1\\ j_3\neq j\\ j_3\neq j_2}}^{J}\left[x_{j_2}^{(1)}x_{j_3}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_2},\hat{\beta}_{j_3}\right)\right]\\
&\quad=\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)-2\left[x_j^{(1)}\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}x_{j_1}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_{j_1}\right)\right]\\
&\quad+\sum_{j=1}^{J}\left[\left(x_j^{(1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\right]+\sum_{j_1=1}^{J}\sum_{\substack{j_2=1\\ j_1\neq j_2}}^{J}\left[x_{j_1}^{(1)}x_{j_2}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)\right]
\end{aligned} \quad \text{(B42)}$$

Replacing and simplifying (B42) by (B35):

$$\begin{aligned}
&\widehat{\mathrm{Var}}\left[\left(x_j^{(1)}-1\right)\hat{\beta}_j+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}\hat{\beta}_{j_1}x_{j_1}^{(1)}\right]\\
&\quad=\widehat{\mathrm{Var}}\left(\hat{\beta}_0\right)+\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)-2\left[x_j^{(1)}\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}x_{j_1}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_{j_1}\right)\right]
\end{aligned} \quad \text{(B43)}$$

Substituting (B39) and (B43) into (B38), and simplifying:

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_0,\hat{\beta}_j\right)=-\left[x_j^{(1)}\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}x_{j_1}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_{j_1}\right)\right],\ \ j=1,2,\ldots J \quad \text{(B44)}$$

The ML solution is on the lower bond of parameter space.

Theorem B4

Suppose that the ML solution is on the lower bound of parameter space, which means the minimum fitted probability is equal to zero, $\min\left(\hat{\mu}_i\right)=0$ *for* $i=1,2,\ldots n$ *. Assume that there are* $R\geq 1$ *distinct sets of covariate vectors with the fitted probability attained zero referred to as boundary vectors. Denote the* r^{th} $(r=1,2,\ldots R)$ *boundary vector as* $\mathbf{x}^{(r)}=\left(1,x_1^{(r)}.x_2^{(r)},\ldots x_J^{(r)}\right)$ *which shares covariate values with* n_r *observations. Then, the constant and first* $R-1$ *covariates in the likelihood function* (B4) *of the identity-link binomial model evaluated at the maximum likelihood solution* $\hat{\boldsymbol{\beta}}$ *can be reparametrized as following*:

$$l\left(\hat{\boldsymbol{\beta}}\right)=\prod_{Y_i=1}\mu_i^{*}\prod_{\substack{Y_i=0,x_{ij}\neq x_j^{(1)},\\ x_{ij}\neq x_j^{(2)},\cdots\\ x_{ij}\neq x_j^{(R)}}}1-\mu_i^{*} \tag{B45}$$

where $\mu_i^{*}=\sum_{j=R}^{J}z_{ij}^{(R)}\hat{\beta}_j$ *. The function of each* $z_{ij}^{(r)}$ *is:*

$$z_{ij}^{(1)}=x_{ij}-x_j^{(1)},\ i=1,2,...n \text{ and } j=0,1,2,...J\ . \tag{B46}$$

$$z_{ij}^{(r)}=z_{ij}^{(r-1)}-\left(\frac{t_j^{(r)}}{t_{r-1}^{(r)}}\right)z_{i,r-1}^{(r-1)} \text{ where } t_j^{(r)}=z_{ij}^{(r-1)}\Big|_{x_{ij}=x_j^{(r)}},\ r=2,3,...R\ .$$

<u>Proof of Theorem 4</u>

Let $\mathbf{x}^{(1)}$ is the first boundary vector in the model, which makes the probability of observation equals to zero under ML solution. Then, the likelihood function of identity-link binomial model is rewritten as:

$$\begin{aligned} l\left(\hat{\boldsymbol{\beta}}\right)&=\prod_{Y_i=1}\left(\hat{\beta}_0+\sum_{j=1}^{J}x_{ij}\hat{\beta}_j\right)\prod_{Y_i=0}\left[1-\left(\hat{\beta}_0+\sum_{j=1}^{J}x_{ij}\hat{\beta}_j\right)\right]\\ &=\left\{\prod_{Y_i=1}\left(\hat{\beta}_0+\sum_{j=1}^{J}x_{ij}\hat{\beta}_j\right)\prod_{\substack{Y_i=0,\\ x_{ij}\neq x_j^{(1)}}}\left[1-\left(\hat{\beta}_0+\sum_{j=1}^{J}x_{ij}\hat{\beta}_j\right)\right]\right\}\left[1-\left(\hat{\beta}_0+\sum_{j=1}^{J}x_j^{(1)}\hat{\beta}_j\right)\right] \end{aligned} \tag{B47}$$

Since $\hat{\beta}_0+\sum_{j=1}^{J}x_j^{(1)}\hat{\beta}_j=0$ and $1-\left(\hat{\beta}_0+\sum_{j=1}^{J}x_j^{(1)}\hat{\beta}_j\right)=1$, we have:

$$\hat{\beta}_0=-\sum_{j=1}^{J}x_j^{(1)}\hat{\beta}_j\ , \tag{B48}$$

the (B47) can be rewritten as:

$$\begin{aligned} l\left(\hat{\boldsymbol{\beta}}\right)&=\prod_{Y_i=1}\left(\hat{\beta}_0+\sum_{j=1}^{J}x_{ij}\hat{\beta}_j\right)\prod_{Y_i=0}\left[1-\left(\hat{\beta}_0+\sum_{j=1}^{J}x_{ij}\hat{\beta}_j\right)\right]\\ &=\prod_{Y_i=1}\left(-\sum_{j=1}^{J}x_j^{(1)}\hat{\beta}_j+\sum_{j=1}^{J}x_{ij}\hat{\beta}_j\right)\prod_{\substack{Y_i=0,\\ x_{ij}\neq x_j^{(1)}}}\left[1-\left(-\sum_{j=1}^{J}x_j^{(1)}\hat{\beta}_j+\sum_{j=1}^{J}x_{ij}\hat{\beta}_j\right)\right]\\ &=\prod_{Y_i=1}\left[\sum_{j=1}^{J}\left(x_{ij}-x_j^{(1)}\right)\hat{\beta}_j\right]\prod_{\substack{Y_i=0,\\ x_{ij}\neq x_j^{(1)}}}\left[1-\left\{\sum_{j=1}^{J}\left(x_{ij}-x_j^{(1)}\right)\hat{\beta}_j\right\}\right] \end{aligned} \tag{B49}$$

Let $z_{ij}^{(1)}=x_{ij}-x_j^{(1)}$. Then, rewrite (B49) to:

$$l\left(\hat{\boldsymbol{\beta}}\right)=\prod_{Y_i=1}\left(\sum_{j=1}^{J}z_{ij}^{(1)}\hat{\beta}_j\right)\prod_{\substack{Y_i=0,\\ x_{ij}\neq x_j^{(1)}}}\left(1-\sum_{j=1}^{J}z_{ij}^{(1)}\hat{\beta}_j\right). \tag{B50}$$

By the second boundary vector $\mathbf{t}^{(2)}=\mathbf{x}^{(2)}-\mathbf{x}^{(1)}$, the likelihood function (B50) is rewritten as:

$$l\left(\hat{\boldsymbol{\beta}}\right)=\left\{\prod_{Y_i=1}\left(\sum_{j=1}^{J}z_{ij}^{(1)}\hat{\beta}_j\right)\prod_{\substack{Y_i=0,\\ x_{ij}\neq x_j^{(1)},\\ x_{ij}\neq x_j^{(2)}}}\left[1-\left(\sum_{j=1}^{J}z_{ij}^{(1)}\hat{\beta}_j\right)\right]\right\}\left(1-\sum_{j=1}^{J}t_j^{(2)}\hat{\beta}_j\right) \tag{B51}$$

Since $1-t_1^{(2)}\hat{\beta}_1-\sum_{j=2}^{J}t_j^{(2)}\hat{\beta}_j=1$ and $t_1^{(2)}\hat{\beta}_1+\sum_{j=2}^{J}t_j^{(2)}\hat{\beta}_j=0$, rearrange it as:

$$\hat{\beta}_1=-\sum_{j=2}^{J}\frac{t_j^{(2)}}{t_1^{(2)}}\hat{\beta}_j \tag{B52}$$

Substitute (B52) into (B51). It is reformed as:

$$\begin{aligned}l\left(\hat{\boldsymbol{\beta}}\right)&=\prod_{Y_i=1}\left(-\sum_{j=2}^{J}\frac{t_j^{(2)}z_{i1}^{(1)}}{t_1^{(2)}}\hat{\beta}_1+\sum_{j=2}^{J}z_{ij}^{(1)}\hat{\beta}_j\right)\prod_{\substack{Y_i=0,x_{ij}\neq x_j^{(1)},\\ x_{ij}\neq x_j^{(2)}}}\left[1-\left(-\sum_{j=2}^{J}\frac{t_j^{(2)}z_{i1}^{(1)}}{t_1^{(2)}}\hat{\beta}_1+\sum_{j=2}^{J}z_{ij}^{(1)}\hat{\beta}_j\right)\right]\\&=\prod_{Y_i=1}\left[\sum_{j=2}^{J}\left(z_{ij}^{(1)}-\frac{t_j^{(2)}}{t_1^{(2)}}z_{i1}^{(1)}\right)\hat{\beta}_j\right]\prod_{\substack{Y_i=0,x_{ij}\neq x_j^{(1)},\\ x_{ij}\neq x_j^{(2)}}}\left\{1-\left[\sum_{j=2}^{J}\left(z_{ij}^{(1)}-\frac{t_j^{(2)}}{t_1^{(2)}}z_{i1}^{(1)}\right)\hat{\beta}_j\right]\right\}\end{aligned} \tag{B53}$$

Let $z_{ij}^{(2)}=z_{ij}^{(1)}-\frac{t_j^{(2)}}{t_1^{(2)}}z_{i1}^{(1)}$. Then, rewrite (B53) as:

$$l\left(\hat{\boldsymbol{\beta}}\right)=\prod_{Y_i=1}\left(\sum_{j=2}^{J}z_{ij}^{(2)}\hat{\beta}_j\right)\prod_{\substack{Y_i=0,x_{ij}\neq x_j^{(1)},\\ x_{ij}\neq x_j^{(2)}}}\left(1-\sum_{j=2}^{J}z_{ij}^{(2)}\hat{\beta}_j\right) \tag{B54}$$

By repeating the previous step on the rest of each $R-2$ boundary vectors, the likelihood function is restructured by

$$\hat{\beta}_r=-\sum_{j=r+1}^{J}\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\hat{\beta}_j,\ \text{for } r=1,2,\ldots R-1 \tag{B55}$$

to:

$$l\left(\hat{\boldsymbol{\beta}}\right)=\prod_{Y_i=1}\left(\sum_{j=R}^{J}z_{ij}^{(R)}\hat{\beta}_j\right)\prod_{\substack{Y_i=0,x_{ij}\neq x_j^{(1)},\\ x_{ij}\neq x_j^{(2)},\cdots\\ x_{ij}\neq x_j^{(R)}}}\left(1-\sum_{j=R}^{J}z_{ij}^{(R)}\hat{\beta}_j\right). \tag{B56}$$

After the reparameterization procedure, the reformed likelihood function is:

$$l\left(\hat{\boldsymbol{\beta}}\right)=\prod_{Y_i=1}\mu_i^{*}\prod_{\substack{Y_i=0,x_{ij}\neq x_j^{(1)},\\ x_{ij}\neq x_j^{(2)},\cdots\\ x_{ij}\neq x_j^{(R)}}}1-\mu_i^{*} \tag{B57}$$

where $\mu_i^{*}=\sum_{j=R}^{J}z_{ij}^{(R)}\hat{\beta}_j$.

Theorem B5

For an identity-link binomial model with $J\geq 1$ independent non-constant covariates fitted by the exact method to data having R $(1\leq R\leq J)$ distinct sets of values of the covariates (including the constant) for which the estimated outcome probability evaluated at the ML solution is zero, the estimated coefficients and the corresponding standard errors of the $R-1$ eliminated non-constant are given by:

$$\hat{\beta}_r=-\sum_{j=r+1}^{J}\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\hat{\beta}_j,\ \text{for } r=1,2,\ldots R-1\,. \tag{B58}$$

and

$$\widehat{\mathrm{SE}}\left(\hat{\beta}_r\right)=\sqrt{\sum_{j=r+1}^{J}\left(\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)+\sum_{\substack{j_1=r+1\\ j_1\neq j_2}}^{J}\sum_{j_2=r+1}^{J}\left\{\frac{t_{j_1}^{(r+1)}t_{j_2}^{(r+1)}}{\left(t_r^{(r+1)}\right)^2}\right\}\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)} \tag{B59}$$

and the estimated covariances between the exact estimates of the coefficients of the $R-1$ eliminated non-constant covariates and the exact estimates of the coefficients of the $J-R+1$ remaining non-constant covariates are given by:

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_r,\hat{\beta}_s\right)=-\left\{\frac{t_s^{(r+1)}}{t_r^{(r+1)}}\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right)+\sum\nolimits_{\substack{j=r+1\\ j\neq s}}^{J}\left[\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_s\right)\right]\right\}, \tag{B60}$$

for $r=1,2,...R-1$ and $s=r+1,\ldots J$.

Proof of Theorem B5

The formula (B58) has been proved through the proof of Theorem 4. The proof is focusing on the (B59) and (B60).

Since $\hat{\beta}_r=-\sum_{j=r+1}^{J}\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\hat{\beta}_j$, we get the variance function of $\hat{\beta}_r$ as:

$$\widehat{\mathrm{Var}}\left(\hat{\beta}_r\right)=\widehat{\mathrm{Var}}\left(-\sum_{j=r+1}^{J}\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\hat{\beta}_j\right) \tag{B61}$$

Expanding the (B61):

$$\widehat{\mathrm{Var}}\left(\hat{\beta}_r\right)=\sum_{j=r+1}^{J}\left(\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)+\sum_{\substack{j_1=r+1\\ j_1\neq j_2}}^{J}\sum_{j_2=r+1}^{J}\left[\frac{t_{j_1}^{(r+1)}t_{j_2}^{(r+1)}}{\left(t_r^{(r+1)}\right)^2}\right]\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right) \qquad \text{(B62)}$$

The formula of standard error for the first $R-1$ estimate of non-constant eliminating coefficients that did not achieve from the fitting procedure of model is obtained by square rooting both sides of (B62):

$$\widehat{\mathrm{SE}}\left(\hat{\beta}_r\right)=\sqrt{\sum_{j=r+1}^{J}\left(\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)+\sum_{\substack{j_1=r+1\\ j_1\neq j_2}}^{J}\sum_{j_2=r+1}^{J}\left[\frac{t_{j_1}^{(r+1)}t_{j_2}^{(r+1)}}{\left(t_r^{(r+1)}\right)^2}\right]\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)}. \qquad \text{(B63)}$$

To obtain the formula of covariance between the first $R-1$ estimate of non-constant eliminating coefficients that did not achieve from the fitting procedure of model and the rest of the coefficients in the model, we begin with restructuring (B58):

$$t_r^{(r+1)}\hat{\beta}_r=-\sum\nolimits_{j=r+1}^{J}\hat{\beta}_j t_j^{(r+1)},\ r=1,2,\ldots R-1 \qquad \text{(B64)}$$

and adding each side of the equation the estimated coefficient of one of the $J-R+1$ covariates remaining in the model:

$$t_r^{(r+1)}\hat{\beta}_r+\hat{\beta}_s=-\left[\left(t_s^{(r+1)}-1\right)\hat{\beta}_s+\sum\nolimits_{\substack{j=r+1\\ j\neq s}}^{J}\hat{\beta}_j t_j^{(r+1)}\right],\ r=1,2,\ldots,R-1;\ s=r+1,\ldots J$$

Thus:

$$\begin{aligned}&\widehat{\mathrm{Var}}\left(t_r^{(r+1)}\hat{\beta}_r+\hat{\beta}_s\right)\\&\quad=\widehat{\mathrm{Var}}\left\{-\left[\left(t_s^{(r+1)}-1\right)\hat{\beta}_s+\sum_{\substack{j=r+1\\ j\neq s}}^{J}\hat{\beta}_j t_j^{(r+1)}\right]\right\},\ r=1,2,\ldots,R-1;\ s=r+1,\ldots J\end{aligned} \qquad \text{(B65)}$$

Expanding the left-hand side of (B65):

$$\widehat{\mathrm{Var}}\left(t_r^{(r+1)}\hat{\beta}_r+\hat{\beta}_s\right)=\left(t_r^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_r\right)+2t_r^{(r+1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_r,\hat{\beta}_s\right)+\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right) \qquad \text{(B66)}$$

Expanding the right-hand side of (B26):

$$\begin{aligned}
&\widehat{\mathrm{Var}}\left\{-\left[\left(t_s^{(r+1)}-1\right)\hat{\beta}_s+\sum_{\substack{j=r+1\\ j\neq s}}^{J}\hat{\beta}_j t_j^{(r+1)}\right]\right\}\\
&=\left(t_s^{(r+1)}-1\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right)+2\sum_{\substack{j=r+1\\ j\neq s}}^{J}\left[t_j^{(r+1)}\left(t_s^{(r+1)}-1\right)\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_s\right)\right]+\widehat{\mathrm{Var}}\left[\sum_{\substack{j=r+1\\ j\neq s}}^{J}\hat{\beta}_j t_j^{(r+1)}\right]\\
&=\left[\left(t_s^{(r+1)}\right)^2-2t_s^{(r+1)}+1\right]\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right)+2\sum_{\substack{j=r+1\\ j\neq s}}^{J}\left[\left(t_j^{(r+1)}t_s^{(r+1)}-t_j^{(r+1)}\right)\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_s\right)\right]\\
&+\sum_{\substack{j=r+1\\ j\neq s}}^{J}\left[\left(t_j^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\right]+\sum_{j_1=r+1}^{J}\sum_{\substack{j_2=r+1\\ j_1,j_2\neq s\\ j_2\neq j_1}}^{J}\left[\left(t_{j_1}^{(r+1)}t_{j_2}^{(r+1)}\right)\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)\right]\\
&=\left(1-2t_s^{(r+1)}\right)\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right)-2\sum_{\substack{j=r+1\\ j\neq s}}^{J}\left[t_j^{(r+1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_s\right)\right]\\
&+\sum_{j=r+1}^{J}\left[\left(t_j^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\right]+\sum_{j_1=r+1}^{J}\sum_{\substack{j_2=r+1\\ j_2\neq j_1}}^{J}\left[\left(t_{j_1}^{(r+1)}t_{j_2}^{(r+1)}\right)\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)\right]
\end{aligned}\tag{B67}$$

By expanding $\widehat{\mathrm{Var}}\left[\hat{\beta}_r t_r^{(r+1)}\right]=\widehat{\mathrm{Var}}\left[-\sum_{j=r+1}^{J}\hat{\beta}_j t_j^{(r+1)}\right]$, $r=1,2,\ldots R-1$, we have:

$$\left(t_r^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_r\right)=\sum_{j=r+1}^{J}\left[\left(t_j^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\right]+\sum_{j_1=r+1}^{J}\sum_{\substack{j_2=r+1\\ j_2\neq j_1}}^{J}\left[\left(t_{j_1}^{(r+1)}t_{j_2}^{(r+1)}\right)\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)\right]\tag{B68}$$

Substituting for $\left(t_r^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_r\right)$ in (B67):

$$\begin{aligned}
&\widehat{\mathrm{Var}}\left\{-\left[\left(t_s^{(r+1)}-1\right)\hat{\beta}_s+\sum_{\substack{j=r+1\\ j\neq s}}^{J}\hat{\beta}_j t_j^{(r+1)}\right]\right\}\\
&\quad=\left(1-2t_s^{(r+1)}\right)\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right)+\left(t_r^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_r\right)-2\sum_{\substack{j=r+1\\ j\neq s}}^{J}t_j^{(r+1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_s\right)
\end{aligned}\tag{B69}$$

Equating (B66) and (B69), the two sides of (B65), and simplifying:

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_r,\hat{\beta}_s\right)=-\left\{\frac{t_s^{(r+1)}}{t_r^{(r+1)}}\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right)+\sum_{\substack{j=r+1\\ j\neq s}}^{J}\left[\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_s\right)\right]\right\},$$

for $r=1,2,\ldots R-1$ and $s=r+1,\ldots J$

<u>*Theorem* B6</u>

For an identity-link binomial model with $J\geq 1$ independent non-constant covariates fitted by the exact method to data having R $(1\leq R\leq J)$ distinct sets of values of the covariates (including the constant) for which the estimated outcome probability evaluated at the ML solution is zero, the estimated coefficient and the corresponding standard error of the constant covariate are given by:

$$\hat{\beta}_0 = -\sum_{j=1}^{J} x_j^{(1)} \hat{\beta}_j \tag{B70}$$

and

$$\widehat{\mathrm{SE}}\left(\hat{\beta}_0\right) = \sqrt{\sum_{j=1}^{J}\left(x_j^{(1)}\right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=1 \\ j_1 \neq j_2}}^{J} \sum_{j_2=1}^{J} x_{j_1}^{(1)} x_{j_2}^{(1)} \widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1}, \hat{\beta}_{j_2}\right)} \tag{B71}$$

and the estimated covariances between the exact estimate of the coefficient of the constant covariate and the exact estimates of the coefficients of the J non-constant covariates are given by:

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_0, \hat{\beta}_j\right) = -\left[x_j^{(1)} \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=1 \\ j_1 \neq j}}^{J} x_{j_1}^{(1)} \widehat{\mathrm{Cov}}\left(\hat{\beta}_j, \hat{\beta}_{j_1}\right) \right] \tag{B72}$$

for $j = 1, 2, \ldots J$.

Proof of Theorem B6

The formula (B70) has been proved through the proof of Theorem 1. The proof is focusing on the (B71) and (B72).

Since $\hat{\beta}_0 = -\sum_{j=1}^{J} x_j^{(1)} \hat{\beta}_j$, we get the variance function of $\hat{\beta}_0$ as:

$$\begin{aligned} \widehat{\mathrm{Var}}\left(\hat{\beta}_0\right) &= \widehat{\mathrm{Var}}\left(-\sum_{j=1}^{J} x_j^{(1)} \hat{\beta}_j \right) \\ &= \widehat{\mathrm{Var}}\left(\sum_{j=1}^{J} x_j^{(1)} \hat{\beta}_j \right) \end{aligned} \tag{B73}$$

Expanding the (B73):

$$\widehat{\mathrm{Var}}\left(\hat{\beta}_0\right) = \sum_{j=1}^{J}\left(x_j^{(1)}\right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=1 \\ j_1 \neq j_2}}^{J} \sum_{j_2=1}^{J} x_{j_1}^{(1)} x_{j_2}^{(1)} \widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1}, \hat{\beta}_{j_2}\right) \tag{B74}$$

The formula of standard error for the estimate of constant is obtained by square rooting both sides of (B74):

$$\widehat{\mathrm{SE}}\left(\hat{\beta}_0\right) = \sqrt{\sum_{j=1}^{J}\left(x_j^{(1)}\right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=1 \\ j_1 \neq j_2}}^{J} \sum_{j_2=1}^{J} x_{j_1}^{(1)} x_{j_2}^{(1)} \widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1}, \hat{\beta}_{j_2}\right)} \tag{B75}$$

To obtain the formula of covariance between the estimate of constant and the estimate of coefficients in the model, we begin with adding a $\hat{\beta}_j$ to both sides of $\hat{\beta}_0 = -\sum_{j=1}^{J} x_j^{(1)} \hat{\beta}_j$:

$$\hat{\beta}_0 + \hat{\beta}_j = -\left[\left(x_j^{(1)} - 1\right)\hat{\beta}_j + \sum_{\substack{j_1=1 \\ j_1 \neq j}}^{J} \hat{\beta}_{j_1} x_{j_1}^{(1)} \right], \quad j = 1, 2, \ldots J \tag{B76}$$

Thus:

$$\begin{aligned}\widehat{\mathrm{Var}}\left(\hat{\beta}_0+\hat{\beta}_j\right)&=\widehat{\mathrm{Var}}\left\{-\left[\left(x_j^{(1)}-1\right)\hat{\beta}_j+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}\hat{\beta}_{j_1}x_{j_1}^{(1)}\right]\right\}\\&=\widehat{\mathrm{Var}}\left[\left(x_j^{(1)}-1\right)\hat{\beta}_j+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}\hat{\beta}_{j_1}x_{j_1}^{(1)}\right]\end{aligned}\tag{B77}$$

Expanding the left-hand side of (B77):

$$\widehat{\mathrm{Var}}\left(\hat{\beta}_0+\hat{\beta}_j\right)=\widehat{\mathrm{Var}}\left(\hat{\beta}_0\right)+2\widehat{\mathrm{Cov}}\left(\hat{\beta}_0,\hat{\beta}_j\right)+\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\tag{B78}$$

Expanding the right-hand side of (B77):

$$\begin{aligned}&\widehat{\mathrm{Var}}\left[\left(x_j^{(1)}-1\right)\hat{\beta}_j+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}\hat{\beta}_{j_1}x_{j_1}^{(1)}\right]\\&=\left(x_j^{(1)}-1\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)+2\left(x_j^{(1)}-1\right)\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}x_{j_1}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_{j_1}\right)+\widehat{\mathrm{Var}}\left(\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}\hat{\beta}_{j_1}x_{j_1}^{(1)}\right)\end{aligned}\tag{B79}$$

Since:

$$\widehat{\mathrm{Var}}\left(\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}\hat{\beta}_{j_1}x_{j_1}^{(1)}\right)=\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}\left[\left(x_{j_1}^{(1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_{j_1}\right)\right]+\sum_{\substack{j_2=1\\ j_2\neq j}}^{J}\sum_{\substack{j_3=1\\ j_3\neq j\\ j_3\neq j_2}}^{J}\left[x_{j_2}^{(1)}x_{j_3}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_2},\hat{\beta}_{j_3}\right)\right],\tag{B80}$$

expand (B79) by (B80) as:

$$\begin{aligned}&\widehat{\mathrm{Var}}\left[\left(x_j^{(1)}-1\right)\hat{\beta}_j+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}\hat{\beta}_{j_1}x_{j_1}^{(1)}\right]\\&=\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)-2\left[x_j^{(1)}\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}x_{j_1}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_{j_1}\right)\right]+\left(x_j^{(1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\\&+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}\left[\left(x_{j_1}^{(1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_{j_1}\right)\right]+2\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}x_j^{(1)}x_{j_1}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_{j_1}\right)\\&+\sum_{\substack{j_2=1\\ j_2\neq j}}^{J}\sum_{\substack{j_3=1\\ j_3\neq j\\ j_3\neq j_2}}^{J}\left[x_{j_2}^{(1)}x_{j_3}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_2},\hat{\beta}_{j_3}\right)\right]\\&=\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)-2\left[x_j^{(1)}\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}x_{j_1}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_{j_1}\right)\right]\\&+\sum_{j=1}^{J}\left[\left(x_j^{(1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\right]+\sum_{j_1=1}^{J}\sum_{\substack{j_2=1\\ j_1\neq j_2}}^{J}\left[x_{j_1}^{(1)}x_{j_2}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)\right]\end{aligned}\tag{B81}$$

Replacing and simplifying (B81) by (B74):

$$\begin{aligned}&\widehat{\mathrm{Var}}\left[\left(x_j^{(1)}-1\right)\hat{\beta}_j+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}\hat{\beta}_{j_1}x_{j_1}^{(1)}\right]\\&\quad=\widehat{\mathrm{Var}}\left(\hat{\beta}_0\right)+\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)-2\left[x_j^{(1)}\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}x_{j_1}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_{j_1}\right)\right]\end{aligned}\tag{B82}$$

Substituting (B78) and (B82) into (B77), and simplifying as:

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_0,\hat{\beta}_j\right)=-\left[x_j^{(1)}\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)+\sum_{\substack{j_1=1\\ j_1\neq j}}^{J}x_{j_1}^{(1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_{j_1}\right)\right],\quad j=1,2,\ldots J\tag{B83}$$

Table B1: Simulation results (with results that the standard error is not missing for the additive binomial method shown for comparison)

Setting		All simulations (n = 20,000)			Simulations with a boundary vector							
		Exact method			Exact method				Additive binomial[§]			
		Bias[*]	MSE[†]	Coverage[‡]	n	Bias[*]	MSE[†]	Coverage[‡]	n	Bias[*]	MSE[†]	Coverage[‡]
1	β_0	–0.482	0.136	95.5	10,683	0.100	0.122	94.8	4,636	0.475	0.115	92.8
	β_1	–6.957	0.108	95.1	10,683	–2.334	0.101	94.7	4,636	–7.664	0.101	92.9
	β_2	–1.248	0.002	97.0	10,683	–0.490	0.001	96.6	4,636	–1.126	0.001	96.3
2	β_0	–0.465	0.135	95.6	10,662	0.103	0.119	95.3	9,428	0.142	0.115	93.9
	β_1	–7.086	0.107	95.4	10,662	–2.224	0.098	95.2	9,428	–4.482	0.096	94.1
	β_2	–1.245	0.002	97.2	10,662	–0.485	0.001	97.0	9,428	–0.597	0.001	96.1
3	β_0	–0.720	0.321	95.7	10,739	–0.143	0.266	95.1	4,773	–0.084	0.246	95.2
	β_1	–7.067	0.108	95.0	10,739	–1.546	0.099	94.9	4,773	–7.231	0.098	93.5
	β_2	–1.248	0.002	96.8	10,739	–0.547	0.001	96.5	4,773	–1.149	0.002	96.9
4	β_0	–0.665	0.317	95.6	10,579	–0.103	0.267	94.8	9,349	–0.121	0.262	93.9
	β_1	–7.987	0.108	94.9	10,579	–2.087	0.101	94.3	9,349	–4.280	0.099	93.0
	β_2	–1.187	0.002	97.0	10,579	–0.509	0.001	96.5	9,349	–0.630	0.001	95.8
5	β_0	–0.774	0.598	96.0	10,262	–0.087	0.478	95.0	4,738	–0.257	0.464	95.2
	β_1	–7.994	0.108	95.1	10,262	–3.210	0.101	94.7	4,738	–8.782	0.101	93.5
	β_2	–1.167	0.002	97.0	10,262	–0.405	0.001	96.4	4,738	–1.056	0.002	96.5
6	β_0	–0.798	0.599	96.3	10,285	–0.114	0.473	95.5	9,052	–0.168	0.464	94.7
	β_1	–7.647	0.108	95.3	10,285	–2.889	0.100	94.7	9,052	–4.914	0.097	93.3
	β_2	–1.181	0.002	97.0	10,285	–0.422	0.001	96.5	9,052	–0.553	0.001	95.7
7	β_0	–0.923	0.983	96.2	11,744	–0.268	0.765	95.8	4,630	–0.462	0.755	95.7
	β_1	–7.413	0.108	94.8	11,744	–2.717	0.100	94.8	4,630	–8.359	0.100	93.3
	β_2	–1.236	0.002	96.9	11,744	–0.480	0.001	96.7	4,630	–1.114	0.002	96.5
8	β_0	–0.910	0.977	96.3	11,567	–0.265	0.771	95.9	10,056	–0.361	0.767	95.4
	β_1	–7.655	0.108	94.9	11,567	–2.759	0.100	94.9	10,056	–4.592	0.097	93.9
	β_2	–1.222	0.002	97.1	11,567	–0.482	0.001	96.7	10,056	–0.641	0.001	96.1

[*] Average value of percent small-sample bias: $(100/n)\sum_{k=1}^{n}\left(\hat{\beta}_{jk}-\beta_j\right)\Big/\beta_j, \quad j=0,1,2; k=1,2,...n$

[†] Average value of mean squared error: $(100/n)\sum_{k=1}^{n}\left[\left(\hat{\beta}_{jk}-\beta_j\right)^2+\widehat{\mathrm{Var}}\left(\hat{\beta}_{jk}\right)\right], \quad j=0,1,2; k=1,2,...n$.

[‡] Percentage of 95 percent confidence intervals covering the design value of the coefficient.

[§] The simulation results with a boundary vector in the additive binomial only include the samples that the standard error was reported.

Table B2: Simulation results

Setting		Simulations with a boundary vector on the lower bound				Simulations with a boundary vector on the upper bound			
		n	Bias*	MSE†	Coverage‡	n	Bias*	MSE†	Coverage‡
1	β_0	5167	–0.264	0.131	96.7	5516	0.441	0.113	93.1
	β_1	5167	0.956	0.100	96.0	5516	–5.415	0.101	93.5
	β_2	5167	–0.094	0.001	97.2	5516	–0.861	0.001	96.0
2	β_0	5188	–0.256	0.128	97.1	5474	0.442	0.111	93.6
	β_1	5188	1.255	0.097	96.2	5474	–5.521	0.099	94.2
	β_2	5188	–0.082	0.001	97.7	5474	–0.868	0.001	96.3
3	β_0	5091	–0.291	0.298	95.7	5648	–0.009	0.237	94.5
	β_1	5091	2.566	0.100	95.5	5648	–5.253	0.099	94.2
	β_2	5091	–0.177	0.001	96.8	5648	–0.881	0.001	96.4
4	β_0	5054	–0.264	0.294	96.4	5525	0.044	0.242	93.4
	β_1	5054	2.397	0.099	95.3	5525	–6.189	0.101	93.3
	β_2	5054	–0.150	0.001	97.3	5525	–0.838	0.001	95.8
5	β_0	4645	–0.024	0.523	95.9	5617	–0.139	0.440	94.2
	β_1	4645	0.901	0.101	95.6	5617	–6.610	0.101	94.0
	β_2	4645	0.062	0.001	96.9	5617	–0.791	0.001	96.1
6	β_0	4728	–0.065	0.517	96.1	5557	–0.155	0.426	95.0
	β_1	4728	1.025	0.101	95.3	5557	–6.219	0.100	94.1
	β_2	4728	0.021	0.001	97.0	5557	–0.798	0.001	96.0
7	β_0	6240	–0.222	0.819	96.7	5504	–0.319	0.705	94.8
	β_1	6240	0.186	0.100	95.4	5504	–6.008	0.099	94.2
	β_2	6240	–0.153	0.001	97.4	5504	–0.851	0.001	96.1
8	β_0	6048	–0.241	0.820	96.7	5519	–0.292	0.717	94.9
	β_1	6048	0.529	0.099	95.6	5519	–6.362	0.101	94.1
	β_2	6048	–0.172	0.001	97.2	5519	–0.823	0.001	96.2

* Average value of percent bias: $(100/n)\sum_{k=1}^{n}\left[\left(\hat{\beta}_{jk}-\beta_j\right)/\beta_j\right]$, $j=0,1,2; k=1,2,...n$

† Average value of mean squared error: $(100/n)\sum_{k=1}^{n}\left[\left(\hat{\beta}_{jk}-\beta_j\right)^2+\widehat{\mathrm{Var}}\left(\hat{\beta}_{jk}\right)\right]$ for $j=0,1,2$ and $k=1,2,\ldots n$.

‡ Percentage of 95 percent confidence intervals covering the design value of the coefficient.

Appendix

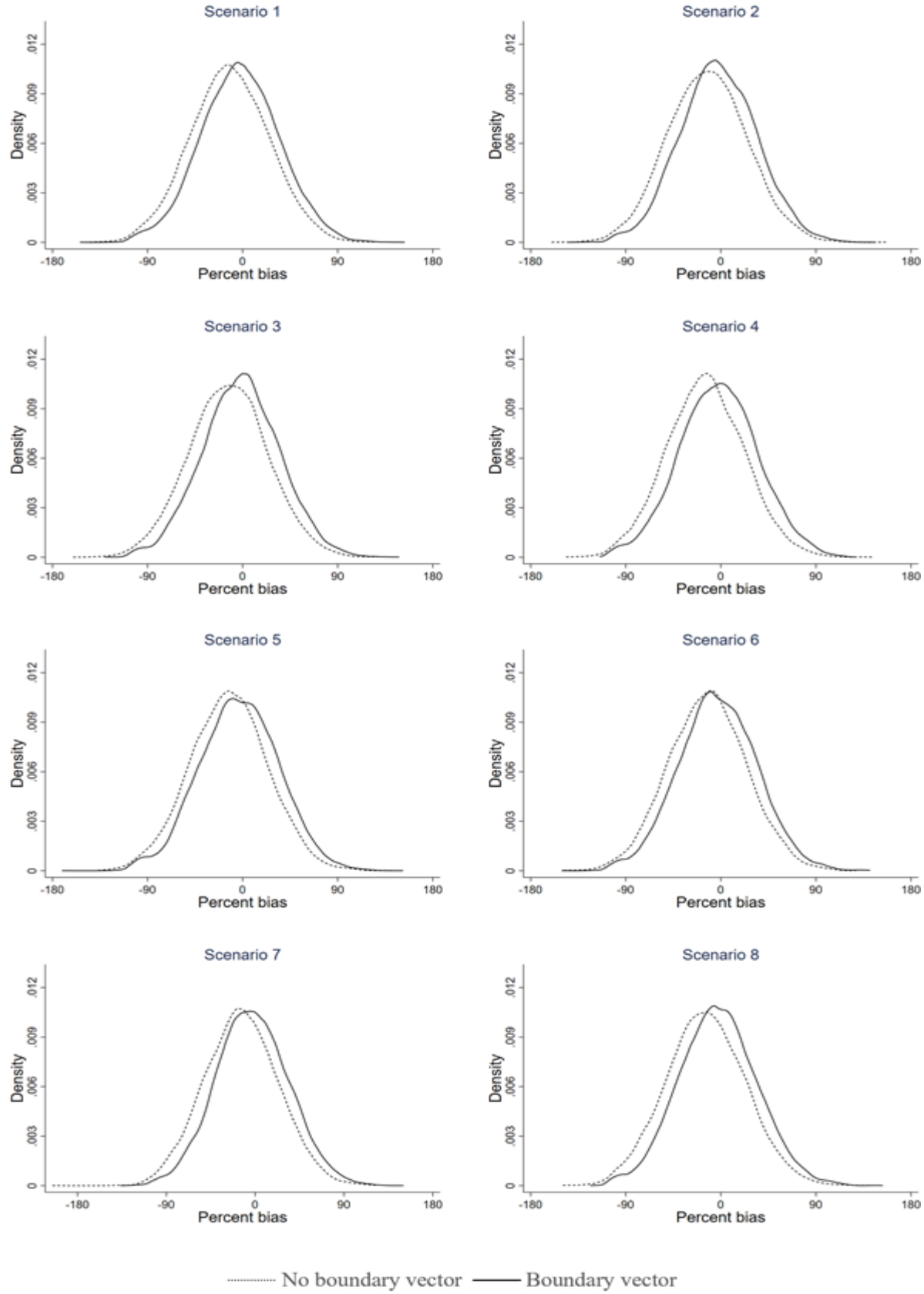

Figure B1: Density plots of the sampling distributions (n=1000, 20000 replications, 8 settings) of the binary covariate β_1. The solid lines represent replications with a boundary vector. The dotted lines represent replications without a boundary vector.

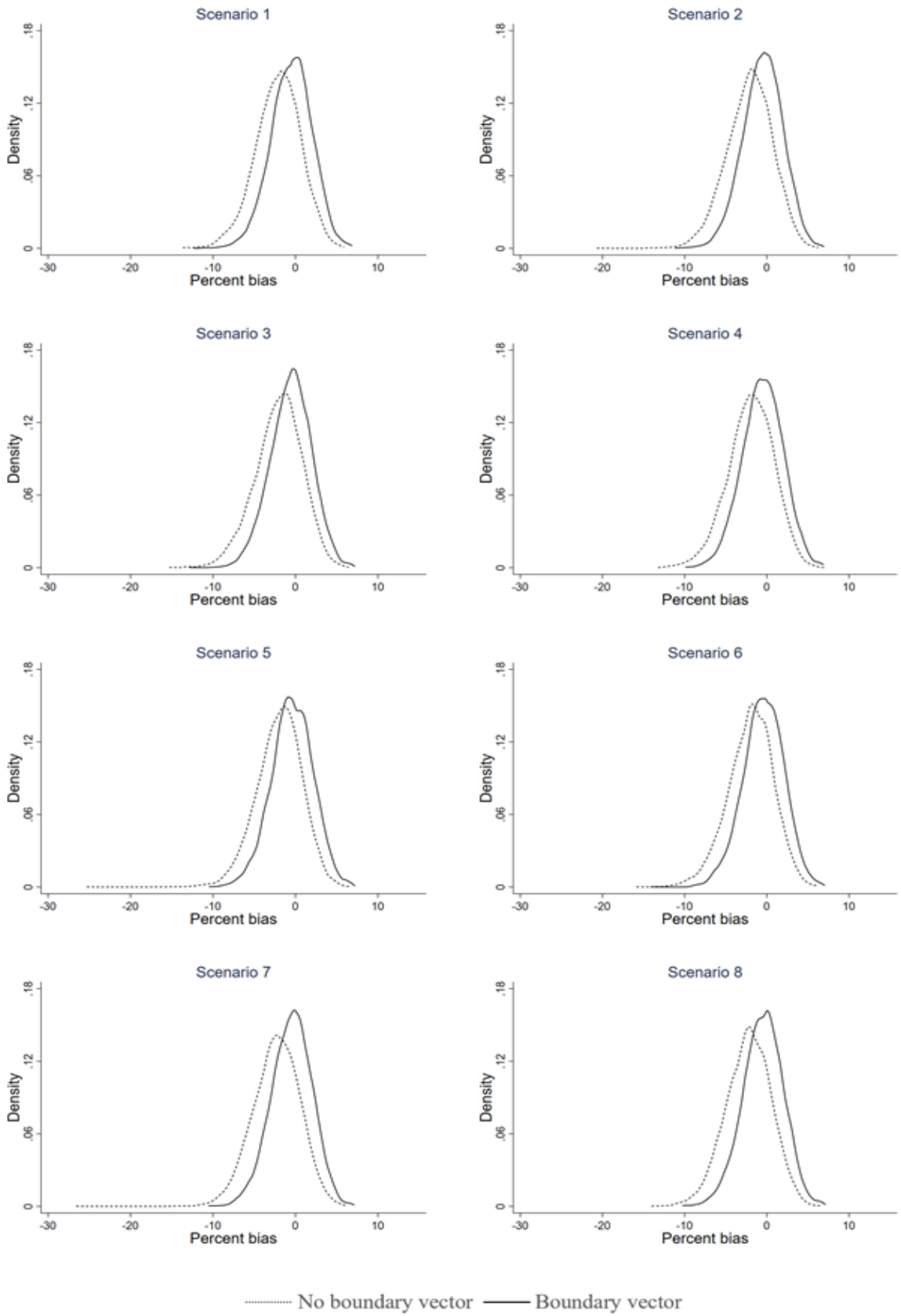

Figure B2: Density plots of the sampling distributions (n=1000, 20000 replications, 8 settings) of the binary covariate β_2. The solid lines represent replications with a boundary vector. The dotted lines represent replications without a boundary vector.

Appendix C

Supplementary materials of the marginal log binomial model estimated by the generalised estimating equation using the exact method

The implementation of the exact method and the proofs of Theorem C1 and C2

Suppose that there are K clusters/subjects in data. There are n_i observations included in the cluster $i,\ i=1,2,...K$. We denote a response variable of p^{th} observation in the cluster i as $y_{ip},\ p=1,2,...n_i$. The relationship between the mean μ_{ip} of y_{ip} and the linear combination of covariates vector $\mathbf{x}_{ip}=\left(x_{ip1},x_{ip2},...x_{ipJ}\right)$ and the marginal model regression parameters $\boldsymbol{\beta}=\left(\beta_0,\beta_1,\beta_2,...\beta_J\right)$ is written as:

$$\log\left(\mu_{ip}\right)=\mathbf{x}_{ip}'\boldsymbol{\beta}=\beta_0+\beta_1x_{ip1}+\beta_2x_{ip2}+...+\beta_Jx_{ipJ}\,.$$

Assume that there are $R\ \left(0<R\le J\right)$ distinct covariate vectors of observations in data with the estimated mean evaluated at the estimates $\hat{\boldsymbol{\beta}}$ equal to 1, which are defined as the boundary vectors. Denote the $r^{\text{th}}\ \left(r=1,2,...R\right)$ boundary vector as $\mathbf{x}^{(r)}$, which shares the covariate values with n_r observations. The fitting procedure of the exact method can be summarised in seven steps:

1. Eliminating the constant by subtracting from the constant and each non-constant covariate its respective value in the boundary vector:

$$z_{ijp}^{(1)}=x_{ijp}-x_j^{(1)},\ \ j=0,1,2,...J\ ,$$

where *i* denotes cluster, *j* denotes the position of a covariate, and *p* denotes the position in the cluster.

2. When there are multiple boundary vectors $(R>1)$, eliminating the first $R-1$ non-constant covariates by re-parametrising the covariates according to the scheme:

$$z_{ijp}^{(r)}=z_{ijp}^{(r-1)}-\left(\frac{t_j^{(r)}}{t_{r-1}^{(r)}}\right)z_{ip,r-1}^{(r-1)}\ \ \text{where}\ t_j^{(r)}=z_{ijp}^{(r-1)}\Big|_{x_{ijp}=x_j^{(r)}},\quad r=2,3,...R$$

3. Drop the observations with covariate vectors $\mathbf{x}^{(1)},\mathbf{x}^{(2)},...\mathbf{x}^{(R)}$ respectively, which make no contribution to the estimation, and fitting the model $\mu_{ip}=\exp\left(\sum_{j=R}^{J}\beta_jz_{ijp}^{(R)}\right)$ without a constant and with $J-R+1$ covariates to the remaining $n-n_1-n_2-...-n_R$ observations to obtain the estimates $\hat{\beta}_R,\hat{\beta}_{R+1},...\hat{\beta}_J$ of the coefficients of the non-eliminated non-constant covariates and their estimated variances $\widehat{\text{Var}}\left(\hat{\beta}_j\right),\ j=R,R+1,...J$ and covariances $\widehat{\text{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right),\ j_1\ne j_2,\ j_1$ and $j_2=R,R+1,...J$.

4. Estimate the coefficients $\hat{\beta}_r,\ r=1,2,...R-1$ of the $R-1$ eliminated covariates as:

$$\hat{\beta}_r=-\frac{\sum_{j=r+1}^{J}\hat{\beta}_jt_j^{(r+1)}}{t_r^{(r+1)}}$$

5. Estimate the standard error of the estimates of the $R-1$ eliminated covariates as:

$$\widehat{\text{SE}}\left(\hat{\beta}_r\right)=\sqrt{\sum_{j=r+1}^{J}\left\{\left[\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\right]^2\widehat{\text{Var}}\left(\hat{\beta}_j\right)\right\}+\sum_{j_1=r+1}^{J}\sum_{\substack{j_2=r+1\\j_2\ne j_1}}^{J}\left(\left\{\frac{t_{j_1}^{(r+1)}t_{j_2}^{(r+1)}}{\left[t_r^{(r+1)}\right]^2}\right\}\widehat{\text{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)\right)},$$

for $r = 1, 2, ...R-1$, where $\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)$ denotes the estimate of variance for $\hat{\beta}_j$, and $\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1}, \hat{\beta}_{j_2}\right)$ denotes the estimated covariance between the estimated coefficients $\hat{\beta}_{j_1}$ and $\hat{\beta}_{j_2}$;

6. Estimate the coefficient of the constant covariate from the boundary condition as:

$$\hat{\beta}_0 = -\sum_{j=1}^{J} \hat{\beta}_j x_j^{(1)}$$

7. Estimate the standard error of constant as:

$$\widehat{SE}\left(\hat{\beta}_0\right) = \sqrt{\sum_{j=1}^{J}\left\{\left[x_j^{(1)}\right]^2 \widehat{Var}\left(\hat{\beta}_j\right)\right\} + \sum_{j_1=1}^{J}\sum_{\substack{j_2=1 \\ j_2 \neq j_1}}^{J}\left[x_{j_1}^{(1)} x_{j_2}^{(1)} \widehat{Cov}\left(\hat{\beta}_{j_1}, \hat{\beta}_{j_2}\right)\right]}.$$

The derivation and proofs of the formulas involved in those seven steps were available in the paper by Petersen and Deddens [9]. However, the formula to estimate the covariances of the estimates of coefficients $\hat{\beta}_r$, $r = 0, 1, 2, ...R-1$ were missing. We provide the following two theorems to fix the deficiency.

Theorem C1

For a marginal log binomial model estimated by generalised estimating equations with $J \geq 1$ independent non-constant covariates fitted by the exact method to data having R ($1 \leq R \leq J$) distinct sets of values of the covariates (including the constant) for which the estimated outcome probability evaluated at the solution is unity, the estimated covariances between the estimates of the coefficients of the $R-1$ eliminated non-constant covariates and the estimates of the coefficients of the other non-constant covariates are given by:

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_r, \hat{\beta}_s\right) = -\left\{\frac{t_s^{(r+1)}}{t_r^{(r+1)}}\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right) + \sum_{\substack{j=r+1 \\ j \neq s}}^{J}\left[\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j, \hat{\beta}_s\right)\right]\right\},$$

for $r = 1, 2, ...R-1$ and $s = r+1, ...J$.

Proof of Theorem C1

In successive re-parameterizations, the boundary conditions require $\sum_{j=0}^{J} \beta_j z_{ijp}^{(1)} = 0$, $\sum_{j=1}^{J} \beta_j t_j^{(2)} = 0$, $\sum_{j=2}^{J} \beta_j t_j^{(3)} = 0$ and, in general for $R > 1$:

$$\sum_{j=r}^{J} \beta_j t_j^{(r+1)} = 0, \quad r = 1, 2, ...R-1 \qquad \text{(C1)}$$

for all observations that share those covariate values. Hence:

$$t_r^{(r+1)} \hat{\beta}_r = -\sum_{j=r+1}^{J} \hat{\beta}_j t_j^{(r+1)}, \quad r = 1, 2, ...R-1 \qquad \text{(C2)}$$

Adding to each side of (C2) the estimated coefficient of one of the $J-R+1$ covariates remaining in the model:

$$t_r^{(r+1)} \hat{\beta}_r + \hat{\beta}_s = -\left[\left(t_s^{(r+1)} - 1\right)\hat{\beta}_s + \sum_{\substack{j=r+1 \\ j \neq s}}^{J} \hat{\beta}_j t_j^{(r+1)}\right], \quad r = 1, 2, ..., R-1; \; s = r+1, \ldots J$$

Thus:

$$\widehat{\mathrm{Var}}\left(t_r^{(r+1)}\hat{\beta}_r+\hat{\beta}_s\right) = \widehat{\mathrm{Var}}\left\{-\left[\left(t_s^{(r+1)}-1\right)\hat{\beta}_s+\sum_{\substack{j=r+1\\ j\neq s}}^{J}\hat{\beta}_j t_j^{(r+1)}\right]\right\},\ r=1,2,...,R-1;\ s=r+1,\ldots J \tag{C3}$$

Expanding the left-hand side of (C3):

$$\widehat{\mathrm{Var}}\left(t_r^{(r+1)}\hat{\beta}_r+\hat{\beta}_s\right)=\left(t_r^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_r\right)+2t_r^{(r+1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_r,\hat{\beta}_s\right)+\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right) \tag{C4}$$

Expanding the right-hand side of (C3):

$$\begin{aligned}
&\widehat{\mathrm{Var}}\left\{-\left[\left(t_s^{(r+1)}-1\right)\hat{\beta}_s+\sum_{\substack{j=r+1\\ j\neq s}}^{J}\hat{\beta}_j t_j^{(r+1)}\right]\right\}\\
&=\left(t_s^{(r+1)}-1\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right)+2\sum_{\substack{j=r+1\\ j\neq s}}^{J}\left\{t_j^{(r+1)}\left(t_s^{(r+1)}-1\right)\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_s\right)\right\}+\widehat{\mathrm{Var}}\left(\sum_{\substack{j=r+1\\ j\neq s}}^{J}\hat{\beta}_j t_j^{(r+1)}\right)\\
&=\left[\left(t_s^{(r+1)}\right)^2-2t_s^{(r+1)}+1\right]\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right)+2\sum_{\substack{j=r+1\\ j\neq s}}^{J}\left[\left(t_j^{(r+1)}t_s^{(r+1)}-t_j^{(r+1)}\right)\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_s\right)\right]\\
&+\sum_{\substack{j=r+1\\ j\neq s}}^{J}\left[\left(t_j^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\right]+\sum_{j_1=r+1}^{J}\sum_{\substack{j_2=r+1\\ j_1,j_2\neq s\\ j_2\neq j_1}}^{J}\left[\left(t_{j_1}^{(r+1)}t_{j_2}^{(r+1)}\right)\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)\right]\\
&=\left(1-2t_s^{(r+1)}\right)\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right)-2\sum_{\substack{j=r+1\\ j\neq s}}^{J}\left[t_j^{(r+1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_s\right)\right]\\
&+\sum_{j=r+1}^{J}\left[\left(t_j^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\right]+\sum_{j_1=r+1}^{J}\sum_{\substack{j_2=r+1\\ j_2\neq j_1}}^{J}\left[\left(t_{j_1}^{(r+1)}t_{j_2}^{(r+1)}\right)\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)\right]
\end{aligned} \tag{C5}$$

From (C2), $\widehat{\mathrm{Var}}\left(\hat{\beta}_r t_r^{(r+1)}\right)=\widehat{\mathrm{Var}}\left(-\sum_{j=r+1}^{J}\hat{\beta}_j t_j^{(r+1)}\right),\ r=1,2,...R-1$ and hence:

$$\left(t_r^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_r\right)=\sum_{j=r+1}^{J}\left[\left(t_j^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_j\right)\right]+\sum_{j_1=r+1}^{J}\sum_{\substack{j_2=r+1\\ j_2\neq j_1}}^{J}\left[\left(t_{j_1}^{(r+1)}t_{j_2}^{(r+1)}\right)\widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_1},\hat{\beta}_{j_2}\right)\right] \tag{C6}$$

Substituting for $\left(t_r^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_r\right)$ in (C5):

$$\begin{aligned}
&\widehat{\mathrm{Var}}\left\{-\left[\left(t_s^{(r+1)}-1\right)\hat{\beta}_s+\sum_{\substack{j=r+1\\ j\neq s}}^{J}\hat{\beta}_j t_j^{(r+1)}\right]\right\}\\
&=\left(1-2t_s^{(r+1)}\right)\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right)-2\sum_{\substack{j=r+1\\ j\neq s}}^{J}\left[t_j^{(r+1)}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_s\right)+\left(t_r^{(r+1)}\right)^2\widehat{\mathrm{Var}}\left(\hat{\beta}_r\right)\right]
\end{aligned} \tag{C7}$$

Equating (C4) and (C7), the two sides of (C3), and simplifying:

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_r,\hat{\beta}_s\right)=-\left\{\frac{t_s^{(r+1)}}{t_r^{(r+1)}}\widehat{\mathrm{Var}}\left(\hat{\beta}_s\right)+\sum_{\substack{j=r+1\\ j\neq s}}^{J}\left[\frac{t_j^{(r+1)}}{t_r^{(r+1)}}\widehat{\mathrm{Cov}}\left(\hat{\beta}_j,\hat{\beta}_s\right)\right]\right\},\ r=1,2,...R-1;\ s=r+1,\ldots J$$

Theorem C2

For a marginal log binomial model estimated by generalised estimating equations with $J \geq 1$ independent non-constant covariates fitted by the exact method to data having R $(1 \leq R \leq J)$ distinct sets of values of the covariates (including the constant) for which the estimated outcome probability evaluated at the solution is unity, the estimated covariances between the exact estimate of the coefficient of the constant covariate and exact estimates of the coefficients of the J non-constant covariates are given by:

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_0, \hat{\beta}_j\right) = -\left[x_j^{(1)} \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=1 \\ j_1 \neq j}}^{J} x_{j_1}^{(1)} \widehat{\mathrm{Cov}}\left(\hat{\beta}_j, \hat{\beta}_{j_1}\right) \right], \; j = 1, 2, \ldots J \, .$$

Proof of Theorem C2

Adding the estimated coefficient $\hat{\beta}_j$, $j = 1, 2, \ldots J$ of any of the non-constant covariates to each side of the $\hat{\beta}_0 = -\sum_{j=1}^{J} \hat{\beta}_j x_j^{(1)}$ yields:

$$\hat{\beta}_0 + \hat{\beta}_j = -\left[\left(x_j^{(1)} - 1 \right) \hat{\beta}_j + \sum_{\substack{j_1=1 \\ j_1 \neq j}}^{J} \hat{\beta}_{j_1} x_{j_1}^{(1)} \right], \; j = 1, 2, \ldots J$$

It follows that:

$$\widehat{\mathrm{Var}}\left(\hat{\beta}_0 + \hat{\beta}_j\right) = \widehat{\mathrm{Var}}\left\{ -\left[\left(x_j^{(1)} - 1 \right) \hat{\beta}_j + \sum_{\substack{j_1=1 \\ j_1 \neq j}}^{J} \hat{\beta}_{j_1} x_{j_1}^{(1)} \right] \right\} \tag{C8}$$

Expanding the left-hand side of (C8):

$$\widehat{\mathrm{Var}}\left(\hat{\beta}_0 + \hat{\beta}_j\right) = \widehat{\mathrm{Var}}\left(\hat{\beta}_0\right) + 2\widehat{\mathrm{Cov}}\left(\hat{\beta}_0, \hat{\beta}_j\right) + \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) \tag{C9}$$

Expanding the right-hand side of (C8):

$$\begin{aligned} &\widehat{\mathrm{Var}}\left\{ -\left[\left(x_j^{(1)} - 1 \right) \hat{\beta}_j + \sum_{\substack{j_1=1 \\ j_1 \neq j}}^{J} \hat{\beta}_{j_1} x_{j_1}^{(1)} \right] \right\} \\ &\quad = \left(x_j^{(1)} - 1 \right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + 2\left(x_j^{(1)} - 1 \right) \sum_{\substack{j_1=1 \\ j_1 \neq j}}^{J} x_{j_1}^{(1)} \widehat{\mathrm{Cov}}\left(\hat{\beta}_j, \hat{\beta}_{j_1}\right) + \widehat{\mathrm{Var}}\left(\sum_{\substack{j_1=1 \\ j_1 \neq j}}^{J} \hat{\beta}_{j_1} x_{j_1}^{(1)} \right) \end{aligned}$$

and gathering like terms:

$$\begin{aligned} &\widehat{\mathrm{Var}}\left\{ -\left[\left(x_j^{(1)} - 1 \right) \hat{\beta}_j + \sum_{\substack{j_1=1 \\ j_1 \neq j}}^{J} \hat{\beta}_{j_1} x_{j_1}^{(1)} \right] \right\} \\ &\quad = \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) - 2\left[x_j^{(1)} \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=1 \\ j_1 \neq j}}^{J} x_{j_1}^{(1)} \widehat{\mathrm{Cov}}\left(\hat{\beta}_j, \hat{\beta}_{j_1}\right) \right] + \left(x_j^{(1)} \right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) \\ &\quad + \sum_{\substack{j_1=1 \\ j_1 \neq j}}^{J} \left[\left(x_{j_1}^{(1)} \right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_{j_1}\right) \right] + 2\sum_{\substack{j_1=1 \\ j_1 \neq j}}^{J} x_j^{(1)} x_{j_1}^{(1)} \widehat{\mathrm{Cov}}\left(\hat{\beta}_j, \hat{\beta}_{j_1}\right) \\ &\quad + \sum_{\substack{j_2=1 \\ j_2 \neq j_1}}^{J} \sum_{\substack{j_3=1 \\ j_3 \neq j_1 \\ j_3 \neq j_2}}^{J} \left[x_{j_2}^{(1)} x_{j_3}^{(1)} \widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_2}, \hat{\beta}_{j_3}\right) \right] \end{aligned} \tag{C10}$$

Since $\hat{\beta}_0 = -\sum_{j_2=1}^{J} \hat{\beta}_{j_2} x_{j_2}^{(1)}$, it follows that $\widehat{\mathrm{Var}}\left(\hat{\beta}_0\right) = \widehat{\mathrm{Var}}\left[-\sum_{j_2=1}^{J} \hat{\beta}_{j_2} x_{j_2}^{(1)}\right]$. Thus:

$$\widehat{\mathrm{Var}}\left(\hat{\beta}_0\right) = \sum_{j_1=1}^{J}\left[\left(x_{j_1}^{(1)}\right)^2 \widehat{\mathrm{Var}}\left(\hat{\beta}_{j_1}\right)\right] + \sum_{j_2=1}^{J}\sum_{\substack{j_3=1 \\ j_3 \neq j_2}}^{J}\left[x_{j_2}^{(1)} x_{j_3}^{(1)} \widehat{\mathrm{Cov}}\left(\hat{\beta}_{j_2}, \hat{\beta}_{j_3}\right)\right]$$

Substituting for $\widehat{\mathrm{Var}}\left(\hat{\beta}_0\right)$ in (C10):

$$\begin{aligned}\widehat{\mathrm{Var}}&\left\{-\left[\left(x_j^{(1)} - 1\right)\hat{\beta}_j + \sum_{\substack{j_1=1 \\ j_1 \neq j}}^{J} \hat{\beta}_{j_1} x_{j_1}^{(1)}\right]\right\} \\ &= \widehat{\mathrm{Var}}\left(\hat{\beta}_0\right) + \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) - 2\left[x_j^{(1)} \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=1 \\ j_1 \neq j}}^{J} x_{j_1}^{(1)} \widehat{\mathrm{Cov}}\left(\hat{\beta}_j, \hat{\beta}_{j_1}\right)\right]\end{aligned} \tag{C11}$$

Substituting (C9) and (C11) into (C8), and simplifying:

$$\widehat{\mathrm{Cov}}\left(\hat{\beta}_0, \hat{\beta}_j\right) = -\left[x_j^{(1)} \widehat{\mathrm{Var}}\left(\hat{\beta}_j\right) + \sum_{\substack{j_1=1 \\ j_1 \neq j}}^{J} x_{j_1}^{(1)} \widehat{\mathrm{Cov}}\left(\hat{\beta}_j, \hat{\beta}_{j_1}\right)\right], \quad j = 1, 2, \ldots J .$$

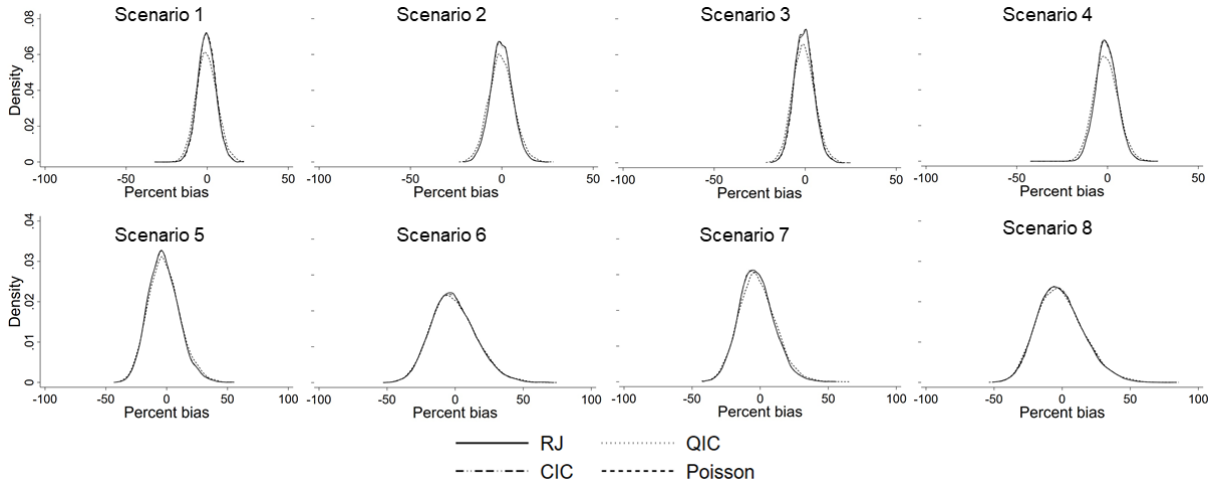

Figure C1: Density plots of the sampling distributions (10000 replications, eight scenarios) of the continuous covariate β_2. The density plots are drawn in ascending order of the average percent bias of all replications in each scenario. The density curves of the exact method under three criteria in each scenario are close together and do not spread out as much as the curve of Poisson.