May 2020

Biomarker Development for Use in Regression Calibration

Yiwen Zhang
University of Wisconsin-Milwaukee

BIOMARKER DEVELOPMENT FOR USE IN REGRESSION CALIBRATION

by

Yiwen Zhang

A Dissertation Submitted in
Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
in Public Health

at

The University of Wisconsin-Milwaukee

May 2020

ABSTRACT

BIOMARKER DEVELOPMENT FOR USE IN REGRESSION CALIBRATION

by

Yiwen Zhang

The University of Wisconsin-Milwaukee, 2020
Under the Supervision of Professor Cheng Zheng

It is challenging to alleviate systematic measurement error in self-reported data when studying the associations between dietary intakes and chronic disease risk. The regression calibration method has been used for this purpose when an objectively measured biomarker that satisfies a classical measurement error assumption is available. The requirement for the biomarkers needs to be quite strong and very few dietary intake biomarkers as such have been developed. Feeding studies provide opportunities to develop such potential biomarkers using regression methods with a much larger variety of dietary variables. However, the measurement error for the resulting biomarkers will be of Berkson type and these biomarkers are not suitable to the existing regression calibration method. Ignoring the violation of the classical measurement error assumption can lead to severe biases in disease association estimates. In this project, we propose three ways to obtain consistent estimates of such associations under rare disease assumption. The asymptotics of the proposed estimators is derived. Theoretical and numerical analyses were performed to compare these estimators. Estimation procedures are applied to the Women's Health Initiative (WHI) data to re-examine the associations between dietary intakes and cardiovascular diseases.

To

my parents,

and my husband

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to sincerely thank my advisor, Professor Cheng Zheng, for his tremendous mentor, immense knowledge and support during the past 4 years on my research and for allowing me grow to be a biostatistician. I consider myself to be extremely privileged to have been his student. I am very grateful to him for all his time spent, great patience and valuable guidance for this dissertation.

I would also like to thank Professor Chiang-Ching Huang, Professor Paul L. Auer, Professor Youngjoo Cho for kindly agreeing to be my committee members. I want to thank them for their continued support, brilliant comments and suggestions on my thesis.

I am immensely thankful to Professor Chiang-Ching Huang for his helpful support, instruction and encouragement throughout my PhD journey. I would like to thank all my professors who taught me and shared their invaluable knowledge.

I would like to specially thank my parents and my husband for their love and encouragement through the years.

Chapter 1: Overview

1.1 Introduction

There is an urgent need to obtain reliable information on dietary patterns that can reduce the risk of various chronic diseases, such as cancer, cardiovascular diseases (CVD), and diabetes. Although the positive association between obesity and cancer risk is well established [1], most epidemiology studies have not shown convincing evidence that key energy balance factors, such as total energy intake, are risk factors for various chronic diseases [2, 3, 4]. A likely cause of this apparent discrepancy is bias in dietary assessment, which is known to be challenging [5]. There is strong evidence [6] that the misreporting of dietary energy intake is related to individual characteristics (for example, body mass index (BMI)). The problem due to classical measurement errors which are assumed to be randomly distributed around true values with mean zero can be attenuated and overcame by increasing the sample size. However, the assumption of classical measurement errors can be violated in our study. The systematic measurement error will lead to bias that cannot be automatically corrected [7]. Previous methodology work and its application to the Women's Health Initiative (WHI) [6, 8, 9] have shown the validity and value of using (joint) regression calibration approaches to tackle this issue when objective measurements are available to be used as biomarkers of intakes. These biomarkers are used to build calibration equations using the self-reported measurements of the exposures of interest. Calibrated intake estimates are subsequently used to estimate associations between these dietary exposures and the risk of various diseases.

There remains a significant research gap that for many nutritional and physical activity variables, it is challenging to build satisfactory biomarkers with only single objective measurement. Therefore, regression models have been used to build biomarkers with multiple objective measurements. For example, in the WHI Nutrition and Physical Activity Assessment Study (NPAAS), to correct systematic measurement error of the self-reported food frequency questionnaire (FFQ) data from the full cohort (with 161,808 subjects), blood and urine measurements were collected for a subgroup (450 subjects) of the cohort [6]. In addition, a feeding study was performed on another smaller subgroup (153 subjects) where both blood and urine measurements and the assessed dietary intake information are collected [10]. The standard regression-based biomarker development procedure is as follows: First, using the 153 subjects regress the feeding study provided dietary intakes on the blood and urine measurements to predict dietary intake using these biospecimens along with study object characteristics [10]. Then the putative biomarker values can be calculated for the 450 subjects using their blood and urine measurements to build a calibration equation by regressing the calculated biomarker on the self-reported dietary intakes. Previous application of the regression calibration methods (i,e. [11]) has used an externally developed biomarker that plausibly satisfies a classical measurement error assumption, which, however, is known to be violated by this feeding study based biomarker development procedure. Ignoring this violation of classical measurement error assumption causes the subsequent estimates of the association between the dietary variable in question and disease risk to be biased. To tackle this issue, in this project, we aim to develop new methods for building biomarkers for regression calibration purposes. Using our new methods, we can establish consistent estimators for diet-disease associations under rare disease assumption and incorporate variation in the biomarker construction step when estimating the asymptotic variance and building confidence intervals for disease association parameters.

1.2 Motivating Data

1.2.1 WHI Study

Data used in this study is based on the cohort of the US Women's Health Initiative (WHI). From the year 1993 to 1998, 48,835 women enrolled in the WHI Dietary Modification Trial (DM), and 93,676 women enrolled in the prospective WHI Observational Study (OS). The age range of the participants were from 50 to 79 and all of them were postmenopausal [12]. Self-reported food frequency questionnaire (FFQ), which is subject to systematic bias, was collected at baseline and administered in the first year in the DM trial. Then the subsequent administrations were followed approximately every three years till April 2005 [13, 14]. The average follow-up time for OS and DM is nine years. The information related to the risk factors of CVD, including age, race, family history of premature CVD, educational level, diabetes, smoking status, use of statin, use of aspirin, use of postmenopausal hormone therapy previously, and an estimate of recreational physical activity, were collected at baseline. Women in the DM were followed after the year-1 visit. Women in the OS were followed at the enrollment. The study ended if either specific CVD outcomes or September 30, 2010, occurred first.

1.2.2 NPAAS Study

To start building a calibration equation, a sample of 450 women from Nutrition and Physical Activity Assessment Study (NPAAS) who were recruited from the WHI OS during 2007-2009 were used [6]. Women in NPAAS were overrepresented minorities and/or had elevated BMIs. Two clinical visits separated by two weeks are required in the study protocol. A 4-day food record, three 24-hour nutrient recalls were conducted. In addition, WHI

FFQ information, protein consumptions, and urinary nitrogen assessments of energy were obtained in the NPAAS study [6].

1.2.3 NPAAS Feeding Study

An ancillary study, called Nutrient and Physical Activity Assessment Feeding Study (NPAAS-FS), with 153 women recruited from 2011 to 2013 from WHI was developed to assess dietary intake information for a 2-week period [10]. Information collected from NPAAS-FS includes individual characteristics such as age, height, weight, BMI, race, medical history, education level, pregnancy history, family history, personal habits, and samples of fasting blood. Age, height, weight, and BMI were measured in NPAAS-FS at the study entry, whereas all other information listed above was collected at the time of enrollment in WHI study [10].

Blood and urine collection are essential information needed for building the calibration equation. Regarding serum metabolite measurements, a targeted liquid chromatography-tandem mass spectrometry (LC-MS/MS) assay, based on an Agilent 1260 HPLC/- Sciex 5500 QTRAP-MS platform, was used to measure metabolites from the 172 serum samples. To monitor the assay performance throughout the 2-day analysis period, 19 split sample blinded duplicate repeat QC specimens were also used. The global lipidomic analysis was performed on an Agilent 6520 QTOF-MS platform after spiking with a standard mixture of C17 lipids. By searching against data embedded in MPP and LIPID MAPS Structure Database, 200 to 400 lipids among 2000 MS features were identified across the sample set under high-resolution measurements.

Sample preparation for urine metabolite measurement includes urease treatment, methoxylation, and derivatization using MSTFA. Then with 1H nuclear magnetic resonance (NMR) spectroscopy at 800 MHz and untargeted gas chromatography-mass spectrometry (GC-MS), metabolite profiles of 172 (both spot and 24-hour) urine samples (including 19 split sample blinded duplicate samples) from the feeding study participants were obtained. Though some potential measurement errors in metabolite data may exist, the bias for GC-MS data was minimized through pre-processing and NMR data were normalized. Furthermore, the non-differential measurement error within the metabolite measurements are under our consideration and would not lead to unexpected bias.

1.3 Literature Review

With a biomarker developed with regression calibration, the association between dietary intake and disease will be examined and analyzed. Continuous, binary, and time-to-event endpoints are each selected depending on specific types of outcomes. With continuous and binary endpoints, generalized linear models will be used, whereas, with time-to-event endpoints, the Cox regression model will be used in this study. In addition, regression calibration build upon a multivariate regression model needs to be developed when multiple exposures are measured. The following subsections provided a brief introduction of all models/methods mentioned above, including a multivariate regression model, generalized linear model, the Cox regression model, measurement error model, and regression calibration method.

1.3.1 Multivariate Multiple Linear Regression Model

With multiple exposures, a multivariate multiple linear regression model (MMLR) can be considered. MMLR is similar to multiple linear regression (MLR) analysis. The difference is more than one response variable is involved in MMLR. When models are set up independently regarding each dependent variables, MMLR and MLR provide the same estimations on coefficients computationally [15, 16]. However, since multiple response variables could be correlated in general, using MMLR considering the correlation between different exposures is more appropriate compared to MLR. Suppose we have sample size of n, the response variable $\boldsymbol{Y} \in \mathbb{R}^{n\times K}$ contains K variables for n individuals. Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, is a design matrix. Then the multivariate multiple regression model can be formulated as:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \epsilon_i \overset{i.i.d}{\sim} N(0, \Sigma),$$

where $\boldsymbol{\beta} \in \mathbb{R}^{p\times K}$ is the coefficient matrix and $\boldsymbol{\epsilon} \in \mathbb{R}^{n\times K}$ denote matrix of residuals. Both $\boldsymbol{\beta}$ and Σ are unknown parameters [17, 18, 19]. A Gaussian model is assumed here with continuous exposure. Under the low-dimensional setting, the estimated coefficient can be derived by maximizing the likelihood determined by both the coefficient matrix $\boldsymbol{\beta}$ and the covariance matrix Σ

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{argmin}\{(\boldsymbol{Y} - \boldsymbol{\beta X})^T \Sigma (\boldsymbol{Y} - \beta \boldsymbol{X})\}.$$

1.3.2 Generalized Linear Model

In the simple linear model, we assumed that Y_i follows a normal distribution with mean μ_i and variance σ^2

$$Y_i \sim N(\mu_i, \sigma^2),$$

and the expected value, μ_i, is assumed to be a linear function of p predictors. To be more specific, μ_i takes values $x_i = (x_{i1}, ..., x_{ip})$ for the ith case and we have $\mu_i = x_i\beta$, where β is a vector of unknown parameters. In practice, there are different types of measurements with non-normal error distribution. To accommodate various data types, generalized linear model (GLM) was developed as a generalization of classical linear models by linking response variable with linear models via a link function and extends the linear model to the general exponential family [20].

We introduce the link function, $g(\mu_i)$, which is a one-to-one continuous differentiable transformed function. A transformed mean is assumed to follow a linear model with a form of $\eta_i = x_i\beta$, where η_i is called the linear predictor. With a one-to-one link function, we can further obtain $\mu_i = g^{-1}(x_i\beta)$. Note that the expected value, μ_i, is the one we should transform. There are various link functions such as identity, log, reciprocal, logit and probit, etc.

Under the GLM framework, the probability density function of Y is taking the form:

$$f(y_i) = exp\{(y_i\theta_i - b(\theta_i))/a(\phi) + c(y_i, \phi)\}.$$

Here θ_i and ϕ are parameters and $a_i(\phi)$, $b(\theta_i)$ and $c(y_i, \phi)$ are known functions.The parameters θ_i and ϕ are essentially location and scale parameters. It can be shown that if Y_i has a distribution in the exponential family then it has mean and variance $E(Y_i) = \mu_i = b'(\theta_i); var(Y_i) = \sigma_i^2 = b''(\theta_i)a_i(\phi)$ where $b'(\theta_i)$ and $b''(\theta_i)$ are the first and second derivatives of $b(\theta_i)$. The exponential family provides uniform parameterization for the parametric family of distributions including normal, binomial, Poisson, exponential, gamma and inverse Gaussian distributions. Here, θ is related to the mean of the distribution, if $b(\theta)$ is an identity function, we say that we have a canonical link.

One challenge under the framework of GLM is the estimated coefficient usually does not come with a closed-form solution. The estimation and inference are driven by the maximum likelihood approach. GLM can be fit to the data using the algorithm, called iteratively re-weighted least squares (IRLS), and we will introduce this algorithm in this section.

(1) Choose an initial value $\hat{\beta}^{(0)}$, we calculated the estimated linear predictor $\hat{\eta}_i = x_i\beta$. Then we can obtain the value for $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$.

(2) With the above calculated values, we can obtain the adjusted response, $z_i = \hat{\eta}_i + (y_i - \hat{\mu}_i)\frac{d\eta_i}{d\mu_i}$.

(3) Next, iterative weight can be calculated: $w_i = (\frac{d\mu_i}{d\eta_i})^2 / b''(\theta_i)$.

(4) This gives the maximum likelihood estimates: $\hat{\beta}^{(1)} = (X^T W X)^{(-1)} X^T W z$.

(5) Iterate above four steps until $\hat{\beta}$ converges.

1.3.3 Cox Regression Model

Since the time-to-event endpoint is also of our interest, we will introduce background information regarding the Cox regression model that we applied for the time-to-event endpoint.

The survivor function, $S(t)$, accounts for a patient's probability of survival from the time of origin (i.e., diagnosis of cancer) to a specified time, t, in the future. The survival probabilities at each time point, t, summarize the survival experience overall. The hazard, denoted by $\lambda(t)$, is the incident event rate for the patient who has already survived until time t. To be more specific, $\lambda(t)$ is the probability encountering an event under the observation at time t. In contrast to the survivor function, which focuses on the cumulative survival probability until time t, hazard function focuses on the instantaneous rate of an event occurring [21, 22, 23].

Cox regression model, also known as the proportional hazard model, is a method used to investigate the effect of risk factors or exposures upon a time an event occurs simultaneously [21, 22]. The measure of effect in the Cox regression model is the hazard rate, and it can be expressed as below:

$$\lambda(t) = \lambda_0(t) exp(\boldsymbol{X}\theta),$$

where $\lambda(t)$ is a hazard function determined by a set of p covariates, $\boldsymbol{X}$, and θ is a $p \times 1$ vector of unknown parameters. The term $\lambda_0(.)$ is called the baseline hazard. It corresponds to the value of the hazard if all the x_i are equal to zero. The estimating equation for coefficient θ can be expressed as:

$$U(\theta) = \sum_i \int_0^\tau \left[\boldsymbol{X}_i - \sum_j \frac{Y_j(t)\exp\{\boldsymbol{X}_j\theta\}}{\sum_k Y_k(t)\exp\{\boldsymbol{X}_k\theta\}} \boldsymbol{X}_j \right] dN_i(t),$$

where $N_i(t) = I(\Delta_i = 1, Y_i \leq t)$ and $Y_i(t) = I(Y_i \geq t)$. The asymptotic normality of $\hat{\theta}$ can be proved with the above estimating equation.

1.3.4 Measurement Error Model

We will first introduce two types of measurement error.

(1) Classical measurement error

Classical measurement error is the most common assumption made in measurement error literature [7]. Suppose $\boldsymbol{Z}$ is the true dietary intake, $\boldsymbol{X}$ is the corresponding biomarker and $\boldsymbol{\epsilon}$ be the measurement error, then classical measurement error model is

$$\boldsymbol{X} = \boldsymbol{Z} + \boldsymbol{\epsilon},$$

where this model states $E(\boldsymbol{\epsilon}|\boldsymbol{Z}) = 0$ and usually the error structure of $\boldsymbol{\epsilon}$ is constant variance [7]. Classical measurement error are independent of the true exposure with mean zero.

(2) Berkson error

With the same symbols in classical measurement error, Berkson error model states that

$$\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{\epsilon},$$

indicating the true dietary intakes contain more variability than biomarker [7]. In this model, $E(\boldsymbol{\epsilon}|\boldsymbol{X}) = 0$. In other words, the approximate exposure is followed by many subjects, while the true exposure varies randomly around the approximate exposure with Berkson error. Regression calibration is one of the most common methods to accommodate measurement error. Other than regression calibration, various other strategies have also been studied and proposed to reduce or eliminate the bias due to measurement error. Stefanski & Carroll (1987) [24] proposed the conditional score method. Nakamura & Tsuyoshi (1990) [25] proposed the corrected score method. The expected estimating equation proposed by Wang et al. (2000) [26] is a technique to attenuate bias when repeated mismeasured variables or surrogate variables are available. Stefanski & Cook (1995) [27] introduced a simulation-based method called simulation extrapolation (SIMEX) without requirement in making an assumption on the distribution of true covariates. SIMEX has been extensively studied and implemented under both parametric and non-parametric statistical problems. The moment reconstruction proposed by Freedman et al. (2004) [28] is another substitution for regression calibration to correct measurement error in univariate continuous exposures. One advantage with the moment reconstruction approach is the covariate measurement error is allowed to be differential.

1.3.5 Regression Calibration

There are generally two types of regression calibration to accommodate measurement error. One is regression calibration with repeated measurements using the same instrument. With repeated measurements, the conditional expectations of the true values given observed values can be estimated. This technique can be used to attenuate random error under the classical measurement error model.

The other is regression calibration with the validation set. Measurement error correction using the validation set usually compares results obtained between a dietary instrument such as FFQ and another instrument with more accurate measurements [29]. The calibration regression study is usually based on a much smaller cohort with consumed dietary intakes and other observed variables with no error. Usually, a bias-corrected matrix can be computed with the coefficients from the linear regression of actual exposures on observed values. This technique is often used to deal with the systematic error. Sugar et al. (2007) [30] developed statistical estimation methods for odds ratio estimation under the framework of the measurement error model proposed by Prentice et al. (2002) [31]. Prentice (1982) [8] developed a failure time regression model for parameter estimation with measurement errors in covariates. To be more specific, an induced hazard function model for failure time developed by Prentice (1982) [8] is as shown below. The hazard function with true covariates, $\boldsymbol{Z} \in R^P$, can be written as:

$$\lambda(t, \boldsymbol{Z}) = \lambda_0(t) exp(\boldsymbol{Z}\theta),$$

with baseline hazard function, $\lambda_0(.)$. With measurement errors in observed covariates, $\boldsymbol{X} \in R^P$, we have biased estimations on parameters, θ. The hazard functions of $\lambda(t, \boldsymbol{Z})$ and $\lambda(t, \boldsymbol{X})$ can induce a new hazard function to alleviate measurement errors in covariates based on the assumption that $\{\boldsymbol{X}, T \geq t\}$ is conditional independent given $\boldsymbol{Z}$. Under the assumptions, we have:

$$\lambda\{t; \boldsymbol{Z}, \boldsymbol{X}\} = \lambda\{t; \boldsymbol{Z}\}.$$

Then $\lambda\{t; \boldsymbol{X}\}$ can be written as the conditional expectation of $\lambda\{t; \boldsymbol{Z}, \boldsymbol{X}\}$ given $T \geq t$ and $\boldsymbol{X}$ where T is the failure time, that is:

$$\lambda\{t; \boldsymbol{X}\} = E\left[\lambda\{t; \boldsymbol{Z}, \boldsymbol{X}\} | T \geq t, \boldsymbol{X}\right],$$

$$\lambda\{t; \boldsymbol{X}\} = E\left[\lambda\{t; \boldsymbol{Z}\} | T \geq t, \boldsymbol{X}\right].$$

Therefore, an induced hazard function can be derived as:

$$\lambda(t;\boldsymbol{X}) = \lambda_0(t)E\left[exp(\boldsymbol{Z}\theta)|T \geq t, \boldsymbol{X}\right].$$

Under rare disease assumption where $P(T < t|\boldsymbol{Z}) \approx 0$, the induced hazard function can be approximated as:

$$\lambda(t;\boldsymbol{X}) \approx \lambda_0(t)E\left[exp(\boldsymbol{Z}\theta)|\boldsymbol{X}\right].$$

This approximation indicates that $\boldsymbol{Z}$ given $\{\boldsymbol{X}, T \geq t\}$ is constant over time. With normally distributed covariates, the conditional distribution of $(\boldsymbol{X}|\boldsymbol{Z})$ is normal with mean $E(\boldsymbol{Z}|\boldsymbol{X})$ and variance $Var(\boldsymbol{Z}|\boldsymbol{X})$. Hence the induced hazard function can be further written as:

$$\lambda(t;\boldsymbol{X}) = \lambda_0(t)exp\left[E(\boldsymbol{Z}|\boldsymbol{X})\theta + \frac{1}{2}\theta^T Var(\boldsymbol{Z}|\boldsymbol{X})\theta\right],$$

$$\lambda(t;\boldsymbol{X}) = \lambda_0^*(t)exp\left[E(\boldsymbol{Z}|\boldsymbol{X})\theta\right],$$

where $\lambda_0^*(t) = \lambda_0(t)exp(\frac{1}{2}\theta^T Var(\boldsymbol{Z}|\boldsymbol{X})\theta)$.

A regression calibration model with failure time regression analysis is conducted by Wang et al. (1997) [32] when data on covariates are missing or inaccurately measured. Shaw & Prentice (2012) [9] applied three estimation procedures for hazard ratio with measurement error data structure. Gorfine, Hsu & Prentice (2004) [33] developed a non-parametric correction approach for covariate measurement error in a stratified Cox model. Liao et al. (2011) [34] proposed a risk set regression calibration in survival analysis with time-varying covariates.

1.4 General Notation and Framework

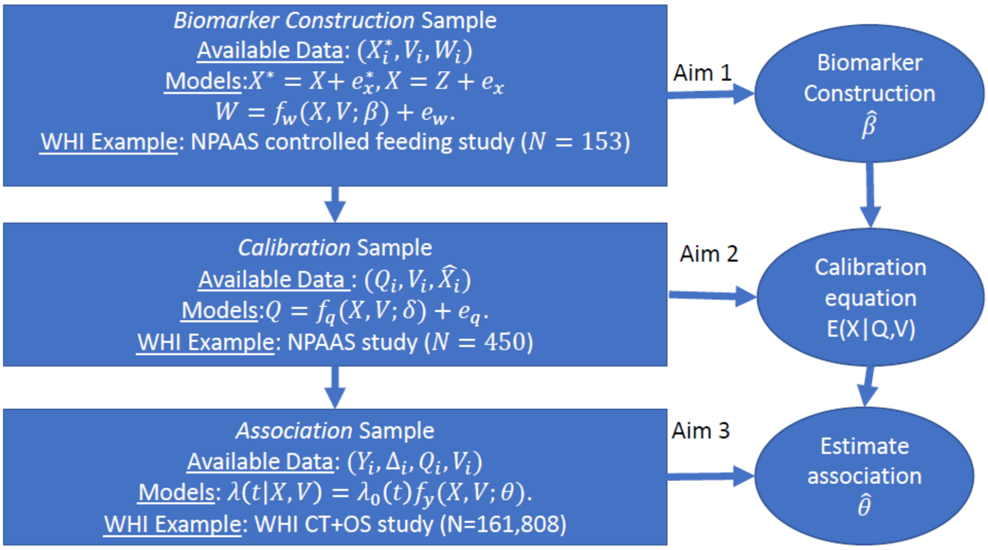
The proposed association study between specified dietary variables and chronic diseases is composed of 3 stages (Figure 1: the biomarker construction, the calibration, and the association assessment) with three corresponding aims: (1) develop a valid biomarker that could be further used in calibration regression study; (2) develop a valid calibration equation for the self-reported dietary intake; (3) achieve a valid estimation of the association between the dietary intake and disease risks.

The more detailed procedure is as follows. In stage 1, we build a model to establish biomarkers using a group of subjects from a controlled feeding study. This model can be built by regressing the consumed dietary intakes measured from the controlled feeding study onto (i) blood/urine measurements combined with personal characteristics; (ii) blood/urine measurements, personal characteristics and self-reported dietary intake; or (iii) personal characteristics and self-reported dietary intake. From the NPAAS, there are 153 individuals whose consumed dietary intakes over a 2-week controlled feeding period were collected. In stage 2, using data from a different group of individuals, we build a calibration equation using the self-reported dietary intakes, and low dimensional personal characteristics to predict the true dietary intake if (i) or (ii) is used in stage 1. If (iii) is used in stage 1, then the calibration equation was already established and stage 2 can be omitted. We have 450 NPAAS samples to use in this stage. In stage 3, for a much larger group of individuals, we only have information on the self-reported dietary intake, the (low dimensional) personal characteristics, and a composite survival outcome. We use the calibration equation developed in stage 2 to calibrate the self-reported dietary intake for the large cohort and perform the association studies with various diseases. In this thesis, we will compare the existing biomarker development method and three newly proposed methods.

First, we list all notations involved in these 3 stages. Denote the sample sizes of these three stages as n_1, n_2 and n_3. When there is no sample overlapping among these three sets, we can assume they are mutually independent and we will handle this scenario when deriving asymptotic results. Without loss of generality, we assume $i = 1, \cdots, n_1$ are from the first sample, $i = n_1 + 1, \cdots, n_1 + n_2$ are from the second sample and $i = n_1 + n_2 + 1, \cdots, n_1 + n_2 + n_3$ are from the third sample.

We denote the underlying long term dietary intakes as $\boldsymbol{Z} \in R^K$, which are primary exposures of interest. For example, $\boldsymbol{Z}$ may be the average daily sodium to potassium intake ratio over a specified one-year period. We denote personal characteristics as $\boldsymbol{V} \in R^q$. We denote the short-term dietary intakes over the feeding period as $\boldsymbol{X} \in R^K$ and assume the classical measurement error model that $\boldsymbol{X} = \boldsymbol{Z} + \epsilon_x$, where $\epsilon_x \sim N(0, \boldsymbol{\Sigma}_x)$ is independent of $\boldsymbol{Z}$, $\boldsymbol{V}$. We denote the assessed dietary intake in the feeding study as $\boldsymbol{X}^*$, where $\boldsymbol{X}^* = \boldsymbol{X} + \epsilon_x^*$, where $\epsilon_x^* \sim N(0, \boldsymbol{\Sigma}_x^*)$ is independent of ϵ_x, $\boldsymbol{Z}$, $\boldsymbol{V}$. The self-reported dietary intakes are denoted as $\boldsymbol{Q} \in R^K$ and are assumed to follow a parametric model $\boldsymbol{Q} = (1, \boldsymbol{Z}^T, \boldsymbol{V}^T)\boldsymbol{A} + \epsilon_q$, where $\epsilon_q \sim N(0, \boldsymbol{\Sigma}_q)$ is independent of ϵ_x, ϵ_x^*, $\boldsymbol{Z}$ and $\boldsymbol{V}$ and $\boldsymbol{A} \in R^{(1+K+q)\times K}$ are unknown parameters. The objective blood and urine measurements are denoted as $\boldsymbol{W} \in R^p$ and are assumed to follow a parametric model $\boldsymbol{W} = (1, \boldsymbol{X}^T, \boldsymbol{V}^T)\boldsymbol{B} + \epsilon_w$, where $\epsilon_w \sim N(0, \boldsymbol{\Sigma}_w)$ is independent of ϵ_x, ϵ_x^*, ϵ_q, $\boldsymbol{Z}$, $\boldsymbol{V}$ and $\boldsymbol{B} \in R^{(1+K+q)\times p}$ are unknown parameters. For continuous outcome or binary outcome, we denote it as Y, which are assumed to follow a generalized linear model $g(E[Y|\boldsymbol{Z}, \boldsymbol{V}, \boldsymbol{Q}, \boldsymbol{X}, \boldsymbol{X}^*, \boldsymbol{W}]) = (1, \boldsymbol{Z}^T, \boldsymbol{V}^T)\theta$, where $\theta \in R^{1+K+q}$ is parameter of interest. With time-to-event endpoint, the composite outcome are denoted as $(Y = T \wedge C, \Delta = I(T \leq C))$ where T is the time to disease occurrence which is assumed to follow a Cox model with hazard specified as $\lambda(t|\boldsymbol{Z}, \boldsymbol{V}, \boldsymbol{Q}, \boldsymbol{X}, \boldsymbol{X}^*, \boldsymbol{W}) = \lambda_0(t)\exp((\boldsymbol{Z}^T, \boldsymbol{V}^T)\theta)$, where $\theta \in R^{K+q}$ is parameter of interest. The censoring time C is assumed to be independent of T given $(\boldsymbol{Z}, \boldsymbol{V})$. For the biomarker construction sample, researcher can observe $(\boldsymbol{X}^*, \boldsymbol{W}, \boldsymbol{V})$ and possibly $\boldsymbol{Q}$. For the calibration set, researcher can observe $(\boldsymbol{Q}, \boldsymbol{W}, \boldsymbol{V})$ and for the cohort, researcher can observes $(Y, \boldsymbol{Q}, \boldsymbol{V})$ or $(Y, \Delta, \boldsymbol{Q}, \boldsymbol{V})$ depending on the type of outcomes. For distribution theory development, we use traditional counting process notation $N_i(t) = I(\Delta_i = 1, Y_i \leq t)$ and $Y_i(t) = I(Y_i \geq t)$.

Figure 1: Flow chart of the whole procedure from biomarker construction to association assessment


Biomarker Construction Sample
Available Data: (X_i^*, V_i, W_i)
Models: $X^* = X + e_x^*, X = Z + e_x$
$W = f_w(X, V; \beta) + e_w.$
WHI Example: NPAAS controlled feeding study ($N = 153$)
Aim 1
Biomarker Construction $\hat{\beta}$
Calibration Sample
Available Data : $(Q_i, V_i, \widehat{X}_i)$
Models: $Q = f_q(X, V; \delta) + e_q.$
WHI Example: NPAAS study ($N = 450$)
Aim 2
Calibration equation E(X|Q,V)
Association Sample
Available Data: $(Y_i, \Delta_i, Q_i, V_i)$
Models: $\lambda(t|X, V) = \lambda_0(t) f_y(X, V; \theta).$
WHI Example: WHI CT+OS study (N=161,808)
Aim 3
Estimate association $\hat{\theta}$

Chapter 2: Low-dimensional Setting with Single Exposure

2.1 Introduction

Cardiovascular disease (CVD), which encompasses a class of diseases that involves several conditions affecting major blood vessels and heart, is a leading cause of mortality worldwide [35]. The associations between dietary consumption and CVD had been indicated in many recent studies [36, 37].

In the US Dietary Guidelines of the year 2015 [38], the sodium intake is recommended to be limited within 2,300 mg/day for CVD prevention. Self-reported intakes are heavily relied for studying these associations in epidemiological studies. However, substantial biases in self-reported intakes that are correlated with personal characteristics (i.e., Body mass index (BMI, measured as weight/height)) were found in many prior studies. More reliable assessments for sodium intakes are needed, but it is extremely hard to estimate sodium intake with standard dietary assessment methods since over 70% of packaged and processed foods contain sodium. The assessment with standard self-report methods is particularly difficult to accurately assess the amounts of hidden sodium in these processed foods. Studies with constructed biomarkers involving 24-hour urine measurements are more reliable.

Based on current literatures, it has been found that a high sodium-to-potassium ratio is directly associated with systolic blood pressure [39, 40]. A positive association between sodium-to-potassium excretion ratio in 24 hours and the incidence of CVD is reported in a small study with 193 CVD incident events [14]. Though 24-hour urinary excretion is an ideal approach, it is impractical to assess the actual 24-hour urinary collections for an extensive epidemiological cohort study based on the Prospective Urban Rural Epidemiology Study. Hence large cohort studies typically do not collect the 24-hour urine measurements. Moderate-sized subcohort with 24-hour urine collected in WHI was utilized to construct calibration equations to alleviate the biases in self-reported dietary data [14]. Then calibrated estimates for disease association parameters can be obtained through calibration equations on self-reported dietary data. In many studies [8, 32], the main assumption of the regression calibration to follow is the biomarker estimated dietary intakes equal log-transformed true dietary intakes plus some errors. This is the so-called classical measurement model. With a classical measurement model, the objective dietary intakes can be different from the true dietary intakes and in a manner that is random and independent among subjects. For instance, an exposure during a 2-week period with some noises is independent of the true dietary intake and of other individual characteristics. However, when true dietary intakes contain more variability with random errors, the classical measurement error assumption can be violated and lead to large bias. This is the so-called Berkson error. Under such cases, we developed a bias factor to correct the bias. In this chapter, we evaluated the associations of the risk of cardiovascular diseases and nutrient intakes with regression calibration.

2.2 Methods

We first consider the case for a single exposure of interest ($K = 1$) where $\Sigma_X^* = \sigma_x^{*2}$ is known. We propose methods to estimate σ_x^{*2} in Section 2.6 in this Chapter. In the real data analysis where σ_x^{*2} is not available, we vary this parameter to perform sensitivity analysis. The notations listed in Chapter 1 is followed in the rest of the chapters. Recall that the number of variables and sample size in the biomarker construction step is denoted by p and n_1, respectively. In this chapter, we focus on low-dimensional data ($p < n_1$).

2.2.1 Method 1: The naïve three-step approach

We first consider the naïve three-step approach. We first perform a linear regression among n_1 subjects in the biomarker discovery sample of consumed diet (X^*) on blood and urine measurements ($\boldsymbol{W}$) as well as subject characteristics ($\boldsymbol{V}$) to obtain:

$$\hat{\beta}_1 = \left\{ \sum_{i=1}^{n_1} (1, \boldsymbol{W}_i^T, \boldsymbol{V}_i^T)^T (1, \boldsymbol{W}_i^T, \boldsymbol{V}_i^T) \right\}^{-1} \left\{ \sum_{i=1}^{n_1} (1, \boldsymbol{W}_i^T, \boldsymbol{V}_i^T)^T X_i^* \right\}.$$

Then we compute $\hat{X}_{1i} = (1, \boldsymbol{W}_i^T, \boldsymbol{V}_i^T)\hat{\beta}_1$ for $i = n_1 + 1, \cdots, n_1 + n_2$ to predict the long-term dietary intake (Z) among the n_2 calibration samples and run a regression of $\hat{X}_1$ on self-reported food frequency questionnaire data (Q) and subject characteristics ($\boldsymbol{V}$) to build calibration equation using the n_2 calibration samples with

$$\hat{\gamma}_1 = \left\{ \sum_{i=n_1+1}^{n_1+n_2} (1, Q_i, \boldsymbol{V}_i^T)^T (1, Q_i, \boldsymbol{V}_i^T) \right\}^{-1} \left\{ \sum_{i=n_1+1}^{n_1+n_2} (1, Q_i, \boldsymbol{V}_i^T)^T \hat{X}_{1i} \right\}.$$

Finally, we predict the exposure by $\hat{Z}_{1i} = (1, Q_i, \boldsymbol{V}_i^T)\hat{\gamma}_1$ for $i = n_1 + n_2 + 1, \cdots, n_1 + n_2 + n_3$ in the n_3 association sample. For continuous endpoint, a linear model of Y on $\hat{Z}_1$ and $\boldsymbol{V}$ is performed to estimate the association parameter by solving the estimating equation:

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \left[(1, \hat{Z}_{1i}, \boldsymbol{V}_i^T)^T \left\{ Y_i - (1, \hat{Z}_{1i}, \boldsymbol{V}_i^T)\theta \right\} \right].$$

For binary endpoint, a logistic model is used to estimate the association parameter based on the estimating equation as below:

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \left[(1, \hat{Z}_{1i}, \boldsymbol{V}_i^T)^T \left\{ Y_i - \frac{exp((1, \hat{Z}_{1i}, \boldsymbol{V}_i^T)\theta)}{1 + exp((1, \hat{Z}_{1i}, \boldsymbol{V}_i^T)\theta)} \right\} \right].$$

For time-to-event endpoint, the score function for a Cox model is used:

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \int_0^{\tau} \left[\begin{pmatrix} \hat{Z}_{1i} \\ \boldsymbol{V}_i \end{pmatrix} - \sum_j \frac{Y_j(t) \exp\left\{(\hat{Z}_{1j}, \boldsymbol{V}_j^T)\theta\right\}}{\sum_k Y_k(t) \exp\left\{(\hat{Z}_{1k}, \boldsymbol{V}_k^T)\theta\right\}} \begin{pmatrix} \hat{Z}_{1j} \\ \boldsymbol{V}_j \end{pmatrix} \right] dN_i(t), \tag{2.1}$$

where τ is a pre-specified large number and we assume $P(C > \tau) > 0$.

We show in Appendix A (Theorem 4) that $E(\hat{Z}_1|Q, \boldsymbol{V}) = BF \times E(Z|Q, \boldsymbol{V}) + (1 - BF) \times E(Z|\boldsymbol{V}) \neq E(Z|Q, \boldsymbol{V})$, where the bias factor (BF) is defined as:

$$BF = 1 - \frac{Var(X|\boldsymbol{W}, \boldsymbol{V})}{Var(X|\boldsymbol{V})} = R^2_{1|\boldsymbol{V}}.$$

Here $R^2_{1|\boldsymbol{V}}$ represents the multiple partial correlation coefficient from stage 1. Such BF will lead to bias in the estimation of association parameter θ. If we further assume $E(X|\boldsymbol{V}) = (1, \boldsymbol{V}^T)\delta$, we show in Appendix A (Theorem 4) that the estimator $\hat{\theta}_1 \to \theta_1^*$ as $n \to \infty$, with $\theta^*_{1z} \approx \frac{\theta_z}{BF}$ and $\theta^*_{1v} \approx \theta_v - \frac{1-BF}{BF}\theta_{1z}$ where $(\theta^*_0, \theta^*_{1z}, \theta^{*T}_{1v})^T = \theta^*_1$ and $(\theta_0, \theta_z, \theta_v^T)^T = theta$. When the outcome is continuous and linear regression model is used, there is no approximation error. When the outcome is binary or time-to-event, the approximation error is ignorable only under rare disease assumption. That is, $P(T < t|Z, \boldsymbol{V}) \to 0$ when $n \to \infty$ for $t \in [0, \tau]$ under time-to-event endpoint or $P(Y = 1|Z, \boldsymbol{V}) \to 0$ under binary endpoint for all levels of Z and $\boldsymbol{V}$.

2.2.2 Method 2: Three-step with Bias Correction

As shown in Method 1, we have a bias factor in $\hat{Z}_1$ when using $\hat{X}_1$, so we propose a bias-corrected estimator $\hat{X}_{2i} = \hat{X}_{1i}\widehat{BF}^{-1}$ where,

$$\widehat{BF} = \hat{R}^2_{1|\boldsymbol{V}} = 1 - \frac{\widehat{Var}(X^*|\boldsymbol{W}, \boldsymbol{V}) - \sigma_x^{*2}}{\widehat{Var}(X^*|\boldsymbol{V}) - \sigma_x^{*2}},$$

is an estimated version of the bias factor. Then we have:

$$\hat{\gamma}_2 = \left\{ \sum_{i=n_1+1}^{n_1+n_2} (1, Q_i, \boldsymbol{V}_i^T)^T (1, Q_i, \boldsymbol{V}_i^T) \right\}^{-1} \sum_{i=n_1+1}^{n_1+n_2} \left\{ (1, Q_i, \boldsymbol{V}_i^T)^T \hat{X}_{2i} \right\}.$$

Finally, we predict the exposure by $\hat{Z}_{2i} = (1, Q_i, \boldsymbol{V}_i^T)\hat{\gamma}_2$ for $i = n_1 + n_2 + 1, \cdots, n_1 + n_2 + n_3$ and we have $\hat{\theta}_2$ by solving the following estimating equations with respect to continuous, binary and time-to-event endpoints:

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \left[(1, \hat{Z}_{2i}, \boldsymbol{V}_i^T)^T \left\{ Y_i - (1, \hat{Z}_{2i}, \boldsymbol{V}_i^T)\theta \right\} \right],$$

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \left[(1, \hat{Z}_{2i}, \boldsymbol{V}_i^T)^T \left\{ Y_i - \frac{exp((1, \hat{Z}_{2i}, \boldsymbol{V}_i^T)\theta)}{1 + exp((1, \hat{Z}_{2i}, \boldsymbol{V}_i^T)\theta)} \right\} \right],$$

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \int_0^\tau \left[\begin{pmatrix} \hat{Z}_{2i} \\ \boldsymbol{V}_i \end{pmatrix} - \sum_j \frac{Y_j(t) \exp\left\{ (\hat{Z}_{2j}, \boldsymbol{V}_j^T)\theta \right\}}{\sum_k Y_k(t) \exp\left\{ (\hat{Z}_{2k}, \boldsymbol{V}_k^T)\theta \right\}} \begin{pmatrix} \hat{Z}_{2j} \\ \boldsymbol{V}_j \end{pmatrix} \right] dN_i(t). \tag{2.2}$$

This method does not require the self-reported dietary intake data (Q) to be collected in the feeding study. As a remark, even if the self-reported data is available in the feeding study, the correlation structure between the self-reported and the actual dietary intake in the feeding study might be different from that correlation structure in the cohort because of, (1) the modification on the dietary pattern during the controlled feeding study or (2) the potential change in dietary preference in the period of the feeding study. This method is robust to such an association difference since we have not directly included Q in stage 1 for biomarker construction. We will next propose two methods that require the availability of the self-reported dietary intake in the feeding study and assume the association between the self-reported dietary intake and the actual dietary intake to be the same among all three samples.

2.2.3 Method 3: Three-step with self-reported data

When the self-reported dietary intake Q is available from the feeding study and we believe that the distribution of $(Q|Z, \boldsymbol{V})$ are the same between controlled feeding study and the cohort, then the bias in the naïve estimator can be corrected simply by including Q in the biomarker development equation since the inclusion of Q guarantees that $E[\hat{Z}|Q, \boldsymbol{V}] = E[E[Z|\boldsymbol{W}, Q, \boldsymbol{V}]|Q, \boldsymbol{V}] = E[Z|Q, \boldsymbol{V}]$.

The three steps of the first method remain the same, but in the first step regression model, the log-transformed self-reported food frequency questionnaire data (Q) is added. That is, for the first step, we regress X^* on $\boldsymbol{W}$, $\boldsymbol{V}$ and Q to build the biomarker, and then use $\boldsymbol{W}$, $\boldsymbol{V}$ and Q to predict Z in the second step. Mathematically, we have:

$$\hat{\beta}_3 = \left\{ \sum_{i=1}^{n_1} (1, \boldsymbol{W}_i^T, Q_i, \boldsymbol{V}_i^T)^T (1, \boldsymbol{W}_i^T, Q_i, \boldsymbol{V}_i^T) \right\}^{-1} \left\{ \sum_{i=1}^{n_1} (1, \boldsymbol{W}_i^T, Q_i, \boldsymbol{V}_i^T)^T X_i^* \right\},$$

$\hat{X}_{3i} = (1, \boldsymbol{W}_i^T, Q_i, \boldsymbol{V}_i^T)\hat{\beta}_3$ for $i = n_1 + 1, \cdots, n_1 + n_2$, and then,

$$\hat{\gamma}_3 = \left\{ \sum_{i=n_1+1}^{n_1+n_2} (1, Q_i, \boldsymbol{V}_i^T)^T (1, Q_i, \boldsymbol{V}_i^T) \right\}^{-1} \left\{ \sum_{i=n_1+1}^{n_1+n_2} (1, Q_i, \boldsymbol{V}_i^T)^T \hat{X}_{3i} \right\},$$

and $\hat{Z}_{3i} = (1, Q_i, \boldsymbol{V}_i^T)\hat{\gamma}_3$ for $i = n_1 + n_2 + 1, \cdots, n_1 + n_2 + n_3$. We obtain $\hat{\theta}_3$ by solving the following estimating equations with respect to continuous, binary and time-to-event endpoints:

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \left[(1, \hat{Z}_{3i}, \boldsymbol{V}_i^T)^T \left\{Y_i - (1, \hat{Z}_{3i}, \boldsymbol{V}_i^T)\theta\right\}\right],$$

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \left[(1, \hat{Z}_{3i}, \boldsymbol{V}_i^T)^T \left\{Y_i - \frac{exp((1, \hat{Z}_{3i}, \boldsymbol{V}_i^T)\theta)}{1 + exp((1, \hat{Z}_{3i}, \boldsymbol{V}_i^T)\theta)}\right\}\right],$$

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \int_0^\tau \left[\begin{pmatrix} \hat{Z}_{3i} \\ \boldsymbol{V}_i \end{pmatrix} - \sum_j \frac{Y_j(t) \exp\left\{(\hat{Z}_{3j}, \boldsymbol{V}_j^T)\theta\right\}}{\sum_k Y_k(t) \exp\left\{(\hat{Z}_{3k}, \boldsymbol{V}_k^T)\theta\right\}} \begin{pmatrix} \hat{Z}_{3j} \\ \boldsymbol{V}_j \end{pmatrix}\right] dN_i(t). \tag{2.3}$$

2.2.4 Method 4: Direct Estimation

When Q is available from the feeding study, another possibility is to ignore the second dataset and directly build the estimating equation by regressing X^* on Q and $\boldsymbol{V}$ in the first step and directly apply it to the third step. All other steps remain the same as the third method, except that we ignore the n_2 calibration samples and directly build the calibration equation using the feeding study by regressing X^* on $\boldsymbol{V}$ and Q. Then we use the calibration equation to predict Z and perform a regression of Y on Z and $\boldsymbol{V}$ in the full cohort to estimate the association parameter. In other words, we have:

$$\hat{\gamma}_4 = \left\{\sum_{i=1}^{n_1} (1, Q_i, \boldsymbol{V}_i^T)^T (1, Q_i, \boldsymbol{V}_i^T)\right\}^{-1} \left\{\sum_{i=1}^{n_1} (1, Q_i, \boldsymbol{V}_i^T)^T X_i^*\right\},$$

with $\hat{Z}_{4i} = (1, Q_i, \boldsymbol{V}_i^T)\hat{\gamma}_4$ for $i = n_1 + n_2 + 1, \cdots, n_1 + n_2 + n_3$ and $\hat{\theta}_4$ by solving the following estimating equations for continuous, binary and time-to-event endpoints, respectively:

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \left[(1, \hat{Z}_{4i}, \boldsymbol{V}_i^T)^T \left\{ Y_i - (1, \hat{Z}_{4i}, \boldsymbol{V}_i^T)\theta \right\} \right],$$

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \left[(1, \hat{Z}_{4i}, \boldsymbol{V}_i^T)^T \left\{ Y_i - \frac{exp((1, \hat{Z}_{4i}, \boldsymbol{V}_i^T)\theta)}{1 + exp((1, \hat{Z}_{4i}, \boldsymbol{V}_i^T)\theta)} \right\} \right],$$

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \int_0^{\tau} \left[\begin{pmatrix} \hat{Z}_{4i} \\ \boldsymbol{V}_i \end{pmatrix} - \sum_j \frac{Y_j(t) \exp\left\{ (\hat{Z}_{4j}, \boldsymbol{V}_j^T)\theta \right\}}{\sum_k Y_k(t) \exp\left\{ (\hat{Z}_{4k}, \boldsymbol{V}_k^T)\theta \right\}} \begin{pmatrix} \hat{Z}_{4j} \\ \boldsymbol{V}_j \end{pmatrix} \right] dN_i(t). \tag{2.4}$$

2.3 Theory

The asymptotic distributions of the four estimators were derived and show that Method 1 tends to give biased result while the other three methods provide consistent estimators under rare disease assumption ($P(T < t|Z, \boldsymbol{V}) \to 0$ when $n \to \infty$ for $t \in [0, \tau]$ or $P(Y = 1|Z, \boldsymbol{V}) \to 0$). In practice, the violation of rare disease assumption will lead to bias in the estimators from Method 2-4 for binary and time-to-event outcome models (i.e., $\theta_2^*, \theta_3^*, \theta_4^*$ as shown in theorems below can be different from θ), but the scale of the bias from Method 2-4 is usually smaller than that of Method 1 based on our numerical studies. Intuitively, Method 4 can be less efficient compared with Method 2 and 3 since it ignores the information contained in $\boldsymbol{W}$. Also, Method 2 can be less efficient than Method 3 when the controlled feeding study does not modify individuals' self-reported behavior as it ignores the information of Q in the first sample. However, if the relationships between the self-reported dietary intake and the short-term dietary intake are different before and after the controlled feeding study, then both Method 3 and Method 4 will yield biased results, while Method 2 can still produce a valid estimator. We will illustrate these properties via simulation studies in the next section. Here we summarize the asymptotic results in the following theorems with the proofs given in Appendix A.

Theorem A1: With $\frac{n_3}{n_2} \to C_2$ and $\frac{n_2}{n_1} \to C_1$, we have $\sqrt{n_3}(\hat{\theta}_1 - \theta_1^*) \to N(0, \Sigma_{\theta 1})$ where $\Sigma_{\theta 1} = I_{\theta 1}^{-1}(I_{\theta 1} + C_2 I_{\gamma 1}\Sigma_{\gamma 1} I_{\gamma 1}^T) I_{\theta 1}^{-T}$ can be consistently estimated by $\hat{\Sigma}_{\theta 1} = \hat{I}_{\theta 1}^{-1}(\hat{I}_{\theta 1} + \frac{n_3}{n_2}\hat{I}_{\gamma 1}\hat{\Sigma}_{\gamma 1}\hat{I}_{\gamma 1}^T)\hat{I}_{\theta 1}^{-T}$ where the detail expression of $I_{\theta 1}$ and $I_{\gamma 1}$ are different for different types of regression models. The detailed expressions are defined in the proofs in Appendix A (Theorem 5 and 6).

Theorem A2: With $\frac{n_3}{n_2} \to C_2$ and $\frac{n_2}{n_1} \to C_1$, we have $\sqrt{n_3}(\hat{\theta}_2 - \theta_2^*) \to N(0, \Sigma_{\theta 2})$ where $\Sigma_{\theta 2} = I_{\theta 2}^{-1}(I_{\theta 2} + C_2 I_{\gamma 2}\Sigma_{\gamma 2} I_{\gamma 2}^T) I_{\theta 2}^{-T}$ can be consistently estimated by $\hat{\Sigma}_{\theta 2} = \hat{I}_{\theta 2}^{-1}(\hat{I}_{\theta 2} + \frac{n_3}{n_2}\hat{I}_{\gamma 2}\hat{\Sigma}_{\gamma 2}\hat{I}_{\gamma 2}^T)\hat{I}_{\theta 2}^{-T}$ where the detail expression of $I_{\theta 2}$ and $I_{\gamma 2}$ are different for different types of regression models. The detailed expressions are defined in the proofs in Appendix A (Theorem 5 and 6).

Theorem A3: With $\frac{n_3}{n_2} \to C_2$ and $\frac{n_2}{n_1} \to C_1$, we have $\sqrt{n_3}(\hat{\theta}_3 - \theta_3^*) \to N(0, \Sigma_{\theta 3})$ where $\Sigma_{\theta 3} = I_{\theta 2}^{-1}(I_{\theta 2} + C_2 I_{\gamma 3}\Sigma_{\gamma 3} I_{\gamma 3}^T) I_{\theta 2}^{-T}$ can be consistently estimated by $\hat{\Sigma}_{\theta 3} = \hat{I}_{\theta 3}^{-1}(\hat{I}_{\theta 3} + \frac{n_3}{n_2}\hat{I}_{\gamma 3}\hat{\Sigma}_{\gamma 3}\hat{I}_{\gamma 3}^T)\hat{I}_{\theta 3}^{-T}$ where the detail expression of $I_{\theta 3}$ and $I_{\gamma 3}$ are different for different types of regression models. The detailed expressions are defined in the proofs in Appendix A (Theorem 5 and 6).

Theorem A4: With $\frac{n_3}{n_2} \to C_2$ and $\frac{n_2}{n_1} \to C_1$, we have $\sqrt{n_3}(\hat{\theta}_4 - \theta_4^*) \to N(0, \Sigma_{\theta 4})$ where $\Sigma_{\theta 4} = I_{\theta 2}^{-1}(I_{\theta 2} + C_1 C_2 I_{\gamma 4}\Sigma_{\gamma 4} I_{\gamma 4}^T) I_{\theta 2}^{-T}$ can be consistently estimated by $\hat{\Sigma}_{\theta 4} =$

$\hat{I}_{\theta 4}^{-1}(\hat{I}_{\theta 4} + \frac{n_3}{n_1}\hat{I}_{\gamma 4}\hat{\Sigma}_{\gamma 4}\hat{I}_{\gamma 4}^{T})\hat{I}_{\theta 4}^{-T}$ where the detail expression of $I_{\theta 4}$ and $I_{\gamma 4}$ are different for different types of regression models. The detailed expressions are defined in the proofs in Appendix A (Theorem 5 and 6).

Here we compare the efficiency of the estimators from Method 3 and Method 4. For simplicity, we assume all variables are centered and only compare the efficiency in estimating $\hat{Z}$ because the variance of $\hat{\theta}$ is a monotone function of the variance of $\hat{Z}$. We compare the expected variance under fixed design. For Method 4,

$$E[Var(\hat{Z}_4|Q,\boldsymbol{V})] = E[n_1^{-1}(1,Q,\boldsymbol{V}^T)\left\{(1,Q,\boldsymbol{V}^T)^T(1,Q,\boldsymbol{V}^T)\right\}^{-1}(1,Q,\boldsymbol{V}^T)^T Var(X^*|Q,\boldsymbol{V})],$$

and for Method 3,

$$\begin{aligned} & E[Var(\hat{Z}_3|Q,\boldsymbol{V})] \\ = \; & E[n_2^{-1}(1,Q,\boldsymbol{V}^T)\left\{(1,Q,\boldsymbol{V}^T)^T(1,Q,\boldsymbol{V}^T)\right\}^{-1}(1,Q,\boldsymbol{V}^T)^T Var(X^*|Q,\boldsymbol{V})] \\ & +(n_1^{-1}-n_2^{-1})(1,Q,\boldsymbol{V}^T)\left\{(1,Q,\boldsymbol{V}^T)^T(1,Q,\boldsymbol{V}^T)\right\}^{-1}(1,Q,\boldsymbol{V}^T)^T Var(X^*|Q,\boldsymbol{V},\boldsymbol{W}) \\ = \; & n_1^{-1}(1,Q,\boldsymbol{V}^T)\left\{(1,Q,\boldsymbol{V}^T)^T(1,Q,\boldsymbol{V}^T)\right\}^{-1}(1,Q,\boldsymbol{V}^T)^T Var(X^*|Q,\boldsymbol{V}) \\ & \times\left\{\frac{n_1}{n_2}+(1-\frac{n_1}{n_2})(1-R^2_{(X^*,\boldsymbol{W})|Q,\boldsymbol{V}})\right\}. \end{aligned}$$

Therefore the relative efficiency between the two methods is

$$\frac{n_1}{n_2}+(1-\frac{n_1}{n_2})(1-R^2_{(X,\boldsymbol{W})|Q,\boldsymbol{V}}).$$

This shows that the key quantity to quantify the usefulness of biomarker $\boldsymbol{W}$ is $R^2_{(X,\boldsymbol{W})|Q,\boldsymbol{V}}$. The closer the value of $R^2_{(X,\boldsymbol{W})|Q,\boldsymbol{V}}$ is towards 0, the ‘weaker’ the biomarker is; the closer it is towards 1, the ‘stronger’ the biomarker is.

Let's look at two extreme examples: (1) when $R^2_{(X,\boldsymbol{W})|Q,\boldsymbol{V}} = 0$, the relative efficiency is 1; Method 3 does not have any efficiency gain compared with Method 4; the biomarker is completely useless. (2) when $R^2_{(X,\boldsymbol{W})|Q,\boldsymbol{V}} = 1$, the relative efficiency is $\frac{n_1}{n_2}$. In such a case, we have observed all X information in NPAAS dataset and the efficiency gain is proportional to the sample size gain. The asymptotic efficiencies comparing $\hat{\theta}_3$ and $\hat{\theta}_4$ are always smaller than 1, which indicates the loss of efficiency due to ignoring the second dataset.

2.4 Simulation

We performed simulations to study the finite sample behavior of our proposed estimators. We generate exposure and covariates from the following models:

$$
\begin{aligned}
&(Z, V) \sim N\left(0, \begin{pmatrix} 1-\sigma_x^2 & \rho \\ \rho & 1 \end{pmatrix}\right), \\
&W = b_0 + b_1 X + b_2 V + \epsilon_w, \\
&X = Z + \epsilon_x, \\
&X^* = X + \epsilon_x^*, \\
&Q = a_0 + a_1 Z + a_2 V + \epsilon_q,
\end{aligned}
$$

where ϵ_w, ϵ_x, ϵ_x^* and ϵ_q are independently sampled from normal distributions with mean zero and standard deviations σ_w, σ_x, σ_x^* and σ_q.

Then we generate the final outcome from linear, logistic and Cox regression model with continuous, binary and time-to-event endpoints, respectively. With continuous endpoint, we have:

$$Y = \theta_0 + \theta_z Z + \theta_v V + \epsilon_y,$$

where ϵ_y is drawn from a $N(0, 1.8)$.

With binary endpoint , we have:

$$logit(P(Y = 1|Z, V)) = \theta_0 + \theta_z Z + \theta_v V.$$

With time-to-event endpoint, we have:

$$\lambda(t|Z, X, V, W, Q) = \lambda(t|Z, V) = \lambda_0(t) \exp(\theta_z Z + \theta_v V).$$

For all three models, sample size with $n_1 = 150$, $n_2 = 300$, $n_3 = 5150$ denoted by N_1 and $n_1 = 300$, $n_2 = 600$, $n_3 = 10300$ denoted by N_2 are used to simulate the data. We set $b_0 = 5$, $b_2 = 1$, $\sigma_x = 0.2$, $\sigma_x^* = 0.5$, $\theta_0 = 1$, $\theta_z = 0.4$, $\theta_v = 0.6$ and $\sigma_w = 1$. Specifically, the censoring time is sampled from a mixture of $Unif(0, 10)$ and a point mass at 10 with equal probability and we set $\lambda_0(t) = 0.002t$ when we have the time-to-event endpoint . Then we change the values of b_1, ρ, a_1, a_2 and σ_q to change the range of the measurements including multiple coefficient of determination on long-term dietary intake quantified by biomarker, personal characteristic and long-term dietary intake quantified by biomarker given personal characteristic in the biomarker construction step, mathematically, $R^2_{ZWV} = 1 - \frac{Var(Z|W,V)}{Var(Z)}$, $R^2_{ZW|V} = 1 - \frac{Var(Z|W,V)}{Var(Z|V)}$; multiple coefficient of determination on long-term dietary intake by self-reported data, personal characteristic and long-term dietary intake by self-reported data given personal characteristic, mathematically, $R^2_{ZQV} = 1 - \frac{Var(Z|Q,V)}{Var(Z)}$, $R^2_{ZQ|V} = 1 - \frac{Var(Z|Q,V)}{Var(Z|V)}$; multiple coefficient of determination on long-term dietary intake by self-reported data, biomarker, personal characteristic and long-term dietary intake by self-reported data and biomarker given personal characteristic, mathematically, $R^2_{ZQWV} = 1 - \frac{Var(Z|W,Q,V)}{Var(Z)}$, $R^2_{ZQW|V} = 1 - \frac{Var(Z|W,Q,V)}{Var(Z|V)}$; multiple coefficient of determination on consumed dietary intake by biomaker, personal characteristic and consumed dietary intake by biomarker given personal characteristic, mathematically, $R^2_{X^*WV} = 1 - \frac{Var(X^*|W,V)}{Var(X^*)}$, $R^2_{X^*W|V} = 1 - \frac{Var(X^*|W,V)}{Var(X^*|V)}$; multiple coefficient of determination on consumed dietary intake by biomaker, self-reported data, personal characteristic and consumed dietary intake by biomarker and self-reported data given personal characteristic, mathematically, $R^2_{X^*WQV} = 1 - \frac{Var(X^*|W,Q,V)}{Var(X^*)}$, $R^2_{X^*WQ|V} = 1 - \frac{Var(X^*|W,Q,V)}{Var(X^*|V)}$, multiple coefficient of determination on estimated dietary intake with Method 2 by self-reported data, personal characteristic and estimated dietary intake by biomarker and self-reported data given personal characteristic, mathematically, $R^2_{\hat{X}_2QV} = 1 - \frac{Var(\hat{X}_2|Q,V)}{Var(\hat{X}_2)}$ and $R^2_{\hat{X}_2Q|V} = 1 - \frac{Var(\hat{X}_2|Q,V)}{Var(\hat{X}_2|V)}$. All such R^2 listed above are calculated to describe the strength of different variables for each method in different steps. For example, $R^2_{ZW|V}$ is related to the strength of biomarker on long-term dietary intake given personal characteristics for Method 2 in the biomarker construction step; $R^2_{X^*W|V}$ is related to the strength of biomarker for Method 2 on consumed dietary intake given personal characteristics in stage 1;$R^2_{\hat{X}_2QV}$ is related to the strength of FFQ on estimated dietary intake given personal characteristics in the regression calibration step for Method 2; $R^2_{ZWQ|V}$ and $R^2_{ZQ|V}$ are two quantities that are related to the strength of FFQ data and biomarker for Method 3 and only FFQ data for Method 4 in the biomarker construction step. Based on different levels of such quantities, six settings are selected.

In setting 1, 2 and 3, we fixed the effect on Q by setting $a_0 = 4$, $a_1 = 1.5$ and $\sigma_q = 3$. In setting 4, 5 and 6, we set $a_0 = 0.4$, $a_1 = 2$ and $\sigma_q = 4$. In addition, we decrease the coefficient of X on W from 1.3 to 0.8 in the first three settings while we decrease the coefficient of X on W from 1.1 to 0.5 in the last three settings. Table 1 displayed all different types of R^2 mentioned above for all six settings. To be more specific, by fixing the strength of self-reported data and the correlation between true dietary intake and subject characteristics, we gradually decreased the strength of the biomarker in the first three settings. In the last three settings, the correlation between true dietary intake and personal characteristic is set to be 0. The strength of self-reported data is again fixed but at a different level compared to the first three settings. We again decreased the strength of biomarker gradually in the last three settings. Below is the list of the six settings with varying parameters.

$b_1 = 1.3$, $\rho = 0.6$, $a_0 = 4$, $a_1 = 1.5$, $\sigma_q = 3$ (setting 1);
$b_1 = 1.1$, $\rho = 0.6$, $a_0 = 4$, $a_1 = 1.5$, $\sigma_q = 3$ (setting 2);
$b_1 = 0.8$, $\rho = 0.6$, $a_0 = 4$, $a_1 = 1.5$, $\sigma_q = 3$ (setting 3);
$b_1 = 1.1$, $\rho = 0$, $a_0 = 0.4$, $a_1 = 2$, $\sigma_q = 4$ (setting 4);
$b_1 = 0.8$, $\rho = 0$, $a_0 = 0.4$, $a_1 = 2$, $\sigma_q = 4$ (setting 5);
$b_1 = 0.5$, $\rho = 0$, $a_0 = 0.4$, $a_1 = 2$, $\sigma_q = 4$ (setting 6).

Table 2, 3, and 4 summarized simulation results comparing different methods' performance when the correlation structures between FFQ and true dietary intakes of the controlled feeding study and the full cohort are the same with continuous endpoint, binary endpoint and time-to-event endpoint, respectively for all six settings. The bias, mean estimated standard error (SE), empirical standard deviation (SD), and coverage rate (CR) of 95% nominal confidence interval for all four methods with original sample size (N_1) and enlarged sample size (N_2) from 1000 simulations have listed in all tables. In general, the results showed that our proposed estimators with Methods 2-4 behave well, while Method 1 tends to be biased even when the R^2 and partial R^2 is high in all cases. When the sample size is N_1, the CR for Method 1 is acceptable in many settings when endpoint-type is binary and time-to-event (i.e., left panel in Table 3 and 4). This may be due to the large variance with relatively small sample size. When the sample size is doubled, the variance becomes smaller and the bias dominates the error for Method 1, where lower

CR has been shown (i.e., right panel in Table 3 and 4). In general, when the correlation structure between Q and Z is the same in controlled feeding study and full cohort, bias is well controlled with Method 2-4. With relatively strong FFQ information (Q) shown in setting 4-6, Method 3 and Method 4 tend to provide more efficient results with smaller SD compared with Method 2. On the other hand, for settings with relatively weak FFQ information (setting 1-3), Method 2 is more efficient than Method 3 and 4. In addition, with a strong biomarker (i.e., setting 1 and 4), the efficiency of estimated parameters with Method 2 has shown to be comparable or even better with Method 3 and 4. This indicates the stronger biomarker employed in the model, the better efficiency with Method 2. Furthermore, Method 3 performs better than Method 4 under enlarged sample size, N_2, which is in accordance with our theoretical result for the asymptotic distributions shown in Appendix A. The performance of Method 3 and 4 are less sensitive to the strength of W compared with Method 2.

To evaluate the relationship of bias versus $R^2_{X^*WV}$ and bias versus $R^2_{X^*W|V}$, Figure 2 is displayed under the framework of setting 1 by varying the parameter, ρ, from 0 to 0.6 and setting $\sigma_w = 1$ and $\sigma_w = 1.7$, respectively. The plot of estimated bias for Method 1 with respect to the squared multiple correlation coefficient from the first stage ($R^2_{X^*WV}$) and squared multiple partial correlation coefficient given covariate V from the first stage ($R^2_{X^*W|V}$) is shown in Figure 2. Based on the plot, we can see that the bias is not a monotonic decreasing function of $R^2_{X^*WV}$, but is a decreasing function of $R^2_{X^*W|V}$, which is consistent with our theoretical derivation. Figure 2 suggests that the requirement $R^2 > 0.36$ (Lampe, 2017) is insufficient as one criterion to decide whether a biomarker is useful or not. In particular, the partial R^2 after given the effect of subject characteristics is an important factor influencing the bias of the current biomarker-based regression calibration for the association study.

Figure 3 shows the relationship between SD of the association parameter upon bias-corrected estimator and the $R^2_{X^*WV}$ and $R^2_{X^*W|V}$. Based on Figure 3, we can see that the SD is a decreasing function of $R^2_{X^*W|V}$ rather than $R^2_{X^*WV}$ which indicates that partial R^2 again is an essential factor in affecting SD rather than R^2. Therefore, to evaluate whether the calibration equation is useful, we should also focus on the partial R^2 instead of the R^2. Similar patterns for the other two estimators by adopting self-reported dietary data based on feeding study (Method 3 and Method 4) are shown in Figure 4. In conclusion, the precision is affected by partial R^2 ($R^2_{X^*W|V}$) rather than R^2 ($R^2_{X^*WV}$) itself.

Table 5, 6 and 7 summarized simulation results comparing different methods' performance when the correlation structures between FFQ and true dietary intakes of the controlled feeding study and the full cohort are different with continuous endpoint, binary endpoint and time-to-event endpoint, respectively under settings 1 and 4. With the correlation unequal between controlled feeding study and full cohort, Method 2 with BF involved gives more robust results in controlling bias compared with Method 3 and 4. Though bias is well controlled with Method 2, The under-coverage rates with less than 0.9 were shown in a few cases under binary endpoint and enlarged sample size (i.e., setting 4 in Table 6). This may be due to poor approximation with regression calibration on empirical SE, leading the bias-variance trade-off problem with relatively small SD. In general, as the difference of association between Q and Z increase from 10% to 50%, the performance of both Method 3 and Method 4 become worse (large bias) while the performance in controlling the bias of the estimator derived based on Method 2 is consistently good in most cases under different types of endpoints. Furthermore, we noticed that Method 2 attains more efficiency in estimated association parameters compared with Method 3 and 4 in most cases. Overall, the performance of Method 2 is adequate even when partial R^2s (i.e., Setting 3: $R^2_{X^*W|V} = 0.21$; Setting 6: $R^2_{X^*W|V} = 0.16$) from the biomarker construction step and the calibration equation building step are both low, which suggests that in the real application of Method 2, one may not need to be too stringent on the threshold of partial R^2.

2.5 Data Analysis

We illustrate our methods with the WHI NPAAS feeding study ($n = 153$), NPAAS biomarker study ($n = 450$) and the full WHI cohort data ($n = 161,808$). These three datasets are not mutually exclusive. Asymptotic normality is still followed and is not relying on the mutually exclusive assumption. In addition, bootstrap was utilized to obtain valid SE. The log-transformed self-reported sodium and potassium intakes from FFQ were used as Q. Variables including age, BMI, race/ethnicity, education level, self-reported physical activity, and smoking status are set as $\boldsymbol{V}$; the 24-hour urine sodium and potassium measurements are set as $\boldsymbol{W}$ and the disease outcomes are different types of CVD, including total coronary heart disease (CHD) and its myocardial infarction (MI) and coronary death components, total stroke and its hemorrhagic and ischemic components, total CVD comprised of CHD and stroke, CABG and PCI, and total CVD that also includes CABG and PCI, and heart failure. Using the log-transformed sodium-to-potassium ratio as a single predictor, we obtained $R^2 = 0.36$, which increased to $R^2 = 0.38$ when we used the log sodium and the log potassium as separate predictors. Therefore, we used these two measurements as two predictors in analyzing the data. The further inclusion of personal characteristics increased the R^2 to 0.45 with a partial R^2 conditional on personal characteristics to be 0.37.

For the feeding study, moderate measurement errors in the assessed consumed dietary data exist. So we differentiated the short term dietary intake, X, and the observed consumed dietary intake, X^*. Specifically, the adjusted bias factor can be estimated by $\widehat{BF} = 1 - \frac{\widehat{Var}(X^*|\boldsymbol{W},\boldsymbol{V}) - \hat{\sigma}_x^{*2}}{\widehat{Var}(X^*|\boldsymbol{V}) - \hat{\sigma}_x^{*2}}$, where $\hat{\sigma}_x^{*2}$ was treated as a sensitivity parameter since there is not enough replication data to provide accurate estimations. Hence, $\hat{\sigma}_x^{*2}$ was set at several different levels and the most conservative estimate, $\hat{\sigma}_x^{*2} = 0$, was used to illustrate the potential bias.

The estimated hazard ratio (HR), according to a 20% increase in the sodium-to-potassium ratio, are shown in Table 8. From the result, we found that the naïve three-step approach (Method 1) over-estimated the HR. The bias factor was estimated as about 0.37 as the partial R^2 under the assumption that there was no measurement error in the consumed dietary data in the controlled feeding study, therefore it is the most conservative HR estimate. The HR estimated from Method 3 (three-step with FFQ approach) is about the same as the estimate using Method 2 with a BF around 0.78, which is equivalent to the case where approximately 50% of the variation in the estimated consumed diet data is from noise. This noise level is about the same across all disease outcomes and is consistent with our estimation from the total energy expenditure. So the three-step with FFQ approach provides an estimator which is very similar to the estimator from the three-step BF correction (Method 2) when we assume $\hat{\sigma}_x^{*2} = 0.5 \times Var(X^*)$. Comparing with the results from Prentice et al. (2017), our findings can be mostly matched, where the sodium-to-potassium ratio is positively associated with the risk of CHD, nonfatal MI, coronary death, ischemic stroke, total CVD, coronary revascularization, non-revascularization are negatively associated with Hemorrhagic stroke. However, the conservative lower bounds of such effects are much smaller than presented before. Method 3 gives a slightly wider confidence interval compared with those in Prentice et al. (2017) [14], however, the difference with respect to point estimate is not statistically significant. Method 2 with 50% error assumption provides a point estimate and a confidence interval that is closed to the results with Method 3. This is consistent with what we have observed from our simulation. For Method 2 with no error assumption, the confidence interval is narrower than Method 3 because of the shrinkage effect on over-estimated BF. However, since the assumption of no error is obviously implausible, it leads to biased results towards the null. In such case, the comparison of efficiency is not of much interest. Method 4 has a wider confidence interval compared with Method 3, indicating that the urine biomarker is 'strong' and provides independent information beyond the self-reported data.

2.6 Discussion

In this study, we carefully examined the requirement for a valid biomarker for regression calibration purposes. Specifically, we showed that the methods without bias correction (Method 1, 3, 4) could lead to severe bias when the $R^2_{1|\boldsymbol{V}}$ is low or when the association between the true dietary intake and the self-reported diet is very different from the feeding study to the cohort. Our proposed BF corrected biomarker can solve this problem and lead to consistent association estimation when regression calibration is used to handle the systematic measurement error.

In conclusion, Method 1 should not be used due to its large bias. All rest three methods have their advantages/disadvantages. Method 4 is the simplest one in design and requires fewest assumptions. However, it depends on the availability of a strong dietary instrument in the feeding study with large sample size to accurately characterize the association between the dietary instrument and the true dietary intake. Method 3 is a three-step approach and allows the use of biomarker information efficiently. It is robust to the measurement error in the assessed diet from the controlled feeding study. This method works well under the assumption that the dietary instrument and the true dietary intake between the cohort and the controlled feeding study subgroup in the WHI data, where the self-reported dietary information (served as the dietary instrument) was collected long before the controlled feeding study. On the other hand, if such dietary instruments are not available, we need to consider Method 2. Method 2 does not use dietary instrument information in the biomarker development stage, and it depends more on the biological association between the biomarker and the dietary intakes. Therefore, as long as the model is correctly specified, the same biomarker can be used to build calibration equations for the dietary instruments that are available for the cohort but are not necessarily measured in the controlled feeding study (such as 4-day food record (4DFR)).

One caveat for Method 2 is that unlike the other methods (Method 3 and 4), which only require the measurement error of the assessed diet to be mean zero, this method further requires the variation of noise in the assessed diet (σ_x^{*2}) to be identifiable. The major portion of this variation might be from the inaccuracy of the records from the nutritional database (a bag of chips labeled with 100 cal might be 90 cal or 110 cal). Ideally, if this variation information can be added to the nutritional database, then the problem will be solved. This might be done by analyzing multiple samples for each type of food used in the feeding study during the nutritional analysis stage and reporting the standard error for each food type.

The σ_x^* was set to be 0.5 and fixed in our simulation settings. However, in a real-word example, a true σ_x^* is usually unknown. One advantage with Method 3 is including crucial FFQ in the controlled feeding study. A BF in Method 2 was derived to supplement the lack of information on FFQ in the controlled feeding study. Hence by setting $\hat{\theta}_1 = \hat{\theta}_3/BF$, we can get an estimated σ_x^* by solving the equation. We know BF is a function of σ_x^*. With some transformations, we have

$$\hat{\sigma_x^*} = \sqrt{\frac{\hat{\theta}_1}{\hat{\theta}_3}\left(Var(X^*|\boldsymbol{W},\boldsymbol{V}) - Var(X^*|\boldsymbol{V})(1-\frac{\hat{\theta}_3}{\hat{\theta}_1})\right)}$$

One concern using $\hat{\sigma_x^*}$ instead of using the true σ_{x^*} in our simulation setting is Method 2 may generate large variance leading poor efficiency as compared with Method 3 in the end. As an extension, we also investigated the performance of Method 2 with $\hat{\sigma_x^*}$. Setting 1 and 4 were selected since partial R^2 of X^* on $\boldsymbol{W}$ given $\boldsymbol{V}$ are greater than 0.36 in these two settings. Then we modified σ_q while keeping other parameters unchanged. To better understand, we denote FFQ in the controlled feeding study as Q_1 and FFQ in the full cohort as Q_2. The efficiency of Method 2 with $\hat{\sigma_x^*}$ involved and Method 3 with Q_1 involved were compared. Table 9 displayed the bias, SD, R^2_{ZQV} and $R^2_{ZQ|V}$ under different σ_q and corresponding R^2 and partial R^2_{ZQV} for Method 2 and 3 with the time-to-event endpoint. Based on Table 9, we can see partial R^2 of Z on Q given $\boldsymbol{V}$ ($R^2_{ZQ|V}$) is apparently lower than R^2 of Z on Q and V (R^2_{ZQV}) in setting 1 while R^2_{ZQV} and $R^2_{ZQ|V}$ are approximately equal to each other in setting 4, giving two different levels of association between $\boldsymbol{V}$ and Z. Method 2 can provide better efficiency (smaller SD) than Method 3 on the estimated associated parameter when the strength of Q_2 is large enough. Specifically, with $R^2_{ZQ|V} = 0.25$ in Method 2 and $R^2_{ZQ|V} = 0.13$ in Method 3, SD of Method 2 is 0.222, which is lower than SD of Method 3, 0.399. Similar trends on SD can be found under other scenarios. This indicates the efficiency of Method 2 can be improved and comparable with Method 3 using $\hat{\sigma_x^*}$.

TABLE 1: List of R^2 and partial R^2 among different measurements under 6 simulation settings

	Setting 1	Setting 2	Setting 3	Setting 4	Setting 5	Setting 6
R^2_{ZWV}	0.68	0.64	0.55	0.53	0.38	0.19
$R^2_{ZW\|V}$	0.49	0.41	0.28	0.53	0.38	0.19
R^2_{ZQV}	0.46	0.46	0.46	0.20	0.20	0.20
$R^2_{ZQ\|V}$	0.13	0.13	0.13	0.20	0.20	0.20
R^2_{ZQWV}	0.71	0.67	0.60	0.58	0.46	0.33
$R^2_{ZQW\|V}$	0.53	0.46	0.35	0.58	0.46	0.33
$R^2_{X^*WV}$	0.56	0.52	0.44	0.44	0.32	0.16
$R^2_{X^*W\|V}$	0.38	0.32	0.21	0.44	0.32	0.16
$R^2_{X^*WQV}$	0.58	0.54	0.48	0.48	0.38	0.26
$R^2_{X^*WQ\|V}$	0.40	0.35	0.26	0.48	0.38	0.26
$R^2_{\hat{X}_2QV}$	0.59	0.92	0.98	0.11	0.08	0.04
$R^2_{\hat{X}_2Q\|V}$	0.07	0.06	0.04	0.11	0.08	0.04

Table 2: Simulation results comparing different methods' performance when the correlation structures between the FFQ and true dietary intakes of the controlled feeding study and the full cohort are the same with continuous endpoint

Setting	Method	N_1				N_2			
		Bias	SE	SD	CR	Bias	SE	SD	CR
1	1	0.51	0.336	0.338	0.85	0.40	0.192	0.226	0.46
	2	0.06	0.171	0.176	0.94	0.01	0.098	0.101	0.93
	3	0.06	0.181	0.197	0.95	0.01	0.097	0.090	0.96
	4	0.07	0.189	0.182	0.94	0.01	0.104	0.101	0.94
2	1	0.71	0.448	0.448	0.81	0.56	0.247	0.301	0.25
	2	0.07	0.190	0.196	0.95	0.01	0.105	0.108	0.92
	3	0.07	0.187	0.204	0.94	0.01	0.098	0.092	0.96
	4	0.07	0.189	0.182	0.94	0.01	0.104	0.101	0.94
3	1	0.45	0.577	0.560	0.97	1.08	0.464	0.612	0.09
	2	0.03	0.293	0.294	0.96	0.01	0.126	0.129	0.91
	3	0.05	0.321	0.428	0.97	0.01	0.100	0.094	0.96
	4	0.04	0.315	0.347	0.97	0.01	0.104	0.101	0.94
4	1	0.41	0.209	0.212	0.52	0.35	0.132	0.162	0.14
	2	0.03	0.112	0.117	0.95	0.00	0.071	0.078	0.93
	3	0.03	0.104	0.111	0.93	0.01	0.067	0.064	0.96
	4	0.03	0.112	0.113	0.97	0.01	0.073	0.074	0.94
5	1	0.77	0.360	0.361	0.34	0.66	0.219	0.273	0.01
	2	0.04	0.135	0.141	0.95	0.00	0.082	0.089	0.91
	3	0.03	0.107	0.115	0.92	0.01	0.068	0.065	0.96
	4	0.03	0.112	0.113	0.97	0.01	0.073	0.074	0.94
6	1	2.14	1.282	1.384	0.43	1.77	0.635	0.809	0.01
	2	0.08	0.233	0.256	0.93	0.00	0.116	0.122	0.90
	3	0.03	0.111	0.117	0.95	0.01	0.070	0.068	0.96
	4	0.03	0.112	0.113	0.97	0.01	0.073	0.074	0.94

Table 3: Simulation results comparing different methods' performance when the correlation structures between the FFQ and true dietary intakes of the controlled feeding study and the full cohort are the same with binary endpoint

Setting	Method	N_1				N_2			
		Bias	SE	SD	CR	Bias	SE	SD	CR
1	1	0.48	0.407	0.407	0.98	0.34	0.234	0.235	0.77
	2	0.04	0.206	0.195	1.00	-0.02	0.120	0.112	0.93
	3	0.04	0.211	0.202	0.99	-0.01	0.121	0.121	0.92
	4	0.05	0.222	0.206	0.96	0.00	0.127	0.141	0.89
2	1	0.67	0.530	0.547	0.97	0.49	0.293	0.300	0.69
	2	0.05	0.223	0.215	1.00	-0.02	0.125	0.116	0.93
	3	0.04	0.216	0.209	0.98	-0.01	0.122	0.123	0.91
	4	0.05	0.222	0.206	0.96	0.00	0.127	0.141	0.89
3	1	0.64	0.727	0.701	0.97	0.98	0.513	0.565	0.51
	2	0.04	0.308	0.311	0.96	-0.02	0.141	0.130	0.93
	3	0.06	0.330	0.453	0.97	-0.01	0.123	0.127	0.93
	4	0.04	0.315	0.347	0.97	0.00	0.127	0.141	0.89
4	1	0.37	0.245	0.273	0.82	0.32	0.157	0.179	0.45
	2	0.01	0.131	0.139	0.96	-0.01	0.084	0.089	0.89
	3	0.00	0.123	0.124	0.96	0.00	0.082	0.087	0.91
	4	0.01	0.131	0.139	0.94	0.00	0.087	0.102	0.93
5	1	0.72	0.405	0.464	0.74	0.63	0.248	0.285	0.18
	2	0.02	0.151	0.166	0.95	-0.01	0.094	0.098	0.89
	3	0.00	0.126	0.131	0.96	0.00	0.083	0.090	0.94
	4	0.01	0.131	0.139	0.94	0.00	0.087	0.102	0.93
6	1	2.06	1.371	1.695	0.77	1.70	0.667	0.782	0.01
	2	0.06	0.247	0.291	0.93	-0.01	0.123	0.124	0.85
	3	0.01	0.129	0.139	0.96	0.00	0.085	0.094	0.94
	4	0.01	0.131	0.139	0.94	0.00	0.087	0.102	0.93

Table 4: Simulation results comparing different methods' performance when the correlation structures between the FFQ and true dietary intakes of the controlled feeding study and the full cohort are the same with time-to-event endpoint

Setting	Method	N_1				N_2			
		Bias	SE	SD	CR	Bias	SE	SD	CR
1	1	0.48	0.564	0.558	0.97	0.40	0.362	0.374	0.87
	2	0.05	0.290	0.286	0.97	0.01	0.187	0.191	0.96
	3	0.07	0.470	0.649	0.97	0.01	0.186	0.191	0.97
	4	0.05	0.299	0.299	0.98	0.02	0.194	0.204	0.97
2	1	0.67	0.721	0.727	0.97	0.56	0.445	0.460	0.84
	2	0.06	0.309	0.309	0.97	0.01	0.192	0.196	0.97
	3	0.07	0.625	0.838	0.97	0.01	0.187	0.192	0.97
	4	0.05	0.299	0.299	0.98	0.02	0.194	0.204	0.97
3	1	1.38	1.816	1.967	0.98	1.09	0.739	0.763	0.79
	2	0.10	0.502	0.547	0.96	0.02	0.210	0.212	0.96
	3	0.07	0.545	0.732	0.97	0.01	0.189	0.195	0.97
	4	0.05	0.299	0.299	0.98	0.02	0.194	0.204	0.97
4	1	0.38	0.354	0.342	0.91	0.35	0.238	0.244	0.75
	2	0.03	0.192	0.190	0.96	0.01	0.129	0.130	0.96
	3	0.02	0.188	0.193	0.97	0.00	0.126	0.129	0.95
	4	0.02	0.192	0.187	0.97	0.01	0.130	0.135	0.95
5	1	0.73	0.551	0.551	0.87	0.66	0.357	0.369	0.58
	2	0.03	0.211	0.217	0.96	0.01	0.137	0.138	0.96
	3	0.02	0.190	0.197	0.97	0.00	0.127	0.130	0.96
	4	0.02	0.192	0.187	0.97	0.01	0.130	0.135	0.95
6	1	2.17	2.306	2.865	0.91	1.77	0.861	0.900	0.42
	2	0.09	0.452	0.595	0.94	0.02	0.165	0.166	0.96
	3	0.02	0.192	0.198	0.97	0.01	0.129	0.132	0.96
	4	0.02	0.192	0.187	0.97	0.01	0.130	0.135	0.95

Table 5: Simulation results comparing different methods' performance when the correlation structures between the short term dietary intake and true dietary intakes of the controlled feeding study and the full cohort are different with continuous endpoint

Setting	Difference	Method	N_1				N_2			
			Bias	SE	SD	CR	Bias	SE	SD	CR
1	0.1	1	0.51	0.336	0.338	0.85	0.40	0.192	0.226	0.46
	0.1	2	0.06	0.171	0.176	0.94	0.01	0.098	0.101	0.93
	0.1	3	0.09	0.205	0.224	0.95	0.03	0.104	0.098	0.96
	0.1	4	0.12	0.237	0.226	0.96	0.05	0.121	0.119	0.97
	0.3	1	0.51	0.336	0.338	0.85	0.40	0.192	0.226	0.46
	0.3	2	0.06	0.171	0.176	0.94	0.01	0.098	0.101	0.93
	0.3	3	0.16	0.284	0.315	0.99	0.07	0.125	0.118	0.99
	0.3	4	0.32	0.519	0.470	1.00	0.16	0.188	0.185	0.99
	0.5	1	0.51	0.336	0.338	0.85	0.40	0.192	0.226	0.46
	0.5	2	0.06	0.171	0.176	0.94	0.01	0.098	0.101	0.93
	0.5	3	0.29	0.537	0.566	1.00	0.13	0.158	0.151	1.00
	0.5	4	0.50	5.548	2.509	1.00	0.42	0.429	0.431	1.00
4	0.1	1	0.41	0.209	0.212	0.52	0.35	0.132	0.162	0.14
	0.1	2	0.03	0.112	0.117	0.95	0.00	0.071	0.078	0.93
	0.1	3	0.04	0.113	0.122	0.95	0.02	0.071	0.068	0.96
	0.1	4	0.06	0.132	0.132	0.98	0.04	0.084	0.085	0.95
	0.3	1	0.41	0.209	0.212	0.52	0.35	0.132	0.162	0.14
	0.3	2	0.03	0.112	0.117	0.95	0.00	0.071	0.078	0.93
	0.3	3	0.09	0.138	0.149	0.99	0.06	0.085	0.080	0.98
	0.3	4	0.17	0.210	0.205	1.00	0.14	0.124	0.124	0.99
	0.5	1	0.41	0.209	0.212	0.52	0.35	0.132	0.162	0.14
	0.5	2	0.03	0.112	0.117	0.95	0.00	0.071	0.078	0.93
	0.5	3	0.15	0.180	0.196	1.00	0.11	0.105	0.099	0.97
	0.5	4	0.44	0.503	0.452	1.00	0.33	0.239	0.234	0.99

Table 6: Simulation results comparing different methods' performance when the correlation structures between the short term dietary intake and true dietary intakes of the controlled feeding study and the full cohort are different with binary endpoint

Setting	Difference	Method	N_1				N_2			
			Bias	SE	SD	CR	Bias	SE	SD	CR
1	0.1	1	0.48	0.407	0.407	0.98	0.34	0.234	0.235	0.77
	0.1	2	0.04	0.206	0.195	1.00	-0.02	0.120	0.112	0.93
	0.1	3	0.06	0.233	0.225	0.99	0.01	0.129	0.129	0.95
	0.1	4	0.10	0.269	0.247	0.96	0.03	0.146	0.162	0.94
	0.3	1	0.48	0.407	0.407	0.98	0.34	0.234	0.235	0.77
	0.3	2	0.04	0.206	0.195	1.00	-0.02	0.120	0.112	0.93
	0.3	3	0.13	0.305	0.301	1.00	0.05	0.150	0.150	0.98
	0.3	4	0.29	0.531	0.454	0.99	0.14	0.214	0.235	0.99
	0.5	1	0.48	0.407	0.407	0.98	0.34	0.234	0.235	0.77
	0.5	2	0.04	0.206	0.195	1.00	-0.02	0.120	0.112	0.93
	0.5	3	0.24	0.523	0.508	1.00	0.11	0.183	0.183	0.98
	0.5	4	0.52	5.052	2.307	1.00	0.39	0.455	0.484	0.99
4	0.1	1	0.37	0.245	0.273	0.82	0.32	0.157	0.179	0.45
	0.1	2	0.01	0.131	0.139	0.96	-0.01	0.084	0.089	0.89
	0.1	3	0.02	0.131	0.133	0.97	0.01	0.087	0.092	0.93
	0.1	4	0.04	0.150	0.158	0.96	0.03	0.098	0.115	0.93
	0.3	1	0.37	0.245	0.273	0.82	0.32	0.157	0.179	0.45
	0.3	2	0.01	0.131	0.139	0.96	-0.01	0.084	0.089	0.89
	0.3	3	0.06	0.155	0.155	0.99	0.04	0.100	0.106	0.96
	0.3	4	0.14	0.225	0.229	1.00	0.12	0.139	0.161	0.97
	0.5	1	0.37	0.245	0.273	0.82	0.32	0.157	0.179	0.45
	0.5	2	0.01	0.131	0.139	0.96	-0.01	0.084	0.089	0.89
	0.5	3	0.12	0.193	0.191	0.99	0.09	0.120	0.126	0.97
	0.5	4	0.39	0.500	0.460	1.00	0.32	0.255	0.285	1.00

Table 7: Simulation results comparing different methods' performance when the correlation structures between the short term dietary intake and true dietary intakes of the controlled feeding study and the full cohort are different with time-to-event endpoint

Setting	Difference	Method	N_1				N_2			
			Bias	SE	SD	CR	Bias	SE	SD	CR
1	0.1	1	0.48	0.564	0.558	0.97	0.40	0.362	0.374	0.87
	0.1	2	0.05	0.290	0.286	0.97	0.01	0.187	0.191	0.96
	0.1	3	0.04	0.988	1.203	0.97	0.03	0.197	0.203	0.97
	0.1	4	0.10	0.352	0.357	0.98	0.06	0.223	0.240	0.97
	0.3	1	0.48	0.564	0.558	0.97	0.40	0.362	0.374	0.87
	0.3	2	0.05	0.290	0.286	0.97	0.01	0.187	0.191	0.96
	0.3	3	0.01	9.056	3.701	0.99	0.08	0.227	0.235	0.98
	0.3	4	0.25	0.662	0.684	0.99	0.18	0.326	0.367	0.99
	0.5	1	0.48	0.564	0.558	0.97	0.40	0.362	0.374	0.87
	0.5	2	0.05	0.290	0.286	0.97	0.01	0.187	0.191	0.96
	0.5	3	0.28	3.033	1.903	0.99	0.15	0.275	0.292	0.98
	0.5	4	0.60	12.144	4.627	1.00	0.37	17.903	7.286	1.00
4	0.1	1	0.38	0.354	0.342	0.91	0.35	0.238	0.244	0.75
	0.1	2	0.03	0.192	0.190	0.96	0.01	0.129	0.130	0.96
	0.1	3	0.04	0.199	0.209	0.98	0.02	0.132	0.135	0.96
	0.1	4	0.06	0.215	0.216	0.98	0.04	0.144	0.149	0.97
	0.3	1	0.38	0.354	0.342	0.91	0.35	0.238	0.244	0.75
	0.3	2	0.03	0.192	0.190	0.96	0.01	0.129	0.130	0.96
	0.3	3	0.08	0.232	0.269	0.99	0.06	0.148	0.152	0.97
	0.3	4	0.18	0.370	0.522	1.00	0.14	0.193	0.202	0.97
	0.5	1	0.38	0.354	0.342	0.91	0.35	0.238	0.244	0.75
	0.5	2	0.03	0.192	0.190	0.96	0.01	0.129	0.130	0.96
	0.5	3	0.16	0.351	0.570	0.99	0.11	0.173	0.178	0.97
	0.5	4	0.47	1.425	1.364	1.00	0.42	3.896	2.252	0.98

Table 8: Association between 20% increase in sodium-to-potassium ratio with various cardiovascular diseases

Outcome	Method 1		Method 3		Method 4	
	HR	95% CI	HR	95% CI	HR	95% CI
CHD	1.20	(1.06, 1.36)	1.16	(1.04, 1.28)	1.21	(1.03,1.43)
Nonfatal MI	1.22	(1.04, 1.42)	1.16	(1.04, 1.30)	1.19	(1.01,1.41)
Coronary death	1.21	(1.07, 1.38)	1.17	(1.03, 1.33)	1.27	(0.98,1.64)
Stroke	1.11	(1.02, 1.21)	1.10	(1.00, 1.20)	1.17	(1.02,1.34)
Ischemic Stroke	1.18	(1.06, 1.31)	1.15	(1.02, 1.29)	1.24	(1.05,1.46)
Hemorrhagic Stroke	0.86	(0.67, 1.10)	0.90	(0.73, 1.11)	0.92	(0.63,1.35)
Total CVD	1.15	(1.08, 1.22)	1.12	(1.03, 1.21)	1.17	(1.06,1.29)
Revascularization	1.21	(1.06, 1.37)	1.15	(1.04, 1.28)	1.18	(1.01,1.39)
Non-Revascularization	1.15	(1.05, 1.26)	1.12	(1.03, 1.22)	1.18	(1.04,1.34)
Heart Failure	1.06	(0.94, 1.19)	1.03	(0.94, 1.14)	1.00	(0.84,1.18)
Outcome	Method 2 (no Error)		Method 2 (40% Error)		Method 2 (50% Error)	
	HR	95% CI	HR	95% CI	HR	95% CI
CHD	1.07	(1.02, 1.12)	1.13	(1.04, 1.23)	1.16	(1.04,1.30)
Nonfatal MI	1.08	(1.03, 1.13)	1.14	(1.03, 1.25)	1.17	(1.03,1.34)
Coronary death	1.07	(1.02, 1.14)	1.14	(1.02, 1.26)	1.17	(1.03,1.34)
Stroke	1.04	(1.00, 1.08)	1.07	(1.00, 1.15)	1.09	(1.00,1.18)
Ischemic Stroke	1.06	(1.01, 1.11)	1.11	(1.02, 1.21)	1.14	(1.03,1.27)
Hemorrhagic Stroke	0.94	(0.86, 1.03)	0.90	(0.77, 1.06)	0.88	(0.72,1.08)
Total CVD	1.05	(1.02, 1.09)	1.10	(1.03, 1.16)	1.12	(1.04,1.21)
Revascularization	1.07	(1.02, 1.12)	1.13	(1.03, 1.25)	1.17	(1.03,1.32)
Non-Revascularization	1.05	(1.02, 1.09)	1.10	(1.03, 1.17)	1.12	(1.03,1.22)
Heart Failure	1.02	(0.98, 1.07)	1.04	(0.96, 1.12)	1.05	(0.94,1.16)

Table 9: Simulation results comparing efficiency between Method 2 and Method 3 with $\hat{\sigma}_{x^*}$ used for Method 2 with time-to-event endpoint

					N_1		N_2	
Setting	$(\sigma_{q1}, \sigma_{q2})$	Method	R^2_{ZQV}	$R^2_{ZQ\|V}$	Bias	SD	Bias	SD
1	(2,3)	2	0.54	0.25	-0.01	0.222	0.01	0.167
		3	0.46	0.13	0.08	0.399	0.02	0.187
4	(2,3)	2	0.34	0.34	-0.01	0.196	0.00	0.170
		3	0.18	0.18	0.08	0.387	0.03	0.204

FIGURE 2: Bias from Method 1 in relation with R^2 and partial R^2

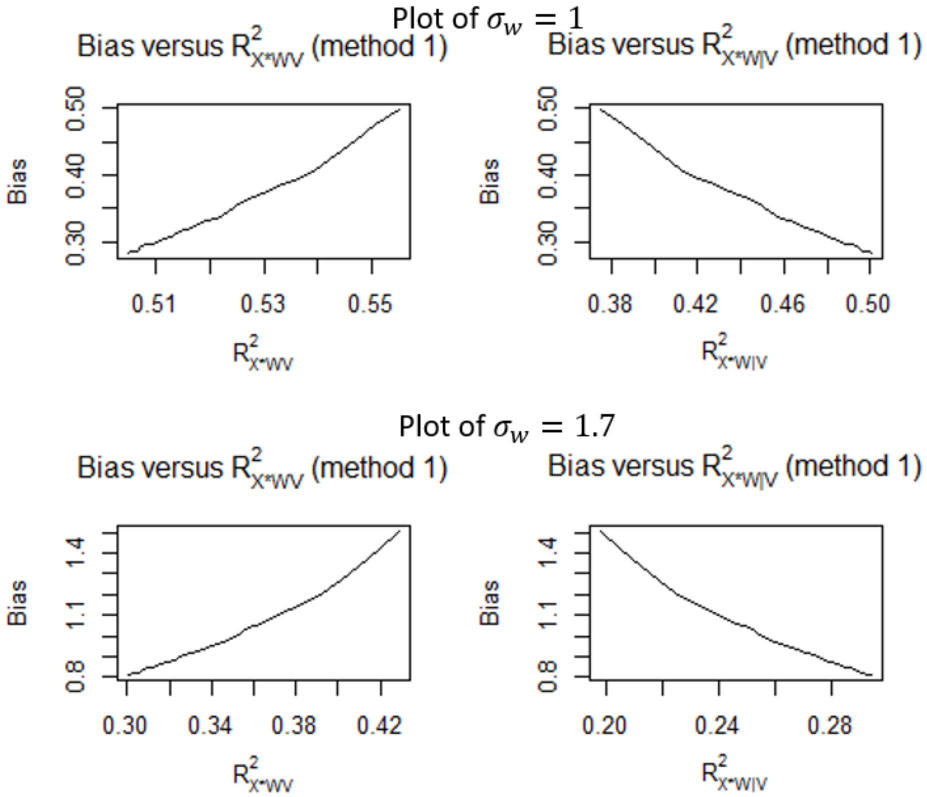


FIGURE 3: SD of Method 1 and Method 2 in relation with R^2 and partial R^2

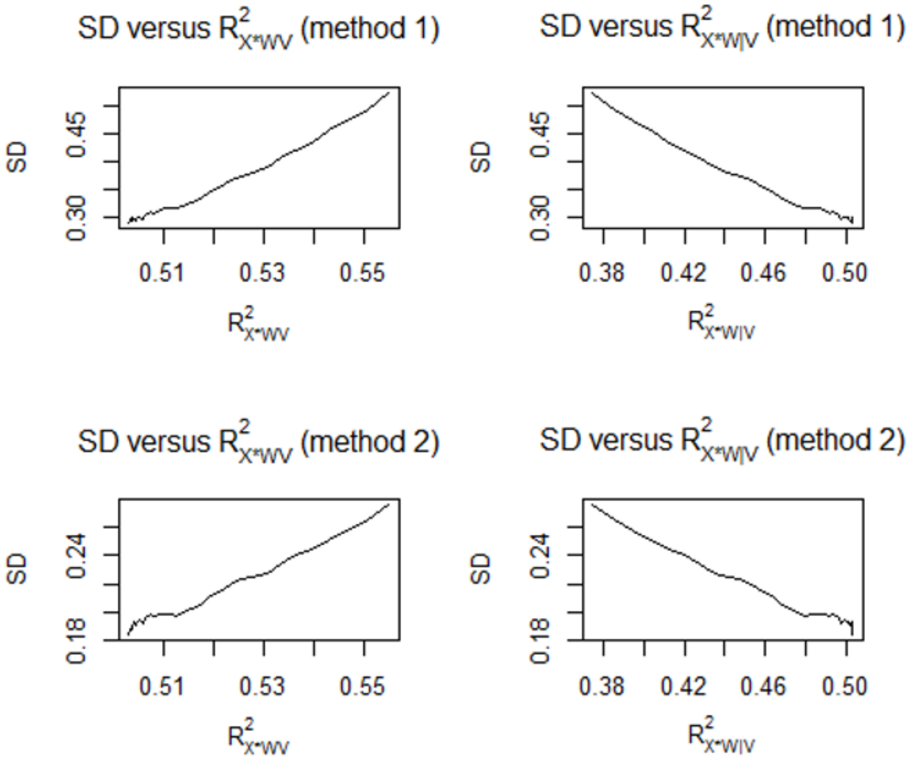
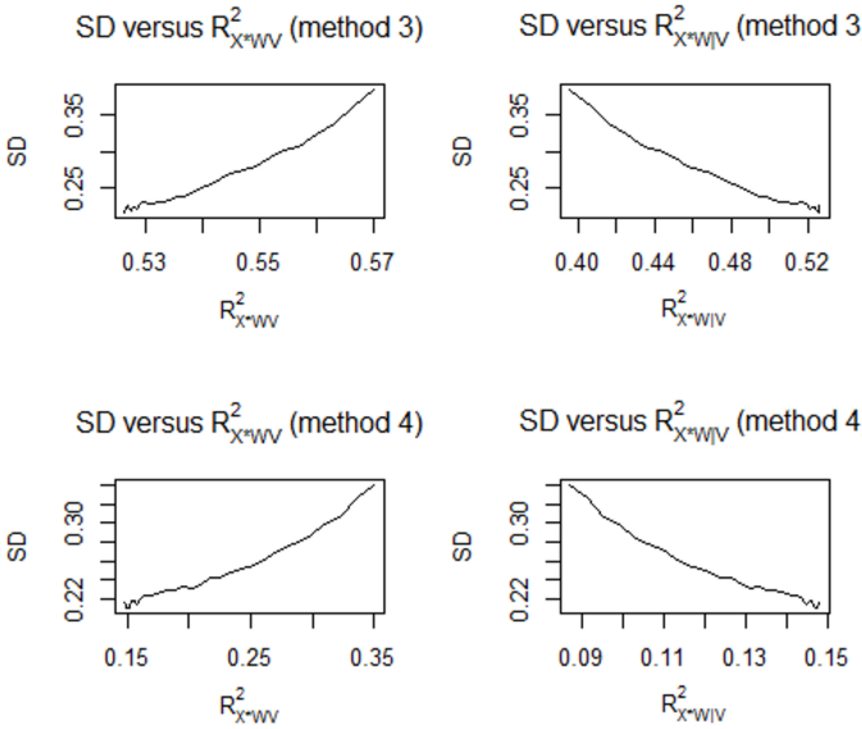


FIGURE 4: SD of Method 3 and Method 4 in relation with R^2 and partial R^2



Chapter 3: Low-dimensional Setting with Multiple Exposures

3.1 Introduction

In Chapter 2, we developed the calibration equation by focusing on a single exposure, sodium-to-potassium ratio. The association between sodium-to-potassium ratio and CVD risks has been evaluated in a recent WHI study using the regression calibration method where the nutrient intake was calibrated by a single measurement biomarker (from a single 24-hour urine collection) [14, 41, 42]. The previous study suggested positive associations between CVD and sodium-to-potassium ratio [43]. However, the single measurement is suboptimal for the performance of the regression calibration approach, as it has a low correlation with the true dietary intakes [43]. Sodium and potassium jointly can be set as two exposures and are found to have a different direction of associations with CVD in several studies. Recent study [44, 45, 46] showed systolic and diastolic blood pressure are positively associated with sodium and negatively associated with potassium. A J-shaped association was found between major cardiovascular events and sodium while no significant association with potassium. O'Donnell et al. (2014) [47] found individuals with 3-6 g of sodium excretion per day have a reduced risk of CVD. Lower risk of hypertension has been identified with a higher level of potassium intake and lower level of sodium intake in many observational studies and randomized trials [43, 48]. Long-term potassium substitution for sodium or sodium intake reduction may also lead to a lower risk of CVD. Other than the sodium-to-potassium ratio, the joint effect of sodium and potassium on CVD is also of particular interest. In this chapter, we will show that simply combining the univariate biomarkers developed from Chapter 2 is not appropriate. We developed biomarkers that can be used in the multivariate regression calibration method for multiple exposures.

3.2 Methods

With multiple exposures, a matrix form for the variance of consumed dietary intake, Σ_x^*, is considered. Similar to Chapter 2, we first consider the case where Σ_x^* is known. In the real data analysis where Σ_x^* is not available, we vary the parameters to perform sensitivity analysis. In this chapter, we conducted both multivariate analysis and univariate analysis for comparison. Detailed information is provided for each method below.

3.2.1 Method 1: The naïve three-step approach with multiple exposures

(i) **Multivariate approach for Method 1**

Considering the Method 1 with multiple exposures using the multivariate approach, we first perform a linear regression among n_1 subjects in the biomarker discovery sample of consumed diet ($\boldsymbol{X^*}$) on blood and urine measurements ($\boldsymbol{W}$) as well as subject characteristics ($\boldsymbol{V}$) to obtain:

$$\hat{\beta}_1 = \left\{ \sum_{i=1}^{n_1} (1, \boldsymbol{W}_i^T, \boldsymbol{V}_i^T)^T (1, \boldsymbol{W}_i^T, \boldsymbol{V}_i^T) \right\}^{-1} \left\{ \sum_{i=1}^{n_1} (1, \boldsymbol{W}_i^T, \boldsymbol{V}_i^T)^T \boldsymbol{X_i^*}^T \right\}.$$

Then we compute $\hat{\boldsymbol{X}}_{\boldsymbol{1i}} = [(\boldsymbol{1}, \boldsymbol{W}_i^T, \boldsymbol{V}_i^T)\hat{\beta}_1]^T$ for $i = n_1 + 1, \cdots, n_1 + n_2$ to predict the long-term dietary intake ($\boldsymbol{Z}$) among the n_2 calibration sample to predict the long-term dietary intake ($\boldsymbol{Z}$) among the n_2 calibration samples and run a regression of $\hat{\boldsymbol{X}}_{\boldsymbol{1}}$ on self-reported food frequency questionnaire data ($\boldsymbol{Q}$) and subject characteristics

$(\boldsymbol{V})$ to build calibration equation using the n_2 calibration samples with:

$$\hat{\gamma}_1 = \left\{ \sum_{i=n_1+1}^{n_1+n_2} (1, \boldsymbol{Q}_i^T, \boldsymbol{V}_i^T)^T (1, \boldsymbol{Q}_i^T, \boldsymbol{V}_i^T) \right\}^{-1} \left\{ \sum_{i=n_1+1}^{n_1+n_2} (1, \boldsymbol{Q}_i^T, \boldsymbol{V}_i^T)^T \hat{\boldsymbol{X}}_{\mathbf{1i}}^T \right\}.$$

Finally, we predict the exposure by $\hat{\boldsymbol{Z}}_{\mathbf{1i}} = [(1, \boldsymbol{Q}_i^T, \boldsymbol{V}_i^T)\hat{\gamma}_1]^T$ for $i = n_1 + n_2 + 1, \cdots, n_1 + n_2 + n_3$ in the n_3 association sample. Then a linear model, a logistic model or a Cox model of Y on $\hat{\boldsymbol{Z}}_{\mathbf{1}}$ and $\boldsymbol{V}$ are performed to estimate the association parameter $\hat{\theta}_1$ by solving the following score equations with respect to continuous, binary or time-to-event endpoint:

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \left[(\mathbf{1}, \hat{\boldsymbol{Z}}_{\mathbf{1i}}^T, \boldsymbol{V}_i^T)^T \left\{ Y_i - (\mathbf{1}, \hat{\boldsymbol{Z}}_{\mathbf{1i}}^T, \boldsymbol{V}_i^T)\theta \right\} \right],$$

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \left[(1, \hat{\boldsymbol{Z}}_{\mathbf{1i}}^T, \boldsymbol{V}_i^T)^T \left\{ Y_i - \frac{exp((1, \hat{\boldsymbol{Z}}_{\mathbf{1i}}^T, \boldsymbol{V}_i^T)\theta)}{1 + exp((1, \hat{\boldsymbol{Z}}_{\mathbf{1i}}^T, \boldsymbol{V}_i^T)\theta)} \right\} \right],$$

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \int_0^{\tau} \left[\begin{pmatrix} \hat{\boldsymbol{Z}}_{\mathbf{1i}} \\ \boldsymbol{V}_i \end{pmatrix} - \sum_j \frac{Y_j(t) \exp\left\{ (\hat{\boldsymbol{Z}}_{\mathbf{1j}}^T, \boldsymbol{V}_j^T)\theta \right\}}{\sum_k Y_k(t) \exp\left\{ (\hat{\boldsymbol{Z}}_{\mathbf{1k}}^T, \boldsymbol{V}_k^T)\theta \right\}} \begin{pmatrix} \hat{\boldsymbol{Z}}_{\mathbf{1j}} \\ \boldsymbol{V}_j \end{pmatrix} \right] dN_i(t), \tag{3.1}$$

where τ is a pre-specified large number and we assume $P(C > \tau) > 0$.

We show in Appendix B (Theorem 7) that $E(\hat{\boldsymbol{Z}}_{\mathbf{1}}|\boldsymbol{Q}, \boldsymbol{V}) = \boldsymbol{BF} \times E(\boldsymbol{Z}|\boldsymbol{Q}, \boldsymbol{V}) + (1 - \boldsymbol{BF}) \times E(\boldsymbol{Z}|\boldsymbol{V}) \neq E(\boldsymbol{Z}|\boldsymbol{Q}, \boldsymbol{V})$ for multiple exposures, where the bias factor $(\boldsymbol{BF})$ is defined as:

$$\boldsymbol{BF} = I_K - Var(\boldsymbol{X}|\boldsymbol{V})^{-1} Var(\boldsymbol{X}|\boldsymbol{V}, \boldsymbol{W}).$$

Such BF will lead to bias in the estimation of association parameter and if we further assume $E(\boldsymbol{X}|\boldsymbol{V}) = \boldsymbol{V}\delta$, we show in Appendix B (Theorem 7) that the estimator $\hat{\theta}_1 \to \theta_1^*$ as $n \to \infty$, with $\theta_{1z}^* = (\boldsymbol{BF})^{-1}\theta_{1z}$ and $\theta_{1v}^* = \theta_{1v} - \boldsymbol{BF}^{-1}(I_K - \boldsymbol{BF})\theta_{1z}$.

(ii) **Univariate approach for Method 1**

Instead of performing multivariate analysis, we can also perform univariate linear regression regarding each element of $\boldsymbol{W}$ and $\boldsymbol{V}$ on $\boldsymbol{X}^*$ in the first step and similar procedures in the second step. Suppose $\boldsymbol{X^*} = (X_1^*...X_K^*)^T$, $\boldsymbol{Q} = (Q_1...Q_K)^T$ are K-dimensional, $\boldsymbol{W} = (W_1, ...W_p)^T$ is p-dimensional and $\boldsymbol{V} = (V_1, ..., V_q)^T$ is q-dimensional. For ease of interpretation and explanation, k is used to denote each element in the K-dimensional space and we used the same settings for the univariate approach in the rest of method sections. Then each element in $\hat{\beta}_1 = (\hat{\beta}_{11}, ...\hat{\beta}_{1K})^T$ can be estimated as below:

$$\hat{\beta}_{1k} = \left\{\sum_{i=1}^{n_1}(1, \boldsymbol{W}_i^T, \boldsymbol{V}_i{}^T)^T(1, \boldsymbol{W}_i^T, \boldsymbol{V}_i{}^T)\right\}^{-1}\left\{\sum_{i=1}^{n_1}(1, \boldsymbol{W}_i^T, \boldsymbol{V}_i{}^T)^T X_{ki}{}^*\right\},$$

$$\hat{X}_{1ki} = (1, \boldsymbol{W}_i^T, \boldsymbol{V}_i{}^T)\hat{\beta}_{1k},$$

for $i = n_1 + 1, ..., n1 + n2$. Then we can derive $\hat{\gamma}_1$, that is,

$$\hat{\gamma}_{1k} = \left\{\sum_{i=n_1+1}^{n_1+n_2}(1, Q_{ki}, \boldsymbol{V}_i{}^T)^T(1, Q_{ki}, \boldsymbol{V_i}^T)\right\}^{-1}\left\{\sum_{i=1}^{n_1}(1, Q_{ki}, \boldsymbol{V_i}^T)^T \hat{X}_{1ki}\right\},$$

$$\hat{Z}_{1ki} = (1, Q_{ki}, \boldsymbol{V}_i{}^T)\hat{\gamma}_{1k},$$

$$\boldsymbol{\hat{Z}_{1i}} = (\hat{Z}_{11i}, ...\hat{Z}_{1Ki})^T,$$

for i in the full cohort. Finally, we estimate the association parameter based on estimating equation 3.1 with respect to different types of endpoints.

3.2.2 Method 2: Three-step with Bias Correction

(i) **Multivariate approach for Method 2**

As shown in method 1, we have a bias factor in $\hat{\boldsymbol{Z}}_{\mathbf{1}}$ when using $\hat{\boldsymbol{X}}_{\mathbf{1}}$, so we propose a bias-corrected estimator $\hat{\boldsymbol{X}}_{\mathbf{2}} = \hat{\boldsymbol{X}}_{\mathbf{1}}\widehat{\boldsymbol{BF}}^{-1}$ using the multivariate approach where,

$$\widehat{\boldsymbol{BF}} = I_K - \left\{\widehat{Var}(\boldsymbol{X}^*|\boldsymbol{V}) - \Sigma_x^*\right\}^{-1}\left\{\widehat{Var}(\boldsymbol{X}^*|\boldsymbol{W},\boldsymbol{V}) - \Sigma_x^*\right\},$$

is an estimated version of the bias factor. Then we have:

$$\hat{\boldsymbol{X}}_{\mathbf{2}} = \hat{\boldsymbol{X}}_{\mathbf{1}}\widehat{\boldsymbol{BF}}^{\mathbf{-1}},$$

$$\hat{\gamma}_{\mathbf{2}} = \left\{\sum_{i=n_1+1}^{n_1+n_2}(1,\boldsymbol{Q}_i^T,\boldsymbol{V}_i^T)^T(1,\boldsymbol{Q}_i^T,\boldsymbol{V}_i^T)\right\}^{-1}\left\{\sum_{i=n_1+1}^{n_1+n_2}(1,\boldsymbol{Q}_i^T,\boldsymbol{V}_i^T)^T\hat{\boldsymbol{X}}_{\mathbf{2i}}^T\right\}.$$

Finally, we predict the exposure by $\hat{\boldsymbol{Z}}_{\mathbf{2i}} = [(1,\boldsymbol{Q}_{\boldsymbol{i}}^T,\boldsymbol{V}_i^T)\hat{\gamma}_2]^T$ for $i = n_1 + n_2 + 1, \cdots, n_1 + n_2 + n_3$ in the n_3 association sample and we have $\hat{\theta}_2$ by solving:

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3}\left[(1,\hat{\boldsymbol{Z}}_{\mathbf{2i}}^T,\boldsymbol{V}_i^T)^T\left\{Y_i - (1,\hat{\boldsymbol{Z}}_{\mathbf{2i}}^T,\boldsymbol{V}_i^T)\theta\right\}\right],$$

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3}\left[(1,\hat{\boldsymbol{Z}}_{\mathbf{2i}}^T,\boldsymbol{V}_i^T)^T\left\{Y_i - \frac{exp((1,\hat{\boldsymbol{Z}}_{\mathbf{2i}}^T,{\boldsymbol{V}_i}^T)\theta)}{1+exp((1,\hat{\boldsymbol{Z}}_{\mathbf{2i}}^T,{\boldsymbol{V}_i}^T)\theta)}\right\}\right],$$

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3}\int_0^{\tau}\left[\begin{pmatrix}\hat{\boldsymbol{Z}}_{\mathbf{2i}}\\ \boldsymbol{V}_i\end{pmatrix} - \sum_j\frac{Y_j(t)\exp\left\{(\hat{\boldsymbol{Z}}_{\mathbf{2j}}^T,\boldsymbol{V}_j^T)\theta\right\}}{\sum_k Y_k(t)\exp\left\{(\hat{\boldsymbol{Z}}_{\mathbf{2k}}^T,{\boldsymbol{V}_k}^T)\theta\right\}}\begin{pmatrix}\hat{\boldsymbol{Z}}_{\mathbf{2j}}\\ \boldsymbol{V}_j^T\end{pmatrix}\right]dN_i(t), \tag{3.2}$$

for the continuous, binary and time-to-event endpoint, respectively. Again, Method 2 does not require the self-reported dietary intake data ($\boldsymbol{Q}$) in the feeding study, where we have multiple exposures. We will next propose two methods that require the availability of self-reported data in the feeding study and assume the association between the self-reported and the actual dietary intake to be the same among all studies.

(ii) **Univariate approach for Method 2**

Similar to in the univariate approach for Method 1, suppose $\boldsymbol{W}$ is p-dimensional, $\boldsymbol{V}$ is q-dimensional, $\boldsymbol{X}^* = (X_1^* ... X_K^*)$ and $\boldsymbol{Q} = (Q_1 ... Q_K)$ are multivariate in K-dimensional space, then we have:

$$\widehat{BF_k} = 1 - \frac{\widehat{Var}(\boldsymbol{X_k^*}|\boldsymbol{W}, \boldsymbol{V}) - \sigma_{xk}^{*2}}{\widehat{Var}(\boldsymbol{X_k^*}|\boldsymbol{V}) - \sigma_{xk}^{*2}},$$

where σ_{xk}^{*2} denote the kth element along the diagonal of Σ_x^*. With $\widehat{BF_k}$, the bias-corrected estimator can be calculated, that is:

$$\hat{\boldsymbol{X}}_{\boldsymbol{2k}} = \frac{\hat{\boldsymbol{X}}_{\boldsymbol{1k}}}{\widehat{BF_k}}.$$

Then we have:

$$\hat{\gamma}_{2k} = \left\{ \sum_{i=n_1+1}^{n_1+n_2} (1, Q_{ki}, \boldsymbol{V_i}^T)^T (1, Q_{ki}, \boldsymbol{V_i}^T) \right\}^{-1} \left\{ \sum_{i=n_1+1}^{n_1+n_2} (1, Q_{ki}, \boldsymbol{V_i}^T)^T \hat{X}_{2ki} \right\}.$$

Finally, we predict the exposure by:

$$\hat{Z}_{2ki} = (1, Q_{ki}, \boldsymbol{V_i}^T)\hat{\gamma}_{2k},$$

for $i = n_1 + n_2 + 1, \cdots, n_1 + n_2 + n_3$ in the full cohort. Then we have:

$$\hat{\boldsymbol{Z}}_{\boldsymbol{2i}} = (\hat{Z}_{21i}, ...\hat{Z}_{2Ki})^T,$$

for i in the full cohort. Finally, we can obtain $\hat{\theta}_2$ by solving estimating equation 3.2 with respect to different types of endpoints.

3.2.3 Method 3: Three-step with self-reported data

(i) **Multivariate approach for Method 3**

When the self-reported data $\boldsymbol{Q}$ is available from the feeding study and we believe that the distribution of $(\boldsymbol{Q}|\boldsymbol{Z}, \boldsymbol{V})$ are the same between controlled feeding study and the cohort, then the bias in the naïve estimator can be corrected simply by including $\boldsymbol{Q}$ in the biomarker development equation because the inclusion of $\boldsymbol{Q}$ guarantee that $E[\hat{\boldsymbol{Z}}|\boldsymbol{Q}, \boldsymbol{V}] = E[E[\boldsymbol{Z}|\boldsymbol{Q}, \boldsymbol{V}, \boldsymbol{W}]|\boldsymbol{Q}, \boldsymbol{V}] = E[\boldsymbol{Z}|\boldsymbol{Q}, \boldsymbol{V}]$. The three steps of the first method remain the same, but in the first step regression model, the log-transformed self-reported food frequency questionnaire data $(\boldsymbol{Q})$ is added. That is, for the first step, we regress X^* on $\boldsymbol{W}$, $\boldsymbol{V}$ and Q to build the biomarker, and then use $\boldsymbol{W}$, $\boldsymbol{V}$ and $\boldsymbol{Q}$ to predict Z in the second step. Mathematically, we have:

$$\hat{\beta}_3 = \left\{ \sum_{i=1}^{n_1} (1, \boldsymbol{W}_i^T, \boldsymbol{Q}_i^T, \boldsymbol{V}_i^T)^T (1, \boldsymbol{W}_i^T \boldsymbol{Q}_i^T, \boldsymbol{V}_i^T) \right\}^{-1} \left\{ \sum_{i=1}^{n_1} (1, \boldsymbol{W}_i^T, \boldsymbol{Q}_i^T, \boldsymbol{V}_i^T)^T \boldsymbol{X}_{\boldsymbol{i}}^{*\boldsymbol{T}} \right\},$$

$\hat{\boldsymbol{X}}_{3i} = [(1, \boldsymbol{W}_i^T, \boldsymbol{Q}_i^T, \boldsymbol{V}_i^T)\hat{\beta}_3]^T$ for $i = n_1 + 1, \cdots, n_1 + n_2$, and then,

$$\hat{\gamma}_3 = \left\{ \sum_{i=n_1+1}^{n_1+n_2} (1, \boldsymbol{Q}_i^T, \boldsymbol{V}_i^T)^T (1, \boldsymbol{Q}_i^T, \boldsymbol{V}_i^T) \right\}^{-1} \left\{ \sum_{i=n_1+1}^{n_1+n_2} (1, \boldsymbol{Q}_i^T, \boldsymbol{V}_i^T)^T \hat{\boldsymbol{X}}_{\boldsymbol{3i}}^T \right\},$$

and $\hat{\boldsymbol{Z}}_{\boldsymbol{3i}} = [(1, \boldsymbol{Q}_i^T, \boldsymbol{V}_i^T)\hat{\gamma}_3]^T$ for $i = n_1 + n_2 + 1, \cdots, n_1 + n_2 + n_3$. We obtain $\hat{\theta}_3$ by solving the following estimating equations:

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \left[(1, \hat{\boldsymbol{Z}}_{\boldsymbol{3i}}^T, \boldsymbol{V}_i^T)^T \left\{ Y_i - (1, \hat{\boldsymbol{Z}}_{\boldsymbol{3i}}^T, \boldsymbol{V}_i^T)\theta \right\} \right],$$

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \left[(1, \hat{\boldsymbol{Z}}_{\boldsymbol{3i}}^T, \boldsymbol{V}_i^T)^T \left\{ Y_i - \frac{exp((1, \hat{\boldsymbol{Z}}_{\boldsymbol{3i}}^T, \boldsymbol{V}_i^T)\theta)}{1 + exp((1, \hat{\boldsymbol{Z}}_{\boldsymbol{3i}}^T, \boldsymbol{V}_i^T)\theta)} \right\} \right],$$

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \int_0^{\tau} \left[\begin{pmatrix} \hat{\boldsymbol{Z}}_{\boldsymbol{3i}} \\ \boldsymbol{V}_i \end{pmatrix} - \sum_j \frac{Y_j(t) \exp\left\{ (\hat{\boldsymbol{Z}}_{\boldsymbol{3j}}^T, \boldsymbol{V}_j^T)\theta \right\}}{\sum_k Y_k(t) \exp\left\{ (\hat{\boldsymbol{Z}}_{\boldsymbol{3k}}^T, \boldsymbol{V}_k^T)\theta \right\}} \begin{pmatrix} \hat{\boldsymbol{Z}}_{\boldsymbol{3j}} \\ \boldsymbol{V}_j \end{pmatrix} \right] dN_i(t), \tag{3.3}$$

for the continuous, binary and time-to-event endpoint, respectively.

(ii) **Univariate approach for Method 3**

The univariate approach for Method 3 is similar as the univariate approach for Method 1 except for the first step, where we add $\boldsymbol{Q} = (\boldsymbol{Q}_1, ..., \boldsymbol{Q}_K)$ as an estimator in the first linear model to build the biomarker, that is,

$$\hat{\beta}_{3k} = \left\{ \sum_{i=1}^{n_1} (1, \boldsymbol{W}_i^T, Q_{ki}, \boldsymbol{V}_i^T)^T (1, \boldsymbol{W}_i^T, Q_{ki}, \boldsymbol{V}_i^T) \right\}^{-1} \left\{ \sum_{i=1}^{n_1} (1, \boldsymbol{W}_i^T, Q_{ki}, \boldsymbol{V}_i^T)^T X_{ki}^* \right\}.$$

Then we have:

$$\hat{X}_{3ki} = (1, \boldsymbol{W}_i^T, Q_{ki}, \boldsymbol{V}_i{}^T)\hat{\beta}_{3k},$$

$$\hat{\gamma}_{3k} = \left\{ \sum_{i=n_1+1}^{n_1+n_2} (1, Q_{ki}, \boldsymbol{V}_i{}^T)^T (1, Q_{ki}, \boldsymbol{V}_i{}^T) \right\}^{-1} \left\{ \sum_{i=n_1+1}^{n_1+n_2} (1, Q_{ki}, \boldsymbol{V}_i{}^T)^T \hat{X}_{3ki} \right\}.$$

Finally, we predict the exposure by:

$$\hat{Z}_{3ki} = (1, Q_{ki}, \boldsymbol{V}_i{}^T)\hat{\gamma}_{3k},$$

for $i = n_1 + n_2 + 1, \cdots, n_1 + n_2 + n_3$ in the full cohort.

$$\hat{Z}_{3ki} = (1, Q_{ki}, \boldsymbol{V}_i{}^T)\hat{\gamma}_{3k},$$

$$\boldsymbol{\hat{Z}_{3i}} = (\hat{Z}_{31i}, ...\hat{Z}_{3Ki})^T,$$

for i in the full cohort. Finally, we can obtain $\hat{\theta}_3$ by solving estimating equation 3.3 with respect to different types of endpoints.

3.2.4 Method 4: Direct Estimation

(i) **Multivariate approach for Method 4**

When $\boldsymbol{Q}$ is available from the feeding study, another possibility is to ignore the second dataset and directly build the estimating equation by regressing $\boldsymbol{X}^*$ on $\boldsymbol{Q}$ and $\boldsymbol{V}$ in the first step and directly apply it to the third step. All other steps remain the same as the third method, except for the ignorance of the n_2 calibration samples. Instead, we directly build the calibration equation using the feeding study by regressing $\boldsymbol{X}^*$ on $\boldsymbol{V}$ and $\boldsymbol{Q}$ and use the calibration equation to predict $\boldsymbol{Z}$ and perform a linear, logistic and Cox regression of Y on $\boldsymbol{Z}$ and $\boldsymbol{V}$ in the full cohort to estimate the association parameter based on different types of outcomes. In other words, we have:

$$\hat{\gamma}_4 = \left\{ \sum_{i=1}^{n_1} (1, \boldsymbol{Q}_i^T, \boldsymbol{V}_i^T)^T (1, \boldsymbol{Q}_i^T, \boldsymbol{V}_i^T) \right\}^{-1} \left\{ \sum_{i=1}^{n_1} (1, \boldsymbol{Q}_i^T, \boldsymbol{V}_i^T)^T \boldsymbol{X}^{*T} \right\},$$

$\hat{\boldsymbol{Z}}_{4i} = [(1, \boldsymbol{Q}_i^T, \boldsymbol{V}_i^T)\hat{\gamma}_4]^T$ for $i = n_1 + n_2 + 1, \cdots, n_1 + n_2 + n_3$ in the full cohort and $\hat{\theta}_4$ by solving:

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \left[(1, \hat{\boldsymbol{Z}}_{4i}^T, \boldsymbol{V}_i^T)^T \left\{Y_i - (1, \hat{\boldsymbol{Z}}_{4i}^T, \boldsymbol{V}_i^T)\theta\right\}\right],$$

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \left[(1, \hat{\boldsymbol{Z}}_{4i}^T, \boldsymbol{V}_i^T)^T \left\{Y_i - \frac{\exp((1, \hat{\boldsymbol{Z}}_{4i}^T, \boldsymbol{V}_i^T)\theta)}{1 + \exp((1, \hat{\boldsymbol{Z}}_{4i}^T, \boldsymbol{V}_i^T)\theta)}\right\}\right],$$

$$0 = \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \int_0^\tau \left[\begin{pmatrix} \hat{\boldsymbol{Z}}_{4i} \\ \boldsymbol{V}_i \end{pmatrix} - \sum_j \frac{Y_j(t) \exp\left\{(\hat{\boldsymbol{Z}}_{4j}^T, \boldsymbol{V}_j^T)\theta\right\}}{\sum_k Y_k(t) \exp\left\{(\hat{\boldsymbol{Z}}_{4k}^T, \boldsymbol{V}_k^T)\theta\right\}} \begin{pmatrix} \hat{\boldsymbol{Z}}_{4j} \\ \boldsymbol{V}_j \end{pmatrix}\right] dN_i(t), \tag{3.4}$$

with respect to the continuous, binary and time-to-event endpoints.

(ii) **Univariate approach for Method 4**

In the univariate approach, we perform univariate linear regression regarding each element in exposures to build biomarker. Below are the steps to obtain $\hat{\boldsymbol{Z}}_{\boldsymbol{4}}$ in the univariate approach for Method 4.

$$\hat{\gamma}_{\boldsymbol{4k}} = \left\{\sum_{i=1}^{n_1}(1, Q_{ki}, \boldsymbol{V}_i^T)^T(1, Q_{ki}, \boldsymbol{V}_i^T)\right\}^{-1} \left\{\sum_{i=1}^{n_1}(1, Q_{ki}, \boldsymbol{V}_i^T)^T X_{ki}^*\right\},$$

Then the long-term dietary intake can be estimated by:

$$\hat{Z}_{4ki} = (1, Q_{ki}, \boldsymbol{V_i}^T)\hat{\gamma}_{4k},$$

for $i = n1 + n2 + 1, \cdots, n1 + n2 + n3$. Then,

$$\hat{\mathbf{Z}}_{\mathbf{4i}} = (\hat{Z}_{41i}, ...\hat{Z}_{4Ki})^T,$$

for i in the full cohort. Finally, we can obtain $\hat{\theta}_4$ by solving estimating equation 3.4 with respect to different types of endpoints.

3.3 Theory for Multivariate Approaches

Theoretical proof for multivariate approaches can be found in Appendix B.

Theorem B1: With $\frac{n_3}{n_2} \to C_2$ and $\frac{n_2}{n_1} \to C_1$, we have $\sqrt{n_3}(\hat{\theta}_1 - \theta_1^*) \to N(0, \Sigma_{\theta 1})$ where $\Sigma_{\theta 1} = I_{\theta 1}^{-1}(I_{\theta 1} + C_2 I_{\gamma 1} \Sigma_{\gamma 1} I_{\gamma 1}^T) I_{\theta 1}^{-T}$ can be consistently estimated by $\hat{\Sigma}_{\theta 1} = \hat{I}_{\theta 1}^{-1}(\hat{I}_{\theta 1} + \frac{n_3}{n_2} \hat{I}_{\gamma 1} \hat{\Sigma}_{\gamma 1} \hat{I}_{\gamma 1}^T) \hat{I}_{\theta 1}^{-T}$ where the detail expression of I_θ and I_γ are different for different types of regression models. The detailed expressions are defined in the proofs in Appendix B (Theorem 8 and 9).

Theorem B2: With $\frac{n_3}{n_2} \to C_2$ and $\frac{n_2}{n_1} \to C_1$, we have $\sqrt{n_3}(\hat{\theta}_2 - \theta_2^*) \to N(0, \Sigma_{\theta 2})$ where $\Sigma_{\theta 2} = I_{\theta 2}^{-1}(I_{\theta 2} + C_2 I_{\gamma 2} \Sigma_{\gamma 2} I_{\gamma 2}^T) I_{\theta 2}^{-T}$ can be consistently estimated by $\hat{\Sigma}_{\theta 2} = \hat{I}_{\theta 2}^{-1}(\hat{I}_{\theta 2} + \frac{n_3}{n_2} \hat{I}_{\gamma 2} \hat{\Sigma}_{\gamma 2} \hat{I}_{\gamma 2}^T) \hat{I}_{\theta 2}^{-T}$ where the detail expression of I_θ and I_γ are different for different types of regression models. The detailed expressions are defined in the proofs in Appendix B (Theorem 8 and 9).

Theorem B3: With $\frac{n_3}{n_2} \to C_2$ and $\frac{n_2}{n_1} \to C_1$, we have $\sqrt{n_3}(\hat{\theta}_3 - \theta_3^*) \to N(0, \Sigma_{\theta 3})$ where $\Sigma_{\theta 3} = I_{\theta 2}^{-1}(I_{\theta 2} + C_2 I_{\gamma 3} \Sigma_{\gamma 3} I_{\gamma 3}^T) I_{\theta 2}^{-T}$ can be consistently estimated by $\hat{\Sigma}_{\theta 3} = \hat{I}_{\theta 3}^{-1}(\hat{I}_{\theta 3} + \frac{n_3}{n_2} \hat{I}_{\gamma 3} \hat{\Sigma}_{\gamma 3} \hat{I}_{\gamma 3}^T) \hat{I}_{\theta 3}^{-T}$ where the detail expression of I_θ and I_γ are different for different types of regression models. The detailed expressions are defined in the proofs in Appendix B

(Theorem 8 and 9).

Theorem B4: With $\frac{n_3}{n_2} \to C_2$ and $\frac{n_2}{n_1} \to C_1$, we have $\sqrt{n_3}(\hat{\theta}_4 - \theta_4^*) \to N(0, \Sigma_{\theta 4})$ where $\Sigma_{\theta 4} = I_{\theta 2}^{-1}(I_{\theta 2} + C_1 C_2 I_{\gamma 4} \Sigma_{\gamma 4} I_{\gamma 4}^T) I_{\theta 2}^{-T}$ can be consistently estimated by $\hat{\Sigma}_{\theta 4} = \hat{I}_{\theta 4}^{-1}(\hat{I}_{\theta 4} + \frac{n_3}{n_1} \hat{I}_{\gamma 4} \hat{\Sigma}_{\gamma 4} \hat{I}_{\gamma 4}^T) \hat{I}_{\theta 4}^{-T}$ where the detail expression of I_θ and I_γ are different for different types of regression models. The detailed expressions are defined in the proofs in Appendix B (Theorem 8 and 9).

3.4 Simulation

We performed simulations to study the finite sample behavior of our proposed two log-transformed parameter estimators. We generate data from Cox, logistic and linear model with time-to-event, binary and continuous endpoints, respectively as listed below:

$$
\begin{aligned}
(\boldsymbol{Z}, V) &\sim N\left(0, \begin{pmatrix} 1-\sigma_x^2 & \rho_z & \rho_{z_1 v} \\ \rho_z & 1-\sigma_x^2 & \rho_{z_2 v} \\ \rho_{z_1 v} & \rho_{z_2 v} & 1 \end{pmatrix}\right), \\
\boldsymbol{W} &= (1, \boldsymbol{Z}^T, V)\boldsymbol{B} + \epsilon_w, \\
\boldsymbol{X} &= \boldsymbol{Z} + \epsilon_x, \\
\boldsymbol{X}^* &= \boldsymbol{X} + \epsilon_x^*, \\
\boldsymbol{Q} &= (1, \boldsymbol{Z}^T, V)\boldsymbol{A} + \epsilon_q.
\end{aligned}
$$

With continuous endpoint, we have:

$$Y = (\mathbf{1}, \boldsymbol{Z}^T, V)\theta + \epsilon_y.$$

With binary endpoint, we have:

$$logit(P(Y = 1|\boldsymbol{Z}, V)) = (1, \boldsymbol{Z}^T, V)\theta.$$

With time-to-event endpoint, we have:

$$\lambda(t|\boldsymbol{Z}, \boldsymbol{X}, V, \boldsymbol{W}, \boldsymbol{Q}) = \lambda(t|\mathbf{1}, \boldsymbol{Z}, V) = \lambda_0(t)\exp((\boldsymbol{Z}^T, V)\theta),$$

where $\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{Q}$ are bivariate and V is a single covariate in our simulation. Hence,

$$\boldsymbol{B} = \begin{pmatrix} b_0 & b_1 & b_2 & b_3 \\ b'_0 & b'_1 & b'_2 & b'_3 \end{pmatrix},$$

$$\boldsymbol{A} = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 \\ a'_0 & a'_1 & a'_2 & a'_3 \end{pmatrix},$$

and we have $\theta = (\theta_0, \theta_1, \theta_2, \theta_3)^T$, $\theta = (\theta'_0, \theta_1, \theta_2, \theta_3)^T$ and $\theta = (\theta_1, \theta_2, \theta_3)^T$ for the continuous, binary and time-to-event endpoint, respectively. The error terms, ϵ_x, ϵ_x^* and ϵ_w are independently sampled from multivariate normal distributions with mean zero and covariate matrix of $\sigma_x^2 I$, $\sigma_x^{*2} I$ and $\sigma_w^2 I$. For Cox model with time-to-event endpoint, censoring time is sampled from a mixture of $Unif(0, 10)$ and a point mass at 10 with equal probability. For linear model with continuous endpoint, a random noise is drawn from a N(0,1.8). We fixed the sample size at $n_1 = 300$, $n_2 = 600$, $n_3 = 5150$ for all settings. We set $b_0 = b'_0 = 5$, $b_3 = b'_3 = 1$, $a_3 = a'_3 = 1$, $\sigma_x = 0.2$, $\sigma_x^* = 0.5$, $\sigma_w = 1$,

$\rho_{z_1 v} = 0.3$, $\rho_{z_2 v} = 0.4$, $\theta_0 = 1$, $\theta_0' = -4$, $\theta_1 = 0.4$, $\theta_2 = 0.6$ and $\theta_3 = 0.4$. Two types of data structure are assumed in our simulation, that is, multivariate and univariate data. Specifically, we set $b_2 = 0$, $b_1' = 0$, $a_2 = 0$, $a_1' = 0$ under univariate cases, and $b_2 = 0.7$, $b_1' = 0.9$, $a_2 = 0.6$, $a_1' = 0.4$ under multivariate cases. With time-to-event endpoint, we set $\lambda_0(t) = 0.002t$. We performed analysis under univariate and multivariate assumptions. For univariate cases, we simulate ϵ_q from multivariate normal distribution with mean 0 and covariance matrix $\sigma_q^2 I$. For multivariate cases, we simulate ϵ_q from multivariate normal distribution as below:

$$\epsilon_q \sim N\left(0, \begin{pmatrix} \sigma_q^2 & 0.2 \\ 0.2 & \sigma_q^2 \end{pmatrix}\right).$$

Then we change the values of b_1, b_2', a_0, a_0', a_1, a_2', ρ_z and σ_q to change the range of R^2 listed in Table 10 under both multivariate and univariate assumptions. Eight representative settings are selected for multivariate data type and another eight settings are selected for univariate data type to make the multiple coefficient of determination comparable between them. Below are eight settings selected under multivariate assumption:

$b_1 = 1.8$, $b_2' = 2.2$, $a_0 = a_0' = 4$, $a_1 = 1.4$, $a_2' = 1.6$, $\rho_z = 0.12$, $\sigma_q = 4$ (setting 1);
$b_1 = b_2' = 1.2$, $a_0 = a_0' = 4$, $a_1 = 1.4$, $a_2' = 1.6$, $\rho_z = 0.12$, $\sigma_q = 4$ (setting 2);
$b_1 = 1.8$, $b_2' = 2.2$, $a_0 = a_0' = 4$, $a_1 = a_2' = 1.1$, $\rho_z = 0.12$, $\sigma_q = 5$ (setting 3);
$b_1 = b_2' = 1.2$, $a_0 = a_0' = 4$, $a_1 = a_2' = 1.1$, $\rho_z = 0.12$, $\sigma_q = 5$ (setting 4);
$b_1 = 1.8$, $b_2' = 2.2$, $a_0 = a_0' = 4$, $a_1 = 1.7$, $a_2' = 2$, $\rho_z = -0.1$, $\sigma_q = 4$ (setting 5);
$b_1 = 1.3$, $b_2' = 1$, $a_0 = a_0' = 4$, $a_1 = 1.7$, $a_2' = 2$, $\rho_z = -0.1$, $\sigma_q = 4$ (setting 6);
$b_1 = 1.8$, $b_2' = 2.2$, $a_0 = a_0' = 4$, $a_1 = a_2' = 1.1$, $\rho_z = -0.1$, $\sigma_q = 5$ (setting 7);
$b_1 = 1.3$, $b_2' = 1$, $a_0 = a_0' = 4$, $a_1 = a_2' = 1.1$, $\rho_z = -0.1$, $\sigma_q = 5$ (setting 8).

Eight settings under univariate assumption are listed below:

$b_1 = 1.5$, $b_2' = 2$, $a_0 = a_0' = 4$, $a_1 = 1.3$, $a_2' = 1.6$, $\rho_z = 0.12$, $\sigma_q = 4$ (setting 1);
$b_1 = b_2' = 0.9$, $a_0 = a_0' = 4$, $a_1 = 1.3$, $a_2' = 1.6$, $\rho_z = 0.12$, $\sigma_q = 4$ (setting 2);

$b_1 = 1.5$, $b_2' = 2$, $a_0 = a_0' = 4$, $a_1 = a_2' = 1.1$, $\rho_z = 0.12$, $\sigma_q = 5$ (setting 3);
$b_1 = b_2' = 0.9$, $a_0 = a_0' = 4$, $a_1 = a_2' = 1.1$, $\rho_z = 0.12$, $\sigma_q = 5$ (setting 4);
$b_1 = 1.5$, $b_2' = 2$, $a_0 = a_0' = 4$, $a_1 = 1.6$, $a_2' = 1.8$, $\rho_z = -0.1$, $\sigma_q = 4$ (setting 5);
$b_1 = b_2' = 0.7$, $a_0 = a_0' = 4$, $a_1 = 1.6$, $a_2' = 1.8$, $\rho_z = -0.1$, $\sigma_q = 4$ (setting 6);
$b_1 = 1.5$, $b_2' = 2$, $a_0 = a_0' = 4$, $a_1 = a_2' = 1$, $\rho_z = -0.1$, $\sigma_q = 5$ (setting 7);
$b_1 = b_2' = 0.7$, $a_0 = a_0' = 4$, $a_1 = a_2' = 1$, $\rho_z = -0.1$, $\sigma_q = 5$ (setting 8).

For both multivariate and univariate data types, the correlation between bivariate $\boldsymbol{Z}$, ρ_z, is 0.12, which is independent conditioning on $\boldsymbol{V}$ in the first four settings and ρ_z is -0.1 in the last four settings. For every two sequential settings, we simulated data with biomarkers in the first setting to be stronger than the second one. The strength of self-reported data is fixed in every two sequential settings and showed different levels of strength among every two settings. More detailed information on the explained variation of true, consumed and estimated dietary intakes by biomarkers and FFQ information can be found in Table 10. Based on Table 10, we can see the lowest and highest coefficient determination of true dietary intakes ($\boldsymbol{Z}$) on FFQ ($\boldsymbol{Q}$) given personal characteristics ($\boldsymbol{V}$) is (0.13, 0.14) and (0.35, 0.41) with the multivariate data type, respectively. The consumed dietary intake ($\boldsymbol{X}^*$) can be explained by biomarkers ($\boldsymbol{W}$) given personal characteristics ($\boldsymbol{V}$) as large as (0.53,0.58) and as low as (0.31, 0.21). Calculated multiple coefficient of determination for univariate data types is showed in the lower part in Table 10 and showed similar trends through different settings.

The bias, SE, SD and CR of 95% nominal confidence interval from 100 simulations are calculated for the multivariate approach while bias and SD are calculated for the univariate approach. For data simulated under the multivariate assumption, results are summarized in Tables 11, 12 and 13 for each type of endpoint. First, we focus on results with the multivariate approach, which are displayed in the left panel in all tables. Though the CR with the multivariate approach for Method 1 is above 0.9 in several cases, we can see

Method 1 showed significant bias in most cases. The bias was greatly attenuated and controlled under the absolute value of 0.02 with Method 2 in most cases, especially when biomarkers are relatively strong. In addition, coverage rates are consistently above 0.9 with Method 2. However, With relatively weak biomarkers and FFQ information, Method 2 did not show good estimation on empirical SE and generated relatively large bias (i.e., setting 4 and 8 in Table 11). Method 3 and 4 provide consistent good estimations on association parameters with good CR. However, the efficiency of Method 2 performed better when we have weak FFQ data and strong biomarkers. For example, Method 2 showed smaller SD compared with Method 3 and 4 in settings 3 and 7 based on Tables 11, 12 and 13. With the univariate approach (right panel), we can see bias is higher compared with those in the multivariate approach in all settings with all different methods. The SD shown with the univariate approach is smaller compared with the multivariate approach in most cases. With a small increase in bias and lower SD, the mean squared error (MSE) is calculated to be smaller with the univariate approach compared with the multivariate approach in several cases. When there is a relatively large amount of increase in bias, the MSE with the univariate approach appeared to be larger than the MSE with the multivariate approach.

For data simulated under the univariate assumption, results are summarized in Tables 14, 15 and 16. Method 1 again generated a large bias in most cases under all three types of endpoints. Focusing on the multivariate approach, Method 2, 3 and 4 showed adequate estimations on association parameters and comparable to each other in all settings and all different endpoints. With relatively strong biomarkers and weak FFQ, we can see Method 2 still tended to provide more efficient results with smaller SD compared with Method 3 and 4. When a relatively small amount of variation of true dietary intake can be explained by biomarkers and FFQ information (setting 8 of Table 14), SD and bias of Method 2 slightly increased with the binary endpoint. When the univariate data structure is simulated, Method 2, 3 and 4 with the univariate approach controlled the bias under 0.02 in the first

4 settings, when bivariate true dietary intakes are independent conditioning on personal characteristics. At the same time, apparent bias is showed in the last four settings when bivariate true dietary intakes are not independent conditioning on individual characteristics.

3.5 Data Analysis

Similar to Chapter 2, we illustrate our methods with the WHI NPAAS feeding study($n = 150$), NPAAS biomarker study ($n = 450$) and the full WHI cohort data in this section. Variables including age, BMI, race/ethnicity, education level, self-reported physical activity and smoking status are included and used as $\boldsymbol{V}$; the disease outcomes are different types of CVD, including total CHD and its myocardial infarction and coronary death components, total stroke and its hemorrhagic and ischemic components, total CVD comprised of CHD and stroke, CABG and PCI, and total CVD that also includes CABG and PCI, and heart failure. Data are analyzed based on two forms of exposures, including the ratio of sodium and calories, ratio of potassium and calories on logarithm base, as well as sodium and potassium in milligram (mg) per day on logarithm base. For ease of interpretation, the ratio of sodium and potassium with calories are expressed as percentages for the rest of this section. Biomarkers ($\boldsymbol{W}$) and FFQ information ($\boldsymbol{Q}$) are also corresponding to the units of exposures measured in percentage and mg per day. To be specific, biomarkers ($\boldsymbol{W}$) adopted in this section include the 24-hour urine sodium and potassium measured in percentage on logarithm base and sodium and potassium measured in total mg per day. Furthermore, log-transformed self-reported sodium and potassium intake in percentage and total log-transformed self-reported sodium and potassium in mg per day are set as $\boldsymbol{Q}$ with corresponding biomarkers and exposures to generate different tables, respectively. Multivariate and univariate analysis with regards to two exposures are both performed.

With multiple exposures, the adjusted bias factors are estimated by

$$\widehat{\boldsymbol{BF}} = I_K - (\widehat{Var}(\boldsymbol{X}^*|\boldsymbol{V}) - \hat{\Sigma}_x^*)^{-1}(\widehat{Var}(\boldsymbol{X}^*|\boldsymbol{W},\boldsymbol{V}) - \hat{\Sigma}_x^*)$$

where $\hat{\Sigma}_x^*$ can be treated as a sensitivity parameter under multivariate analysis. The most conservative estimate on $\hat{\Sigma}_x^*$, a zero matrix, is utilized to illustrate the potential bias. The adjusted bias factor under univariate analysis can be estimated by $\widehat{BF_k} = 1 - \frac{\widehat{Var}(X_k^*|\boldsymbol{W},\boldsymbol{V}) - \hat{\sigma}_{xk}^{*2}}{\widehat{Var}(X_k^*|\boldsymbol{V}) - \hat{\sigma}_{xk}^{*2}}$ where k denotes each element in the diagonal matrix of multiple exposures.

The estimated hazard ratio (HR) according to a 20% increase in the sodium and potassium is shown in Table 17 and 18 for measurements in percentage and mg per day under multivariate analysis, respectively. Based on the result, we found that the naïve three-step approach (Method 1) provided the highest HR for sodium and the lowest HR for potassium among all 4 methods. The HR estimated from Method 3 (three-step with FFQ approach) is similar to the estimate using Method 2 in both Table 17 and 18. To be more specific, the point estimated HR with Method 2 is slightly lower for sodium and slightly higher for potassium and Method 2 generated a narrower confidence interval compared with Method 3 in most cases. Overall, we found the point estimate of HR with Method 2 and 3 when total dietary intake is measured in mg per day shown in Table 18 is similar to the results shown in Table 17 when the unit of measurements is in percentage per day. Comparing to the results in Prentice et al. (2017) [14], sodium is positively associated with the risk of CVD, while potassium is negatively associated with cardiovascular diseases, which is consistent with the results in Prentice et al. (2017) [14]. In addition, Method 3 gives a slightly narrower confidence interval compared with Prentice et al. (2017) [14].

Tables 19 and 20 displayed the results of the estimated hazard ratio according to a 20% increase in sodium and potassium for measurements in percentage and mg per day under univariate analysis, respectively. The confidence interval shown in Table 19 is slightly narrower with univariate analysis compared with multivariate analysis in most cases in Table

17. The results are comparable with those shown in Prentice et al. (2017) [14]. Comparing with the previous three tables, higher estimated HR on sodium and lower estimated HR on potassium with broader confidence intervals are generated by Method 2, 3 and 4 under univariate analysis and showed in Table 20 for measurements in mg per day. This may be due to the strong correlation among different multivariate estimated assessed dietary intakes given personal characteristics found in such cases. In general, the multivariate approach seemed to provide more efficient results with narrower CI in estimating the association of CVD with sodium and potassium compared to the univariate approach.

3.6 Discussion

In this chapter, we examined the requirement for a valid biomarker for regression calibration purposes with multiple exposures. Multivariate and univariate analyses are both performed to test the performance in controlling bias. The multivariate approach is more complex to implement compared with the univariate approach. The bias is well controlled with the univariate approach only when the bivariate long-term dietary intakes are independent conditioning on personal characteristics in our simulation settings. When multiple exposures are correlated conditioning on personal characteristics, the univariate approach can lead to large bias even when the data structure is univariate. The multivariate approach provides results with small bias and good CR in most settings with Method 2, 3 and 4. However, when bias increased by a small amount while SD is substantially low, MSE with the univariate approach can be smaller compared with MSE for the corresponding multivariate approach. Hence a bias and variance trade-off problem need to be considered. In general, the multivariate approach generated consistent and more robust estimations on association parameters compared with the univariate approach. The multivariate approach is recommended, especially when the correlation between multiple dietary intakes conditioning on personal characteristics exists. Under such cases, large bias and worse MSE are usually shown with the univariate approach. The univariate analysis can be considered for the occasions when long-term dietary intakes are independent given personal characteristics, where comparable or even smaller MSE are shown with the univariate approach.

With the multivariate approach, we noticed estimation on association parameters with Method 2 could be affected by weak biomarker and weak FFQ information. Few extreme estimated values on asymptotic SE could be generated among 100 simulations under such settings and lead inaccurate mean estimated SE in the end. Such an issue could be solved by increasing sample size in the biomarker construction and calibration building steps. The performance of Method 1 in controlling bias is not good in most settings and should not be used. Method 3 and 4 provided consistent estimations and have shown more efficient results when FFQ information is strongly associated with long-term dietary intakes. However, in reality, the association between FFQ data and true dietary intakes may be much smaller than the value shown in such settings. Under such cases, Method 2 can give better efficiency.

In order to derive empirical SE for Method 2 under multivariate analysis, Delta method was used to approximate the Σ_{β_2}. More detailed information can be found in appendix B, however, due to the imprecision by our Delta method, an alternative way through the bootstrap approach was found to provide a more accurate estimator and was utilized in our programming to obtain $\hat{\Sigma}_{\beta_2}$. More precise methods can be further considered in our future analysis.

TABLE 10: List of R^2 among different measurements for multivariate and univariate data types

Data Type	Type of R2	Setting 1		Setting 2		Setting 3		Setting 4		Setting 5		Setting 6		Setting 7		Setting 8	
Multivariate	R^2_{ZWV}	0.68	0.76	0.48	0.53	0.68	0.76	0.48	0.53	0.66	0.75	0.43	0.38	0.66	0.75	0.43	0.38
	$R^2_{ZW_V}$	0.64	0.71	0.42	0.44	0.64	0.71	0.42	0.44	0.63	0.69	0.37	0.26	0.63	0.69	0.37	0.26
	R^2_{ZQV}	0.36	0.45	0.36	0.45	0.25	0.31	0.25	0.31	0.41	0.51	0.41	0.51	0.21	0.29	0.21	0.29
	$R^2_{ZQ_V}$	0.29	0.34	0.29	0.34	0.17	0.18	0.17	0.18	0.35	0.41	0.35	0.41	0.13	0.14	0.13	0.14
	R^2_{ZQWV}	0.71	0.78	0.56	0.61	0.69	0.77	0.51	0.56	0.72	0.79	0.58	0.59	0.68	0.75	0.47	0.43
	$R^2_{ZQW_V}$	0.68	0.74	0.51	0.54	0.66	0.72	0.46	0.47	0.69	0.75	0.54	0.50	0.64	0.70	0.41	0.32
	$R^2_{X^*WV}$	0.56	0.63	0.39	0.44	0.56	0.63	0.39	0.44	0.55	0.62	0.36	0.32	0.55	0.62	0.36	0.32
	$R^2_{X^*W_V}$	0.53	0.58	0.35	0.36	0.53	0.58	0.35	0.36	0.51	0.56	0.31	0.21	0.51	0.56	0.31	0.21
	$R^2_{\hat{X}_2QV}$	0.23	0.35	0.06	0.21	0.17	0.26	0.03	0.19	0.28	0.39	0.08	0.29	0.16	0.23	0.04	0.27
	$R^2_{\hat{X}_2Q_V}$	0.17	0.23	0.06	0.07	0.10	0.11	0.03	0.03	0.20	0.28	0.06	0.05	0.08	0.09	0.02	0.02
Univariate	R^2_{ZWV}	0.67	0.77	0.46	0.48	0.67	0.77	0.46	0.48	0.68	0.78	0.38	0.41	0.68	0.78	0.38	0.41
	$R^2_{ZW_V}$	0.63	0.73	0.41	0.38	0.63	0.73	0.41	0.38	0.65	0.73	0.31	0.28	0.65	0.73	0.31	0.28
	R^2_{ZQV}	0.34	0.45	0.34	0.45	0.26	0.30	0.26	0.30	0.43	0.51	0.43	0.51	0.24	0.30	0.24	0.30
	$R^2_{ZQ_V}$	0.27	0.34	0.27	0.34	0.18	0.16	0.18	0.16	0.37	0.41	0.37	0.41	0.16	0.15	0.16	0.15
	R^2_{ZQWV}	0.71	0.80	0.56	0.61	0.69	0.78	0.53	0.54	0.73	0.81	0.56	0.60	0.70	0.79	0.45	0.47
	$R^2_{ZQW_V}$	0.68	0.76	0.52	0.53	0.66	0.74	0.47	0.45	0.71	0.77	0.51	0.52	0.67	0.74	0.39	0.36
	$R^2_{X^*WV}$	0.55	0.64	0.38	0.40	0.55	0.64	0.38	0.40	0.56	0.64	0.31	0.33	0.56	0.64	0.31	0.33
	$R^2_{X^*W_V}$	0.52	0.59	0.33	0.31	0.52	0.59	0.33	0.31	0.53	0.59	0.25	0.23	0.53	0.59	0.25	0.23
	$R^2_{\hat{X}_2QV}$	0.28	0.40	0.28	0.41	0.23	0.29	0.25	0.36	0.36	0.45	0.41	0.53	0.25	0.31	0.36	0.49
	$R^2_{\hat{X}_2Q_V}$	0.17	0.25	0.11	0.13	0.11	0.12	0.07	0.06	0.23	0.29	0.11	0.11	0.09	0.10	0.04	0.04

Table 11: Simulation results under multivariate assumption comparing multivariate and univariate approaches with continuous endpoint

Setting	Method	Multivariate				Univariate	
		Bias	SE	SD	CR	Bias	SD
1	1	0.05	0.143	0.163	0.91	0.36	0.147
		0.11	0.130	0.134	0.87	0.21	0.097
	2	0.01	0.085	0.099	0.9	0.04	0.084
		0.00	0.085	0.082	0.96	-0.03	0.073
	3	0.01	0.091	0.100	0.95	0.10	0.091
		-0.01	0.089	0.088	0.94	0.00	0.078
	4	0.02	0.105	0.117	0.94	0.11	0.096
		0.00	0.101	0.107	0.95	0.01	0.093
2	1	-0.46	0.817	0.727	0.98	0.63	0.197
		0.78	0.804	0.688	0.85	0.69	0.183
	2	0.01	0.130	0.110	0.94	-0.05	0.070
		0.00	0.136	0.099	0.98	-0.10	0.075
	3	0.01	0.095	0.101	0.96	0.11	0.093
		-0.01	0.092	0.092	0.94	0.01	0.086
	4	0.02	0.105	0.117	0.94	0.11	0.096
		0.00	0.101	0.107	0.95	0.01	0.093
3	1	0.06	0.259	0.283	0.93	0.49	0.196
		0.10	0.248	0.248	0.98	0.25	0.139
	2	0.02	0.144	0.161	0.94	0.11	0.112
		0.00	0.151	0.144	0.97	-0.01	0.101
	3	0.02	0.159	0.164	0.96	0.20	0.126
		-0.02	0.165	0.153	0.96	0.05	0.117
	4	0.03	0.205	0.205	0.96	0.21	0.133
		-0.01	0.211	0.199	0.96	0.06	0.147
4	1	-0.86	103.783	5.071	1.00	0.77	0.245
		1.24	118.929	5.703	0.98	0.68	0.236
	2	-0.01	5.708	0.362	0.98	0.00	0.087
		0.04	7.705	0.429	0.98	-0.11	0.092
	3	0.02	0.170	0.167	0.96	0.21	0.128
		-0.01	0.178	0.158	0.95	0.05	0.130
	4	0.03	0.205	0.205	0.96	0.21	0.133
		-0.01	0.211	0.199	0.96	0.06	0.147

Setting	Method	Multivariate				Univariate	
		Bias	SE	SD	CR	Bias	SD
5	1	0.06	0.102	0.095	0.94	0.25	0.127
		0.13	0.089	0.095	0.74	0.16	0.097
	2	0.00	0.071	0.066	0.97	-0.05	0.065
		0.00	0.070	0.066	0.95	-0.09	0.063
	3	0.00	0.073	0.068	0.97	-0.03	0.067
		-0.01	0.069	0.071	0.96	-0.07	0.069
	4	0.00	0.080	0.088	0.93	-0.02	0.073
		0.00	0.078	0.084	0.95	-0.06	0.082
6	1	-0.73	0.557	0.433	0.88	0.92	0.301
		1.51	0.714	0.557	0.26	1.08	0.324
	2	0.00	0.225	0.097	0.98	-0.14	0.061
		-0.01	0.241	0.087	0.97	-0.17	0.079
	3	0.00	0.075	0.072	0.97	-0.02	0.070
		0.00	0.072	0.074	0.96	-0.06	0.079
	4	0.00	0.080	0.088	0.93	-0.02	0.073
		0.00	0.078	0.084	0.95	-0.06	0.082
7	1	0.07	0.196	0.191	0.98	0.49	0.242
		0.13	0.177	0.190	0.91	0.22	0.185
	2	0.00	0.124	0.118	0.98	0.08	0.123
		0.00	0.126	0.126	0.95	-0.05	0.121
	3	0.00	0.135	0.121	0.98	0.14	0.131
		-0.01	0.140	0.148	0.93	0.01	0.149
	4	-0.01	0.166	0.174	0.96	0.15	0.155
		0.03	0.192	0.207	0.97	0.05	0.224
8	1	-2.47	1021.523	10.996	1.00	1.11	0.399
		3.63	1215.715	12.912	0.96	1.01	0.453
	2	-0.08	75.236	0.585	0.99	-0.10	0.080
		0.26	169.272	1.560	0.97	-0.18	0.113
	3	-0.01	0.146	0.131	0.97	0.15	0.140
		0.00	0.157	0.162	0.93	0.02	0.188
	4	-0.01	0.166	0.174	0.96	0.15	0.155
		0.03	0.192	0.207	0.97	0.05	0.224

Table 12: Simulation results under multivariate assumption comparing multivariate and univariate approaches with binary endpoint

Setting	Method	Multivariate				Univariate	
		Bias	SE	SD	CR	Bias	SD
1	1	0.04	0.234	0.244	0.96	0.35	0.199
		0.09	0.213	0.231	0.91	0.19	0.167
	2	0.00	0.137	0.139	0.97	0.03	0.110
		-0.01	0.135	0.146	0.93	-0.05	0.123
	3	0.00	0.140	0.140	0.98	0.09	0.126
		-0.02	0.136	0.142	0.93	-0.01	0.128
	4	0.01	0.151	0.153	0.97	0.10	0.128
		-0.02	0.146	0.145	0.96	-0.01	0.130
2	1	-0.47	1.303	1.242	0.99	0.62	0.291
		0.75	1.288	1.233	0.94	0.66	0.278
	2	0.00	0.181	0.164	0.99	-0.05	0.090
		-0.01	0.195	0.163	0.99	-0.12	0.117
	3	0.00	0.143	0.144	0.98	0.10	0.127
		-0.02	0.140	0.145	0.93	-0.01	0.133
	4	0.01	0.151	0.153	0.97	0.10	0.128
		-0.02	0.146	0.145	0.96	-0.01	0.130
3	1	0.04	0.418	0.429	0.98	0.47	0.241
		0.08	0.400	0.430	0.98	0.22	0.238
	2	0.00	0.230	0.232	0.98	0.10	0.134
		-0.02	0.241	0.258	0.97	-0.02	0.173
	3	0.00	0.238	0.228	1.00	0.18	0.160
		-0.03	0.248	0.243	0.95	0.03	0.188
	4	0.01	0.280	0.272	0.99	0.19	0.164
		-0.02	0.292	0.276	0.97	0.04	0.198
4	1	-1.67	179.173	9.380	1.00	0.74	0.333
		2.01	205.153	10.349	0.99	0.64	0.369
	2	-0.12	10.005	0.808	1.00	-0.01	0.104
		0.11	13.354	0.822	1.00	-0.12	0.151
	3	0.00	0.253	0.240	1.00	0.19	0.162
		-0.02	0.265	0.253	0.96	0.03	0.198
	4	0.01	0.280	0.272	0.99	0.19	0.164
		-0.02	0.292	0.276	0.97	0.04	0.198

Setting	Method	Multivariate				Univariate	
		Bias	SE	SD	CR	Bias	SD
5	1	0.08	0.171	0.178	0.90	0.27	0.208
		0.11	0.147	0.151	0.89	0.13	0.163
	2	0.01	0.115	0.120	0.95	-0.04	0.109
		-0.01	0.112	0.118	0.93	-0.11	0.113
	3	0.00	0.116	0.111	0.96	-0.02	0.111
		-0.02	0.111	0.114	0.91	-0.08	0.117
	4	0.00	0.121	0.114	0.96	-0.02	0.113
		-0.01	0.118	0.124	0.89	-0.08	0.123
6	1	-0.69	0.808	0.796	0.95	0.95	0.455
		1.42	1.025	0.959	0.76	1.03	0.411
	2	0.01	0.154	0.145	0.95	-0.13	0.096
		-0.02	0.153	0.138	0.90	-0.18	0.110
	3	0.00	0.118	0.114	0.96	-0.02	0.114
		-0.02	0.113	0.117	0.91	-0.08	0.122
	4	0.00	0.121	0.114	0.96	-0.02	0.113
		-0.01	0.118	0.124	0.89	-0.08	0.123
7	1	0.08	0.326	0.324	0.95	0.49	0.340
		0.11	0.293	0.288	0.96	0.20	0.280
	2	0.01	0.203	0.204	0.96	0.08	0.177
		-0.01	0.205	0.207	0.92	-0.06	0.194
	3	0.00	0.210	0.201	0.98	0.14	0.196
		-0.02	0.214	0.211	0.94	-0.01	0.216
	4	-0.01	0.238	0.218	0.99	0.14	0.213
		0.02	0.263	0.262	0.94	0.03	0.266
8	1	-1.12	1243.604	13.833	1.00	1.13	0.611
		1.85	1480.000	16.263	0.99	0.96	0.584
	2	0.03	91.595	0.765	0.98	-0.10	0.133
		0.00	206.071	1.963	0.94	-0.19	0.157
	3	0.00	0.223	0.212	0.99	0.15	0.219
		-0.01	0.235	0.228	0.93	0.01	0.244
	4	-0.01	0.238	0.218	0.99	0.14	0.213
		0.02	0.263	0.262	0.94	0.03	0.266

Table 13: Simulation results under multivariate assumption comparing multivariate and univariate approaches with time-to-event endpoint

Setting	Method	Multivariate				Univariate	
		Bias	SE	SD	CR	Bias	SD
1	1	0.04	0.266	0.286	0.92	0.36	0.235
		0.11	0.237	0.245	0.93	0.21	0.182
	2	0.00	0.155	0.169	0.92	0.04	0.138
		0.00	0.148	0.159	0.92	-0.03	0.138
	3	0.01	0.157	0.163	0.93	0.10	0.154
		0.00	0.145	0.147	0.95	0.00	0.137
	4	0.01	0.169	0.177	0.96	0.11	0.165
		0.00	0.154	0.164	0.92	0.01	0.149
2	1	-0.59	8.130	1.353	1.00	0.63	0.319
		0.89	8.426	1.320	1.00	0.69	0.290
	2	0.00	1.357	0.190	0.99	-0.05	0.116
		0.00	1.164	0.185	0.98	-0.10	0.133
	3	0.01	0.161	0.165	0.95	0.10	0.159
		0.00	0.149	0.151	0.95	0.01	0.143
	4	0.01	0.169	0.177	0.96	0.11	0.165
		0.00	0.154	0.164	0.92	0.01	0.149
3	1	0.02	0.486	0.466	0.97	0.48	0.284
		0.13	0.479	0.442	0.97	0.25	0.259
	2	0.00	0.265	0.263	0.96	0.11	0.166
		0.01	0.292	0.272	0.95	0.00	0.190
	3	0.00	0.271	0.265	0.97	0.19	0.192
		0.00	0.278	0.261	0.95	0.05	0.200
	4	0.01	0.338	0.332	0.97	0.21	0.210
		0.01	0.351	0.330	0.95	0.06	0.232
4	1	-1.24	92898.200	5.495	1.00	0.75	0.370
		1.61	106425.736	6.035	1.00	0.69	0.390
	2	-0.04	7616.258	0.504	1.00	0.00	0.136
		0.05	10241.661	0.601	0.99	-0.10	0.169
	3	0.00	0.292	0.277	0.96	0.20	0.200
		0.01	0.301	0.274	0.96	0.06	0.209
	4	0.01	0.338	0.332	0.97	0.21	0.210
		0.01	0.351	0.330	0.95	0.06	0.232

Setting	Method	Multivariate				Univariate	
		Bias	SE	SD	CR	Bias	SD
5	1	0.07	0.203	0.205	0.94	0.25	0.246
		0.13	0.161	0.178	0.87	0.15	0.206
	2	0.00	0.137	0.136	0.96	-0.05	0.127
		0.00	0.122	0.128	0.96	-0.10	0.135
	3	0.00	0.134	0.137	0.94	-0.02	0.143
		-0.01	0.117	0.127	0.94	-0.08	0.140
	4	0.00	0.140	0.137	0.96	-0.02	0.141
		0.00	0.122	0.140	0.94	-0.07	0.150
6	1	-0.74	3.412	0.830	1.00	0.93	0.537
		1.51	5.366	1.077	1.00	1.06	0.486
	2	0.00	0.520	0.161	0.99	-0.14	0.110
		-0.01	1.607	0.139	1.00	-0.17	0.131
	3	0.00	0.137	0.140	0.93	-0.02	0.144
		-0.01	0.119	0.129	0.95	-0.07	0.145
	4	0.00	0.140	0.137	0.96	-0.02	0.141
		0.00	0.122	0.140	0.94	-0.07	0.150
7	1	0.06	0.390	0.399	0.96	0.48	0.397
		0.13	0.355	0.361	0.96	0.21	0.360
	2	-0.01	0.246	0.242	0.97	0.07	0.204
		0.00	0.254	0.238	0.96	-0.06	0.235
	3	0.00	0.241	0.246	0.95	0.14	0.256
		-0.01	0.234	0.242	0.95	0.00	0.265
	4	-0.02	0.276	0.249	0.97	0.14	0.253
		0.03	0.296	0.301	0.95	0.04	0.343
8	1	1.71	10032598.422	24.923	1.00	1.12	0.716
		-1.45	11950680.950	29.647	1.00	0.99	0.734
	2	0.14	627708.499	1.922	0.99	-0.10	0.150
		-0.43	1401293.651	4.105	1.00	-0.19	0.197
	3	0.00	0.258	0.255	0.96	0.15	0.263
		0.01	0.256	0.256	0.94	0.02	0.302
	4	-0.02	0.276	0.249	0.97	0.14	0.253
		0.03	0.296	0.301	0.95	0.04	0.343

Table 14: Simulation results under univariate assumption comparing multivariate and univariate approaches with continuous endpoint

Setting	Method	Multivariate				Univariate	
		Bias	SE	SD	CR	Bias	SD
1	1	0.23	0.137	0.154	0.70	0.18	0.125
		0.20	0.116	0.113	0.59	0.19	0.097
	2	0.00	0.092	0.102	0.92	-0.02	0.083
		0.00	0.089	0.082	0.93	0.00	0.075
	3	0.01	0.091	0.099	0.93	-0.02	0.079
		-0.01	0.084	0.086	0.96	-0.01	0.078
	4	0.02	0.102	0.111	0.93	-0.01	0.084
		0.00	0.095	0.101	0.94	0.00	0.092
2	1	0.58	0.295	0.304	0.49	0.51	0.228
		0.94	0.302	0.293	0.03	0.92	0.239
	2	0.01	0.122	0.126	0.94	-0.01	0.093
		-0.01	0.120	0.107	0.93	0.00	0.097
	3	0.01	0.097	0.102	0.94	-0.02	0.081
		0.00	0.090	0.097	0.92	-0.01	0.086
	4	0.02	0.102	0.111	0.93	-0.01	0.084
		0.00	0.095	0.101	0.94	0.00	0.092
3	1	0.24	0.178	0.202	0.77	0.19	0.162
		0.21	0.168	0.165	0.85	0.19	0.142
	2	0.01	0.118	0.131	0.92	-0.01	0.107
		0.00	0.127	0.118	0.94	0.00	0.107
	3	0.02	0.122	0.129	0.93	-0.01	0.102
		-0.01	0.130	0.128	0.94	-0.01	0.114
	4	0.02	0.140	0.145	0.93	-0.01	0.106
		0.00	0.152	0.154	0.95	0.00	0.140
4	1	0.61	0.378	0.397	0.67	0.53	0.297
		0.96	0.432	0.401	0.28	0.93	0.335
	2	0.02	0.155	0.158	0.95	0.00	0.122
		0.00	0.169	0.143	0.93	0.01	0.135
	3	0.02	0.130	0.133	0.94	-0.01	0.104
		-0.01	0.140	0.143	0.94	-0.01	0.126
	4	0.02	0.140	0.145	0.93	-0.01	0.106
		0.00	0.152	0.154	0.95	0.00	0.140

Setting	Method	Multivariate				Univariate	
		Bias	SE	SD	CR	Bias	SD
5	1	0.28	0.134	0.120	0.47	0.02	0.083
		0.28	0.122	0.127	0.34	0.07	0.081
	2	0.00	0.084	0.073	0.96	-0.12	0.056
		0.00	0.087	0.081	0.95	-0.09	0.058
	3	0.00	0.081	0.075	0.96	-0.12	0.056
		0.00	0.082	0.084	0.95	-0.09	0.064
	4	0.00	0.089	0.093	0.97	-0.12	0.062
		0.01	0.092	0.098	0.94	-0.08	0.075
6	1	1.45	0.576	0.517	0.05	0.50	0.218
		1.99	0.630	0.618	0.00	1.17	0.310
	2	0.00	0.147	0.129	0.96	-0.11	0.068
		-0.01	0.153	0.132	0.93	-0.09	0.082
	3	0.00	0.087	0.087	0.96	-0.12	0.060
		0.00	0.089	0.093	0.93	-0.09	0.070
	4	0.00	0.089	0.093	0.97	-0.12	0.062
		0.01	0.092	0.098	0.94	-0.08	0.075
7	1	0.29	0.218	0.191	0.85	-0.03	0.126
		0.28	0.207	0.208	0.81	0.05	0.144
	2	0.00	0.135	0.116	0.97	-0.15	0.085
		0.00	0.145	0.138	0.94	-0.11	0.103
	3	0.00	0.141	0.120	0.97	-0.15	0.084
		0.00	0.158	0.162	0.92	-0.11	0.119
	4	0.01	0.172	0.160	0.97	-0.15	0.095
		0.04	0.215	0.220	0.93	-0.08	0.160
8	1	1.55	1.070	0.875	0.70	0.41	0.331
		2.11	1.223	1.032	0.31	1.16	0.533
	2	0.02	0.256	0.201	0.98	-0.14	0.105
		0.01	0.290	0.241	0.94	-0.09	0.146
	3	0.01	0.161	0.143	0.95	-0.15	0.090
		0.02	0.194	0.201	0.93	-0.09	0.136
	4	0.01	0.172	0.160	0.97	-0.15	0.095
		0.04	0.215	0.220	0.93	-0.08	0.160

Table 15: Simulation results under univariate assumption comparing multivariate and univariate approaches with binary endpoint

Setting	Method	Multivariate				Univariate	
		Bias	SE	SD	CR	Bias	SD
1	1	0.21	0.202	0.200	0.79	0.17	0.172
		0.18	0.168	0.174	0.83	0.16	0.161
	2	0.00	0.134	0.132	0.95	-0.02	0.114
		-0.02	0.128	0.135	0.94	-0.02	0.127
	3	0.00	0.133	0.128	0.96	-0.03	0.112
		-0.03	0.124	0.124	0.94	-0.03	0.121
	4	0.01	0.142	0.136	0.96	-0.02	0.113
		-0.02	0.132	0.126	0.97	-0.02	0.123
2	1	0.56	0.376	0.379	0.72	0.49	0.286
		0.88	0.382	0.372	0.33	0.86	0.327
	2	0.00	0.155	0.153	0.94	-0.02	0.118
		-0.03	0.150	0.154	0.91	-0.02	0.143
	3	0.00	0.137	0.132	0.97	-0.03	0.113
		-0.03	0.128	0.127	0.94	-0.03	0.121
	4	0.01	0.142	0.136	0.96	-0.02	0.113
		-0.02	0.132	0.126	0.97	-0.02	0.123
3	1	0.21	0.256	0.246	0.93	0.17	0.211
		0.17	0.244	0.251	0.91	0.15	0.236
	2	0.00	0.169	0.160	0.97	-0.02	0.140
		-0.02	0.185	0.193	0.95	-0.03	0.184
	3	0.00	0.171	0.160	0.98	-0.03	0.140
		-0.04	0.184	0.179	0.95	-0.04	0.176
	4	0.01	0.186	0.174	0.97	-0.02	0.141
		-0.03	0.202	0.190	0.96	-0.03	0.186
4	1	0.56	0.478	0.47	0.79	0.50	0.351
		0.89	0.554	0.53	0.71	0.86	0.483
	2	0.00	0.197	0.18	0.96	-0.02	0.146
		-0.03	0.216	0.21	0.95	-0.02	0.206
	3	0.00	0.177	0.17	0.99	-0.02	0.143
		-0.04	0.191	0.18	0.93	-0.04	0.178
	4	0.01	0.186	0.17	0.97	-0.02	0.141
		-0.03	0.202	0.19	0.96	-0.03	0.186

Setting	Method	Multivariate				Univariate	
		Bias	SE	SD	CR	Bias	SD
5		0.29	0.205	0.214	0.76	0.03	0.154
	1	0.26	0.184	0.183	0.72	0.05	0.143
	2	0.01	0.128	0.132	0.95	-0.11	0.103
		-0.01	0.129	0.133	0.91	-0.11	0.110
	3	0.00	0.125	0.117	0.95	-0.11	0.098
		-0.02	0.126	0.125	0.93	-0.11	0.109
	4	0.00	0.131	0.120	0.98	-0.11	0.097
		-0.01	0.134	0.137	0.91	-0.10	0.115
6	1	1.48	0.699	0.730	0.33	0.54	0.363
		1.93	0.751	0.685	0.06	1.11	0.447
	2	0.01	0.182	0.182	0.95	-0.10	0.113
		-0.03	0.179	0.172	0.90	-0.10	0.132
	3	0.01	0.129	0.121	0.97	-0.11	0.098
		-0.01	0.131	0.132	0.92	-0.10	0.112
	4	0.00	0.131	0.120	0.98	-0.11	0.097
		-0.01	0.134	0.137	0.91	-0.10	0.115
7	1	0.31	0.338	0.346	0.91	-0.01	0.229
		0.27	0.318	0.305	0.93	0.02	0.236
	2	0.01	0.209	0.213	0.95	-0.14	0.155
		-0.01	0.221	0.217	0.92	-0.12	0.183
	3	0.01	0.209	0.192	0.98	-0.14	0.151
		-0.01	0.228	0.215	0.92	-0.13	0.183
	4	0.01	0.233	0.199	0.98	-0.14	0.155
		0.02	0.276	0.266	0.94	-0.10	0.207
8	1	1.59	1.323	1.250	0.89	0.46	0.537
		2.05	1.461	1.195	0.75	1.10	0.746
	2	0.04	0.323	0.299	0.95	-0.13	0.172
		0.00	0.344	0.300	0.89	-0.10	0.226
	3	0.02	0.225	0.199	0.98	-0.14	0.153
		0.01	0.255	0.241	0.91	-0.11	0.195
		0.01	0.233	0.199	0.98	-0.14	0.155
	4	0.02	0.276	0.266	0.94	-0.10	0.207

Table 16: Simulation results under univariate assumption comparing multivariate and univariate approaches with time-to-event endpoint

Setting	Method	Multivariate				Univariate	
		Bias	SE	SD	CR	Bias	SD
1		0.16	0.204	0.205	0.90	0.18	0.222
	1	0.16	0.171	0.172	0.84	0.18	0.171
	2	0.01	0.152	0.152	0.94	-0.01	0.149
		0.00	0.137	0.145	0.93	-0.01	0.137
	3	0.01	0.143	0.145	0.94	-0.02	0.144
		-0.01	0.128	0.130	0.96	-0.02	0.127
	4	0.01	0.153	0.151	0.93	-0.01	0.150
		0.00	0.133	0.144	0.93	-0.01	0.143
2	1	0.33	0.343	0.290	0.90	0.51	0.355
		0.51	0.341	0.269	0.77	0.90	0.358
	2	0.01	0.196	0.165	0.96	-0.01	0.155
		-0.01	0.192	0.155	0.95	0.00	0.154
	3	0.01	0.146	0.147	0.94	-0.02	0.147
		-0.01	0.129	0.135	0.95	-0.01	0.135
	4	0.01	0.153	0.151	0.93	-0.01	0.150
		0.00	0.133	0.144	0.93	-0.01	0.143
3	1	0.16	0.288	0.266	0.93	0.19	0.274
		0.17	0.279	0.258	0.97	0.18	0.254
	2	0.01	0.215	0.197	0.96	-0.01	0.183
		0.00	0.227	0.210	0.94	0.00	0.199
	3	0.01	0.197	0.192	0.95	-0.02	0.178
		-0.01	0.195	0.190	0.95	-0.02	0.186
	4	0.02	0.217	0.206	0.95	-0.01	0.189
		0.00	0.215	0.222	0.91	0.00	0.219
4	1	0.33	0.536	0.372	0.96	0.53	0.439
		0.53	0.682	0.405	0.98	0.93	0.530
	2	0.01	0.320	0.211	0.96	0.00	0.190
		0.00	0.411	0.222	0.96	0.01	0.218
	3	0.01	0.203	0.196	0.95	-0.01	0.183
		-0.01	0.200	0.198	0.95	-0.01	0.199
		0.02	0.217	0.206	0.95	-0.01	0.189
	4	0.00	0.215	0.222	0.91	0.00	0.219

Setting	Method	Multivariate				Univariate	
		Bias	SE	SD	CR	Bias	SD
5	1	0.20	0.235	0.195	0.95	0.02	0.178
		0.21	0.191	0.189	0.84	0.06	0.179
	2	0.00	0.166	0.137	0.98	-0.12	0.118
		0.00	0.143	0.137	0.98	-0.10	0.133
	3	0.00	0.141	0.134	0.94	-0.12	0.120
		-0.01	0.124	0.135	0.94	-0.10	0.131
	4	0.00	0.150	0.134	0.95	-0.12	0.122
		0.00	0.130	0.148	0.94	-0.09	0.138
6	1	0.51	0.652	0.299	1.00	0.50	0.400
		0.91	0.816	0.383	0.98	1.16	0.543
	2	0.00	0.327	0.163	0.97	-0.11	0.129
		-0.01	0.337	0.148	0.98	-0.09	0.150
	3	0.00	0.146	0.136	0.95	-0.12	0.123
		0.00	0.126	0.139	0.95	-0.09	0.135
	4	0.00	0.150	0.134	0.95	-0.12	0.122
		0.00	0.130	0.148	0.94	-0.09	0.138
7	1	0.21	0.437	0.293	1.00	-0.02	0.270
		0.22	0.417	0.316	0.99	0.05	0.310
	2	0.00	0.320	0.205	0.97	-0.15	0.180
		0.00	0.327	0.229	0.97	-0.11	0.230
	3	0.00	0.241	0.205	0.95	-0.15	0.187
		0.00	0.237	0.234	0.96	-0.11	0.230
	4	0.01	0.292	0.205	0.97	-0.15	0.190
		0.03	0.313	0.277	0.96	-0.08	0.266
8	1	0.53	1.745	0.469	1.00	0.41	0.601
		0.95	2.550	0.641	1.00	1.15	0.931
	2	0.01	1.053	0.244	0.98	-0.14	0.193
		0.00	1.299	0.245	0.98	-0.09	0.256
	3	0.01	0.263	0.212	0.96	-0.15	0.193
		0.01	0.266	0.251	0.96	-0.10	0.245
	4	0.01	0.292	0.205	0.97	-0.15	0.190
		0.03	0.313	0.277	0.96	-0.08	0.266

Table 17: Summarized result of multivariate analysis for association between 20% increase in sodium and potassium on logrithm base with various cardiovascular diseases (percentages)

Outcome		Method 1		Method 2		Method 3		Method 4	
		HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI
CHD	Sodium	1.29	(0.92,1.80)	1.08	(1.01,1.14)	1.14	(0.99,1.31)	1.12	(1.00,1.25)
	Potassium	0.94	(0.58,1.51)	0.95	(0.88,1.01)	0.92	(0.84,1.02)	0.91	(0.84,0.98)
Nonfatal MI	sodium	1.17	(0.76,1.79)	1.07	(1.00,1.15)	1.12	(0.96,1.29)	1.10	(0.96,1.25)
	Potassium	0.72	(0.41,1.24)	0.89	(0.80,1.00)	0.88	(0.81,0.95)	0.87	(0.80,0.94)
Coronary death	Sodium	1.77	(0.90,3.50)	1.12	(1.00,1.25)	1.23	(1.01,1.50)	1.20	(1.05,1.38)
	Potassium	1.72	(0.75,3.95)	1.06	(0.93,1.21)	1.00	(0.86,1.16)	0.97	(0.85,1.12)
Stroke	Sodium	0.97	(0.75,1.26)	1.01	(0.96,1.06)	1.01	(0.92,1.11)	1.01	(0.94,1.08)
	Potassium	0.80	(0.51,1.25)	0.95	(0.87,1.03)	0.95	(0.90,1.00)	0.95	(0.90,1.00)
Ischemic Stroke	Sodium	1.02	(0.71,1.48)	1.03	(0.97,1.10)	1.05	(0.91,1.21)	1.04	(0.93,1.15)
	Potassium	0.74	(0.42,1.28)	0.92	(0.81,1.04)	0.92	(0.85,0.99)	0.91	(0.84,0.98)
Hemorrhagic Stroke	Sodium	0.89	(0.39,2.05)	0.94	(0.84,1.05)	0.90	(0.74,1.11)	0.92	(0.74,1.15)
	Potassium	1.49	(0.54,4.10)	1.13	(0.92,1.39)	1.14	(0.99,1.32)	1.16	(1.03,1.31)
Total CVD	Sodium	1.31	(1.00,1.72)	1.07	(1.01,1.13)	1.12	(1.02,1.24)	1.11	(1.02,1.20)
	Potassium	1.11	(0.77,1.60)	0.99	(0.92,1.06)	0.96	(0.89,1.03)	0.94	(0.89,1.00)
Revascularization	Sodium	1.79	(1.11,2.89)	1.13	(1.01,1.27)	1.26	(1.04,1.53)	1.23	(1.04,1.44)
	Potassium	1.47	(0.81,2.67)	1.01	(0.90,1.14)	0.95	(0.84,1.08)	0.93	(0.84,1.03)
Non-Revascularization	Sodium	1.11	(0.86,1.42)	1.04	(0.99,1.09)	1.07	(0.97,1.17)	1.06	(0.98,1.14)
	Potassium	0.86	(0.62,1.19)	0.94	(0.88,1.02)	0.94	(0.89,0.99)	0.93	(0.89,0.97)
Heart Failure	Sodium	1.72	(0.98,3.01)	1.10	(0.96,1.27)	1.21	(0.98,1.49)	1.18	(1.02,1.37)
	Potassium	1.81	(0.87,3.76)	1.08	(0.95,1.23)	1.02	(0.92,1.13)	1.00	(0.89,1.11)

Table 18: Summarized result of multivariate analysis for association between 20% increase in sodium and potassium on logrithm base with various cardiovascular diseases (milligrams)

Outcome		Method 1		Method 2		Method 3		Method 4	
		HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI
CHD	Sodium	1.23	(1.03,1.45)	1.08	(1.01,1.15)	1.11	(0.97,1.28)	1.12	(0.94,1.33)
	Potassium	0.87	(0.61,1.24)	0.95	(0.85,1.06)	0.91	(0.83,0.99)	0.90	(0.83,0.98)
Nonfatal MI	Sodium	1.33	(1.02,1.73)	1.11	(1.02,1.21)	1.18	(1.00,1.39)	1.21	(0.97,1.50)
	Potassium	1.01	(0.64,1.57)	1.00	(0.85,1.17)	0.93	(0.82,1.04)	0.92	(0.82,1.02)
Coronary death	Sodium	1.13	(0.89,1.43)	1.05	(0.96,1.14)	1.05	(0.90,1.22)	1.02	(0.81,1.29)
	Potassium	0.70	(0.42,1.16)	0.89	(0.76,1.04)	0.87	(0.77,0.99)	0.87	(0.78,0.97)
Stroke	Sodium	0.99	(0.80,1.23)	1.00	(0.92,1.08)	0.97	(0.84,1.12)	0.94	(0.81,1.08)
	Potassium	0.70	(0.47,1.05)	0.89	(0.76,1.04)	0.90	(0.82,0.99)	0.91	(0.83,0.99)
Ischemic Stroke	Sodium	1.07	(0.81,1.42)	1.03	(0.94,1.12)	1.01	(0.86,1.19)	0.98	(0.86,1.13)
	Potassium	0.68	(0.42,1.11)	0.88	(0.76,1.02)	0.88	(0.79,0.97)	0.88	(0.80,0.97)
Hemorrhagic Stroke	Sodium	0.69	(0.49,0.96)	0.87	(0.76,1.00)	0.78	(0.59,1.04)	0.74	(0.49,1.11)
	Potassium	0.70	(0.29,1.66)	0.90	(0.68,1.18)	1.00	(0.81,1.24)	1.02	(0.82,1.27)
Total CVD	Sodium	1.13	(0.98,1.30)	1.05	(1.01,1.09)	1.06	(0.97,1.16)	1.06	(0.95,1.18)
	Potassium	0.85	(0.61,1.16)	0.94	(0.85,1.05)	0.92	(0.86,0.98)	0.92	(0.87,0.97)
Revascularization	Sodium	1.31	(1.00,1.72)	1.11	(1.02,1.20)	1.17	(1.01,1.35)	1.20	(0.97,1.48)
	Potassium	1.00	(0.57,1.75)	0.99	(0.85,1.17)	0.93	(0.83,1.04)	0.92	(0.82,1.03)
Non-Revascularization	Sodium	1.11	(0.93,1.31)	1.04	(0.98,1.10)	1.04	(0.94,1.16)	1.03	(0.91,1.16)
	Potassium	0.80	(0.59,1.08)	0.93	(0.83,1.03)	0.91	(0.85,0.98)	0.91	(0.86,0.97)
	Sodium	1.24	(0.98,1.57)	1.08	(0.98,1.19)	1.16	(0.96,1.40)	1.21	(0.99,1.48)
Heart Failure	Potassium	1.37	(0.80,2.34)	1.10	(0.94,1.30)	1.03	(0.91,1.17)	1.02	(0.92,1.13)

TABLE 19: Summarized result of univariate analysis for association between 20% increase in sodium and potassium on logrithm base with various cardiovascular diseases (percentages)

Outcome		Method 1		Method 2		Method 3		Method 4	
		HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI
CHD	Sodium	1.36	(0.91,2.05)	1.07	(0.98,1.16)	1.10	(0.99,1.23)	1.09	(0.99,1.19)
	Potassium	0.62	(0.40,0.96)	0.88	(0.79,0.98)	0.90	(0.84,0.97)	0.90	(0.85,0.96)
Nonfatal MI	Sodium	1.23	(0.84,1.79)	1.04	(0.97,1.12)	1.07	(0.94,1.20)	1.06	(0.95,1.18)
	Potassium	0.50	(0.28,0.89)	0.84	(0.73,0.95)	0.86	(0.80,0.92)	0.86	(0.81,0.92)
Coronary death	Sodium	1.94	(1.01,3.73)	1.14	(0.99,1.32)	1.23	(1.00,1.51)	1.19	(1.05,1.37)
	Potassium	0.84	(0.53,1.35)	0.96	(0.85,1.08)	0.96	(0.86,1.07)	0.96	(0.87,1.07)
Stroke	Sodium	0.98	(0.71,1.34)	1.00	(0.94,1.05)	0.99	(0.90,1.10)	0.99	(0.91,1.08)
	Potassium	0.78	(0.58,1.04)	0.94	(0.86,1.01)	0.95	(0.89,1.00)	0.95	(0.90,1.00)
Ischemic Stroke	Sodium	1.04	(0.76,1.44)	1.01	(0.96,1.07)	1.01	(0.91,1.13)	1.01	(0.92,1.11)
	Potassium	0.64	(0.39,1.05)	0.89	(0.79,1.01)	0.91	(0.84,0.97)	0.91	(0.85,0.97)
Hemorrhagic Stroke	Sodium	0.85	(0.38,1.89)	0.97	(0.85,1.11)	0.95	(0.77,1.18)	0.96	(0.80,1.15)
	Potassium	2.05	(0.98,4.28)	1.20	(0.98,1.47)	1.17	(1.03,1.33)	1.17	(1.03,1.32)
Total CVD	Sodium	1.38	(0.96,1.99)	1.07	(0.98,1.16)	1.11	(1.02,1.20)	1.09	(1.02,1.17)
	Potassium	0.75	(0.59,0.95)	0.93	(0.87,0.98)	0.94	(0.90,0.97)	0.94	(0.91,0.97)
Revascularization	Sodium	1.98	(1.08,3.63)	1.15	(1.01,1.31)	1.24	(1.04,1.47)	1.20	(1.02,1.41)
	Potassium	0.67	(0.48,0.94)	0.90	(0.82,0.99)	0.92	(0.88,0.96)	0.92	(0.88,0.96)
Non-Revascularization	Sodium	1.14	(0.88,1.48)	1.03	(0.97,1.09)	1.04	(0.97,1.13)	1.04	(0.97,1.11)
	Potassium	0.69	(0.51,0.94)	0.91	(0.84,0.98)	0.92	(0.88,0.97)	0.92	(0.88,0.97)
Heart Failure	Sodium	1.86	(0.94,3.68)	1.14	(0.98,1.31)	1.21	(1.01,1.46)	1.18	(1.03,1.36)
	Potassium	0.95	(0.68,1.32)	0.99	(0.91,1.07)	0.99	(0.92,1.06)	0.99	(0.92,1.06)

Table 20: Summarized result of univariate analysis for association between 20% increase in sodium and potassium on logrithm base with various cardiovascular diseases (milligrams)

Outcome		Method 1		Method 2		Method 3		Method 4	
		HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI
CHD	Sodium	3.24	(0.81,12.96)	1.46	(0.98,2.18)	1.48	(1.07,2.06)	1.57	(1.02,2.41)
	Potassium	0.39	(0.18,0.81)	0.73	(0.56,0.95)	0.78	(0.68,0.89)	0.79	(0.68,0.91)
Nonfatal MI	Sodium	4.31	(0.71,26.25)	1.60	(0.97,2.65)	1.63	(1.01,2.64)	1.75	(1.02,3.01)
	Potassium	0.36	(0.15,0.88)	0.72	(0.53,0.97)	0.76	(0.65,0.89)	0.78	(0.66,0.92)
Coronary death	Sodium	2.60	(0.73,9.28)	1.36	(0.93,1.99)	1.38	(0.88,2.15)	1.44	(0.90,2.31)
	Potassium	0.37	(0.14,0.98)	0.72	(0.53,0.98)	0.77	(0.63,0.94)	0.78	(0.66,0.92)
Stroke	Sodium	1.33	(0.59,3.01)	1.10	(0.86,1.39)	1.10	(0.87,1.39)	1.11	(0.85,1.47)
	Potassium	0.59	(0.34,1.04)	0.84	(0.70,1.02)	0.87	(0.77,0.99)	0.88	(0.78,0.99)
Ischemic Stroke	Sodium	2.01	(0.66,6.11)	1.25	(0.88,1.79)	1.26	(0.90,1.77)	1.31	(0.89,1.92)
	Potassium	0.43	(0.18,1.03)	0.76	(0.58,1.00)	0.80	(0.68,0.94)	0.81	(0.69,0.95)
Hemorrhagic Stroke	Sodium	0.20	(0.02,1.69)	0.60	(0.31,1.15)	0.58	(0.30,1.14)	0.54	(0.28,1.06)
	Potassium	2.23	(0.79,6.33)	1.30	(0.92,1.84)	1.24	(0.98,1.57)	1.22	(0.93,1.61)
Total CVD	Sodium	2.22	(0.85,5.76)	1.29	(0.96,1.74)	1.30	(1.02,1.66)	1.36	(1.00,1.85)
	Potassium	0.49	(0.29,0.83)	0.79	(0.66,0.96)	0.83	(0.75,0.91)	0.84	(0.76,0.92)
Revascularization	Sodium	4.01	(0.68,23.62)	1.56	(0.96,2.56)	1.59	(1.00,2.53)	1.70	(1.05,2.75)
	Potassium	0.38	(0.17,0.82)	0.73	(0.56,0.95)	0.77	(0.67,0.88)	0.79	(0.69,0.89)
Non-Revascularization	Sodium	2.06	(0.81,5.21)	1.26	(0.97,1.65)	1.27	(0.97,1.67)	1.32	(1.00,1.74)
	Potassium	0.49	(0.30,0.81)	0.79	(0.66,0.96)	0.83	(0.74,0.92)	0.84	(0.75,0.93)
Heart Failure	Sodium	2.29	(0.74,7.04)	1.31	(0.93,1.84)	1.32	(0.93,1.87)	1.37	(0.94,2.00)
	Potassium	0.74	(0.37,1.51)	0.91	(0.72,1.14)	0.92	(0.76,1.12)	0.93	(0.77,1.11)

Chapter 4: High-dimensional Setting

4.1 Introduction

In this chapter, we focused on high-dimensional measurements for a single exposure of interest, where the sample size is less than the dimension of the variables in building a biomarker model. Nowadays, variable selection problems under high-dimensional data structure encompass a majority of areas in statistics. There are extensive studies that were developed and are devoted to understanding the pros and cons of different variable selection techniques with high-dimensional data. Frank & Friedman (1993) [49] first proposed a technique called bridge regression. Then the nonnegative garrote for shrinkage estimation and the variable selection was mentioned in Breiman & Leo (1995) [50]. Least absolute shrinkage and selection operator (Lasso) with L-1 regularized least squares was studied and introduced by Tibshirani & Robert (1996) [51]. Fan & Li (2001) [52] and Fan & Peng (2004) [53] proposed nonconcave penalized likelihood estimators such as smoothly clipped absolute deviation (SCAD). In particular, Fan & Li (2001) [52] proposed a unified algorithm for optimizing nonconcave penalized likelihood. Efron et al. (2004) [54] introduced the least angle regression for variable selection and presented an algorithm named as LARS. Zou & Li (2008) [55] proposed one-step sparse estimates for nonconcave penalized likelihood models and introduced the LLA algorithm for optimizing nonconcave penalized likelihood.

Predictive performance is important to build a biomarker model with high dimensional sparse data. One problem performing high dimensional model is collinearity among covariates. Collinearity can lead to a wrong model with spurious correlations among variables [56]. Penalized regressions such as Lasso and SCAD have been extensively studied and discussed by many researchers to deal with high-dimensional sparse data. Furthermore, random forest

(RF), which is based on the ranking of predictive power, is another approach for variable selection [57, 58]. In addition, under high-dimensional sparse feature space, the BF build upon variance estimation in our proposed Method 2 needs to be carefully considered and generated. In high-dimensional linear regression, ordinary linear regression does not work properly. A recent study [59] implies that spurious variables can be easily selected to enter the model leading to severe underestimation on error variance. Advanced estimating techniques were developed that can greatly attenuate the bias, including k-fold cross-validation, naive two-stage estimator, and refitted cross-validation (RCV) [59]. The rest of this chapter is organized as below. First, we introduced different methods and detailed variance estimation procedures for BF construction. Second, We will perform extensive simulations to show that the conclusion we obtained from asymptotic results above also holds for finite sample cases. Last, we will extend our methods to real data analysis.

4.2 Methods

Similar to Chapter 2, we consider the case where σ_x^{*2} is known. We propose methods to estimate σ_x^{*2} in the discussion section. In the real data analysis where σ_x^{*2} is not available, we vary this parameter to perform sensitivity analysis.

With high-dimensional data on urine measurements ($\boldsymbol{W}$), we first need to obtain estimated coefficients among n_1 subjects in the biomarker discovery sample of consumed diet (X^*) on high-dimensional blood and urine measurements ($\boldsymbol{W}$) as well as subject characteristics ($\boldsymbol{V}$). In this chapter, three different approaches including Lasso, SCAD, and RF are used to conduct variable selection in high-dimensional statistical inference. We will describe each approach explicitly for every method in the following subsections.

4.2.1 Method 1: The naïve three-step approach with multiple exposures

In the first step, we need to fit a linear regression of X^* on $\boldsymbol{W}$ and $\boldsymbol{V}$. That is:

$$X^* = (1, \boldsymbol{W}^T, \boldsymbol{V}^T)\beta_1 + \epsilon_{X^*}.$$

With the Lasso approach, the coefficients, $\hat{\beta}_1$, minimize the penalized least squares (PL_{Lasso}) as below:

$$PL_{Lasso} = ||X^* - (1, \boldsymbol{W}^T, \boldsymbol{V}^T)\beta_1||^2 + \lambda \sum_{j=1}^{p} |\beta_{1j}|.$$

The Lasso performs variable selection by shrinking coefficients estimates towards zero leading to a sparse model in the end. λ is selected through cross-validation based on the smallest MSE.

With the SCAD approach, a nonconvex penalty is given by:

$$PL_{SCAD}(\beta_{1j}) = \begin{cases} \lambda|\beta_1| \quad if\, |\beta_{1j}| < \lambda \\ 2a\lambda|\beta_1|^2 - 2a\lambda|\beta_1| \quad if\, \lambda < |\beta_{1j}| < a\lambda \\ (a+1)\lambda^2/2 \quad if\, |\beta_{1j}| > a\lambda \end{cases} .$$

The first derivatives of $PL_{SCAD}(\beta_{1j})$ is continuous and is given by:

$$PL_{SCAD}(\beta_1)' = \lambda\{I(\beta_1 < \lambda) + \frac{(a\lambda - \beta_1)_+}{(a-1)\lambda} I(\beta_1 > \lambda)\},$$

for some a>2 and $\beta_1 > 0$. Similar to Lasso, λ in SCAD is selected through cross-validation based on the smallest MSE whereas a is set to be 3.7 based on simulation results and Bayesian statistical point of view from Fan & Li (2001) [52]. Other than penalized regression as we described above, RF is another choice for variable selection.

The basic concept is to grow a regression tree. The general form of a regression tree is as below:

$$X^* = \sum_{m=1}^{M} c_m 1_{(\boldsymbol{W},\boldsymbol{V}) \in R_m} ,$$

where $R_1, ..., R_M$ denotes a partition of feature space. Then we can repeat this procedure to build the RF by considering the approximately square root of the total number of predictors each time. The advantages of RF is we can see the contribution of each variable to the regression tree and their relative importance.

For each method, we did the direct-regression selection and post-regression selection. For direct-regression, we applied the estimated model from each approach to predict the long-term dietary intake directly. For post-regression, we performed linear regression afterward with selected variables from each approach. Specifically, for Lasso and SCAD, we have:

$$\hat{\beta}_1 = argmin\{||X^* - (1, \boldsymbol{W}^T_{selected}, \boldsymbol{V}^T_{selected})\beta_1||^2_2\},$$

$$\beta_1 \in R^P \ and \ \hat{\beta}_{1j} = 0 \ \forall \ j \notin \hat{S}.$$

For RF, the 10 most important variables will be considered as final selected variables. For both direct and post-regression, we thought two ways dealing with $\boldsymbol{W}$ and $\boldsymbol{V}$, where one is considering both $\boldsymbol{W}$ and $\boldsymbol{V}$ in the variable selection procedure, while the other is considering only $\boldsymbol{W}$. To be more specific, in Lasso and SCAD, the penalization will be applied to ($\boldsymbol{W}$, $\boldsymbol{V}$) and to only $\boldsymbol{W}$, respectively. In RF, the decision trees will be built by considering ($\boldsymbol{W}$, $\boldsymbol{V}$) and only $\boldsymbol{W}$, respectively.

With estimated $\hat{\beta}_1$ we had in the prior step, we can then compute $\hat{X}_1 = (1, \boldsymbol{W}^T, \boldsymbol{V}^T)\hat{\beta}_1$ to predict the long-term dietary intake (Z) among the n_2 calibration samples and run a regression of $\hat{X}_1$ on self-reported food frequency questionnaire data (Q) and $\boldsymbol{V}$ to build calibration equation using the n_2 calibration samples same in Chapter 2. Finally, the association parameter $\hat{\theta}_1$ can be estimated by solving the score equations 2.1 in Chapter 2 with respect to continuous, binary, and time-to-event endpoints.

4.2.2 Method 2: Three-step with Bias Correction

As shown in method 1, we have a bias factor in $\hat{Z}_1$ when using $\hat{X}_1$, so we propose a bias-corrected estimator $\hat{X}_2 = \hat{X}_1 \widehat{BF}^{-1}$ where,

$$\widehat{BF} = \hat{R}^2_{1|\boldsymbol{V}} = 1 - \frac{\widehat{Var}(X^*|\boldsymbol{W},\boldsymbol{V}) - \sigma_x^{*2}}{\widehat{Var}(X^*|\boldsymbol{V}) - \sigma_x^{*2}},$$

is an estimated version of the bias factor.

When we perform direct-selection, we used mean cross-validated errors as $\widehat{Var}(X^*|\boldsymbol{W},\boldsymbol{V})$ in penalized regression and RF to obtain $\widehat{BF}$.

Regarding post-selection, we first obtain selected variables from Lasso, SCAD, or RF. Then coefficients can be estimated by refitting a linear regression. For ease of interpretation, we consider both $\boldsymbol{W}$ and $\boldsymbol{V}$ in variable selection for the rest of this subsection, we have:

$$X^* = (1, \boldsymbol{W}_s^T, \boldsymbol{V}_s^T)\beta^{PS}_{WV} + \epsilon_{X_k^*}.$$

Second, we can fit a low dimensional model as below:

$$X^* = (1, \boldsymbol{V}^T)\beta_V + \epsilon'_{X_k^*},$$

where $\boldsymbol{W}_s$ and $\boldsymbol{V}_s$ represent for selected $\boldsymbol{W}$ and $\boldsymbol{V}$, β^{PS}_{WV}, β_V are the corresponding coefficients and $\epsilon_{X_k^*}$, $\epsilon'_{X_k^*}$ are the corresponding error terms in above equations. From there, BF can be estimated as:

$$\widehat{BF} = 1 - \frac{\widehat{Var}(X^*|\boldsymbol{W},\boldsymbol{V}) - \sigma_x^{*2}}{\widehat{Var}(X^*|\boldsymbol{V}) - \sigma_x^{*2}},$$

where

$$\widehat{Var}(X^*|\boldsymbol{W},\boldsymbol{V}) = n^{-1}\sum_{i=1}^{n_1}(X_i^* - (1, \boldsymbol{W_{si}^T}, \boldsymbol{V_{si}^T})\hat{\beta}_{WV}^{PS})^2,$$

$$\widehat{Var}(X^*|\boldsymbol{V}) = n^{-1}\sum_{i=1}^{n_1}(X_i^* - \boldsymbol{V_i^T}\hat{\beta}_V)^2.$$

As we can see above, a good estimation of $\widehat{Var}(X^*|\boldsymbol{W},\boldsymbol{V})$ is the key to get a good estimation of BF. Chatterjee & Jafarov (2015) [60] showed that estimator, $\widehat{Var}(X^*|\boldsymbol{W},\boldsymbol{V})$, listed above lead to downward bias with Lasso, So we decide to calculate and compare three types of $\widehat{Var}(X^*|\boldsymbol{W},\boldsymbol{V})$ in our study with post-selection. We will describe each of them below:

(i) K-fold cross-validation

We need to fit penalized regression or RF with a training dataset and get predicted X^* with selected $(\boldsymbol{W}, \boldsymbol{V})$ for each fold. Denote the selected subset as S_k for each training set X^*_{-k}.

Denote $\boldsymbol{W}_{S_k}$ as selected $\boldsymbol{W}$ and $\boldsymbol{V}_{S_k}$ as selected $\boldsymbol{V}$ in the $(K$-1) training dataset for each fold, then we can fit a linear regression with $\boldsymbol{W}_{S_k}$, $\boldsymbol{V}_{S_k}$, $\boldsymbol{V}_k$ and X_k^* in the kth test dataset as below:

$$X_k^* = (1, \boldsymbol{W}_{S_k}^T, \boldsymbol{V}_{S_k}^T)\beta_{W_{S_k}V_{S_k}} + \epsilon_{X_k^*}. \qquad (4.1)$$

The estimated value with (4.1) is denoted as $\widehat{X_{1k}^*}$

$$X_k^* = (1, \boldsymbol{V}_k^T)\beta_{V_k} + \epsilon'_{X_k^*}. \qquad (4.2)$$

The corresponding regression coefficients and error terms for each fold are $\beta_{W_{S_k}V_{S_k}}$, β_{V_k}, and $\epsilon_{X_k^*}$, $\epsilon'_{X_k^*}$ in the above equations, respectively. The estimated value with (4.2) is denoted as $\widehat{X_{2k}^*}$. With (4.1) and (4.2), we get the estimated values of X^* for the whole sample 1, that is,

$$\widehat{X_1^*} = (\widehat{X_{11}^*}, \widehat{X_{12}^*} ... \widehat{X_{1K}^*}),$$

$$\widehat{X_2^*} = (\widehat{X_{21}^*}, \widehat{X_{22}^*} ... \widehat{X_{2K}^*}).$$

With $\widehat{X_1^*}$ and $\widehat{X_2^*}$, $\widehat{BF}$ can be calculated,

$$\widehat{BF} = 1 - \frac{\sum_{i=1}^{n_1} (X_i^* - \widehat{X_{1i}^*})^2}{\sum_{i=1}^{n_1} (X_i^* - \widehat{X_{2i}^*})^2}.$$

(ii) Naive two-stage estimator

When penalized regression for variable selection was applied for variable selection, the choice of λ is an essential factor in determining whether accurately informative estimators can enter the model. A very large λ can lead to a failure to include all correct contributed variables into the model and an upward bias usually appeared in variance estimation with more significant signal appeared in each selected variable. On the other hand, when a too small λ is applied, unnecessary variables with many noises may enter the model leading to a downward bias in variance estimation. Hence the size of λ is to some extent determines the number of variables entering into the model and the degree of shrinkage towards zero of the estimated coefficient for each variable. Based on Ried, Tibshirani & Friedman (2014)'s [61] paper, selecting the appropriate λ to maintain the balance is suggested. That is:

$$\widehat{Var^*}(X^* | \boldsymbol{W}, \boldsymbol{V}) = (n - \hat{s}_\lambda)^{-1} \sum_i (X_i^* - (1, \boldsymbol{W}_{si}^T, \boldsymbol{V}_{si}^T) \hat{\beta}_{WV}^{PS})^2,$$

where $\hat{s}_\lambda$ is the number of nonzero elements in $\hat{b}$ at the regulation parameter λ selected with K-fold (usually 5 to 10) cross-validation. Then we have:

$$\widehat{BF} = 1 - \frac{\widehat{Var^*}(X^*|\boldsymbol{W}, \boldsymbol{V})}{\widehat{Var}(X^*|\boldsymbol{V})}.$$

(iii) Refitted cross-validation estimator (RCV).

This estimator is derived by the refitted cross-validation (RCV) procedure suggested by Fan et al. (2012) [59]. We first split the dataset into roughly equal two parts, $(X^{*(1)}, \boldsymbol{W}^{(1)}, \boldsymbol{V}^{(1)})$ and $(X^{*(2)}, \boldsymbol{W}^{(2)}, \boldsymbol{V}^{(2)})$. Then on the first part, penalized regression and RF need to be performed. With penalized regression, we fit Lasso or SCAD on $\boldsymbol{W}$ and $\boldsymbol{V}$ with cross-validated $\hat{\lambda}_1$ to obtain the non-zero estimated coefficients for $\boldsymbol{W}$ and $\boldsymbol{V}$. With RF, the 10 most important variables are selected based on the residual sum of squares (RSS). Then we refit the model with selected $\boldsymbol{W}$ and $\boldsymbol{V}$ to get the post-selected estimations on their coefficients, $\hat{\beta}_{WV}^{PS(1)}$. Next, with selected $\boldsymbol{W}$ and $\boldsymbol{V}$ in $\boldsymbol{W}^{(2)}$ and $\boldsymbol{V}^{(2)}$, we can obtain the following variance estimate on the second part:

$$\widehat{Var_1^{**}}(X^*|\boldsymbol{W}, \boldsymbol{V}) = (n - \hat{s}^{(1)})^{-1} \sum_i (X_i^{*(2)} - (1, \boldsymbol{W}_{si}^{T(2)}, \boldsymbol{V}_{si}^{T(2)}) \hat{\beta}_{WV}^{PS(1)})^2,$$

where $\hat{s}^{(1)}$ is the number of selected variables in the first part. Repeating the mirror image procedure on $(X^{*(2)}, \boldsymbol{W}_s^{(2)}, \boldsymbol{V}_s^{(2)})$, we can obtain $\hat{\lambda}_2$, selected W obtained from Lasso in the second part and $\widehat{Var}_2(X^*|\boldsymbol{W}, \boldsymbol{V})$. Last, $\widehat{BF}$ can be derived as below:

$$\widehat{Var^{**}}(X^*|\boldsymbol{W}, \boldsymbol{V}) = \frac{1}{2}(\widehat{Var_1^{**}}(X^*|\boldsymbol{W}, \boldsymbol{V}) + \widehat{Var_2^{**}}(X^*|\boldsymbol{W}, \boldsymbol{V})),$$

$$\widehat{BF} = 1 - \frac{\widehat{Var^{**}}(X^*|\boldsymbol{W}, \boldsymbol{V})}{\widehat{Var}(X^*|\boldsymbol{V})}.$$

With $\widehat{BF}$, we have $\hat{X}_2 = \hat{X}_1/\widehat{BF}$. Then we can follow the steps illustrated in Method 2 section of Chapter 2 and estimate the $\hat{\theta}_2$ by solving estimating equations 2.2 for continuous, binary, and time-to-event points, respectively.

This method does not require the self-reported dietary intake data (Q) in the feeding study. As a remark, even if the self-reported data is available in the feeding study, the association between the self-reported and the actual dietary intake in the feeding study might be different from that association from the cohort because of, (1) the modification on the dietary pattern during the controlled feeding study or (2) the potential change in dietary preference in the period of the feeding study. This method is robust to the difference in the association since we have not directly included Q in stage 1 for biomarker construction. We will next propose two methods that require the availability of self-reported data in the feeding study and assume the association between the self-reported and the actual dietary intake to be the same among all studies.

4.2.3 Method 3: Three-step with self-reported data

The three steps of the first method remain the same, but in the first step regression model, the log-transformed self-reported food frequency questionnaire data (Q) is added. That is, for the first step, predictors, $\boldsymbol{W}$, $\boldsymbol{V}$ and Q are used here to build the biomarker, and then in the second step, we use $\boldsymbol{W}$, $\boldsymbol{V}$ and Q to predict Z. Lasso, SCAD, and RF previously described by considering both direct-selection and post-selection are all applied in Method 3 to perform variable selection and estimating effects in high-dimensional statistical inference. Then with estimated $\hat{\beta}_3$ in the first step, $\hat{X}_3 = (1, \boldsymbol{W}^T, Q, \boldsymbol{V}^T)\hat{\beta}_3$, and then we can follow the steps illustrated in Method 3 section of Chapter 2 and estimate the $\hat{\theta}_3$ by solving estimating equations 2.3 for continuous, binary, and time-to-event points, respectively.

4.2.4 Method 4: Direct Estimation

Since the predictors in the first step in Method 4 are $\boldsymbol{V}$ and Q, which are low-dimensional, the procedure for Method 4 in this chapter is the same as in Chapter 2. We build the estimating equation by regressing X^* on Q and $\boldsymbol{V}$ in the first step and directly apply it to the third step. Then we build the calibration equation using the feeding study by regressing X^* on $\boldsymbol{V}$ and Q and use the calibration equation to predict Z and perform a Cox regression of Y on Z and $\boldsymbol{V}$ in the full cohort to estimate the association parameter. In other words, we have $\hat{\gamma}_4 = \sum_{i=1}^{n_1} \left\{(1, Q_i, \boldsymbol{V}_i^T)^T(1, Q_i, \boldsymbol{V}_i^T)\right\}^{-1} \sum_{i=1}^{n_1} \left\{(1, Q_i, \boldsymbol{V}_i^T)^T X_i^*\right\}$, $\hat{Z}_4 = (1, Q, \boldsymbol{V}^T)\hat{\gamma}_4$ and $\hat{\theta}_4$ by solving estimating equations 2.4 for continuous, binary and time-to-event points, respectively.

4.3 Simulation

We simulate data with different sparsity, effect size, and shape within high-dimensional statistical inference. We are intended to explore how the sparsity, effect size, and shape among different measurements play an important role in the bias and variance of different estimators. The bias, empirical standard error (SD), estimated standard error (SE), and coverage rate for nominal 95% confidence interval (CR) will be compared with varying sample sizes, effect shape, effect sizes, and correlation structures. We study both settings whether the penalty was given or not on the personal characteristics V during the first stage penalized regression. Data are generated from Cox, logistic and linear models with time-to-event, binary and continuous endpoints, respectively.

$$
\begin{aligned}
&(Z, V) \sim N\left(0, \begin{pmatrix} 1-\sigma_x^2 & \rho \\ \rho & 1 \end{pmatrix}\right), \\
&\boldsymbol{W} = b_0 + b_1 X + b_2 V + \epsilon_w \ , \\
&X = Z + \epsilon_x \ , \\
&X^* = X + \epsilon_x^* \ , \\
&Q = a_0 + a_1 Z + a_2 V + \epsilon_q \ ,
\end{aligned}
$$

where Z, V, X and Q are all in one-dimensional space while $\boldsymbol{W}$ is in high-dimensional space. In our simulation, we set $\boldsymbol{W}$ in 100-dimensional space. That is, $\boldsymbol{b}_0 = (b_0^1, ..., b_0^{100})^T$, $\boldsymbol{b}_1 = (b_1^1, ..., b_1^{100})^T$ and $\boldsymbol{b}_2 = (b_2^1, ..., b_2^{100})^T$. In the last step of the simulation, data are generated based on continuous, binary, and time-to-event endpoints. With continuous endpoint, we have:

$$Y = \theta_0 + \theta_z Z + \theta_v V + \epsilon_y \ .$$

With binary endpoint, we have:

$$logit(P(Y = 1|Z, V)) = \theta_0 + \theta_z Z + \theta_v V.$$

With time-to-event endpoint, we have:

$$\lambda(t|Z, X, V, \boldsymbol{W}, Q) = \lambda(t|Z, V) = \lambda_0(t)\exp(\theta_z Z + \theta_v V),$$

where ϵ_x and ϵ_q are independently sampled from normal distributions with mean zero and standard deviations σ_x and σ_q. For the Cox model with time-to-event endpoint, censoring time is sampled from a mixture of $Unif(0, 10)$ and a point mass at 10 with equal probability and we set $\lambda_0(t) = 0.002t$. For the linear model with continuous endpoint, a random noise is drawn from a $N(0, 1.8)$. We fixed the sample size at $n_1 = 150$, $n_2 = 300$, $n_3 = 5150$. We simulate $\boldsymbol{b}_0$ randomly based on $Unif(4,5)$. Furthermore, we set $\boldsymbol{b}_2 = (1, .., 1)$, $\sigma_x = 0.2$, $\sigma_{x^*} = 0.5$, $\theta_z = 0.4$, $\theta_v = 0.6$, $\sigma_w = 1$. Then we change the values of $||\boldsymbol{b}_1||_2$, ρ, a_1, a_2 and σ_q to change the range of R^2. All types of R^2 are shown in Table 21 with respect to different patterns, effect sizes, and levels of the sparsity of X on $\boldsymbol{W}$ in this chapter. Three representative settings were selected with the sparsity of $\boldsymbol{W}$ equal to 2, 5, and 10. Under each size of sparsity, three different forms on the effect size of $\boldsymbol{W}$ are also generated and compared, including an equivalent effect size of X on $\boldsymbol{W}$, random pattern of effect size on $\boldsymbol{W}$ and decreasing effect size on $\boldsymbol{W}$. Below are the three settings selected for simulation in this chapter:

$||\boldsymbol{b}_1||_2 = 1.3$, $\rho = 0.6$, $a_0 = 4$, $a_1 = 1.5$, $\sigma_q = 3$ (setting 1);
$||\boldsymbol{b}_1||_2 = 1.1$, $\rho = 0$, $a_0 = 0.4$, $a_1 = 2$, $\sigma_q = 4$ (setting 2);
$||\boldsymbol{b}_1||_2 = 2$, $\rho = 0.6$, $a_0 = 4$, $a_1 = 1.5$, $\sigma_q = 3$ (setting 3);
where $||\boldsymbol{b}_1||_2$ shown in above settings is the total effect size of X on $\boldsymbol{W}$, that is, $||\boldsymbol{b}_1||_2 = \sqrt{\sum_{i=1}^{100} (b_1^i)^2}$. For example, if we assume the pattern of the effect of X on $\boldsymbol{W}$ is equally distributed with a sparse size of 5 using setting 1, then we have:

$$\boldsymbol{b}_1 = (1.1/\sqrt{5}, 1.1/\sqrt{5}, 1.1/\sqrt{5}, 1.1/\sqrt{5}, 1.1/\sqrt{5}, 0..., 0)^T.$$

Based on Table 21, almost all types of R^2 are similar to each other for different patterns, sparsity within each setting except for $R^2_{\hat{X}QV}$ and $R^2_{\hat{X}Q|V}$ under the random pattern of effect size. Specifically, the strength of FFQ on long term dietary intake given personal characteristics in stage 1 is controlled to be relatively low ($R^2_{ZQ|V}$=0.13) in setting 1 and

3 while increased to some extent with $R^2_{ZQ|V}$=0.19 in setting 2. The strength of the biomarker on consumed dietary intakes in stage 1 given personal characteristics generally follow an increasing trend from setting 1 to setting 3 with $R^2_{ZW|V}$ from 0.49 to 0.68.

The bias, mean estimated standard error (SE), empirical standard deviation (SD), and coverage rate (CR) of 95% nominal confidence interval for all four methods from 100 simulations are listed in Tables 22-33 with Lasso penalized regression. The performance of our proposed method with BF, Method 2, varied under different scenarios. Under sparsity of 2, 5, and 10, three patterns of effect size for W are generated, which includes equivalent, random, and decreasing patterns on the sized of effect. Post-selection or direct-selection are performed for variable selection. Within each type of selection, fixed and non-fixed personal characteristics (V) are applied. With weak strength on FFQ (Q) in setting 1 and 3, bias showed to some extent non-stabilized with Method 2. To be more specific, the bias is large especially when the equivalent effect size is taken under sparsity of 5 or 10 with Method 2. The direct-selection with non-fixed personal characteristics considered for filtering showed more stable results compared with direct-selection with fixed personal characteristics considered in Method 2. Three different approaches of variance estimation, shown as 2.1, 2.2, and 2.3 in all tables consequently, were applied for BF construction within Method 2 when variables are post selected under high-dimensional space. The third approach, named as RCV estimation under post-selection within Method 2, performs the best among all three approaches with the smallest bias and good CR as shown in tables under all scenarios for all types of endpoints. Under both post and direct-selections, there is an increasing trend on SD as sparsity increase for equivalent effect form when personal characteristics are not fixed for penalization. On the other hand, the change in SD is not following an obvious trend with personal characteristics fixed. Comparing different settings, we can see setting 1, where weak FFQ and biomarkers were applied, is generally showed the largest bias compared with setting 2 and 3 and SD of setting 1 is also the highest compared with the other two settings. Method 3 and 4 both provided promising and stable results in most cases. However, when the strength of biomarker is strong and strength of FFQ is relatively weak (i,e., setting 3), we can see Method 2 generally generated the most efficient result compared with Method 3 and 4, which is consistent with our summarized results found under low-dimensional data space in Chapter 2.

Table 34-45 displayed results for SCAD penalized regression. Compared with Lasso, a similar trend on bias control and SD can be found among different patterns of effect size and settings with SCAD. Focusing on the change of SD, we can see SD increased as sparsity increased under equivalent and random effect size with personal characteristics both fixed and non-fixed for variable filtering. Furthermore, comparing three approaches on variance estimation for BF construction of Method 2 when post-selection were applied, RCV still performs the best in controlling bias and gives the most efficient result. With direct-selection using SCAD in Method 2, the bias is generally well controlled with promising CR when personal characteristics are not fixed for variable filtering and the results are comparable with those gained with Lasso. On the other hand, when variables are post selected, the performance of SCAD is not as well as Lasso since the bias is not well controlled under many scenarios, especially when the sparsity is large under binary and time-to-event endpoints. Overall, the performance of Lasso is better in variance estimation and controlling bias compared with SCAD.

RF provides another way of building a biomarker prediction model in the second stage. Table 46-51 displayed results with RF. With the 10 most important variables selected directly with RF, the estimated bias is noticeably large in most cases. With post-selection on variables using RF, results with variance estimation 2.2 and 2.3 both showed small bias with promising CR. The results are comparable to those obtained with Lasso for Method 2 with RCV estimation (2.3). In general, RF did not provide a good estimation of the associated parameter with direct-selection and performed similarly as Lasso with post-selection.

Overall, Lasso generally provides a consistent estimator in most cases and largely attenuates the bias compared with SCAD and RF. The Lasso post-selection with RCV variance estimation for BF construction provides a consistent estimation of associated parameters with stable CR and is recommended especially when we have a sparse high-dimensional data structure.

4.4 Data Analysis

The estimated HR and corresponding 95% confidence interval according to a 20% increase in the sodium-to-potassium ratio are shown in Table 52 for Lasso, SCAD, and RF with WHI data. High-dimensional measurements in 24-hour NMR are used as W and analyzed in the WHI NPAAS feeding study. After excluding measurements with missing values exceeding the extent of 20%, 59 NMR measurements were included. Such measurements were normalized and log-transformed. Missing observations were also imputed with the median values for each corresponding NMR measurement. Log-transformed self-reported dietary intake from FFQ is set as Q. Variables related to personal characteristics, V, include age, BMI, race/ethnicity, educational level, self-reported physical activity, and smoking status. The disease outcomes applied are total CVD. Log transformed sodium-to-potassium ratio is again set as a single predictor in this section.

The proportion of variance that can be explained for each selected variable is of our interests and is further explored in the feeding study. Based on the ANOVA table, the calculated sodium-to-potassium ratio based on the existing biomarkers were found to account for the highest proportion of explained variance for short-term assessed sodium-to-potassium ratio compared to all other variables. To be more specific, 34% and 35% of variance were explained by the sodium-to-potassium ratio, with respect to Method 2 and 3 under post-lasso selection. Other selected NMR variables include acetone (7%), allantoin (3%), methyl-guanidine (3%) and methy-benzyl-alcohol (3%) for Method 2, and acetone (7%), methyl-guanidline (4%), methyl-benzyl-alcohol (3%), tyrosine (2%), self-reported sodium-to-potassium ratio (2%) and hippurate and trimethylamine-N-oxide (1%) for Method 3. Furthermore, we found Race is the only variable selected among individual characteristics and it accounts for 4% and 3% of explained variance for Method 2 and 3.

Based on Table 52, we can see the estimated HR is greater than 1 in all cases, indicating a higher risk of CVD with a higher level of sodium-to-potassium ratio regardless of different approaches in high-dimensional space, which is consistent with the results in chapter 2. The most conservative estimate on $\hat{\sigma}_x^{*2}$, 0, is utilized to construct the BF in Method 2. In addition, RCV variance estimation was adopted to construct the BF for the post-selection approach with Method 2. The estimation of the association parameter derived from Method 2 shrinks to a small scale compared with Method 1 and is comparable with Method 3 and Method 4. We noticed that the 95% CI does not embrace HR of 1 with Lasso and RF in most cases, indicating the significant association between calibrated dietary intake and risk of CVD. On the other hand, the 95% CI with SCAD showed less efficient results with larger variance indicating a non-significant association between calibrated dietary intake and risk of CVD in many cases.

4.5 Discussion

In this chapter, we examined the requirement for a valid biomarker for regression calibration purposes in high-dimensional space. Different methods to handle high-dimensional data (i.e., Lasso, SCAD, and RF), as well as different approaches on variable selections (i.e., direct and post-selection) are applied and compared across different scenarios such as sparse level, patterns of effect size. This chapter provides researchers a comprehensive picture dealing with high-dimensional data for calibrated regression study. The challenge of dealing with high-dimensional data when building linear regression models is the ease of overfitting to samples and not generic enough for estimation. Multicollinearity is also an issue in building linear regression models under high-dimensional space. To identify the most effective measurements associated with consumed dietary intakes in the feeding study, Lasso, SCAD, and RF in the high dimensional data set were applied for variable selection in this chapter. Overall, Lasso represents more stable results for variable selection compared to the other two approaches. Method 2 with BF constructed under RCV estimation of Lasso post-selection approach achieves our expectation with consistent good estimation in most cases. Generally speaking, various factors such as ways of obtaining tuning parameters and filtering conditions can affect the accuracy of the biomarker prediction model with penalized regression methods and RF. Depending on these choices, accuracy on the estimation of the association parameter can be substantially different.

Identifying the effective measurements associated with consumed dietary intakes is a very important step for biomarker construction. Statistical inference is a challenging issue with the penalized estimator. In this chapter, a bootstrapping approach was utilized for variance estimation in high-dimensional data for penalized regression and RF. There are other approaches for variance estimation under high-dimensional data with penalized regression we can consider in future analysis. One issue of the estimated covariance matrix is related to the zero components. Specifically, with coefficients equal to zero, the approximate covariance matrix produces zero for the estimated variance. The estimation on non-zero components is robust, however, the signs of zero-components may take negative or positive values. The same issue also exists with the sandwich formula of the covariance matrix developed by Fan & Lv (2008) [56]. A two-stage procedure was studied by Wasserman & Roeder (2009) [62] for validation. With their methods, the whole data was first randomly split into two parts, that is, a training dataset and a testing dataset. Then the penalized regression technique (i.e., Lasso) was applied for variable selection in the training data. Next, in the testing data, standard errors can be obtained based on ordinary least squares with variables selected from training data. One needs to pay attention to the way how to split the data with the single-split approach. Meinshausen et al. (2009) [63] extended the single-split to multi-split, where the same procedure with the single-split is repeating multiple times. Furthermore, Lockhart et al. (2014) [64] provided a test statistics to understand the significance of the variables selected by Lasso. For cases under ultra-high dimensional space where the sample size is equal or smaller than the variable dimension, sure independent screening (SIS) technique proposed by Fan & Lv (2008) [56] can be considered for variable screening in our future work. Variance estimation under high-dimensional data is still an open question and needs to be further considered.

TABLE 21: List of R^2 for the three settings under different patterns and sparsity

		Setting 1			Setting 2			Setting 3		
Pattern	Type of R^2	S2	S5	S10	S2	S5	S10	S2	S5	S10
Same	R^2_{ZWV}	0.68	0.68	0.67	0.52	0.51	0.50	0.80	0.79	0.79
	$R^2_{ZW\|V}$	0.49	0.48	0.47	0.52	0.51	0.50	0.67	0.67	0.66
	R^2_{ZQV}	0.46	0.46	0.46	0.19	0.19	0.19	0.46	0.46	0.46
	$R^2_{ZQ\|V}$	0.13	0.13	0.13	0.19	0.19	0.19	0.13	0.13	0.13
	R^2_{ZWQV}	0.70	0.70	0.70	0.57	0.56	0.56	0.81	0.80	0.80
	$R^2_{ZWQ\|V}$	0.52	0.51	0.51	0.57	0.56	0.56	0.69	0.68	0.68
	$R^2_{X^*WV}$	0.56	0.55	0.55	0.44	0.43	0.42	0.66	0.66	0.65
	$R^2_{X^*W\|V}$	0.37	0.37	0.36	0.44	0.43	0.42	0.52	0.51	0.50
	$R^2_{X^*WQV}$	0.57	0.57	0.56	0.47	0.46	0.46	0.66	0.66	0.66
	$R^2_{X^*WQ\|V}$	0.40	0.39	0.38	0.47	0.46	0.46	0.52	0.52	0.51
	$R^2_{\hat{X}QV}$	0.50	0.51	0.52	0.03	0.07	0.07	0.46	0.47	0.48
	$R^2_{\hat{X}Q\|V}$	0.05	0.06	0.06	0.03	0.07	0.07	0.07	0.08	0.08
Random	R^2_{ZWV}	0.68	0.68	0.69	0.52	0.53	0.53	0.80	0.80	0.80
	$R^2_{ZW\|V}$	0.49	0.49	0.49	0.52	0.53	0.53	0.67	0.68	0.68
	R^2_{ZQV}	0.46	0.46	0.46	0.19	0.19	0.19	0.46	0.46	0.46
	$R^2_{ZQ\|V}$	0.13	0.13	0.13	0.19	0.19	0.19	0.13	0.13	0.13
	R^2_{ZWQV}	0.71	0.71	0.71	0.57	0.58	0.58	0.81	0.81	0.81
	$R^2_{ZWQ\|V}$	0.52	0.53	0.53	0.57	0.58	0.58	0.69	0.69	0.69
	$R^2_{X^*WV}$	0.56	0.56	0.56	0.44	0.44	0.44	0.66	0.66	0.66
	$R^2_{X^*W\|V}$	0.37	0.37	0.37	0.44	0.44	0.44	0.52	0.52	0.52
	$R^2_{X^*WQV}$	0.57	0.57	0.57	0.47	0.47	0.47	0.66	0.66	0.66
	$R^2_{X^*WQ\|V}$	0.40	0.40	0.40	0.47	0.47	0.47	0.53	0.52	0.53
	$R^2_{\hat{X}QV}$	0.48	0.29	0.23	0.00	0.01	0.02	0.43	0.14	0.07
	$R^2_{\hat{X}Q\|V}$	0.04	0.00	0.00	0.00	0.01	0.02	0.05	0.00	0.01
Decreasing	R^2_{ZWV}	0.68	0.68	0.68	0.52	0.52	0.51	0.80	0.80	0.79
	$R^2_{ZW\|V}$	0.49	0.48	0.48	0.52	0.52	0.51	0.67	0.67	0.67
	R^2_{ZQV}	0.46	0.46	0.46	0.19	0.19	0.19	0.46	0.46	0.46
	$R^2_{ZQ\|V}$	0.13	0.13	0.13	0.19	0.19	0.19	0.13	0.13	0.13
	R^2_{ZWQV}	0.71	0.70	0.70	0.57	0.57	0.57	0.81	0.81	0.80
	$R^2_{ZWQ\|V}$	0.52	0.52	0.52	0.57	0.57	0.57	0.69	0.69	0.68
	$R^2_{X^*WV}$	0.56	0.56	0.55	0.44	0.43	0.43	0.66	0.66	0.65
	$R^2_{X^*W\|V}$	0.37	0.37	0.37	0.44	0.43	0.43	0.52	0.51	0.51
	$R^2_{X^*WQV}$	0.57	0.57	0.57	0.47	0.47	0.46	0.66	0.66	0.66
	$R^2_{X^*WQ\|V}$	0.40	0.39	0.39	0.47	0.47	0.46	0.53	0.52	0.52
	$R^2_{\hat{X}QV}$	0.50	0.51	0.51	0.02	0.03	0.03	0.46	0.47	0.47
	$R^2_{\hat{X}Q\|V}$	0.05	0.06	0.06	0.02	0.03	0.03	0.07	0.08	0.07

Table 22: Simulation results with direct-Lasso selection for non-fixed personal characteristics under continuous endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.75	0.530	0.536	0.75	0.60	0.375	0.301	0.52	0.315	0.263	0.264	0.82
		2	0.07	0.209	0.220	0.92	0.04	0.147	0.133	0.94	0.063	0.171	0.166	0.90
		3	0.01	0.156	0.163	0.89	0.02	0.119	0.107	0.91	0.021	0.153	0.161	0.92
		4	0.04	0.217	0.197	0.90	0.02	0.121	0.114	0.90	0.042	0.217	0.197	0.90
	5	1	1.22	2.932	0.735	0.78	0.76	0.810	0.405	0.66	0.374	0.357	0.297	0.83
		2	0.07	0.504	0.253	0.94	0.02	0.186	0.151	0.93	0.049	0.191	0.171	0.93
		3	0.01	0.161	0.172	0.92	0.01	0.122	0.109	0.93	0.018	0.167	0.173	0.92
		4	0.04	0.217	0.197	0.90	0.02	0.121	0.114	0.90	0.042	0.217	0.197	0.90
	10	1	1.31	1.644	1.228	0.85	0.73	4.259	6.772	0.88	0.439	0.570	3.076	0.91
		2	-0.03	0.202	0.424	0.94	-0.04	0.559	2.234	0.93	0.054	0.911	0.535	0.95
		3	0.01	0.166	0.186	0.93	0.01	0.130	0.128	0.92	0.028	0.199	0.263	0.94
		4	0.04	0.217	0.197	0.90	0.02	0.121	0.114	0.90	0.042	0.217	0.197	0.90
Random	2	1	0.61	0.495	0.437	0.76	0.55	0.319	0.277	0.56	0.255	0.248	0.224	0.85
		2	0.03	0.255	0.185	0.91	0.05	0.155	0.125	0.91	0.035	0.182	0.146	0.90
		3	-0.01	0.185	0.132	0.89	0.01	0.138	0.099	0.91	-0.005	0.186	0.130	0.89
		4	0.05	0.252	0.175	0.88	0.02	0.127	0.107	0.86	0.052	0.252	0.175	0.88
	5	1	0.67	0.514	0.507	0.73	0.59	0.358	0.302	0.56	0.273	0.240	0.240	0.80
		2	0.02	0.187	0.213	0.95	0.05	0.140	0.132	0.94	0.037	0.164	0.156	0.93
		3	-0.01	0.159	0.148	0.87	0.01	0.116	0.101	0.88	-0.007	0.154	0.143	0.87
		4	0.05	0.241	0.181	0.89	0.02	0.125	0.108	0.93	0.052	0.241	0.181	0.89
	10	1	0.80	0.577	0.911	0.69	0.58	0.618	0.631	0.87	0.276	0.362	0.384	0.89
		2	0.02	0.204	0.343	0.91	0.05	0.235	0.350	0.91	0.031	0.192	0.330	0.91
		3	-0.01	0.173	0.154	0.88	0.03	0.153	0.120	0.92	0.044	0.227	0.196	0.93
		4	0.05	0.232	0.224	0.92	0.02	0.124	0.116	0.92	0.050	0.232	0.224	0.92
Decreasing	2	1	0.74	0.479	0.497	0.67	0.61	0.346	0.286	0.47	0.320	0.254	0.255	0.77
		2	0.07	0.207	0.213	0.93	0.05	0.141	0.129	0.92	0.063	0.168	0.162	0.92
		3	0.01	0.155	0.164	0.91	0.02	0.113	0.108	0.92	0.022	0.157	0.158	0.92
		4	0.04	0.217	0.197	0.90	0.02	0.121	0.114	0.90	0.042	0.217	0.197	0.90
	5	1	0.82	0.513	0.584	0.71	0.66	0.385	0.332	0.51	0.348	0.266	0.271	0.78
		2	0.06	0.205	0.242	0.95	0.04	0.148	0.136	0.91	0.053	0.166	0.164	0.93
		3	0.01	0.158	0.171	0.91	0.01	0.113	0.110	0.92	0.020	0.160	0.162	0.92
		4	0.04	0.217	0.197	0.90	0.02	0.121	0.114	0.90	0.042	0.217	0.197	0.90
	10	1	0.86	0.534	0.656	0.69	0.64	0.594	0.728	0.81	0.251	0.257	0.509	0.94
		2	0.06	0.210	0.224	0.94	0.09	0.322	0.242	0.91	0.145	1.152	0.229	0.89
		3	0.01	0.160	0.289	0.91	0.02	0.133	0.163	0.93	0.044	0.188	0.238	0.93
		4	0.04	0.217	0.197	0.90	0.02	0.121	0.114	0.90	0.042	0.217	0.197	0.90

Table 23: Simulation results with direct-Lasso selection for fixed personal characteristics under continuous endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.80	0.596	0.647	0.75	0.61	0.369	0.294	0.46	0.34	0.288	0.259	0.79
		2	0.08	0.224	0.331	0.91	0.05	0.145	0.129	0.89	0.07	0.176	0.162	0.91
		3	0.03	0.175	0.163	0.89	0.01	0.116	0.099	0.89	0.03	0.164	0.151	0.89
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	5	1	0.76	2.018	0.983	0.79	0.73	0.637	0.425	0.63	0.40	0.410	0.305	0.80
		2	0.64	6.101	0.282	0.94	0.02	0.160	0.144	0.94	0.05	0.184	0.176	0.95
		3	0.02	0.186	0.174	0.88	0.01	0.124	0.104	0.89	0.02	0.188	0.155	0.90
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	10	1	67.19	659.071	78.369	0.88	0.98	1.159	0.634	0.65	0.46	0.441	0.433	0.80
		2	-2.36	23.215	10.844	0.97	-0.03	0.142	0.177	0.97	0.01	0.157	0.227	0.96
		3	0.02	0.183	0.188	0.89	0.01	0.118	0.110	0.89	0.02	0.177	0.167	0.89
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
Random	2	1	0.73	0.873	0.494	0.76	0.56	0.319	0.282	0.57	0.30	0.296	0.246	0.79
		2	0.09	0.359	0.203	0.91	0.05	0.160	0.127	0.92	0.06	0.216	0.157	0.91
		3	0.29	2.707	0.172	0.89	0.01	0.150	0.101	0.91	0.47	4.596	0.159	0.89
		4	0.05	0.252	0.245	0.89	0.02	0.127	0.111	0.92	0.05	0.252	0.245	0.89
	5	1	0.74	0.570	0.502	0.73	0.58	0.341	0.292	0.60	0.30	0.254	0.238	0.78
		2	0.06	0.188	0.200	0.91	0.05	0.130	0.130	0.94	0.05	0.164	0.153	0.91
		3	0.03	0.190	0.178	0.89	0.01	0.114	0.108	0.90	0.02	0.180	0.163	0.90
		4	0.05	0.241	0.207	0.86	0.02	0.125	0.119	0.92	0.05	0.241	0.207	0.86
	10	1	0.88	0.629	0.587	0.67	0.67	0.380	0.330	0.48	0.35	0.289	0.256	0.72
		2	0.06	0.223	0.214	0.93	0.05	0.145	0.135	0.92	0.06	0.183	0.157	0.89
		3	0.04	0.215	0.175	0.91	0.02	0.120	0.103	0.91	0.04	0.201	0.156	0.91
		4	0.05	0.232	0.195	0.92	0.02	0.124	0.109	0.88	0.05	0.232	0.195	0.92
Decreasing	2	1	0.80	0.534	0.511	0.70	0.61	0.366	0.289	0.49	0.34	0.281	0.254	0.78
		2	0.09	0.240	0.203	0.90	0.06	0.140	0.129	0.86	0.07	0.181	0.159	0.90
		3	0.03	0.175	0.168	0.90	0.01	0.111	0.100	0.94	0.03	0.167	0.152	0.89
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	5	1	0.90	0.589	0.603	0.73	0.68	0.407	0.329	0.49	0.37	0.285	0.270	0.77
		2	0.08	0.241	0.214	0.91	0.05	0.146	0.134	0.87	0.07	0.183	0.162	0.91
		3	0.03	0.176	0.170	0.90	0.01	0.111	0.102	0.92	0.03	0.168	0.153	0.89
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	10	1	0.96	0.645	0.650	0.71	0.74	0.454	0.344	0.44	0.39	0.305	0.279	0.77
		2	0.07	0.251	0.225	0.90	0.05	0.148	0.137	0.89	0.06	0.185	0.164	0.91
		3	0.03	0.177	0.172	0.88	0.01	0.110	0.102	0.93	0.03	0.168	0.155	0.90
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90

Table 24: Simulation results with post-Lasso selection for non-fixed personal characteristics under continuous endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.52	0.486	0.471	0.86	0.41	0.282	0.229	0.57	0.21	0.270	0.214	0.85
		2.1	-0.20	0.192	0.194	0.73	-0.08	0.123	0.117	0.86	-0.08	0.136	0.134	0.84
		2.2	0.22	0.327	0.317	0.92	0.11	0.169	0.142	0.86	0.11	0.221	0.183	0.86
		2.3	0.05	0.229	0.241	0.92	0.01	0.152	0.129	0.89	0.03	0.196	0.156	0.88
		3	0.02	0.171	0.181	0.91	0.02	0.131	0.105	0.90	0.02	0.168	0.167	0.88
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	5	1	0.47	0.746	0.558	0.87	0.53	0.577	0.302	0.66	0.23	0.346	0.232	0.83
		2.1	-0.22	0.194	0.217	0.74	-0.15	0.128	0.136	0.73	-0.08	0.153	0.140	0.79
		2.2	0.16	0.417	0.345	0.92	0.16	0.303	0.172	0.88	0.10	0.249	0.190	0.91
		2.3	-0.03	0.229	0.240	0.93	-0.04	0.159	0.150	0.94	0.00	0.184	0.159	0.92
		3	0.02	0.197	0.189	0.90	0.01	0.139	0.110	0.95	0.04	0.313	0.165	0.91
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	10	1	0.70	0.921	2.072	0.88	0.82	1.161	0.549	0.70	0.33	0.397	0.342	0.86
		2.1	-0.29	0.319	0.316	0.73	-0.28	0.254	0.204	0.70	-0.13	0.135	0.175	0.80
		2.2	0.21	0.423	1.124	0.95	0.26	0.502	0.284	0.87	0.15	0.254	0.257	0.93
		2.3	-0.06	0.222	0.816	0.98	-0.09	0.163	0.206	0.92	-0.01	0.199	0.190	0.89
		3	0.04	0.208	0.228	0.94	0.01	0.130	0.120	0.91	0.04	0.185	0.215	0.92
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
Random	2	1	0.50	0.517	0.440	0.83	0.39	0.292	0.226	0.65	0.19	0.237	0.213	0.86
		2.1	-0.12	0.208	0.173	0.77	-0.10	0.136	0.118	0.74	-0.06	0.160	0.128	0.80
		2.2	0.17	0.343	0.264	0.88	0.12	0.206	0.146	0.85	0.09	0.201	0.172	0.90
		2.3	0.05	0.309	0.215	0.91	0.02	0.189	0.127	0.91	0.02	0.191	0.152	0.92
		3	0.04	0.387	0.187	0.93	0.02	0.159	0.111	0.87	0.04	0.319	0.166	0.91
		4	0.05	0.252	0.245	0.89	0.02	0.127	0.111	0.92	0.05	0.252	0.245	0.89
	5	1	0.47	0.386	0.439	0.85	0.38	0.245	0.236	0.70	0.19	0.212	0.206	0.86
		2.1	-0.16	0.186	0.182	0.77	-0.13	0.124	0.122	0.76	-0.07	0.142	0.129	0.84
		2.2	0.14	0.241	0.269	0.93	0.11	0.154	0.156	0.89	0.08	0.180	0.167	0.91
		2.3	0.00	0.209	0.224	0.90	0.00	0.135	0.131	0.94	0.01	0.164	0.148	0.91
		3	0.02	0.195	0.314	0.89	0.01	0.129	0.116	0.92	0.03	0.190	0.184	0.90
		4	0.05	0.241	0.207	0.86	0.02	0.125	0.119	0.92	0.05	0.241	0.207	0.86
	10	1	0.57	0.519	0.451	0.78	0.43	0.294	0.254	0.62	0.23	0.254	0.215	0.85
		2.1	-0.18	0.168	0.184	0.74	-0.14	0.133	0.132	0.75	-0.07	0.141	0.136	0.83
		2.2	0.17	0.300	0.267	0.89	0.13	0.174	0.164	0.85	0.10	0.202	0.174	0.90
		2.3	0.02	0.249	0.220	0.93	0.01	0.150	0.140	0.93	0.03	0.176	0.153	0.90
		3	0.04	0.272	0.204	0.91	0.02	0.140	0.112	0.92	0.04	0.246	0.170	0.90
		4	0.05	0.232	0.195	0.92	0.02	0.124	0.109	0.88	0.05	0.232	0.195	0.92
Decreasing	2	1	0.54	0.570	0.461	0.84	0.43	0.269	0.232	0.52	0.21	0.231	0.215	0.86
		2.1	-0.19	0.197	0.190	0.78	-0.09	0.113	0.117	0.83	-0.08	0.139	0.133	0.81
		2.2	0.24	0.399	0.310	0.91	0.12	0.164	0.145	0.86	0.11	0.199	0.184	0.87
		2.3	0.05	0.266	0.223	0.91	0.00	0.128	0.129	0.89	0.02	0.164	0.154	0.90
		3	0.02	0.180	0.179	0.93	0.02	0.128	0.106	0.92	0.02	0.173	0.199	0.89
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	5	1	0.62	0.814	0.581	0.83	0.48	0.306	0.255	0.55	0.23	0.246	0.223	0.85
		2.1	-0.22	0.222	0.208	0.72	-0.11	0.128	0.123	0.76	-0.09	0.146	0.135	0.82
		2.2	0.27	0.571	0.372	0.92	0.13	0.175	0.155	0.84	0.12	0.205	0.188	0.88
		2.3	0.04	0.377	0.259	0.91	0.00	0.139	0.135	0.89	0.01	0.161	0.157	0.92
		3	0.03	0.176	0.182	0.90	0.02	0.122	0.107	0.91	0.02	0.166	0.195	0.91
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	10	1	0.61	0.632	0.619	0.81	0.51	0.324	0.266	0.53	0.24	0.241	0.233	0.81
		2.1	-0.24	0.217	0.218	0.73	-0.12	0.126	0.126	0.80	-0.10	0.142	0.138	0.80
		2.2	0.26	0.439	0.388	0.91	0.14	0.185	0.159	0.86	0.12	0.195	0.193	0.90
		2.3	0.03	0.279	0.271	0.90	0.00	0.138	0.138	0.89	0.01	0.163	0.160	0.91
		3	0.03	0.182	0.189	0.89	0.02	0.118	0.107	0.92	0.02	0.170	0.195	0.90
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90

Table 25: Simulation results with post-Lasso selection for fixed personal characteristics under continuous endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.55	0.657	0.533	0.86	0.43	0.350	0.252	0.60	0.22	0.249	0.223	0.89
		2.1	0.22	0.327	0.317	0.92	0.11	0.169	0.142	0.86	0.11	0.221	0.183	0.86
		2.2	0.24	0.401	0.302	0.87	0.15	0.213	0.159	0.85	0.12	0.211	0.186	0.90
		2.3	0.06	0.280	0.271	0.92	0.03	0.163	0.139	0.91	0.04	0.177	0.161	0.87
		3	0.04	0.194	0.292	0.93	0.02	0.134	0.119	0.89	0.04	0.197	0.200	0.93
		4	0.04	0.217	0.197	0.90	0.02	0.121	0.114	0.90	0.04	0.217	0.197	0.90
	5	1	0.50	0.721	0.774	0.82	0.48	0.512	0.293	0.74	0.24	0.296	0.249	0.86
		2.1	0.16	0.417	0.345	0.92	0.16	0.303	0.172	0.88	0.10	0.249	0.190	0.91
		2.2	0.18	0.416	0.888	0.91	0.15	0.278	0.175	0.86	0.11	0.220	0.196	0.91
		2.3	-0.04	0.223	0.306	0.89	-0.03	0.153	0.149	0.91	0.00	0.169	0.167	0.92
		3	0.03	0.198	0.283	0.94	0.01	0.149	0.119	0.93	0.03	0.216	0.202	0.93
		4	0.04	0.217	0.197	0.90	0.02	0.121	0.114	0.90	0.04	0.217	0.197	0.90
	10	1	0.65	1.111	1.306	0.89	0.71	1.050	0.447	0.71	0.32	0.374	0.310	0.87
		2.1	0.21	0.423	1.124	0.95	0.26	0.502	0.284	0.87	0.15	0.254	0.257	0.93
		2.2	0.21	0.515	0.620	0.93	0.21	0.483	0.225	0.85	0.13	0.241	0.239	0.92
		2.3	-0.10	0.176	0.439	0.92	-0.07	0.174	0.187	0.89	-0.03	0.152	0.185	0.90
		3	0.02	0.208	0.272	0.91	0.01	0.142	0.122	0.92	0.02	0.179	0.215	0.95
		4	0.04	0.217	0.197	0.90	0.02	0.121	0.114	0.90	0.04	0.217	0.197	0.90
Random	2	1	0.58	1.062	0.402	0.81	0.39	0.307	0.230	0.68	0.23	0.411	0.202	0.84
		2.1	0.17	0.343	0.264	0.88	0.12	0.206	0.146	0.85	0.09	0.201	0.172	0.90
		2.2	0.28	0.835	0.292	0.86	0.13	0.224	0.156	0.85	0.13	0.389	0.177	0.86
		2.3	0.11	0.675	0.206	0.87	0.02	0.185	0.129	0.90	0.05	0.340	0.147	0.86
		3	-0.03	0.537	0.183	0.91	0.03	0.220	0.108	0.92	-0.01	0.370	0.174	0.92
		4	0.05	0.252	0.175	0.88	0.02	0.127	0.107	0.86	0.05	0.252	0.175	0.88
	5	1	0.46	0.417	0.725	0.80	0.38	0.245	0.246	0.71	0.19	0.221	0.219	0.84
		2.1	0.14	0.241	0.269	0.93	0.11	0.154	0.156	0.89	0.08	0.180	0.167	0.91
		2.2	0.17	0.251	0.294	0.93	0.12	0.158	0.163	0.90	0.10	0.191	0.179	0.91
		2.3	0.02	0.205	0.319	0.93	0.01	0.130	0.135	0.95	0.02	0.169	0.159	0.93
		3	0.03	0.194	0.256	0.89	0.01	0.114	0.113	0.91	0.03	0.204	0.193	0.92
		4	0.05	0.241	0.181	0.89	0.02	0.125	0.108	0.93	0.05	0.241	0.181	0.89
	10	1	0.55	0.528	0.990	0.78	0.43	0.275	0.262	0.69	0.22	0.243	0.215	0.83
		2.1	0.17	0.300	0.267	0.89	0.13	0.174	0.164	0.85	0.10	0.202	0.174	0.90
		2.2	0.22	0.334	0.297	0.88	0.13	0.182	0.166	0.87	0.11	0.206	0.180	0.87
		2.3	0.03	0.241	0.423	0.89	0.01	0.160	0.138	0.96	0.03	0.183	0.149	0.87
		3	0.05	0.252	0.273	0.92	0.02	0.139	0.116	0.91	0.05	0.228	0.189	0.92
		4	0.05	0.232	0.224	0.92	0.02	0.124	0.116	0.92	0.05	0.232	0.224	0.92
Decreasing	2	1	0.56	0.522	0.613	0.78	0.43	0.298	0.241	0.59	0.25	0.258	0.241	0.87
		2.1	0.24	0.399	0.310	0.91	0.12	0.164	0.145	0.86	0.11	0.199	0.184	0.87
		2.2	0.26	0.372	0.382	0.86	0.15	0.190	0.156	0.81	0.12	0.212	0.182	0.87
		2.3	0.06	0.240	0.308	0.93	0.02	0.142	0.135	0.92	0.02	0.164	0.164	0.91
		3	0.04	0.205	0.260	0.93	0.02	0.127	0.121	0.94	0.04	0.182	0.212	0.94
		4	0.04	0.217	0.197	0.90	0.02	0.121	0.114	0.90	0.04	0.217	0.197	0.90
	5	1	0.61	0.634	0.510	0.77	0.46	0.327	0.263	0.65	0.24	0.259	0.236	0.88
		2.1	0.27	0.571	0.372	0.92	0.13	0.175	0.155	0.84	0.12	0.205	0.188	0.88
		2.2	0.26	0.424	0.371	0.90	0.15	0.192	0.168	0.86	0.13	0.220	0.186	0.89
		2.3	0.03	0.222	0.243	0.93	0.00	0.138	0.141	0.90	0.02	0.166	0.163	0.92
		3	0.03	0.180	0.274	0.96	0.02	0.121	0.125	0.95	0.04	0.187	0.210	0.93
		4	0.04	0.217	0.197	0.90	0.02	0.121	0.114	0.90	0.04	0.217	0.197	0.90
	10	1	0.64	0.597	0.704	0.76	0.49	0.357	0.278	0.63	0.25	0.258	0.241	0.87
		2.1	0.26	0.439	0.388	0.91	0.14	0.185	0.159	0.86	0.12	0.195	0.193	0.90
		2.2	0.26	0.367	0.362	0.90	0.16	0.206	0.174	0.87	0.13	0.218	0.188	0.90
		2.3	0.03	0.216	0.321	0.94	0.00	0.132	0.142	0.94	0.02	0.164	0.164	0.91
		3	0.03	0.178	0.304	0.96	0.01	0.116	0.124	0.94	0.04	0.182	0.212	0.94
		4	0.04	0.217	0.197	0.90	0.02	0.121	0.114	0.90	0.04	0.217	0.197	0.90

Table 26: Simulation results for direct-Lasso selection with non-fixed personal characteristics under binary endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.74	0.557	0.688	0.83	0.56	0.314	0.329	0.68	0.32	0.305	0.318	0.88
		2	0.08	0.248	0.292	0.94	0.04	0.139	0.150	0.97	0.07	0.203	0.207	0.93
		3	0.01	0.192	0.189	0.90	0.00	0.115	0.121	0.93	0.02	0.192	0.187	0.9
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	5	1	1.12	2.191	0.954	0.85	0.70	0.630	0.451	0.74	0.38	0.391	0.365	0.87
		2	0.07	0.401	0.320	0.94	0.01	0.146	0.162	0.97	0.05	0.211	0.216	0.93
		3	0.01	0.208	0.195	0.91	0.00	0.117	0.125	0.93	0.02	0.208	0.193	0.93
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	10	1	1.40	1.536	2.393	0.85	1.00	1.542	3.564	0.86	0.45	0.449	2.111	0.84
		2	-0.01	0.222	0.913	0.93	-0.06	0.147	0.204	0.97	0.02	0.196	0.236	0.95
		3	0.01	0.202	0.207	0.92	0.00	0.114	0.133	0.97	0.02	0.202	0.208	0.94
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
Random	2	1	0.63	0.473	0.530	0.87	0.53	0.338	0.318	0.67	0.27	0.248	0.296	0.9
		2	0.03	0.213	0.224	0.96	0.04	0.143	0.146	0.95	0.04	0.169	0.193	0.98
		3	-0.01	0.149	0.183	0.96	0.00	0.110	0.126	0.95	0.00	0.154	0.181	0.94
		4	0.05	0.208	0.259	0.93	0.00	0.097	0.133	0.96	0.05	0.208	0.259	0.93
	5	1	0.60	0.672	0.565	0.89	0.53	0.317	0.334	0.71	0.23	0.276	0.289	0.94
		2	-0.02	0.183	0.224	0.94	0.03	0.137	0.147	0.92	0.00	0.178	0.184	0.94
		3	-0.04	0.178	0.170	0.89	-0.01	0.123	0.124	0.92	-0.03	0.173	0.168	0.9
		4	0.02	0.247	0.255	0.92	0.00	0.124	0.136	0.93	0.02	0.247	0.255	0.92
	10	1	0.84	0.658	0.691	0.74	0.65	0.367	0.378	0.62	0.33	0.295	0.317	0.8
		2	0.04	0.221	0.265	0.91	0.04	0.148	0.158	0.96	0.05	0.186	0.198	0.94
		3	0.01	0.192	0.184	0.92	0.01	0.147	0.126	0.89	0.01	0.181	0.178	0.93
		4	0.06	0.265	0.245	0.92	0.01	0.144	0.132	0.89	0.06	0.265	0.245	0.92
Decreasing	2	1	0.73	0.529	0.579	0.82	0.57	0.313	0.321	0.58	0.31	0.292	0.313	0.86
		2	0.08	0.233	0.259	0.94	0.03	0.138	0.148	0.96	0.06	0.200	0.204	0.93
		3	0.01	0.200	0.192	0.90	0.01	0.118	0.122	0.91	0.02	0.200	0.187	0.91
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	5	1	0.81	0.595	0.699	0.81	0.65	0.404	0.359	0.58	0.34	0.317	0.334	0.87
		2	0.07	0.244	0.364	0.93	0.03	0.147	0.154	0.95	0.06	0.202	0.209	0.93
		3	0.01	0.208	0.193	0.91	0.00	0.119	0.123	0.92	0.02	0.208	0.190	0.91
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	10	1	0.86	0.638	0.826	0.79	0.69	0.390	0.381	0.59	0.36	0.332	0.349	0.83
		2	0.06	0.250	0.276	0.92	0.03	0.150	0.157	0.96	0.06	0.207	0.211	0.94
		3	0.01	0.212	0.194	0.91	0.00	0.119	0.124	0.92	0.03	0.212	0.192	0.9
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91

TABLE 27: Simulation results for direct-Lasso selection with fixed personal characteristics under binary endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.78	0.603	0.719	0.85	0.57	0.317	0.326	0.69	0.33	0.312	0.325	0.86
		2	0.08	0.247	0.358	0.91	0.04	0.138	0.148	0.96	0.07	0.205	0.208	0.92
		3	0.03	0.202	0.206	0.93	0.00	0.111	0.120	0.96	0.03	0.195	0.193	0.90
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	5	1	0.89	1.699	1.013	0.85	0.69	0.509	0.446	0.73	0.39	0.400	0.370	0.80
		2	0.01	0.418	0.312	0.93	0.01	0.142	0.158	0.97	0.05	0.208	0.214	0.94
		3	0.03	0.216	0.213	0.93	0.00	0.115	0.124	0.94	0.03	0.220	0.197	0.94
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	10	1	43.64	422.313	59.716	0.90	0.94	0.987	0.662	0.75	0.46	0.437	0.502	0.84
		2	-1.19	11.572	9.610	0.95	-0.03	0.139	0.189	0.99	0.02	0.188	0.259	0.96
		3	0.02	0.216	0.231	0.93	0.00	0.114	0.131	0.95	0.02	0.209	0.211	0.94
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
Random	2	1	0.73	0.629	0.564	0.82	0.54	0.349	0.323	0.70	0.30	0.286	0.310	0.88
		2	0.09	0.351	0.237	0.97	0.04	0.138	0.147	0.95	0.06	0.192	0.202	0.97
		3	0.41	3.930	0.217	0.97	0.00	0.117	0.126	0.95	0.34	3.197	0.208	0.96
		4	0.05	0.208	0.259	0.93	0.00	0.097	0.133	0.96	0.05	0.208	0.259	0.93
	5	1	0.65	0.587	0.595	0.87	0.53	0.312	0.338	0.70	0.25	0.290	0.302	0.93
		2	0.02	0.197	0.232	0.93	0.02	0.134	0.149	0.92	0.02	0.186	0.190	0.94
		3	0.00	0.205	0.204	0.94	-0.01	0.119	0.125	0.92	0.00	0.192	0.189	0.93
		4	0.02	0.247	0.255	0.92	0.00	0.124	0.136	0.93	0.02	0.247	0.255	0.92
	10	1	0.87	0.699	0.754	0.75	0.66	0.418	0.377	0.62	0.36	0.312	0.339	0.80
		2	0.07	0.224	0.267	0.93	0.04	0.147	0.156	0.95	0.07	0.189	0.206	0.95
		3	0.05	0.243	0.223	0.92	0.01	0.144	0.125	0.87	0.04	0.221	0.202	0.92
		4	0.06	0.265	0.245	0.92	0.01	0.144	0.132	0.89	0.06	0.265	0.245	0.92
Decreasing	2	1	0.78	0.551	0.609	0.81	0.57	0.318	0.319	0.61	0.33	0.306	0.320	0.86
		2	0.08	0.234	0.247	0.94	0.04	0.136	0.146	0.97	0.07	0.200	0.204	0.93
		3	0.03	0.215	0.213	0.93	0.00	0.113	0.121	0.93	0.03	0.208	0.196	0.91
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	5	1	0.89	0.670	0.721	0.81	0.64	0.368	0.361	0.65	0.36	0.329	0.341	0.84
		2	0.07	0.246	0.261	0.93	0.04	0.145	0.152	0.95	0.06	0.204	0.207	0.93
		3	0.03	0.223	0.216	0.93	0.00	0.116	0.122	0.94	0.03	0.221	0.198	0.92
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	10	1	0.94	0.697	0.775	0.80	0.68	0.383	0.377	0.62	0.38	0.346	0.352	0.83
		2	0.07	0.250	0.270	0.94	0.04	0.148	0.154	0.95	0.06	0.207	0.209	0.93
		3	0.03	0.225	0.219	0.93	0.00	0.116	0.123	0.94	0.03	0.219	0.201	0.92
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91

Table 28: Simulation results for post-Lasso selection with non-fixed personal characteristics under binary endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.51	0.492	0.548	0.90	0.38	0.245	0.258	0.72	0.20	0.271	0.273	0.89
		2.1	-0.19	0.210	0.213	0.76	-0.09	0.116	0.127	0.86	-0.08	0.167	0.164	0.87
		2.2	0.22	0.343	0.371	0.97	0.09	0.153	0.164	0.91	0.11	0.230	0.231	0.93
		2.3	0.05	0.256	0.281	0.95	-0.01	0.130	0.142	0.96	0.03	0.197	0.198	0.92
		3	0.03	0.218	0.224	0.91	0.01	0.126	0.125	0.93	0.03	0.199	0.214	0.90
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	5	1	0.51	0.722	0.637	0.89	0.49	0.439	0.333	0.76	0.23	0.317	0.287	0.92
		2.1	-0.22	0.227	0.233	0.77	-0.16	0.114	0.143	0.81	-0.08	0.172	0.167	0.83
		2.2	0.18	0.418	0.394	0.94	0.13	0.227	0.193	0.96	0.10	0.242	0.233	0.93
		2.3	-0.02	0.266	0.272	0.94	-0.06	0.132	0.159	0.96	-0.01	0.191	0.194	0.92
		3	0.03	0.224	0.236	0.89	0.00	0.127	0.129	0.95	0.04	0.262	0.205	0.95
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	10	1	0.77	0.995	2.192	0.90	0.77	0.891	0.586	0.80	0.34	0.381	0.405	0.89
		2.1	-0.31	0.254	0.333	0.76	-0.28	0.222	0.207	0.74	-0.13	0.157	0.197	0.80
		2.2	0.23	0.421	1.223	0.94	0.23	0.383	0.311	0.94	0.15	0.262	0.305	0.92
		2.3	-0.07	0.208	0.909	0.98	-0.09	0.152	0.214	0.96	-0.02	0.184	0.224	0.95
		3	0.05	0.235	0.276	0.92	0.00	0.129	0.140	0.95	0.04	0.226	0.235	0.94
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
Random	2	1	0.50	0.495	0.494	0.89	0.38	0.285	0.262	0.78	0.19	0.229	0.268	0.96
		2.1	-0.12	0.182	0.191	0.84	-0.11	0.130	0.126	0.79	-0.05	0.152	0.161	0.93
		2.2	0.16	0.294	0.306	0.93	0.10	0.188	0.170	0.92	0.09	0.191	0.219	0.97
		2.3	0.03	0.241	0.249	0.92	0.00	0.156	0.145	0.93	0.02	0.173	0.192	0.95
		3	0.04	0.280	0.223	0.94	0.00	0.129	0.134	0.95	0.03	0.239	0.214	0.95
		4	0.05	0.208	0.259	0.93	0.00	0.097	0.133	0.96	0.05	0.208	0.259	0.93
	5	1	0.40	0.435	0.483	0.94	0.34	0.238	0.273	0.81	0.15	0.240	0.255	0.96
		2.1	-0.19	0.175	0.189	0.75	-0.14	0.116	0.128	0.73	-0.10	0.143	0.152	0.88
		2.2	0.10	0.258	0.297	0.98	0.08	0.155	0.178	0.96	0.04	0.197	0.205	0.96
		2.3	-0.03	0.199	0.235	0.95	-0.02	0.132	0.147	0.91	-0.02	0.172	0.177	0.94
		3	-0.01	0.213	0.308	0.92	-0.01	0.131	0.133	0.93	0.00	0.197	0.204	0.92
		4	0.02	0.247	0.255	0.92	0.00	0.124	0.136	0.93	0.02	0.247	0.255	0.92
	10	1	0.57	0.498	0.509	0.79	0.42	0.301	0.286	0.73	0.24	0.282	0.272	0.86
		2.1	-0.18	0.174	0.199	0.75	-0.15	0.141	0.139	0.76	-0.07	0.152	0.165	0.90
		2.2	0.18	0.285	0.297	0.88	0.12	0.176	0.185	0.90	0.10	0.218	0.217	0.93
		2.3	0.02	0.257	0.241	0.93	-0.01	0.155	0.154	0.91	0.03	0.179	0.190	0.95
		3	0.05	0.272	0.244	0.94	0.02	0.159	0.132	0.88	0.05	0.250	0.214	0.91
		4	0.06	0.265	0.245	0.92	0.01	0.144	0.132	0.89	0.06	0.265	0.245	0.92
Decreasing	2	1	0.51	0.454	0.544	0.84	0.40	0.246	0.264	0.68	0.21	0.262	0.271	0.89
		2.1	-0.19	0.201	0.212	0.79	-0.10	0.106	0.127	0.85	-0.08	0.169	0.162	0.87
		2.2	0.22	0.314	0.368	0.91	0.10	0.149	0.168	0.92	0.11	0.223	0.231	0.95
		2.3	0.04	0.227	0.269	0.95	-0.01	0.133	0.145	0.97	0.02	0.191	0.196	0.92
		3	0.03	0.218	0.231	0.93	0.01	0.130	0.125	0.92	0.03	0.214	0.278	0.89
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	5	1	0.58	0.566	0.655	0.84	0.46	0.299	0.288	0.67	0.23	0.278	0.281	0.90
		2.1	-0.22	0.211	0.228	0.79	-0.12	0.124	0.132	0.83	-0.09	0.168	0.165	0.86
		2.2	0.24	0.382	0.424	0.91	0.11	0.171	0.176	0.91	0.12	0.229	0.236	0.94
		2.3	0.03	0.263	0.303	0.93	-0.01	0.141	0.149	0.96	0.02	0.192	0.197	0.93
		3	0.03	0.223	0.239	0.93	0.01	0.130	0.127	0.92	0.03	0.220	0.278	0.89
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	10	1	0.59	0.504	0.715	0.86	0.49	0.308	0.299	0.66	0.24	0.285	0.291	0.89
		2.1	-0.24	0.216	0.236	0.73	-0.13	0.123	0.135	0.82	-0.10	0.174	0.165	0.85
		2.2	0.24	0.334	0.455	0.92	0.12	0.174	0.180	0.90	0.12	0.232	0.242	0.93
		2.3	0.02	0.233	0.316	0.94	-0.02	0.135	0.150	0.98	0.01	0.196	0.197	0.92
		3	0.04	0.238	0.244	0.93	0.01	0.128	0.127	0.93	0.03	0.222	0.275	0.90
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91

Table 29: Simulation results for post-Lasso selection with fixed personal characteristics under binary endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.52	0.536	0.523	0.91	0.40	0.292	0.270	0.73	0.21	0.268	0.278	0.89
		2.1	-0.16	0.213	0.200	0.79	-0.13	0.137	0.131	0.77	-0.07	0.175	0.165	0.85
		2.2	0.23	0.349	0.352	0.94	0.13	0.181	0.181	0.90	0.12	0.231	0.236	0.94
		2.3	0.04	0.252	0.269	0.96	0.01	0.137	0.151	0.95	0.03	0.198	0.201	0.94
		3	0.04	0.217	0.298	0.92	0.01	0.128	0.128	0.92	0.04	0.212	0.239	0.93
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	5	1	0.53	0.656	1.727	0.9	0.44	0.393	0.311	0.77	0.24	0.300	0.297	0.90
		2.1	-0.20	0.205	0.351	0.79	-0.17	0.112	0.141	0.75	-0.08	0.165	0.167	0.87
		2.2	0.19	0.390	1.049	0.94	0.12	0.208	0.194	0.91	0.11	0.235	0.242	0.93
		2.3	-0.04	0.238	0.684	0.95	-0.04	0.114	0.159	0.99	-0.01	0.185	0.195	0.96
		3	0.03	0.232	0.239	0.93	0.00	0.126	0.130	0.96	0.04	0.233	0.210	0.91
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	10	1	0.73	1.110	1.225	0.89	0.66	0.826	0.445	0.78	0.32	0.367	0.387	0.90
		2.1	-0.33	0.292	0.346	0.74	-0.26	0.181	0.174	0.63	-0.13	0.150	0.188	0.84
		2.2	0.25	0.533	0.707	0.93	0.18	0.371	0.253	0.94	0.13	0.251	0.289	0.93
		2.3	-0.09	0.216	0.382	0.93	-0.07	0.158	0.183	0.94	-0.04	0.173	0.218	0.95
		3	0.03	0.223	0.262	0.92	-0.01	0.122	0.138	0.96	0.02	0.211	0.222	0.94
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
Random	2	1	0.57	0.884	0.494	0.88	0.37	0.299	0.265	0.77	0.23	0.332	0.266	0.92
		2.1	-0.14	0.318	0.193	0.80	-0.14	0.130	0.131	0.79	-0.06	0.198	0.157	0.89
		2.2	0.27	0.660	0.333	0.90	0.12	0.201	0.178	0.90	0.13	0.303	0.226	0.94
		2.3	0.10	0.500	0.254	0.93	0.01	0.162	0.147	0.91	0.06	0.256	0.193	0.93
		3	-0.01	0.410	0.266	0.94	0.01	0.164	0.136	0.94	0.00	0.296	0.233	0.95
		4	0.05	0.208	0.259	0.93	0.00	0.097	0.133	0.96	0.05	0.208	0.259	0.93
	5	1	0.40	0.486	0.506	0.92	0.34	0.238	0.276	0.83	0.15	0.243	0.260	0.94
		2.1	-0.21	0.174	0.195	0.68	-0.16	0.113	0.130	0.74	-0.11	0.145	0.150	0.83
		2.2	0.13	0.287	0.342	0.94	0.09	0.160	0.184	0.94	0.06	0.207	0.217	0.96
		2.3	-0.01	0.210	0.252	0.93	-0.02	0.134	0.150	0.94	-0.01	0.179	0.182	0.93
		3	0.00	0.216	0.294	0.93	-0.01	0.122	0.134	0.94	0.00	0.203	0.209	0.93
		4	0.02	0.247	0.255	0.92	0.00	0.124	0.136	0.93	0.02	0.247	0.255	0.92
	10	1	0.57	0.683	0.504	0.84	0.42	0.303	0.285	0.79	0.22	0.248	0.275	0.87
		2.1	-0.20	0.261	0.208	0.74	-0.14	0.141	0.137	0.77	-0.08	0.149	0.164	0.90
		2.2	0.23	0.417	0.329	0.94	0.12	0.184	0.187	0.91	0.11	0.205	0.227	0.91
		2.3	0.02	0.208	0.243	0.96	-0.01	0.150	0.154	0.90	0.03	0.173	0.194	0.95
		3	0.06	0.254	0.241	0.93	0.02	0.161	0.133	0.88	0.05	0.230	0.214	0.95
		4	0.06	0.265	0.245	0.92	0.01	0.144	0.132	0.89	0.06	0.265	0.245	0.92
Decreasing	2	1	0.54	0.483	0.615	0.84	0.40	0.266	0.266	0.70	0.22	0.264	0.273	0.90
		2.1	-0.16	0.227	0.215	0.74	-0.15	0.139	0.132	0.73	-0.07	0.172	0.163	0.85
		2.2	0.24	0.334	0.408	0.94	0.13	0.173	0.179	0.90	0.12	0.228	0.231	0.92
		2.3	0.05	0.244	0.290	0.96	0.00	0.137	0.147	0.94	0.03	0.194	0.197	0.94
		3	0.05	0.223	0.473	0.93	0.01	0.125	0.128	0.92	0.05	0.220	1.415	0.92
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	5	1	0.59	0.572	0.641	0.86	0.43	0.307	0.292	0.70	0.24	0.287	0.284	0.88
		2.1	-0.18	0.227	0.226	0.80	-0.16	0.126	0.138	0.72	-0.07	0.183	0.164	0.88
		2.2	0.24	0.381	0.421	0.93	0.13	0.175	0.190	0.87	0.13	0.242	0.236	0.92
		2.3	0.03	0.257	0.301	0.95	-0.01	0.136	0.153	0.96	0.02	0.198	0.198	0.93
		3	0.03	0.216	0.257	0.93	0.01	0.123	0.129	0.94	0.04	0.228	0.225	0.93
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	10	1	0.62	0.552	0.626	0.91	0.46	0.333	0.306	0.72	0.25	0.290	0.288	0.89
		2.1	-0.20	0.229	0.224	0.75	-0.17	0.132	0.143	0.73	-0.09	0.181	0.166	0.85
		2.2	0.25	0.344	0.407	0.93	0.14	0.186	0.198	0.93	0.13	0.240	0.237	0.93
		2.3	0.02	0.241	0.286	0.95	-0.01	0.134	0.157	0.96	0.02	0.197	0.197	0.93
		3	0.04	0.219	0.415	0.92	0.00	0.122	0.130	0.94	0.04	0.232	0.311	0.93
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91

Table 30: Simulation results with direct-Lasso selection with non-fixed personal characteristics with time-to-event endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.65	0.730	0.814	0.89	0.52	0.459	0.455	0.82	0.26	0.437	0.454	0.94
		2	0.03	0.301	0.339	0.96	0.01	0.192	0.206	0.95	0.02	0.278	0.287	0.97
		3	-0.03	0.238	0.265	0.97	-0.02	0.157	0.184	0.97	-0.02	0.244	0.269	0.95
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	5	1	1.14	3.476	1.526	0.93	0.69	0.945	0.596	0.86	0.32	0.558	0.518	0.95
		2	0.03	0.524	0.515	0.95	-0.01	0.222	0.215	0.97	0.01	0.286	0.299	0.95
		3	-0.04	0.243	0.275	0.96	-0.02	0.163	0.187	0.97	-0.02	0.264	0.277	0.95
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	10	1	1.16	2.053	6.957	0.94	0.98	2.178	0.909	0.92	0.36	0.605	1.228	0.94
		2	-0.07	0.253	1.764	0.96	-0.08	0.165	0.233	0.98	-0.03	0.247	0.302	0.95
		3	-0.04	0.243	0.281	0.95	-0.03	0.156	0.193	0.96	-0.03	0.259	0.287	0.95
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
Random	2	1	0.54	0.621	0.713	0.91	0.51	0.422	0.456	0.83	0.21	0.378	0.421	0.94
		2	-0.01	0.277	0.299	0.97	0.03	0.208	0.214	0.97	0.00	0.256	0.275	0.95
		3	-0.06	0.213	0.254	0.96	-0.01	0.178	0.182	0.97	-0.05	0.222	0.252	0.96
		4	-0.01	0.258	0.324	0.97	0.00	0.177	0.191	0.96	-0.01	0.258	0.324	0.97
	5	1	0.62	0.657	0.829	0.91	0.58	0.461	0.503	0.88	0.25	0.369	0.458	0.95
		2	-0.01	0.241	0.336	0.99	0.05	0.193	0.222	0.95	0.01	0.235	0.294	1.00
		3	-0.04	0.203	0.265	0.96	0.00	0.150	0.193	0.96	-0.03	0.202	0.264	0.97
		4	0.01	0.250	0.323	0.97	0.01	0.161	0.206	0.98	0.01	0.250	0.323	0.97
	10	1	0.88	1.015	0.878	0.86	0.61	0.518	0.513	0.82	0.34	0.499	0.458	0.88
		2	0.03	0.313	0.325	0.94	0.03	0.203	0.215	0.95	0.05	0.294	0.281	0.90
		3	-0.01	0.240	0.254	0.93	-0.01	0.172	0.182	0.93	0.00	0.245	0.252	0.94
		4	0.05	0.309	0.330	0.95	-0.01	0.173	0.188	0.93	0.05	0.309	0.330	0.95
Decreasing	2	1	0.62	0.644	0.759	0.87	0.52	0.417	0.450	0.82	0.26	0.418	0.451	0.94
		2	0.02	0.291	0.325	0.97	0.01	0.183	0.205	0.97	0.02	0.271	0.287	0.96
		3	-0.04	0.240	0.265	0.96	-0.02	0.150	0.184	0.97	-0.03	0.247	0.268	0.96
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	5	1	0.70	0.722	0.868	0.87	0.59	0.467	0.491	0.82	0.28	0.448	0.475	0.94
		2	0.02	0.292	0.372	0.97	0.00	0.185	0.208	0.97	0.02	0.274	0.290	0.96
		3	-0.04	0.238	0.268	0.95	-0.02	0.149	0.186	0.98	-0.03	0.246	0.272	0.94
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	10	1	0.73	0.744	0.939	0.87	0.63	0.474	0.519	0.82	0.30	0.453	0.493	0.94
		2	0.01	0.293	0.339	0.98	0.00	0.191	0.209	0.97	0.02	0.277	0.292	0.96
		3	-0.04	0.238	0.269	0.96	-0.02	0.149	0.187	0.97	-0.03	0.246	0.273	0.94
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94

Table 31: Simulation results with direct-Lasso selection with fixed personal characteristics with time-to-event endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.69	0.844	0.867	0.94	0.53	0.469	0.461	0.85	0.28	0.453	0.462	0.94
		2	0.03	0.298	0.362	0.96	0.02	0.187	0.208	0.97	0.02	0.276	0.288	0.96
		3	-0.02	0.249	0.287	0.96	-0.02	0.155	0.182	0.97	-0.02	0.245	0.276	0.97
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	5	1	0.30	5.454	1.294	0.95	0.66	0.761	0.596	0.87	0.34	0.613	0.528	0.95
		2	-0.06	0.375	0.402	0.95	-0.01	0.209	0.215	0.96	0.00	0.287	0.297	0.95
		3	-0.03	0.259	0.300	0.97	-0.02	0.167	0.186	0.98	-0.03	0.264	0.283	0.96
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	10	1	78.75	776.634	107.201	0.94	0.89	1.328	0.819	0.88	0.38	0.662	0.625	0.94
		2	-3.72	36.126	2.578	0.95	-0.07	0.175	0.229	0.99	-0.04	0.236	0.307	0.95
		3	-0.03	0.259	0.312	0.95	-0.03	0.160	0.192	0.97	-0.03	0.257	0.292	0.94
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
Random	2	1	0.63	0.770	0.760	0.90	0.52	0.420	0.459	0.88	0.23	0.399	0.440	0.93
		2	0.04	0.369	0.324	0.96	0.04	0.212	0.217	0.97	0.02	0.268	0.289	0.94
		3	0.00	0.356	0.291	0.96	-0.01	0.178	0.182	0.97	0.04	0.746	0.278	0.96
		4	-0.01	0.258	0.324	0.97	0.00	0.177	0.191	0.96	-0.01	0.258	0.324	0.97
	5	1	0.69	0.708	0.850	0.88	0.57	0.459	0.506	0.9	0.27	0.381	0.474	0.97
		2	0.03	0.258	0.350	1.00	0.04	0.184	0.225	0.97	0.03	0.242	0.304	0.99
		3	-0.01	0.227	0.303	0.97	0.00	0.149	0.194	0.96	-0.01	0.219	0.290	0.97
		4	0.01	0.250	0.323	0.97	0.01	0.161	0.206	0.98	0.01	0.250	0.323	0.97
	10	1	0.92	1.022	0.955	0.87	0.61	0.503	0.513	0.78	0.37	0.520	0.486	0.89
		2	0.07	0.322	0.349	0.93	0.03	0.206	0.214	0.94	0.07	0.304	0.294	0.91
		3	0.04	0.287	0.296	0.95	-0.01	0.172	0.183	0.94	0.03	0.278	0.281	0.95
		4	0.05	0.309	0.330	0.95	-0.01	0.173	0.188	0.93	0.05	0.309	0.330	0.95
Decreasing	2	1	0.66	0.673	0.816	0.89	0.53	0.425	0.455	0.84	0.27	0.421	0.461	0.93
		2	0.02	0.286	0.322	0.98	0.02	0.185	0.206	0.96	0.02	0.272	0.288	0.97
		3	-0.02	0.250	0.289	0.95	-0.02	0.149	0.183	0.97	-0.02	0.251	0.277	0.96
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	5	1	0.75	0.769	0.922	0.91	0.59	0.463	0.501	0.83	0.30	0.453	0.486	0.94
		2	0.01	0.284	0.331	0.97	0.01	0.187	0.210	0.96	0.02	0.272	0.290	0.96
		3	-0.02	0.251	0.293	0.95	-0.02	0.148	0.185	0.97	-0.02	0.247	0.282	0.97
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	10	1	0.81	0.806	0.979	0.90	0.63	0.471	0.520	0.82	0.31	0.459	0.500	0.94
		2	0.02	0.295	0.336	0.97	0.01	0.198	0.211	0.95	0.02	0.274	0.291	0.96
		3	-0.02	0.251	0.294	0.95	-0.02	0.148	0.186	0.97	-0.02	0.246	0.282	0.97
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94

Table 32: Simulation results with post-Lasso selection with non-fixed personal characteristics with time-to-event endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.42	0.597	0.672	0.94	0.34	0.346	0.366	0.90	0.16	0.396	0.378	0.93
		2.1	-0.22	0.200	0.227	0.72	-0.11	0.145	0.164	0.91	-0.11	0.202	0.213	0.89
		2.2	0.15	0.394	0.446	0.95	0.06	0.208	0.230	0.93	0.06	0.324	0.314	0.94
		2.3	0.00	0.305	0.334	0.94	-0.02	0.186	0.194	0.94	-0.01	0.280	0.267	0.93
		3	-0.02	0.252	0.303	0.97	-0.01	0.168	0.187	0.97	-0.02	0.256	0.286	0.95
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	5	1	0.38	0.941	0.800	0.95	0.47	0.676	0.454	0.88	0.18	0.495	0.408	0.94
		2.1	-0.26	0.183	0.244	0.78	-0.17	0.144	0.172	0.78	-0.12	0.206	0.218	0.88
		2.2	0.09	0.522	0.493	0.97	0.11	0.349	0.266	0.93	0.06	0.364	0.327	0.95
		2.3	-0.08	0.320	0.335	0.95	-0.06	0.216	0.205	0.93	-0.03	0.306	0.268	0.93
		3	-0.02	0.296	0.326	0.95	-0.02	0.182	0.189	0.98	0.00	0.418	0.294	0.97
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	10	1	0.59	1.237	1.937	0.97	0.72	1.267	0.731	0.88	0.27	0.580	0.510	0.92
		2.1	-0.31	0.363	0.317	0.75	-0.29	0.268	0.217	0.68	-0.17	0.175	0.226	0.89
		2.2	0.15	0.566	1.066	0.98	0.20	0.546	0.388	0.93	0.10	0.384	0.382	0.94
		2.3	-0.08	0.343	0.748	0.93	-0.09	0.351	0.247	0.94	-0.04	0.275	0.284	0.91
		3	-0.01	0.284	0.357	0.95	-0.02	0.169	0.200	0.96	-0.01	0.269	0.381	0.94
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
Random	2	1	0.41	0.568	0.645	0.93	0.36	0.366	0.370	0.86	0.14	0.336	0.379	0.96
		2.1	-0.15	0.212	0.224	0.79	-0.11	0.166	0.166	0.8	-0.09	0.208	0.217	0.94
		2.2	0.11	0.372	0.395	0.94	0.09	0.240	0.240	0.94	0.04	0.276	0.307	0.95
		2.3	-0.01	0.298	0.312	0.94	0.00	0.209	0.203	0.95	-0.02	0.245	0.267	0.95
		3	-0.01	0.298	0.300	0.96	0.00	0.186	0.188	0.97	-0.01	0.278	0.280	0.96
		4	-0.01	0.258	0.324	0.97	0.00	0.177	0.191	0.96	-0.01	0.258	0.324	0.97
	5	1	0.43	0.522	0.694	0.94	0.37	0.349	0.409	0.88	0.16	0.323	0.408	0.96
		2.1	-0.18	0.199	0.255	0.85	-0.13	0.139	0.171	0.88	-0.09	0.185	0.236	0.95
		2.2	0.11	0.311	0.431	0.95	0.10	0.219	0.266	0.96	0.06	0.261	0.329	0.98
		2.3	-0.03	0.239	0.337	0.99	-0.01	0.178	0.213	0.96	-0.01	0.230	0.285	0.98
		3	-0.01	0.234	0.547	0.94	0.00	0.156	0.199	0.96	-0.01	0.230	0.312	0.95
		4	0.01	0.250	0.323	0.97	0.01	0.161	0.206	0.98	0.01	0.250	0.323	0.97
	10	1	0.59	0.738	0.683	0.89	0.38	0.396	0.392	0.88	0.24	0.433	0.399	0.87
		2.1	-0.19	0.203	0.242	0.79	-0.16	0.161	0.166	0.75	-0.07	0.223	0.227	0.89
		2.2	0.18	0.415	0.405	0.92	0.10	0.242	0.253	0.92	0.11	0.335	0.315	0.91
		2.3	0.02	0.316	0.313	0.93	-0.02	0.187	0.202	0.92	0.04	0.286	0.272	0.92
		3	0.04	0.298	0.320	0.95	0.00	0.182	0.189	0.95	0.04	0.297	0.293	0.95
		4	0.05	0.309	0.330	0.95	-0.01	0.173	0.188	0.93	0.05	0.309	0.330	0.95
Decreasing	2	1	0.43	0.606	0.666	0.91	0.36	0.343	0.367	0.85	0.15	0.364	0.381	0.94
		2.1	-0.22	0.210	0.224	0.79	-0.12	0.147	0.163	0.87	-0.11	0.215	0.213	0.9
		2.2	0.17	0.417	0.443	0.93	0.08	0.214	0.231	0.95	0.06	0.301	0.318	0.94
		2.3	0.01	0.299	0.321	0.96	-0.02	0.181	0.192	0.94	-0.01	0.264	0.268	0.93
		3	-0.02	0.255	0.302	0.97	-0.01	0.163	0.188	0.96	-0.02	0.252	0.304	0.97
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	5	1	0.50	0.770	0.782	0.92	0.41	0.366	0.396	0.82	0.17	0.391	0.394	0.94
		2.1	-0.24	0.231	0.240	0.78	-0.13	0.153	0.163	0.84	-0.12	0.209	0.213	0.87
		2.2	0.19	0.511	0.502	0.92	0.09	0.213	0.242	0.93	0.07	0.311	0.324	0.95
		2.3	0.01	0.331	0.350	0.95	-0.02	0.188	0.195	0.96	-0.02	0.265	0.268	0.93
		3	-0.02	0.265	0.310	0.95	-0.01	0.164	0.190	0.96	-0.02	0.255	0.324	0.96
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	10	1	0.49	0.636	0.824	0.91	0.44	0.383	0.408	0.80	0.18	0.385	0.407	0.94
		2.1	-0.26	0.226	0.242	0.72	-0.14	0.151	0.162	0.82	-0.13	0.210	0.211	0.86
		2.2	0.17	0.406	0.519	0.94	0.09	0.222	0.245	0.94	0.06	0.307	0.331	0.96
		2.3	-0.01	0.293	0.363	0.96	-0.02	0.191	0.195	0.93	-0.02	0.263	0.272	0.95
		3	-0.02	0.265	0.314	0.96	-0.02	0.160	0.189	0.97	-0.02	0.260	0.326	0.95
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94

Table 33: Simulation results with post-Lasso selection with fixed personal characteristics with time-to-event endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.46	0.813	0.658	0.93	0.36	0.421	0.375	0.91	0.16	0.376	0.387	0.94
		2.1	-0.20	0.226	0.217	0.79	-0.15	0.154	0.160	0.76	-0.103	0.208	0.216	0.90
		2.2	0.18	0.480	0.436	0.93	0.10	0.262	0.248	0.93	0.07	0.308	0.322	0.95
		2.3	0.02	0.374	0.329	0.94	0.00	0.219	0.203	0.94	-0.01	0.266	0.273	0.94
		3	-0.01	0.260	0.404	0.97	-0.02	0.167	0.190	0.97	0.00	0.274	0.336	0.97
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	5	1	0.39	0.858	1.286	0.95	0.41	0.592	0.434	0.92	0.18	0.448	0.417	0.95
		2.1	-0.24	0.181	0.296	0.80	-0.18	0.131	0.169	0.81	-0.115	0.206	0.217	0.89
		2.2	0.09	0.470	0.792	0.96	0.11	0.319	0.267	0.91	0.06	0.333	0.335	0.94
		2.3	-0.09	0.277	0.524	0.95	-0.04	0.222	0.205	0.97	-0.04	0.259	0.269	0.93
		3	-0.02	0.280	0.348	0.96	-0.02	0.187	0.190	0.98	-0.01	0.303	0.303	0.97
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	10	1	0.49	1.302	1.341	0.95	0.62	1.157	0.563	0.89	0.26	0.542	0.501	0.93
		2.1	-0.32	0.406	0.316	0.73	-0.27	0.199	0.191	0.72	-0.170	0.174	0.222	0.89
		2.2	0.12	0.609	0.771	0.95	0.16	0.528	0.317	0.93	0.08	0.355	0.374	0.93
		2.3	-0.14	0.285	0.407	0.94	-0.08	0.321	0.216	0.93	-0.06	0.269	0.275	0.92
		3	-0.03	0.269	0.334	0.95	-0.03	0.181	0.196	0.98	-0.03	0.256	0.304	0.96
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
Random	2	1	0.46	0.740	0.653	0.93	0.36	0.373	0.381	0.87	0.16	0.384	0.379	0.94
		2.1	-0.18	0.263	0.221	0.75	-0.15	0.155	0.162	0.79	-0.102	0.214	0.211	0.92
		2.2	0.19	0.554	0.438	0.94	0.11	0.255	0.257	0.94	0.08	0.336	0.321	0.95
		2.3	0.04	0.432	0.325	0.95	0.00	0.208	0.209	0.95	0.00	0.293	0.272	0.94
		3	-0.05	0.390	0.335	0.97	0.00	0.210	0.190	0.96	-0.04	0.311	0.301	0.96
		4	-0.01	0.258	0.324	0.97	0.00	0.177	0.191	0.96	-0.01	0.258	0.324	0.97
	5	1	0.42	0.532	0.702	0.94	0.37	0.353	0.413	0.90	0.16	0.323	0.410	0.97
		2.1	-0.20	0.184	0.251	0.85	-0.15	0.142	0.173	0.81	-0.102	0.184	0.232	0.93
		2.2	0.14	0.335	0.479	0.98	0.11	0.223	0.276	0.97	0.07	0.266	0.345	0.98
		2.3	0.00	0.268	0.349	0.95	0.00	0.184	0.221	0.97	0.00	0.232	0.290	0.98
		3	-0.01	0.241	0.480	0.96	0.00	0.152	0.201	0.96	-0.01	0.232	0.317	0.95
		4	0.01	0.250	0.323	0.97	0.01	0.161	0.206	0.98	0.01	0.250	0.323	0.97
	10	1	0.57	0.792	0.675	0.89	0.37	0.383	0.388	0.86	0.23	0.414	0.401	0.91
		2.1	-0.20	0.269	0.244	0.77	-0.16	0.158	0.161	0.75	-0.077	0.221	0.220	0.88
		2.2	0.23	0.492	0.442	0.94	0.10	0.249	0.253	0.92	0.12	0.336	0.329	0.91
		2.3	0.01	0.297	0.319	0.96	-0.02	0.188	0.201	0.94	0.03	0.277	0.277	0.94
		3	0.04	0.285	0.325	0.95	0.00	0.181	0.190	0.96	0.04	0.282	0.300	0.94
		4	0.05	0.309	0.330	0.95	-0.01	0.173	0.188	0.93	0.05	0.309	0.330	0.95
Decreasing	2	1	0.45	0.596	0.724	0.91	0.36	0.356	0.369	0.86	0.17	0.363	0.386	0.94
		2.1	-0.21	0.217	0.228	0.79	-0.16	0.157	0.161	0.77	-0.109	0.211	0.216	0.90
		2.2	0.18	0.413	0.477	0.92	0.10	0.232	0.245	0.93	0.07	0.305	0.321	0.95
		2.3	0.01	0.287	0.351	0.96	-0.01	0.178	0.198	0.94	-0.01	0.257	0.271	0.96
		3	-0.01	0.265	0.376	0.97	-0.01	0.158	0.189	0.97	-0.01	0.263	0.607	0.95
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	5	1	0.48	0.603	0.725	0.95	0.38	0.367	0.401	0.87	0.18	0.390	0.400	0.93
		2.1	-0.22	0.203	0.230	0.80	-0.17	0.147	0.165	0.75	-0.113	0.212	0.214	0.89
		2.2	0.17	0.383	0.470	0.96	0.10	0.233	0.259	0.93	0.07	0.312	0.329	0.95
		2.3	-0.01	0.273	0.336	0.97	-0.02	0.183	0.204	0.96	-0.01	0.263	0.271	0.95
		3	-0.01	0.255	0.343	0.98	-0.02	0.150	0.191	0.99	-0.01	0.260	0.316	0.96
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	10	1	0.50	0.587	0.758	0.94	0.41	0.376	0.415	0.89	0.19	0.379	0.406	0.94
		2.1	-0.25	0.195	0.233	0.71	-0.18	0.146	0.167	0.74	-0.129	0.200	0.212	0.89
		2.2	0.17	0.370	0.483	0.96	0.11	0.238	0.264	0.94	0.07	0.299	0.328	0.97
		2.3	-0.01	0.268	0.339	0.98	-0.02	0.176	0.205	0.97	-0.02	0.254	0.270	0.97
		3	-0.01	0.258	0.383	0.97	-0.02	0.150	0.193	0.99	-0.01	0.262	0.339	0.97
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94

Table 34: Simulation results for direct-SCAD selection with non-fixed personal characteristics under continuous endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.53	0.383	0.422	0.75	0.50	0.337	0.283	0.58	0.24	0.235	0.229	0.80
		2	0.01	0.206	0.194	0.91	-0.02	0.135	0.123	0.88	0.02	0.152	0.152	0.89
		3	0.04	0.185	0.173	0.90	0.02	0.131	0.105	0.89	0.05	0.170	0.171	0.90
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	5	1	1.46	8.304	1.091	0.89	0.91	1.073	1.110	0.86	0.76	4.091	0.381	0.85
		2	-0.06	0.245	0.333	0.90	-0.05	0.212	0.183	0.85	-0.01	0.241	0.181	0.90
		3	0.02	0.207	0.185	0.90	0.01	0.119	0.118	0.93	0.05	0.245	0.191	0.89
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	10	1	1.05	8.676	2.904	0.84	3.25	4.641	4.266	0.86	0.63	0.625	0.649	0.84
		2	-0.11	0.245	0.721	0.94	-0.10	0.381	1.145	0.86	-0.05	0.161	0.226	0.94
		3	0.00	0.210	0.181	0.89	0.01	0.128	0.121	0.92	0.04	0.239	0.217	0.90
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
Random	2	1	0.47	0.402	0.411	0.77	0.43	0.249	0.248	0.65	0.19	0.212	0.208	0.86
		2	-0.02	0.186	0.175	0.91	0.00	0.135	0.117	0.93	-0.01	0.159	0.138	0.91
		3	0.00	0.194	0.159	0.90	0.02	0.158	0.104	0.89	0.01	0.206	0.154	0.89
		4	0.05	0.252	0.245	0.89	0.02	0.127	0.111	0.92	0.05	0.252	0.245	0.89
	5	1	0.54	0.437	0.462	0.78	0.46	0.261	0.268	0.65	0.22	0.226	0.211	0.83
		2	-0.04	0.180	0.179	0.88	-0.01	0.117	0.118	0.92	-0.01	0.154	0.138	0.88
		3	0.00	0.195	0.157	0.88	0.01	0.122	0.111	0.92	0.01	0.194	0.157	0.86
		4	0.05	0.241	0.207	0.86	0.02	0.125	0.119	0.92	0.05	0.241	0.207	0.86
	10	1	0.68	0.612	0.661	0.81	0.58	0.400	0.341	0.66	0.26	0.272	0.230	0.77
		2	-0.05	0.220	0.292	0.85	-0.02	0.132	0.128	0.91	-0.01	0.161	0.141	0.89
		3	0.00	0.184	0.154	0.90	0.02	0.126	0.107	0.92	0.01	0.196	0.152	0.89
		4	0.05	0.232	0.195	0.92	0.02	0.124	0.109	0.88	0.05	0.232	0.195	0.92
Decreasing	2	1	0.58	0.403	0.417	0.70	0.55	0.326	0.269	0.50	0.26	0.231	0.225	0.80
		2	0.02	0.179	0.207	0.89	0.01	0.134	0.124	0.88	0.03	0.150	0.150	0.87
		3	0.05	0.182	0.178	0.91	0.02	0.118	0.104	0.94	0.06	0.184	0.177	0.92
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	5	1	0.67	0.474	0.549	0.70	0.65	0.393	0.310	0.49	0.30	0.254	0.245	0.81
		2	0.00	0.180	0.249	0.93	0.00	0.141	0.129	0.88	0.03	0.150	0.156	0.88
		3	0.04	0.187	0.180	0.91	0.02	0.119	0.105	0.94	0.06	0.183	0.178	0.92
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	10	1	0.70	0.476	0.503	0.72	0.68	0.410	0.330	0.52	0.32	0.258	0.255	0.81
		2	0.01	0.185	0.213	0.92	0.00	0.141	0.133	0.88	0.03	0.152	0.157	0.91
		3	0.04	0.181	0.178	0.91	0.02	0.119	0.106	0.94	0.05	0.183	0.179	0.92
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90

Table 35: Simulation results for direct-SCAD selection with fixed personal characteristics under continuous endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.62	0.985	0.590	0.79	0.44	0.281	0.244	0.59	0.24	0.249	0.228	0.84
		2	0.01	0.269	0.202	0.90	-0.01	0.114	0.117	0.91	0.01	0.149	0.147	0.90
		3	0.03	0.173	0.173	0.91	0.02	0.131	0.103	0.89	0.03	0.162	0.158	0.91
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	5	1	2.78	15.643	1.988	0.87	0.79	1.074	0.589	0.80	0.95	6.071	0.369	0.85
		2	-0.13	0.612	0.307	0.91	-0.05	0.250	0.159	0.87	-0.03	0.250	0.175	0.87
		3	0.03	0.200	0.186	0.92	0.01	0.120	0.112	0.90	0.03	0.246	0.185	0.90
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	10	1	3.05	6.185	8.215	0.88	2.14	2.764	1.898	0.73	0.58	0.607	0.740	0.87
		2	-3.65	34.085	1.791	0.90	-0.09	0.213	0.345	0.90	-0.08	0.171	0.214	0.90
		3	0.03	0.213	0.245	0.88	0.01	0.123	0.116	0.90	0.03	0.221	0.193	0.88
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
Random	2	1	0.50	0.522	0.404	0.81	0.41	0.246	0.224	0.59	0.21	0.245	0.211	0.82
		2	0.02	0.334	0.180	0.91	0.00	0.133	0.114	0.93	0.01	0.180	0.141	0.92
		3	0.24	2.235	0.181	0.89	0.02	0.165	0.104	0.91	0.16	1.474	0.166	0.89
		4	0.05	0.252	0.245	0.89	0.02	0.127	0.111	0.92	0.05	0.252	0.245	0.89
	5	1	0.57	0.463	0.451	0.79	0.44	0.261	0.243	0.63	0.22	0.223	0.216	0.82
		2	-0.01	0.169	0.174	0.91	-0.01	0.113	0.115	0.92	0.00	0.149	0.135	0.90
		3	0.02	0.190	0.186	0.90	0.01	0.117	0.111	0.91	0.02	0.174	0.171	0.88
		4	0.05	0.241	0.207	0.86	0.02	0.125	0.119	0.92	0.05	0.241	0.207	0.86
	10	1	0.83	0.891	0.643	0.79	0.56	0.384	0.323	0.64	0.28	0.270	0.242	0.81
		2	-0.02	0.200	0.200	0.90	-0.03	0.133	0.124	0.92	0.00	0.164	0.144	0.89
		3	0.04	0.244	0.194	0.91	0.02	0.123	0.108	0.92	0.04	0.219	0.167	0.92
		4	0.05	0.232	0.195	0.92	0.02	0.124	0.109	0.88	0.05	0.232	0.195	0.92
Decreasing	2	1	0.59	0.455	0.388	0.73	0.49	0.303	0.237	0.52	0.25	0.247	0.216	0.83
		2	0.02	0.203	0.191	0.90	0.00	0.118	0.117	0.90	0.02	0.161	0.145	0.89
		3	0.05	0.292	0.176	0.89	0.01	0.109	0.103	0.92	0.03	0.184	0.164	0.91
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	5	1	0.71	0.565	0.455	0.67	0.57	0.352	0.270	0.51	0.29	0.269	0.233	0.81
		2	0.02	0.224	0.247	0.89	0.00	0.126	0.124	0.92	0.02	0.167	0.150	0.89
		3	0.05	0.296	0.176	0.87	0.02	0.113	0.104	0.93	0.04	0.186	0.165	0.89
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	10	1	0.76	0.600	0.593	0.69	0.61	0.375	0.287	0.49	0.30	0.268	0.244	0.81
		2	0.02	0.229	0.249	0.88	-0.01	0.127	0.126	0.93	0.02	0.165	0.151	0.89
		3	0.06	0.306	0.177	0.89	0.02	0.112	0.105	0.93	0.03	0.185	0.165	0.90
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90

Table 36: Simulation results for post-SCAD selection with non-fixed personal characteristics under continuous endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.50	0.763	1.490	0.92	0.46	0.388	0.400	0.84	0.22	0.275	0.285	0.90
		2.1	-0.23	0.185	0.303	0.83	-0.04	0.291	0.240	0.89	-0.10	0.143	0.165	0.86
		2.2	0.23	0.464	0.329	0.86	0.14	0.221	0.158	0.85	0.11	0.211	0.180	0.87
		2.3	0.05	0.292	0.516	0.97	0.00	0.153	0.208	0.97	0.04	0.237	0.198	0.90
		3	0.02	0.182	0.170	0.91	0.04	0.259	0.114	0.90	0.03	0.170	0.163	0.90
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	5	1	0.75	1.313	2.103	0.96	0.68	0.471	1.598	0.84	0.31	0.365	0.467	0.90
		2.1	-0.26	0.376	0.273	0.72	-0.03	0.456	0.308	0.89	-0.08	0.212	0.180	0.85
		2.2	-0.31	4.515	0.383	0.95	0.18	0.365	0.384	0.90	0.11	0.305	0.263	0.93
		2.3	0.15	1.539	1.546	0.95	0.00	0.261	0.643	0.98	0.02	0.290	0.314	0.92
		3	0.02	0.247	0.191	0.92	0.01	0.148	0.115	0.96	0.03	0.287	0.172	0.91
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	10	1	0.87	1.722	13.960	0.94	1.31	1.676	8.397	0.94	0.43	0.431	1.003	0.89
		2.1	-0.39	0.453	2.120	0.80	0.15	4.395	0.524	0.86	-0.15	0.189	0.310	0.81
		2.2	0.35	0.890	1.256	0.93	0.26	0.602	0.286	0.90	0.15	0.324	0.272	0.90
		2.3	-0.07	0.521	17.152	0.95	-0.03	0.457	1.129	0.99	-0.01	0.229	0.513	0.95
		3	0.02	0.197	0.228	0.92	0.01	0.132	0.124	0.89	0.02	0.197	0.181	0.88
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
Random	2	1	0.58	0.641	2.877	0.86	0.43	0.325	0.423	0.79	0.22	0.268	0.287	0.84
		2.1	-0.17	0.281	0.986	0.83	-0.11	0.189	0.180	0.83	-0.09	0.153	0.244	0.81
		2.2	0.19	0.351	0.319	0.90	0.12	0.202	0.156	0.86	0.11	0.278	0.177	0.88
		2.3	0.06	0.391	0.532	0.92	0.10	0.558	0.218	0.91	0.04	0.225	0.205	0.91
		3	4.50	44.974	0.202	0.93	0.03	0.212	0.115	0.90	-0.08	0.895	0.188	0.93
		4	0.05	0.252	0.245	0.89	0.02	0.127	0.111	0.92	0.05	0.252	0.245	0.89
	5	1	0.57	0.781	1.780	0.90	0.73	2.646	0.521	0.79	0.22	0.331	0.271	0.90
		2.1	-0.24	0.225	0.289	0.71	-0.12	0.307	0.170	0.79	-0.11	0.184	0.155	0.77
		2.2	0.17	0.250	0.312	0.90	0.11	0.150	0.162	0.90	0.09	0.188	0.184	0.89
		2.3	0.03	0.310	0.627	0.90	0.04	0.235	0.301	0.92	0.01	0.176	0.233	0.89
		3	0.01	0.182	0.228	0.90	0.01	0.124	0.120	0.93	0.02	0.187	0.180	0.91
		4	0.05	0.241	0.207	0.86	0.02	0.125	0.119	0.92	0.05	0.241	0.207	0.86
	10	1	0.58	3.020	4.510	0.87	0.57	0.468	1.275	0.82	0.31	0.458	0.448	0.86
		2.1	-0.22	0.300	0.305	0.72	-0.17	0.203	0.177	0.75	-0.10	0.168	0.149	0.76
		2.2	0.21	0.344	0.335	0.89	0.14	0.187	0.187	0.84	0.10	0.208	0.186	0.88
		2.3	0.04	0.400	1.043	0.90	0.02	0.203	1.188	0.92	0.01	0.194	0.206	0.90
		3	0.03	0.229	0.217	0.92	0.03	0.141	0.117	0.91	0.03	0.216	0.174	0.91
		4	0.05	0.232	0.195	0.92	0.02	0.124	0.109	0.88	0.05	0.232	0.195	0.92
Decreasing	2	1	0.51	0.583	0.800	0.89	0.54	0.518	0.468	0.79	0.23	0.298	0.266	0.92
		2.1	-0.23	0.190	0.311	0.72	-0.10	0.217	0.178	0.85	-0.11	0.146	0.143	0.82
		2.2	0.23	0.416	0.310	0.88	0.15	0.200	0.164	0.85	0.12	0.216	0.186	0.87
		2.3	0.04	0.267	0.677	0.93	0.03	0.188	0.270	0.91	0.04	0.205	0.196	0.87
		3	0.02	0.185	0.193	0.89	0.03	0.131	0.115	0.93	0.03	0.181	0.257	0.87
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	5	1	0.58	0.600	1.582	0.87	0.59	0.518	0.515	0.82	0.25	0.301	0.320	0.89
		2.1	-0.26	0.207	0.310	0.70	-0.12	0.206	0.192	0.87	-0.12	0.144	0.151	0.76
		2.2	0.24	0.430	0.359	0.90	0.15	0.207	0.173	0.87	0.12	0.224	0.190	0.87
		2.3	0.05	0.312	0.724	0.87	0.02	0.200	0.343	0.95	0.03	0.192	0.216	0.92
		3	0.02	0.189	0.332	0.90	0.02	0.127	0.116	0.91	0.03	0.181	0.281	0.85
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	10	1	0.73	1.231	1.763	0.90	0.69	0.889	0.573	0.84	0.26	0.312	0.451	0.89
		2.1	-0.28	0.196	0.281	0.72	-0.13	0.216	0.192	0.82	-0.13	0.147	0.143	0.74
		2.2	0.25	0.389	0.406	0.91	0.15	0.219	0.175	0.84	0.12	0.222	0.196	0.88
		2.3	0.12	0.753	0.491	0.91	0.05	0.236	0.438	0.92	0.04	0.221	0.231	0.88
		3	0.02	0.180	0.179	0.92	0.02	0.133	0.116	0.88	0.03	0.182	0.408	0.87
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90

Table 37: Simulation results for post-SCAD selection with fixed personal characteristics under continuous endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.58	0.727	1.931	0.93	0.56	0.969	0.514	0.86	0.21	0.247	0.346	0.93
		2.1	-0.21	0.172	0.193	0.76	-0.16	0.138	0.139	0.73	-0.10	0.137	0.159	0.81
		2.2	0.23	0.335	0.351	0.86	0.15	0.237	0.161	0.85	0.12	0.217	0.184	0.86
		2.3	-0.14	2.671	0.734	0.91	0.14	0.360	0.242	0.89	0.06	0.217	0.241	0.91
		3	0.04	0.189	0.266	0.92	0.02	0.132	0.125	0.91	0.04	0.187	0.209	0.92
		4	0.04	0.217	0.197	0.90	0.02	0.121	0.114	0.90	0.04	0.217	0.197	0.90
	5	1	0.69	1.003	5.576	0.96	0.79	1.035	2.048	0.86	0.33	0.365	0.470	0.91
		2.1	-0.27	0.280	0.315	0.72	-0.16	0.177	0.179	0.80	-0.10	0.177	0.152	0.77
		2.2	0.59	4.056	0.641	0.93	0.21	0.808	0.299	0.91	0.15	0.358	0.248	0.93
		2.3	-0.39	3.419	2.246	0.93	-0.08	1.543	1.159	0.90	0.04	0.319	0.263	0.91
		3	0.03	0.228	0.417	0.96	0.02	0.188	0.123	0.93	0.05	0.350	0.234	0.95
		4	0.04	0.217	0.197	0.90	0.02	0.121	0.114	0.90	0.04	0.217	0.197	0.90
	10	1	1.39	2.844	20.341	0.95	0.73	4.259	6.772	0.88	0.44	0.570	3.076	0.91
		2.1	-0.40	0.693	1.970	0.82	-0.26	0.354	0.290	0.85	-0.16	0.186	0.231	0.79
		2.2	0.26	0.534	0.970	0.96	0.23	0.470	0.263	0.88	0.15	0.230	0.309	0.91
		2.3	0.62	4.502	7.014	0.94	-0.04	0.559	2.234	0.93	0.05	0.911	0.535	0.95
		3	0.03	0.204	0.248	0.91	0.01	0.130	0.128	0.92	0.03	0.199	0.263	0.94
		4	0.04	0.217	0.197	0.90	0.02	0.121	0.114	0.90	0.04	0.217	0.197	0.90
Random	2	1	0.66	1.568	1.433	0.87	0.44	0.398	0.410	0.74	0.24	0.448	0.262	0.87
		2.1	-0.22	0.231	0.319	0.71	-0.16	0.136	0.126	0.66	-0.10	0.154	0.219	0.74
		2.2	0.07	1.134	0.287	0.88	0.13	0.219	0.158	0.89	0.13	0.382	0.180	0.90
		2.3	-0.12	2.584	1.194	0.90	0.04	0.198	0.201	0.93	0.05	0.218	0.248	0.91
		3	0.00	0.306	0.208	0.94	0.04	0.366	0.112	0.93	0.00	0.363	0.197	0.94
		4	0.05	0.252	0.176	0.88	0.02	0.127	0.107	0.86	0.05	0.252	0.175	0.88
	5	1	0.86	1.848	3.574	0.90	0.49	0.477	1.433	0.81	0.25	0.363	0.348	0.86
		2.1	-0.23	0.176	0.259	0.73	-0.17	0.131	0.139	0.72	-0.11	0.137	0.136	0.77
		2.2	0.18	0.271	0.309	0.90	0.12	0.155	0.166	0.91	0.10	0.196	0.183	0.90
		2.3	0.10	0.306	0.975	0.89	0.04	0.166	0.320	0.95	0.04	0.188	0.317	0.91
		3	0.03	0.200	0.259	0.92	0.01	0.125	0.118	0.91	0.03	0.201	0.211	0.92
		4	0.05	0.241	0.181	0.89	0.02	0.125	0.108	0.93	0.05	0.241	0.181	0.89
	10	1	0.74	2.369	2.541	0.94	0.58	0.618	0.631	0.87	0.28	0.362	0.384	0.89
		2.1	-0.24	0.213	0.246	0.77	-0.17	0.178	0.150	0.74	-0.11	0.150	0.140	0.81
		2.2	0.22	0.304	0.337	0.86	0.14	0.187	0.175	0.85	0.12	0.217	0.189	0.89
		2.3	1.18	10.787	3.298	0.93	0.05	0.235	0.350	0.91	0.03	0.192	0.330	0.91
		3	0.06	0.322	0.275	0.93	0.03	0.153	0.120	0.92	0.04	0.227	0.196	0.93
		4	0.05	0.232	0.224	0.92	0.02	0.124	0.116	0.92	0.05	0.232	0.224	0.92
Decreasing	2	1	0.67	1.116	2.411	0.92	0.58	0.960	0.497	0.80	0.24	0.271	0.304	0.89
		2.1	-0.24	0.270	0.210	0.67	-0.17	0.119	0.149	0.71	-0.11	0.140	0.133	0.81
		2.2	0.26	0.452	0.312	0.86	0.16	0.199	0.163	0.84	0.14	0.248	0.187	0.84
		2.3	-0.30	2.477	0.599	0.91	0.08	0.220	0.241	0.92	0.10	0.452	0.251	0.90
		3	0.05	0.208	0.250	0.90	0.03	0.129	0.125	0.94	0.05	0.196	0.626	0.90
		4	0.04	0.217	0.197	0.90	0.02	0.121	0.114	0.90	0.04	0.217	0.197	0.90
	5	1	0.65	0.705	1.661	0.91	0.58	0.558	1.143	0.81	0.25	0.265	0.389	0.92
		2.1	-0.27	0.210	0.228	0.70	-0.19	0.129	0.146	0.72	-0.13	0.142	0.137	0.77
		2.2	0.28	0.458	0.419	0.87	0.15	0.203	0.173	0.87	0.14	0.249	0.196	0.87
		2.3	-0.01	0.719	1.321	0.92	0.07	0.263	0.247	0.92	0.10	0.577	0.274	0.92
		3	0.05	0.217	0.274	0.92	0.02	0.139	0.137	0.91	0.04	0.188	0.245	0.92
		4	0.04	0.217	0.197	0.90	0.02	0.121	0.114	0.90	0.04	0.217	0.197	0.90
	10	1	1.01	2.962	2.768	0.91	0.64	0.594	0.728	0.81	0.25	0.257	0.509	0.94
		2.1	-0.29	0.261	0.227	0.66	-0.20	0.141	0.158	0.70	-0.12	0.142	0.139	0.77
		2.2	0.31	0.491	0.351	0.89	0.18	0.230	0.177	0.84	0.14	0.240	0.203	0.86
		2.3	0.01	0.592	1.150	0.90	0.09	0.322	0.242	0.91	0.15	1.152	0.229	0.89
		3	0.05	0.215	0.315	0.93	0.02	0.133	0.163	0.93	0.04	0.188	0.238	0.93
		4	0.04	0.217	0.197	0.90	0.02	0.121	0.114	0.90	0.04	0.217	0.197	0.90

Table 38: Simulation results for direct-SCAD selection with non-fixed personal characteristics under binary endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.55	0.565	0.496	0.86	0.47	0.372	0.311	0.72	0.24	0.268	0.289	0.88
		2	0.01	0.219	0.228	0.92	-0.03	0.141	0.137	0.93	0.02	0.189	0.193	0.94
		3	0.05	0.219	0.217	0.90	0.01	0.123	0.125	0.94	0.05	0.209	0.218	0.92
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	5	1	0.88	1.392	1.059	0.87	0.90	1.068	0.962	0.88	0.39	0.602	0.434	0.94
		2	-0.03	0.431	0.344	0.91	-0.07	0.257	0.184	0.93	-0.01	0.234	0.210	0.90
		3	0.03	0.228	0.221	0.88	0.03	0.397	0.135	0.94	0.05	0.235	0.233	0.93
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	10	1	2.54	5.805	2.784	0.86	3.28	5.094	4.131	0.83	0.72	0.844	0.715	0.87
		2	-0.07	0.605	0.805	0.92	-0.11	0.380	0.864	0.88	-0.04	0.241	0.263	0.96
		3	0.01	0.234	0.228	0.89	0.00	0.118	0.142	0.96	0.04	0.237	0.266	0.92
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
Random	2	1	0.49	0.401	0.474	0.88	0.42	0.277	0.283	0.76	0.21	0.224	0.265	0.93
		2	-0.02	0.189	0.202	0.94	-0.01	0.130	0.134	0.94	0.00	0.158	0.177	0.97
		3	0.00	0.173	0.202	0.95	0.00	0.125	0.128	0.94	0.01	0.177	0.199	0.94
		4	0.05	0.208	0.259	0.93	0.00	0.097	0.133	0.96	0.05	0.208	0.259	0.93
	5	1	0.44	0.394	0.507	0.91	0.42	0.253	0.311	0.74	0.17	0.242	0.263	0.93
		2	-0.07	0.163	0.193	0.90	-0.03	0.108	0.135	0.95	-0.04	0.159	0.167	0.92
		3	-0.03	0.192	0.185	0.92	-0.01	0.125	0.128	0.92	-0.02	0.184	0.186	0.91
		4	0.02	0.247	0.255	0.92	0.00	0.124	0.136	0.93	0.02	0.247	0.255	0.92
	10	1	0.73	0.634	0.716	0.76	0.54	0.376	0.376	0.72	0.27	0.269	0.288	0.83
		2	-0.05	0.204	0.250	0.90	-0.03	0.146	0.142	0.90	-0.01	0.164	0.176	0.97
		3	0.01	0.197	0.198	0.90	0.01	0.152	0.129	0.90	0.02	0.216	0.193	0.92
		4	0.06	0.265	0.245	0.92	0.01	0.144	0.132	0.89	0.06	0.265	0.245	0.92
Decreasing	2	1	0.57	0.451	0.499	0.83	0.52	0.306	0.301	0.65	0.26	0.273	0.287	0.89
		2	0.01	0.206	0.241	0.92	-0.01	0.124	0.139	0.95	0.02	0.183	0.189	0.93
		3	0.05	0.228	0.229	0.91	0.01	0.121	0.124	0.93	0.06	0.230	0.231	0.89
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	5	1	0.66	0.510	0.623	0.79	0.60	0.362	0.345	0.64	0.30	0.302	0.310	0.88
		2	0.00	0.212	0.277	0.92	-0.01	0.135	0.143	0.94	0.02	0.184	0.195	0.93
		3	0.05	0.225	0.232	0.92	0.01	0.120	0.126	0.93	0.07	0.254	0.229	0.91
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	10	1	0.69	0.527	0.597	0.80	0.65	0.382	0.367	0.62	0.32	0.306	0.322	0.87
		2	0.00	0.216	0.246	0.92	-0.02	0.135	0.147	0.94	0.02	0.187	0.196	0.95
		3	0.05	0.223	0.230	0.92	0.01	0.120	0.127	0.93	0.07	0.252	0.231	0.91
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91

Table 39: Simulation results for direct-SCAD selection with fixed personal characteristics under binary endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.57	0.706	0.654	0.87	0.42	0.286	0.271	0.74	0.22	0.267	0.283	0.92
		2	0.01	0.238	0.236	0.92	-0.02	0.121	0.131	0.98	0.01	0.181	0.186	0.93
		3	0.03	0.214	0.221	0.91	0.00	0.119	0.123	0.95	0.06	0.361	0.204	0.91
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	5	1	2.09	10.022	2.009	0.91	1.26	6.204	0.576	0.81	0.38	0.679	0.419	0.93
		2	-0.61	5.411	0.311	0.89	-0.09	0.163	0.163	0.94	-0.04	0.193	0.201	0.90
		3	0.03	0.224	0.224	0.90	0.00	0.112	0.130	0.95	0.04	0.242	0.228	0.92
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	10	1	2.96	14.735	9.056	0.85	1.82	2.688	1.819	0.82	0.65	1.089	0.775	0.92
		2	-5.77	44.229	1.814	0.86	-0.11	0.219	0.324	0.92	-0.08	0.233	0.238	0.93
		3	0.04	0.254	0.270	0.91	0.00	0.122	0.138	0.96	0.03	0.237	0.236	0.93
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
Random	2	1	0.52	0.450	0.462	0.85	0.39	0.255	0.260	0.75	0.22	0.244	0.267	0.90
		2	0.01	0.216	0.206	0.96	-0.01	0.121	0.130	0.95	0.01	0.167	0.180	0.98
		3	-0.26	2.784	0.227	0.97	0.00	0.129	0.128	0.95	-0.90	9.170	0.215	0.96
		4	0.05	0.208	0.259	0.93	0.00	0.097	0.133	0.96	0.05	0.208	0.259	0.93
	5	1	0.50	0.492	0.506	0.90	0.39	0.241	0.284	0.72	0.18	0.251	0.270	0.93
		2	-0.05	0.163	0.199	0.93	-0.04	0.110	0.132	0.95	-0.03	0.157	0.168	0.93
		3	0.00	0.203	0.212	0.93	-0.01	0.117	0.129	0.92	-0.01	0.187	0.199	0.93
		4	0.02	0.247	0.255	0.92	0.00	0.124	0.136	0.93	0.02	0.247	0.255	0.92
	10	1	0.76	0.626	0.730	0.80	0.54	0.384	0.368	0.71	0.28	0.280	0.319	0.83
		2	-0.02	0.195	0.259	0.92	-0.04	0.136	0.140	0.88	0.00	0.167	0.184	0.98
		3	0.05	0.287	0.238	0.96	0.01	0.150	0.130	0.89	0.05	0.242	0.213	0.95
		4	0.06	0.265	0.245	0.92	0.01	0.144	0.132	0.89	0.06	0.265	0.245	0.92
Decreasing	2	1	0.58	0.471	0.469	0.79	0.46	0.271	0.269	0.67	0.24	0.265	0.274	0.88
		2	0.01	0.205	0.225	0.89	-0.01	0.117	0.132	0.98	0.01	0.177	0.183	0.91
		3	0.04	0.210	0.225	0.93	0.00	0.113	0.123	0.94	0.04	0.210	0.213	0.91
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	5	1	0.68	0.556	0.545	0.77	0.53	0.320	0.303	0.65	0.29	0.301	0.294	0.85
		2	0.01	0.215	0.273	0.89	-0.02	0.121	0.137	0.97	0.01	0.181	0.186	0.90
		3	0.04	0.217	0.228	0.92	0.00	0.114	0.124	0.95	0.04	0.215	0.210	0.91
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	10	1	0.72	0.580	0.669	0.80	0.57	0.359	0.319	0.63	0.30	0.310	0.307	0.84
		2	0.01	0.217	0.276	0.89	-0.02	0.123	0.139	0.98	0.01	0.182	0.188	0.90
		3	0.04	0.215	0.230	0.93	0.01	0.114	0.125	0.95	0.04	0.219	0.216	0.91
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91

Table 40: Simulation results for post-SCAD selection with non-fixed personal characteristics under binary endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.36	2.452	1.387	0.88	0.53	0.506	0.440	0.81	0.22	0.287	0.352	0.92
		2.1	-0.23	0.177	0.253	0.83	-0.06	0.256	0.239	0.93	-0.10	0.163	0.193	0.86
		2.2	0.22	0.384	0.368	0.91	0.12	0.183	0.177	0.90	0.10	0.224	0.228	0.93
		2.3	0.03	0.619	0.519	0.96	0.06	0.244	0.216	0.90	0.04	0.331	0.230	0.94
		3	0.03	0.208	0.216	0.92	0.02	0.203	0.131	0.93	0.03	0.203	0.210	0.89
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	5	1	2.64	13.017	2.078	0.93	0.91	1.620	1.597	0.89	0.33	0.422	0.572	0.94
		2.1	-0.27	0.379	0.285	0.73	-0.06	0.363	0.299	0.94	-0.09	0.199	0.196	0.87
		2.2	-0.41	5.638	0.414	0.94	0.15	0.275	0.273	0.93	0.11	0.268	0.292	0.93
		2.3	0.05	1.161	1.158	0.94	0.00	0.247	0.655	0.96	0.02	0.272	0.320	0.94
		3	0.02	0.241	0.229	0.92	0.00	0.131	0.133	0.95	0.03	0.251	0.208	0.93
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	10	1	-0.03	14.375	15.977	0.91	1.42	1.773	5.189	0.89	0.48	0.583	0.837	0.95
		2.1	-0.43	0.395	2.591	0.77	0.00	3.300	0.490	0.85	-0.16	0.186	0.317	0.84
		2.2	0.35	0.789	1.628	0.94	0.23	0.456	0.311	0.94	0.15	0.306	0.323	0.92
		2.3	-0.09	0.583	6.039	0.92	-0.24	1.618	1.223	0.95	0.01	0.346	0.595	0.93
		3	0.03	0.239	0.274	0.93	0.00	0.129	0.143	0.95	0.02	0.225	0.230	0.94
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
Random	2	1	0.56	0.732	2.096	0.87	0.46	0.450	0.433	0.82	0.21	0.293	0.335	0.89
		2.1	-0.17	0.285	0.837	0.82	-0.13	0.161	0.181	0.83	-0.09	0.167	0.314	0.88
		2.2	0.19	0.302	0.358	0.92	0.11	0.184	0.181	0.91	0.11	0.226	0.227	0.98
		2.3	0.05	0.330	0.602	0.92	0.02	0.186	0.247	0.90	0.03	0.212	0.262	0.93
		3	3.15	31.379	0.243	0.96	0.01	0.159	0.139	0.96	-0.05	0.647	0.240	0.97
		4	0.05	0.208	0.259	0.93	0.00	0.097	0.133	0.96	0.05	0.208	0.259	0.93
	5	1	-1.01	13.643	1.743	0.91	0.43	0.429	0.487	0.86	0.15	0.281	0.314	0.92
		2.1	-0.26	0.203	0.285	0.70	-0.13	0.231	0.172	0.72	-0.14	0.147	0.170	0.82
		2.2	0.13	0.277	0.341	0.94	0.08	0.143	0.183	0.96	0.05	0.200	0.216	0.95
		2.3	-0.13	0.666	0.637	0.94	0.00	0.221	0.306	0.89	-0.02	0.200	0.250	0.94
		3	-0.01	0.207	0.236	0.92	-0.01	0.122	0.136	0.92	-0.01	0.186	0.198	0.94
		4	0.02	0.247	0.255	0.92	0.00	0.124	0.136	0.93	0.02	0.247	0.255	0.92
	10	1	-11.09	116.219	4.708	0.91	0.59	0.582	1.403	0.80	0.28	0.352	0.381	0.86
		2.1	-0.23	0.298	0.352	0.74	-0.17	0.186	0.185	0.74	-0.11	0.163	0.176	0.80
		2.2	0.22	0.408	0.349	0.90	0.13	0.181	0.203	0.89	0.11	0.208	0.228	0.95
		2.3	0.03	0.333	1.295	0.91	0.02	0.211	0.712	0.93	0.02	0.192	0.231	0.97
		3	0.04	0.232	0.256	0.92	0.02	0.157	0.138	0.89	0.03	0.209	0.215	0.94
		4	0.06	0.265	0.245	0.92	0.01	0.144	0.132	0.89	0.06	0.265	0.245	0.92
Decreasing	2	1	0.55	0.556	0.915	0.90	0.48	0.380	0.443	0.77	0.24	0.314	0.327	0.90
		2.1	-0.24	0.189	0.310	0.77	-0.11	0.217	0.181	0.87	-0.10	0.175	0.164	0.81
		2.2	0.20	0.321	0.364	0.94	0.13	0.176	0.182	0.91	0.11	0.220	0.233	0.92
		2.3	0.20	2.120	0.521	0.93	0.02	0.168	0.266	0.97	0.04	0.270	0.232	0.89
		3	0.03	0.220	0.266	0.90	0.02	0.135	0.131	0.92	0.04	0.227	0.399	0.88
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	5	1	0.61	0.661	1.555	0.89	0.55	0.424	0.544	0.81	0.26	0.335	0.396	0.87
		2.1	-0.26	0.207	0.310	0.78	-0.13	0.208	0.197	0.92	-0.13	0.163	0.169	0.81
		2.2	0.21	0.334	0.431	0.91	0.13	0.184	0.193	0.92	0.12	0.236	0.238	0.94
		2.3	-0.18	2.140	0.758	0.90	0.01	0.207	0.349	0.98	0.04	0.331	0.260	0.95
		3	0.03	0.234	0.663	0.92	0.01	0.133	0.133	0.95	0.04	0.226	0.506	0.90
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	10	1	0.64	0.674	2.585	0.89	0.66	1.038	0.564	0.85	0.26	0.322	0.404	0.90
		2.1	-0.28	0.195	0.255	0.74	-0.13	0.229	0.197	0.86	-0.12	0.172	0.164	0.81
		2.2	0.23	0.333	0.456	0.91	0.14	0.196	0.197	0.88	0.12	0.236	0.245	0.93
		2.3	-0.10	1.219	0.507	0.91	0.01	0.222	0.448	0.96	0.03	0.291	0.269	0.96
		3	0.03	0.233	0.232	0.91	0.01	0.140	0.134	0.92	0.04	0.232	0.758	0.89
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91

Table 41: Simulation results for post-SCAD selection with fixed personal characteristics under binary endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.27	2.642	1.800	0.89	0.41	1.316	0.670	0.79	0.29	0.396	0.421	0.87
		2.1	-0.21	0.179	0.207	0.72	-0.16	0.139	0.140	0.73	-0.09	0.162	0.190	0.83
		2.2	0.22	0.354	0.363	0.93	0.13	0.195	0.181	0.90	0.12	0.226	0.231	0.94
		2.3	-0.06	1.358	0.734	0.90	0.04	0.651	0.314	0.92	0.09	0.285	0.282	0.87
		3	0.04	0.226	0.235	0.91	0.01	0.127	0.130	0.93	0.04	0.211	0.216	0.91
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	5	1	0.94	13.501	4.728	0.89	0.97	1.654	1.672	0.84	0.39	0.513	0.660	0.92
		2.1	-0.27	0.292	0.328	0.78	-0.17	0.162	0.176	0.86	-0.10	0.173	0.176	0.82
		2.2	0.70	5.086	0.574	0.95	0.17	0.597	0.240	0.94	0.14	0.294	0.285	0.95
		2.3	0.53	3.058	1.573	0.91	0.15	0.621	0.465	0.89	0.05	0.321	0.361	0.94
		3	0.03	0.227	0.247	0.94	0.00	0.154	0.133	0.96	0.05	0.281	0.315	0.94
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	10	1	2.62	10.651	19.600	0.90	1.12	1.591	3.408	0.91	0.79	1.888	1.062	0.88
		2.1	-0.46	0.625	2.274	0.85	-0.28	0.313	0.287	0.87	-0.16	0.193	0.246	0.81
		2.2	0.30	0.566	1.160	0.94	0.20	0.365	0.288	0.90	0.15	0.273	0.340	0.94
		2.3	0.24	1.639	4.222	0.93	0.04	0.378	1.005	0.94	0.13	0.768	0.489	0.94
		3	0.03	0.240	0.336	0.94	0.00	0.125	0.141	0.97	0.03	0.223	0.359	0.94
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
Random	2	1	0.62	0.836	0.985	0.89	0.41	0.367	0.437	0.87	0.22	0.284	0.311	0.94
		2.1	-0.22	0.200	0.259	0.74	-0.17	0.125	0.131	0.70	-0.10	0.143	0.182	0.86
		2.2	0.11	0.833	0.338	0.93	0.12	0.189	0.183	0.89	0.12	0.289	0.231	0.95
		2.3	0.13	0.429	0.416	0.93	0.03	0.165	0.195	0.94	0.04	0.196	0.222	0.98
		3	0.01	0.257	0.282	0.95	0.02	0.263	0.140	0.93	0.01	0.294	0.249	0.95
		4	0.05	0.208	0.259	0.93	0.00	0.097	0.133	0.96	0.05	0.208	0.259	0.93
	5	1	0.46	0.468	3.239	0.94	0.43	0.391	0.474	0.86	0.18	0.273	0.338	0.96
		2.1	-0.25	0.172	0.275	0.64	-0.18	0.126	0.137	0.71	-0.13	0.143	0.151	0.77
		2.2	0.14	0.308	0.352	0.95	0.09	0.150	0.187	0.90	0.06	0.202	0.222	0.94
		2.3	0.00	0.214	1.086	0.95	0.00	0.159	0.238	0.93	0.00	0.182	0.231	0.97
		3	0.00	0.205	0.313	0.94	-0.01	0.121	0.138	0.94	0.00	0.204	0.218	0.94
		4	0.02	0.247	0.255	0.92	0.00	0.124	0.136	0.93	0.02	0.247	0.255	0.92
	10	1	0.88	0.978	2.938	0.89	0.57	0.539	0.767	0.85	0.31	0.385	0.407	0.88
		2.1	-0.25	0.232	0.262	0.69	-0.17	0.171	0.158	0.68	-0.10	0.162	0.167	0.81
		2.2	0.22	0.326	0.355	0.89	0.13	0.187	0.194	0.90	0.12	0.207	0.233	0.91
		2.3	0.11	0.365	1.028	0.93	0.03	0.231	0.290	0.93	0.06	0.240	0.257	0.93
		3	0.07	0.298	0.263	0.94	0.02	0.169	0.137	0.90	0.05	0.241	0.230	0.93
		4	0.06	0.265	0.245	0.92	0.01	0.144	0.132	0.89	0.06	0.265	0.245	0.92
Decreasing	2	1	0.72	0.917	1.247	0.89	0.53	0.443	0.436	0.78	0.27	0.339	0.377	0.88
		2.1	-0.24	0.228	0.217	0.71	-0.17	0.119	0.161	0.80	-0.11	0.154	0.156	0.82
		2.2	0.24	0.337	0.372	0.91	0.14	0.178	0.183	0.89	0.13	0.228	0.234	0.93
		2.3	0.16	0.511	0.583	0.93	0.08	0.251	0.225	0.91	0.07	0.242	0.258	0.91
		3	0.05	0.225	0.262	0.92	0.02	0.126	0.132	0.94	0.05	0.226	0.646	0.92
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	5	1	3.36	25.498	1.978	0.90	0.62	0.550	1.052	0.75	0.34	0.442	0.421	0.90
		2.1	-0.27	0.218	0.233	0.68	-0.19	0.131	0.150	0.71	-0.13	0.157	0.159	0.78
		2.2	0.26	0.377	0.492	0.93	0.13	0.178	0.193	0.89	0.13	0.244	0.245	0.93
		2.3	1.26	10.901	0.848	0.92	0.09	0.318	0.523	0.89	0.10	0.324	0.279	0.89
		3	0.05	0.246	0.322	0.92	0.01	0.135	0.130	0.92	0.04	0.228	0.501	0.92
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	10	1	0.92	2.425	7.479	0.89	0.63	0.510	0.730	0.73	0.36	0.435	0.571	0.85
		2.1	-0.28	0.261	0.236	0.69	-0.21	0.144	0.162	0.67	-0.12	0.160	0.159	0.79
		2.2	0.28	0.403	0.410	0.89	0.16	0.204	0.199	0.90	0.14	0.244	0.252	0.93
		2.3	0.24	1.417	2.477	0.95	0.07	0.230	0.340	0.89	0.10	0.302	0.370	0.87
		3	0.05	0.232	0.293	0.92	0.01	0.132	0.133	0.91	0.05	0.232	0.836	0.91
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91

Table 42: Simulation results for direct-SCAD selection with non-fixed personal characteristics under time-to-event endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.52	0.941	0.643	0.91	0.45	0.578	0.420	0.84	0.19	0.379	0.408	0.95
		2	-0.04	0.237	0.283	0.95	-0.05	0.195	0.184	0.95	-0.03	0.240	0.262	0.93
		3	0.00	0.275	0.300	0.96	-0.01	0.179	0.188	0.97	0.01	0.262	0.307	0.96
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	5	1	0.73	1.848	1.324	0.87	0.91	1.238	1.124	0.89	0.37	1.066	0.566	0.97
		2	-0.10	0.322	0.405	0.92	-0.10	0.226	0.210	0.91	-0.04	0.345	0.277	0.92
		3	-0.02	0.282	0.317	0.94	-0.03	0.157	0.195	0.97	0.01	0.329	0.330	0.95
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	10	1	1.75	2.790	3.470	0.90	2.78	4.566	4.244	0.87	0.56	0.959	0.940	0.93
		2	-0.14	0.636	0.684	0.92	-0.12	0.343	0.462	0.86	-0.08	0.268	0.316	0.92
		3	-0.05	0.246	0.293	0.94	-0.02	0.164	0.202	0.96	-0.02	0.263	0.337	0.96
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
Random	2	1	0.42	0.555	0.629	0.90	0.42	0.392	0.408	0.88	0.15	0.349	0.380	0.95
		2	-0.05	0.244	0.270	0.95	-0.01	0.189	0.193	0.95	-0.04	0.233	0.248	0.94
		3	-0.04	0.223	0.274	0.96	0.00	0.189	0.184	0.96	-0.04	0.231	0.267	0.95
		4	-0.01	0.258	0.324	0.97	0.00	0.177	0.191	0.96	-0.01	0.258	0.324	0.97
	5	1	0.54	0.670	0.755	0.92	0.46	0.369	0.453	0.91	0.19	0.331	0.413	0.98
		2	-0.06	0.221	0.293	0.93	-0.01	0.169	0.197	0.96	-0.02	0.253	0.265	0.98
		3	-0.04	0.210	0.281	0.95	0.00	0.152	0.197	0.96	-0.03	0.217	0.283	0.94
		4	0.01	0.250	0.323	0.97	0.01	0.161	0.206	0.98	0.01	0.250	0.323	0.97
	10	1	0.70	0.874	0.936	0.87	0.53	0.564	0.496	0.81	0.28	0.464	0.427	0.88
		2	-0.04	0.273	0.347	0.89	-0.05	0.173	0.188	0.94	0.00	0.265	0.249	0.90
		3	0.01	0.281	0.270	0.94	-0.01	0.172	0.185	0.93	0.01	0.258	0.269	0.94
		4	0.05	0.309	0.330	0.95	-0.01	0.173	0.188	0.93	0.05	0.309	0.330	0.95
Decreasing	2	1	0.49	0.591	0.654	0.91	0.46	0.379	0.423	0.83	0.22	0.406	0.413	0.93
		2	-0.02	0.273	0.301	0.96	-0.02	0.179	0.189	0.95	-0.01	0.263	0.269	0.94
		3	0.01	0.299	0.312	0.93	-0.02	0.155	0.186	0.97	0.02	0.302	0.322	0.93
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	5	1	0.55	0.635	0.763	0.91	0.54	0.420	0.476	0.80	0.25	0.426	0.441	0.91
		2	-0.03	0.268	0.325	0.96	-0.03	0.185	0.192	0.93	-0.01	0.264	0.274	0.94
		3	0.00	0.287	0.310	0.93	-0.02	0.153	0.187	0.96	0.02	0.303	0.318	0.93
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	10	1	0.59	0.671	0.772	0.90	0.57	0.435	0.500	0.80	0.26	0.432	0.454	0.91
		2	-0.02	0.274	0.314	0.96	-0.03	0.185	0.195	0.94	-0.01	0.264	0.274	0.93
		3	0.00	0.286	0.307	0.93	-0.02	0.152	0.189	0.96	0.02	0.301	0.319	0.93
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94

Table 43: Simulation results for direct-SCAD selection with fixed personal characteristics under time-to-event endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.53	1.219	0.715	0.94	0.39	0.463	0.384	0.89	0.17	0.359	0.398	0.94
		2	-0.02	0.356	0.277	0.95	-0.04	0.179	0.180	0.95	-0.04	0.230	0.252	0.94
		3	-0.02	0.248	0.306	0.97	-0.01	0.176	0.185	0.96	-0.02	0.266	0.290	0.96
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	5	1	1.00	2.047	2.359	0.91	1.48	8.942	0.734	0.91	0.37	1.263	0.538	0.95
		2	-0.12	0.255	0.387	0.93	-0.07	0.257	0.201	0.93	-0.06	0.326	0.262	0.91
		3	-0.02	0.259	0.317	0.97	-0.03	0.156	0.193	0.98	-0.01	0.331	0.319	0.96
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	10	1	3.15	5.880	8.839	0.90	1.56	2.116	2.040	0.82	0.55	0.927	0.971	0.92
		2	-8.89	62.117	1.686	0.88	-0.12	0.239	0.357	0.88	-0.12	0.230	0.297	0.93
		3	-0.02	0.272	0.338	0.95	-0.02	0.165	0.197	0.95	-0.02	0.272	0.312	0.95
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
Random	2	1	0.42	0.572	0.623	0.93	0.39	0.371	0.380	0.86	0.16	0.357	0.383	0.93
		2	-0.03	0.281	0.281	0.95	-0.02	0.179	0.190	0.96	-0.03	0.242	0.257	0.94
		3	-0.25	2.306	0.299	0.97	0.00	0.190	0.185	0.95	-0.18	1.627	0.284	0.96
		4	-0.01	0.258	0.324	0.97	0.00	0.177	0.191	0.96	-0.01	0.258	0.324	0.97
	5	1	0.52	0.618	0.742	0.89	0.43	0.356	0.424	0.91	0.20	0.339	0.421	0.97
		2	-0.03	0.231	0.300	0.97	-0.02	0.162	0.196	0.98	-0.02	0.221	0.265	0.98
		3	-0.01	0.226	0.311	0.97	0.00	0.149	0.197	0.96	-0.01	0.221	0.299	0.96
		4	0.01	0.250	0.323	0.97	0.01	0.161	0.206	0.98	0.01	0.250	0.323	0.97
	10	1	0.81	1.084	0.923	0.87	0.51	0.499	0.471	0.81	0.29	0.459	0.446	0.89
		2	-0.01	0.291	0.310	0.94	-0.06	0.162	0.185	0.95	0.01	0.265	0.259	0.92
		3	0.04	0.328	0.309	0.95	-0.01	0.171	0.185	0.93	0.03	0.292	0.294	0.96
		4	0.05	0.309	0.330	0.95	-0.01	0.173	0.188	0.93	0.05	0.309	0.330	0.95
Decreasing	2	1	0.49	0.595	0.641	0.90	0.41	0.354	0.389	0.83	0.19	0.376	0.396	0.92
		2	-0.02	0.266	0.287	0.93	-0.03	0.171	0.184	0.93	-0.02	0.250	0.258	0.93
		3	-0.01	0.267	0.301	0.93	-0.02	0.153	0.185	0.96	-0.01	0.264	0.299	0.93
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	5	1	0.58	0.660	0.727	0.89	0.49	0.402	0.432	0.80	0.22	0.398	0.423	0.92
		2	-0.03	0.272	0.312	0.92	-0.04	0.174	0.189	0.93	-0.02	0.252	0.261	0.92
		3	-0.01	0.265	0.303	0.93	-0.02	0.154	0.187	0.96	-0.01	0.264	0.294	0.94
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	10	1	0.62	0.695	0.823	0.89	0.52	0.416	0.452	0.80	0.24	0.405	0.434	0.92
		2	-0.02	0.272	0.316	0.93	-0.03	0.178	0.190	0.93	-0.02	0.253	0.262	0.93
		3	-0.01	0.270	0.304	0.92	-0.02	0.154	0.188	0.96	-0.01	0.267	0.297	0.94
		4	-0.01	0.280	0.355	0.94	-0.01924	0.158	0.194	0.95	-0.01	0.280	0.355	0.94

Table 44: Simulation results for post-SCAD selection with non-fixed personal characteristics under time-to-event endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.76	3.202	2.103	0.92	0.47	0.576	0.526	0.86	0.17	0.405	0.439	0.92
		2.1	-0.25	0.187	0.421	0.75	-0.06	0.339	0.284	0.91	-0.12	0.203	0.226	0.86
		2.2	0.18	0.590	0.446	0.93	0.10	0.277	0.246	0.93	0.06	0.304	0.312	0.95
		2.3	0.09	0.605	0.602	0.94	0.01	0.287	0.276	0.94	0.00	0.306	0.325	0.93
		3	-0.02	0.260	0.293	0.94	0.01	0.301	0.193	0.97	-0.02	0.247	0.292	0.96
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	5	1	0.80	1.438	2.574	0.91	0.82	1.209	1.697	0.93	0.26	0.521	0.657	0.94
		2.1	-0.27	0.290	0.279	0.74	-0.06	0.493	0.339	0.90	-0.12	0.276	0.269	0.88
		2.2	-0.30	3.900	0.527	0.94	0.14	0.430	0.486	0.93	0.07	0.419	0.363	0.95
		2.3	-0.10	0.782	2.144	0.93	-0.07	2.098	0.720	0.92	-0.01	0.332	0.470	0.95
		3	-0.02	0.346	0.327	0.96	-0.02	0.191	0.193	0.97	-0.01	0.380	0.309	0.97
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	10	1	1.14	2.148	16.145	0.96	1.27	2.461	6.402	0.91	0.42	0.774	1.518	0.92
		2.1	-0.41	0.532	0.622	0.80	0.15	4.712	0.523	0.88	-0.19	0.214	0.303	0.88
		2.2	0.32	1.334	1.225	0.96	0.20	0.661	0.381	0.95	0.10	0.441	0.402	0.94
		2.3	-0.13	0.842	10.394	0.96	-0.19	2.013	1.073	0.93	0.00	0.493	0.646	0.92
		3	-0.03	0.282	0.340	0.95	-0.02	0.174	0.203	0.96	-0.02	0.272	0.304	0.95
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
Random	2	1	-0.16	5.530	3.078	0.94	0.43	0.741	0.543	0.87	0.16	0.411	0.444	0.95
		2.1	-0.17	0.320	1.417	0.82	-0.13	0.190	0.220	0.85	-0.11	0.211	0.288	0.92
		2.2	0.13	0.370	0.432	0.94	0.11	0.255	0.254	0.94	0.06	0.305	0.318	0.92
		2.3	0.03	0.444	0.690	0.95	0.03	0.254	0.291	0.93	-0.01	0.275	0.337	0.95
		3	2.55	25.860	0.321	0.96	0.01	0.212	0.192	0.96	-0.08	0.569	0.298	0.96
		4	-0.01	0.258	0.324	0.97	0.00	0.177	0.191	0.96	-0.01	0.258	0.324	0.97
	5	1	0.52	0.641	2.078	0.93	0.46	0.500	0.612	0.92	0.19	0.372	0.459	0.95
		2.1	-0.27	0.213	0.319	0.83	-0.13	0.285	0.213	0.82	-0.13	0.186	0.254	0.90
		2.2	0.14	0.337	0.489	0.96	0.10	0.206	0.274	0.97	0.06	0.261	0.348	1.00
		2.3	0.04	0.483	0.735	0.96	0.01	0.225	0.377	0.98	0.00	0.229	0.394	0.99
		3	-0.02	0.226	0.362	0.96	0.00	0.160	0.203	0.96	-0.01	0.225	0.308	0.97
		4	0.01	0.250	0.323	0.97	0.01	0.161	0.206	0.98	0.01	0.250	0.323	0.97
	10	1	1.27	3.524	5.354	0.90	0.25	3.285	1.265	0.94	0.32	0.559	0.598	0.93
		2.1	-0.22	0.326	0.422	0.68	-0.18	0.200	0.204	0.79	-0.09	0.245	0.225	0.81
		2.2	0.23	0.513	0.483	0.93	0.11	0.260	0.267	0.92	0.11	0.333	0.327	0.92
		2.3	0.16	1.085	1.196	0.91	-0.11	1.113	1.551	0.95	0.01	0.288	0.318	0.92
		3	0.02	0.273	0.325	0.94	0.00	0.183	0.192	0.95	0.02	0.267	0.296	0.95
		4	0.05	0.309	0.330	0.95	-0.01	0.173	0.188	0.93	0.05	0.309	0.330	0.95
Decreasing	2	1	0.44	0.650	1.072	0.93	0.45	0.441	0.557	0.85	0.18	0.419	0.442	0.93
		2.1	-0.25	0.186	0.306	0.76	-0.12	0.237	0.208	0.83	-0.13	0.206	0.208	0.85
		2.2	0.16	0.410	0.454	0.95	0.11	0.273	0.250	0.92	0.07	0.315	0.318	0.95
		2.3	0.02	0.350	1.006	0.95	0.02	0.251	0.340	0.95	-0.01	0.288	0.315	0.93
		3	-0.02	0.258	0.345	0.97	-0.01	0.170	0.194	0.96	-0.01	0.256	0.371	0.94
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	5	1	0.54	1.081	2.390	0.90	0.52	0.537	0.637	0.90	0.22	0.495	0.484	0.91
		2.1	-0.27	0.228	0.281	0.75	-0.15	0.204	0.224	0.86	-0.15	0.182	0.211	0.83
		2.2	0.16	0.411	0.480	0.95	0.11	0.269	0.262	0.93	0.06	0.320	0.328	0.95
		2.3	0.03	0.391	0.871	0.92	0.05	0.312	0.403	0.94	0.00	0.293	0.336	0.94
		3	-0.02	0.259	0.854	0.96	-0.02	0.158	0.195	0.97	-0.02	0.256	0.605	0.96
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	10	1	0.64	1.334	2.174	0.87	0.22	3.825	0.689	0.86	0.25	0.555	0.787	0.93
		2.1	-0.30	0.193	0.352	0.76	-0.15	0.240	0.222	0.80	-0.15	0.201	0.204	0.82
		2.2	0.18	0.446	0.514	0.94	0.11	0.275	0.263	0.95	0.07	0.322	0.334	0.95
		2.3	0.03	0.400	0.604	0.91	0.06	0.340	0.461	0.90	0.00	0.293	0.372	0.95
		3	-0.03	0.255	0.309	0.96	-0.01	0.168	0.196	0.94	-0.02	0.262	0.703	0.96
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94

Table 45: Simulation results for post-SCAD selection with fixed personal characteristics under time-to-event endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.49	0.940	1.289	0.92	0.46	0.591	0.514	0.86	0.18	0.420	0.443	0.95
		2.1	-0.23	0.180	0.227	0.78	-0.18	0.154	0.166	0.74	-0.12	0.201	0.208	0.87
		2.2	0.15	0.393	0.470	0.94	0.11	0.293	0.251	0.93	0.07	0.308	0.321	0.95
		2.3	0.08	0.500	2.590	0.93	0.01	0.268	0.670	0.93	0.02	0.329	1.304	0.94
		3	-0.01	0.259	0.331	0.97	-0.01	0.171	0.194	0.96	0.00	0.262	0.315	0.97
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	5	1	0.46	2.725	11.428	0.93	0.42	2.520	3.700	0.93	0.25	0.732	0.667	0.91
		2.1	-0.29	0.312	0.309	0.71	-0.18	0.175	0.201	0.79	-0.13	0.237	0.224	0.86
		2.2	0.46	3.503	0.759	0.95	0.17	0.884	0.406	0.93	0.10	0.464	0.361	0.96
		2.3	0.41	4.576	2.323	0.91	-0.21	1.818	1.132	0.87	-0.01	0.315	0.650	0.94
		3	-0.01	0.322	0.352	0.96	-0.01	0.229	0.195	0.98	0.01	0.450	0.462	0.97
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	10	1	1.12	2.994	8.838	0.94	1.08	3.364	3.596	0.88	0.44	0.690	1.498	0.90
		2.1	-0.43	0.773	0.858	0.83	-0.27	0.319	0.307	0.78	-0.19	0.208	0.268	0.86
		2.2	0.16	0.628	1.148	0.96	0.18	0.534	0.357	0.95	0.09	0.334	0.436	0.95
		2.3	0.05	0.629	2.045	0.95	-0.01	0.893	2.983	0.93	0.10	1.633	1.124	0.95
		3	-0.02	0.276	0.412	0.94	-0.02	0.174	0.200	0.95	-0.02	0.269	0.357	0.95
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
Random	2	1	0.45	0.641	2.098	0.97	0.44	0.717	0.674	0.87	0.20	0.614	0.432	0.96
		2.1	-0.25	0.191	0.254	0.72	-0.18	0.147	0.161	0.77	-0.13	0.198	0.227	0.86
		2.2	0.06	0.742	0.431	0.96	0.11	0.263	0.258	0.91	0.07	0.341	0.323	0.91
		2.3	0.06	0.429	0.839	0.94	0.03	0.243	0.321	0.95	0.01	0.286	0.362	0.94
		3	-0.03	0.288	0.348	0.96	0.02	0.294	0.191	0.96	-0.03	0.313	0.309	0.95
		4	-0.01	0.258	0.324	0.97	0.00	0.177	0.191	0.96	-0.01	0.258	0.324	0.97
	5	1	0.64	1.151	3.383	0.92	0.46	0.497	0.657	0.90	0.21	0.381	0.476	0.97
		2.1	-0.24	0.183	0.433	0.77	-0.18	0.150	0.181	0.75	-0.12	0.181	0.222	0.91
		2.2	0.15	0.342	0.504	0.97	0.11	0.214	0.278	0.97	0.07	0.268	0.351	0.99
		2.3	0.06	0.409	0.811	0.97	0.02	0.231	0.327	0.98	0.02	0.229	0.364	0.99
		3	-0.01	0.230	0.476	0.97	0.00	0.162	0.205	0.96	-0.01	0.231	0.324	0.95
		4	0.01	0.250	0.323	0.97	0.01	0.161	0.206	0.98	0.01	0.250	0.323	0.97
	10	1	0.96	1.940	4.993	0.92	0.84	2.196	1.202	0.86	0.35	0.762	0.525	0.95
		2.1	-0.24	0.247	0.300	0.73	-0.19	0.171	0.175	0.78	-0.10	0.216	0.217	0.84
		2.2	0.25	0.480	0.465	0.88	0.11	0.256	0.262	0.94	0.12	0.339	0.337	0.93
		2.3	0.24	1.034	1.989	0.93	0.04	0.348	0.374	0.92	0.08	0.442	0.338	0.92
		3	0.05	0.313	0.346	0.94	0.00	0.192	0.192	0.94	0.04	0.289	0.318	0.95
		4	0.05	0.309	0.330	0.95	-0.01	0.173	0.188	0.93	0.05	0.309	0.330	0.95
Decreasing	2	1	0.49	0.970	1.139	0.92	0.50	0.541	0.756	0.87	0.20	0.432	0.459	0.95
		2.1	-0.26	0.291	0.248	0.76	-0.18	0.132	0.161	0.72	-0.14	0.183	0.202	0.85
		2.2	0.18	0.419	0.460	0.94	0.11	0.268	0.253	0.95	0.08	0.321	0.326	0.94
		2.3	0.07	0.387	1.354	0.92	0.03	0.216	0.284	0.95	0.02	0.296	0.340	0.94
		3	0.00	0.275	0.355	0.95	-0.01	0.166	0.195	0.97	0.00	0.257	0.429	0.94
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	5	1	0.68	1.095	1.853	0.91	0.46	0.818	0.648	0.86	0.26	0.607	0.472	0.93
		2.1	-0.28	0.217	0.276	0.71	-0.20	0.150	0.168	0.71	-0.15	0.180	0.204	0.83
		2.2	0.20	0.485	0.519	0.96	0.11	0.259	0.260	0.94	0.08	0.318	0.336	0.94
		2.3	0.08	0.426	0.885	0.91	0.03	0.244	0.350	0.94	0.01	0.293	0.354	0.95
		3	0.00	0.277	0.380	0.97	-0.01	0.165	0.193	0.95	-0.01	0.256	0.606	0.95
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	10	1	0.52	1.644	2.553	0.92	0.61	0.660	0.785	0.85	0.24	0.482	0.584	0.93
		2.1	-0.30	0.278	0.260	0.70	-0.22	0.152	0.191	0.68	-0.16	0.182	0.202	0.84
		2.2	0.22	0.510	0.485	0.92	0.13	0.273	0.267	0.94	0.08	0.317	0.341	0.93
		2.3	0.05	0.594	2.005	0.91	0.07	0.386	0.360	0.95	0.02	0.303	0.411	0.93
		3	0.00	0.274	0.378	0.96	-0.02	0.162	0.196	0.95	-0.01	0.257	1.088	0.95
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94

Table 46: Simulation results for direct-RF selection with non-fixed personal characteristics under continuous endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	13.75	83.708	49.914	0.89	2.65	1.615	1.764	0.67	5.07	8.947	20.723	0.82
		2	0.28	4.015	5.307	0.90	0.32	0.510	0.674	0.91	2.92	22.900	19.597	0.87
		3	2.08	1.613	1.599	0.74	1.07	0.643	0.581	0.66	2.09	1.429	2.714	0.71
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	5	1	1.29	12.309	5.297	0.89	2.17	1.158	1.071	0.51	1.91	1.510	1.820	0.73
		2	0.11	2.915	0.878	0.93	0.02	0.238	0.283	0.94	0.56	1.181	0.703	0.91
		3	1.27	0.889	0.861	0.72	0.80	0.492	0.418	0.59	1.09	0.783	0.709	0.68
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	10	1	2.99	4.391	4.585	0.78	2.63	1.265	1.633	0.54	1.33	0.870	0.834	0.66
		2	-0.02	0.666	0.596	0.93	-0.10	0.250	0.338	0.96	0.12	0.308	0.318	0.95
		3	1.24	0.923	0.835	0.68	0.77	0.512	0.477	0.66	0.78	0.572	0.495	0.72
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
Random	2	1	16.07	394.761	62.468	0.89	-21.41	138.365	106.093	0.80	4.38	72.415	141.059	0.93
		2	-911.48	9676.152	11.556	0.91	4.11	35.568	24.707	0.85	-15.06	116.688	46.685	0.89
		3	3.60	9.578	23.215	0.90	5.77	12.644	25.224	0.77	-3.92	68.685	41.728	0.87
		4	0.05	0.252	0.245	0.89	0.02	0.127	0.111	0.92	0.05	0.252	0.245	0.89
	5	1	-4.79	39.379	125.926	0.93	14.04	114.374	132.458	0.90	3.68	47.631	168.435	0.91
		2	-6.73	102.610	35.202	0.92	-0.78	8.158	48.380	0.94	12.43	323.695	49.053	0.89
		3	8.14	58.304	33.231	0.92	7.31	20.757	39.956	0.75	-0.92	48.350	37.917	0.83
		4	0.05	0.241	0.207	0.86	0.02	0.125	0.119	0.92	0.05	0.241	0.207	0.86
	10	1	-0.11	61.604	84.512	0.89	87.33	545.204	72.314	0.81	37.02	346.778	199.022	0.88
		2	0.56	14.928	20.384	0.91	-1.38	31.678	596.967	0.86	-2.68	61.533	48.963	0.90
		3	3.41	8.619	15.703	0.89	3.92	5.766	116.173	0.73	0.96	17.661	58.359	0.82
		4	0.05	0.232	0.195	0.92	0.02	0.124	0.109	0.88	0.05	0.232	0.195	0.92
Decreasing	2	1	-13.64	240.522	47.341	0.95	3.28	3.028	3.094	0.78	12.37	49.144	81.954	0.88
		2	1.06	5.112	10.203	0.92	0.55	0.798	0.758	0.88	2.96	6.030	23.782	0.87
		3	1.02	11.224	3.359	0.75	1.22	0.646	0.714	0.56	1.66	7.590	6.387	0.79
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	5	1	4.12	15.492	41.684	0.92	2.34	1.490	1.800	0.78	2.47	10.550	11.472	0.90
		2	-0.53	12.867	9.951	0.88	0.18	0.395	0.488	0.92	1.50	2.714	11.044	0.91
		3	1.17	8.021	9.943	0.79	1.15	0.720	0.669	0.63	2.14	3.569	1.741	0.73
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	10	1	4.10	9.681	10.136	0.89	2.50	2.156	1.741	0.72	3.64	5.107	7.213	0.85
		2	0.35	1.708	62.527	0.91	0.15	0.369	0.464	0.92	2.69	16.974	5.411	0.89
		3	1.82	1.624	1.979	0.75	1.03	0.614	0.618	0.63	1.80	1.334	1.352	0.75
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90

Table 47: Simulation results for post-RF selection with non-fixed personal characteristics under continuous endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.55	0.486	0.420	0.79	0.45	0.285	0.244	0.58	0.23	0.238	0.216	0.82
		2	0.05	0.171	0.202	0.97	0.02	0.130	0.137	0.93	0.04	0.150	0.153	0.93
		3	0.04	0.176	0.182	0.92	0.03	0.119	0.109	0.93	0.03	0.164	0.162	0.91
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	5	1	0.78	0.987	0.675	0.78	0.65	0.404	0.392	0.62	0.29	0.295	0.260	0.87
		2	0.14	1.101	0.246	0.96	-0.05	0.141	0.161	0.94	0.05	0.176	0.173	0.92
		3	0.06	0.241	0.214	0.92	0.04	0.160	0.115	0.92	0.05	0.232	0.196	0.91
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	10	1	1.18	1.063	1.462	0.81	1.38	2.559	0.954	0.72	0.67	0.429	0.528	0.79
		2	0.03	0.341	0.402	0.95	-0.05	0.229	0.257	0.91	0.15	0.244	0.286	0.94
		3	0.11	0.450	0.266	0.93	0.05	0.153	0.130	0.93	0.31	1.595	7.903	0.92
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
Random	2	1	0.49	0.487	0.519	0.84	0.41	0.243	0.237	0.64	0.21	0.244	0.209	0.83
		2	0.01	0.282	0.216	0.95	0.03	0.149	0.128	0.92	0.03	0.186	0.147	0.92
		3	0.55	5.254	0.187	0.93	0.03	0.160	0.110	0.90	0.07	0.547	0.175	0.92
		4	0.05	0.252	0.245	0.89	0.02	0.127	0.111	0.92	0.05	0.252	0.245	0.89
	5	1	0.54	0.458	0.492	0.77	0.43	0.271	0.240	0.64	0.21	0.217	0.204	0.84
		2	-0.02	0.194	0.381	0.92	0.00	0.128	0.129	0.92	0.02	0.153	0.148	0.92
		3	0.05	0.223	0.206	0.94	0.03	0.131	0.119	0.92	0.04	0.204	0.177	0.92
		4	0.05	0.241	0.207	0.86	0.02	0.125	0.119	0.92	0.05	0.241	0.207	0.86
	10	1	0.73	1.205	1.237	0.83	0.51	0.346	0.274	0.58	0.26	0.280	0.228	0.82
		2	-0.08	0.280	0.524	0.90	-0.01	0.148	0.138	0.94	0.02	0.180	0.150	0.90
		3	0.07	0.262	0.200	0.89	0.04	0.138	0.114	0.88	0.05	0.212	0.171	0.90
		4	0.05	0.232	0.195	0.92	0.02	0.124	0.109	0.88	0.05	0.232	0.195	0.92
Decreasing	2	1	0.55	0.402	0.406	0.75	0.48	0.314	0.242	0.55	0.23	0.233	0.213	0.81
		2	0.06	0.201	0.203	0.89	0.02	0.138	0.131	0.89	0.04	0.154	0.151	0.92
		3	0.04	0.195	0.210	0.93	0.03	0.126	0.111	0.92	0.04	0.171	0.170	0.94
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	5	1	0.61	0.458	0.487	0.76	0.54	0.327	0.273	0.56	0.25	0.245	0.235	0.82
		2	0.04	0.178	0.218	0.92	0.01	0.136	0.139	0.92	0.04	0.160	0.159	0.92
		3	0.05	0.194	0.380	0.89	0.03	0.128	0.112	0.90	0.04	0.185	0.178	0.93
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90
	10	1	0.67	0.493	0.633	0.74	0.58	0.352	0.300	0.56	0.28	0.244	0.246	0.79
		2	0.06	0.191	0.279	0.95	0.01	0.167	0.143	0.91	0.04	0.166	0.165	0.93
		3	0.07	0.238	0.201	0.92	0.04	0.126	0.114	0.91	0.05	0.242	0.189	0.91
		4	0.04	0.217	0.278	0.90	0.02	0.121	0.112	0.89	0.04	0.217	0.278	0.90

Table 48: Simulation results for direct-RF selection with non-fixed personal characteristics under binary endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	33.03	273.602	61.995	0.88	3.00	2.387	1.817	0.69	56.60	505.605	20.671	0.87
		2	0.34	5.633	5.762	0.90	0.31	0.421	0.662	0.95	1.11	11.717	23.462	0.86
		3	2.33	2.567	1.847	0.77	1.00	0.662	0.624	0.70	2.30	2.055	2.488	0.71
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	5	1	2.35	4.031	6.584	0.90	2.06	1.095	1.192	0.56	1.96	1.775	2.081	0.80
		2	3.14	29.145	0.950	0.97	0.05	0.287	0.295	0.96	0.58	0.827	0.855	0.88
		3	1.33	1.174	1.033	0.78	0.73	0.398	0.473	0.68	1.09	0.835	0.814	0.79
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	10	1	2.57	2.534	5.136	0.85	2.54	1.190	1.688	0.61	1.36	1.042	1.008	0.75
		2	0.04	1.126	0.737	0.94	-0.09	0.282	0.338	0.95	0.13	0.347	0.373	0.93
		3	1.16	0.900	1.092	0.81	0.65	0.484	0.507	0.80	0.84	0.689	0.602	0.80
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
Random	2	1	30.68	357.349	65.180	0.90	0.98	27.909	129.066	0.79	-2.04	75.479	113.112	0.94
		2	0.88	7.825	11.333	0.94	-0.27	6.173	21.101	0.85	-0.13	95.495	49.973	0.89
		3	2.38	21.579	20.772	0.90	0.44	20.100	23.888	0.80	-9.71	132.058	50.378	0.87
		4	0.05	0.208	0.259	0.93	0.00	0.097	0.133	0.96	0.05	0.208	0.259	0.93
	5	1	-2.46	34.511	151.095	0.93	0.77	27.647	107.275	0.93	-22.71	218.174	155.232	0.91
		2	0.72	18.439	29.193	0.91	-0.59	6.882	37.220	0.91	-2.38	20.999	45.157	0.91
		3	-1.00	20.855	30.381	0.90	2.35	22.093	44.985	0.75	4.93	37.426	37.994	0.90
		4	0.02	0.247	0.255	0.92	0.00	0.124	0.136	0.93	0.02	0.247	0.255	0.92
	10	1	-8.90	110.268	73.586	0.90	-8.44	143.818	73.183	0.81	160.87	1255.172	181.844	0.90
		2	-0.95	5.841	18.265	0.94	-2.32	15.466	577.132	0.88	-58.22	531.525	46.320	0.92
		3	5.23	19.693	15.052	0.90	2.84	9.116	116.126	0.73	2.13	13.510	61.641	0.81
		4	0.06	0.265	0.245	0.92	0.01	0.144	0.132	0.89	0.06	0.265	0.245	0.92
Decreasing	2	1	4.26	25.375	54.085	0.91	3.17	3.616	3.047	0.78	6.21	16.774	71.849	0.92
		2	0.69	4.378	9.580	0.93	0.52	0.617	0.812	0.90	-4.30	49.297	20.526	0.93
		3	2.15	1.606	3.691	0.85	1.28	1.076	0.770	0.68	0.86	16.189	7.996	0.83
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	5	1	-2.02	46.173	43.455	0.92	2.32	1.561	1.856	0.71	3.45	3.817	11.415	0.93
		2	0.11	2.763	7.207	0.95	0.20	0.350	0.519	0.95	1.67	2.370	8.920	0.87
		3	2.65	9.169	16.529	0.81	1.04	0.710	0.707	0.70	1.84	1.808	1.979	0.84
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	10	1	10.42	60.651	13.573	0.91	2.60	3.870	1.654	0.74	3.24	2.616	7.806	0.86
		2	0.27	3.182	56.971	0.94	0.11	0.315	0.487	0.96	1.21	3.301	5.442	0.92
		3	1.83	1.383	2.044	0.78	1.10	0.724	0.660	0.64	1.69	1.446	1.567	0.80
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91

Table 49: Simulation results for post-RF selection with non-fixed personal characteristics under binary endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.55	0.515	0.507	0.86	0.43	0.260	0.280	0.75	0.22	0.272	0.277	0.90
		2.1	-0.06	0.187	0.200	0.89	-0.08	0.126	0.127	0.88	-0.03	0.168	0.173	0.92
		2.2	0.04	0.244	0.241	0.95	0.05	0.143	0.156	0.97	0.02	0.185	0.189	0.95
		2.3	0.07	0.247	0.244	0.93	0.01	0.130	0.152	0.97	0.05	0.207	0.198	0.94
		3	0.05	0.235	0.234	0.90	0.02	0.132	0.131	0.96	0.03	0.205	0.206	0.90
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	5	1	0.85	1.357	0.683	0.84	0.62	0.434	0.428	0.73	0.29	0.308	0.331	0.89
		2.1	-0.06	0.259	0.222	0.91	-0.15	0.117	0.157	0.84	-0.03	0.172	0.185	0.91
		2.2	0.08	0.390	0.272	0.93	0.07	0.172	0.199	0.96	0.02	0.179	0.201	0.94
		2.3	0.08	0.413	0.300	0.96	-0.06	0.143	0.169	0.93	0.04	0.198	0.217	0.95
		3	0.07	0.289	0.269	0.93	0.02	0.129	0.138	0.95	0.06	0.273	0.241	0.94
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	10	1	1.24	1.140	1.812	0.88	1.29	1.134	0.974	0.73	0.66	0.532	0.614	0.87
		2.1	-0.12	0.268	0.373	0.88	-0.28	0.283	0.231	0.79	-0.01	0.193	0.249	0.94
		2.2	0.12	0.321	0.563	0.96	0.20	0.316	0.363	0.90	0.09	0.225	0.284	0.96
		2.3	0.01	0.299	0.437	0.98	-0.03	0.513	0.255	0.89	0.16	0.291	0.331	0.92
		3	0.11	0.345	0.327	0.92	0.04	0.151	0.152	0.95	0.25	0.938	5.596	0.94
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
Random	2	1	0.50	0.444	0.591	0.88	0.40	0.270	0.275	0.79	0.21	0.248	0.268	0.90
		2.1	-0.06	0.190	0.234	0.93	-0.07	0.115	0.127	0.90	-0.01	0.166	0.170	0.93
		2.2	0.07	0.239	0.297	0.95	0.04	0.139	0.153	0.97	0.05	0.185	0.194	0.94
		2.3	0.03	0.250	0.260	0.96	0.01	0.120	0.144	0.97	0.03	0.186	0.190	0.97
		3	-0.36	3.954	0.235	0.96	0.02	0.131	0.136	0.95	0.08	0.555	0.230	0.95
		4	0.05	0.208	0.259	0.93	0.00	0.097	0.133	0.96	0.05	0.208	0.259	0.93
	5	1	0.51	0.548	0.540	0.90	0.38	0.248	0.277	0.75	0.17	0.261	0.260	0.92
		2.1	-0.11	0.164	0.189	0.83	-0.09	0.113	0.124	0.88	-0.06	0.148	0.159	0.93
		2.2	0.03	0.205	0.247	0.95	0.03	0.136	0.149	0.94	0.00	0.170	0.182	0.94
		2.3	-0.04	0.219	0.251	0.90	-0.01	0.128	0.145	0.95	-0.01	0.175	0.179	0.93
		3	0.01	0.212	0.225	0.94	0.00	0.124	0.135	0.93	0.00	0.189	0.203	0.93
		4	0.02	0.247	0.255	0.92	0.00	0.124	0.136	0.93	0.02	0.247	0.255	0.92
	10	1	0.74	1.017	1.227	0.82	0.47	0.305	0.306	0.74	0.25	0.270	0.285	0.85
		2.1	-0.08	0.219	0.272	0.89	-0.09	0.130	0.138	0.83	-0.02	0.154	0.170	0.95
		2.2	0.11	0.295	0.378	0.91	0.06	0.150	0.161	0.94	0.05	0.179	0.197	0.94
		2.3	-0.02	0.290	0.405	0.90	-0.04	0.131	0.150	0.89	0.01	0.180	0.189	0.94
		3	0.08	0.275	0.246	0.90	0.04	0.171	0.136	0.87	0.06	0.235	0.215	0.91
		4	0.06	0.265	0.245	0.92	0.01	0.144	0.132	0.89	0.06	0.265	0.245	0.92
Decreasing	2	1	0.56	0.481	0.485	0.85	0.45	0.282	0.274	0.67	0.23	0.268	0.274	0.87
		2.1	-0.06	0.196	0.198	0.88	-0.08	0.114	0.125	0.87	-0.02	0.167	0.172	0.92
		2.2	0.05	0.228	0.237	0.93	0.06	0.153	0.152	0.91	0.02	0.185	0.190	0.93
		2.3	0.05	0.241	0.247	0.95	0.01	0.143	0.148	0.97	0.04	0.190	0.195	0.93
		3	0.06	0.255	0.265	0.91	0.03	0.129	0.131	0.94	0.04	0.211	0.215	0.91
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	5	1	0.62	0.485	0.571	0.86	0.52	0.372	0.314	0.67	0.26	0.276	0.296	0.90
		2.1	-0.07	0.182	0.212	0.89	-0.09	0.145	0.133	0.87	-0.02	0.167	0.177	0.93
		2.2	0.04	0.214	0.256	0.94	0.07	0.175	0.166	0.91	0.02	0.180	0.194	0.95
		2.3	0.05	0.235	0.260	0.94	0.00	0.146	0.154	0.93	0.04	0.201	0.200	0.91
		3	0.07	0.306	0.594	0.90	0.03	0.139	0.135	0.94	0.05	0.250	0.227	0.94
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91
	10	1	0.68	0.551	0.674	0.85	0.59	0.365	0.342	0.65	0.29	0.307	0.312	0.87
		2.1	-0.08	0.194	0.219	0.89	-0.10	0.135	0.139	0.85	-0.02	0.175	0.183	0.90
		2.2	0.06	0.233	0.287	0.93	0.08	0.173	0.175	0.92	0.03	0.196	0.199	0.93
		2.3	0.06	0.257	0.276	0.95	0.00	0.144	0.162	0.97	0.05	0.209	0.206	0.94
		3	0.07	0.269	0.246	0.94	0.03	0.127	0.135	0.92	0.06	0.277	0.243	0.91
		4	0.05	0.245	0.247	0.91	0.01	0.118	0.133	0.95	0.05	0.245	0.247	0.91

Table 50: Simulation results for direct-RF selection with non-fixed personal characteristics under time-to-event endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	1.08	20.056	42.699	0.91	2.29	1.361	2.066	0.90	6.28	17.557	22.674	0.91
		2	-0.07	7.706	4.795	0.93	0.35	0.528	0.645	0.93	3.56	17.387	15.925	0.91
		3	1.62	1.506	2.256	0.91	1.03	0.724	0.785	0.84	1.94	2.153	3.978	0.9
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	5	1	2.77	3.167	5.799	0.90	1.90	1.236	1.366	0.76	1.67	1.840	2.236	0.88
		2	0.17	0.753	1.028	0.93	0.00	0.242	0.324	0.98	0.62	1.384	0.919	0.87
		3	1.16	1.282	1.335	0.89	0.69	0.520	0.587	0.84	0.87	0.876	1.067	0.93
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	10	1	2.34	2.588	4.669	0.88	2.40	1.475	2.011	0.86	1.18	1.374	1.195	0.88
		2	-0.18	3.913	0.618	0.95	-0.13	0.254	0.393	0.93	0.07	0.370	0.463	0.95
		3	0.99	1.141	1.293	0.87	0.60	0.486	0.641	0.90	0.67	0.765	0.835	0.89
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
Random	2	1	-2.68	31.746	61.244	0.92	1.29	25.138	109.942	0.86	-80.83	641.046	101.489	0.94
		2	-1.85	9.158	14.467	0.95	-0.89	5.915	19.942	0.88	19.35	184.338	46.897	0.9
		3	3.53	6.489	25.208	0.98	2.56	23.515	24.412	0.92	-0.63	42.632	40.913	0.94
		4	-0.01	0.258	0.324	0.97	0.00	0.177	0.191	0.96	-0.01	0.258	0.324	0.97
	5	1	32.47	339.650	88.785	0.91	-0.50	66.632	120.634	0.88	-16.21	142.496	168.673	0.92
		2	8.23	83.762	26.403	0.92	-0.06	8.969	53.692	0.91	-0.43	22.138	67.965	0.94
		3	-1.26	20.206	24.392	0.91	1.39	23.664	33.013	0.82	1.70	39.330	52.827	0.87
		4	0.01	0.250	0.323	0.97	0.01	0.161	0.206	0.98	0.01	0.250	0.323	0.97
	10	1	4.92	81.304	82.072	0.89	6.89	58.844	71.207	0.89	-5.83	125.518	375.281	0.88
		2	0.01	7.096	21.133	0.98	48.46	496.413	595.578	0.90	2.63	31.842	43.618	0.91
		3	-14.44	127.206	13.957	0.89	2.56	4.880	172.054	0.84	4.08	17.941	31.275	0.89
		4	0.05	0.309	0.330	0.95	-0.01	0.173	0.188	0.93	0.05	0.309	0.330	0.95
Decreasing	2	1	0.34	23.387	49.286	0.92	3.18	2.517	3.349	0.84	-7.41	175.609	87.754	0.9
		2	1.08	10.908	9.594	0.95	0.45	0.683	0.852	0.93	3.97	22.815	22.529	0.94
		3	1.88	1.814	4.161	0.92	1.10	0.721	0.962	0.83	2.43	2.704	5.270	0.94
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	5	1	5.13	13.529	44.103	0.92	2.36	1.939	2.105	0.83	4.25	7.419	14.172	0.95
		2	0.69	1.402	9.374	0.96	0.19	0.491	0.557	0.94	1.50	4.588	5.078	0.94
		3	1.48	2.188	12.181	0.89	0.94	0.735	0.851	0.82	1.48	1.431	2.309	0.88
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	10	1	4.11	5.183	11.691	0.93	2.21	1.591	2.153	0.84	3.16	3.567	8.330	0.92
		2	0.56	2.171	67.359	0.95	0.13	0.420	0.543	0.94	0.89	1.266	5.552	0.94
		3	1.50	1.652	2.237	0.89	0.90	0.746	0.840	0.82	1.46	1.608	1.736	0.93
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94

Table 51: Simulation results for post-RF selection with non-fixed personal characteristics under time-to-event endpoint

			Setting 1				Setting 2				Setting 3			
Pattern	Sparsity	Method	Bias	SD	SE	CR	Bias	SD	SE	CR	Bias	SD	SE	CR
Same	2	1	0.47	0.606	0.655	0.88	0.41	0.383	0.394	0.83	0.17	0.373	0.391	0.93
		2.1	-0.09	0.217	0.248	0.90	-0.09	0.153	0.167	0.91	-0.07	0.218	0.238	0.94
		2.2	0.00	0.275	0.306	0.97	0.04	0.199	0.217	0.96	-0.02	0.246	0.264	0.95
		2.3	0.00	0.279	0.324	0.97	-0.01	0.189	0.205	0.96	0.00	0.259	0.275	0.95
		3	-0.02	0.260	0.308	0.96	-0.01	0.170	0.193	0.97	-0.01	0.262	0.291	0.95
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	5	1	0.72	1.298	0.974	0.90	0.61	0.569	0.564	0.85	0.25	0.483	0.463	0.92
		2.1	-0.10	0.266	0.297	0.91	-0.14	0.161	0.187	0.84	-0.06	0.236	0.253	0.94
		2.2	0.02	0.354	0.384	0.97	0.06	0.229	0.263	0.94	-0.01	0.262	0.282	0.94
		2.3	0.17	1.820	0.366	0.96	-0.07	0.180	0.209	0.96	0.01	0.285	0.299	0.96
		3	0.00	0.293	0.345	0.98	0.00	0.183	0.201	0.97	0.03	0.365	0.337	0.93
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	10	1	1.02	1.184	2.059	0.91	1.20	1.336	1.205	0.86	0.59	0.736	0.818	0.90
		2.1	-0.17	0.229	0.412	0.92	-0.24	0.299	0.244	0.73	-0.05	0.269	0.317	0.96
		2.2	0.04	0.338	0.646	0.97	0.17	0.416	0.444	0.92	0.04	0.309	0.379	0.95
		2.3	-0.01	0.335	0.517	0.95	-0.07	0.268	0.275	0.90	0.09	0.372	0.431	0.91
		3	0.06	0.524	0.395	0.94	0.01	0.188	0.213	0.97	0.25	1.677	13.383	0.93
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
Random	2	1	0.43	0.570	0.651	0.91	0.39	0.367	0.395	0.88	0.15	0.341	0.381	0.93
		2.1	-0.09	0.230	0.248	0.95	-0.07	0.164	0.177	0.92	-0.06	0.222	0.238	0.97
		2.2	0.03	0.313	0.328	0.96	0.04	0.203	0.224	0.97	0.00	0.254	0.275	0.95
		2.3	-0.02	0.282	0.308	0.97	0.01	0.199	0.210	0.96	-0.01	0.251	0.269	0.97
		3	0.05	0.740	0.316	0.97	0.01	0.188	0.194	0.96	-0.25	2.308	0.293	0.96
		4	-0.01	0.258	0.324	0.97	0.00	0.177	0.191	0.96	-0.01	0.258	0.324	0.97
	5	1	0.68	1.701	0.743	0.93	0.41	0.350	0.421	0.87	0.18	0.325	0.408	0.98
		2.1	-0.08	0.380	0.257	0.91	-0.07	0.146	0.177	0.96	-0.05	0.202	0.249	0.98
		2.2	0.08	0.459	0.351	0.97	0.05	0.189	0.228	0.96	0.01	0.232	0.290	1.00
		2.3	-0.04	0.238	0.461	0.95	0.00	0.176	0.210	0.99	-0.01	0.223	0.280	1.00
		3	0.01	0.236	0.334	0.97	0.02	0.159	0.204	0.97	0.00	0.225	0.305	0.96
		4	0.01	0.250	0.323	0.97	0.01	0.161	0.206	0.98	0.01	0.250	0.323	0.97
	10	1	0.70	2.558	1.365	0.87	0.45	0.443	0.425	0.84	0.26	0.431	0.421	0.92
		2.1	-0.05	0.377	0.310	0.87	-0.11	0.167	0.174	0.85	-0.02	0.247	0.240	0.91
		2.2	0.13	0.507	0.444	0.94	0.04	0.213	0.221	0.94	0.05	0.291	0.284	0.91
		2.3	0.06	0.797	0.579	0.87	-0.03	0.206	0.195	0.93	0.02	0.271	0.263	0.91
		3	0.06	0.298	0.319	0.96	0.00	0.172	0.195	0.94	0.05	0.283	0.296	0.95
		4	0.05	0.309	0.330	0.95	-0.01	0.173	0.188	0.93	0.05	0.309	0.330	0.95
Decreasing	2	1	0.43	0.545	0.644	0.94	0.41	0.371	0.387	0.83	0.17	0.365	0.391	0.93
		2.1	-0.10	0.217	0.251	0.92	-0.09	0.158	0.169	0.88	-0.06	0.223	0.239	0.95
		2.2	-0.02	0.255	0.309	0.96	0.03	0.198	0.214	0.95	-0.02	0.245	0.265	0.95
		2.3	0.01	0.271	0.324	0.96	0.00	0.184	0.199	0.97	0.00	0.261	0.274	0.96
		3	-0.02	0.264	0.317	0.96	0.00	0.168	0.195	0.97	-0.01	0.262	0.290	0.94
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	5	1	0.52	0.643	0.762	0.91	0.46	0.394	0.432	0.86	0.19	0.378	0.422	0.95
		2.1	-0.11	0.208	0.270	0.93	-0.11	0.149	0.173	0.90	-0.06	0.223	0.246	0.95
		2.2	0.00	0.276	0.342	0.96	0.04	0.207	0.227	0.95	-0.03	0.241	0.274	0.95
		2.3	0.02	0.289	0.336	0.96	-0.01	0.203	0.206	0.97	0.00	0.259	0.284	0.96
		3	-0.02	0.273	0.404	0.95	0.00	0.170	0.200	0.96	-0.02	0.258	0.303	0.94
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94
	10	1	0.56	0.680	1.025	0.88	0.50	0.397	0.456	0.83	0.21	0.394	0.437	0.94
		2.1	-0.12	0.213	0.302	0.90	-0.12	0.150	0.174	0.89	-0.06	0.224	0.249	0.95
		2.2	0.00	0.276	0.431	0.95	0.05	0.204	0.232	0.94	-0.02	0.246	0.277	0.94
		2.3	-0.01	0.274	0.467	0.96	-0.02	0.180	0.213	0.96	0.00	0.266	0.288	0.95
		3	-0.01	0.281	0.323	0.94	0.00	0.172	0.200	0.95	0.00	0.281	0.311	0.96
		4	-0.01	0.280	0.355	0.94	-0.02	0.158	0.194	0.95	-0.01	0.280	0.355	0.94

Table 52: Association between 20% increase in sodium-to-potassium ratio with total CVD in high-dimensional space (NMR measurements) under Lasso, SCAD and RF approaches

Approach	Method	Lasso		SCAD		Random Forest	
		HR	95 CI	HR	95 CI	HR	95 CI
Direct-selction with V not fixed	1	1.30	(1.16,1.46)	1.18	(1.04,1.34)	1.39	(0.61,3.18)
	2	1.09	(1.04,1.15)	1.05	(0.99,1.11)	1.10	(0.71,1.70)
	3	1.08	(1.00,1.17)	1.08	(0.97,1.21)	1.28	(1.00,1.65)
	4	1.08	(1.03,1.13)	1.08	(1.03,1.13)	1.08	(1.03,1.13)
Direct-selction with V fixed	1	1.21	(1.04,1.4)	1.18	(0.80,1.73)	-	-
	2	1.06	(1.00,1.14)	1.05	(1.00,1.11)	-	-
	3	1.08	(1.00,1.18)	1.08	(0.88,1.33)	-	-
	4	1.08	(1.03,1.13)	1.08	(1.03,1.13)	-	-
Post-selection with V not fixed	1	1.20	(1.08,1.33)	1.17	(0.81,1.69)	1.19	(1.06,1.34)
	2.3	1.08	(1.01,1.15)	1.07	(0.87,1.31)	1.08	(1.03,1.13)
	3	1.09	(0.98,1.21)	1.10	(0.97,1.25)	1.11	(1.05,1.16)
	4	1.08	(1.03,1.13)	1.08	(1.03,1.13)	1.08	(1.03,1.13)
Post-selection with V fixed	1	1.16	(1.03,1.31)	1.15	(0.87,1.52)	-	-
	2.3	1.06	(0.99,1.14)	1.05	(0.90,1.22)	-	-
	3	1.08	(1.00,1.17)	1.08	(0.97,1.2)	-	-
	4	1.08	(1.03,1.13)	1.08	(1.02,1.14)	-	-

Chapter 5: Discussion

5.1 Summary

In this study, we developed a regression calibration model under the systematic measurement error assumption. A valid biomarker with bias correction was developed for regression calibration with low-dimensional data, multivariate exposures, and high-dimensional data, respectively. Four methods were examined and compared through simulations and implemented with WHI data. Overall, our proposed BF corrected biomarker model leads to consistent estimation of the association parameter between disease and dietary intakes through regression calibration under various settings.

Due to severe bias in most cases with Method 1, Method 1 is not suggested to be used. The advantages and disadvantages exist in the other three methods. Method 4 is the simplest one in design and requires fewest assumptions. However, it depends on the availability of a strong dietary instrument among the cohort and a large number of subjects in the feeding study to accurately characterize the association between the dietary instrument and the true dietary intake. The three-step approach in Method 3 allows the efficient utilization of biomarker information and is robust to the measurement error in the assessed diet in the first stage. This method works well when the dietary instrument is available in the biomarker development stage. On the other hand, if such dietary instruments are not available, Method 2 overperforms Method 3 since Method 2 does not require dietary instrument information in the biomarker development stage and it depends more on the biological association between the biomarker and the dietary intakes. Therefore, calibration equations can be built based on the same biomarker with dietary instruments that are available for the cohort but are not necessary for the controlled feeding study with Method 2.

With multiple exposures, we can see Method 3 and 4 generally provided consistent estimations and have shown efficient results, especially when there is a strong association between FFQ information and long-term dietary intake. Method 2 also gave relatively consistent estimations overall. When the association between FFQ data and true dietary intakes is weak, more efficient results have shown with Method 2 than Method 3 and 4. More importantly, the multivariate approach is shown to give robust estimations compared with the univariate approach especially when multiple exposures are correlated conditioning on personal characteristics. Under conditions when multiple exposures are independent given personal characteristics, the univariate approach is suggested because of the ease of implementation.

Under sparse high-dimensional data space, Lasso, SCAD, and RF were utilized to handle the sparse data and conduct variable selection at the first stage. In general, we found both Lasso direct-selection and Lasso post-selection with RCV variance estimation on BF generally provide consistent estimator with relatively stable CR in most cases and largely attenuate the bias compared with SCAD and RF with Method 2.

5.2 Extensions

Our developed biomarker for calibration in this paper can be generalized in other studies with samples from a similar population and disease outcome such as kidney cancer. The independence assumption naturally holds and we only need to know the form of the biomarker. In this study, the biomarker construction is based on the form of a linear regression model. Regression coefficients and associated asymptotic variance of the coefficients from the first stage can be estimated with the form of a linear regression model in the first stage. Then the asymptotic variance for the estimated association parameter between calibrated dietary intake and disease can be computed in the full cohort. When samples are overlapped across datasets, the association estimator still follows asymptotic normality, but the variance estimation of the association parameter is more complicated since datasets are not independent. In such a case, a bootstrap method can be utilized for variance estimation to avoid the violation of the assumption of independence across datasets. On the other hand, with high dimensional data, variance estimation using the bootstrap method may not be computationally efficient. Hence more computationally efficient methods need to be considered for variance estimation when high-dimensional data is of our interest.

5.3 Design

In this paper, we have shown the biased estimation caused by the systematic error in a nutrient cohort study can be largely attenuated by regression calibration with valid developed biomarkers. Method 3 with FFQ information included in the controlled feeding study has also mitigated the bias on the estimated association parameter between disease and dietary intakes, especially when the strength of biomarker is weak. Therefore, when the strength of the biomarker is found to be weaker than expected, FFQ information is suggested to be collected and included in the feeding study. In this study, repeated measurements using the same instruments are not available. With repeated measurements, $\hat{\sigma}_{x^*}$ needed for BF construction can be easily calculated in our proposed Method 2. In the future study design, repeated measurements using the same instruments can be considered if possible.

Bibliography

[1] Kenneth F Adams, Arthur Schatzkin, Tamara B Harris, Victor Kipnis, Traci Mouw, Rachel Ballard-Barbash, Albert Hollenbeck, and Michael F Leitzmann. Overweight, obesity, and mortality in a large prospective cohort of persons 50 to 71 years old. *New England Journal of Medicine*, 355(8):763–778, 2006.

[2] Annie S Anderson, Timothy J Key, Teresa Norat, Chiara Scoccianti, Michele Cecchini, Franco Berrino, Marie-Christine Boutron-Ruault, Carolina Espina, Michael Leitzmann, Hilary Powers, et al. European code against cancer 4th edition: obesity, body fatness and cancer. *Cancer Epidemiology*, 39(1):S34–S45, 2015.

[3] World Health Organization et al. *Global status report on noncommunicable diseases 2014*. Number WHO/NMH/NVI/15.1. World Health Organization, 2014.

[4] World Cancer Research Fund and American Institute for Cancer Research. Food, nutrition, physical activity, and the prevention of cancer: a global perspective, 2007.

[5] Sahasporn Paeratakul, Barry M Popkin, Lenore Kohlmeier, Irva Hertz-Picciotto, Xinxin Guo, and LJ Edwards. Measurement error in dietary data: implications for the epidemiologic study of the diet–disease relationship. *European Journal of Clinical Nutrition*, 52(10):722–727, 1998.

[6] Ross L Prentice and Ying Huang. Measurement error modeling and nutritional epidemiology association analyses. *Canadian Journal of Statistics*, 39(3):498–509, 2011.

[7] Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. *Measurement error in nonlinear models: a modern perspective*. CRC press, 2006.

[8] Ross L Prentice. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69(2):331–342, 1982.

[9] Pamela A Shaw and Ross L Prentice. Hazard ratio estimation for biomarker-calibrated dietary exposures. *Biometrics*, 68(2):397–407, 2012.

[10] Johanna W Lampe, Ying Huang, Marian L Neuhouser, Lesley F Tinker, Xiaoling Song, Dale A Schoeller, Soyoung Kim, Daniel Raftery, Chongzhi Di, Cheng Zheng, et al. Dietary biomarker evaluation in a controlled feeding study in women from the Women's Health Initiative cohort. *The American Journal of Clinical Nutrition*, 105(2):466–475, 2017.

[11] Cheng Zheng, Shirley A Beresford, Linda Van Horn, Lesley F Tinker, Cynthia A Thomson, Marian L Neuhouser, Chongzhi Di, JoAnn E Manson, Yasmin Mossavar-Rahmani, Rebecca Seguin, et al. Simultaneous association of total energy consumption and activity-related energy expenditure with risks of cardiovascular disease, cancer, and diabetes among postmenopausal women. *American Journal of Epidemiology*, 180(5):526–535, 2014.

[12] The Women's Health Initiative Study et al. Design of the Women's Health Initiative clinical trial and observational study. *Controlled Clinical Trials*, 19(1):61–109, 1998.

[13] Ruth E Patterson, Alan R Kristal, Lesley F Tinker, Rachel A Carter, Mary P Bolton, and Tanya Agurs-Collins. Measurement characteristics of the Women's Health Initiative food frequency questionnaire. *Annals of Epidemiology*, 9(3):178–187, 1999.

[14] Ross L Prentice, Ying Huang, Marian L Neuhouser, JoAnn E Manson, Yasmin Mossavar-Rahmani, Fridtjof Thomas, Lesley F Tinker, Matthew Allison, Karen C Johnson, Sylvia Wassertheil-Smoller, et al. Associations of biomarker-calibrated

sodium and potassium intakes with cardiovascular disease risk among postmenopausal women. *American Journal of Epidemiology*, 186(9):1035–1043, 2017.

[15] Leo Breiman and Jerome H Friedman. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(1):3–54, 1997.

[16] Sridhar Baldava. *Seismic performance prediction of steel structures using multiple intensity measures*. PhD thesis, University of Texas at Austin, 2004.

[17] Peter R Monge. Multivariate multiple regression in communication research. 1977.

[18] J Gary Lutz and Tanya L Eckert. The relationship between canonical correlation analysis and multivariate multiple regression. *Educational and Psychological Measurement*, 54(3):666–675, 1994.

[19] Ravindra Khattree and Dayanand N Naik. *Applied multivariate statistics with SAS software*. SAS Institute Inc., 2018.

[20] John A Nelder and Robert W Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

[21] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[22] Per K Andersen and Richard D Gill. Cox's regression model for counting processes: a large sample study. *Annals of Statistics*, 10(4):1100–1120, 1982.

[23] Norman E Breslow. Contribution to the discussion on the paper by Dr Cox, Regression models and life-tables. *Journal of Royal Statistical Society, Series B*, 34:216–217, 1972.

[24] Leonard A Stefanski and Raymond J Carroll. Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika*, 74(4):703–716, 1987.

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

[26] Ching-Yun Wang and Margaret S Pepe. Expected estimating equations to accommodate covariate measurement error. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(3):509–524, 2000.

[27] Leonard A Stefanski and James R Cook. Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association*, 90(432):1247–1256, 1995.

[28] Laurence S Freedman, Vitaly Fainberg, Victor Kipnis, Douglas Midthune, and Raymond J Carroll. A new method for dealing with measurement error in explanatory variables of regression models. *Biometrics*, 60(1):172–181, 2004.

[29] Frances E Thompson, Sharon I Kirkpatrick, Amy F Subar, Jill Reedy, TusaRebecca E Schap, Magdalena M Wilson, and Susan M Krebs-Smith. The national cancer institute's dietary assessment primer: A resource for diet research. *Journal of the Academy of Nutrition and Dietetics*, 115(12):1986–1995, 2015.

[30] Elizabeth A Sugar, Ching-Yun Wang, and Ross L Prentice. Logistic regression with exposure biomarkers and flexible measurement error. *Biometrics*, 63(1):143–151, 2007.

[31] Ross L Prentice, Elizabeth Sugar, Ching-Yun Wang, Marian Neuhouser, and Ruth Patterson. Research strategies and the use of nutrient biomarkers in studies of diet and chronic disease. *Public Health Nutrition*, 5(6a):977–984, 2002.

[32] Ching-Yun Wang, Li Hsu, Ziding Feng, and Ross L Prentice. Regression calibration in failure time regression. *Biometrics*, 53(1):131–145, 1997.

[33] Malka Gorfine, Li Hsu, and Ross L Prentice. Nonparametric correction for covariate measurement error in a stratified cox model. *Biostatistics*, 5(1):75–87, 2004.

[34] Xiaomei Liao, David M Zucker, Yi Li, and Donna Spiegelman. Survival analysis with error-prone time-varying covariates: A risk set calibration approach. *Biometrics*, 67(1):50–58, 2011.

[35] George A Mensah and David W Brown. An overview of cardiovascular disease burden in the united states. *Health Affairs*, 26(1):38–48, 2007.

[36] Sonia S Anand, Corinna Hawkes, Russell J De Souza, Andrew Mente, Mahshid Dehghan, Rachel Nugent, Michael A Zulyniak, Tony Weis, Adam M Bernstein, Ronald M Krauss, et al. Food consumption and its impact on cardiovascular disease: importance of solutions focused on the globalized food system: a report from the workshop convened by the world heart federation. *Journal of the American College of Cardiology*, 66(14):1590–1614, 2015.

[37] Najlaa Aljefree and Faruk Ahmed. Association between dietary pattern and risk of cardiovascular disease among adults in the middle east and North Africa region: a systematic review. *Food & Nutrition Research*, 59(1):27486, 2015.

[38] Dietary Guidelines Advisory Committee et al. *Dietary guidelines for Americans 2015-2020*. Government Printing Office, 2015.

[39] Vanessa Perez and Ellen T Chang. Sodium-to-potassium ratio and blood pressure, hypertension, and related factors. *Advances in Nutrition*, 5(6):712–741, 2014.

[40] Catherine E Huggins, Sharleen O'Reilly, Maree Brinkman, Allison Hodge, Graham G Giles, Dallas R English, and Caryl A Nowson. Relationship of urinary sodium and sodium-to-potassium ratio to blood pressure in older adults in Australia. *Medical Journal of Australia*, 195(3):128–132, 2011.

[41] Ross L Prentice, Lesley F Tinker, Ying Huang, and Marian L Neuhouser. Calibration of self-reported dietary measures using biomarkers: an approach to enhancing nutritional epidemiology reliability. *Current Atherosclerosis Reports*, 15(9):353, 2013.

[42] Ying Huang, Linda Van Horn, Lesley F Tinker, Marian L Neuhouser, Laura Carbone, Yasmin Mossavar-Rahmani, Fridtjof Thomas, and Ross L Prentice. Measurement error corrected sodium and potassium intake estimation using 24-hour urinary excretion. *Hypertension*, 63(2):238–244, 2014.

[43] Nancy R Cook, Eva Obarzanek, Jeffrey A Cutler, Julie E Buring, Kathryn M Rexrode, Shiriki K Kumanyika, Lawrence J Appel, and Paul K Whelton. Joint effects of sodium and potassium intake on subsequent cardiovascular disease: the trials of hypertension prevention follow-up study. *Archives of Internal Medicine*, 169(1):32–40, 2009.

[44] Carmelle Mizéhoun-Adissoda, Dismand Houinato, Corine Houehanou, Thierry Chianea, François Dalmay, André Bigot, Victor Aboyans, Pierre-Marie Preux, Pascal Bovet, and Jean-Claude Desport. Dietary sodium and potassium intakes: Data from urban and rural areas. *Nutrition*, 33:35–41, 2017.

[45] Lu Xi, Yong-Chen Hao, Jing Liu, Wei Wang, Miao Wang, Guo-Qi Li, Yue Qi, Fan Zhao, Wu-Xiang Xie, Yan Li, et al. Associations between serum potassium and sodium levels and risk of hypertension: a community-based cohort study. *Journal of Geriatric Cardiology: JGC*, 12(2):119–126, 2015.

[46] Nicolas Glatz, Aline Chappuis, David Conen, Paul Erne, Antoinette Péchère-Bertschi, Idris Guessous, Valentina F Ogna, Luca Gabutti, Franco Muggli, Augusto Gallino, et al. Associations of sodium, potassium and protein intake with blood pressure and hypertension in Switzerland. *Swiss Medical Weekly*, 147:w14411, 2017.

[47] Martin O'Donnell, Andrew Mente, Sumathy Rangarajan, Matthew J McQueen, Xingyu Wang, Lisheng Liu, Hou Yan, Shun F Lee, Prem Mony, Anitha Devanath, et al. Urinary sodium and potassium excretion, mortality, and cardiovascular events. *New England Journal of Medicine*, 371(7):612–623, 2014.

[48] Johanna M Geleijnse, Frans J Kok, and Diederick E Grobbee. Impact of dietary and lifestyle factors on the prevalence of hypertension in western populations. *The European Journal of Public Health*, 14(3):235–239, 2004.

[49] LLdiko E Frank and Jerome H Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

[50] Leo Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995.

[51] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[52] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[53] Jianqing Fan and Heng Peng. Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, 32(3):928–961, 2004.

[54] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[55] Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509–1533, 2008.

[56] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[57] Tin K Ho. Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1:278–282, 1995.

[58] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[59] Jianqing Fan, Shaojun Guo, and Ning Hao. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):37–65, 2012.

[60] Sourav Chatterjee and Jafar Jafarov. Prediction error of cross-validated lasso. *arXiv preprint arXiv:1502.06291*, 2015.

[61] Stephen Reid, Robert Tibshirani, and Jerome Friedman. A study of error variance estimation in lasso regression. *Statistica Sinica*, 26:35–67, 2016.

[62] Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of Statistics*, 37(5A):2178–2201, 2009.

[63] Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

[64] Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of Statistics*, 42(2):413–468, 2014.

Appendix A: Technical Details for Chapter 2

Theorem 1. *(Uniform weak law of large numbers (ULLN)): Suppose Θ is compact, $g(x,\theta)$ is continuous function at each $\theta \in \Theta$ with probability one, $g(x,\theta)$ is dominated by a function $G(X)$, i.e. $|g(x,\theta)| \leq G(x)$, and $EG(X) < \infty$. Then:*

$$\sup_{\theta \in \Theta} \left| n^{-1} \sum_i g(X_i,\theta) - Eg(X_i,\theta) \right| \xrightarrow{p} 0.$$

Theorem 2. *(Continuous mapping theorem (CM)): Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables and X another random variable, all taking values in the same metric space x. Let y be a metric space and $f : x \to y$ a measurable function. Define:*

$$C_f = \{x :\ g\, is\, continuous\, at\, x\}.$$

Suppose that $X_n \xrightarrow{D} X$ and $P(X \in C_f) = 1$, then $f(X_n) \xrightarrow{D} f(X)$.
Suppose that $X_n \xrightarrow{P} X$ and $P(X \in C_f) = 1$, then $f(X_n) \xrightarrow{P} f(X)$.
Suppose that $X_n \xrightarrow{a.s.} X$ and $P(X \in C_f) = 1$, then $f(X_n) \xrightarrow{a.s.} f(X)$.

Theorem 3. *(Multivariate Lindeberg-Feller Central limit theorem (CLT)): Suppose $y_{ni} \in \mathbb{R}^k$ are independent but not necessarily identically distributed with finte means $\mu_{ni} = E(y_{ni})$ and variance matrices $V_{ni} = E((y_{ni} - \mu_{ni})(y_{ni} - \mu_{ni})^T)$. Set $\overline{V_n} = n^{-1} \sum_{i=1}^{n} V_{ni}$ and $\nu_n^2 = \lambda_{min}(\overline{V_n})$. If $\nu_n^2 > 0$ and for all $\epsilon > 0$,*

$$\lim_{n\to\infty} \frac{1}{n\nu_n^2} \sum_{i=1}^{n} \mathbb{E}\left(||y_{ni} - \mu_{ni}||^2 1(||y_{ni} - \mu_{ni}||^2 \geqslant \epsilon n \nu_n^2) \right) = 0.$$

Lemma 1. *Assume $n_3/n_2 \to C_2 < \infty$ and θ^* is the unique solution to: $EU(\theta,\gamma^*) = 0$ for an estimating function $U(\theta,\gamma)$. If $\hat{\theta}$ solve the estimating equation: $0 = n_3^{-1} \sum_{i=1}^{n_3} U_i(\theta,\hat{\gamma})$,* *where* $\sqrt{n_2}(\hat{\gamma}-\gamma^*) \to N(0,\Sigma_\gamma)$ *and* $\hat{\gamma}$ *is independent of* $U_i(\theta,\gamma)$*, then we have* $\sqrt{n_3}(\hat{\theta}-\theta^*) \to N(0,\Sigma_\theta)$ *where,*

$$\Sigma_\theta = I_\theta^{-1}(J_\theta + C_2 I_\gamma \Sigma_\gamma I_\gamma^T) I_\theta^{-T},$$

where $I_\theta = E\left(-\frac{\partial U(\theta,\gamma)}{\partial\theta}|_{\theta^*,\gamma^*}\right)$, $I_\gamma = E\left(-\frac{\partial U(\theta,\gamma)}{\partial\gamma}|_{\theta^*,\gamma^*}\right)$ *and* $J_\theta = Var(U(\theta^*,\gamma^*))$. *This variance can be consistently estimated by:*

$$\hat{\Sigma}_\theta = \hat{I}_\theta^{-1}\left(\hat{J}_\theta + \frac{n_3}{n_2}\hat{I}_\gamma\hat{\Sigma}_\gamma\hat{I}_\gamma^T\right)\hat{I}_\theta^{-T},$$

where $\hat{I}_\theta = n_3^{-1}\sum_{i=1}^{n_3}\left(-\frac{\partial U_i(\theta,\gamma)}{\partial\theta}|_{\hat{\theta},\hat{\gamma}}\right)$, $\hat{I}_\gamma = n_3^{-1}\sum_{i=1}^{n_3}\left(-\frac{\partial U_i(\theta,\gamma)}{\partial\gamma}|_{\hat{\theta},\hat{\gamma}}\right)$, $\hat{J}_\theta = n_3^{-1}\sum_{i=1}^{n_3} U_i(\hat{\theta},\hat{\gamma})U_i^T(\hat{\theta},\hat{\gamma})$ *and* $\hat{\Sigma}_\gamma$ *is a consistent estimator of* Σ_γ.

Proof. We can derive the asymptotic for $\hat{\theta}$ as below:

$$\begin{aligned} 0 &= \sum_{i=1}^{n_3} U_i(\hat{\theta},\hat{\gamma}) \\ &= \sum_{i=1}^{n_3}\left\{U_i(\theta^*,\gamma^*) + \frac{\partial U_i(\theta,\gamma)}{\partial\theta}|_{\theta^*,\gamma^*}(\hat{\theta}-\theta^*) + \frac{\partial U_i(\theta,\gamma)}{\partial\gamma}|_{\theta^*,\gamma^*}(\hat{\gamma}-\gamma^*) + o(||\hat{\theta}-\theta^*||,||\hat{\gamma}-\gamma^*||)\right\}. \end{aligned}$$

With Theorem 1, we have:

$$n_3^{-1}\sum_{i=1}^{n_3}\frac{\partial U_i(\theta,\gamma)}{\partial\theta}|_{\theta^*,\gamma^*} = E\left(\frac{\partial U(\theta,\gamma)}{\partial\theta}|_{\theta^*,\gamma^*}\right) + o_p(1) = -I_\theta + o_p(1).$$

Similarly, we have:

$$n_3^{-1}\sum_{i=1}^{n_3}\frac{\partial U_i(\theta,\gamma)}{\partial\gamma}|_{\theta^*,\gamma^*} = -I_\gamma + o_p(1).$$

Plug in the Taylor expansion, notice that $EU(\theta^*, \gamma^*) = 0$, we have:

$$\begin{aligned} 0 &= n_3^{-1} \sum_{i=1}^{n_3} \{U_i(\theta^*, \gamma^*) - EU(\theta^*, \gamma^*)\} - I_\theta(\hat{\theta} - \theta^*) - I_\gamma(\hat{\gamma} - \gamma^*) \\ & \quad + o(||\hat{\theta} - \theta^*||, ||\hat{\gamma} - \gamma^*||) + o_p(1) * ||\hat{\gamma} - \gamma^*|| + o_p(1) * ||\hat{\theta} - \theta^*||. \end{aligned}$$

So we have:

$$\sqrt{n_3}(\hat{\theta} - \theta^*) = I_\theta^{-1} \left\{ n_3^{-1/2} \sum_{i=1}^{n_3} U_i(\theta^*, \gamma^*) - I_\gamma \sqrt{n_3}(\hat{\gamma} - \gamma^*) \right\} + o_p(1).$$

By Theorem 3, we have:

$$n_3^{-1/2} \sum_{i=1}^{n_3} \{U_i(\theta^*, \gamma^*) - EU(\theta^*, \gamma^*)\} \Rightarrow_d N(0, J_\theta).$$

So we have:

$$\Sigma_\theta = I_\theta^{-1} \left\{ J_\theta + C_2 I_\gamma \Sigma_\gamma I_\gamma^T \right\} I_\theta^{-T}.$$

By Theorem 1 and Theorem 2, we have $\hat{I}_\theta = I_\theta + o_p(1)$, $\hat{I}_\gamma = I_\gamma + o_p(1)$ and $\hat{J}_\theta = J_\theta + o_p(1)$. By assumption $\hat{\Sigma}_\gamma = \Sigma_\gamma + o_p(1)$ and $n_3/n_2 \to C_2 < \infty$, using Theorem 2, we have $\hat{\Sigma}_\theta = \Sigma_\theta + o_p(1)$.

□

Lemma 2. *Assume $n_3/n_2 \to C_2 < \infty$ and θ^* is the unique solution to:*

$$0 = U(\theta, \gamma^*) = E \int_0^\tau \left[\begin{pmatrix} Z^* \\ \boldsymbol{V} \end{pmatrix} - \frac{E\left[\begin{pmatrix} Z^* \\ \boldsymbol{V} \end{pmatrix} Y(t) \exp\left\{ (Z^*, \boldsymbol{V}^T)\theta \right\} \right]}{E\left[Y(t) \exp\left\{ (Z^*, \boldsymbol{V}^T)\theta \right\} \right]} \right] dN(t).$$

where $Z^* = \mathbb{X}\gamma^*$. If $\hat{\theta}$ solve the estimating equation:

$$0 = n_3^{-1} \sum_{i=1}^{n_3} \int_0^{\tau} \left[\begin{pmatrix} \hat{Z}_i \\ \boldsymbol{V}_i \end{pmatrix} - \sum_{j=1}^{n_3} \frac{Y_j(t) \exp\left\{(\hat{Z}_j, \boldsymbol{V}_j^T)\theta\right\}}{\sum_{k=1}^{n_3} Y_k(t) \exp\left\{(\hat{Z}_k, \boldsymbol{V}_k^T)\theta\right\}} \begin{pmatrix} \hat{Z}_j \\ \boldsymbol{V}_j \end{pmatrix} \right] dN_i(t),$$

where $\hat{Z}_i = \mathbb{X}_i\hat{\gamma}$ and $\sqrt{n_2}(\hat{\gamma} - \gamma^*) \to N(0, \Sigma_\gamma)$, then we have $\sqrt{n_3}(\hat{\theta} - \theta^*) \to N(0, \Sigma_\theta)$ where,

$$\Sigma_\theta = I_\theta^{-1}(J_\theta + C_2 I_\gamma \Sigma_\gamma I_\gamma^T) I_\theta^{-T},$$

$$\begin{aligned}
I_\theta &= E\left[Y_i(t) \exp\left\{(Z_i^*, \boldsymbol{V}_i^T)\theta\right\} \begin{pmatrix} Z_i^* \\ V_i \end{pmatrix}^{\otimes 2} \right], \\
J_\theta &= E \int_0^{\tau} \left[\begin{pmatrix} Z_i^* \\ \boldsymbol{V}_i \end{pmatrix} - \frac{s^{(1)}(\theta, t)}{s^{(0)}(\theta, t)} \right]^{\otimes 2} dN_i(t), \\
I_\gamma &= -E\left[\int_0^{\tau} \left\{ \begin{pmatrix} \mathbb{X}_i \\ 0 \end{pmatrix} - \frac{A(\theta, t)}{s^{(0)}(\theta, t)} + \frac{s^{(1)}(\theta, t) G(\theta, t)}{s^{(0)}(\theta, t)^2} \right\} dN(t) \right],
\end{aligned}$$

where $a^{\otimes 2} = aa^T$ and,

$$\begin{aligned}
A(\theta, t) &= E\left[Y_i(t) \begin{pmatrix} 1 + \theta_Z Z_i^* \\ \theta_Z \boldsymbol{V}_i \end{pmatrix} \exp\left\{ (Z_i^*, \boldsymbol{V}_i^T)\theta \right\} \mathbb{X}_i \right], \\
G(\theta, t) &= E\left[Y_i(t) \exp\left\{ (Z_i^*, \boldsymbol{V}_i^T)\theta \right\} \theta_Z \mathbb{X}_i \right], \\
s^{(0)}(\theta, t) &= E\left[Y_i(t) \exp\left\{ (Z_i^*, \boldsymbol{V}_i^T)\theta \right\} \right], \\
s^{(1)}(\theta, t) &= E\left[Y_i(t) \exp\left\{ (Z_i^*, \boldsymbol{V}_i^T)\theta \right\} \begin{pmatrix} Z_i^* \\ V_i \end{pmatrix} \right].
\end{aligned}$$

This variance can be consistently estimated by:

$$\hat{\Sigma}_\theta = \hat{I}_\theta^{-1} \left(\hat{J}_\theta + \frac{n_3}{n_2} \hat{I}_\gamma \hat{\Sigma}_\gamma \hat{I}_\gamma^T \right) \hat{I}_\theta^{-T},$$

where,

$$\begin{aligned}
\hat{I}_\theta &= n_3^{-1} \sum_{i=1}^{n_3} \left[Y_i(t) \exp\left\{ (\hat{Z}_i, \boldsymbol{V}_i^T)\hat{\theta} \right\} \begin{pmatrix} \hat{Z}_i \\ V_i \end{pmatrix}^{\otimes 2} \right], \\
\hat{J}_\theta &= n_3^{-1} \sum_{i=1}^{n_3} \Delta_i \left[\begin{pmatrix} \hat{Z}_i \\ \boldsymbol{V}_i \end{pmatrix} - \sum_j \frac{Y_j(T_i) \exp\left\{ (\hat{Z}_j, \boldsymbol{V}_j^T)\hat{\theta} \right\}}{\sum_k Y_k(T_i \exp\left\{ (\hat{Z}_k, \boldsymbol{V}_k^T)\hat{\theta} \right\}} \begin{pmatrix} \hat{Z}_j \\ \boldsymbol{V}_j \end{pmatrix} \right]^{\otimes 2}, \\
\hat{I}_\gamma &= n_3^{-1} \sum_{i=1}^{n_3} \Delta_i \left\{ \begin{pmatrix} \mathbb{X}_i \\ 0 \end{pmatrix} - \frac{\hat{A}(\hat{\theta}, T_i)}{s^{(0)}(\hat{\theta}, T_i)} + \frac{\hat{s}^{(1)}(\hat{\theta}, T_i)\hat{G}(\hat{\theta}, T_i)}{\hat{s}^{(0)}(\hat{\theta}, T_i)^2} \right\},
\end{aligned}$$

$$
\begin{aligned}
\hat{A}(\theta,t) &= n_3^{-1}\sum_{i=1}^{n_3}\left[Y_i(t)\begin{pmatrix} 1+\theta_Z\hat{Z}_i \\ \theta_Z\boldsymbol{V}_i \end{pmatrix}\exp\left\{(\hat{Z}_i,\boldsymbol{V}_i^T)\theta\right\}\mathbb{X}_i\right], \\
\hat{G}(\theta,t) &= n_3^{-1}\sum_{i=1}^{n_3}\left[Y_i(t)\exp\left\{(\hat{Z}_i,\boldsymbol{V}_i^T)\theta\right\}\theta_Z\mathbb{X}_i\right],
\end{aligned}
$$

$\hat{\Sigma}_\gamma$ is a consistent estimator of Σ_γ and $\frac{n3}{n2} \to C_2$.

Proof. Denote $U_i(\theta,\gamma) = \int_0^\tau \left[\left[\begin{pmatrix} Z_i \\ \boldsymbol{V}_i \end{pmatrix} - \frac{E\left[Y_j(t)\exp\left\{(Z_j,\boldsymbol{V}_j^T)\theta\right\}\begin{pmatrix} Z_j \\ \boldsymbol{V}_j \end{pmatrix}\right]}{E\left[Y_k(t)\exp\left\{(Z_k,\boldsymbol{V}_k^T)\theta\right\}\right]}\right]\right] dN_i(t)$ where $Z_i = \mathbb{X}_i\gamma$. Then by definition of I_θ, I_γ, J_θ in Lemma 1, we can compute the form of these terms as stated above. The convergence of $\hat{\theta}$ and $\hat{\gamma}$ as long as Theorem 1 and 2 ensure that we have:

$$
\begin{aligned}
\hat{A}(\hat{\theta},t) &= A(\theta^*,t) + o_p(1), \\
\hat{G}(\hat{\theta},t) &= G(\theta^*,t) + o_p(1), \\
\hat{s}^{(0)}(\hat{\theta},t) &= s^{(0)}(\theta^*,t) + o_p(1), \\
\hat{s}^{(1)}(\hat{\theta},t) &= s^{(1)}(\theta^*,t) + o_p(1),
\end{aligned}
$$

which lead to the convergence $\hat{I}_\theta = I_\theta + o_p(1)$, $\hat{I}_\gamma = I_\gamma + o_p(1)$ and $\hat{J}_\theta = J_\theta + o_p(1)$. Applying Lemma 1, we have the asymptotic for $\tilde{\theta}$ that solve the equation $0 = n_3^{-1}\sum_{i=1}^{n_3} U_i(\theta,\hat{\gamma})$. Now we just need to show $\tilde{\theta}$ and $\hat{\theta}$ is asymptotically equivalent, which is guaranteed by applying Theorem 1 to get:

$$n_3^{-1}\sum_{i=1}^{n_3} Y_j(t)\exp\left\{(\hat{Z}_j,\boldsymbol{V}_j^T)\theta\right\} = s^{(0)}(\theta,t)+o_p(1),$$

$$n_3^{-1}\sum_{i=1}^{n_3} Y_j(t)\exp\left\{(\hat{Z}_j,\boldsymbol{V}_j^T)\theta\right\}\left(\begin{array}{c}\hat{Z}_j\\ \boldsymbol{V}_j\end{array}\right) = s^{(1)}(\theta,t)+o_p(1).$$

Here we would like to comment that for Method 2-4, Z^* has the same expression $E(Z|Q,\boldsymbol{V})$ and thus the I_θ are the same though their estimated version $\hat{I}_\theta$ are different.

□

Lemma 3. *Assume $n_3/n_2 \to C_2 < \infty$ and θ^* is the unique solution to:*

$$0 = E\left[(1,Z^*,\boldsymbol{V}^T)^T\left\{Y - g^{-1}((1,Z^*,\boldsymbol{V}^T)\theta)\right\}\right],$$

where $Z^* = \mathbb{X}\gamma^*$. If $\hat{\theta}$ solve the estimating equation:

$$0 = n_3^{-1}\sum_{i=1}^{n_3}\left[(1,\hat{Z}_i,\boldsymbol{V}_i^T)^T Y_i - (1,\hat{Z}_i,\boldsymbol{V}_i^T)^T g^{-1}[(1,\hat{Z}_i,\boldsymbol{V}_i^T)\theta]\right],$$

where $\hat{Z}_i = \mathbb{X}_i\hat{\gamma}$ and $\sqrt{n_2}(\hat{\gamma}-\gamma^*) \to N(0,\Sigma_\gamma)$, then we have $\sqrt{n_3}(\hat{\theta}-\theta^*) \to N(0,\Sigma_\theta)$ where,

$$\Sigma_\theta = I_\theta^{-1}(J_\theta + C_2 I_\gamma \Sigma_\gamma I_\gamma^T)I_\theta^{-T},$$

where,

$$
\begin{aligned}
I_\theta &= E\left\{\begin{pmatrix} 1 \\ Z^* \\ \boldsymbol{V} \end{pmatrix} \frac{\partial g^{-1}(\eta)}{\partial(\eta)}(1, Z^*, \boldsymbol{V}^T)\right\}, \\
J_\theta &= E\left[(1, Z^*, \boldsymbol{V}^T)^T \left\{Y - g^{-1}((1, Z^*, \boldsymbol{V}^T)\theta)\right\}\right]^{\otimes}, \\
I_\gamma &= E\begin{pmatrix} 1 \\ Z^* \\ \boldsymbol{V} \end{pmatrix} \frac{\partial g^{-1}(\eta)}{\partial \eta}\theta_z \mathbb{X},
\end{aligned}
$$

where $\eta = (1, Z^*, \boldsymbol{V}^T)\theta$ and g is the corresponding link function. This variance can be consistently estimated by:

$$
\hat{\Sigma}_\theta = \hat{I}_\theta^{-1}\{\hat{J}_\theta + \frac{n_3}{n_2}\hat{I}_\gamma \hat{\Sigma}_\gamma \hat{I}_\gamma^T\}\hat{I}_\theta^{-T},
$$

where,

$$
\begin{aligned}
\hat{I}_\theta &= n_3^{-1}\sum_{i=1}^{n_3}\left\{\begin{pmatrix} 1 \\ \hat{Z}_i \\ \boldsymbol{V}_i \end{pmatrix} \frac{\partial g^{-1}(\eta_i)}{\partial(\eta_i)}(1, \hat{Z}_i, \boldsymbol{V}_i^T)\right\}, \\
\hat{J}_\theta &= n_3^{-1}\sum_{i=1}^{n_3}\left[(1, \hat{Z}_i, \boldsymbol{V}_i^T)^T \left\{Y_i - g^{-1}((1, \hat{Z}_i, \boldsymbol{V}_i^T)\hat{\theta})\right\}\right]^{\otimes}, \\
\hat{I}_\gamma &= n_3^{-1}\sum_{i=1}^{n_3}\begin{pmatrix} 1 \\ \hat{Z}_i \\ \boldsymbol{V}_i \end{pmatrix} \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i}\hat{\theta}_z \mathbb{X}_i,
\end{aligned}
$$

where $\eta_i = \mathbb{X}\hat{\gamma}_i$, $\hat{I}_\theta = -\frac{1}{n_3}\sum_i \frac{\partial U_i}{\partial \theta}$, $\hat{I}_\gamma = -\frac{1}{n_3}\sum_i \frac{\partial U_i}{\partial \gamma}$ *and* $\hat{\Sigma}_\gamma$ *is a consistent estimator of* Σ_γ.

Proof. We apply Lemma 1 with $U(\theta, \gamma) = (1, Z, \boldsymbol{V}^T)^T \left\{Y - g^{-1}((1, Z, \boldsymbol{V}^T)\theta)\right\}$ where $Z = \mathbb{X}\gamma$. With some calculus, we obtain the form of I_θ, I_γ, J_θ and $\hat{I}_\theta$, $\hat{I}_\gamma$, $\hat{J}_\theta$. □

Corollary 1. *Assume* $n_3/n_2 \to C_2 < \infty$ *and* θ^* *is the unique solution to:*

$$0 = U = E\sum_{i=1}\left[(1, Z_i^*, \boldsymbol{V}_i^T)^T Y_i - (1, Z_i^*, \boldsymbol{V}_i^T)^T(1, Z_i^*, \boldsymbol{V}_i^T)\theta\right].$$

where $Z^* = X\gamma^*$. *If* $\hat{\theta}$ *solve the estimating equation:*

$$0 = \sum U_i = n_3^{-1}\sum_{i=1}\left[(1, \hat{Z}_i, \boldsymbol{V}_i^T)^T Y_i - (1, \hat{Z}_i, \boldsymbol{V}_i^T)^T(1, \hat{Z}_i, \boldsymbol{V}_i^T)\theta\right],$$

where $\hat{Z}_i = X_i\hat{\gamma}$ *and* $\sqrt{n_2}(\hat{\gamma} - \gamma^*) \to N(0, \Sigma_\gamma)$, *then we have* $\sqrt{n_3}(\hat{\theta} - \theta^*) \to N(0, \Sigma_\theta)$ *where*

$$\Sigma_\theta = I_\theta^{-1}(I_\theta + C_2 I_\gamma \Sigma_\gamma I_\gamma^T)I_\theta^{-T},$$

where $I_\theta = -E\frac{\partial U}{\partial \theta}$ *and* $I_\gamma = -E\frac{\partial U}{\partial \gamma}$. *This variance can be consistently estimated by:*

$$\hat{\Sigma}_\theta = \hat{I}_\theta^{-1}\{\hat{I}_\theta + \frac{n_3}{n_2}\hat{I}_\gamma\hat{\Sigma}_\gamma\hat{I}_\gamma^T\}\hat{I}_\theta^{-T},$$

where $\hat{I}_\theta = -\frac{1}{n_3}\sum_i \frac{\partial U_i}{\partial \theta}$, $\hat{I}_\gamma = -\frac{1}{n_3}\sum_i \frac{\partial U_i}{\partial \gamma}$ *and* $\hat{\Sigma}_\gamma$ *is a consistent estimator of* Σ_γ.

Proof. Plug in $g(x) = x$ for linear regression form, we have:

$$I_\theta = E\left\{\sum_i \begin{pmatrix} 1 \\ Z_i^* \\ \boldsymbol{V}_i \end{pmatrix}(1, Z_i^*, \boldsymbol{V}_i^T)\right\},$$

$$I_\gamma = E\begin{pmatrix} 1 \\ Z_i^* \\ \boldsymbol{V}_i \end{pmatrix}(\mathbb{X}_i\theta_z),$$

$$\hat{I}_\theta \quad = \quad {n_3}^{-1}\sum_i \begin{pmatrix} 1 \\ \hat{Z}_i \\ \boldsymbol{V}_i \end{pmatrix}(1, \hat{Z}_i, \boldsymbol{V}_i^T),$$

$$\hat{I}_\gamma = {n_3}^{-1}\sum_i \begin{pmatrix} 1 \\ \hat{Z}_i \\ \boldsymbol{V}_i \end{pmatrix}(\mathbb{X}_i\theta_z).$$

Applying Lemma 3, we obtain the asymptotic for θ in linear regression setting. □

Corollary 2. *Assume $n_3/n_2 \to C_2 < \infty$ and θ^* is the unique solution to:*

$$0 = U = \sum_{i=1}\left[(1, Z_i^*, \boldsymbol{V}_i^T)^T Y_i - (1, Z_i^*, \boldsymbol{V}_i^T)^T \frac{exp(\eta_i)}{(1+exp(\eta_i))}\right],$$

where $\eta_i = (1, Z_i^*, \boldsymbol{V}_i^T)\theta$, $Z^* = X\gamma^*$. If $\hat{\theta}$ solve the estimating equation:

$$0 = \sum U_i = n_3^{-1}\sum_{i=1}^{n_3}\left[(1, \hat{Z}_i, \boldsymbol{V}_i^T)^T Y_i - (1, \hat{Z}_i, \boldsymbol{V}_i^T)^T \frac{exp(\eta_i)}{(1+exp(\eta_i))}\right],$$

where $\hat{Z}_i = X_i\hat{\gamma}$ and $\sqrt{n_2}(\hat{\gamma} - \gamma^*) \to N(0, \Sigma_\gamma)$, then we have $\sqrt{n_3}(\hat{\theta} - \theta^*) \to N(0, \Sigma_\theta)$ where,

$$\Sigma_\theta = I_\theta^{-1}(I_\theta + C_2 I_\gamma \Sigma_\gamma I_\gamma^T)I_\theta^{-T},$$

where $I_\theta = -E\frac{\partial U}{\partial \theta}$ and $I_\gamma = -E\frac{\partial U}{\partial \gamma}$. This variance can be consistently estimated by:

$$\hat{\Sigma}_\theta = \hat{I}_\theta^{-1}\{\hat{I}_\theta + \frac{n_3}{n_2}\hat{I}_\gamma \hat{\Sigma}_\gamma \hat{I}_\gamma^T\}\hat{I}_\theta^{-T},$$

where $\hat{I}_\theta = -\frac{1}{n_3}\sum_i \frac{\partial U_i}{\partial \theta}$, $\hat{I}_\gamma = -\frac{1}{n_3}\sum_i \frac{\partial U_i}{\partial \gamma}$ and $\hat{\Sigma}_\gamma$ is a consistent estimator of Σ_γ.

Proof. Plug in $g(x) = \log\left(\frac{x}{1-x}\right)$ for linear regression form, we have:

$$I_\theta = -E\sum_i \begin{pmatrix} 1 \\ Z_i^* \\ \boldsymbol{V}_i \end{pmatrix} \frac{exp(\eta_i)}{(1+exp(\eta_i))^2}(1, Z_i^*, \boldsymbol{V}_i^T),$$

$$I_\gamma = E \begin{pmatrix} 1 \\ Z_i^* \\ \boldsymbol{V}_i \end{pmatrix} \frac{exp(\eta_i)}{(1+exp(\eta_i))^2}\left(\mathbb{X}_i\theta_z\right),$$

$$\hat{I}_\theta \quad = \quad -{n_3}^{-1}\sum_i \begin{pmatrix} 1 \\ \hat{Z}_i \\ \boldsymbol{V}_i \end{pmatrix} \frac{exp(\hat{\eta}_i)}{(1+exp(\hat{\eta}_i))^2}(1, \hat{Z}_i, \boldsymbol{V}_i^T),$$

$$\hat{I}_\gamma = {n_3}^{-1}\sum_i \begin{pmatrix} 1 \\ \hat{Z}_i \\ \boldsymbol{V}_i \end{pmatrix} \frac{exp(\hat{\eta}_i)}{(1+exp(\hat{\eta}_i))^2}\left(\mathbb{X}_i\hat{\theta}_z\right).$$

Applying Lemma 3, we obtain the asymptotic for θ in linear regression setting. □

Lemma 4. *Assume $n_2/n_1 \to C_1 < \infty$ and γ^* is the unique solution to $EU(\gamma, \beta^*) = 0$ for an estimating function $U(\gamma, \beta) = [(1, Q_i, \boldsymbol{V}_i^T)^T \hat{X}_i - (1, Q_i, \boldsymbol{V}_i^T)^T (1, Q_i, \boldsymbol{V}_i^T)\gamma]$. If $\hat{\gamma}$ solve the estimating equation $0 = n_3^{-1} \sum_{i=1}^{n_3} U_i(\gamma, \hat{\beta})$ where, $\sqrt{n_1}(\hat{\beta} - \beta^*) \to N(0, \Sigma_\beta)$, then we have $\sqrt{n_2}(\hat{\gamma} - \gamma^*) \to N(0, \Sigma_\gamma)$ where,*

$$\Sigma_\gamma = I_\gamma^{-1}(I_\gamma + C_1 I_\beta \Sigma_\beta I_\beta^T) I_\gamma^{-T},$$

where $I_\gamma = E\left(-\frac{\partial U(\gamma,\beta)}{\partial \gamma}|_{\gamma^*,\beta^*}\right) = E[(1, Q_i, \boldsymbol{V}_i^T)^T (1, Q_i, \boldsymbol{V}_i^T)]$ and $I_\beta = E\left(-\frac{\partial U(\gamma,\beta)}{\partial \beta}|_{\gamma^*,\beta^*}\right)$. This variance can be consistently estimated by:

$$\hat{\Sigma}_\gamma = \hat{I}_\gamma^{-1}\left(\hat{I}_\gamma + \frac{n_2}{n_1}\hat{I}_\beta \hat{\Sigma}_\beta \hat{I}_\beta^T\right)\hat{I}_\gamma^{-T},$$

where $\hat{I}_\gamma = n_2^{-1}\sum_{i=1}^{n_2}\left(-\frac{\partial U_i(\gamma,\beta)}{\partial \gamma}|_{\hat{\gamma},\hat{\beta}}\right)$, $\hat{I}_\beta = n_2^{-1}\sum_{i=1}^{n_2}\left(-\frac{\partial U_i(\gamma,\beta)}{\partial \beta}|_{\hat{\gamma},\hat{\beta}}\right)$ and $\hat{\Sigma}_\beta$ is a consistent estimator of Σ_β.

Proof. We can derive the asymptotic for $\hat{\gamma}$ as below.

$$\begin{aligned} 0 &= \sum_{i=1}^{n_2} U_i(\hat{\gamma}, \hat{\beta}) \\ &= \sum_{i=1}^{n_2}\left\{U_i(\gamma^*, \beta^*) + \frac{\partial U}{\partial \gamma}|_{\gamma^*,\beta^*}(\hat{\gamma} - \gamma^*) + \frac{\partial U}{\partial \beta}|_{\gamma^*,\beta^*}(\hat{\beta} - \beta^*) + o(||\hat{\gamma} - \gamma^*||, ||\hat{\beta} - \beta^*||)\right\}. \end{aligned}$$

With Theorem 1,

$$n_2^{-1}\sum_{i=1}^{n_2}\frac{\partial U_i(\gamma,\beta)}{\partial \gamma}|_{\hat{\gamma},\hat{\beta}} = E\left(\frac{\partial U}{\partial \gamma}|_{\hat{\gamma},\hat{\beta}}\right) + o_p(1),$$

and with Theorem 2, we have:

$$E\left(\frac{\partial U}{\partial \gamma}|_{\hat{\gamma},\hat{\beta}}\right) = E\left(\frac{\partial U}{\partial \gamma}|_{\gamma^*,\beta^*}\right) + o_p(1).$$

So we have:

$$n_2^{-1}\sum_{i=1}^{n_2}\frac{\partial U_i(\gamma,\beta)}{\partial\gamma}|_{\hat{\gamma},\hat{\beta}} = -I_\gamma + o_p(1).$$

Similarly, we have:

$$n_2^{-1}\sum_{i=1}^{n_2}\frac{\partial U_i(\gamma,\beta)}{\partial\beta}|_{\hat{\gamma},\hat{\beta}} = -I_\beta + o_p(1).$$

Plug in the Taylor expansion, notice that $EU(\gamma^*,\beta^*)=0$, we have:

$$\begin{aligned} 0 &= n_2^{-1}\sum_{i=1}^{n_2}\{U_i(\gamma^*,\beta^*) - EU(\gamma^*,\beta^*)\} - I_\gamma(\hat{\gamma}-\gamma^*) - I_\beta(\hat{\beta}-\beta^*) \\ &\quad + o(||\hat{\gamma}-\gamma^*||, ||\hat{\beta}-\beta^*||) + o_p(1) * ||\hat{\beta}-\beta^*|| + o_p(1) * ||\hat{\gamma}-\gamma^*||. \end{aligned}$$

So we have:

$$\sqrt{n_2}(\hat{\gamma}-\gamma^*) = I_\gamma^{-1}\left\{n_2^{-1/2}\sum_{i=1}^{n_2} U_i(\gamma^*,\beta^*) - I_\beta\sqrt{n_2}(\hat{\beta}-\beta^*)\right\} + o_p(1).$$

By Theorem 3, we have:

$$n_2^{-1/2}\sum_{i=1}^{n_2}\{U_i(\gamma^*,\beta^*) - EU(\gamma^*,\beta^*)\} \Rightarrow_d N(0, Var(U_i(\gamma^*,\beta^*))).$$

So we have:

$$\Sigma_\gamma = I_\gamma^{-1}\left\{Var(U(\gamma,\beta)) + C_2 I_\beta \Sigma_\beta I_\beta^T\right\} I_\gamma^{-T}.$$

As we have shown $\hat{I}_\gamma = I_\gamma + o_p(1)$, $\hat{I}_\beta = I_\beta + o_p(1)$ and by assumption $\hat{\Sigma}_\beta = \Sigma_\beta + o_p(1)$ and $n_2/n_1 \to C_1 < \infty$, using Theorem 2, we have $\hat{\Sigma}_\gamma = \Sigma_\gamma + o_p(1)$. □

Lemma 5. *Assume $n_2/n_1 \to C_1 < \infty$ and γ_1^* is the unique solution to $EU(\gamma_1, \beta_1^*) = 0$ for an estimating function $U(\gamma_1, \beta_1)$. If $\hat{\gamma}_1$ solve the estimating equation $0 = n_2^{-1} \sum_{i=1}^{n_2} U_i(\gamma_1, \hat{\beta}_1)$ where $\sqrt{n_1}(\hat{\beta}_1 - \beta_1^*) \to N(0, \Sigma_{\beta_1})$, then we have $\sqrt{n_2}(\hat{\gamma}_1 - \gamma_1^*) \to N(0, \Sigma_{\gamma_1})$ where,*

$$\Sigma_{\gamma_1} = I_{\gamma_1}^{-1}(I_{\gamma_1} + C_1 I_{\beta_1} \Sigma_{\beta_1} I_{\beta_1}^T) I_{\gamma_1}^{-T},$$

where $I_{\gamma_1} = E\left(-\frac{\partial U(\gamma_1,\beta_1)}{\partial \gamma_1}|_{\gamma_1^*,\beta_1^*}\right)$ and $I_{\beta_1} = E\left(-\frac{\partial U(\gamma_1,\beta_1)}{\partial \beta_1}|_{\gamma_1^*,\beta_1^*}\right)$. This variance can be consistently estimated by:

$$\hat{\Sigma}_{\gamma_1} = \hat{I}_{\gamma_1}^{-1}\left(\hat{I}_{\gamma_1} + \frac{n_2}{n_1} \hat{I}_{\beta_1} \hat{\Sigma}_{\beta_1} \hat{I}_{\beta_1}^T\right) \hat{I}_{\gamma_1}^{-T},$$

where $\hat{I}_{\gamma_1} = n_2^{-1} \sum_{i=1}^{n_2} \left(-\frac{\partial U_i(\gamma_1,\beta_1)}{\partial \gamma_1}|_{\hat{\gamma}_1,\hat{\beta}_1}\right)$, $\hat{I}_{\beta_1} = n_2^{-1} \sum_{i=1}^{n_2} \left(-\frac{\partial U_i(\gamma_1,\beta_1)}{\partial \beta_1}|_{\hat{\gamma}_1,\hat{\beta}_1}\right)$ and $\hat{\Sigma}_{\beta_1}$ is a consistent estimator of Σ_{β_1}.

Proof. To derive the asymptotic for $\hat{\gamma}_1$, we need to derive asymptotic for $\hat{\beta}_1$ and then apply Lemma 4.

$$\begin{aligned} 0 &= \sum_{i=1}^{n_1} U_i(\hat{\beta}_1) \\ &= \sum_{i=1}^{n_1} \left\{ U_i(\beta_1^*) + \frac{\partial U}{\partial \beta_1}|_{\beta_1^*}(\hat{\beta}_1 - \beta_1^*) + o(||\hat{\beta}_1 - \beta_1^*||) \right\}, \end{aligned}$$

where $U_i(\beta_1^*) = (1, \boldsymbol{W}_i^T, \boldsymbol{V}_i^T)^T X_i^* - (1, \boldsymbol{W}_i^T, \boldsymbol{V}_i^T)^T (1, \boldsymbol{W}_i^T, \boldsymbol{V}_i^T) \beta_1^*$,

With Theorem 1,

$$n_1^{-1} \sum_{i=1}^{n_1} \frac{\partial U_i(\beta_1)}{\partial \beta_1}|_{\hat{\beta}_1} = E\left(\frac{\partial U}{\partial \beta_1}|_{\hat{\beta}_1}\right) + o_p(1),$$

and with Theorem 2, we have:

$$E\left(\frac{\partial U}{\partial \beta_1}|_{\hat{\beta}_1}\right) = E\left(\frac{\partial U}{\partial \beta_1}|_{\beta_1^*}\right) + o_p(1).$$

So we have:

$$n_1^{-1}\sum_{i=1}^{n_1}\frac{\partial U_i(\beta_1)}{\partial \beta_1}|_{\hat{\beta}_1} = -I_{\beta_1} + o_p(1).$$

Plug in the Taylor expansion, notice that $EU(\beta_1^*) = 0$, we have:

$$\begin{aligned} 0 &= n_1^{-1}\sum_{i=1}^{n_1}\{U_i(\beta_1^*) - EU(\beta_1^*)\} - I_{\beta_1}(\hat{\beta_1} - \beta_1^*) \\ &\quad + o_p(1)||\hat{\beta}_1 - \beta_1^*||. \end{aligned}$$

So we have:

$$\sqrt{n_1}(\hat{\beta}_1 - \beta_1^*) = I_{\beta_1}^{-1}\left\{n_1^{-1/2}\sum_{i=1}^{n_1}U_i(\beta_1^*)\right\} + o_p(1).$$

By Theorem 3, we have:

$$n_1^{-1/2}\sum_{i=1}^{n_2}\{U_i(\beta_1^*) - EU(\beta_1^*)\} \Rightarrow_d N(0, Var(U_i(\beta_1^*))).$$

So we have:

$$\Sigma_{\beta_1} = I_{\beta_1}^{-1}\{Var(U(\beta_1))\}I_{\beta_1}^{-T}.$$

By assumption, $\hat{\Sigma}_{\beta_1} = \Sigma_{\beta_1} + o_p(1)$ which is a consistent estimator of Σ_{β_1}.

Then by applying Lemma 4, we have:

$$\hat{\Sigma}_{\gamma_1} = \hat{I}_{\gamma_1}^{-1} \left(\hat{I}_{\gamma_1} + \frac{n_2}{n_1} \hat{I}_{\beta_1} \hat{\Sigma}_{\beta_1} \hat{I}_{\beta_1}^T \right) \hat{I}_{\gamma_1}^{-T}.$$

□

Lemma 6. *Assume $n_2/n_1 \to C_1 < \infty$ and γ_2^* is the unique solution to $EU(\gamma_2, \beta_2^*) = 0$ for an estimating function $U(\gamma_2, \beta_2)$. If $\hat{\gamma}_2$ solve the estimating equation $0 = n_2^{-1} \sum_{i=1}^{n_2} U_i(\gamma_2, \hat{\beta}_2)$ where $\sqrt{n_1}(\hat{\beta}_2 - \beta_2^*) \to N(0, \Sigma_{\beta_2})$, then we have $\sqrt{n_2}(\hat{\gamma}_2 - \gamma_2^*) \to N(0, \Sigma_{\gamma_2})$ where,*

$$\Sigma_{\gamma_2} = I_{\gamma_2}^{-1}(I_{\gamma_2} + C_1 I_{\beta_2} \Sigma_{\beta_2} I_{\beta_2}^T) I_{\gamma_2}^{-T},$$

where $I_{\gamma_2} = E\left(-\frac{\partial U(\gamma_2,\beta_2)}{\partial \gamma_2}|_{\gamma_2^*,\beta_2^*}\right)$ and $I_{\beta_2} = E\left(-\frac{\partial U(\gamma_2,\beta_2)}{\partial \beta_2}|_{\gamma_2^*,\beta_2^*}\right)$. This variance can be consistently estimated by:

$$\hat{\Sigma}_{\gamma_2} = \hat{I}_{\gamma_2}^{-1} \left(\hat{I}_{\gamma_2} + \frac{n_2}{n_1} \hat{I}_{\beta_2} \hat{\Sigma}_{\beta_2} \hat{I}_{\beta_2}^T \right) \hat{I}_{\gamma_2}^{-T},$$

where $\hat{I}_{\gamma_2} = n_2^{-1} \sum_{i=1}^{n_2} \left(-\frac{\partial U_i(\gamma_2,\beta_2)}{\partial \gamma_2}|_{\hat{\gamma}_2,\hat{\beta}_2}\right)$, $\hat{I}_{\beta_2} = n_2^{-1} \sum_{i=1}^{n_2} \left(-\frac{\partial U_i(\gamma_2,\beta_2)}{\partial \beta_2}|_{\hat{\gamma}_2,\hat{\beta}_2}\right)$ and $\hat{\Sigma}_{\beta_2}$ is a consistent estimator of Σ_{β_2}.

Proof. The asymptotic of $\hat{\gamma}_2$ can be derived as below.

First note that,

$$Var(X^*|\boldsymbol{V}, \boldsymbol{W}) = \Omega_1 = \frac{1}{n_1} \sum_i \left\{ X_i^* - (1, \boldsymbol{W}_i^T, \boldsymbol{V}_i^T)\beta \right\}^2,$$

$$Var(X^*|\boldsymbol{V}) = \Omega_2 = \frac{1}{n_1} \sum_i \left\{ X_i^* - (1, \boldsymbol{V}_i^T)\beta_t \right\}^2,$$

where we let $\Omega_{1i} = \left\{ X_i^* - (1, \boldsymbol{W}_i^T, \boldsymbol{V}_i^T)\beta \right\}^2$ and $\Omega_{2i} = \left(X_i^* - (1, \boldsymbol{V}_i^T)\beta_t \right)^2$.

Second, the estimating equations considered are:

$$U_{121i} = (1, \boldsymbol{W}_i^T, \boldsymbol{V}_i^T)^T \Omega_1^{-1} (X_i^* - (1, \boldsymbol{W}_i^T, \boldsymbol{V}_i^T)\beta),$$

$$U_{122i} = (1, \boldsymbol{V}_i^T)^T \Omega_2^{-1} (X_i^* - (1, \boldsymbol{V}_i^T)\beta_t),$$

$$U_{123i} = (X_i^* - (1, \boldsymbol{W}_i^T, \boldsymbol{V}_i^T)^T \beta)^2 - \Omega_1,$$

$$U_{124i} = (X_i^* - (1, \boldsymbol{V}_i^T)\beta_t)^2 - \Omega_2.$$

Third, we can derive the asymptotic normal distribution for β, β_t, Ω_1 and Ω_2 as:

$$\sqrt{n_1} \left[\begin{pmatrix} \hat{\beta} \\ \hat{\beta}_t \\ \widehat{\Omega}_1 \\ \widehat{\Omega}_2 \end{pmatrix} - \begin{pmatrix} \beta \\ \beta_t \\ \Omega_1 \\ \Omega_2 \end{pmatrix} \right] \to N(0, I^{-1} J I^{-T}),$$

where J is the variance covariance matrix of the above four estimating equations and I is a matrix composed by the expectation of derivatives of each estimating equation with respect to β, β_t, Ω_1 and Ω_2, respectively. Specifically,

$$I = \begin{pmatrix} \frac{1}{n_1} \sum_i (\mathbb{X}_i)^T (\mathbb{X}_i) & 0 & 0 & 0 \\ 0 & \frac{1}{n_1} \sum_i (\mathbb{X}_{ti})^T (\mathbb{X}_{ti}) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

where $\mathbb{X}_i = (1, \boldsymbol{W}_i^T, \boldsymbol{V}_i^T)$ and $\mathbb{X}_{ti} = (1, \boldsymbol{V}_i^T)$.

Fourth, the asymptotic normal distribution for $\hat{\beta}$ and $\widehat{BF}$ can be derived using delta method.

$$\sqrt{n_1}\left[\begin{pmatrix} \hat{\beta} \\ \widehat{BF} \end{pmatrix} - \begin{pmatrix} \beta \\ BF \end{pmatrix}\right] \rightarrow N(0, CI^{-1}JI^{-T}C^T),$$

where C is a matrix derived by taking derivative of β and BF each with respect to β, β_t, V_1 and V_2 respectively. For example,

$$\frac{\partial BF}{\partial \beta} = 0, \frac{\partial BF}{\partial \beta_t} = 0, \frac{\partial BF}{\partial \Omega_1} = \frac{-1}{\Omega_2 - \sigma_*^2}, \frac{\partial BF}{\partial V_2} = \frac{\Omega_1 - \sigma_*^2}{(\Omega_2 - \sigma_*^2)^2}.$$

Fifth, $\hat{\beta}_2 = \frac{\hat{\beta}}{\hat{BF}}$ can be derived using delta method. That is:

$$\sqrt{n_1}\left(\hat{\beta}_2 \;-\; \beta_2 \right) \rightarrow N(0, C'(CI^{-1}JI^{-T}C^T)C'^T),$$

where C' is a matrix derived by taking derivative of β_2 each with respect to β and BF, respectively. That is:

$$\frac{\partial \beta_2}{\partial \beta} = \frac{1}{BF}, \frac{\partial \beta_2}{\partial BF} = -\frac{\beta}{BF^2}.$$

Then we have estimating equation for β_2 as below:

$$\begin{aligned} 0 &= \sum_{i=1}^{n_1} U_i(\hat{\beta}_2) \\ &= \sum_{i=1}^{n_1} \left\{ U_i(\beta_2^*) + \frac{\partial U}{\partial \beta_2}|_{\beta_2^*}(\hat{\beta}_2 - \beta_2^*) + o(||\hat{\beta}_2 - \beta_2^*||) \right\}. \end{aligned}$$

With Theorem 1,

$$n_1^{-1}\sum_{i=1}^{n_1}\frac{\partial U_i(\beta_2)}{\partial\beta_2}|_{\hat{\beta}_2} = E\left(\frac{\partial U}{\partial\beta_2}|_{\hat{\beta}_2}\right) + o_p(1),$$

and with Theorem 2, we have:

$$E\left(\frac{\partial U}{\partial\beta_1}|_{\hat{\beta}_2}\right) = E\left(\frac{\partial U}{\partial\beta_1}|_{\beta_2^*}\right) + o_p(1).$$

So we have:

$$n_1^{-1}\sum_{i=1}^{n_1}\frac{\partial U_i(\beta_2)}{\partial\beta_1}|_{\hat{\beta}_2} = -I_{\beta_1} + o_p(1).$$

Plug in the Taylor expansion, notice that $EU(\beta_2^*) = 0$, we have:

$$\begin{aligned} 0 &= n_1^{-1}\sum_{i=1}^{n_1}\{U_i(\beta_2^*) - EU(\beta_2^*)\} - I_{\beta_1}(\hat{\beta_2} - \beta_2^*) \\ &\quad + o_p(1)||\hat{\beta}_2 - \beta_2^*||. \end{aligned}$$

So we have:

$$\sqrt{n_1}(\hat{\beta}_2 - \beta_2^*) = I_{\beta_2}^{-1}\left\{n_1^{-1/2}\sum_{i=1}^{n_1}U_i(\beta_2^*)\right\} + o_p(1).$$

By Theorem 3, we have:

$$n_1^{-1/2}\sum_{i=1}^{n_2}\{U_i(\beta_2^*) - EU(\beta_2^*)\} \to N(0, Var(U_i(\beta_2^*))).$$

So we have:

$$\Sigma_{\beta_2} = I_{\beta_2}^{-1}\{Var(U(\beta_2))\}I_{\beta_2}^{-T}.$$

By assumption, $\hat{\Sigma}_{\beta_2} = \Sigma_{\beta_2} + o_p(1)$ which is a consistent estimator of Σ_{β_2}.

Then by applying Lemma 4, we have:

$$\hat{\Sigma}_{\gamma_2} = \hat{I}_{\gamma_2}^{-1}\left(\hat{I}_{\gamma_2} + \frac{n_2}{n_1}\hat{I}_{\beta_2}\hat{\Sigma}_{\beta_2}\hat{I}_{\beta_2}^{T}\right)\hat{I}_{\gamma_2}^{-T}.$$

□

Lemma 7. *Assume $n_2/n_1 \to C_1 < \infty$ and γ_3^* is the unique solution to: $EU(\gamma_3, \beta_3^*) = 0$ for an estimating function $U(\gamma_3, \beta_3)$. If $\hat{\gamma}_3$ solve the estimating equation: $0 = n_2^{-1}\sum_{i=1}^{n_2} U_i(\gamma_3, \hat{\beta}_3)$ where $\sqrt{n_1}(\hat{\beta}_3 - \beta_3^*) \to N(0, \Sigma_{\beta_3})$, then we have $\sqrt{n_2}(\hat{\gamma}_3 - \gamma_3^*) \to N(0, \Sigma_{\gamma_3})$ where,*

$$\Sigma_{\gamma_3} = I_{\gamma_3}^{-1}(I_{\gamma_3} + C_1 I_{\beta_3}\Sigma_{\beta_3}I_{\beta_3}^{T})I_{\gamma_3}^{-T},$$

where $I_{\gamma_3} = E\left(-\frac{\partial U(\gamma_3,\beta_3)}{\partial \gamma_3}|_{\gamma_3^*,\beta_3^*}\right)$ and $I_{\beta_3} = E\left(-\frac{\partial U(\gamma_3,\beta_3)}{\partial \beta_3}|_{\gamma_3^*,\beta_3^*}\right)$. This variance can be consistently estimated by:

$$\hat{\Sigma}_{\gamma_3} = \hat{I}_{\gamma_3}^{-1}\left(\hat{I}_{\gamma_3} + \frac{n_2}{n_1}\hat{I}_{\beta_3}\hat{\Sigma}_{\beta_3}\hat{I}_{\beta_3}^{T}\right)\hat{I}_{\gamma_3}^{-T},$$

where $\hat{I}_{\gamma_3} = n_2^{-1}\sum_{i=1}^{n_2}\left(-\frac{\partial U_i(\gamma_3,\beta_3)}{\partial \gamma_3}|_{\hat{\gamma}_3,\hat{\beta}_3}\right)$, $\hat{I}_{\beta_3} = n_2^{-1}\sum_{i=1}^{n_2}\left(-\frac{\partial U_i(\gamma_3,\beta_3)}{\partial \beta_3}|_{\hat{\gamma}_3,\hat{\beta}_3}\right)$ and $\hat{\Sigma}_{\beta_3}$ is a consistent estimator of Σ_{β_3}.

Proof. The asymptotic of $\hat{\gamma}_3$ can be derived as below.

$$\begin{aligned} 0 &= \sum_{i=1}^{n_1} U_i(\hat{\beta}_3) \\ &= \sum_{i=1}^{n_1}\left\{U_i(\beta_3^*) + \frac{\partial U}{\partial \beta_3}|_{\beta_3^*}(\hat{\beta}_3 - \beta_3^*) + o(||\hat{\beta}_3 - \beta_3^*||)\right\}, \end{aligned}$$

where $U_i(\beta_3^*) = (1, \boldsymbol{W}_i^T, Q_i, \boldsymbol{V}_i^T)^T X_i^* - (1, \boldsymbol{W}_i^T, Q_i, \boldsymbol{V}_i^T)^T (1, \boldsymbol{W}_i^T, Q_i, \boldsymbol{V}_i^T)\beta_3^*$. With Theorem 1,

$$n_1^{-1} \sum_{i=1}^{n_1} \frac{\partial U_i(\beta_3)}{\partial \beta_3}|_{\hat{\beta}_3} = E\left(\frac{\partial U}{\partial \beta_3}|_{\hat{\beta}_3}\right) + o_p(1),$$

and with Theorem 2, we have:

$$E\left(\frac{\partial U}{\partial \beta_3}|_{\hat{\beta}_3}\right) = E\left(\frac{\partial U}{\partial \beta_3}|_{\beta_3^*}\right) + o_p(1).$$

So we have:

$$n_1^{-1} \sum_{i=1}^{n_1} \frac{\partial U_i(\beta_3)}{\partial \beta_3}|_{\hat{\beta}_3} = -I_{\beta_3} + o_p(1).$$

Plug in the Taylor expansion, notice that $EU(\beta_3^*) = 0$, we have:

$$\begin{aligned} 0 &= n_1^{-1} \sum_{i=1}^{n_1} \{U_i(\beta_3^*) - EU(\beta_3^*)\} - I_{\beta_3}(\hat{\beta}_3 - \beta_3^*) \\ &\quad + o_p(1)||\hat{\beta}_3 - \beta_3^*||. \end{aligned}$$

So we have:

$$\sqrt{n_1}(\hat{\beta}_3 - \beta_3^*) = I_{\beta_3}^{-1} \left\{ n_1^{-1/2} \sum_{i=1}^{n_1} U_i(\beta_3^*) \right\} + o_p(1).$$

By Theorem 3, we have:

$$n_1^{-1/2} \sum_{i=1}^{n_2} \{U_i(\beta_3^*) - EU(\beta_3^*)\} \Rightarrow_d N(0, Var(U_i(\beta_3^*))).$$

So we have:

$$\Sigma_{\beta_3} = I_{\beta_3}^{-1} \{Var(U(\beta_3))\} I_{\beta_3}^{-T}.$$

By assumption, $\hat{\Sigma}_{\beta_3} = \Sigma_{\beta_3} + o_p(1)$ which is a consistent estimator of Σ_{β_3}.

By applying Lemma 4, we have:

$$\hat{\Sigma}_{\gamma_3} = \hat{I}_{\gamma_3}^{-1} \left(\hat{I}_{\gamma_3} + \frac{n_2}{n_1} \hat{I}_{\beta_3} \hat{\Sigma}_{\beta_3} \hat{I}_{\beta_3}^{T} \right) \hat{I}_{\gamma_3}^{-T}.$$

□

Lemma 8. *Assume γ_4^* is the unique solution to $EU(\gamma_4) = 0$ for an estimating function $U(\gamma_4)$. Solving the estimating equation $0 = n_1^{-1} \sum_{i=1}^{n_1} U_i(\hat{\gamma}_4)$, we have $\sqrt{n_1}(\hat{\gamma}_4 - \gamma_4^*) \to N(0, \Sigma_{\gamma_4})$ where,*

$$\Sigma_{\gamma_4} = I_{\gamma_4}^{-1}(Var(U(\gamma_4)))I_{\gamma_4}^{-T},$$

where $I_{\gamma_4} = E\left(-\frac{\partial U(\gamma_4)}{\partial \gamma_4}|_{\gamma_4^*}\right)$. This variance can be consistently estimated by:

$$\hat{\Sigma}_{\gamma_4} = \hat{I}_{\gamma_4}^{-1}(Var(U(\gamma_4)))\hat{I}_{\gamma_4}^{-T},$$

where $\hat{I}_{\gamma_4} = n_1^{-1} \sum_{i=1}^{n_1} \left(-\frac{\partial U_i(\gamma_4)}{\partial \gamma_4}|_{\hat{\gamma}_4}\right)$ and $\hat{\Sigma}_{\gamma_4}$ is a consistent estimator of Σ_{γ_4}.

Proof. The asymptotic of $\hat{\gamma}_4$ can be derived as below.

$$\begin{aligned} 0 &= \sum_{i=1}^{n_1} U_i(\hat{\gamma}_4) \\ &= \sum_{i=1}^{n_1} \left\{ U_i(\gamma_4^*) + \frac{\partial U}{\partial \gamma_4}|_{\gamma_4^*}(\hat{\gamma}_4 - \gamma_4^*) + o(||\hat{\gamma}_4 - \gamma_4^*||) \right\}, \end{aligned}$$

where $U_i(\gamma_4^*) = (1, Q_i, \boldsymbol{V}_i^T)^T X_i^* - (1, Q_i, \boldsymbol{V}_i^T)^T(1, Q_i, \boldsymbol{V}_i^T)\gamma_4^*$.

With Theorem 1,

$$n_1^{-1}\sum_{i=1}^{n_1}\frac{\partial U_i(\gamma_4)}{\partial\gamma_4}|_{\hat{\gamma}_4} = E\left(\frac{\partial U}{\partial\gamma_4}|_{\hat{\gamma}_4}\right) + o_p(1),$$

and with Theorem 2, we have:

$$E\left(\frac{\partial U}{\partial\gamma_4}|_{\hat{\gamma}_4}\right) = E\left(\frac{\partial U}{\partial\gamma_4}|_{\gamma_4^*}\right) + o_p(1).$$

So we have:

$$n_1^{-1}\sum_{i=1}^{n_1}\frac{\partial U_i(\gamma_4)}{\partial\gamma_4}|_{\hat{\gamma}_4} = -I_{\gamma_4} + o_p(1).$$

Plug in the Taylor expansion, notice that $EU(\gamma_4^*) = 0$, we have:

$$\begin{aligned}0 \quad = \quad & n_1^{-1}\sum_{i=1}^{n_1}\{U_i(\gamma_4^*) - EU(\gamma_4^*)\} - I_{\gamma_4}(\hat{\gamma}_4 - \gamma_4^*) \\ & + o_p(1)||\hat{\gamma}_4 - \gamma_4^*||.\end{aligned}$$

So we have:

$$\sqrt{n_1}(\hat{\gamma}_4 - \gamma_4^*) = I_{\gamma_4}^{-1}\left\{n_1^{-1/2}\sum_{i=1}^{n_1}U_i(\gamma_4^*)\right\} + o_p(1).$$

By Theorem 3, we have:

$$n_1^{-1/2}\sum_{i=1}^{n_2}\{U_i(\gamma_4^*) - EU(\gamma_4^*)\} \Rightarrow_d N(0, Var(U_i(\gamma_4^*))).$$

So we have:

$$\Sigma_{\gamma_4} = I_{\gamma_4}^{-1}\left\{Var(U(\gamma_4))\right\} I_{\gamma_4}^{-T}.$$

By assumption, $\hat{\Sigma}_{\gamma_4} = \Sigma_{\gamma_4} + o_p(1)$ which is a consistent estimator of Σ_{γ_4}. □

Theorem 4. *The asymptotic bias in Method 1 is associated with ρ and δ with a form such that the bias corrected association parameters, θ_z^* and θ_v^*, are $\theta_z^* = \rho^{-1}\theta_z$ where $\rho = R^2_{X,\boldsymbol{W}|\boldsymbol{V}}$,and $\theta_v^* = \theta_v - \frac{(1-\rho)\delta}{\rho}\theta_z$ when a linear function form, $E(X|\boldsymbol{V}) = \boldsymbol{V}\delta$, exists.*

Proof. To see the asymptotic bias of Method 1 and get the bias corrected parameters, we consider the first step regression model of X^* on $(\boldsymbol{W}, \boldsymbol{V})$, then we have:

$$\begin{aligned}
\hat{X} &= E(X^*|\boldsymbol{W},\boldsymbol{V}) = E(X|\boldsymbol{W},\boldsymbol{V}) \\
&= E(X|\boldsymbol{V}) + \{\boldsymbol{W} - E(\boldsymbol{W}|\boldsymbol{V})\}^T \Sigma^{-1}_{\boldsymbol{WW}|\boldsymbol{V}}\Sigma^{T}_{X\boldsymbol{W}|\boldsymbol{V}}\,.
\end{aligned}$$

Now we compute $E(\hat{X}|Q,\boldsymbol{V})$ to see how it is biased away from $E(X|Q,\boldsymbol{V})$. We have:

$$\begin{aligned}
E(\hat{X}|Q,\boldsymbol{V}) &= E\left\{E\left(\hat{X}|X,Q,\boldsymbol{V}\right)|Q,\boldsymbol{V}\right\} = E\left\{E\left(\hat{X}|X,\boldsymbol{V}\right)|Q,\boldsymbol{V}\right\} \\
&= E\left[E\left[E(X|\boldsymbol{V}) + \{\boldsymbol{W} - E(\boldsymbol{W}|\boldsymbol{V})\}^T\Sigma^{-1}_{\boldsymbol{WW}|\boldsymbol{V}}\Sigma^{T}_{X\boldsymbol{W}|\boldsymbol{V}}|X,\boldsymbol{V}\right]|Q,\boldsymbol{V}\right] \\
&= E\left[E(X|\boldsymbol{V}) + \{E(\boldsymbol{W}|X,\boldsymbol{V}) - E(\boldsymbol{W}|\boldsymbol{V})\}^T\Sigma^{-1}_{\boldsymbol{WW}|\boldsymbol{V}}\Sigma^{T}_{X\boldsymbol{W}|\boldsymbol{V}}|Q,\boldsymbol{V}\right] \\
&= E\left[E(X|\boldsymbol{V}) + \{X - E(X|\boldsymbol{V})\}\Sigma^{-1}_{XX|\boldsymbol{V}}\Sigma_{X\boldsymbol{W}|\boldsymbol{V}}\Sigma^{-1}_{\boldsymbol{WW}|\boldsymbol{V}}\Sigma^{T}_{X\boldsymbol{W}|\boldsymbol{V}}|Q,\boldsymbol{V}\right] \\
&= E\left[E(X|\boldsymbol{V}) + \{X - E(X|\boldsymbol{V})\}\rho_{\boldsymbol{V}}|Q,\boldsymbol{V}\right] \\
&= \rho_{\boldsymbol{V}}E(X|Q,\boldsymbol{V}) + (1-\rho_{\boldsymbol{V}})E(X|\boldsymbol{V}).
\end{aligned}$$

When $\rho_{\boldsymbol{V}}$ is a constant over $\boldsymbol{V}$, we simply denote it as ρ and we have $\rho\theta_z^* = \theta_z$, or $\theta_z^* = \rho^{-1}\theta_z$ with appropriate adjustment for V. Explicitly, we have:

$$\begin{aligned}\rho &= 1 - \Sigma_{XX|\boldsymbol{V}}^{-1}\left(\Sigma_{XX|\boldsymbol{V}} - \Sigma_{X\boldsymbol{W}|\boldsymbol{V}}\Sigma_{\boldsymbol{W}\boldsymbol{W}|\boldsymbol{V}}^{-1}\Sigma_{X\boldsymbol{W}|\boldsymbol{V}}^{T}\right)\\ &= 1 - \frac{Var(X|\boldsymbol{W},\boldsymbol{V})}{Var(X|\boldsymbol{V})} = R^2_{X,\boldsymbol{W}|\boldsymbol{V}}.\end{aligned}$$

If we further have $E(X|\boldsymbol{V}) = \boldsymbol{V}\delta$ is a linear function of V, then $(1-\rho)\delta\theta_z^* + \theta_v^* = \theta_v$, or $\theta_v^* = \theta_v - \frac{(1-\rho)\delta}{\rho}\theta_z$. □

Theorem 5. *With* $\frac{n_3}{n_2} \to C_2$ *and* $\frac{n_2}{n_1} \to C_1$*, we have* $\sqrt{n_3}(\hat{\theta}_1 - \theta_1^*) \to N(0, \Sigma_{\theta_1})$ *where* $\Sigma_{\theta_1} = I_{\theta_1}^{-1}(I_{\theta_1} + C_2 I_{\gamma_1}\Sigma_{\gamma_1}I_{\gamma_1}^T)I_{\theta_1}^{-T}$ *can be consistently estimated by:* $\hat{\Sigma}_{\theta_1} = \hat{I}_{\theta_1}^{-1}(\hat{I}_{\theta_1} + \frac{n_3}{n_2}\hat{I}_{\gamma_1}\hat{\Sigma}_{\gamma_1}\hat{I}_{\gamma_1}^T)\hat{I}_{\theta_1}^{-T}$ *for Method 1.*
With $\frac{n_3}{n_2} \to C_2$ *and* $\frac{n_2}{n_1} \to C_1$*, we have* $\sqrt{n_3}(\hat{\theta}_2 - \theta_2^*) \to N(0, \Sigma_{\theta_2})$ *where* $\Sigma_{\theta_2} = I_{\theta_2}^{-1}(I_{\theta_2} + C_2 I_{\gamma_2}\Sigma_{\gamma_2}I_{\gamma_2}^T)I_{\theta_2}^{-T}$ *can be consistently estimated by:* $\hat{\Sigma}_{\theta_2} = \hat{I}_{\theta_2}^{-1}(\hat{I}_{\theta_2} + \frac{n_3}{n_2}\hat{I}_{\gamma_2}\hat{\Sigma}_{\gamma_2}\hat{I}_{\gamma_2}^T)\hat{I}_{\theta_2}^{-T}$ *for Method 2.*
With $\frac{n_3}{n_2} \to C_2$ *and* $\frac{n_2}{n_1} \to C_1$*, we have* $\sqrt{n_3}(\hat{\theta}_3 - \theta_3^*) \to N(0, \Sigma_{\theta_3})$ *where* $\Sigma_{\theta_3} = I_{\theta_3}^{-1}(I_{\theta_3} + C_2 I_{\gamma_3}\Sigma_{\gamma_3}I_{\gamma_3}^T)I_{\theta_3}^{-T}$ *can be consistently estimated by:* $\hat{\Sigma}_{\theta_3} = \hat{I}_{\theta_3}^{-1}(\hat{I}_{\theta_3} + \frac{n_3}{n_2}\hat{I}_{\gamma_3}\hat{\Sigma}_{\gamma_3}\hat{I}_{\gamma_3}^T)\hat{I}_{\theta_3}^{-T}$ *for Method 3.*
With $\frac{n_3}{n_2} \to C_2$ *and* $\frac{n_2}{n_1} \to C_1$*, we have* $\sqrt{n_3}(\hat{\theta}_4 - \theta_4^*) \to N(0, \Sigma_{\theta_4})$ *where* $\Sigma_{\theta_4} = I_{\theta_4}^{-1}(I_{\theta_4} + C_2 I_{\gamma_4}\Sigma_{\gamma_4}I_{\gamma_4}^T)I_{\theta_4}^{-T}$ *can be consistently estimated by:* $\hat{\Sigma}_{\theta_4} = \hat{I}_{\theta_4}^{-1}(\hat{I}_{\theta_4} + \frac{n_3}{n_2}\hat{I}_{\gamma_4}\hat{\Sigma}_{\gamma_4}\hat{I}_{\gamma_4}^T)\hat{I}_{\theta_4}^{-T}$ *for Method 4.*

Proof. By applying Lemma 2 and Lemma 5, the asymptotic Σ_{θ_1} can be derived as
$\hat{\Sigma}_{\theta_1} = \hat{I}_{\theta_1}^{-1}(\hat{I}_{\theta_1} + \frac{n_3}{n_2}\hat{I}_{\gamma_1}\hat{\Sigma}_{\gamma_1}\hat{I}_{\gamma_1}^T)\hat{I}_{\theta_1}^{-T}$.
By applying Lemma 2 and Lemma 6, the asymptotic Σ_{θ_2} can be derived as
$\hat{\Sigma}_{\theta_2} = \hat{I}_{\theta_2}^{-1}(\hat{I}_{\theta_2} + \frac{n_3}{n_2}\hat{I}_{\gamma_2}\hat{\Sigma}_{\gamma_2}\hat{I}_{\gamma_2}^T)\hat{I}_{\theta_2}^{-T}$.
By applying Lemma 2 and Lemma 7, the asymptotic Σ_{θ_3} can be derived as
$\hat{\Sigma}_{\theta_3} = \hat{I}_{\theta_3}^{-1}(\hat{I}_{\theta_3} + \frac{n_3}{n_2}\hat{I}_{\gamma_3}\hat{\Sigma}_{\gamma_3}\hat{I}_{\gamma_3}^T)\hat{I}_{\theta_3}^{-T}$.

By applying Lemma 2 and Lemma 8, the asymptotic Σ_{θ_4} can be derived as
$\hat{\Sigma}_{\theta_4} = \hat{I}_{\theta_4}^{-1}(\hat{I}_{\theta_4} + \frac{n_3}{n_2}\hat{I}_{\gamma_4}\hat{\Sigma}_{\gamma_4}\hat{I}_{\gamma_4}^{T})\hat{I}_{\theta_4}^{-T}$. □

Theorem 6. *With $\frac{n_3}{n_2} \to C_2$ and $\frac{n_2}{n_1} \to C_1$, we have $\sqrt{n_3}(\hat{\theta}_1 - \theta_1^*) \to N(0, \Sigma_{\theta_1})$ where $\Sigma_{\theta_1} = I_{\theta_1}^{-1}(I_{\theta_1} + C_2 I_{\gamma_1}\Sigma_{\gamma_1}I_{\gamma_1}^{T})I_{\theta_1}^{-T}$ can be consistently estimated by: $\hat{\Sigma}_{\theta_1} = \hat{I}_{\theta_1}^{-1}(\hat{I}_{\theta_1} + \frac{n_3}{n_2}\hat{I}_{\gamma_1}\hat{\Sigma}_{\gamma_1}\hat{I}_{\gamma_1}^{T})\hat{I}_{\theta_1}^{-T}$ for Method 1.*
With $\frac{n_3}{n_2} \to C_2$ and $\frac{n_2}{n_1} \to C_1$, we have $\sqrt{n_3}(\hat{\theta}_2 - \theta_2^*) \to N(0, \Sigma_{\theta_2})$ where $\Sigma_{\theta_2} = I_{\theta_2}^{-1}(I_{\theta_2} + C_2 I_{\gamma_2}\Sigma_{\gamma_2}I_{\gamma_2}^{T})I_{\theta_2}^{-T}$ can be consistently estimated by: $\hat{\Sigma}_{\theta_2} = \hat{I}_{\theta_2}^{-1}(\hat{I}_{\theta_2} + \frac{n_3}{n_2}\hat{I}_{\gamma_2}\hat{\Sigma}_{\gamma_2}\hat{I}_{\gamma_2}^{T})\hat{I}_{\theta_2}^{-T}$ for Method 2.
With $\frac{n_3}{n_2} \to C_2$ and $\frac{n_2}{n_1} \to C_1$, we have $\sqrt{n_3}(\hat{\theta}_3 - \theta_3^*) \to N(0, \Sigma_{\theta_3})$ where $\Sigma_{\theta_3} = I_{\theta_3}^{-1}(I_{\theta_3} + C_2 I_{\gamma_3}\Sigma_{\gamma_3}I_{\gamma_3}^{T})I_{\theta_3}^{-T}$ can be consistently estimated by: $\hat{\Sigma}_{\theta_3} = \hat{I}_{\theta_3}^{-1}(\hat{I}_{\theta_3} + \frac{n_3}{n_2}\hat{I}_{\gamma_3}\hat{\Sigma}_{\gamma_3}\hat{I}_{\gamma_3}^{T})\hat{I}_{\theta_3}^{-T}$ for Method 3.
With $\frac{n_3}{n_2} \to C_2$ and $\frac{n_2}{n_1} \to C_1$, we have $\sqrt{n_3}(\hat{\theta}_4 - \theta_4^*) \to N(0, \Sigma_{\theta_4})$ where $\Sigma_{\theta_4} = I_{\theta_4}^{-1}(I_{\theta_4} + C_2 I_{\gamma_4}\Sigma_{\gamma_4}I_{\gamma_4}^{T})I_{\theta_4}^{-T}$ can be consistently estimated by: $\hat{\Sigma}_{\theta_4} = \hat{I}_{\theta_4}^{-1}(\hat{I}_{\theta_4} + \frac{n_3}{n_2}\hat{I}_{\gamma_4}\hat{\Sigma}_{\gamma_4}\hat{I}_{\gamma_4}^{T})\hat{I}_{\theta_4}^{-T}$ for Method 4.

Proof. By applying Lemma 3 and Lemma 5, the asymptotic Σ_{θ_1} can be derived as
$\hat{\Sigma}_{\theta_1} = \hat{I}_{\theta_1}^{-1}(\hat{I}_{\theta_1} + \frac{n_3}{n_2}\hat{I}_{\gamma_1}\hat{\Sigma}_{\gamma_1}\hat{I}_{\gamma_1}^{T})\hat{I}_{\theta_1}^{-T}$.
By applying Lemma 3 and Lemma 6, the asymptotic Σ_{θ_2} can be derived as
$\hat{\Sigma}_{\theta_2} = \hat{I}_{\theta_2}^{-1}(\hat{I}_{\theta_2} + \frac{n_3}{n_2}\hat{I}_{\gamma_2}\hat{\Sigma}_{\gamma_2}\hat{I}_{\gamma_2}^{T})\hat{I}_{\theta_2}^{-T}$.
By applying Lemma 3 and Lemma 7, the asymptotic Σ_{θ_3} can be derived as
$\hat{\Sigma}_{\theta_3} = \hat{I}_{\theta_3}^{-1}(\hat{I}_{\theta_3} + \frac{n_3}{n_2}\hat{I}_{\gamma_3}\hat{\Sigma}_{\gamma_3}\hat{I}_{\gamma_3}^{T})\hat{I}_{\theta_3}^{-T}$.
By applying Lemma 3 and Lemma 8, the asymptotic Σ_{θ_4} can be derived as
$\hat{\Sigma}_{\theta_4} = \hat{I}_{\theta_4}^{-1}(\hat{I}_{\theta_4} + \frac{n_3}{n_2}\hat{I}_{\gamma_4}\hat{\Sigma}_{\gamma_4}\hat{I}_{\gamma_4}^{T})\hat{I}_{\theta_4}^{-T}$. □

Appendix B: Technical Details for Chapter 3

Lemma 9. *Assume $n_3/n_2 \to C_2 < \infty$ and θ^* is the unique solution to:*

$$0 = U(\theta,\gamma^*) = E\int_0^\tau \left[\begin{pmatrix} \boldsymbol{Z}^* \\ \boldsymbol{V} \end{pmatrix} - \frac{E\left[\begin{pmatrix} \boldsymbol{Z}^* \\ \boldsymbol{V} \end{pmatrix} Y(t)\exp\left\{ (\boldsymbol{Z}^{*T}, \boldsymbol{V}^T)\theta \right\} \right]}{E\left[Y(t)\exp\left\{ (\boldsymbol{Z}^{*T}, \boldsymbol{V}^T)\theta \right\} \right]} \right] dN(t),$$

where $\boldsymbol{Z}^* = \boldsymbol{X}\gamma^*$. If $\hat{\theta}$ solve the estimating equation:

$$0 = n_3^{-1}\sum_{i=1}^{n_3} U_i(\theta,\hat{\gamma}) = n_3^{-1}\sum_{i=1}^{n_3}\int_0^\tau \left[\begin{pmatrix} \hat{\boldsymbol{Z}}_{\boldsymbol{i}} \\ \boldsymbol{V}_i \end{pmatrix} - \sum_j \frac{Y_j(t)\exp\left\{ (\hat{\boldsymbol{Z}}_{\boldsymbol{j}}^T, \boldsymbol{V}_j{}^T)\theta \right\}}{\sum_k Y_k(t)\exp\left\{ (\hat{\boldsymbol{Z}}_{\boldsymbol{k}}^T, \boldsymbol{V}_k{}^T)\theta \right\}} \begin{pmatrix} \hat{\boldsymbol{Z}}_{\boldsymbol{j}} \\ \boldsymbol{V}_j \end{pmatrix} \right] dN_i(t),$$

where $\hat{\boldsymbol{Z}}_{\boldsymbol{i}} = \boldsymbol{X}_{\boldsymbol{i}}\hat{\gamma}$ and $\sqrt{n_2}(vec(\hat{\gamma}) - vec(\gamma^*)) \to N(0,\Sigma_\gamma)$, then we have $\sqrt{n_3}(\hat{\theta} - \theta^*) \to N(0,\Sigma_\theta)$ where,

$$\Sigma_\theta = I_\theta^{-1}(I_\theta + C_2 I_\gamma \Sigma_\gamma I_\gamma^T) I_\theta^{-T},$$

$I_\theta = -E\left(\frac{\partial U(\theta,\gamma)}{\partial\theta}|_{\theta^*,\gamma^*}\right)$ and $I_\gamma = -E\left(\frac{\partial U(\theta,\gamma)}{\partial vec(\gamma)}|_{\theta^*,\gamma^*}\right)$. This variance can be consistently estimated by:

$$\hat{\Sigma}_\theta = \hat{I}_\theta^{-1}\left(\hat{I}_\theta + \frac{n_3}{n_2}\hat{I}_\gamma\hat{\Sigma}_\gamma\hat{I}_\gamma^T\right)\hat{I}_\theta^{-T},$$

where $\hat{I}_\theta = -n_3^{-1}\sum_{i=1}^{n_3}\frac{\partial U_i(\theta,\gamma)}{\partial\theta}|_{\hat{\theta},\hat{\gamma}}$, $\hat{I}_\gamma = -n_3^{-1}\sum_{i=1}^{n_3}\frac{\partial U_i(\theta,\gamma)}{\partial vec(\gamma)}|_{\hat{\theta},\hat{\gamma}}$, $\hat{\Sigma}_\gamma$ is a consistent estimator of Σ_γ and $\frac{n3}{n2} \to C_2$.

Proof. For this specific estimating function from Cox regression, we have:

$$I_\theta \;=\; E\int_0^\tau \left[\begin{pmatrix} \boldsymbol{Z}_{\boldsymbol{i}}^* \\ \boldsymbol{V}_i \end{pmatrix} - \sum_j \frac{Y_j(t)\exp\left\{(\boldsymbol{Z}_j^{*T}, \boldsymbol{V}_j{}^T)\theta\right\}}{\sum_k Y_k(t)\exp\left\{(\boldsymbol{Z}_k^{*T}, \boldsymbol{V}_k{}^T)\theta\right\}} \begin{pmatrix} \boldsymbol{Z}_j^* \\ \boldsymbol{V}_j \end{pmatrix}\right]^{\otimes 2} dN_i(t),$$

where $a^{\otimes 2} = aa^T$.

$$I_\gamma \;=\; -E\left[\int_0^\tau \left\{\begin{pmatrix} I_m \otimes \mathbb{X} \\ 0 \end{pmatrix} - \frac{A(\theta,t)}{s^{(0)}(\theta,t)} + \frac{s^{(1)}(\theta,t)G(\theta,t)}{s^{(0)}(\theta,t)^2}\right\} dN(t)\right].$$

Here we would like to comment that for Method 2-4, $\boldsymbol{Z}^*$ has the same expression $E(\boldsymbol{Z}|\boldsymbol{Q},\boldsymbol{V})$ and thus, the I_θ are the same though their estimated version $\hat{I}_\theta$ are different.

So the expectation can be consistently estimated by:

$$\hat{I}_\theta = -n_3^{-1}\sum_i \Delta_i \left[\begin{pmatrix} \hat{\boldsymbol{Z}}_{\boldsymbol{i}} \\ \boldsymbol{V}_i \end{pmatrix} - \sum_j \frac{Y_j(T_i)\exp\left\{(\hat{\boldsymbol{Z}}_{\boldsymbol{j}}^T, \boldsymbol{V}_j{}^T)\theta\right\}}{\sum_k Y_k(T_i \exp\left\{(\hat{\boldsymbol{Z}}_{\boldsymbol{k}}^T, \boldsymbol{V}_k{}^T)\theta\right\}} \begin{pmatrix} \hat{\boldsymbol{Z}}_{\boldsymbol{j}} \\ \boldsymbol{V}_j \end{pmatrix}\right]^{\otimes 2}.$$

As for $\hat{I}_\gamma$, we have:

$$E\left(\frac{\partial U}{\partial vec(\gamma)}\right) = E\left[\int_0^\tau \left\{\begin{pmatrix} I_m \otimes \mathbb{X} \\ 0 \end{pmatrix} - \frac{A(\theta,t)}{s^{(0)}(\theta,t)} + \frac{s^{(1)}(\theta,t)G(\theta,t)}{s^{(0)}(\theta,t)^2}\right\} dN(t)\right],$$

where,

$$
\begin{aligned}
A(\theta, t) &= E\left[Y(t)\begin{pmatrix} I_m + \boldsymbol{Z}^*\theta_Z^T \\ \boldsymbol{V}\theta_Z^T \end{pmatrix}\exp\left\{(\boldsymbol{Z}^{*T}, \boldsymbol{V}^T)\theta\right\} \otimes \mathbb{X}\right], \\
G(\theta, t) &= E\left[Y(t)\exp\left\{(\boldsymbol{Z}^{*T}, \boldsymbol{V}^T)\theta\right\}\theta_Z^T \otimes \mathbb{X}\right], \\
s^{(0)}(\theta, t) &= n_3^{-1}\sum_i Y_i(t)\exp\left\{(\boldsymbol{Z}_i^{*T}, \boldsymbol{V}_i{}^T)\theta\right\}, \\
s^{(1)}(\theta, t) &= n_3^{-1}\sum_i Y_i(t)\exp\left\{(\boldsymbol{Z}_i^{*T}, \boldsymbol{V}_i{}^T)\theta\right\}\begin{pmatrix} \boldsymbol{Z}_i^* \\ \boldsymbol{V}_i \end{pmatrix},
\end{aligned}
$$

which can be consistently estimated by:

$$
\begin{aligned}
\hat{A}(\theta, t) &= n_3^{-1}\sum_i\left[Y_i(t)\begin{pmatrix} I_m + \hat{\boldsymbol{Z}}_i\theta_Z^T \\ \boldsymbol{V}_i\theta_Z^T \end{pmatrix}\exp\left\{(\hat{\boldsymbol{Z}}_i^T, \boldsymbol{V}_i{}^T)\theta\right\} \otimes \mathbb{X}_i\right], \\
\hat{G}(\theta, t) &= n_3^{-1}\sum_i\left[Y_i(t)\exp\left\{(\hat{\boldsymbol{Z}}_i^T, \boldsymbol{V}_i{}^T)\theta\right\}\theta_Z^T \otimes \mathbb{X}_i\right],
\end{aligned}
$$

and,

$$
\hat{I}_\gamma = -E\left(\frac{\partial U}{\partial vec(\gamma)}\right) = -n_3^{-1}\sum_i \Delta_i\left\{\begin{pmatrix} I_m \otimes \mathbb{X}_i \\ 0 \end{pmatrix} - \frac{\hat{A}(\hat{\theta}, T_i)}{s^{(0)}(\hat{\theta}, T_i)} + \frac{\hat{s}^{(1)}(\hat{\theta}, T_i)\hat{G}(\hat{\theta}, T_i)}{\hat{s}^{(0)}(\hat{\theta}, T_i)^2}\right\},
$$

where,

$$\begin{aligned}\hat{s}^{(0)}(\theta,t) &= n_3^{-1}\sum_i Y_i(t)\exp\left\{(\hat{\boldsymbol{Z}}_i^T,\boldsymbol{V}_i{}^T)\theta\right\},\\ \hat{s}^{(1)}(\theta,t) &= n_3^{-1}\sum_i Y_i(t)\exp\left\{(\hat{\boldsymbol{Z}}_i^T,\boldsymbol{V}_i{}^T)\theta\right\}\begin{pmatrix}\hat{\boldsymbol{Z}}_i\\ \boldsymbol{V}_i\end{pmatrix}.\end{aligned}$$

So by Lemma 1, we obtain the asymptotic result for θ. □

Lemma 10. *Assume $n_3/n_2 \to C_2 < \infty$ and θ^* is the unique solution to:*

$$0 = U = E\left[(1,\boldsymbol{Z}^{*T},\boldsymbol{V}^T)^T Y - (1,\boldsymbol{Z}^{*T},\boldsymbol{V}^T)^T g^{-1}[(1,\boldsymbol{Z}^{*T},\boldsymbol{V}^T)\theta]\right].$$

where $Z^* = X\gamma^*$. If $\hat{\theta}$ solve the estimating equation:

$$0 = 1/n_3\sum_i U_i = n_3^{-1}\sum_{i=1}^{n_3}\left[(1,\hat{\boldsymbol{Z}}_i^T,\boldsymbol{V}_i{}^T)^T Y_i - (1,\hat{\boldsymbol{Z}}_i^T,\boldsymbol{V}_i{}^T)^T g^{-1}[(1,\hat{\boldsymbol{Z}}_i^T,\boldsymbol{V}_i{}^T)\theta]\right],$$

where $\hat{\boldsymbol{Z}}_i = \boldsymbol{X_i}\hat{\gamma}$ and $\sqrt{n_2}(vec(\hat{\gamma}) - vec(\gamma^*)) \to N(0,\Sigma_\gamma)$, then we have $\sqrt{n_3}(\hat{\theta} - \theta^*) \to N(0,\Sigma_\theta)$ where,

$$\Sigma_\theta = I_\theta^{-1}(I_\theta + C_2 I_\gamma \Sigma_\gamma I_\gamma^T) I_\theta^{-T},$$

where $I_\theta = -E\frac{\partial U}{\partial\theta}$ and $I_\gamma = -E\frac{\partial U}{\partial vec(\gamma)}$. This variance can be consistently estimated by:

$$\hat{\Sigma}_\theta = \hat{I}_\theta^{-1}\{\hat{I}_\theta + \frac{n_3}{n_2}\hat{I}_\gamma\hat{\Sigma}_\gamma\hat{I}_\gamma^T\}\hat{I}_\theta^{-T},$$

where $\hat{I}_\theta = -\frac{1}{n_3}\sum_i\frac{\partial U_i}{\partial\theta}$, $\hat{I}_\gamma = -\frac{1}{n_3}\sum_i\frac{\partial U_i}{\partial vec(\gamma)}$ and $\hat{\Sigma}_\gamma$ is a consistent estimator of Σ_γ.

Proof. For this specific estimating function from generalized linear model (GLM), we have:

$$I_\theta = -E\frac{\partial U}{\partial \theta} \quad = \quad -E\sum_i \left\{ \begin{pmatrix} \mathbf{1} \\ \boldsymbol{Z}_i^* \\ \boldsymbol{V}_i \end{pmatrix} \frac{\partial g^{-1}(\eta_i)}{\partial(\eta_i)} (1, \boldsymbol{Z}_i^{*T}, \boldsymbol{V}_i{}^T) \right\},$$

$$\begin{aligned} I_\gamma &= -E\frac{\partial U}{\partial vec(\gamma)}, \\ &= -E\sum_i \left\{ \begin{pmatrix} 0 \\ I_m \otimes \mathbb{X}_i(Y_i - g^{-1}(\eta_i)) \\ 0 \end{pmatrix} - \begin{pmatrix} \mathbf{1} \\ \boldsymbol{Z}_i^* \\ \boldsymbol{V}_i \end{pmatrix} \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} \theta_z^T \otimes \mathbb{X}_i \right\}, \\ &= \begin{pmatrix} \mathbf{1} \\ \boldsymbol{Z}_i^* \\ \boldsymbol{V}_i \end{pmatrix} \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} \theta_z^T \otimes \mathbb{X}_i, \end{aligned}$$

where $\eta_i = (1, \boldsymbol{Z}_i^{*T}, \boldsymbol{V}_i^T)\theta$ and g is the corresponding link function.

$$\hat{I}_\theta \quad = \quad -n_3{}^{-1}\sum_i \left\{ \begin{pmatrix} \mathbf{1} \\ \hat{\boldsymbol{Z}}_i \\ \boldsymbol{V}_i \end{pmatrix} \frac{\partial g^{-1}(\hat{\eta}_i)}{\partial \hat{\eta}_i} \frac{\partial \hat{\eta}_i}{\partial \theta} \right\},$$

$$\hat{I}_\gamma = n_3{}^{-1}\sum_i \begin{pmatrix} \mathbf{1} \\ \hat{\boldsymbol{Z}}_i \\ \boldsymbol{V}_i \end{pmatrix} \frac{\partial g^{-1}(\hat{\eta}_i)}{\partial \eta_i} \hat{\theta}_z^T \otimes \mathbb{X}_i,$$

where $\hat{\eta}_i = (1, \hat{\boldsymbol{Z}}_i^T, \boldsymbol{V}_i^T)\theta$ and g is the corresponding link function. So by Lemma 1, we obtain the asymptotic result for θ. □

Corollary 3. *Assume $n_3/n_2 \rightarrow C_2 < \infty$ and θ^* is the unique solution to:*

$$0 = U = E\left[(1, \boldsymbol{Z}_i^{*T}, \boldsymbol{V}_i{}^T)^T Y_i - (1, \boldsymbol{Z}_i^{*T}, \boldsymbol{V}_i{}^T)^T (1, \boldsymbol{Z}_i^{*T}, \boldsymbol{V}_i{}^T)\theta\right],$$

where $Z^* = X\gamma^*$. If $\hat{\theta}$ solve the estimating equation:

$$0 = \sum U_i = n_3^{-1} \sum_{i=1} \left[(1, \hat{\boldsymbol{Z}}_i^T, \boldsymbol{V}_i{}^T)^T Y_i - (1, \hat{\boldsymbol{Z}}_i^T, \boldsymbol{V}_i{}^T)^T (1, \hat{\boldsymbol{Z}}_i^T, \boldsymbol{V}_i{}^T)\theta\right],$$

where $\hat{\boldsymbol{Z}}_i = \boldsymbol{X}_i\hat{\gamma}$ and $\sqrt{n_2}(vec(\hat{\gamma}) - vec(\gamma^*)) \rightarrow N(0, \Sigma_\gamma)$, then we have $\sqrt{n_3}(\hat{\theta} - \theta^*) \rightarrow N(0, \Sigma_\theta)$ where,

$$\Sigma_\theta = I_\theta^{-1}(I_\theta + C_2 I_\gamma \Sigma_\gamma I_\gamma^T) I_\theta^{-T},$$

where $I_\theta = -E\frac{\partial U}{\partial \theta}$ and $I_\gamma = -E\frac{\partial U}{\partial vec(\gamma)}$. This variance can be consistently estimated by

$$\hat{\Sigma}_\theta = \hat{I}_\theta^{-1}\{\hat{I}_\theta + \frac{n_3}{n_2}\hat{I}_\gamma \hat{\Sigma}_\gamma \hat{I}_\gamma^T\}\hat{I}_\theta^{-T},$$

where $\hat{I}_\theta = -\frac{1}{n_3}\sum_i \frac{\partial U_i}{\partial \theta}$, $\hat{I}_\gamma = -\frac{1}{n_3}\sum_i \frac{\partial U_i}{\partial vec(\gamma)}$ and $\hat{\Sigma}_\gamma$ is a consistent estimator of Σ_γ.

Proof. Plug in $g(x) = x$ for linear regression form, we have:

$$I_\theta = -E\left\{\begin{pmatrix} \mathbf{1} \\ \boldsymbol{Z}_i^* \\ \boldsymbol{V}_i \end{pmatrix} (\mathbf{1}, \boldsymbol{Z}_i^{*T}, \boldsymbol{V}_i{}^T)\right\},$$

$$I_\gamma = E\begin{pmatrix} \mathbf{1} \\ \boldsymbol{Z}^* \\ \boldsymbol{V}_i \end{pmatrix} \theta_z^T \otimes \mathbb{X},$$

$$\hat{I}_\theta \;=\; -{n_3}^{-1}\sum_i \begin{pmatrix} \mathbf{1} \\ \hat{\boldsymbol{Z}}_i \\ \boldsymbol{V}_i \end{pmatrix} (1, \hat{Z}_i^T, {\boldsymbol{V}_i}^T),$$

$$\hat{I}_\gamma = {n_3}^{-1}\sum_i \begin{pmatrix} \mathbf{1} \\ \hat{\boldsymbol{Z}}_i \\ \boldsymbol{V}_i \end{pmatrix} \hat{\theta}_z^T \otimes \mathbb{X}_i.$$

By applying Lemma 3, we obtain the asymptotic for θ in linear regression setting. □

Corollary 4. *Assume $n_3/n_2 \to C_2 < \infty$ and θ^* is the unique solution to:*

$$0 = U = \sum_{i=1}\left[(1, {Z_i^*}^T, {\boldsymbol{V}_i}^T)^T Y_i - (1, {Z_i^*}^T, {\boldsymbol{V}_i}^T)^T \frac{exp(\eta_i)}{(1+exp(\eta_i))}\right],$$

where $Z^* = X\gamma^*$. If $\hat{\theta}$ solve the estimating equation:

$$0 = \sum U_i = n_3^{-1}\sum_{i=1}^{n_3}\left[(1, \hat{Z}_i^T, {\boldsymbol{V}_i}^T)^T Y_i - (1, \hat{Z}_i^T, {\boldsymbol{V}_i}^T)^T \frac{exp(\eta_i)}{(1+exp(\eta_i))}\right],$$

where $\hat{Z}_i = X_i\hat{\gamma}$ and $\sqrt{n_2}(vec(\hat{\gamma}) - vec(\gamma^*)) \to N(0, \Sigma_\gamma)$, then we have $\sqrt{n_3}(\hat{\theta} - \theta^*) \to N(0, \Sigma_\theta)$ where,

$$\Sigma_\theta = I_\theta^{-1}(I_\theta + C_2 I_\gamma \Sigma_\gamma I_\gamma^T) I_\theta^{-T},$$

where $I_\theta = -E\frac{\partial U}{\partial \theta}$ *and* $I_\gamma = -E\frac{\partial U}{\partial vec(\gamma)}$*. This variance can be consistently estimated by:*

$$\hat{\Sigma}_\theta = \hat{I}_\theta^{-1}\{\hat{I}_\theta + \frac{n_3}{n_2}\hat{I}_\gamma \hat{\Sigma}_\gamma \hat{I}_\gamma^T\}\hat{I}_\theta^{-T},$$

where $\hat{I}_\theta = -\frac{1}{n_3}\sum_i \frac{\partial U_i}{\partial \theta}$*,* $\hat{I}_\gamma = -\frac{1}{n_3}\sum_i \frac{\partial U_i}{\partial vec(\gamma)}$ *and* $\hat{\Sigma}_\gamma$ *is a consistent estimator of* Σ_γ*.*

Proof. Plug in $g(x) = \log\left(\frac{x}{1-x}\right)$ for linear regression form, we have:

$$I_\theta = -E\sum_i \begin{pmatrix} \mathbf{1} \\ \boldsymbol{Z}_i^* \\ \boldsymbol{V}_i \end{pmatrix} \frac{exp(\eta_i)}{(1+exp(\eta_i))^2}(\mathbf{1}, \boldsymbol{Z}_i^{*T}, \boldsymbol{V}_i{}^T),$$

$$I_\gamma = E \begin{pmatrix} \mathbf{1} \\ \boldsymbol{Z}^* \\ \boldsymbol{V}_i \end{pmatrix} \frac{exp(\eta_i)}{(1+exp(\eta_i))^2}\theta_z^T \otimes \mathbb{X},$$

$$\hat{I}_\theta \quad = \quad -n_3{}^{-1}\sum_i \begin{pmatrix} \mathbf{1} \\ \hat{\boldsymbol{Z}}_i \\ \boldsymbol{V}_i \end{pmatrix} \frac{exp(\hat{\eta}_i)}{(1+exp(\hat{\eta}_i))^2}(1, \hat{\boldsymbol{Z}}_i^T, \boldsymbol{V}_i{}^T),$$

$$\hat{I}_\gamma = n_3{}^{-1}\sum_i \begin{pmatrix} \mathbf{1} \\ \hat{\boldsymbol{Z}}_i \\ \boldsymbol{V}_i \end{pmatrix} \frac{exp(\hat{\eta}_i)}{(1+exp(\hat{\eta}_i))^2}\hat{\theta}_z^T \otimes \mathbb{X}_i.$$

Applying Lemma 3, we obtain the asymptotic for θ in linear regression setting. □

Lemma 11. *Assume* $n_2/n_1 \to C_1 < \infty$ *and* γ^* *is the unique solution from the general least square regression of* $X_1\beta^*$ *on* X_2*, i.e., solve the estimating equation below:*

$$0 = U = E\left[vec((\mathbf{1}, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T)^T \hat{X}_i - (\mathbf{1}, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T)^T (1, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T)\gamma)\right],$$

where $\hat{X}_i = (1, \boldsymbol{W}_i{}^T, \boldsymbol{V}_i{}^T)^T\beta^*$*. Also assume* $\sqrt{n_1}(vec(\hat{\beta}) - vec(\beta^*)) \to N(0, \Sigma_\beta)$*. Then we have* $\sqrt{n_2}(vec(\hat{\gamma}) - vec(\gamma^*)) \to N(0, \Sigma_\gamma)$ *where,*

$$\Sigma_\gamma = I_\gamma^{-1}(I_\gamma + C_1 I_\beta \Sigma_\beta I_\beta^T) I_\gamma^{-T},$$

where $I_\gamma = -E\frac{\partial U}{\partial vec(\gamma)}$ *and* $I_\beta = -E\frac{\partial U}{\partial vec(\beta)}$*. This variance can be consistently estimated by:*

$$\hat{\Sigma}_\gamma = \hat{I}_\gamma^{-1}\{\hat{I}_\gamma + \frac{n_2}{n_1}\hat{I}_\beta \hat{\Sigma}_\beta \hat{I}_\beta^T\}\hat{I}_\gamma^{-T},$$

where $\hat{I}_\gamma = -\frac{1}{n_2}\sum_i \frac{\partial U_i}{\partial vec(\gamma)}$*,* $\hat{I}_\beta = -\frac{1}{n_3}\sum_i \frac{\partial U_i}{\partial vec(\beta)}$ *and* $\hat{\Sigma}_\beta$ *is a consistent estimator of* Σ_β*.*

Proof. We can derive the asymptotic for $vec(\hat{\gamma})$ as below:

$$0 = U = E\left[vec((1, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T)^T \hat{\boldsymbol{X}}_i - (1, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T)^T (1, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T)\gamma)\right],$$

where $\hat{\boldsymbol{X}}_i = (1, \boldsymbol{W}_i{}^T, \boldsymbol{V}_i{}^T)^T\beta^*$. If $\hat{\gamma}$ solve the estimating equation:

$$0 = \sum U_i = n_3^{-1}\sum_{i=1}\left[vec((\mathbf{1}, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T)^T \hat{\boldsymbol{X}}_i - (1, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T)^T (\mathbf{1}, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T)\gamma)\right],$$

$$\begin{aligned}
0 &= \sum_{i=1}^{n_2} U_i(\hat{\gamma}, \hat{\beta}) \\
&= \sum_{i=1}^{n_2} U_i(\gamma^*, \beta^*) + \frac{\partial U}{\partial vec(\gamma)}|_{\gamma^*,\beta^*}(vec(\hat{\gamma}) - vec(\gamma^*)) + \\
&\frac{\partial U}{\partial vec(\beta)}|_{\gamma^*,\beta^*}(vec(\hat{\beta}) - vec(\beta^*)) + o(||vec(\hat{\gamma}) - vec(\gamma^*)||, ||vec(\hat{\beta}) - vec(\beta^*)||).
\end{aligned}$$

With Theorem 1,

$$n_2^{-1} \sum_{i=1}^{n_2} \frac{\partial U_i(\gamma, \beta)}{\partial vec(\gamma)}|_{\hat{\gamma},\hat{\beta}} = E\left(\frac{\partial U}{\partial vec(\gamma)}|_{\hat{\gamma},\hat{\beta}}\right) + o_p(1).$$

and with Theorem 2, we have:

$$E\left(\frac{\partial U}{\partial vec(\gamma)}|_{\hat{\gamma},\hat{\beta}}\right) = E\left(\frac{\partial U}{\partial vec(\gamma)}|_{\gamma^*,\beta^*}\right) + o_p(1).$$

So we have:

$$n_2^{-1} \sum_{i=1}^{n_2} \frac{\partial U_i(\gamma, \beta)}{\partial vec(\gamma)}|_{\hat{\gamma},\hat{\beta}} = -I_\gamma + o_p(1).$$

Similarly, we have:

$$n_2^{-1} \sum_{i=1}^{n_2} \frac{\partial U_i(\gamma, \beta)}{\partial vec(\beta)}|_{\hat{\gamma},\hat{\beta}} = -I_\beta + o_p(1).$$

Plug in the Taylor expansion, notice that $EU(\gamma^*, \beta^*) = 0$, we have:

$$\begin{aligned} 0 &= n_2^{-1} \sum_{i=1}^{n_2} \{U_i(vec(\gamma^*), vec(\beta^*)) - EU(vec(\gamma^*), vec(\beta^*))\} \\ & -I_\gamma(vec(\hat{\gamma}) - vec(\gamma^*)) - I_\beta(vec(\hat{\beta}) - vec(\beta^*)) \\ & +o(||vec(\hat{\gamma}) - vec(\gamma^*)||, ||vec(\hat{\beta}) - vec(\beta^*)||) + o_p(1) * ||vec(\hat{\beta}) - vec(\beta^*)|| \\ & +o_p(1) * ||vec(\hat{\gamma}) - vec(\gamma^*)||. \end{aligned}$$

So we have:

$$\sqrt{n_2}(vec(\hat{\gamma}) - vec(\gamma^*)) = I_\gamma^{-1} \left\{ n_2^{-1/2} \sum_{i=1}^{n_2} U_i(\gamma^*, \beta^*) - I_\beta \sqrt{n_2}(vec(\hat{\beta}) - vec(\beta^*)) \right\} + o_p(1).$$

By Theorem 3, we have:

$$n_2^{-1/2} \sum_{i=1}^{n_2} \{U_i(\gamma^*, \beta^*) - EU(\gamma^*, \beta^*)\} \to N(0, Var(U_i(\gamma^*, \beta^*))).$$

So we have:

$$\Sigma_\gamma = I_\gamma^{-1} \left\{ Var(U(\gamma, \beta)) + C_2 I_\beta \Sigma_\beta I_\beta^T \right\} I_\gamma^{-T}.$$

As we have shown $\hat{I}_\gamma = I_\gamma + o_p(1)$, $\hat{I}_\beta = I_\beta + o_p(1)$ and by assumption $\hat{\Sigma}_\beta = \Sigma_\beta + o_p(1)$ and $n_2/n_1 \to C_1 < \infty$, using Theorem 2, we have $\hat{\Sigma}_\gamma = \Sigma_\gamma + o_p(1)$. where $I_\gamma = -E\frac{\partial U}{\partial vec(\gamma)}$ and $I_\beta = -E\frac{\partial U}{\partial vec(\beta)}$. Specifically,

$$\begin{aligned} I_\gamma &= E\left\{ (I_m \otimes \mathbb{X}_i)^T (I_m \otimes \mathbb{X}_i) \right\}, \\ I_\beta &= E\left\{ (I_m \otimes \mathbb{X}_{\mathbb{1}i})^T (I_m \otimes \mathbb{X}_i) \right\}, \end{aligned}$$

where $\mathbb{X}_i = (1, Q_i, \boldsymbol{V}_i)$ and $\mathbb{X}_{\mathbb{1}i}$ are matrix needed in the prior step with respect to different methods. More detailed information can be found in the following lemmas. □

Lemma 12. *Assume $n_2/n_1 \to C_1 < \infty$ and γ_1^* is the unique solution to $EU(\gamma_1, \beta_1^*) = 0$ for an estimating function $U(\gamma_1, \beta_1)$. If $\hat{\gamma}_1$ solve the estimating equation $0 = n_2^{-1} \sum_{i=1}^{n_2} U_i(\gamma_1, \hat{\beta}_1)$ where $\sqrt{n_1}(vec(\hat{\beta}_1) - vec(\beta_1^*)) \to N(0, \Sigma_{\beta_1})$, then we have $\sqrt{n_2}(vec(\hat{\gamma}_1) - vec(\gamma_1^*)) \to N(0, \Sigma_{\gamma_1})$ where,*

$$0 = U_i = E\left[vec((\mathbf{1}, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T)^T \hat{X}_i - (\mathbf{1}, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T)^T (\mathbf{1}, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T)\gamma_1)\right],$$

$$\hat{X}_i = (\mathbf{1}, \boldsymbol{W}_i{}^T, \boldsymbol{V}_i{}^T)\beta^*.$$

Then we have:

$$\Sigma_{\gamma_1} = I_{\gamma_1}^{-1}(I_{\gamma_1} + C_1 I_{\beta_1} \Sigma_{\beta_1} I_{\beta_1}^T) I_{\gamma_1}^{-T},$$

where $I_{\gamma_1} = E\left\{(I_m \otimes \mathbb{X}_i)^T (I_m \otimes \mathbb{X}_i)\right\}$ and $I_{\beta_1} = E\left\{(I_m \otimes \mathbb{X}_{\mathbb{1}i})^T (I_m \otimes \mathbb{X}_i)\right\}$. This variance can be consistently estimated by:

$$\hat{\Sigma}_{\gamma_1} = \hat{I}_{\gamma_1}^{-1}\left(\hat{I}_{\gamma_1} + \frac{n_2}{n_1} \hat{I}_{\beta_1} \hat{\Sigma}_{\beta_1} \hat{I}_{\beta_1}^T\right) \hat{I}_{\gamma_1}^{-T},$$

where $\hat{I}_{\gamma_1} = n_2^{-1} \sum_{i=1}^{n_2} \left\{(I_m \otimes \mathbb{X}_i)^T (I_m \otimes \mathbb{X}_i)\right\}$, $\hat{I}_{\beta_1} = n_2^{-1} \sum_{i=1}^{n_2} \left\{(I_m \otimes \mathbb{X}_{\mathbb{1}i})^T (I_m \otimes \mathbb{X}_i)\right\}$, and $\mathbb{X}_{\mathbb{1}i} = (1, \boldsymbol{W}_i^T, \boldsymbol{V}_i^T)$ in the second sample. Moreover, $\hat{\Sigma}_{\beta_1}$ is a consistent estimator of Σ_{β_1}.

Proof. To derive asymptotic for $vec(\hat{\gamma}_1)$, we need to derive asymptotic for $vec(\hat{\beta}_1)$ and then apply Lemma 11.

$$\begin{aligned} 0 &= \sum_{i=1}^{n_1} U_i(\hat{\beta}_1) \\ &= \sum_{i=1}^{n_1} \left\{ U_i(\beta_1^*) + \frac{\partial U}{\partial vec(\beta_1)}|_{\beta_1^*}(vec(\hat{\beta}_1) - vec(\beta_1^*)) + o(||vec(\hat{\beta}_1) - vec(\beta_1^*)||) \right\}. \end{aligned}$$

where $U_i(\beta_1^*) = vec((1, \boldsymbol{W}_i{}^T, \boldsymbol{V}_i{}^T)^T \boldsymbol{X}_i^* - (1, \boldsymbol{W}_i{}^T, \boldsymbol{V}_i{}^T)^T (1, \boldsymbol{W}_i{}^T, \boldsymbol{V}_i{}^T)\beta_1^*)$.
With Theorem 1,

$$n_1^{-1} \sum_{i=1}^{n_1} \frac{\partial U_i(\beta_1)}{\partial vec(\beta_1)}|_{\hat{\beta}_1} = E\left(\frac{\partial U}{\partial vec(\beta_1)}|_{\hat{\beta}_1}\right) + o_p(1),$$

and with Theorem 2, we have:

$$E\left(\frac{\partial U}{\partial vec(\beta_1)}|_{\hat{\beta}_1}\right) = E\left(\frac{\partial U}{\partial vec(\beta_1)}|_{\beta_1^*}\right) + o_p(1).$$

So we have:

$$n_1^{-1} \sum_{i=1}^{n_1} \frac{\partial U_i(\beta_1)}{\partial vec(\beta_1)}|_{\hat{\beta}_1} = -I_{\beta_1} + o_p(1).$$

Plug in the Taylor expansion, notice that $EU(\beta_1^*) = 0$, we have:

$$\begin{aligned} 0 &= n_1^{-1} \sum_{i=1}^{n_1} \{U_i(\beta_1^*) - EU(\beta_1^*)\} - I_{\beta_1}(vec(\hat{\beta}_1) - vec(\beta_1^*)) \\ &\quad + o_p(1)||vec(\hat{\beta}_1) - vec(\beta_1^*)||. \end{aligned}$$

where $I_{\beta_1} = -E\frac{\partial U_i(\beta_1)}{\partial vec(\beta_1)}$.
So we have:

$$\sqrt{n_1}(\hat{\beta}_1 - \beta_1^*) = I_{\beta_1}^{-1} \left\{ n_1^{-1/2} \sum_{i=1}^{n_1} U_i(\beta_1^*) \right\} + o_p(1).$$

By Theorem 3, we have:

$$n_1^{-1/2} \sum_{i=1}^{n_2} \{U_i(\beta_1^*) - EU(\beta_1^*)\} \to N(0, Var(U_i(\beta_1^*))).$$

So we have :

$$\Sigma_{\beta_1} = I_{\beta_1}^{-1} \left\{ Var(U(\beta_1)) \right\} I_{\beta_1}^{-T}.$$

By assumption, $\hat{\Sigma}_{\beta_1} = \Sigma_{\beta_1} + o_p(1)$ which is a consistent estimator of Σ_{β_1}.

Then by applying Lemma 11, we have:

$$\hat{\Sigma}_{\gamma_1} = \hat{I}_{\gamma_1}^{-1} \left(\hat{I}_{\gamma_1} + \frac{n_2}{n_1} \hat{I}_{\beta_1} \hat{\Sigma}_{\beta_1} \hat{I}_{\beta_1}^{T} \right) \hat{I}_{\gamma_1}^{-T}.$$

□

Lemma 13. *Assume $n_2/n_1 \to C_1 < \infty$ and γ_2^* is the unique solution to $EU(\gamma_2, \beta_2^*) = 0$ for an estimating function $U(\gamma_2, \beta_2)$. If $\hat{\gamma}_2$ solve the estimating equation $0 = n_2^{-1} \sum_{i=1}^{n_2} U_i(\gamma_2, \hat{\beta}_2)$ where $\sqrt{n_1}(vec(\hat{\beta}_2) - vec(\beta_2^*)) \to N(0, \Sigma_{\beta_2})$, then we have $\sqrt{n_2}(vec(\hat{\gamma}_2) - vec(\gamma_2^*)) \to N(0, \Sigma_{\gamma_2})$ where,*

$$\Sigma_{\gamma_2} = I_{\gamma_2}^{-1}(I_{\gamma_2} + C_1 I_{\beta_2} \Sigma_{\beta_2} I_{\beta_2}^{T}) I_{\gamma_2}^{-T},$$

where $I_{\gamma_2} = E\left(-\frac{\partial U(\gamma_2,\beta_2)}{\partial \gamma_2}|_{\gamma_2^*,\beta_2^*}\right)$ and $I_{\beta_2} = E\left(-\frac{\partial U(\gamma_2,\beta_2)}{\partial \beta_2}|_{\gamma_2^*,\beta_2^*}\right)$. This variance can be consistently estimated by:

$$\hat{\Sigma}_{\gamma_2} = \hat{I}_{\gamma_2}^{-1} \left(\hat{I}_{\gamma_2} + \frac{n_2}{n_1} \hat{I}_{\beta_2} \hat{\Sigma}_{\beta_2} \hat{I}_{\beta_2}^{T} \right) \hat{I}_{\gamma_2}^{-T},$$

where $\hat{I}_{\gamma_2} = n_2^{-1} \sum_{i=1}^{n_2} \left(-\frac{\partial U_i(\gamma_2,\beta_2)}{\partial \gamma_2}|_{\hat{\gamma}_2,\hat{\beta}_2}\right)$, $\hat{I}_{\beta_2} = n_2^{-1} \sum_{i=1}^{n_2} \left(-\frac{\partial U_i(\gamma_2,\beta_2)}{\partial \beta_2}|_{\hat{\gamma}_2,\hat{\beta}_2}\right)$ and $\hat{\Sigma}_{\beta_2}$ is a consistent estimator of Σ_{β_2}.

Proof. The asymptotic for $\hat{\gamma}_2$ can be derived as below.

First note that,

$$Var(\boldsymbol{X}^*|\boldsymbol{V},\boldsymbol{W}) = \Omega_1 = \frac{1}{n_1}\sum_i \left\{ (\boldsymbol{X}_i^* - (\mathbf{1}, \boldsymbol{W}_i{}^T, \boldsymbol{V}_i{}^T)\beta)^T (\boldsymbol{X}_i^* - (\mathbf{1}, \boldsymbol{W}_i{}^T, \boldsymbol{V}_i{}^T)\beta) \right\},$$

$$Var(\boldsymbol{X}^*|\boldsymbol{V}) = \Omega_2 = \frac{1}{n_1}\sum_i \left\{ (\boldsymbol{X}_i^* - (1, \boldsymbol{V}_i{}^T)\beta_t)^T (\boldsymbol{X}_i^* - (\mathbf{1}, \boldsymbol{V}_i{}^T)\beta_t) \right\},$$

where we let $\Omega_{1i} = \left\{ \boldsymbol{X}_i^* - (\mathbf{1}, \boldsymbol{W}_i{}^T, \boldsymbol{V}_i{}^T)\beta \right\}^T \left\{ \boldsymbol{X}_i^* - (\mathbf{1}, \boldsymbol{W}_i{}^T, \boldsymbol{V}_i{}^T)\beta \right\}$ and
$\Omega_{2i} = \left(\boldsymbol{X}_i^* - (\mathbf{1}, \boldsymbol{V}_i{}^T)\beta_t \right)^T \left(\boldsymbol{X}_i^* - (\mathbf{1}, \boldsymbol{V}_i{}^T)\beta_t \right)$.

Second, the estimating equations considered are:

$$U_{121i} = vec \left\{ (\mathbf{1}, \boldsymbol{W}_i{}^T, \boldsymbol{V}_i{}^T)^T (\boldsymbol{X}_i^* - (1, \boldsymbol{W}_i{}^T, \boldsymbol{V}_i{}^T)\beta) \right\},$$

$$U_{122i} = vec \left\{ (\mathbf{1}, \boldsymbol{V}_i{}^T)^T (\boldsymbol{X}_i^* - (\mathbf{1}, \boldsymbol{V}_i{}^T)\beta_t) \right\}.$$

$$U_{123i} = vec \left\{ (\boldsymbol{X}_i^* - (\mathbf{1}, \boldsymbol{W}_i{}^T, \boldsymbol{V}_i{}^T)\beta)^T (\mathbf{1}, \boldsymbol{W}_i{}^T, \boldsymbol{V}_i{}^T)\beta) - \Omega_1 \right\},$$

$$U_{124i} = vec \left\{ (\boldsymbol{X}_i^* - (\mathbf{1}, \boldsymbol{V}_i{}^T)\beta_t)^T (\boldsymbol{X}_i^* - (\mathbf{1}, \boldsymbol{V}_i{}^T)\beta_t) - \Omega_2 \right\}.$$

Third, we can derive the asymptotic normal distribution for $vec(\beta)$, $vec(\beta_t)$, $vec(\Omega_1)$ and $vec(\Omega_2)$ as:

$$\sqrt{n_1}\left[\begin{pmatrix} vec(\hat{\beta}) \\ vect(\hat{\beta}_t) \\ vec(\widehat{\Omega}_1) \\ vec(\widehat{\Omega}_2) \end{pmatrix} - \begin{pmatrix} vec(\beta) \\ vec(\beta_t) \\ vec(\Omega_1) \\ vec(\Omega_2) \end{pmatrix} \right] \rightarrow N(0, I^{-1}JI^{-T}),$$

where J is the variance covariance matrix of the above four estimating equations and I is a matrix composed by the expectation of derivatives of each estimating equation with respect to vec(β), vec(β_t), vec(Ω_1) and vec(Ω_2), respectively. Specifically,

$$I = \begin{pmatrix} \frac{1}{n_1}\sum_i (\mathbb{X}_i)^T(\mathbb{X}_i) & 0 & 0 & 0 \\ 0 & \frac{1}{n_1}\sum_i (\mathbb{X}_{ti})^T(\mathbb{X}_{ti}) & 0 & 0 \\ 0 & 0 & I_{m^2} & 0 \\ 0 & 0 & 0 & I_{m^2} \end{pmatrix},$$

where $\mathbb{X}_i = (1, \boldsymbol{W}_i^T, \boldsymbol{V}_i^T)$ and $\mathbb{X}_{ti} = (1, \boldsymbol{V}_i^T)$.

Fourth, $\hat{\beta}_2 = \hat{\beta}\widehat{BF}^{-1}$ can be derived using delta method.

$$\sqrt{n_1}\Bigg(\; vec(\hat{\beta}_2) \; - \; vec(\beta_2) \; \Bigg) \rightarrow N(0, CI^{-1}JI^{-T}C^T),$$

where C is a matrix derived by taking derivative of β_2 each with respect to β and BF, respectively. That is:

$$C_1 = \begin{pmatrix} I_{mp} & 0 & 0 \\ 0 & 0 & {\Sigma_2}^{-1} \otimes {\Sigma_2}^{-1} \\ 0 & I_{m^2} & 0 \end{pmatrix},$$

$$C_2 = \begin{pmatrix} I_{mp} & 0 & 0 \\ 0 & I_m \otimes {\Sigma_2}^{-1} & \Sigma_1 \otimes I_m \end{pmatrix},$$

$$C_3 = \begin{pmatrix} I_{mp} & 0 \\ 0 & I_{m^2} \end{pmatrix},$$

$$\begin{aligned} C_4 &= \begin{pmatrix} I_{mp} & 0 \\ 0 & (I - {\Sigma_2}^{-1}\Sigma_1)^{-1} \otimes (I - {\Sigma_2}^{-1}\Sigma_1)^{-1} \end{pmatrix}, \\ C_5 &= \begin{pmatrix} I_{mp} \otimes (I - {\Sigma_2}^{-1}\Sigma_1)^{-1} & \hat{\beta} \otimes I_m \end{pmatrix}, \end{aligned}$$

and $C = C_5 C_4 C_3 C_2 C_1 C_0$. Then we have estimating equation for β_2 as below:

$$\begin{aligned} 0 &= \sum_{i=1}^{n_1} U_i(\hat{\beta}_2) = \sum_{i=1}^{n_1} \left\{ (1, \boldsymbol{W}_i{}^T, \boldsymbol{V}_i{}^T)^T \boldsymbol{X}_i^* - (1, \boldsymbol{W}_i{}^T, \boldsymbol{V}_i{}^T)^T (1, \boldsymbol{W}_i{}^T, \boldsymbol{V}_i{}^T) vec(\hat{\beta}_2) \right\} \\ &= \sum_{i=1}^{n_1} \left\{ U_i(\beta_2^*) + \frac{\partial U}{\partial vec(\beta_2)}|_{\beta_2^*} (vec(\hat{\beta}_2) - vec(\beta_2^*)) + o(||vec(\hat{\beta}_2) - vec(\beta_2^*)||) \right\}. \end{aligned}$$

With Theorem 1,

$$n_1^{-1} \sum_{i=1}^{n_1} \frac{\partial U_i(\beta_2)}{\partial vec(\beta_2)}|_{\hat{\beta}_2} = E\left(\frac{\partial U}{\partial vec(\beta_2)}|_{\hat{\beta}_2} \right) + o_p(1),$$

and with Theorem 2, we have:

$$E\left(\frac{\partial U}{\partial vec(\beta_2)}|_{\hat{\beta}_2} \right) = E\left(\frac{\partial U}{\partial vec(\beta_2)}|_{\beta_2^*} \right) + o_p(1).$$

So we have:

$$n_1^{-1} \sum_{i=1}^{n_1} \frac{\partial U_i(\beta_2)}{\partial vec(\beta_2)}|_{\hat{\beta}_2} = -I_{\beta_2} + o_p(1).$$

Plug in the Taylor expansion, notice that $EU(\beta_2^*) = 0$, we have:

$$\begin{aligned} 0 &= n_1^{-1} \sum_{i=1}^{n_1} \{ U_i(\beta_2^*) - EU(\beta_2^*) \} - I_{\beta_1} (vec(\hat{\beta}_2) - vec(\beta_2^*)) \\ & \quad + o_p(1) ||vec(\hat{\beta}_2) - vec(\beta_2^*)||. \end{aligned}$$

So we have:

$$\sqrt{n_1}(vec(\hat{\beta}_2) - vec(\beta_2^*)) = I_{\beta_2}^{-1} \left\{ n_1^{-1/2} \sum_{i=1}^{n_1} U_i(\beta_2^*) \right\} + o_p(1).$$

By Theorem 3, we have:

$$n_1^{-1/2} \sum_{i=1}^{n_2} \{U_i(\beta_2^*) - EU(\beta_2^*)\} \to N(0, Var(U_i(\beta_2^*))).$$

So we have:

$$\Sigma_{\beta_2} = I_{\beta_2}^{-1} \{Var(U(\beta_2))\} I_{\beta_2}^{-T}.$$

By assumption, $\hat{\Sigma}_{\beta_2} = \Sigma_{\beta_2} + o_p(1)$ which is a consistent estimator of Σ_{β_2}.

Then by applying Lemma 4, we have:

$$\hat{\Sigma}_{\gamma_2} = \hat{I}_{\gamma_2}^{-1} \left(\hat{I}_{\gamma_2} + \frac{n_2}{n_1} \hat{I}_{\beta_2} \hat{\Sigma}_{\beta_2} \hat{I}_{\beta_2}^{T} \right) \hat{I}_{\gamma_2}^{-T}.$$

□

Lemma 14. *Assume $n_2/n_1 \to C_1 < \infty$ and γ_3^* is the unique solution to: $EU(\gamma_3, \beta_3^*) = 0$ for an estimating function $U(\gamma_3, \beta_3)$. If $\hat{\gamma}_1$ solve the estimating equation: $0 = n_2^{-1} \sum_{i=1}^{n_2} U_i(\gamma_3, \hat{\beta}_3)$ where $\sqrt{n_1}(vec(\hat{\beta}_3) - vec(\beta_3^*)) \to N(0, \Sigma_{\beta_3})$, then we have $\sqrt{n_2}(vec(\hat{\gamma}_3) - vec(\gamma_3^*)) \to N(0, \Sigma_{\gamma_3})$ where,*

$$0 = U_i = E\left[vec((\mathbf{1}, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T)^T \hat{\boldsymbol{X}}_i - (1, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T)^T (1, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T) \gamma_3) \right],$$

$$\hat{X}_i = (\mathbf{1}, \boldsymbol{W}_i{}^T, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T) \beta^*.$$

Then we have:

$$\Sigma_{\gamma_3} = I_{\gamma_3}^{-1}(I_{\gamma_3} + C_1 I_{\beta_3}\Sigma_{\beta_1}I_{\beta_3}^T)I_{\gamma_3}^{-T},$$

where $I_{\gamma_3} = E\left(-\frac{\partial U(\gamma_3,\beta_3)}{\partial vec(\gamma_3)}|_{\gamma_3^*,\beta_3^*}\right)$ *and* $I_{\beta_3} = E\left(-\frac{\partial U(\gamma_1,\beta_3)}{\partial vec(\beta_3)}|_{\gamma_3^*,\beta_3^*}\right)$. *This variance can be consistently estimated by:*

$$\hat{\Sigma}_{\gamma_3} = \hat{I}_{\gamma_3}^{-1}\left(\hat{I}_{\gamma_3} + \frac{n_2}{n_1}\hat{I}_{\beta_3}\hat{\Sigma}_{\beta_3}\hat{I}_{\beta_3}^T\right)\hat{I}_{\gamma_3}^{-T},$$

where $\hat{I}_{\gamma_3} = n_2^{-1}\sum_{i=1}^{n_2}\left(-\frac{\partial U_i(\gamma_3,\beta_3)}{\partial vec(\gamma_3)}|_{\hat{\gamma}_3,\hat{\beta}_3}\right)$, $\hat{I}_{\beta_3} = n_2^{-1}\sum_{i=1}^{n_2}\left(-\frac{\partial U_i(\gamma_3,\beta_3)}{\partial vec(\beta_3)}|_{\hat{\gamma}_3,\hat{\beta}_3}\right)$ *and* $\hat{\Sigma}_{\beta_3}$ *is a consistent estimator of* Σ_{β_3}.

Proof. To prove Lemma 14, we need to derive asymptotic for $vec(\hat{\beta}_3)$ and then apply Lemma 11. The asymptotic of $vec(\hat{\beta}_3)$ can be derived as below:

$$\begin{aligned} 0 &= \sum_{i=1}^{n_1} U_i(\hat{\beta}_3) \\ &= \sum_{i=1}^{n_1}\left\{U_i(\beta_3^*) + \frac{\partial U}{\partial vec(\beta_3)}|_{\beta_3^*}(vec(\hat{\beta}_3) - vec(\beta_3^*)) + o(||vec(\hat{\beta}_3) - vec(\beta_3^*)||)\right\}. \end{aligned}$$

where $U_i(\beta_3^*) = vec((1, \boldsymbol{W}_i{}^T, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T)^T\boldsymbol{X}_i^* - (1, \boldsymbol{W}_i{}^T, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T)^T(1, \boldsymbol{W}_i{}^T, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T))\beta_1^*)$. With Theorem 1,

$$n_1^{-1}\sum_{i=1}^{n_1}\frac{\partial U_i(\beta_3)}{\partial vec(\beta_3)}|_{\hat{\beta}_3} = E\left(\frac{\partial U}{\partial vec(\beta_3)}|_{\hat{\beta}_3}\right) + o_p(1),$$

and with Theorem 2, we have:

$$E\left(\frac{\partial U}{\partial vec(\beta_3)}|_{\hat{\beta}_3}\right) = E\left(\frac{\partial U}{\partial vec(\beta_3)}|_{\beta_3^*}\right) + o_p(1).$$

So we have:

$$n_1^{-1}\sum_{i=1}^{n_1}\frac{\partial U_i(\beta_3)}{\partial vec(\beta_3)}|_{\hat{\beta}_3} = -I_{\beta_3} + o_p(1).$$

Plug in the Taylor expansion, notice that $EU(\beta_3^*) = 0$, we have:

$$\begin{aligned} 0 &= n_1^{-1}\sum_{i=1}^{n_1}\{U_i(\beta_3^*) - EU(\beta_3^*)\} - I_{\beta_3}(vec(\hat{\beta}_3) - vec(\beta_3^*)) \\ &\quad + o_p(1)||vec(\hat{\beta}_3) - vec(\beta_3^*)||. \end{aligned}$$

where $I_{\beta_3} = -E\frac{\partial U_i(\beta_3)}{\partial vec(\beta_3)}$.

So we have:

$$\sqrt{n_1}(\hat{\beta}_3 - \beta_3^*) = I_{\beta_3}^{-1}\left\{n_1^{-1/2}\sum_{i=1}^{n_1}U_i(\beta_3^*)\right\} + o_p(1).$$

By Theorem 3, we have:

$$n_1^{-1/2}\sum_{i=1}^{n_2}\{U_i(\beta_3^*) - EU(\beta_3^*)\} \to N(0, Var(U_i(\beta_3^*))).$$

So we have:

$$\Sigma_{\beta_3} = I_{\beta_3}^{-1}\{Var(U(\beta_3))\}\, I_{\beta_3}^{-T}.$$

By assumption, $\hat{\Sigma}_{\beta_3} = \Sigma_{\beta_3} + o_p(1)$ which is a consistent estimator of Σ_{β_3}.

Then by applying Lemma 11, we have:

$$\hat{\Sigma}_{\gamma_3} = \hat{I}_{\gamma_3}^{-1}\left(\hat{I}_{\gamma_3} + \frac{n_2}{n_1}\hat{I}_{\beta_3}\hat{\Sigma}_{\beta_3}\hat{I}_{\beta_1}^{T}\right)\hat{I}_{\gamma_3}^{-T}.$$

□

Lemma 15. *Assume* γ_4^* *is the unique solution to:* $EU(\gamma_4) = 0$ *for an estimating function* $U(\gamma_4)$. *Solving the estimating equation* $0 = n_1^{-1}\sum_{i=1}^{n_1} U_i(\hat{\gamma}_4)$, *we have* $\sqrt{n_1}(vec(\hat{\gamma}_4) - vec(\gamma_4^*)) \to N(0, \Sigma_{\gamma_4})$ *where,*

$$U_i(\gamma_4^*) = vec((1, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T)^T \boldsymbol{X}_i^* - (1, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T)^T (1, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T)\gamma_4^*).$$

and,

$$\Sigma_{\gamma_4} = I_{\gamma_4}^{-1}(Var(U(\gamma_4)))I_{\gamma_4}^{-T},$$

where $I_{\gamma_4} = E\left(-\frac{\partial U(\gamma_4)}{\partial vec(\gamma_4)}|_{\gamma_4^*}\right)$. *This variance can be consistently estimated by:*

$$\hat{\Sigma}_{\gamma_4} = \hat{I}_{\gamma_4}^{-1}(Var(U(\gamma_4)))\hat{I}_{\gamma_4}^{-T},$$

where $\hat{I}_{\gamma_4} = n_1^{-1}\sum_{i=1}^{n_1}\left(-\frac{\partial U_i(\gamma_4)}{\partial vec(\gamma_4)}|_{\hat{\gamma}_4}\right)$ *and* $\hat{\Sigma}_{\gamma_4}$ *is a consistent estimator of* Σ_{γ_4}.

Proof. The asymptotic for $\hat{\gamma}_4$ can be derived as below.

$$\begin{aligned} 0 &= \sum_{i=1}^{n_1} U_i(\hat{\gamma}_4) = \sum_{i=1}^{n_1}\left\{vec((1, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T)^T \boldsymbol{X}_i^* - (1, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T)^T (1, \boldsymbol{Q}_i{}^T, \boldsymbol{V}_i{}^T)\hat{\gamma}_4)\right\} \\ &= \sum_{i=1}^{n_1}\left\{U_i(\gamma_4^*) + \frac{\partial U}{\partial vec(\gamma_4)}|_{\gamma_4^*}(vec(\hat{\gamma}_4) - vec(\gamma_4^*)) + o(||vec(\hat{\gamma}_4) - vec(\gamma_4^*)||)\right\}. \end{aligned}$$

With Theorem 1,

$$n_1^{-1}\sum_{i=1}^{n_1}\frac{\partial U_i(\gamma_4)}{\partial vec(\gamma_4)}|_{\hat{\gamma}_4} = E\left(\frac{\partial U}{\partial vec(\gamma_4)}|_{\hat{\gamma}_4}\right) + o_p(1),$$

and with Theorem 2, we have:

$$E\left(\frac{\partial U}{\partial vec(\gamma_4)}|_{\hat{\gamma}_4}\right) = E\left(\frac{\partial U}{\partial vec(\gamma_4)}|_{\gamma_4^*}\right) + o_p(1).$$

So we have:

$$n_1^{-1}\sum_{i=1}^{n_1}\frac{\partial U_i(\gamma_4)}{\partial vec(\gamma_4)}|_{\hat{\gamma}_4} = -I_{\gamma_4} + o_p(1).$$

Plug in the Taylor expansion, notice that $EU(\gamma_4^*) = 0$, we have:

$$\begin{aligned} 0 &= n_1^{-1}\sum_{i=1}^{n_1}\{U_i(\gamma_4^*) - EU(\gamma_4^*)\} - I_{\gamma_4}(vec(\hat{\gamma}_4) - vec(\gamma_4^*)) \\ &\quad + o_p(1)||vec(\hat{\gamma}_4) - vec(\gamma_4^*)||. \end{aligned}$$

So we have:

$$\sqrt{n_1}(vec(\hat{\gamma}_4) - vec(\gamma_4^*)) = I_{\gamma_4}^{-1}\left\{n_1^{-1/2}\sum_{i=1}^{n_1}U_i(\gamma_4^*)\right\} + o_p(1).$$

By Theorem 3, we have:

$$n_1^{-1/2}\sum_{i=1}^{n_2}\{U_i(\gamma_4^*) - EU(\gamma_4^*)\} \to N(0, Var(U_i(\gamma_4^*))).$$

So we have:

$$\Sigma_{\gamma_4} = I_{\gamma_4}^{-1}\{Var(U(\gamma_4))\}I_{\gamma_4}^{-T}.$$

By assumption, $\hat{\Sigma}_{\gamma_4} = \Sigma_{\gamma_4} + o_p(1)$ which is a consistent estimator of Σ_{γ_4}. □

Theorem 7. *With $X \in R^K$ and $Q \in R^K$, the asymptotic bias in Method 1 with K exposures is associated with ρ and δ with a form such that the bias corrected association parameters, θ_z^* and θ_v^*, are $\theta_z^* = \rho^{-1}\theta_z$ where $\rho = I_K - Var(\boldsymbol{X}|\boldsymbol{V})^{-1}Var(\boldsymbol{X}|\boldsymbol{W},\boldsymbol{V})$, and $\theta_v^* = \theta_v - \rho^{-1}(1-\rho)\delta\theta_z$ when a linear function form, $E(X|\boldsymbol{V}) = \boldsymbol{V}\delta$, exists.*

Proof. To see the asymptotic bias of Method 1, we consider the first step regression model of $\boldsymbol{X}^*$ on $(\boldsymbol{W}, \boldsymbol{V})$, then we have:

$$\begin{aligned}\hat{X} &= E(\boldsymbol{X}^*|\boldsymbol{W}, \boldsymbol{V}) = E(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{V}) \\ &= E(\boldsymbol{X}|\boldsymbol{V}) + \{\boldsymbol{W} - E(\boldsymbol{W}|\boldsymbol{V})\}^T \Sigma^{-1}_{\boldsymbol{W}\boldsymbol{W}|\boldsymbol{V}} \Sigma^T_{\boldsymbol{X}\boldsymbol{W}|\boldsymbol{V}}.\end{aligned}$$

Now we compute $E(\hat{X}|\boldsymbol{Q}, \boldsymbol{V})$ to see how it is biased away from $E(\boldsymbol{X}|\boldsymbol{Q}, \boldsymbol{V})$. We have:

$$\begin{aligned}E(\hat{X}|\boldsymbol{Q}, \boldsymbol{V}) &= E\left\{E\left(\hat{X}|\boldsymbol{X}, \boldsymbol{Q}, \boldsymbol{V}\right)|\boldsymbol{Q}, \boldsymbol{V}\right\} = E\left\{E\left(\hat{X}|\boldsymbol{X}, \boldsymbol{V}\right)|\boldsymbol{Q}, \boldsymbol{V}\right\} \\ &= E\left[E\left[E(\boldsymbol{X}|\boldsymbol{V}) + \{\boldsymbol{W} - E(\boldsymbol{W}|\boldsymbol{V})\}^T \Sigma^{-1}_{\boldsymbol{W}\boldsymbol{W}|\boldsymbol{V}} \Sigma^T_{\boldsymbol{X}\boldsymbol{W}|\boldsymbol{V}}|\boldsymbol{X}, \boldsymbol{V}\right]|\boldsymbol{Q}, \boldsymbol{V}\right] \\ &= E\left[E(\boldsymbol{X}|\boldsymbol{V}) + \{E(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{V}) - E(\boldsymbol{W}|\boldsymbol{V})\}^T \Sigma^{-1}_{\boldsymbol{W}\boldsymbol{W}|\boldsymbol{V}} \Sigma^T_{\boldsymbol{X}\boldsymbol{W}|\boldsymbol{V}}|\boldsymbol{Q}, \boldsymbol{V}\right] \\ &= E\left[E(\boldsymbol{X}|\boldsymbol{V}) + \{\boldsymbol{X} - E(\boldsymbol{X}|\boldsymbol{V})\} \Sigma^{-1}_{\boldsymbol{X}\boldsymbol{X}|\boldsymbol{V}} \Sigma_{\boldsymbol{X}\boldsymbol{W}|\boldsymbol{V}} \Sigma^{-1}_{\boldsymbol{W}\boldsymbol{W}|\boldsymbol{V}} \Sigma^T_{\boldsymbol{X}\boldsymbol{W}|\boldsymbol{V}}|\boldsymbol{Q}, \boldsymbol{V}\right] \\ &= E\left[E(\boldsymbol{X}|\boldsymbol{V}) + \{\boldsymbol{X} - E(\boldsymbol{X}|\boldsymbol{V})\} \rho_{\boldsymbol{V}}|\boldsymbol{Q}, \boldsymbol{V}\right] \\ &= \rho_{\boldsymbol{V}} E(\boldsymbol{X}|\boldsymbol{Q}, \boldsymbol{V}) + (I_K - \rho_{\boldsymbol{V}}) E(\boldsymbol{X}|\boldsymbol{V}).\end{aligned}$$

When $\rho_{\boldsymbol{V}}$ is a constant over $\boldsymbol{V}$, we simply denote it as ρ and we have $\rho\theta_z^* = \theta_z$, or $\theta_z^* = \rho^{-1}\theta_z$ with appropriate adjustment for V. Explicitly, we have:

$$\begin{aligned}\rho &= I_K - \Sigma^{-1}_{\boldsymbol{X}\boldsymbol{X}|\boldsymbol{V}}\left(\Sigma_{\boldsymbol{X}\boldsymbol{X}|\boldsymbol{V}} - \Sigma_{\boldsymbol{X}\boldsymbol{W}|\boldsymbol{V}} \Sigma^{-1}_{\boldsymbol{W}\boldsymbol{W}|\boldsymbol{V}} \Sigma^T_{\boldsymbol{X}\boldsymbol{W}|\boldsymbol{V}}\right) \\ &= I_K - Var(\boldsymbol{X}|\boldsymbol{V})^{-1} Var(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{V}).\end{aligned}$$

If we further have $E(\boldsymbol{X}|\boldsymbol{V}) = \boldsymbol{V}\delta$ is a linear function of V, then $(I_K - \rho)\delta\theta_z^* + \theta_v^* = \theta_v$, or $\theta_v^* = \theta_v - \rho^{-1}(I_K - \rho)\delta\theta_z$. □

Theorem 8. *With* $\frac{n_3}{n_2} \to C_2$ *and* $\frac{n_2}{n_1} \to C_1$*, we have* $\sqrt{n_3}(\hat{\theta}_1 - \theta_1^*) \to N(0, \Sigma_{\theta_1})$ *where*

$\Sigma_{\theta_1} = I_{\theta_1}^{-1}(I_{\theta_1} + C_2 I_{\gamma_1}\Sigma_{\gamma_1}I_{\gamma_1}^T)I_{\theta_1}^{-T}$ can be consistently estimated by $\hat{\Sigma}_{\theta_1} = \hat{I}_{\theta_1}^{-1}(\hat{I}_{\theta_1} + \frac{n_3}{n_2}\hat{I}_{\gamma_1}\hat{\Sigma}_{\gamma_1}\hat{I}_{\gamma_1}^T)\hat{I}_{\theta_1}^{-T}$ for Method 1.

With $\frac{n_3}{n_2} \to C_2$ and $\frac{n_2}{n_1} \to C_1$, we have $\sqrt{n_3}(\hat{\theta}_2 - \theta_2^*) \to N(0, \Sigma_{\theta_2})$ where $\Sigma_{\theta_2} = I_{\theta_2}^{-1}(I_{\theta_2} + C_2 I_{\gamma_2}\Sigma_{\gamma_2}I_{\gamma_2}^T)I_{\theta_2}^{-T}$ can be consistently estimated by $\hat{\Sigma}_{\theta_2} = \hat{I}_{\theta_2}^{-1}(\hat{I}_{\theta_2} + \frac{n_3}{n_2}\hat{I}_{\gamma_2}\hat{\Sigma}_{\gamma_2}\hat{I}_{\gamma_2}^T)\hat{I}_{\theta_2}^{-T}$ for Method 2.

With $\frac{n_3}{n_2} \to C_2$ and $\frac{n_2}{n_1} \to C_1$, we have $\sqrt{n_3}(\hat{\theta}_3 - \theta_3^*) \to N(0, \Sigma_{\theta_3})$ where $\Sigma_{\theta_3} = I_{\theta_3}^{-1}(I_{\theta_3} + C_2 I_{\gamma_3}\Sigma_{\gamma_3}I_{\gamma_3}^T)I_{\theta_3}^{-T}$ can be consistently estimated by $\hat{\Sigma}_{\theta_3} = \hat{I}_{\theta_3}^{-1}(\hat{I}_{\theta_3} + \frac{n_3}{n_2}\hat{I}_{\gamma_3}\hat{\Sigma}_{\gamma_3}\hat{I}_{\gamma_3}^T)\hat{I}_{\theta_3}^{-T}$ for Method 3.

With $\frac{n_3}{n_2} \to C_2$ and $\frac{n_2}{n_1} \to C_1$, we have $\sqrt{n_3}(\hat{\theta}_4 - \theta_4^*) \to N(0, \Sigma_{\theta_4})$ where $\Sigma_{\theta_4} = I_{\theta_4}^{-1}(I_{\theta_4} + C_2 I_{\gamma_4}\Sigma_{\gamma_4}I_{\gamma_4}^T)I_{\theta_4}^{-T}$ can be consistently estimated by $\hat{\Sigma}_{\theta_4} = \hat{I}_{\theta_4}^{-1}(\hat{I}_{\theta_4} + \frac{n_3}{n_2}\hat{I}_{\gamma_4}\hat{\Sigma}_{\gamma_4}\hat{I}_{\gamma_4}^T)\hat{I}_{\theta_4}^{-T}$ for Method 4.

Proof. By applying Lemma 2 and Lemma 5, the asymptotic Σ_{θ_1} can be derived as
$\hat{\Sigma}_{\theta_1} = \hat{I}_{\theta_1}^{-1}(\hat{I}_{\theta_1} + \frac{n_3}{n_2}\hat{I}_{\gamma_1}\hat{\Sigma}_{\gamma_1}\hat{I}_{\gamma_1}^T)\hat{I}_{\theta_1}^{-T}$.
By applying Lemma 2 and Lemma 6, the asymptotic Σ_{θ_2} can be derived as
$\hat{\Sigma}_{\theta_2} = \hat{I}_{\theta_2}^{-1}(\hat{I}_{\theta_2} + \frac{n_3}{n_2}\hat{I}_{\gamma_2}\hat{\Sigma}_{\gamma_2}\hat{I}_{\gamma_2}^T)\hat{I}_{\theta_2}^{-T}$.
By applying Lemma 2 and Lemma 7, the asymptotic Σ_{θ_3} can be derived as
$\hat{\Sigma}_{\theta_3} = \hat{I}_{\theta_3}^{-1}(\hat{I}_{\theta_3} + \frac{n_3}{n_2}\hat{I}_{\gamma_3}\hat{\Sigma}_{\gamma_3}\hat{I}_{\gamma_3}^T)\hat{I}_{\theta_3}^{-T}$.
By applying Lemma 2 and Lemma 8, the asymptotic Σ_{θ_4} can be derived as
$\hat{\Sigma}_{\theta_4} = \hat{I}_{\theta_4}^{-1}(\hat{I}_{\theta_4} + \frac{n_3}{n_2}\hat{I}_{\gamma_4}\hat{\Sigma}_{\gamma_4}\hat{I}_{\gamma_4}^T)\hat{I}_{\theta_4}^{-T}$. □

Theorem 9. *With $\frac{n_3}{n_2} \to C_2$ and $\frac{n_2}{n_1} \to C_1$, we have $\sqrt{n_3}(\hat{\theta}_1 - \theta_1^*) \to N(0, \Sigma_{\theta_1})$ where $\Sigma_{\theta_1} = I_{\theta_1}^{-1}(I_{\theta_1} + C_2 I_{\gamma_1}\Sigma_{\gamma_1}I_{\gamma_1}^T)I_{\theta_1}^{-T}$ can be consistently estimated by $\hat{\Sigma}_{\theta_1} = \hat{I}_{\theta_1}^{-1}(\hat{I}_{\theta_1} + \frac{n_3}{n_2}\hat{I}_{\gamma_1}\hat{\Sigma}_{\gamma_1}\hat{I}_{\gamma_1}^T)\hat{I}_{\theta_1}^{-T}$ for Method 1.*

With $\frac{n_3}{n_2} \to C_2$ and $\frac{n_2}{n_1} \to C_1$, we have $\sqrt{n_3}(\hat{\theta}_2 - \theta_2^*) \to N(0, \Sigma_{\theta_2})$ where $\Sigma_{\theta_2} = I_{\theta_2}^{-1}(I_{\theta_2} + C_2 I_{\gamma_2}\Sigma_{\gamma_2}I_{\gamma_2}^T)I_{\theta_2}^{-T}$ can be consistently estimated by $\hat{\Sigma}_{\theta_2} = \hat{I}_{\theta_2}^{-1}(\hat{I}_{\theta_2} + \frac{n_3}{n_2}\hat{I}_{\gamma_2}\hat{\Sigma}_{\gamma_2}\hat{I}_{\gamma_2}^T)\hat{I}_{\theta_2}^{-T}$ for Method 2.

With $\frac{n_3}{n_2} \to C_2$ and $\frac{n_2}{n_1} \to C_1$, we have $\sqrt{n_3}(\hat{\theta}_3 - \theta_3^*) \to N(0, \Sigma_{\theta_3})$ where $\Sigma_{\theta_3} = I_{\theta_3}^{-1}(I_{\theta_3} +$

$C_2 I_{\gamma_3} \Sigma_{\gamma_3} I_{\gamma_3}^T) I_{\theta_3}^{-T}$ can be consistently estimated by $\hat{\Sigma}_{\theta_3} = \hat{I}_{\theta_3}^{-1}(\hat{I}_{\theta_3} + \frac{n_3}{n_2} \hat{I}_{\gamma_3} \hat{\Sigma}_{\gamma_3} \hat{I}_{\gamma_3}^T) \hat{I}_{\theta_3}^{-T}$ for Method 3.

With $\frac{n_3}{n_2} \rightarrow C_2$ and $\frac{n_2}{n_1} \rightarrow C_1$, we have $\sqrt{n_3}(\hat{\theta}_4 - \theta_4^*) \rightarrow N(0, \Sigma_{\theta_4})$ where $\Sigma_{\theta_4} = I_{\theta_4}^{-1}(I_{\theta_4} + C_2 I_{\gamma_4} \Sigma_{\gamma_4} I_{\gamma_4}^T) I_{\theta_4}^{-T}$ can be consistently estimated by $\hat{\Sigma}_{\theta_4} = \hat{I}_{\theta_4}^{-1}(\hat{I}_{\theta_4} + \frac{n_3}{n_2} \hat{I}_{\gamma_4} \hat{\Sigma}_{\gamma_4} \hat{I}_{\gamma_4}^T) \hat{I}_{\theta_4}^{-T}$ for Method 4.

Proof. By applying Lemma 3 and Lemma 5, the asymptotic Σ_{θ_1} can be derived as
$\hat{\Sigma}_{\theta_1} = \hat{I}_{\theta_1}^{-1}(\hat{I}_{\theta_1} + \frac{n_3}{n_2} \hat{I}_{\gamma_1} \hat{\Sigma}_{\gamma_1} \hat{I}_{\gamma_1}^T) \hat{I}_{\theta_1}^{-T}$.
By applying Lemma 3 and Lemma 6, the asymptotic Σ_{θ_2} can be derived as
$\hat{\Sigma}_{\theta_2} = \hat{I}_{\theta_2}^{-1}(\hat{I}_{\theta_2} + \frac{n_3}{n_2} \hat{I}_{\gamma_2} \hat{\Sigma}_{\gamma_2} \hat{I}_{\gamma_2}^T) \hat{I}_{\theta_2}^{-T}$.
By applying Lemma 3 and Lemma 7, the asymptotic Σ_{θ_3} can be derived as
$\hat{\Sigma}_{\theta_3} = \hat{I}_{\theta_3}^{-1}(\hat{I}_{\theta_3} + \frac{n_3}{n_2} \hat{I}_{\gamma_3} \hat{\Sigma}_{\gamma_3} \hat{I}_{\gamma_3}^T) \hat{I}_{\theta_3}^{-T}$.
By applying Lemma 3 and Lemma 8, the asymptotic Σ_{θ_4} can be derived as
$\hat{\Sigma}_{\theta_4} = \hat{I}_{\theta_4}^{-1}(\hat{I}_{\theta_4} + \frac{n_3}{n_2} \hat{I}_{\gamma_4} \hat{\Sigma}_{\gamma_4} \hat{I}_{\gamma_4}^T) \hat{I}_{\theta_4}^{-T}$. □

CURRICULUM VITAE

Yiwen Zhang

Education

M.S. in Biostatistcs Aug 2012-Sep 2014

Unversity of Minnesota-Twin Cities

GPA: 3.72/4.00

B.S. in Applied Mathematics Aug 2009-May 2012

State University of New York at Fredonia, Fredonia, NY

GPA: 3.84/4.00

Dissertation Title

Biomarker Development For Use In Regression Calibration.

Work Experience

Grad Intern – R & D (Biostatistics), Amgen Inc., Thousand Oaks, CA

Jun 2019 – Aug 2019

Biostatistics Intern, Agios Pharmaceutical, Cambridge, MA

Jun 2017 – Aug 2017

Research Programmer, Novartis Pharmaceutical, Shanghai, China

Jul 2015 – Aug 2016

Research Analyst, Minnesota Department of Health (MDH), Saint Paul, MN

Aug 2014 - May 2015

Research Experience

Research Assistant for Prof. Cheng Zheng, University of Wisconsin – Milwaukee

Aug 2017 - Present

Project: Biomarker Development for Use in Regression Calibration

Research Assistant for Prof. Chiang-Ching Huang, University of Wisconsin – Milwaukee

July 2018 – Sep 2018

Project: Breast Cancer Treatment Patterns in Women Age greater than 80: A Report from the National Cancer Database

Research Assistant for Prof. Xianghua Luo, University of Minnesota-Twin Cities

Dec 2013 – Oct 2015

Project: Analysis of Missing Data for a Smoking Cessation Study