OPTIMAL SAMPLING STRATEGIES USING CASE-CONTROL STUDIES FOR BINARY SECONDARY OUTCOMES UNDER BUDGET CONSTRAINTS

By

Liang Wang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Biostatistics—Doctor of Philosophy

2024

ABSTRACT

A case-control study is efficient for investigating the association between outcomes and exposures. After conducting the primary outcome analysis, researchers can utilize the existing case-control study data to perform a secondary outcome analysis. Several methods have been proposed for analyzing secondary outcomes in case-control studies over the past few decades, but few of them have focused on the study design aspect. We propose optimal sampling strategies under a budget constraint for case-control studies with binary and Poisson secondary outcomes. We then extend our optimal sampling strategy by considering a confounder and derive the parameter of interest using doubly-weighted estimating equations. The term "optimal" refers to minimizing the variance of the estimator of the parameter of interest. We elucidate our proposed methods by developing the asymptotic variance of the estimator of the coefficient using weighted estimating equations and doubly-weighted estimating equations. Furthermore, we derive the optimal sampling ratio formulas through the Lagrange multiplier method based on certain monetary constraints. We verify our proposed methods through Monte Carlo simulation studies. Additionally, we apply our methods to empirical epidemiological studies that motivated the method development.

I dedicate this dissertation to my family and friends.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my dissertation committee members: Dr. Zhehui Luo, Dr. Chenxi Li, Dr. Honglei Chen, and Dr. Yuehua Cui for their guidance throughout my dissertation research.

I would also like to thank Michigan State University Institute of Health Policy for their support of my PhD graduate assistantship.

I would like to extend my appreciation to my family and friends for their love and emotional support during my PhD journey.

TABLE OF CONTENTS

Chapter 1 Introduction

The case-control study is designed to efficiently investigate the association between rare outcomes and exposures. To form a case group, a sample of individuals with the disease is randomly selected from the target population, while a sample of individuals without the disease is selected to form a control group. The primary outcome is the disease by which the caseness is defined. Researchers can use an existing case-control study dataset to conduct secondary outcome analysis and examine the association between secondary outcomes and covariates.

Numerous methods have been proposed for analyzing secondary outcomes in case-control studies over the past few decades. These methods include analyzing controls only (Nagelkerke et al., 1995) or cases only (Li et al., 2010), and conducting joint analysis of cases and controls while adjusting for the primary outcome (Lee et al., 1997). However, some of these methods have been considered naive as they are valid only under certain circumstances, such as a rare disease assumption (Li et al., 2010), and may produce invalid inferences. When considering the funding limitation for the secondary outcome analysis, the data for the secondary outcome analysis needs to be sampled from the cohort. Because the sampling for the secondary outcome analysis is taken from the cohort, the associations between exposure and the secondary outcome based on these samples can differ from those in the general population (Lin and Zeng, 2009). To overcome this issue, more methods that can provide valid inference have been proposed, including likelihood methods (Lin and Zeng, 2009; Jiang et al., 2006; Ghosh et al., 2013; Brownstein et al., 2022), weighted estimating equations methods (Monsees et al., 2009; Xing et al., 2016; Song et al., 2016; Sofer et al., 2017), bias correction methods (Wang and Shete, 2012; Chen et al., 2013), and semi-parametric methods (Wei et al., 2013; Tchetgen Tchetgen, 2014; Ma and Carroll, 2016).

While current methodological studies on secondary outcomes focus on inferential procedures, the corresponding sampling strategies using case-control study for secondary outcome analysis have not been investigated. In Chapter 2, we proposed an optimal sampling strategy for a case-control study with a binary secondary outcome. We derived the variance of the estimator of the exposure effect for the inverse probability weighted estimating equations and minimized the vari-

ance to obtain an optimal sampling ratio between the sample size of controls to cases, considering study cost constraints.

In Chapter 2, the secondary outcome is presented as a binary variable. However, in reality, there are situations where the secondary outcome of interest is not binary, but a count variable. For instance, in a study by Tchetgen Tchetgen (2014), the secondary outcome of interest was the number of live births, which is typically considered a count outcome. Similarly, in the Pesticides and Sense of Smell (PASS) Study (Shrestha et al., 2019), the cognitive decline scores in the survey could also be treated as a count variable.

Sample size calculation for count data has been extensively investigated. For instance, Lou et al. (2017) derived an analytic sample size formula for comparing rates of change between multiple treatment groups with repeatedly measured count outcomes using generalized estimating equations. Amatya et al. (2013) provided simple sample size expressions for determining the number of clusters in the context of multi-center randomized clinical trials. Zhu and Lakkis (2014) developed an explicit sample size calculation formula based on the likelihood function of the negative binomial model. Wang et al. (2020) provide a closed-form sample size formula accounting for the variability in cluster size in cluster randomized studies. In addition to deriving analytic sample size calculation formulas, simulation studies have also been used to determine sample size for count data, as demonstrated in the works of Lyles et al. (2007); Aban et al. (2009); Rettiganti and Nagaraja (2012). However, the sampling allocation estimation in the analysis of count secondary outcomes in case-control studies remains understudied. This situation motivated us to explore a sampling strategy for secondary case-control studies with count outcomes in Chapter 3.

In addition, our proposed sampling strategy formulas were derived by minimizing the variance of the estimator of the exposure effect using inverse probability weighted estimating equations in Chapter 2 and Chapter 3. The weights in these two chapters in estimating equations were design-based weights, representing the sampling probability for the secondary outcome analysis from the cohort. In chapter 4, we incorporate the propensity score weights into the estimating equations to create doubly-weighted estimating equations. The propensity score weights can be estimated using

cohort data, which captures the probability of exposure given the confounder. Consequently, the weights in doubly-weighted estimating equations were the product of the design weights and the propensity score weights. The general propensity score assumptions and the inference of doubly-weighted estimating equations can be found in Negi (2024). It is important to note that in Chapter 4, our focus is not on inference on the marginal effect; rather, we aim to provide an optimal sampling designs for case-control studies with binary secondary outcomes given the exposure and a confounder using doubly-weighted estimating equations.

Chapter 2 Optimal Sampling Strategies Using Case-Control Studies for Binary Secondary Outcomes under Budget Constraints

2.1 Background

Optimal sampling strategies have been fruitfully studied in a variety of epidemiology study designs. In the case-control study with a binary outcome, Demidenko (2006) found that the optimal ratio of controls to cases for fixed power is equal to the square root of the alternative odds ratio. In another paper related to the unmatched case-control study design, Demidenko (2008) gave the optimal control–case ratio for the test of an interaction between two binary covariates. Morgenstern and Winn (1983) proposed that the optimal sampling ratio is a function of the expected frequency of exposure among controls, odd ratio, and the unit cost ratio. Nam and Fears (1990) investigated optimal allocation for stratified case-control studies.

The design consideration is immaterial when we only use an existing case-control sample for secondary analysis. However, in the situation where the "true" outcome of interest is not measured in the target population or too costly to ascertain, but a proxy of the outcome or another variable strongly associated with the true outcome is routinely collected in the target population or ascertainable with little cost, then it may be more efficient to sample observations based on the proxy as if the proxy is the primary outcome in a case-control study and the true outcome of interest is the secondary outcome. The design aspect of secondary outcome analysis in case-control studies has received less attention compared to the inference procedures. In this chapter, we propose a new method to determine the optimal sampling ratio for a secondary case-control study when estimating coefficients of interest using inverse probability weighted estimating equations. The purpose of this paper is twofold. First, we provide a detailed explanation of a novel Lagrange multiplier method by developing the asymptotic variance-coariance matrix of estimators of coefficients obtained from the weighted estimating equations. Second, we showcase the optimal allocation design for a secondary outcome analysis using the method, taking into account monetary constraints.

2.2 Notation and estimation

2.2.1 Notation

Suppose the study cohort (target population) has $i = 1, ..., N$ independent subjects. Let D_i be the binary primary disease status for subject i in the cohort, where $D_i = 1$ indicates the presence of the disease, $D_i = 0$ indicates its absence. Denote by Y_i the binary secondary outcome of interest, where $Y_i = 1$ indicates the presence of a secondary disease, and $Y_i = 0$ otherwise. Let S_i be the sampling indicator, with $S_i = 1$ indicating the inclusion of ith subject in the secondary outcome analysis, and $S_i = 0$ otherwise. Assume that there are N_1 known subjects in the case group ($D_i = 1$) and N_0 known subjects in the control group ($D_i = 0$), where $N_1 + N_0 = N$. Let n_1 be the unknown sample size selected from the case group and let n_0 be the unknown sample size selected from the control group. The total number of subjects selected among N is $n = n_1 + n_0$.

Denote by $\boldsymbol{X}_i = (1, W_i)^T$ the 2×1 covariates vector for subject i, where W_i denotes the binary exposure status for subject i, with $W_i = 1$ indicating exposed, and $W_i = 0$ otherwise. Assume that the sampling probability from the study cohort for the secondary outcome analysis depends only on D_i, where $\Pr(S_i = 1 \mid D_i = d, Y_i, \boldsymbol{X}_i) = \Pr(S_i = 1 \mid D_i = d) = \pi(D_i = d) = \frac{n_d}{N_d}$, $d = 0, 1$. The goal is to find the optimal sampling ratio n_0/n_1 when we aim to examine the association between Y_i and W_i in the target population without conditioning on D_i. Let $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ be a 2×1 vector in the conditional expectation $\mathbb{E}(Y_i \mid \boldsymbol{X}_i; \boldsymbol{\beta}) = \mu(\boldsymbol{X}_i; \boldsymbol{\beta})$, where $g(\mu(\boldsymbol{X}_i; \boldsymbol{\beta})) = \boldsymbol{X}_i^T \boldsymbol{\beta}$. Since Y_i is a binary outcome, we can use the logit link function for $g(\cdot)$. The mean model $\mathbb{E}(Y_i \mid \boldsymbol{X}_i; \boldsymbol{\beta})$ can be expressed as $\mathbb{E}(Y_i \mid \boldsymbol{X}_i; \boldsymbol{\beta}) = \frac{e^{\boldsymbol{X}_i^T \boldsymbol{\beta}}}{1 + e^{\boldsymbol{X}_i^T \boldsymbol{\beta}}}$.

A valid estimator of $\boldsymbol{\beta}$ can be obtained by solving the inverse of sampling probability weighted estimating equations:

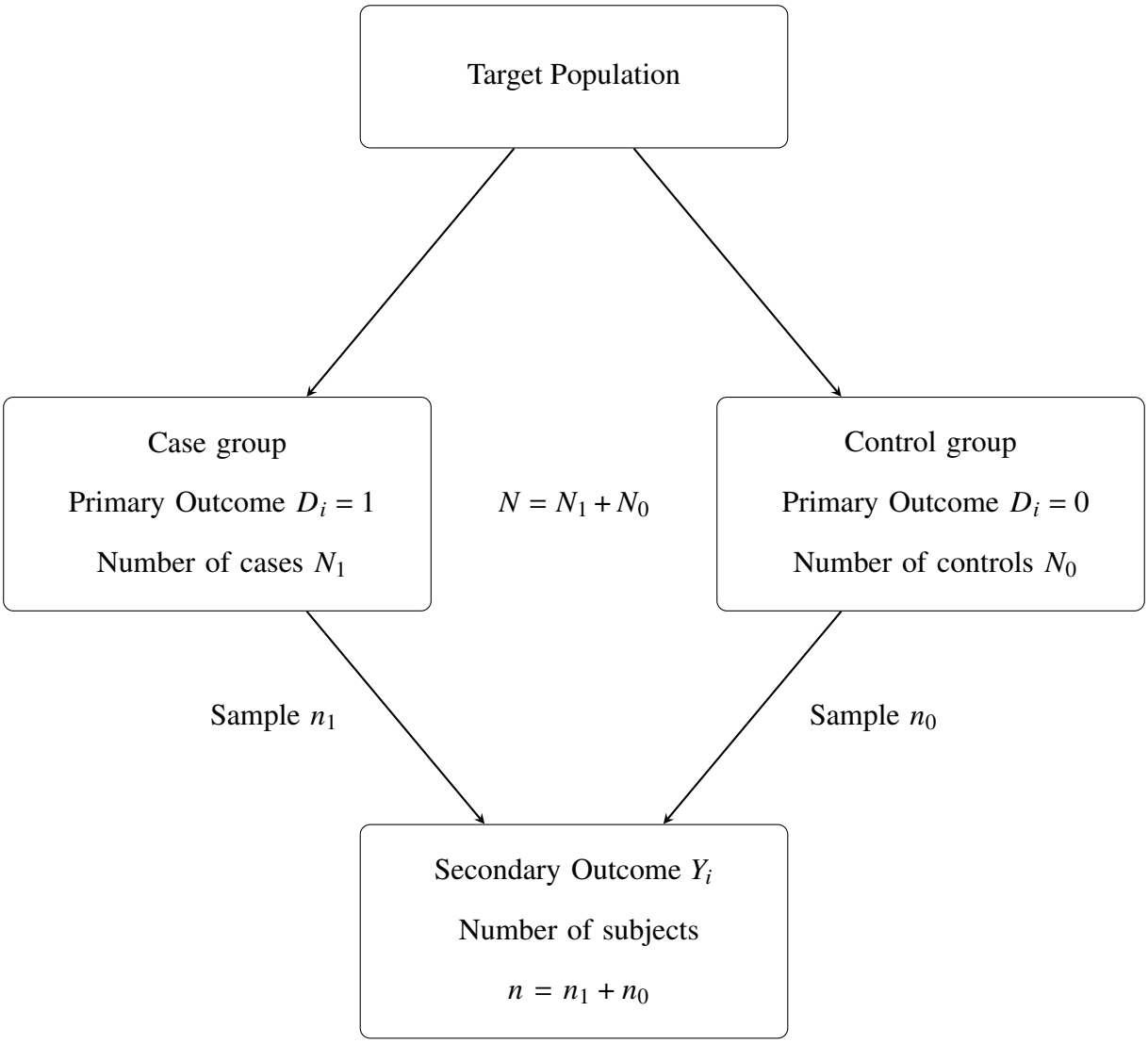
$$\sum_{i=1}^{N} U_i(\boldsymbol{\beta}) = 0,$$

where

$$U_i(\boldsymbol{\beta}) = \frac{\boldsymbol{X}_i S_i}{\pi(D_i)} \left[Y_i - \mu(\boldsymbol{X}_i) \right].$$

Figure 1 shows the secondary case-control study design flowchart.

Figure 1. Secondary Case-control Study Design Flow Chart



2.2.2 Variance estimation

In order to determine the optimal sampling ratio for the secondary outcome in a case-control design, we develop the variance of the estimator of the exposure effect, denoted as $Var(\hat{\beta}_1)$.

Let $\boldsymbol{\beta}^*$ represents the true parameters, and let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)^T$ denote the solution of estimating equations. Under some regulatory conditions, it follows that $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}^*$, asymptotically normal $N\left(\boldsymbol{\beta}^*, \frac{\boldsymbol{V}(\boldsymbol{\beta}^*)}{N}\right)$, where $\boldsymbol{V}(\boldsymbol{\beta}^*) = \boldsymbol{A}(\boldsymbol{\beta}^*)^{-1}\boldsymbol{B}(\boldsymbol{\beta}^*)\left[\boldsymbol{A}(\boldsymbol{\beta}^*)^{-1}\right]^T$. Here, $\boldsymbol{A}(\boldsymbol{\beta}^*) = \mathbb{E}\left[-\frac{\partial}{\partial \boldsymbol{\beta}^T} U(\boldsymbol{\beta}^*)\right]$, and $\boldsymbol{B}(\boldsymbol{\beta}^*) = \mathbb{E}\left[U(\boldsymbol{\beta}^*)U(\boldsymbol{\beta}^*)^T\right]$. Since $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}^*$, we can consistently estimate $\boldsymbol{V}(\boldsymbol{\beta})$ with $\boldsymbol{V}(\hat{\boldsymbol{\beta}})$. In the current study settings, we only consider the exposure variable in the estimating equation for simplicity. Therefore, it is straightforward to see that $\frac{\boldsymbol{V}(\boldsymbol{\beta})}{N}$ is a 2×2 variance-covariance matrix, and the bottom-right element of $\frac{\boldsymbol{V}(\boldsymbol{\beta})}{N}$ represents $Var(\hat{\beta}_1)$. We obtain $Var(\hat{\beta}_1)$ by computing $\frac{\boldsymbol{A}(\boldsymbol{\beta})^{-1}\boldsymbol{B}(\boldsymbol{\beta})\left[\boldsymbol{A}(\boldsymbol{\beta})^{-1}\right]^T}{N}$ directly. It is easy to verify that

$$\begin{aligned}\boldsymbol{A}(\boldsymbol{\beta}) &= \mathbb{E}\left[\frac{S_i}{\pi(D_i)} \times \frac{e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}{\left(1+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}\right)^2}\boldsymbol{X}_i\boldsymbol{X}_i^T\right] \\ &= \mathbb{E}\left[\frac{e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}{\left(1+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}\right)^2}\boldsymbol{X}_i\boldsymbol{X}_i^T\right].\end{aligned}$$

(See Appendix A.1 for the explicit derivation of $\boldsymbol{A}(\boldsymbol{\beta})$.)

We substitute $U_i(\boldsymbol{\beta})$ into $\boldsymbol{B}(\boldsymbol{\beta})$, thus

$$\boldsymbol{B}(\boldsymbol{\beta}) = \mathbb{E}\left[\frac{S_i}{[\pi(D_i)]^2}(Y_i - \mu(\boldsymbol{X}_i))^2\boldsymbol{X}_i\boldsymbol{X}_i^T\right].$$

We further simplify $\boldsymbol{B}(\boldsymbol{\beta})$ to (See Appendix A.2 for the explicit derivation of $\boldsymbol{B}(\boldsymbol{\beta})$.):

$$\boldsymbol{B}(\boldsymbol{\beta}) = \mathbb{E}\left[\frac{(\tilde{\mu}(\boldsymbol{X}_i, D_i) - \mu(\boldsymbol{X}_i))^2 + \tilde{\mu}(\boldsymbol{X}_i, D_i) - \tilde{\mu}(\boldsymbol{X}_i, D_i)^2}{\pi(D_i)}\boldsymbol{X}_i\boldsymbol{X}_i^T\right],$$

where $\tilde{\mu}(\boldsymbol{X}_i, D_i; \boldsymbol{\alpha}) = \mathbb{E}(Y_i|\boldsymbol{X}_i, D_i; \boldsymbol{\alpha}) = \frac{e^{\alpha_0+\alpha_1 w_i+\alpha_2 d_i+\alpha_3 w_i d_i}}{1+e^{\alpha_0+\alpha_1 w_i+\alpha_2 d_i+\alpha_3 w_i d_i}}$ is the expectation of secondary outcome given the primary outcome, the exposure variable and their interactions. $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2\,\alpha_3)^T$, are coefficient vectors associated with $\mu(\boldsymbol{X}_i; \boldsymbol{\beta})$ and $\tilde{\mu}(\boldsymbol{X}_i, D_i; \boldsymbol{\alpha})$, respectively. To obtain the optimal sampling ratio, the parameter set $\boldsymbol{\alpha}$ need to be pre-specified. Because both models are saturated and the parameter of interest is the odds ratio, the functional forms for

the two models do not cause issues of compatibility.

2.3 Optimal sampling strategy

Theorem 1. *Under the above study design settings with the prevalence of the exposure in the cohort, with known* $P(W=w)=p_w$ *and* $P(W=w,D=d)=p_{wd}$*, for* w *and* $d=0$ *or* 1*. Let*

$$A(\boldsymbol{\beta})=\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

and

$$B(\boldsymbol{\beta})=\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}.$$

Then $Var(\hat{\beta}_1)$*, the lower right corner element of variance-covariance matrix* $\frac{\boldsymbol{A}(\boldsymbol{\beta})^{-1}\boldsymbol{B}(\boldsymbol{\beta})\left[\boldsymbol{A}(\boldsymbol{\beta})^{-1}\right]^T}{N}$*, is given by* $\frac{\left[a_{21}^2 b_{11}-2a_{11}b_{12}a_{21}+a_{11}^2 b_{22}\right]}{(detA)^2 N}$*, where*

$$a_{11}=\frac{p_1 e^{\beta_0+\beta_1}}{\left(1+e^{\beta_0+\beta_1}\right)^2}+\frac{(1-p_1)e^{\beta_0}}{\left(1+e^{\beta_0}\right)^2},$$

$$a_{12}=a_{21}=a_{22}=\frac{p_1 e^{\beta_0+\beta_1}}{\left(1+e^{\beta_0+\beta_1}\right)^2},$$

$$b_{11}=\frac{p_{11}N_1\tilde{\tilde{\mu}}(1,1)}{n_1}+\frac{p_{10}N_0\tilde{\tilde{\mu}}(1,0)}{n_0}+\frac{p_{01}N_1\tilde{\tilde{\mu}}(0,1)}{n_1}+\frac{p_{00}N_0\tilde{\tilde{\mu}}(0,0)}{n_0},$$

$$b_{12}=b_{21}=b_{22}=\frac{p_{11}N_1\tilde{\tilde{\mu}}(1,1)}{n_1}+\frac{p_{10}N_0\tilde{\tilde{\mu}}(1,0)}{n_0},$$

where $\tilde{\tilde{\mu}}(\boldsymbol{X}_i,D_i)\equiv(\tilde{\mu}(\boldsymbol{X}_i,D_i)-\mu(\boldsymbol{X}_i))^2+\tilde{\mu}(\boldsymbol{X}_i,D_i)-\tilde{\mu}(\boldsymbol{X}_i,D_i)^2$ *and the shorthands* $\tilde{\tilde{\mu}}(1,1)$*,* $\tilde{\tilde{\mu}}(1,0)$*,* $\tilde{\tilde{\mu}}(0,1)$*,* $\tilde{\tilde{\mu}}(0,0)$ *represent the value of function* $\tilde{\tilde{\mu}}(\cdot,\cdot)$ *given the exposure value* w *and the case-control status* d*, for* w *and* $d=0$ *or* 1*, and* $detA=a_{11}a_{22}-a_{12}a_{21}$*.*

Proof. See Appendix A.3 and Appendix A.4. □

The computation of $Var(\hat{\beta}_1)$ can be performed directly by plugging in (a_{11}, $a_{12}=a_{21}=a_{22}$, b_{11}, $b_{12}=b_{21}=b_{22}$) into the expression $\frac{\left[a_{21}^2 b_{11}-2a_{11}b_{12}a_{21}+a_{11}^2 b_{22}\right]}{(detA)^2 N}$.

Consideration of the cost for data collection is a crucial aspect when conducting epidemiological research. Cost can be viewed as a special constraint in sample size calculation. In this paper, we will focus on a scenario where the total cost of the secondary outcome analysis is fixed and known. We make the assumption that the costs for a case and a control are the same.

Proposition 1. *Denote the known total cost of the secondary outcome study as $Cost$, and let c_{per} represent the known cost per individual for samples from case group or control group. The maximum sample size n for the secondary outcome analysis is given by $\frac{Cost}{c_{per}} = n_0 + n_1 = n$, where n_0 is the sample size of the selected controls , and n_1 is the sample size for selected cases. Let $R_{OD} = \frac{n_0}{n_1}$ denote the optimal design-based sampling ratio. Under the study design settings describe above, the optimal ratio can be determined as follows:*

$$R_{OD} = \frac{n_0}{n_1} = \sqrt{\frac{[\zeta(1,0) + \iota(0,0) - \kappa(1,0) + \varpi(1,0)]}{[\zeta(1,1) + \iota(0,1) - \kappa(1,1) + \varpi(1,1)]}},$$

where,

$\zeta(1,1) = a_{21}^2 p_{11} N_1 \tilde{\tilde{\mu}}(1,1)$,

$\zeta(1,0) = a_{21}^2 p_{10} N_0 \tilde{\tilde{\mu}}(1,0)$,

$\iota(0,1) = a_{21}^2 p_{01} N_1 \tilde{\tilde{\mu}}(0,1)$,

$\iota(0,0) = a_{21}^2 p_{00} N_0 \tilde{\tilde{\mu}}(0,0)$,

$\kappa(1,1) = 2a_{21} a_{11} p_{11} N_1 \tilde{\tilde{\mu}}(1,1)$,

$\kappa(1,0) = 2a_{21} a_{11} p_{10} N_0 \tilde{\tilde{\mu}}(1,0)$,

$\varpi(1,1) = a_{11}^2 p_{11} N_1 \tilde{\tilde{\mu}}(1,1)$,

$\varpi(1,0) = a_{11}^2 p_{10} N_0 \tilde{\tilde{\mu}}(1,0)$.

Proof. See Appendix A.5. □

Remark 1. We define $\tilde{\tilde{\mu}}(\boldsymbol{X}, D)$ term as Quasi Mean Squared Error(QMSE). Then,

$$\tilde{\tilde{\mu}}(\boldsymbol{X}, D) = \underbrace{\left[\underbrace{\widetilde{\mu}(\boldsymbol{X}, D) - \mu(\boldsymbol{X})}_{bias}\right]^2 + \underbrace{\widetilde{\mu}(\boldsymbol{X}, D) - \widetilde{\mu}(\boldsymbol{X}, D)^2}_{variance}}_{MSE_{x,d}}.$$

Let $q_1 = P(D = 1) = \frac{N_1}{N}$, $q_0 = P(D = 0) = \frac{N_0}{N}$. Then,

$$Var(\hat{\beta}_1) = \theta_1 \frac{q_1}{n_1 (detA)^2} + \theta_0 \frac{q_0}{n_0 (detA)^2}.$$

Where,

$\theta_1 = a_{21}^2 (p_{11} QMSE_{11} + p_{01} QMSE_{01}) + \left(a_{11}^2 - 2a_{21}a_{11}\right) p_{11} QMSE_{11}$.

$\theta_0 = a_{21}^2 (p_{10} QMSE_{10} + p_{00} QMSE_{00}) + \left(a_{11}^2 - 2a_{21}a_{11}\right) p_{10} QMSE_{00}$.

Thus, the optimal sampling ratio is $R_{OD} = \sqrt{\frac{\theta_0 q_0}{\theta_1 q_1}}$. Since $q_1 \ll q_0$. If θ_1 is big, θ_0 is small, then $R_{OD} \approx 1$. If θ_1 is small, θ_0 is big, then $R_{OD} < 1$. If $\frac{\theta_1}{\theta_0} \approx 1$, then $R_{OD} \approx \sqrt{\frac{q_0}{q_1}} = \sqrt{\frac{N_0}{N_1}}$.

2.4 Simulation

We denote the derived variance of $\hat{\beta}_1$ in Theorem 1 as $Var_{DER}(\hat{\beta}_1)$. We denote by $Var_{GMM}(\hat{\beta}_1)$ the variance of $\hat{\beta}_1$ using the Stata command "gmm". We define the empirical variance as $Var_{EMP}(\hat{\beta}_1) = \frac{1}{1000-1} \sum_{i=1}^{1000} (\hat{\beta}_1^i - \hat{\beta}_1^{mean})^2$, where $\hat{\beta}_1^i$ is the estimator of the exposure effect on one simulated dataset using the Stata command "gmm". $\hat{\beta}_1^{mean}$ is the average of $\hat{\beta}_1^i$ across 1000 simulation runs.

In this section, we conduct simulation studies to verify our derived variance formula. Specifically, we investigate whether our proposed variance $Var_{DER}(\hat{\beta}_1)$ closely approximates $Var_{EMP}(\hat{\beta}_1)$ and $Var_{GMM}(\hat{\beta}_1)$.

We provide the data generating process as follows. We simulate $i = 1, \ldots, 1,000$ datasets, with each dataset comprising $10,000$ observations (target population, i.e., cohort size N). Within each dataset, we first simulate the binary exposure variable $X \backsim Bernoulli(0.17)$. We then simulate the binary primary outcome D, with $\mathbb{E}(D|\boldsymbol{X}; \boldsymbol{\gamma}) = \frac{e^{\gamma_0 + \gamma_1 w}}{1 + e^{\gamma_0 + \gamma_1 w}}$, where $\boldsymbol{\gamma} = [\gamma_0, \gamma_1] = [-1.4, 0.7]$. Finally, we simulate the binary secondary outcome Y, with $\mathbb{E}(Y|\boldsymbol{X}, D; \boldsymbol{\alpha}) = \frac{e^{\alpha_0 + \alpha_1 w + \alpha_2 d + \alpha_3 wd}}{1 + e^{\alpha_0 + \alpha_1 w + \alpha_2 d + \alpha_3 wd}}$, where

$\boldsymbol{\alpha} = [\alpha_0, \alpha_1, \alpha_2, \alpha_3] = [-1, 0.3, 0.25, 1]$. In each simulated dataset, the prevalence of the primary outcome D and secondary Y is around 0.22 and 0.31, respectively. The number of individuals selected based on the study budget is $\frac{Cost}{c_{per}} = n = 3,000$. Table 2.1 presents a comparison of $Var_{EMP}(\hat{\beta}_1)$, $Var_{GMM}(\hat{\beta}_1)$, and our proposed variance $Var_{DER}(\hat{\beta}_1)$ under different sampling ratios with the given parameters. The first column of the table represents the ratio of n_0 to n_1. The second column displays the empirical variance for each sampling design. The third column shows the average of variances of estimators of coefficients obtained using Stata "gmm" command over 1,000 simulations. The final column is our proposed variances under different sampling designs. The results in the Table 2.1 clearly illustrates that our proposed $Var_{DER}(\hat{\beta}_1)$ is very close to $Var_{EMP}(\hat{\beta}_1)$ and $Var_{GMM}(\hat{\beta}_1)$ across all reasonable sampling ratios. It is evident that the balanced design is not the most efficient choice. By using our proposed optimal sampling formula, we obtained the optimal sampling ratio $R_{OD} = 2.63$.

Table 2.1: The comparison of $Var_{EMP}(\hat{\beta}_1)$, $Var_{GMM}(\hat{\beta}_1)$, and the proposed variance $Var_{DER}(\hat{\beta}_1)$ under different sampling designs.

$n_0 : n_1$	$Var_{EMP}(\hat{\beta}_1)$	$Var_{GMM}(\hat{\beta}_1)$	$Var_{DER}(\hat{\beta}_1)$
	E-2	E-2	E-2
1 : 1	1.239	1.182	1.179
2 : 1	0.981	0.998	0.996
2.63 : 1	0.989	0.984	0.981
3 : 1	0.957	0.989	0.985
4 : 1	0.981	1.019	1.016
1 : 2	1.729	1.657	1.657
1 : 3	2.238	2.164	2.160
1 : 4	2.796	2.680	2.669

2.5 Numerical illustration

We apply our proposed method to the secondary outcome analysis of the Pesticides and Sense of Smell (PASS) Study, which is an add-on study of the Agricultural Health Study (AHS). The PASS Study aims to better understand the relationship between high pesticide exposure events (HPEE) and olfactory impairment (OI). In the target AHS phase-4 cohort, participants were asked if they

had lost their sense of smell. Some literature has shown an association between OI and cognitive decline (Yaffe et al., 2017; Shrestha et al., 2019; Dintica et al., 2019). Thus, the investigators used the self-reported smell loss to define the sampling strata D and mailed the selected participants the Cognitive Function Instrument (CFI) questionnaire, which is used to define a dichotomous outcome of interest, Y, CFI-based cognitive compliant. We utilized our proposed optimal sampling ratio formula to obtain an efficient study design for the analysis of the secondary outcome under some reasonable scenarios of the strength of association between D and Y. Certain parameters were derived from the cohort data, e.g., $p(W = 1) = 0.14$, $\gamma = [-1.86, 0.397]$. The total sample size in case-control cohort is $N = N_0 + N_1 = 15,893 + 2,633 = 18,526$. We use a range of $\boldsymbol{\alpha}$ that are meaningful for the following scenarios: the prevalence of Y is either 0.1 or 0.2. The association between D and Y among $W = 0$, $OR_{YD|W=0} = [1.2, 1.4, 1.6]$; and among $W = 1$, $OR_{YD|W=1} = [1.5, 2.0, 2.5]$. We created Table 2.2 for these scenarios with column 1 for prevalence of Y, column 2 for $OR_{YD|W=0}$, column 3 for $OR_{YD|W=1}$ and the last column is the proposed optimal sampling ratio R_{OD} based on the formula with these given parameters. The budget constraint is $\$25,000$ for data collection, and the unit cost is \$5.40. We can observe that as the association between D and Y increases, the optimal sampling ratio decreases, resulting in fewer subjects being sampled from the control group.

Table 2.2: Optimal sampling ratio R_{OD} varies with the prevalence of the secondary outcome Y, and the association between D and Y, while keeping other parameters fixed.

$P(Y=1)$	$OR_{YD\|W=0}$	$OR_{YD\|W=1}$	β_1	α_0	α_1	α_2	α_3	R_{OD}
0.1	1.2	1.5	0.659	−2.3	0.6	0.182	0.223	4.66
	1.2	2.0	0.731	−2.35	0.6	0.182	0.511	4.35
	1.2	2.5	0.790	−2.36	0.6	0.182	0.734	4.13
	1.4	1.5	0.630	−2.36	0.6	0.336	0.069	4.62
	1.4	2.0	0.712	−2.37	0.6	0.336	0.375	4.27
	1.4	2.5	0.765	−2.39	0.6	0.336	0.580	4.07
	1.6	1.5	0.611	−2.37	0.6	0.470	−0.065	4.57
	1.6	2.0	0.684	−2.39	0.6	0.470	0.223	4.24
	1.6	2.5	0.738	−2.41	0.6	0.470	0.446	4.02
0.2	1.2	1.5	0.656	−1.515	0.6	0.182	0.223	4.89
	1.2	2.0	0.713	−1.525	0.6	0.182	0.511	4.70
	1.2	2.5	0.765	−1.535	0.6	0.182	0.734	4.58
	1.4	1.5	0.631	−1.535	0.6	0.336	0.069	4.84
	1.4	2.0	0.698	−1.545	0.6	0.336	0.375	4.64
	1.4	2.5	0.742	−1.555	0.6	0.336	0.580	4.53
	1.6	1.5	0.610	−1.555	0.6	0.470	−0.065	4.81
	1.6	2.0	0.670	−1.565	0.6	0.470	0.223	4.61
	1.6	2.5	0.720	−1.575	0.6	0.470	0.446	4.49

2.6 Conclusion

The optimal sampling strategies have been discussed for two-stage (Breslow and Chatterjee, 1999; McNamee, 2005), or so called two-phase designs (Reilly, 1996; Breslow and Cain, 1988) for case control studies. Breslow (2005) points out that two-phase designs and two-stage designs are the same study design with different terminologies. A two-stage case-control design involves determining exposure and outcome for a large sample, but covariates are measured only on a subsample (Hanley et al., 2005). The outcome of interest remains the same for both the first stage and the second stage. However, for secondary outcome analysis for case-control study, the primary outcome and the secondary outcome are different. Since the studies are different, our proposed method differs from the method used in the mentioned papers. The inference for secondary outcome analysis in case-control study has been studied in the last decades with numerous methods being

proposed. However, there is no off-the-shelf method for the sampling design when the data are deliberately collected for the secondary outcome. In this paper, we proposed an optimal sampling strategy for the analysis of the secondary outcome using weighted estimating equations. The term "optimal" refers to allocation of cases and controls that minimizes the variance of the estimator of interest given an analytic method and a fixed sample size under a budget constraint. We derive the asymptotic variance-covariance matrix of estimators of coefficients using the "Sandwich" variance-covariance matrix of a weighted estimating equations estimators. Given the variance of the estimator of the coefficient is minimal, the power for the test on exposure effect is maximal. We provide a sampling formula for achieving an efficient study design for a valid estimation strategy of the effect of interest, namely the inverse probability of sampling weighted estimating equations estimators. For different estimation strategies, the optimal sampling ratio might be different. To verify our provided formula, we conduct simulation studies. The results demonstrate that our formula performs well for both common primary outcomes and secondary outcomes. Interestingly, our findings indicate that the widely used balanced design is not always the most efficient choice for secondary outcome analysis study designs. Therefore, researchers should carefully calculate the sampling ratio when the purpose is a secondary outcome analysis.

Chapter 3 Optimal Sampling Strategies Using Case-Control Studies for Poisson Count Secondary Outcomes under Budget Constraints

3.1 Background

In this chapter, we expand upon our sampling methodology from Chapter 2 by including an explicit sampling ratio formula for case-control studies with count secondary outcomes. With this sampling formula, researchers will be able to achieve an efficient study design for their count secondary outcome analysis. The derivation of the optimal sampling ratio is based on estimating the variance of the estimator of the exposure effect using inverse probability weighted estimating equations. Our proposed framework, which incorporates inverse probability estimating equations, offers an optimal formula that can accommodate Poisson distributed count data. This chapter is organized as follows: First, we introduce general notations. Second, we derive the variance formula of the exposure effect in Poisson distributed secondary outcomes. Third, we verify our proposed variance formula through Monte Carlo simulations. Finally, we draw conclusions based on our findings.

3.2 Notation and estimation

3.2.1 Notation

Suppose the study cohort (target population) consists of N independent subjects. Let D_i be the binary case-control primary outcome, where $D_i = 1$ or 0 indicates the presence or absence of the disease. The population size of cases and controls is denoted by N_1 and N_0, respectively. We assume that N_1 and N_0 are known, and $N_1 + N_0 = N$. Let Y_i represent the Poisson distributed count secondary outcome of interest. S_i is the sampling indicator, with $S_i = 1$ indicating the inclusion in the secondary outcome analysis, and $S_i = 0$ otherwise. We define n_1 as the unknown sample size selected from the case group and n_0 as the unknown sample size selected from the control group. The total number of subjects to be selected from the target population, denoted as n, is given by $n = n_1 + n_0$. Our objective is to derive an expression of the sampling ratio n_0/n_1 under the study budget constraint, when we aim to examine the association between Y_i and W_i in the target population without conditioning on D_i. This will allow us to determine the exact sample size of cases and controls for the secondary outcome analysis. Let W_i denotes the binary

exposure status, with $W_i = 1$ indicating exposure, and $W_i = 0$ otherwise. Denote by $\boldsymbol{X}_i = (1, W_i)^T$ the 2×1 covariates vector for each subject i. Similar to Chapter 1, we assume that the sampling probability from the study cohort for the secondary outcome analysis depends only on D_i, where $\Pr(S_i = 1 \mid D_i = d, Y_i, \boldsymbol{X}_i) = \Pr(S_i = 1 \mid D_i = d) = \pi(d) = \frac{n_d}{N_d}$, $d = 0, 1$. Let $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ be the 2×1 coefficients vector. Since Y_i is Poisson distributed random variable, the mean model $\mathbb{E}(Y_i \mid \boldsymbol{X}_i)$ can be expressed as $\mathbb{E}(Y_i \mid \boldsymbol{X}_i) = \mu(\boldsymbol{X}_i) = e^{\boldsymbol{X}_i^T \boldsymbol{\beta}}$. A valid estimator of $\boldsymbol{\beta}$ can be obtained by solving the inverse of sampling probability weighted estimating equations:

$$\sum_{i=1}^{N} U_i(\boldsymbol{\beta}) = 0,$$

where

$$U(\boldsymbol{\beta}) = \frac{\boldsymbol{X}_i S_i}{\pi(D_i)} \left[Y_i - \mu(\boldsymbol{X}_i; \boldsymbol{\beta}) \right].$$

(For the simplicity, we will not include the subscript i in the following derivations).

3.2.2 Variance estimation

In this scenario, we consider one exposure variable in the mean model, therefore, it is straightforward to see the bottom-right element of variance-covariance matrix represents $Var(\hat{\beta}_1)$. The general frame work of the derivation of variance-covariance matrix of the parameters of the inverse probability weighted estimating equations can be found in Chapter 2, Section 2. We obtain $Var(\hat{\beta}_1)$ by computing $\frac{\boldsymbol{A}(\boldsymbol{\beta})^{-1} \boldsymbol{B}(\boldsymbol{\beta}) \left[\boldsymbol{A}(\boldsymbol{\beta})^{-1}\right]^T}{N}$ directly. It is easy to verify that

$$\begin{aligned} \boldsymbol{A}(\boldsymbol{\beta}) &= \mathbb{E}\left[-\frac{\partial}{\partial \boldsymbol{\beta}^T} \frac{\boldsymbol{X} S}{\pi(D)} \left(Y - e^{\boldsymbol{X}^T \boldsymbol{\beta}} \right) \right] \\ &= \mathbb{E}\left[\frac{S}{\pi(D)} \times e^{\boldsymbol{X}^T \boldsymbol{\beta}} \boldsymbol{X} \boldsymbol{X}^T \right]. \end{aligned} \tag{3.2.1}$$

Taking iterated expectations in (3.2.1), then

$\boldsymbol{A}(\boldsymbol{\beta}) = \mathbb{E}\left\{ \mathbb{E}\left[\frac{S}{\pi(D)} \times e^{\boldsymbol{X}^T \boldsymbol{\beta}} \boldsymbol{X} \boldsymbol{X}^T \middle| D \right] \right\}$.

Note that $\pi(D) = \Pr(S = 1 \mid D = d)$ only depends on D, thus

$$\boldsymbol{A}(\boldsymbol{\beta}) = \mathbb{E}\left\{\frac{\mathbb{E}(S|D)}{\pi(D)} \times \mathbb{E}\left[e^{\boldsymbol{X}^T\boldsymbol{\beta}}\boldsymbol{X}\boldsymbol{X}^T \middle| D\right]\right\}.$$

Since S is a binary variable, then $\mathbb{E}(S|D) = \Pr(S = 1 \mid D = d) = \pi(d)$ under the assumption that the sampling is only conditional on the primary outcome D, thus

$$\boldsymbol{A}(\boldsymbol{\beta}) = \mathbb{E}\left\{\mathbb{E}\left[e^{\boldsymbol{X}^T\boldsymbol{\beta}}\boldsymbol{X}\boldsymbol{X}^T \middle| D\right]\right\} = \mathbb{E}\left[e^{\boldsymbol{X}^T\boldsymbol{\beta}}\boldsymbol{X}\boldsymbol{X}^T\right].$$

It is also easy to verify that

$$\boldsymbol{B}(\boldsymbol{\beta}) = \mathbb{E}\left[\frac{S^2}{[\pi(D)]^2}(Y - \mu(\boldsymbol{X}))^2 \boldsymbol{X}\boldsymbol{X}^T\right].$$

With further simplification on $\boldsymbol{B}(\boldsymbol{\beta})$ (See Section Appendix B.1 for the explicit derivation of $\boldsymbol{B}(\boldsymbol{\beta})$),

we have,

$$\boldsymbol{B}(\boldsymbol{\beta}) = \mathbb{E}\left[\frac{(\tilde{\mu}(\boldsymbol{X}, D) - \mu(\boldsymbol{X}))^2 + \tilde{\mu}(\boldsymbol{X}, D)}{\pi(D)}\boldsymbol{X}\boldsymbol{X}^T\right],$$

where $\tilde{\mu}(\boldsymbol{X}, D; \boldsymbol{\alpha}) = \mathbb{E}(Y|\boldsymbol{X}, D; \boldsymbol{\alpha}) = e^{\alpha_0 + \alpha_1 w + \alpha_2 d + \alpha_3 w d}$. The parameter set $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2\, \alpha_3)^T$, are coefficient vectors associated with $\mu(\boldsymbol{X})$ and $\tilde{\mu}(\boldsymbol{X}, D)$, respectively. We assume $\boldsymbol{\alpha}$ are known parameters for the purpose of calculating the optimal sampling ratio.

We further calculate $\boldsymbol{A}(\boldsymbol{\beta})$ and $\boldsymbol{B}(\boldsymbol{\beta})$ by the rule of functional expectation.

$$
\begin{aligned}
A(\boldsymbol{\beta}) &= \mathbb{E}\left[e^{\boldsymbol{X}^T\boldsymbol{\beta}}\boldsymbol{X}\boldsymbol{X}^T\right] \\
&= p_{w=1}e^{\beta_0+\beta_1}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + \left[1-p_{w=1}\right]e^{\beta_0}\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \\
&= \begin{bmatrix} p_{w=1}e^{\beta_0+\beta_1}+\left[1-p_{w=1}\right]e^{\beta_0} & p_{w=1}e^{\beta_0+\beta_1} \\ p_{w=1}e^{\beta_0+\beta_1} & p_{w=1}e^{\beta_0+\beta_1} \end{bmatrix} \\
&\equiv \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}
\end{aligned}
$$

where

$$a_{11} = p_{w=1}e^{\beta_0+\beta_1}+\left[1-p_{w=1}\right]e^{\beta_0},$$

$$a_{12} = a_{21} = a_{22} = p_{w=1}e^{\beta_0+\beta_1},$$

and $p_{w=1}$ is the known prevalence of the exposure in the cohort.

Similar for $B(\boldsymbol{\beta})$, and we let $\tilde{\tilde{\mu}}(\boldsymbol{X}, D) \equiv (\tilde{\mu}(\boldsymbol{X}, D) - \mu(\boldsymbol{X}))^2 + \tilde{\mu}(\boldsymbol{X}, D)$, then

$$
\begin{aligned}
B(\boldsymbol{\beta}) &= \mathbb{E}\left[\frac{1}{\pi(D)}\tilde{\tilde{\mu}}(\boldsymbol{X}, D)\,\boldsymbol{X}\boldsymbol{X}^T\right] \\
&= \frac{p_{11}\tilde{\tilde{\mu}}(1,1)N_1}{n_1}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + \frac{p_{10}\tilde{\tilde{\mu}}(1,0)N_0}{n_0}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \\
&\quad + \frac{p_{01}\tilde{\tilde{\mu}}(0,1)N_1}{n_1}\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \frac{p_{00}\tilde{\tilde{\mu}}(0,0)N_0}{n_0}\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \\
&= \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}
\end{aligned}
$$

$$
b_{11} = \frac{p_{11}N_1\tilde{\tilde{\mu}}(1,1) + p_{01}N_1\tilde{\tilde{\mu}}(0,1)}{n_1} + \frac{p_{10}N_0\tilde{\tilde{\mu}}(1,0) + p_{00}N_0\tilde{\tilde{\mu}}(0,0)}{n_0},
$$

$$
b_{12} = b_{12} = b_{22} = \frac{p_{11}N_1\tilde{\tilde{\mu}}(1,1)}{n_1} + \frac{p_{10}N_0\tilde{\tilde{\mu}}(1,0)}{n_0},
$$

And p_{wd} is the known joint probability between the exposure and primary outcome in the cohort.

We have similar variance-covariance matrix as binary secondary outcome case from Chapter 2. So $Var(\hat{\beta}_1) = \frac{\left[a_{21}^2 b_{11} - 2a_{11}b_{12}a_{21} + a_{11}^2 b_{22}\right]}{(detA)^2 N}$ is the lower corner element of variance-covariance matrix $\frac{\boldsymbol{A}(\boldsymbol{\beta})^{-1}\boldsymbol{B}(\boldsymbol{\beta})\left[\boldsymbol{A}(\boldsymbol{\beta})^{-1}\right]^T}{N}$. Where $detA = a_{11}a_{22} - a_{12}a_{21}$. For details, see Chapter 2, Theorem 1.

3.3 Optimal sampling strategy

We consider cost as a special constraint when we minimize the variance to get the optimal sample ratio. Define the optimal design ratio as R_{OD}. We can write R_{OD} as a function of $var(\hat{\beta}_1)$, that is

$$
var(\hat{\beta}_1) = \frac{f_1 R_{OD}^2 + (f_1 + f_2)R_{OD} + f_2}{f_3 R_{OD}},
$$

where,

$$f_1 = f_1(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, p_{w=1}, p_{wd}, N_1, N_0) = \zeta(1,1) + \iota(0,1) - \kappa(1,1) + \varpi(1,1),$$

$$f_2 = f_2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, p_{w=1}, p_{wd}, N_1, N_0) = \zeta(1,0) + \iota(0,0) - \kappa(1,0) + \varpi(1,0),$$

$$f_3 = (det\boldsymbol{A})^2 nN,$$

$$(w, d = 0, 1),$$

and

$$\zeta(1,1) = a_{21}^2 p_{11} N_1 \tilde{\tilde{\mu}}(1,1),$$

$$\zeta(1,0) = a_{21}^2 p_{10} N_0 \tilde{\tilde{\mu}}(1,0),$$

$$\iota(0,1) = a_{21}^2 p_{01} N_1 \tilde{\tilde{\mu}}(0,1),$$

$$\iota(0,0) = a_{21}^2 p_{00} N_0 \tilde{\tilde{\mu}}(0,0),$$

$$\kappa(1,1) = 2a_{21} a_{11} p_{11} N_1 \tilde{\tilde{\mu}}(1,1),$$

$$\kappa(1,0) = 2a_{21} a_{11} p_{10} N_0 \tilde{\tilde{\mu}}(1,0),$$

$$\varpi(1,1) = a_{11}^2 p_{11} N_1 \tilde{\tilde{\mu}}(1,1),$$

$$\varpi(1,0) = a_{11}^2 p_{10} N_0 \tilde{\tilde{\mu}}(1,0).$$

The definition of $p_{w=1}, p_{wd}, N_1, N_0$ can be found in Chapter 2. By applying simple algebra (Demidenko, 2008), we can determine that the minimum of $var(\hat{\beta}_1)$ is achieved when $R_{OD} = \sqrt{\frac{f_2}{f_1}}$.

3.4 Simulation

To verify our proposed variance formula for the Poisson-distributed count secondary outcomes, we conducted Monte Carlo simulations. The three candidate variances, $Var_{DER}(\hat{\beta}_1)$, $Var_{GMM}(\hat{\beta}_1)$,

and $Var_{EMP}(\hat{\beta}_1)$, have the same definition as in our previous Chapter 2, Section 4.

We start by simulating $1,000$ datasets, with each dataset containing $10,000$ observations (target population, i.e., cohort size N). Within each dataset, we first simulate the binary exposure variable W with a prevalence is 17%. We then simulate the binary primary outcome D, with $\mathbb{E}(D|\boldsymbol{X}) = \frac{e^{\gamma_0+\gamma_1 w}}{1+e^{\gamma_0+\gamma_1 w}}$. The parameters γ_0 and γ_1 are set as -1.4 and 0.7 respectively. So, the prevalence of the primary outcome D in each simulated dataset is around 22%. Additionally, we simulate the Poisson distributed secondary outcome, Y, with $\mathbb{E}(Y|\boldsymbol{X}, D) = e^{\alpha_0+\alpha_1 w+\alpha_2 d+\alpha_3 wd}$. The parameters α_0, α_1,α_2,α_3 are chosen as -0.01, 0.05, 0.035, and 0.11, respectively. Therefore, the expected value $\mathbb{E}(Y)$ and variance $Var(Y)$ are equal to 1 in each simulation dataset. The value of $\hat{\beta}_1$ is 0.094, which can be estimated from a large simulation dataset. Table 3.1 compares $Var_{EMP}(\hat{\beta}_1)$, $Var_{GMM}(\hat{\beta}_1)$, and our proposed variance $Var_{DER}(\hat{\beta}_1)$ under different sampling ratios using the given parameter set. The first column of the table represents the ratio of controls to cases. The second column displays the empirical variance of the estimator of the exposure effect for each sampling ratio. The third column shows the average of variance of the estimator of the exposure effect obtained using the Stata "gmm" command across $1,000$ simulation runs. The last column shows our proposed variance formula. When we calculate $Var_{DER}(\hat{\beta}_1)$, we assume N_1 and N_0 is fixed, with $N_1 = N \times P(D = 1) = 2,205$ and $N_0 = 7,795$. Table 3.1 clearly illustrates that our proposed $Var_{DER}(\hat{\beta}_1)$ is very close to $Var_{EMP}(\hat{\beta}_1)$ and $Var_{GMM}(\hat{\beta}_1)$ across all reasonable sampling designs. By applying our optimal sampling formula with the provided pre-specified parameters, we find that the optimal sampling ratio $R_{OD} = 2.64$. Given that the number of individuals that can be selected is $\frac{Cost}{c_{per}} = n = 3,000$, we can determine the number of controls as $n_0 = 2,176$, and the number of cases as $n_1 = 824$.

3.5 Conclusion

When planning research for a secondary case-control study, an important consideration is how to achieve an efficient design. In this chapter, instead of using simulation studies to estimate the optimal sampling ratio, we provide a close-form optimal sampling formula for case-control studies with Poisson distributed secondary outcomes, which is easy to understand and implement.

Table 3.1: The comparison of empirical variance $Var_{EMP}(\hat{\beta}_1)$, $Var_{GMM}(\hat{\beta}_1)$, and the proposed variance $Var_{DER}(\hat{\beta}_1)$ under the different sampling designs with a Poisson-distributed secondary outcome Y.

$n_0 : n_1$	$Var_{EMP}(\hat{\beta}_1)$	$Var_{GMM}(\hat{\beta}_1)$	$Var_{DER}(\hat{\beta}_1)$
	E-2	E-2	E-2
1 : 1	0.275	0.261	0.261
2 : 1	0.219	0.220	0.220
2.64 : 1	0.217	0.216	0.217
3 : 1	0.213	0.217	0.217
4 : 1	0.214	0.225	0.224
1 : 2	0.366	0.348	0.366
1 : 3	0.476	0.486	0.477
1 : 4	0.539	0.588	0.590

Our proposed sampling formula can assist researchers in achieving efficient epidemiological study designs with count secondary outcomes. We conducted simulation studies to verify our proposed variance formula in accurately approximating the empirical variance and the variance from Stata "gmm" command.

Chapter 4 Optimal Sampling Strategies Using Case-Control Studies for Binary Secondary Outcomes Using Doubly-Weighted Inverse Probability Estimating Equations under Budget Constraints

4.1 Background

In the previous chapters, we presented our optimal sampling strategies for secondary case-control studies with different types of secondary outcomes, aiming to achieve an efficient case-control study design. In those scenarios, we only focused on the association between a single exposure variable with the outcome of interest. However, in this chapter, we extend our proposed optimal sampling strategies to incorporate an additional confounder. The estimators of interest were obtained with doubly-weighted estimating equation, see Chapter 1 for details.

This chapter is organized as follows: First, we derive the variance formula for the estimator of the exposure effect in the context of doubly-weighted estimating equations. Second, we present the optimal sampling formula, considering the minimization of the variance of the estimator of the exposure coefficient while accounting for a budget constraint. Third, we verify our proposed sampling formula through Monte Carlo simulations. Fourth, we apply the derived optimal sampling formula to an empirical study. Finally, we draw conclusions based on our findings.

4.2 Notation and estimation

4.2.1 Notation

This chapter employs similar notations as the previous chapters. Suppose the study cohort (target population) consists of a total N independent subjects. We denote the binary case-control primary outcome as D, where $D = 1$ indicates the presence of the disease, $D = 0$ indicates its absence. The sample size of cases and controls is denoted by N_1 and N_0, respectively. We assume that N_1 and N_0 are known, and $N_1 + N_0 = N$. We let Y_i represents the binary secondary outcome of interest. We let S_i be the sampling indicator, with $S_i = 1$ indicating the inclusion in the secondary outcome analysis, and $S_i = 0$ otherwise. We define n_1 and n_0 as the unknown sample size to be selected from the case and control group, respectively. The total number of subjects being selected from the target population, denoted by n, and $n = n_1 + n_0$. We define the total cost of

the secondary outcome study as $Cost$, and we let c_{per} represents the known cost per individual for samples from case group or control group. Our objective is to derive an expression for the optimal sampling ratio under a study budget constraint, where $n = \frac{Cost}{c_{per}} = n_1 + n_0$. Let W_i denote the binary exposure status, with $W_i = 1$ indicating exposure, and $W_i = 0$ otherwise. Let C_i denote the binary confounder, with $C_i = 0, 1$. We denote by $\boldsymbol{X}_i = (1, W_i, C_i)^T$ the 3×1 vector of covariates for each subject i. Following the approach in Chapter 1, we assume that the sampling probability from the study cohort for the secondary outcome analysis depends only on D_i. Thus we have $\Pr(S_i = 1 \mid D_i = d, Y_i, \boldsymbol{X}_i) = \Pr(S_i = 1 \mid D_i = d) = \pi(d) = \frac{n_d}{N_d}$, $d = 0, 1$. Let $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$ be a 3×1 vector of parameters of interest. Let conditional expectation $\mathbb{E}(Y_i \mid \boldsymbol{X}_i; \boldsymbol{\beta}) = \mu(\boldsymbol{X}_i; \boldsymbol{\beta})$, where $g(\mu(\boldsymbol{X}_i; \boldsymbol{\beta})) = \boldsymbol{X}_i^T \boldsymbol{\beta}$. Since Y_i is a binary outcome, we can use the logit link function for $g(\cdot)$. We let $\mathbb{E}(Y_i \mid \boldsymbol{X}_i; \boldsymbol{\beta}) = \frac{e^{\boldsymbol{X}_i^T \boldsymbol{\beta}}}{1 + e^{\boldsymbol{X}_i^T \boldsymbol{\beta}}}$. We define $P(W_i = w|c)$ as the propensity score, where $\frac{1}{P(W_i = w|c)}$ represents the inverse probability weight for the subjects in the group w for $w = 0, 1$. An appropriate estimator of $\boldsymbol{\beta}$ can be obtained by solving the doubly-weighted estimating equations:

$$\sum_{i=1}^{N} U_i(\boldsymbol{\beta}) = 0,$$

where

$$U_i(\boldsymbol{\beta}) = \frac{\boldsymbol{X}_i S_i}{\pi_i(D_i) \times P(W_i = w_i|c_i)} \left[Y_i - \mu(\boldsymbol{X}_i) \right].$$

(For the simplicity, we will not include the subscript i in the following content)

4.2.2 Variance estimation and optimal sampling ratio formula derivation

To provide the optimal sampling ratio formula for the secondary outcome in a case-control design, we derived the variance of the estimator of the exposure effect for the doubly weighted estimating equations. In this chapter, we extend the variance derivation method to consider both the exposure variable and a binary confounder in the mean model. We build upon the variance derivation method used in the previous chapters. We denote the variance of the estimator of the exposure effect as $Var(\hat{\beta}_1)$. Therefore, the variance-covariance matrix $\frac{V(\beta)}{N}$ is a 3×3 matrix, where the second

element on the diagonal of the variance-covariance matrix represents $Var(\hat{\beta}_1)$. We obtain $Var(\hat{\beta}_1)$ by computing $\frac{\boldsymbol{A}(\boldsymbol{\beta})^{-1}\boldsymbol{B}(\boldsymbol{\beta})\left[\boldsymbol{A}(\boldsymbol{\beta})^{-1}\right]^T}{N}$ directly. It is easy to verify that

$$\boldsymbol{A}(\boldsymbol{\beta}) = \mathbb{E}\left[\frac{e^{\boldsymbol{X}^T\boldsymbol{\beta}}}{P(W=w|c)\left(1+e^{\boldsymbol{X}^T\boldsymbol{\beta}}\right)^2}\boldsymbol{X}\boldsymbol{X}^T\right],$$

$$\boldsymbol{B}(\boldsymbol{\beta}) = \mathbb{E}\left[\frac{(\tilde{\mu}(\boldsymbol{X},D)-\mu(\boldsymbol{X}))^2+\tilde{\mu}(\boldsymbol{X},D)-\tilde{\mu}(\boldsymbol{X},D)^2}{[P(W=w|c)]^2\pi(D)}\boldsymbol{X}\boldsymbol{X}^T\right]$$

(See Appendix C.1 and Appendix C.2 for the explicit derivation of $\boldsymbol{A}(\boldsymbol{\beta})$ and $\boldsymbol{B}(\boldsymbol{\beta})$.)

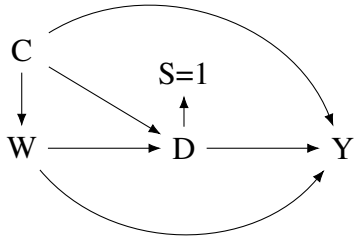
Here $\tilde{\mu}(\boldsymbol{X},D;\boldsymbol{\alpha}) = \mathbb{E}(Y|\boldsymbol{X},D;\boldsymbol{\alpha}) = \frac{e^{\alpha_0+\alpha_1 w+\alpha_2 c+\alpha_3 d+\alpha_4 wd}}{1+e^{\alpha_0+\alpha_1 w+\alpha_2 c+\alpha_3 d+\alpha_4 wd}}$, represents the expectation of secondary outcome given the primary outcome, the exposure variable, and the confounder. We assume $\mathbb{E}(Y|\boldsymbol{X},D;\boldsymbol{\alpha}) = \frac{e^{\alpha_0+\alpha_1 w+\alpha_2 c+\alpha_3 d+\alpha_4 wd}}{1+e^{\alpha_0+\alpha_1 w+\alpha_2 c+\alpha_3 d+\alpha_4 wd}}$ is the true model, and there is no interaction between W and C. We have coefficient vectors $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4)^T$ associated with $\mu(\boldsymbol{X})$ and $\tilde{\mu}(\boldsymbol{X},D)$ respectively. It is important to note that parameter sets, $\boldsymbol{\alpha}$, need to be specified according to the related literature or expert experience in order to calculate the optimal sampling ratio using our method. The details of the optimal sampling ratio R_{OD} derivation steps can be found in Appendix C.3.

4.3 Simulation

We conducted Monte Carlo simulations to evaluate the performance of the proposed variance formula presented in this chapter. The three candidate variances, $Var_{DER}(\hat{\beta}_1)$, $Var_{GMM}(\hat{\beta}_1)$, and $Var_{EMP}(\hat{\beta}_1)$, have the same definitions as in our previous Chapters. We started by simulating 1,000 target population datasets, each containing 10,000 observations. Within each dataset, we first simulated the continuous age variable C, with a mean of 64.76, and a standard deviation of 10.82. Then we simulated the binary exposure variable W conditional on C, and with a prevalence of 17.2%. We then simulated the binary primary outcome D with $\mathbb{E}(D|\boldsymbol{X}) = \frac{e^{\gamma_0+\gamma_1 w+\gamma_2 c}}{1+e^{\gamma_0+\gamma_1 w+\gamma_2 c}}$. The parameters γ_0, and γ_1 are set at -0.1, 1.1, and -0.02 respectively. Thus, the prevalence of the primary outcome D in each simulated dataset is around 24.3%. Additionally, we simulate the binary secondary outcome, Y with $\mathbb{E}(Y|\boldsymbol{X},D) = \frac{e^{\alpha_0+\alpha_1 w+\alpha_2 d+\alpha_3 c+\alpha_4 wd}}{1+e^{\alpha_0+\alpha_1 w+\alpha_2 d+\alpha_3 c+\alpha_4 wd}}$. The parameters α_0, α_1, α_2, α_3, and

α_4 are chosen as -0.65, 0.1, -0.01, 1.001, and -0.368 respectively. Therefore, the prevalence of the secondary outcome is around 26.4% in each simulated dataset. The causal DAG for the data generating process can be found in Figure 2.

Figure 2. The causal DAG for the data generating process



Where W is the exposure, C is the confounder, D is the primary outcome, Y is the secondary outcome, and S is the selection indicator. Our method requires a dichotomous confounder so we categorized age C as a binary variable L, where $L = 0$ if age is less or equal to 63, and $L = 1$ if age greater than 64. The propensity score was estimated with $P(W = 1|L)$ using the cohort data. To obtain the true value of $\hat{\boldsymbol{\beta}}$, we simulated a large dataset with $10,000,000$ observations. We then applied logistic regression for the outcome Y on W and L, resulting in an estimated coefficient $\hat{\beta}_1 = 0.177$. It is interesting to note that our target parameter of interest is not the marginal causal odds ratio, instead, we are interested in the conditional odds ratio with a binary confounder L in the outcome model, and the propensity score is also estimated using the binary confounder.

Table 4.1 compares $Var_{EMP}(\hat{\beta}_1)$, $Var_{GMM}(\hat{\beta}_1)$, and our proposed variance $Var_{DER}(\hat{\beta}_1)$ under different sampling ratios using the given parameter sets. The first column of the table represents the mean of $\hat{\beta}_1$ among $1,000$ simulations. The second column of the table represents the ratio of controls to cases. The third column displays the empirical variance of the estimator of the exposure effect for each sampling ratio. The fourth column shows the average of variance of the estimator of the exposure effect obtained using the Stata "gmm" command across $1,000$ simulation runs. The last column shows the value of using our proposed method. When calculating $Var_{DER}(\hat{\beta}_1)$, we assume N_1 and N_0 is fixed, with $N_1 = N \times P(D = 1) = 2,426$ and $N_0 = N - N_1 = 7,574$. We declare

Table 4.1: The comparison of empirical variance $Var_{EMP}(\hat{\beta}_1)$, $Var_{GMM}(\hat{\beta}_1)$, and the proposed variance $Var_{DER}(\hat{\beta}_1)$ under the different sampling designs with a binary secondary outcome Y.

$n_0 : n_1$	$\hat{\beta}_1^{mean}$	$Var_{EMP}(\hat{\beta}_1)$	$Var_{GMM}(\hat{\beta}_1)$	$Var_{DER}(\hat{\beta}_1)$
1 : 1	0.175	0.013	0.013	0.012
1.9 : 1	0.175	0.011	0.012	0.011
2 : 1	0.177	0.011	0.012	0.011
3 : 1	0.171	0.011	0.012	0.012
4 : 1	0.176	0.014	0.013	0.013
1 : 2	0.176	0.016	0.017	0.017
1 : 3	0.166	0.023	0.022	0.021
1 : 4	0.180	0.028	0.027	0.026

that this method could be equivalent to using each simulated dataset to calculate $Var_{DER}(\hat{\beta}_1)$ and obtaining the average of $Var_{DER}(\hat{\beta}_1)$ among $1,000$ simulation runs. This equivalence arises because N_1 and N_0 are calculated based on the the expected value of D. Table 4.1 clearly illustrates that our proposed $Var_{DER}(\hat{\beta}_1)$ is very close to $Var_{EMP}(\hat{\beta}_1)$ and $Var_{GMM}(\hat{\beta}_1)$ across all reasonable sampling ratios. By applying our optimal sampling formula with the provided pre-specified parameters, we find that the optimal sampling ratio $R_{OD} = 1.90$. Given that the number of individuals that can be selected is $\frac{Cost}{c_{per}} = n = 3,000$, we can determine the number of controls as $n_0 = 1,965$, and the number of cases as $n_1 = 1034$.

4.4 Empirical illustration

We applied our proposed optimal sampling ratio formula to the Pesticides and Sense of Smell (PASS) Study to develop an efficient study design for the purpose of analyzing the association between a binary secondary outcome with the exposure and a confounder. Detailed information about the PASS study can be found in Chapter 2 of our work. In this study, the primary outcome of interest is olfactory impairment, denoted as D, while the secondary outcome of interest is cognitive decline, denoted as Y. The binary exposure variable is the high pesticide exposure estimate (HPEE), denoted as W, and age serves as the confounder, denoted as L. Our purpose is to provide an efficient sampling design that aims to examine the association between Y and W given L in the target population without conditioning on D, using doubly-weighted estimating equations. Our

optimal sampling formula requires certain parameters to be given as priors. Some parameters can be derived from the cohort data. For example, the prevalence of the exposure W in the cohort is 14%. The total sample size in the case-control cohort is $N = N_0 + N_1 = 15,893 + 2,633 = 18,526$. The propensity score and the joint probability between W, L, and D can also be estimated from the cohort data. The prevalence of Y is fixed at 0.1 and 0.2. The study budget is \$25,000 for data collection and the unit cost is \$5.40 for the sampling of cases and controls. We also consider varying the association between the primary outcome D and the secondary outcome Y in the exposure group with $OR_{YD|W=1} = [1.5, 2.0, 2.5]$, and the non-exposure group $OR_{YD|W=0} = [1.2, 1.4, 1.6]$. Table 4.2 lists the optimal sampling ratios and the exact number of samples from cases and controls for the above given parameter scenarios by using our proposed optimal sampling formula. The advantage of the "T-table" liked format is that it provides researchers with an intuitive understanding of the sampling ratio when they have knowledge of a range of parameters. Furthermore, the table demonstrates that a stronger association between Y and D results in a smaller sampling ratio, meaning fewer subjects in the control group will be sampled and more subjects in the case group will be sampled.

Table 4.2: Optimal sample ratio R_{OD} varies with the prevalence of the secondary outcome Y, and the association between D and Y, while keeping other parameters fixed.

$P(Y=1)$	$OR_{YD\|W=0}$	$OR_{YD\|W=1}$	α_3	α_4	R_{OD}	n_0	n_1
0.1	1.200	1.500	0.182	0.223	3.796	3,664	965
	1.200	2.000	0.182	0.511	3.556	3,612	1,016
	1.200	2.500	0.182	0.734	3.401	3,557	1,051
	1.400	1.500	0.336	0.069	3.762	3,657	972
	1.400	2.000	0.336	0.375	3.503	3,601	1,028
	1.400	2.500	0.336	0.58	3.365	3,568	1,061
	1.600	1.500	0.470	−0.065	3.735	3,651	978
	1.600	2.000	0.470	0.223	3.489	3,598	1,031
	1.600	2.500	0.470	0.446	3.331	3,560	1,069
0.2	1.200	1.500	0.182	0.223	3.957	3,695	934
	1.200	2.000	0.182	0.511	3.845	3,674	955
	1.200	2.500	0.182	0.734	3.788	3,662	967
	1.400	1.500	0.336	0.069	3.934	3,691	938
	1.400	2.000	0.336	0.375	3.815	3,667	961
	1.400	2.500	0.336	0.58	3.768	3,658	971
	1.600	1.500	0.470	−0.065	3.915	3,687	942
	1.600	2.000	0.470	0.223	3.794	3,664	965
	1.600	2.500	0.470	0.446	3.737	3,651	977

4.5 Conclusion

In Chapter 2 and Chapter 3, we proposed our optimal sampling formula for binary and count secondary outcomes with one exposure variable in the mean model. In this chapter, we extended our optimal sampling formula by considering a binary confounder in the mean model and using doubly-weighted estimating equations. We derived the variance of the estimator of the exposure effect of the doubly-weighted estimating equations and then minimized the variance formula with the cost as a constraint to obtain the optimal sampling formula. To verify our sampling formula, we conducted Monte Carlo simulations and compared our proposed variance formula with the empirical variance and the variance from the Stata "gmm" package. Our results showed that these candidate variances were very close. Finally, we applied our proposed optimal sampling formula to an empirical study and provided an efficient study design.

Chapter 5 Conclusion

In Chapter 2 and Chapter 3, we proposed our optimal sampling formulas using case-control studies for binary and count secondary outcomes with one exposure variable in the mean model. In chapter 4, we extended our optimal sampling formula by considering a binary confounder in the mean model and using doubly-weighted estimating equations. We derived the variance of the estimator of the exposure effects of weighted estimating equations and doubly-weighted estimating equations. Then, we minimized the variance formulas with the study cost as a constraint to obtain the optimal sampling ratios. To verify our proposed optimal sampling ratio formulas, we conducted Monte Carlo simulations and compared our proposed variance formula with the empirical variance and the variance from the Stata "gmm" package. Our simulation results showed that these candidate variances were very close in all simulations in Chapter 2, Chapter 3 and Chapter 4. Finally, we applied our proposed optimal sampling formulas to empirical studies and provided efficient study designs.

There are several interesting directions for future research. First, our proposed sampling strategy considers the inclusion of one additional confounder in the binary case. However, when adding more confounders, the sampling formula may differ. Second, there are various types of count data, but our provided formula primarily focuses on Poisson count data. Other count outcomes in epidemiology, such as the number of emergency room visits or the number of falls in nursing homes, may exhibit an excess of zero values. Sample size determination formulas have been proposed for zero-inflated Poisson distributed outcomes (Zhou et al., 2022) in cluster randomized trials. Thus, an intriguing direction would be to determine the optimal sampling ratio for zero-inflated count outcomes or hurdle outcomes in secondary case-control studies, particularly for count data that exhibit an excess of zeros. Another interesting extension is to develop an R Shiny app that can automatically calculate the optimal sampling ratio when researchers provide specific parameters. Finally, we assumed that the cost is the same in the case and control groups for the secondary outcome analysis. In practice, there may be situations where the costs of data collection in cases and controls are unequal. Therefore, the cost constraint needs to be re-considered. Meanwhile, in all

three chapters, we consider study cost as a constraint. Thus, the total number of sample sizes that we can select is fixed. In the future, it would also be interesting to consider power as a constraint. We can obtain the optimal sampling ratio under a minimum required power. We can explore how variations in power will affect the sampling ratio.

BIBLIOGRAPHY

Aban, I. B., Cutter, G. R., & Mavinga, N. (2009). Inferences and power analysis concerning two negative binomial distributions with an application to MRI lesion counts data. *Computational Statistics & Data Analysis*, *53*(3), 820–833.

Amatya, A., Bhaumik, D., & Gibbons, R. D. (2013). Sample size determination for clustered count data. *Statistics in Medicine*, *32*(24), 4162–4179.

Breslow, N. E. (2005). Case–Control Study, Two-Phase. In P. Armitage & T. Colton (Eds.), *Encyclopedia of Biostatistics* (1st ed.). Wiley.

Breslow, N. E., & Cain, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, *75*(1), 11–20.

Breslow, N. E., & Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *48*(4), 457–468.

Brownstein, N. C., Cai, J., Smith, S., Diatchenko, L., Slade, G. D., & Bair, E. (2022). Modeling Secondary Phenotypes Conditional on Genotypes in Case–Control Studies. *Stats*, *5*(1), 203–214.

Chen, H. Y., Kittles, R., & Zhang, W. (2013). Bias correction to secondary trait analysis with case-control design. *Statistics in Medicine*, *32*(9), 1494–1508.

Demidenko, E. (2006). Sample size determination for logistic regression revisited. *Statistics in Medicine*, *26*(18), 3385–3397.

Demidenko, E. (2008). Sample size and optimal design for logistic regression with binary interaction. *Statistics in Medicine*, *27*(1), 36–46.

Dintica, C. S., Marseglia, A., Rizzuto, D., Wang, R., Seubert, J., Arfanakis, K., Bennett, D. A., & Xu, W. (2019). Impaired olfaction is associated with cognitive decline and neurodegeneration in the brain. *Neurology*, *92*(7).

Ghosh, A., Wright, F. A., & Zou, F. (2013). Unified Analysis of Secondary Traits in Case–Control Association Studies. *Journal of the American Statistical Association*, *108*(502), 566–576.

Hanley, J. A., Csizmadi, I., & Collet, J.-P. (2005). Two-Stage Case-Control Studies: Precision of Parameter Estimates and Considerations in Selecting Sample Size. *American Journal of Epidemiology*, *162*(12), 1225–1234.

Jiang, Y., Scott, A. J., & Wild, C. J. (2006). Secondary analysis of case-control data. *Statistics in Medicine*, *25*(8), 1323–1339.

Lee, A. J., McMURCHY, L., & Scott, A. J. (1997). RE-USING DATA FROM CASE-CONTROL STUDIES. *Statistics in Medicine*, *16*(12), 1377–1389.

Li, H., Gail, M. H., Berndt, S., & Chatterjee, N. (2010). Using cases to strengthen inference on the association between single nucleotide polymorphisms and a secondary phenotype in genome-wide association studies. *Genetic Epidemiology*, *34*(5), 427–433.

Lin, D. Y., & Zeng, D. (2009). Proper analysis of secondary phenotype data in case-control association studies. *Genetic Epidemiology*, *33*(3), 256–265.

Lou, Y., Cao, J., Zhang, S., & Ahn, C. (2017). Sample size estimation for a two-group comparison of repeated count outcomes using GEE. *Communications in Statistics - Theory and Methods*, *46*(14), 6743–6753.

Lyles, R. H., Lin, H.-M., & Williamson, J. M. (2007). A practical approach to computing power for generalized linear models with nominal, count, or ordinal responses. *Statistics in Medicine*, *26*(7), 1632–1648.

Ma, Y., & Carroll, R. J. (2016). Semiparametric Estimation in the Secondary Analysis of Case–Control Studies. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *78*(1), 127–151.

McNamee, R. (2005). Optimal design and efficiency of two-phase case–control studies with error-prone and error-free exposure measures. *Biostatistics*, *6*(4), 590–603.

Monsees, G. M., Tamimi, R. M., & Kraft, P. (2009). Genome-wide association scans for secondary traits using case-control samples. *Genetic Epidemiology*, *33*(8), 717–728.

Morgenstern, H., & Winn, D. M. (1983). A Method for determining the sampling ratio in epidemiologic studies. *Statistics in Medicine*, *2*(3), 387–396.

Nagelkerke, N. J. D., Moses, S., Plummer, F. A., Brunham, R. C., & Fish, D. (1995). Logistic regression in case-control studies: The effect of using independent as dependent variables. *Statistics in Medicine*, *14*(8), 769–775.

Nam, J.-M. (1973). Optimum Sample Sizes for the Comparison of the Control and Treatment. *Biometrics*, *29*(1), 101.

Negi, A. (2024). Doubly weighted M-estimation for nonrandom assignment and missing outcomes. *Journal of Causal Inference*, *12*(1), 20230016.

Reilly, M. (1996). Optimal Sampling Strategies for Two-Stage Studies. *American Journal of Epidemiology*, *143*(1), 92–100.

Rettiganti, M., & Nagaraja, H. N. (2012). Power Analyses for Negative Binomial Models with Application to Multiple Sclerosis Clinical Trials. *Journal of Biopharmaceutical Statistics*, *22*(2), 237–259.

Shrestha, S., Kamel, F., Umbach, D. M., Freeman, L. E. B., Koutros, S., Alavanja, M., Blair, A., Sandler, D. P., & Chen, H. (2019). High Pesticide Exposure Events and Olfactory Impairment among U.S. Farmers. *Environmental Health Perspectives*, *127*(1), 017005.

Sofer, T., Cornelis, M., Kraft, P., & Tchetgen, T., Eric. (2017). Control function assisted IPW estimation with a secondary outcome in case-control stuies. *Statistica Sinica*, *27*(2), 785–804.

Song, X., Ionita-Laza, I., Liu, M., Reibman, J., & Wei, Y. (2016). A General and Robust Framework for Secondary Traits Analysis. *Genetics*, *202*(4), 1329–1343.

Tchetgen Tchetgen, E. J. (2014). A general regression framework for a secondary outcome in case-control studies. *Biostatistics*, *15*(1), 117–128.

Wang, J., & Shete, S. (2012). Analysis of Secondary Phenotype Involving the Interactive Effect of the Secondary Phenotype and Genetic Variants on the Primary Disease: Secondary Phenotype Analysis. *Annals of Human Genetics*, *76*(6), 484–499.

Wang, J., Zhang, S., & Ahn, C. (2020). Sample size calculation for count outcomes in cluster randomization trials with varying cluster sizes. *Communications in Statistics - Theory and Methods*, *49*(1), 116–124.

Wei, J., Carroll, R. J., Müller, U. U., Keilegom, I. V., & Chatterjee, N. (2013). Robust estimation for homoscedastic regression in the secondary analysis of case-control data: *Secondary Analysis of Case-Control Data*. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*(1), 185–206.

Xing, C., M. McCarthy, J., Dupuis, J., Adrienne Cupples, L., B. Meigs, J., Lin, X., & S. Allen, A. (2016). Robust analysis of secondary phenotypes in case-control genetic association studies: Robust analysis of secondary phenotypes in case-control genetic association studies. *Statistics in Medicine*, *35*(23), 4226–4237.

Yaffe, K., Freimer, D., Chen, H., Asao, K., Rosso, A., Rubin, S., Tranah, G., Cummings, S., & Simonsick, E. (2017). Olfaction and risk of dementia in a biracial cohort of older adults. *Neurology*, *88*(5), 456–462.

Zhou, Z., Li, D., & Zhang, S. (2022). Sample size calculation for cluster randomized trials with zero-inflated count outcomes. *Statistics in Medicine*, *41*(12), 2191–2204.

APPENDIX A PROOFS OF CHAPTER 2

A.1 Derivation of $A(\boldsymbol{\beta})$

Proof. We plug $\frac{X_i S_i}{\pi(D_i)}\left[Y_i - \mu(\boldsymbol{X}_i)\right]$ into $A(\boldsymbol{\beta})$, then we have

$$\begin{aligned} A(\boldsymbol{\beta}) &= \mathbb{E}\left[-\frac{\partial}{\partial \boldsymbol{\beta}^T}\frac{\boldsymbol{X}_i^T S_i}{\pi(D_i)}\left(Y_i - \frac{e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}{1+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}\right)\right] \\ &= \mathbb{E}\left[\frac{S_i}{\pi(D_i)} \times \frac{e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}{\left(1+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}\right)^2}\boldsymbol{X}_i\boldsymbol{X}_i^T\right]. \end{aligned}$$

Taking iterated expectations on it, then

$$A(\boldsymbol{\beta}) = \mathbb{E}\left\{\mathbb{E}\left[\frac{S_i}{\pi(D_i)} \times \frac{e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}{\left(1+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}\right)^2}\boldsymbol{X}_i\boldsymbol{X}_i^T \middle| D_i\right]\right\}.$$

Note that $\pi(D_i) = \Pr(S_i = 1 \mid D_i = j)$ only depends on D_i, thus

$$A(\boldsymbol{\beta}) = \mathbb{E}\left\{\frac{\mathbb{E}(S_i|D_i)}{\pi(D_i)} \times \mathbb{E}\left[\frac{e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}{\left(1+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}\right)^2}\boldsymbol{X}_i\boldsymbol{X}_i^T \middle| D_i\right]\right\}.$$

Since S_i is a binary variable, then $\mathbb{E}(S_i|D_i) = \Pr(S_i = 1 \mid D_i = j) = \pi(D_i)$ under the assumption that the sampling probability from the study cohort for the secondary outcome analysis depends only on the primary outcome. Thus

$$\begin{aligned} A(\boldsymbol{\beta}) &= \mathbb{E}\left\{\mathbb{E}\left[\frac{e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}{\left(1+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}\right)^2}\boldsymbol{X}_i\boldsymbol{X}_i^T \middle| D_i\right]\right\} \\ &= \mathbb{E}\left[\frac{e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}{\left(1+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}\right)^2}\boldsymbol{X}_i\boldsymbol{X}_i^T\right]. \end{aligned}$$

□

A.2 Derivation of $\boldsymbol{B}(\boldsymbol{\beta})$

Proof. Similar to the derivation of $\boldsymbol{A}(\boldsymbol{\beta})$, we first plug $\frac{X_i S_i}{\pi(D_i)}\left[Y_i - \mu(\boldsymbol{X}_i)\right]$ into $\boldsymbol{B}(\boldsymbol{\beta})$, then we have

$$\boldsymbol{B}(\boldsymbol{\beta}) = \mathbb{E}\left[\frac{S_i^2}{\left[\pi(D_i)\right]^2}\left(Y_i - \mu(\boldsymbol{X}_i)\right)^2 \boldsymbol{X}_i \boldsymbol{X}_i^T\right].$$

The sampling indicator variable S_i is a binary variable, thus $S_i^2 = S_i$, then

$$\boldsymbol{B}(\boldsymbol{\beta}) = \mathbb{E}\left[\frac{S_i}{\left[\pi(D_i)\right]^2}\left(Y_i - \mu(\boldsymbol{X}_i)\right)^2 \boldsymbol{X}_i \boldsymbol{X}_i^T\right].$$

Then take the iterated expectations,

$$\begin{aligned}
\boldsymbol{B}(\boldsymbol{\beta}) &= \mathbb{E}\left\{\mathbb{E}\left[\frac{S_i}{\left[\pi(D_i)\right]^2}\left(Y_i - \mu(\boldsymbol{X}_i)\right)^2 \boldsymbol{X}_i \boldsymbol{X}_i^T \middle| \boldsymbol{X}_i, D_i\right]\right\} \\
&= \mathbb{E}\left\{\frac{\mathbb{E}(S_i | D_i)}{\left[\pi(D_i)\right]^2}\mathbb{E}\left[\left(Y_i - \mu(\boldsymbol{X}_i)\right)^2 \boldsymbol{X}_i \boldsymbol{X}_i^T \middle| \boldsymbol{X}_i, D_i\right]\right\} \\
&= \mathbb{E}\left[\frac{1}{\pi(D_i)}\mathbb{E}\left[\left(Y_i - \mu(\boldsymbol{X}_i)\right)^2 \middle| \mathbf{X}_i, D_i\right] \boldsymbol{X}_i \boldsymbol{X}_i^T\right].
\end{aligned}$$

We can simplify the expectation $\mathbb{E}\left[\left(Y_i - \mu(\boldsymbol{X}_i)\right)^2 \middle| \boldsymbol{X}_i, D_i\right]$,

$$\begin{aligned}
\mathbb{E}\left[\left(Y_i - \mu(\boldsymbol{X}_i)\right)^2 \middle| \boldsymbol{X}_i, D_i\right] &= \mathbb{E}\left[Y_i^2 - 2Y_i\mu(\boldsymbol{X}_i) + \mu(\boldsymbol{X}_i)^2 \middle| \boldsymbol{X}_i, D_i\right] \\
&= \mathbb{E}\left(Y_i^2 \middle| \boldsymbol{X}_i, D_i\right) - 2\mathbb{E}\left[Y_i\mu(\boldsymbol{X}_i) | \boldsymbol{X}_i, D_i\right] \\
&\quad + \mathbb{E}\left[\mu(\boldsymbol{X}_i)^2 | \boldsymbol{X}_i, D_i\right] \\
&= \mathbb{E}\left(Y_i^2 \middle| \boldsymbol{X}_i, D_i\right) \\
&\quad - 2\mu(\boldsymbol{X}_i)\mathbb{E}(Y_i | \boldsymbol{X}_i, D_i) + \mu(\boldsymbol{X}_i)^2.
\end{aligned}$$

Under the above study design setting, Y_i is a binary secondary outcome, so

$$\tilde{\mu}(\boldsymbol{X}_i, D_i) \equiv \mathbb{E}\left(Y_i^2 \middle| \boldsymbol{X}_i, D_i\right) = \mathbb{E}(Y_i | \boldsymbol{X}_i, D_i).$$

Plug $\tilde{\mu}(\boldsymbol{X}_i, D_i)$ into $\mathbb{E}\left[(Y_i - \mu(\boldsymbol{X}_i))^2 \middle| \boldsymbol{X}_i, D_i\right]$, we have

$$\begin{aligned}
\mathbb{E}\left[(Y_i - \mu(\boldsymbol{X}_i))^2 \middle| \boldsymbol{X}_i, D_i\right] &= \tilde{\mu}(\boldsymbol{X}_i, D_i) - 2\mu(\boldsymbol{X}_i)\tilde{\mu}(\boldsymbol{X}_i, D_i) + \mu(\boldsymbol{X}_i)^2 \\
&= \tilde{\mu}(\boldsymbol{X}_i, D_i)^2 - 2\mu(\boldsymbol{X}_i)\tilde{\mu}(\boldsymbol{X}_i, D_i) + \mu(\boldsymbol{X}_i)^2 \\
&\quad + \tilde{\mu}(\boldsymbol{X}_i, D_i) - \tilde{\mu}(\boldsymbol{X}_i, D_i)^2 \\
&= [\tilde{\mu}(\boldsymbol{X}_i, D_i) - \mu(\boldsymbol{X}_i)]^2 + \tilde{\mu}(\boldsymbol{X}_i, D_i) - \tilde{\mu}(\boldsymbol{X}_i, D_i)^2 .
\end{aligned}$$

Now substitute $[\tilde{\mu}(\boldsymbol{X}_i, D_i) - \mu(\boldsymbol{X}_i)]^2 + \tilde{\mu}(\boldsymbol{X}_i, D_i) - \tilde{\mu}(\boldsymbol{X}_i, D_i)^2$ into $\boldsymbol{B}(\boldsymbol{\beta})$, so

$$\begin{aligned}
\boldsymbol{B}(\boldsymbol{\beta}) &= \mathbb{E}\left[\frac{1}{\pi(D_i)}\mathbb{E}_i\left\{\left[[\tilde{\mu}(\boldsymbol{X}_i, D_i) - \mu(\boldsymbol{X}_i)]^2 + \tilde{\mu}(\boldsymbol{X}_i, D_i) - \tilde{\mu}(\boldsymbol{X}_i, D_i)^2\right] \middle| \boldsymbol{X}_i, D_i\right\}\boldsymbol{X}_i\boldsymbol{X}_i^T\right] \\
&= \mathbb{E}\left[\frac{1}{\pi(D_i)}\left[(\tilde{\mu}(\boldsymbol{X}_i, D_i) - \mu(\boldsymbol{X}_i))^2 + \tilde{\mu}(\boldsymbol{X}_i, D_i) - \tilde{\mu}(\boldsymbol{X}_i, D_i)^2\right]\boldsymbol{X}_i\boldsymbol{X}_i^T\right].
\end{aligned}$$

Denote $\tilde{\tilde{\mu}}(\boldsymbol{X}_i, D_i) \equiv (\tilde{\mu}(\boldsymbol{X}_i, D_i) - \mu(\boldsymbol{X}_i))^2 + \tilde{\mu}(\boldsymbol{X}_i, D_i) - \tilde{\mu}(\boldsymbol{X}_i, D_i)^2$,

$$\boldsymbol{B}(\boldsymbol{\beta}) = \mathbb{E}\left[\frac{\tilde{\tilde{\mu}}(\boldsymbol{X}_i, D_i)}{\pi(D_i)}\boldsymbol{X}_i\boldsymbol{X}_i^T\right].$$

□

A.3 Derivation of $A^{-1}B\left[A^{-1}\right]^T$

Proof. The inverse of matrix A is $\frac{1}{detA}\begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$, where $detA = a_{11}a_{22} - a_{12}a_{21}$.

Then $[A^{-1}]^T = \frac{1}{detA}\begin{bmatrix} a_{22} & -a_{21} \\ -a_{12} & a_{11} \end{bmatrix}$,

$$A^{-1} \times B = \frac{1}{detA}\begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} \times \begin{bmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{bmatrix}$$

$$= \frac{1}{detA}\begin{bmatrix} a_{22}b_{11} - a_{12}b_{12} & a_{22}b_{12} - a_{12}b_{22} \\ -a_{21}b_{11} + a_{11}b_{12} & -a_{21}b_{12} + a_{11}b_{22} \end{bmatrix},$$

and

$$A^{-1}B[A^{-1}]^T = \frac{1}{detA}\begin{bmatrix} a_{22}b_{11} - a_{12}b_{12} & a_{22}b_{12} - a_{12}b_{22} \\ -a_{21}b_{11} + a_{11}b_{12} & -a_{21}b_{12} + a_{11}b_{22} \end{bmatrix} \times \frac{1}{detA}\begin{bmatrix} a_{22} & -a_{21} \\ -a_{12} & a_{11} \end{bmatrix}$$

$$= \frac{1}{(detA)^2}\begin{bmatrix} \Omega_1 & \Omega_2 \\ \Omega_3 & \Omega_4 \end{bmatrix},$$

where,

$\Omega_1 = (a_{22}b_{11} - a_{12}b_{12})a_{22} - (a_{22}b_{12} - a_{12}b_{22})a_{12}$,

$\Omega_2 = (a_{12}b_{12} - a_{22}b_{11})a_{21} + (a_{22}b_{12} - a_{12}b_{22})a_{11}$,

$\Omega_3 = (a_{11}b_{12} - a_{21}b_{11})a_{22} + (a_{21}b_{12} - a_{11}b_{22})a_{12}$,

$\Omega_4 = (a_{21}b_{11} - a_{11}b_{12})a_{21} + (a_{11}b_{22} - a_{21}b_{12})a_{11}$. □

A.4 Proof of Theory 1

Proof. By the rule of functional expectation $\mathbb{E}(g(W)) = \sum_w g(w) p_W(w)$, we have

$$
\begin{aligned}
A(\boldsymbol{\beta}) &= \mathbb{E}\left[\frac{e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}}{\left(1+e^{\boldsymbol{X}_i^T\boldsymbol{\beta}}\right)^2}\boldsymbol{X}_i\boldsymbol{X}_i^T\right] \\
&= \frac{p_{w=1}e^{\beta_0+\beta_1}}{\left(1+e^{\beta_0+\beta_1}\right)^2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + \frac{[1-p_{w=1}]\,e^{\beta_0}}{\left(1+e^{\beta_0}\right)^2}\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \\
&= \begin{bmatrix} \frac{p_{w=1}e^{\beta_0+\beta_1}}{\left(1+e^{\beta_0+\beta_1}\right)^2} + \frac{[1-p_{w=1}]e^{\beta_0}}{\left(1+e^{\beta_0}\right)^2} & \frac{p_{w=1}e^{\beta_0+\beta_1}}{\left(1+e^{\beta_0+\beta_1}\right)^2} \\ \frac{p_{w=1}e^{\beta_0+\beta_1}}{\left(1+e^{\beta_0+\beta_1}\right)^2} & \frac{p_{w=1}e^{\beta_0+\beta_1}}{\left(1+e^{\beta_0+\beta_1}\right)^2} \end{bmatrix} \\
&\equiv \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}
\end{aligned}
$$

where

$$
a_{11} = \frac{p_{w=1}e^{\beta_0+\beta_1}}{\left(1+e^{\beta_0+\beta_1}\right)^2} + \frac{[1-p_{w=1}]\,e^{\beta_0}}{\left(1+e^{\beta_0}\right)^2},
$$

$$
a_{12} = a_{21} = a_{22} = \frac{p_{w=1}e^{\beta_0+\beta_1}}{\left(1+e^{\beta_0+\beta_1}\right)^2}.
$$

Similarly, using rule of function of expectation on $B(\boldsymbol{\beta})$, we have,

$$
\begin{aligned}
B(\boldsymbol{\beta}) &= \mathbb{E}\left[\frac{1}{\pi(D_i)}\tilde{\tilde{\mu}}(\boldsymbol{X}_i, D_i)\,\boldsymbol{X}_i\boldsymbol{X}_i^T\right] \\
&= \frac{p_{11}\tilde{\tilde{\mu}}(1,1)\,N_1}{n_1}\mathbf{1}\mathbf{1}' + \frac{p_{10}\tilde{\tilde{\mu}}(1,0)\,N_0}{n_0}\mathbf{1}\mathbf{1}' \\
&\quad + \frac{p_{01}\tilde{\tilde{\mu}}(0,1)\,N_1}{n_1}\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \frac{p_{00}\tilde{\tilde{\mu}}(0,0)\,N_0}{n_0}\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \\
&= \begin{bmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{bmatrix}
\end{aligned}
$$

where

$$
b_{11} = \frac{p_{11}N_1\tilde{\tilde{\mu}}(1,1) + p_{01}N_1\tilde{\tilde{\mu}}(0,1)}{n_1} + \frac{p_{10}N_0\tilde{\tilde{\mu}}(1,0) + p_{00}N_0\tilde{\tilde{\mu}}(0,0)}{n_0},
$$

$$
b_{12} = b_{22} = \frac{p_{11}N_1\tilde{\tilde{\mu}}(1,1)}{n_1} + \frac{p_{10}N_0\tilde{\tilde{\mu}}(1,0)}{n_0}.
$$

p_{11}, p_{10}, p_{01}, p_{00}, are the joint probability between the exposure variable W_i and primary outcome D_i. These joint probability can be estimate analytically once we specify $\boldsymbol{\gamma} = [\gamma_0, \gamma_1]$, the parameter between W_i and D_i.

$$
\begin{aligned}
p_{11} &= p(D=1|W=1) \times p_{w=1} = \tfrac{e^{\gamma_0+\gamma_1 w} p_{w=1}}{1+e^{\gamma_0+\gamma_1 w}}, \\
p_{10} &= p(D=0|W=1) \times p_{w=1} = \tfrac{p_{w=1}}{1+e^{\gamma_0+\gamma_1 w}}, \\
p_{01} &= p(D=1|W=0) \times (1-p_{w=1}) = \tfrac{e^{\gamma_0}(1-p_{w=1})}{1+e^{\gamma_0}}, \\
p_{00} &= p(D=0|W=0) \times (1-p_{w=1}) = \tfrac{(1-p_{w=1})}{1+e^{\gamma_0}}.
\end{aligned}
$$

□

We assume N_1 and N_0 are constant with $N_1 = NE(D)$ and $N_0 = N - N_1$. It is easy to see that $Var(\hat{\beta}_1)$ is the lower corner element of covariance matrix $\frac{\boldsymbol{A}(\boldsymbol{\beta})^{-1}\boldsymbol{B}(\boldsymbol{\beta})\left[\boldsymbol{A}(\boldsymbol{\beta})^{-1}\right]^T}{N}$. Combing the derivation of $A^{-1}B\left[A^{-1}\right]^T$ and the condition that $a_{12} = a_{21} = a_{22}$ and $b_{12} = b_{22}$, we have $Var(\hat{\beta}_1) =$

$\frac{\left[a_{21}^2 b_{11} - 2a_{11}b_{12}a_{21} + a_{11}^2 b_{22}\right]}{(detA)^2 N}$.

A.5 Proof of Proposition 1

Proof. We define

$$\zeta(1,1) = a_{21}^2 p_{11} N_1 \tilde{\tilde{\mu}}(1,1),$$
$$\zeta(1,0) = a_{21}^2 p_{10} N_0 \tilde{\tilde{\mu}}(1,0),$$
$$\iota(0,1) = a_{21}^2 p_{01} N_1 \tilde{\tilde{\mu}}(0,1),$$
$$\iota(0,0) = a_{21}^2 p_{00} N_0 \tilde{\tilde{\mu}}(0,0),$$
$$\kappa(1,1) = 2a_{21}a_{11} p_{11} N_1 \tilde{\tilde{\mu}}(1,1),$$
$$\kappa(1,0) = 2a_{21}a_{11} p_{10} N_0 \tilde{\tilde{\mu}}(1,0),$$
$$\varpi(1,1) = a_{11}^2 p_{11} N_1 \tilde{\tilde{\mu}}(1,1),$$
$$\varpi(1,0) = a_{11}^2 p_{10} N_0 \tilde{\tilde{\mu}}(1,0).$$

Then plug in a_{11}, a_{21}, a_{12}, a_{22}, b_{11}, b_{12}, b_{22} into $Var(\hat{\beta}_1)$, we have

$$Var(\hat{\beta_1}) = \frac{\left[\zeta(1,1) + \iota(0,1) - \kappa(1,1) + \varpi(1,1)\right] n_0 + \left[\zeta(1,0) + \iota(0,0) - \kappa(1,0) + \varpi(1,0)\right] n_1}{(detA)^2 n_0 n_1 N}.$$

To obtain the optimal design, we need to minimize $Var(\hat{\beta})$ subject to the constraint

$$Cost = c_{per}(n_1 + n_0),$$

where this constraint is equivalent to $n_0 + n_1 = n$. We can write,

$$R_{OD} \in \underset{n_0+n_1=n}{\operatorname{argmin}} \frac{T_1 n_0 + T_2 n_1}{(detA)^2 n_0 n_1 N}.$$

Where $T_1 = \zeta(1,1) + \iota(0,1) - \kappa(1,1) + \varpi(1,1)$ and $T_2 = \zeta(1,0) + \iota(0,0) - \kappa(1,0) + \varpi(1,0)$.

By using the Lagrange multiplier method, we have

$$\min\left[\frac{\left[\zeta(1,1)+\iota(0,1)-\kappa(1,1)+\varpi(1,1)\right]n_0+\left[\zeta(1,0)+\iota(0,0)-\kappa(1,0)+\varpi(1,0)\right]n_1}{(detA)^2 n_0 n_1 N}\right]$$
$$+\lambda(n_0+n_1-n)$$

Where λ is the Lagrange multiplier, let

$$L=\frac{\left[\zeta(1,1)+\iota(0,1)-\kappa(1,1)+\varpi(1,1)\right]n_0+\left[\zeta(1,0)+\iota(0,0)-\kappa(1,0)+\varpi(1,0)\right]n_1}{(detA)^2 n_0 n_1 N}$$
$$+\lambda(n_0+n_1-n)$$

then

$$\frac{\partial L}{\partial n_0}=-\left[\zeta(1,1)+\iota(0,1)-\kappa(1,1)+\varpi(1,1)\right]n_0^{-2}+\lambda,$$

$$\frac{\partial L}{\partial n_1}=-\left[\zeta(1,0)+\iota(0,0)-\kappa(1,0)+\varpi(1,0)\right]n_1^{-2}+\lambda.$$

Let the above two equations, equal to 0, we have

$n_0=\sqrt{\frac{[\zeta(1,1)+\iota(0,1)-\kappa(1,1)+\varpi(1,1)]}{\lambda}}$, $n_1=\sqrt{\frac{[\zeta(1,0)+\iota(0,0)-\kappa(1,0)+\varpi(1,0)]}{\lambda}}$.

Therefore, $R_{OD}=\frac{n_0}{n_1}=\sqrt{\frac{[\zeta(1,0)+\iota(0,0)-\kappa(1,0)+\varpi(1,0)]}{[\zeta(1,1)+\iota(0,1)-\kappa(1,1)+\varpi(1,1)]}}$. □

APPENDIX B PROOFS OF CHAPTER 3

B.1 Derivation of $\boldsymbol{B}(\boldsymbol{\beta})$

Proof. Similar to the derivation of $\boldsymbol{A}(\boldsymbol{\beta})$, we first plug $\frac{XS}{\pi(D)}[Y-\mu(\boldsymbol{X})]$ into $\boldsymbol{B}(\boldsymbol{\beta})$, then we have

$$\boldsymbol{B}(\boldsymbol{\beta}) = \mathbb{E}\left[\frac{S^2}{[\pi(D)]^2}(Y-\mu(\boldsymbol{X}))^2 \boldsymbol{X}\boldsymbol{X}^T\right].$$

The sampling indicator variable S is a binary variable, thus $S^2 = S$, then

$$\boldsymbol{B}(\boldsymbol{\beta}) = \mathbb{E}\left[\frac{S}{[\pi(D)]^2}(Y-\mu(\boldsymbol{X}))^2 \boldsymbol{X}\boldsymbol{X}^T\right].$$

Then take the iterated expectations,

$$\begin{aligned}
\boldsymbol{B}(\boldsymbol{\beta}) &= \mathbb{E}\left\{\mathbb{E}\left[\frac{S}{[\pi(D)]^2}(Y-\mu(\boldsymbol{X}))^2 \boldsymbol{X}\boldsymbol{X}^T\middle|\boldsymbol{X},D\right]\right\} \\
&= \mathbb{E}\left\{\frac{\mathbb{E}(S|D)}{[\pi(D)]^2}\mathbb{E}\left[(Y-\mu(\boldsymbol{X}))^2 \boldsymbol{X}\boldsymbol{X}^T\middle|\boldsymbol{X},D\right]\right\} \\
&= \mathbb{E}\left[\frac{1}{\pi(D)}\mathbb{E}\left[(Y-\mu(\boldsymbol{X}))^2\middle|\boldsymbol{X},D\right]\boldsymbol{X}\boldsymbol{X}^T\right].
\end{aligned}$$

We can simplify the expectation $\mathbb{E}\left[(Y-\mu(\boldsymbol{X}))^2\middle|\boldsymbol{X},D\right]$,

$$\begin{aligned}
\mathbb{E}\left[(Y-\mu(\boldsymbol{X}))^2\middle|\boldsymbol{X},D\right] &= \mathbb{E}\left[Y^2-2Y\mu(\boldsymbol{X})+\mu(\boldsymbol{X})^2\middle|\boldsymbol{X},D\right] \\
&= \mathbb{E}\left(Y^2\middle|\boldsymbol{X},D\right) - 2\mathbb{E}\left[Y\mu(\boldsymbol{X})|\boldsymbol{X},D\right] \\
&\quad + \mathbb{E}\left[\mu(\boldsymbol{X})^2|\boldsymbol{X},D\right] \\
&= \mathbb{E}\left(Y^2\middle|\boldsymbol{X},D\right) \\
&\quad - 2\mu(\boldsymbol{X})\mathbb{E}(Y|\boldsymbol{X},D) + \mu(\boldsymbol{X})^2
\end{aligned}$$

Under the above study design setting, Y is a Poisson distributed secondary outcome, so let

$$\begin{aligned}\mathbb{E}\left(Y^2\middle|\boldsymbol{X},D\right) &= Var\left(Y|\boldsymbol{X},D\right)+\left(\mathbb{E}\left(Y|\boldsymbol{X},D\right)\right)^2\\ &= \mathbb{E}\left(Y|\boldsymbol{X},D\right)+\left(\mathbb{E}\left(Y|\boldsymbol{X},D\right)\right)^2\end{aligned}$$

We define $\tilde{\mu}\left(\boldsymbol{X},D\right)=\mathbb{E}(Y|\boldsymbol{X},D)$.

We plug $\mathbb{E}\left(Y^2\middle|\boldsymbol{X},D\right)$ into $\mathbb{E}\left[\left(Y-\mu\left(\boldsymbol{X}\right)\right)^2\middle|\boldsymbol{X},D\right]$, we have

$$\begin{aligned}\mathbb{E}\left[\left(Y-\mu\left(\boldsymbol{X}\right)\right)^2\middle|\boldsymbol{X},D\right] &= \mathbb{E}\left(Y^2\middle|\boldsymbol{X},D\right)-2\mu\left(\boldsymbol{X}\right)\mathbb{E}\left(Y|\boldsymbol{X},D\right)+\mu\left(\boldsymbol{X}\right)^2\\ &= \mathbb{E}\left(Y|\boldsymbol{X},D\right)+\left(\mathbb{E}\left(Y|\boldsymbol{X},D\right)\right)^2\\ &\quad -2\mu\left(\boldsymbol{X}\right)\mathbb{E}\left(Y|\boldsymbol{X},D\right)+\mu\left(\boldsymbol{X}\right)^2\\ &= \tilde{\mu}\left(\boldsymbol{X},D\right)+\tilde{\mu}\left(\boldsymbol{X},D\right)^2-2\mu\left(\boldsymbol{X}\right)\tilde{\mu}\left(\boldsymbol{X},D\right)+\mu\left(\boldsymbol{X}\right)^2\\ &= \left[\tilde{\mu}\left(\boldsymbol{X},D\right)-\mu\left(\boldsymbol{X}\right)\right]^2+\tilde{\mu}\left(\boldsymbol{X},D\right).\end{aligned}$$

Then we have

$$\begin{aligned}\boldsymbol{B}\left(\boldsymbol{\beta}\right) &= \mathbb{E}\left[\frac{1}{\pi\left(D\right)}\mathbb{E}\left\{\left[\left[\tilde{\mu}\left(\boldsymbol{X},D\right)-\mu\left(\boldsymbol{X}\right)\right]^2+\tilde{\mu}\left(\boldsymbol{X},D\right)\right]\middle|\boldsymbol{X},D\right\}\boldsymbol{X}\boldsymbol{X}^T\right]\\ &= \mathbb{E}\left[\frac{1}{\pi\left(D\right)}\left[\left(\tilde{\mu}\left(\boldsymbol{X},D\right)-\mu\left(\boldsymbol{X}\right)\right)^2+\tilde{\mu}\left(\boldsymbol{X},D\right)\right]\boldsymbol{X}\boldsymbol{X}^T\right].\end{aligned}$$

Thus,

$$\boldsymbol{B}\left(\boldsymbol{\beta}\right)=\mathbb{E}\left[\frac{\left(\tilde{\mu}\left(\boldsymbol{X},D\right)-\mu\left(\boldsymbol{X}\right)\right)^2+\tilde{\mu}\left(\boldsymbol{X},D\right)}{\pi\left(D\right)}\boldsymbol{X}\boldsymbol{X}^T\right].$$

□

APPENDIX C PROOF OF CHAPTER 4

C.1 Derivation of $A(\boldsymbol{\beta})$

In the main text, we defined W as the binary exposure variable, C as the binary confounder, and let $\boldsymbol{X} = (1, W, C)^T$. $\frac{1}{P(W=w|c)}$ is the inverse probability weight for the subjects in the group w for $w = 0, 1$. Plugging the double weighted expression $\frac{XS}{\pi(D)\times P(W|c)}[Y - \mu(\boldsymbol{X})]$ into $A(\boldsymbol{\beta})$, we have

$$\begin{aligned} A(\boldsymbol{\beta}) &= \mathbb{E}\left[-\frac{\partial}{\partial \boldsymbol{\beta}^T}\frac{\boldsymbol{X}^T S}{\pi(D)\times P(W|c)}\left(Y - \frac{e^{\boldsymbol{X}^T\boldsymbol{\beta}}}{1+e^{\boldsymbol{X}^T\boldsymbol{\beta}}}\right)\right] \\ &= \mathbb{E}\left[\frac{S}{\pi(D)\times P(W|c)}\times\frac{e^{\boldsymbol{X}^T\boldsymbol{\beta}}}{\left(1+e^{\boldsymbol{X}^T\boldsymbol{\beta}}\right)^2}\boldsymbol{X}\boldsymbol{X}^T\right]. \end{aligned}$$

Taking iterated expectations on $A(\boldsymbol{\beta})$, then

$$A(\boldsymbol{\beta}) = \mathbb{E}\left\{\mathbb{E}\left[\frac{S}{\pi(D)\times P(W|c)}\times\frac{e^{\boldsymbol{X}^T\boldsymbol{\beta}}}{\left(1+e^{\boldsymbol{X}^T\boldsymbol{\beta}}\right)^2}\boldsymbol{X}\boldsymbol{X}^T\middle| D\right]\right\}.$$

Note that $\pi(D) = \Pr(S = 1 \mid D)$ only depends on D, thus

$$A(\boldsymbol{\beta}) = \mathbb{E}\left\{\frac{\mathbb{E}(S|D)}{\pi(D)}\times\mathbb{E}\left[\frac{e^{\boldsymbol{X}^T\boldsymbol{\beta}}}{\pi(D)\times P(W|c)\times\left(1+e^{\boldsymbol{X}^T\boldsymbol{\beta}}\right)^2}\boldsymbol{X}\boldsymbol{X}^T\middle| D\right]\right\}.$$

Since S is a binary variable, then $\mathbb{E}(S|D) = \Pr(S = 1 \mid D) = \pi(D)$, thus

$$\begin{aligned} A(\boldsymbol{\beta}) &= \mathbb{E}\left\{\mathbb{E}\left[\frac{e^{\boldsymbol{X}^T\boldsymbol{\beta}}}{\pi(D)\times P(W|c)\left(1+e^{\boldsymbol{X}^T\boldsymbol{\beta}}\right)^2}\boldsymbol{X}\boldsymbol{X}^T\middle| D\right]\right\} \\ &= \mathbb{E}\left[\frac{e^{\boldsymbol{X}^T\boldsymbol{\beta}}}{\pi(D)\times P(W|c)\times\left(1+e^{\boldsymbol{X}^T\boldsymbol{\beta}}\right)^2}\boldsymbol{X}\boldsymbol{X}^T\right]. \end{aligned}$$

C.2 Derivation of $\boldsymbol{B}(\boldsymbol{\beta})$

Similar to the derivation of $\boldsymbol{A}(\boldsymbol{\beta})$, we first plugged the double weighted expression

$$\frac{\boldsymbol{X}S}{\pi(D)\times P(W|c)}\left[Y-\mu(\boldsymbol{X})\right]$$

into $\boldsymbol{B}(\boldsymbol{\beta})$, then we had

$$\boldsymbol{B}(\boldsymbol{\beta})=\mathbb{E}\left[\frac{S^2}{\left[P(W|c)\right]^2\left[\pi(D)\right]^2}\left(Y-\mu(\boldsymbol{X})\right)^2\boldsymbol{X}\boldsymbol{X}^T\right].$$

The sampling indicator variable S is a binary variable, thus $S^2=S$, then

$$\boldsymbol{B}(\boldsymbol{\beta})=\mathbb{E}\left[\frac{S}{\left[P(W|c)\right]^2\left[\pi(D)\right]^2}\left(Y-\mu(\boldsymbol{X})\right)^2\boldsymbol{X}\boldsymbol{X}^T\right].$$

Then take the iterated expectations,

$$\begin{aligned}\boldsymbol{B}(\boldsymbol{\beta})&=\mathbb{E}\left\{\mathbb{E}\left[\frac{S}{\left[P(W|c)\right]^2\left[\pi(D)\right]^2}\left(Y-\mu(\boldsymbol{X})\right)^2\boldsymbol{X}\boldsymbol{X}^T\middle|\boldsymbol{X},D\right]\right\}\\&=\mathbb{E}\left\{\frac{\mathbb{E}(S|D)}{\left[\pi(D)\right]^2\left[P(W|c)\right]^2}\mathbb{E}\left[\left(Y-\mu(\boldsymbol{X})\right)^2\boldsymbol{X}\boldsymbol{X}^T\middle|\boldsymbol{X},D\right]\right\}\\&=\mathbb{E}\left[\frac{1}{\pi(D)\left[P(W|c)\right]^2}\mathbb{E}\left[\left(Y-\mu(\boldsymbol{X})\right)^2\middle|\boldsymbol{X},D\right]\boldsymbol{X}\boldsymbol{X}^T\right].\end{aligned}$$

We can simplify the expectation $\mathbb{E}\left[\left(Y-\mu(\boldsymbol{X})\right)^2\middle|\boldsymbol{X},D\right]$,

$$\begin{aligned}\mathbb{E}\left[\left(Y-\mu(\boldsymbol{X})\right)^2\middle|\boldsymbol{X},D\right]&=\mathbb{E}\left[Y^2-2Y\mu(\boldsymbol{X})+\mu(\boldsymbol{X})^2\middle|\boldsymbol{X},D\right]\\&=\mathbb{E}\left(Y^2\middle|\boldsymbol{X},D\right)-2\mathbb{E}\left[Y\mu(\boldsymbol{X})|\boldsymbol{X},D\right]\\&\quad+\mathbb{E}\left[\mu(\boldsymbol{X})^2|\boldsymbol{X},D\right]\\&=\mathbb{E}\left(Y^2\middle|\boldsymbol{X},D\right)\\&\quad-2\mu(\boldsymbol{X})\mathbb{E}(Y|\boldsymbol{X},D)+\mu(\boldsymbol{X})^2.\end{aligned}$$

Under the above study design setting, Y is a binary secondary outcome.

So let $\tilde{\mu}(\boldsymbol{X}, D) \equiv \mathbb{E}\left(Y^2 \middle| \boldsymbol{X}, D\right) = \mathbb{E}(Y|\boldsymbol{X}, D)$. Then we have

$$\begin{aligned}
\mathbb{E}\left[(Y-\mu(\boldsymbol{X}))^2 \middle| \boldsymbol{X}, D\right] &= \tilde{\mu}(\boldsymbol{X}, D) - 2\mu(\boldsymbol{X})\tilde{\mu}(\boldsymbol{X}, D) + \mu(\boldsymbol{X})^2 \\
&= \tilde{\mu}(\boldsymbol{X}, D)^2 - 2\mu(\boldsymbol{X})\tilde{\mu}(\boldsymbol{X}, D) + \mu(\boldsymbol{X})^2 \\
&\quad + \tilde{\mu}(\boldsymbol{X}, D) - \tilde{\mu}(\boldsymbol{X}, D)^2 \\
&= [\tilde{\mu}(\boldsymbol{X}, D) - \mu(\boldsymbol{X})]^2 + \tilde{\mu}(\boldsymbol{X}, D) - \tilde{\mu}(\boldsymbol{X}, D)^2 .
\end{aligned}$$

So,

$$\boldsymbol{B}(\boldsymbol{\beta}) = \mathbb{E}\left[\frac{\left[(\tilde{\mu}(\boldsymbol{X}, D) - \mu(\mathbf{X}))^2 + \tilde{\mu}(\boldsymbol{X}, D) - \tilde{\mu}(\boldsymbol{X}, D)^2\right]}{[P(W|c)]^2 \pi(D)} \boldsymbol{X}\boldsymbol{X}^T\right].$$

Denote $\tilde{\tilde{\mu}}(\boldsymbol{X}, D) \equiv (\tilde{\mu}(\boldsymbol{X}, D) - \mu(\boldsymbol{X}))^2 + \tilde{\mu}(\boldsymbol{X}, D) - \tilde{\mu}(\boldsymbol{X}, D)^2$,

$$\boldsymbol{B}(\boldsymbol{\beta}) = \mathbb{E}\left[\frac{\tilde{\tilde{\mu}}(\boldsymbol{X}, D)}{[P(W|c)]^2 \pi(D)} \boldsymbol{X}\boldsymbol{X}^T\right].$$

C.3 Derivation of $A^{-1}B\left[A^{-1}\right]^T$

We denote $p(w, c)$ as the joint probability between W and C. By the rule of function expectation,

we have

$$
\begin{aligned}
A(\boldsymbol{\beta}) &= \mathbb{E}\left[\frac{e^{\boldsymbol{X}^T\boldsymbol{\beta}}}{[P(W|c)]\left(1+e^{\boldsymbol{X}^T\boldsymbol{\beta}}\right)^2}\boldsymbol{X}\boldsymbol{X}^T\right] \\
&= \mathbb{E}\left[\frac{e^{\boldsymbol{X}^T\boldsymbol{\beta}}}{[P(W|c)]\left(1+e^{\boldsymbol{X}^T\boldsymbol{\beta}}\right)^2}\begin{bmatrix} 1 & w & c \\ w & w & wc \\ c & wc & c \end{bmatrix}\right] \\
&= \frac{p(w=1,c=1)e^{\beta_0+\beta_1+\beta_2}}{P(w=1|c=1)\left(1+e^{\beta_0+\beta_1+\beta_2}\right)^2}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \\
&\quad + \frac{p(w=1,c=0)e^{\beta_0+\beta_1}}{P(w=1|c=0)\left(1+e^{\beta_0+\beta_1}\right)^2}\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
&\quad + \frac{p(w=0,c=1)e^{\beta_0+\beta_2}}{P(w=0|c=1)\left(1+e^{\beta_0+\beta_2}\right)^2}\begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix}
\end{aligned}
$$

$$+\frac{p(w=0,c=0)e^{\beta_0}}{P(w=0|c=0)\left(1+e^{\beta_0}\right)^2}\begin{bmatrix}1&0&0\\0&0&0\\0&0&0\end{bmatrix}$$

$$=\frac{p(c=1)e^{\beta_0+\beta_1+\beta_2}}{\left(1+e^{\beta_0+\beta_1+\beta_2}\right)^2}\begin{bmatrix}1&1&1\\1&1&1\\1&1&1\end{bmatrix}+\frac{p(c=0)e^{\beta_0+\beta_1}}{\left(1+e^{\beta_0+\beta_1}\right)^2}\begin{bmatrix}1&1&0\\1&1&0\\0&0&0\end{bmatrix}$$

$$+\frac{p(c=1)e^{\beta_0+\beta_2}}{\left(1+e^{\beta_0+\beta_2}\right)^2}\begin{bmatrix}1&0&1\\0&0&0\\1&0&1\end{bmatrix}+\frac{p(c=0)e^{\beta_0}}{\left(1+e^{\beta_0}\right)^2}\begin{bmatrix}1&0&0\\0&0&0\\0&0&0\end{bmatrix}$$

$$
\begin{aligned}
&= \begin{bmatrix}
\frac{p(c=1)e^{\beta_0+\beta_1+\beta_2}}{\left(1+e^{\beta_0+\beta_1+\beta_2}\right)^2} & \frac{p(c=1)e^{\beta_0+\beta_1+\beta_2}}{\left(1+e^{\beta_0+\beta_1+\beta_2}\right)^2} & \frac{p(c=1)e^{\beta_0+\beta_1+\beta_2}}{\left(1+e^{\beta_0+\beta_1+\beta_2}\right)^2} \\
\frac{p(c=1)e^{\beta_0+\beta_1+\beta_2}}{\left(1+e^{\beta_0+\beta_1+\beta_2}\right)^2} & \frac{p(c=1)e^{\beta_0+\beta_1+\beta_2}}{\left(1+e^{\beta_0+\beta_1+\beta_2}\right)^2} & \frac{p(c=1)e^{\beta_0+\beta_1+\beta_2}}{\left(1+e^{\beta_0+\beta_1+\beta_2}\right)^2} \\
\frac{p(c=1)e^{\beta_0+\beta_1+\beta_2}}{\left(1+e^{\beta_0+\beta_1+\beta_2}\right)^2} & \frac{p(c=1)e^{\beta_0+\beta_1+\beta_2}}{\left(1+e^{\beta_0+\beta_1+\beta_2}\right)^2} & \frac{p(c=1)e^{\beta_0+\beta_1+\beta_2}}{\left(1+e^{\beta_0+\beta_1+\beta_2}\right)^2}
\end{bmatrix} + \begin{bmatrix}
\frac{p(c=0)e^{\beta_0+\beta_1}}{\left(1+e^{\beta_0+\beta_1}\right)^2} & \frac{p(c=0)e^{\beta_0+\beta_1}}{\left(1+e^{\beta_0+\beta_1}\right)^2} & 0 \\
\frac{p(c=0)e^{\beta_0+\beta_1}}{\left(1+e^{\beta_0+\beta_1}\right)^2} & \frac{p(c=0)e^{\beta_0+\beta_1}}{\left(1+e^{\beta_0+\beta_1}\right)^2} & 0 \\
0 & 0 & 0
\end{bmatrix} \\
&\quad + \begin{bmatrix}
\frac{p(c=1)e^{\beta_0+\beta_2}}{\left(1+e^{\beta_0+\beta_2}\right)^2} & 0 & \frac{p(c=1)e^{\beta_0+\beta_2}}{\left(1+e^{\beta_0+\beta_2}\right)^2} \\
0 & 0 & 0 \\
\frac{p(c=1)e^{\beta_0+\beta_2}}{\left(1+e^{\beta_0+\beta_2}\right)^2} & 0 & \frac{p(c=1)e^{\beta_0+\beta_2}}{\left(1+e^{\beta_0+\beta_2}\right)^2}
\end{bmatrix} + \begin{bmatrix}
\frac{p(c=0)e^{\beta_0}}{\left(1+e^{\beta_0}\right)^2} & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{bmatrix} \\
&= \begin{bmatrix}
a & b & c \\
b & b & d \\
c & d & c
\end{bmatrix}.
\end{aligned}
$$

Where

$$
\begin{aligned}
a &= \frac{p(c=1)e^{\beta_0+\beta_1+\beta_2}}{\left(1+e^{\beta_0+\beta_1+\beta_2}\right)^2} + \frac{p(c=0)e^{\beta_0+\beta_1}}{\left(1+e^{\beta_0+\beta_1}\right)^2} + \frac{p(c=1)e^{\beta_0+\beta_2}}{\left(1+e^{\beta_0+\beta_2}\right)^2} + \frac{p(c=0)e^{\beta_0}}{\left(1+e^{\beta_0}\right)^2}, \\
b &= \frac{p(c=1)e^{\beta_0+\beta_1+\beta_2}}{\left(1+e^{\beta_0+\beta_1+\beta_2}\right)^2} + \frac{p(c=0)e^{\beta_0+\beta_1}}{\left(1+e^{\beta_0+\beta_1}\right)^2}, \\
c &= \frac{p(c=1)e^{\beta_0+\beta_1+\beta_2}}{\left(1+e^{\beta_0+\beta_1+\beta_2}\right)^2} + \frac{p(c=1)e^{\beta_0+\beta_2}}{\left(1+e^{\beta_0+\beta_2}\right)^2}, \\
d &= \frac{p(c=1)e^{\beta_0+\beta_1+\beta_2}}{\left(1+e^{\beta_0+\beta_1+\beta_2}\right)^2}.
\end{aligned}
$$

We define some simple notations. We define, p_{wcd} as the joint probability of W, C, and D. We

define $p^{wc} = p(w|c)$. Using the rule of function of expectation on $B(\boldsymbol{\beta})$, we have,

$$
\begin{aligned}
B(\boldsymbol{\beta}) &= \mathbb{E}\left[\frac{\tilde{\tilde{\mu}}(W,C,D)}{[P(W|c)]^2 \pi(D)} \boldsymbol{X}\boldsymbol{X}^T\right] \\
&= \mathbb{E}\left[\frac{\tilde{\tilde{\mu}}(W,C,D)}{[P(W|c)]^2 \pi(D)} \begin{bmatrix} 1 & w & c \\ w & w & wc \\ c & wc & c \end{bmatrix}\right] \\
&= \frac{p_{111}\tilde{\tilde{\mu}}(1,1,1)}{\left(p^{11}\right)^2 \pi(1)} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} + \frac{p_{110}\tilde{\tilde{\mu}}(1,1,0)}{\left(p^{11}\right)^2 \pi(0)} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \\
&\quad + \frac{p_{101}\tilde{\tilde{\mu}}(1,0,1)}{\left(p^{10}\right)^2 \pi(1)} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \frac{p_{100}\tilde{\tilde{\mu}}(1,0,0)}{\left(p^{10}\right)^2 \pi(0)} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
&\quad + \frac{p_{011}\tilde{\tilde{\mu}}(0,1,1)}{\left(p^{01}\right)^2 \pi(1)} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix} + \frac{p_{010}\tilde{\tilde{\mu}}(0,1,0)}{\left(p^{01}\right)^2 \pi(0)} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix} \\
&\quad + \frac{p_{001}\tilde{\tilde{\mu}}(0,0,1)}{\left(p^{00}\right)^2 \pi(1)} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \frac{p_{000}\tilde{\tilde{\mu}}(0,0,0)}{\left(p^{00}\right)^2 \pi(0)} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}
\end{aligned}
$$

$$
= \begin{bmatrix}
\frac{p_{111}\tilde{\tilde{\mu}}(1,1,1)}{(p^{11})^2\pi(1)} & \frac{p_{111}\tilde{\tilde{\mu}}(1,1,1)}{(p^{11})^2\pi(1)} & \frac{p_{111}\tilde{\tilde{\mu}}(1,1,1)}{(p^{11})^2\pi(1)} \\
\frac{p_{111}\tilde{\tilde{\mu}}(1,1,1)}{(p^{11})^2\pi(1)} & \frac{p_{111}\tilde{\tilde{\mu}}(1,1,1)}{(p^{11})^2\pi(1)} & \frac{p_{111}\tilde{\tilde{\mu}}(1,1,1)}{(p^{11})^2\pi(1)} \\
\frac{p_{111}\tilde{\tilde{\mu}}(1,1,1)}{(p^{11})^2\pi(1)} & \frac{p_{111}\tilde{\tilde{\mu}}(1,1,1)}{(p^{11})^2\pi(1)} & \frac{p_{111}\tilde{\tilde{\mu}}(1,1,1)}{(p^{11})^2\pi(1)}
\end{bmatrix} + \begin{bmatrix}
\frac{p_{110}\tilde{\tilde{\mu}}(1,1,0)}{(p^{11})^2\pi(0)} & \frac{p_{110}\tilde{\tilde{\mu}}(1,1,0)}{(p^{11})^2\pi(0)} & \frac{p_{110}\tilde{\tilde{\mu}}(1,1,0)}{(p^{11})^2\pi(0)} \\
\frac{p_{110}\tilde{\tilde{\mu}}(1,1,0)}{(p^{11})^2\pi(0)} & \frac{p_{110}\tilde{\tilde{\mu}}(1,1,0)}{(p^{11})^2\pi(0)} & \frac{p_{110}\tilde{\tilde{\mu}}(1,1,0)}{(p^{11})^2\pi(0)} \\
\frac{p_{110}\tilde{\tilde{\mu}}(1,1,0)}{(p^{11})^2\pi(0)} & \frac{p_{110}\tilde{\tilde{\mu}}(1,1,0)}{(p^{11})^2\pi(0)} & \frac{p_{110}\tilde{\tilde{\mu}}(1,1,0)}{(p^{11})^2\pi(0)}
\end{bmatrix}
$$

$$
+ \begin{bmatrix}
\frac{p_{101}\tilde{\tilde{\mu}}(1,0,1)}{(p^{10})^2\pi(1)} & \frac{p_{101}\tilde{\tilde{\mu}}(1,0,1)}{(p^{10})^2\pi(1)} & 0 \\
\frac{p_{101}\tilde{\tilde{\mu}}(1,0,1)}{(p^{10})^2\pi(1)} & \frac{p_{101}\tilde{\tilde{\mu}}(1,0,1)}{(p^{10})^2\pi(1)} & 0 \\
0 & 0 & 0
\end{bmatrix} + \begin{bmatrix}
\frac{p_{100}\tilde{\tilde{\mu}}(1,0,0)}{(p^{10})^2\pi(0)} & \frac{p_{100}\tilde{\tilde{\mu}}(1,0,0)}{(p^{10})^2\pi(0)} & 0 \\
\frac{p_{100}\tilde{\tilde{\mu}}(1,0,0)}{(p^{10})^2\pi(0)} & \frac{p_{100}\tilde{\tilde{\mu}}(1,0,0)}{(p^{10})^2\pi(0)} & 0 \\
0 & 0 & 0
\end{bmatrix}
$$

$$
+ \begin{bmatrix}
\frac{p_{011}\tilde{\tilde{\mu}}(0,1,1)}{(p^{01})^2\pi(1)} & 0 & \frac{p_{011}\tilde{\tilde{\mu}}(0,1,1)}{(p^{01})^2\pi(1)} \\
0 & 0 & 0 \\
\frac{p_{011}\tilde{\tilde{\mu}}(0,1,1)}{(p^{01})^2\pi(1)} & 0 & \frac{p_{011}\tilde{\tilde{\mu}}(0,1,1)}{(p^{01})^2\pi(1)}
\end{bmatrix} + \begin{bmatrix}
\frac{p_{010}\tilde{\tilde{\mu}}(0,1,0)}{(p^{01})^2\pi(0)} & 0 & \frac{p_{010}\tilde{\tilde{\mu}}(0,1,0)}{(p^{01})^2\pi(0)} \\
0 & 0 & 0 \\
\frac{p_{010}\tilde{\tilde{\mu}}(0,1,0)}{(p^{01})^2\pi(0)} & 0 & \frac{p_{010}\tilde{\tilde{\mu}}(0,1,0)}{(p^{01})^2\pi(0)}
\end{bmatrix}
$$

$$
+ \begin{bmatrix}
\frac{p_{001}\tilde{\tilde{\mu}}(0,0,1)}{(p^{00})^2\pi(1)} & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{bmatrix} + \begin{bmatrix}
\frac{p_{000}\tilde{\tilde{\mu}}(0,0,0)}{(p^{00})^2\pi(0)} & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{bmatrix}
$$

$$
= \begin{bmatrix}
f & g & h \\
g & i & j \\
h & j & k
\end{bmatrix}
$$

$$f = \frac{p_{111}\tilde{\tilde{\mu}}(1,1,1)}{(p^{11})^2\pi(1)} + \frac{p_{110}\tilde{\tilde{\mu}}(1,1,0)}{(p^{11})^2\pi(0)} + \frac{p_{101}\tilde{\tilde{\mu}}(1,0,1)}{(p^{10})^2\pi(1)} + \frac{p_{100}\tilde{\tilde{\mu}}(1,0,0)}{(p^{10})^2\pi(0)} + \frac{p_{011}\tilde{\tilde{\mu}}(0,1,1)}{(p^{01})^2\pi(1)} + \frac{p_{010}\tilde{\tilde{\mu}}(0,1,0)}{(p^{01})^2\pi(0)} + \frac{p_{00}\tilde{\tilde{\mu}}(0,0,1)}{(p^{00})^2\pi(1)} + \frac{p_{000}\tilde{\tilde{\mu}}(0,0,0)}{(p^{00})^2\pi(0)}$$

$$g = \frac{p_{111}\tilde{\tilde{\mu}}(1,1,1)}{(p^{11})^2\pi(1)} + \frac{p_{110}\tilde{\tilde{\mu}}(1,1,0)}{(p^{11})^2\pi(0)} + \frac{p_{101}\tilde{\tilde{\mu}}(1,0,1)}{(p^{10})^2\pi(1)} + \frac{p_{100}\tilde{\tilde{\mu}}(1,0,0)}{(p^{10})^2\pi(0)}$$

$$h = \frac{p_{111}\tilde{\tilde{\mu}}(1,1,1)}{(p^{11})^2\pi(1)} + \frac{p_{110}\tilde{\tilde{\mu}}(1,1,0)}{(p^{11})^2\pi(0)} + \frac{p_{011}\tilde{\tilde{\mu}}(0,1,1)}{(p^{01})^2\pi(1)} + \frac{p_{010}\tilde{\tilde{\mu}}(0,1,0)}{(p^{01})^2\pi(0)}$$

$$i = \frac{p_{111}\tilde{\tilde{\mu}}(1,1,1)}{(p^{11})^2\pi(1)} + \frac{p_{110}\tilde{\tilde{\mu}}(1,1,0)}{(p^{11})^2\pi(0)} + \frac{p_{101}\tilde{\tilde{\mu}}(1,0,1)}{(p^{10})^2\pi(1)} + \frac{p_{100}\tilde{\tilde{\mu}}(1,0,0)}{(p^{10})^2\pi(0)}$$

$$j = \frac{p_{111}\tilde{\tilde{\mu}}(1,1,1)}{(p^{11})^2\pi(1)} + \frac{p_{110}\tilde{\tilde{\mu}}(1,1,0)}{(p^{11})^2\pi(0)}$$

$$k = \frac{p_{111}\tilde{\tilde{\mu}}(1,1,1)}{(p^{11})^2\pi(1)} + \frac{p_{110}\tilde{\tilde{\mu}}(1,1,0)}{(p^{11})^2\pi(0)} + \frac{p_{011}\tilde{\tilde{\mu}}(0,1,1)}{(p^{01})^2\pi(1)} + \frac{p_{010}\tilde{\tilde{\mu}}(0,1,0)}{(p^{01})^2\pi(0)}$$

We can see that $i = g$ and $h = k$.

Thus, we can write $A^{-1}B\left[A^{-1}\right]^T = \begin{bmatrix} a & b & c \\ b & b & d \\ c & d & c \end{bmatrix}^{-1} \begin{bmatrix} f & g & h \\ g & g & j \\ h & j & h \end{bmatrix} \left[\begin{bmatrix} a & b & c \\ b & b & d \\ c & d & c \end{bmatrix}^{-1}\right]^T$. Our target is to get the second element of the diagonal of $A^{-1}B\left[A^{-1}\right]^T$. Then, we can divide this element by N to obtain the variance of the estimator of the exposure effect. And we minimize this variance to get the optimal sampling ratio.

To compute $\begin{bmatrix} a & b & c \\ b & b & d \\ c & d & c \end{bmatrix}^{-1}$, we need to get $detA$ and $adj(A)$. See details below.

The minor of matrix A is,

$$\begin{bmatrix} bc-d^2 & bc-cd & bd-bc \\ bc-cd & ac-c^2 & ad-bc \\ bd-bc & ad-bc & ab-b^2 \end{bmatrix}.$$

The cofactor matrix of A is the same as the cofactor matrix, because it is symmetric.

By transposing the cofactor matrix, we obtain the adjoint matrix of A,

$$\begin{bmatrix} bc-d^2 & cd-bc & bd-bc \\ cd-bc & ac-c^2 & bc-ad \\ bd-bc & bc-ad & ab-b^2 \end{bmatrix}.$$

The determinant of A is,

$$detA = abc + 2bcd - ad^2 - b^2c - c^2b.$$

Thus,

$$\begin{bmatrix} a & b & c \\ b & b & d \\ c & d & c \end{bmatrix}^{-1} = \frac{1}{detA} adjA$$

$$= \frac{1}{abc+2bcd-ad^2-b^2c-c^2b} \begin{bmatrix} bc-d^2 & cd-bc & bd-bc \\ cd-bc & ac-c^2 & bc-ad \\ bd-bc & bc-ad & ab-b^2 \end{bmatrix}$$

$$= \begin{bmatrix} E & F & G \\ F & H & I \\ G & I & J \end{bmatrix},$$

where,

$$E = \frac{bc-d^2}{abc+2bcd-ad^2-b^2c-c^2b},$$

$$F = \frac{cd-bc}{abc+2bcd-ad^2-b^2c-c^2b},$$

$$G = \frac{bd-bc}{abc+2bcd-ad^2-b^2c-c^2b},$$

$$H = \frac{ac-c^2}{abc+2bcd-ad^2-b^2c-c^2b},$$

$$I = \frac{bc-ad}{abc+2bcd-ad^2-b^2c-c^2b},$$

$$J = \frac{ab-b^2}{abc+2bcd-ad^2-b^2c-c^2b}.$$

Further,

$$A^{-1}B = \begin{bmatrix} E & F & G \\ F & H & I \\ G & I & J \end{bmatrix} \times \begin{bmatrix} f & g & h \\ g & g & j \\ h & j & h \end{bmatrix}.$$

So by simple computation, the second row of $A^{-1}B$ is

$$\begin{bmatrix} Ff+Hg+Ih, & Fg+Hg+Ij, & Fh+Hj+Ih \end{bmatrix}.$$

Since $\left[A^{-1}\right]^T = \begin{bmatrix} E & F & G \\ F & H & I \\ G & I & J \end{bmatrix}$,

thus, the second element in the diagonal of $A^{-1}B\left[A^{-1}\right]^T$ is

$$F^2f+2FHg+2FIh+H^2g+2HIj+I^2h.$$

So,

$$var(\hat{\beta}_1) = \frac{F^2f+2FHg+2FIh+H^2g+2HIj+I^2h}{N},$$

where

$$
\begin{aligned}
F^2 f &= \frac{\frac{F^2}{(p^{11})^2} p_{111} \tilde{\tilde{\mu}}(1,1,1)}{\pi(1)} + \frac{\frac{F^2}{(p^{11})^2} p_{110} \tilde{\tilde{\mu}}(1,1,0)}{\pi(0)} \\
&\quad + \frac{\frac{F^2}{(p^{10})^2} p_{101} \tilde{\tilde{\mu}}(1,0,1)}{\pi(1)} + \frac{\frac{F^2}{(p^{10})^2} p_{100} \tilde{\tilde{\mu}}(1,0,0)}{\pi(0)} \\
&\quad + \frac{\frac{F^2}{(p^{01})^2} p_{011} \tilde{\tilde{\mu}}(0,1,1)}{\pi(1)} + \frac{\frac{F^2}{(p^{01})^2} p_{010} \tilde{\tilde{\mu}}(0,1,0)}{\pi(0)} \\
&\quad + \frac{\frac{F^2}{(p^{00})^2} p_{001} \tilde{\tilde{\mu}}(0,0,1)}{\pi(1)} + \frac{\frac{F^2}{(p^{00})^2} p_{000} \tilde{\tilde{\mu}}(0,0,0)}{\pi(0)} \\
&= \frac{\frac{F^2}{(p^{11})^2} p_{111} \tilde{\tilde{\mu}}(1,1,1) N_1}{n_1} + \frac{\frac{F^2}{(p^{11})^2} p_{110} \tilde{\tilde{\mu}}(1,1,0) N_0}{n_0} \\
&\quad + \frac{\frac{F^2}{(p^{10})^2} p_{101} \tilde{\tilde{\mu}}(1,0,1) N_1}{n_1} + \frac{\frac{F^2}{(p^{10})^2} p_{100} \tilde{\tilde{\mu}}(1,0,0) N_0}{n_0} \\
&\quad + \frac{\frac{F^2}{(p^{01})^2} p_{011} \tilde{\tilde{\mu}}(0,1,1) N_1}{n_1} + \frac{\frac{F^2}{(p^{01})^2} p_{010} \tilde{\tilde{\mu}}(0,1,0) N_0}{n_0} \\
&\quad + \frac{\frac{F^2}{(p^{00})^2} p_{001} \tilde{\tilde{\mu}}(0,0,1) N_1}{n_1} + \frac{\frac{F^2}{(p^{00})^2} p_{000} \tilde{\tilde{\mu}}(0,0,0) N_0}{n_0} \\
&= \frac{\frac{F^2}{(p^{11})^2} p_{111} \tilde{\tilde{\mu}}(1,1,1) N_1 n_0}{n_1 n_0} + \frac{\frac{F^2}{(p^{11})^2} p_{110} \tilde{\tilde{\mu}}(1,1,0) N_0 n_1}{n_1 n_0} \\
&\quad + \frac{\frac{F^2}{(p^{10})^2} p_{101} \tilde{\tilde{\mu}}(1,0,1) N_1 n_0}{n_1 n_0} + \frac{\frac{F^2}{(p^{10})^2} p_{100} \tilde{\tilde{\mu}}(1,0,0) N_0 n_1}{n_1 n_0} \\
&\quad + \frac{\frac{F^2}{(p^{01})^2} p_{011} \tilde{\tilde{\mu}}(0,1,1) N_1 n_0}{n_1 n_0} + \frac{\frac{F^2}{(p^{01})^2} p_{010} \tilde{\tilde{\mu}}(0,1,0) N_0 n_1}{n_1 n_0} \\
&\quad + \frac{\frac{F^2}{(p^{00})^2} p_{001} \tilde{\tilde{\mu}}(0,0,1) N_1 n_0}{n_1 n_0} + \frac{\frac{F^2}{(p^{00})^2} p_{000} \tilde{\tilde{\mu}}(0,0,0) N_0 n_1}{n_1 n_0}
\end{aligned}
$$

$$
\begin{aligned}
2FHg &= \frac{2FHp_{111}\tilde{\tilde{\mu}}(1,1,1)}{\left(p^{11}\right)^2\pi(1)} + \frac{2FHp_{110}\tilde{\tilde{\mu}}(1,1,0)}{\left(p^{11}\right)^2\pi(0)} \\
&\quad + \frac{2FHp_{101}\tilde{\tilde{\mu}}(1,0,1)}{\left(p^{10}\right)^2\pi(1)} + \frac{2FHp_{100}\tilde{\tilde{\mu}}(1,0,0)}{\left(p^{10}\right)^2\pi(0)} \\
&= \frac{\frac{2FH}{\left(p^{11}\right)^2}p_{111}\tilde{\tilde{\mu}}(1,1,1)N_1n_0}{n_1n_0} + \frac{\frac{2FH}{\left(p^{11}\right)^2}p_{110}\tilde{\tilde{\mu}}(1,1,0)N_0n_1}{n_1n_0} \\
&\quad + \frac{\frac{2FH}{\left(p^{10}\right)^2}p_{101}\tilde{\tilde{\mu}}(1,0,1)N_1n_0}{n_1n_0} + \frac{\frac{2FH}{\left(p^{10}\right)^2}p_{100}\tilde{\tilde{\mu}}(1,0,0)N_0n_1}{n_1n_0}
\end{aligned}
$$

$$
\begin{aligned}
2FIh &= \frac{2FIp_{111}\tilde{\tilde{\mu}}(1,1,1)}{\left(p^{11}\right)^2\pi(1)} + \frac{2FIp_{110}\tilde{\tilde{\mu}}(1,1,0)}{\left(p^{11}\right)^2\pi(0)} \\
&\quad + \frac{2FIp_{011}\tilde{\tilde{\mu}}(0,1,1)}{\left(p^{01}\right)^2\pi(1)} + \frac{2FIp_{010}\tilde{\tilde{\mu}}(0,1,0)}{\left(p^{01}\right)^2\pi(0)} \\
&= \frac{\frac{2FI}{\left(p^{11}\right)^2}p_{111}\tilde{\tilde{\mu}}(1,1,1)N_1n_0}{n_1n_0} + \frac{\frac{2FI}{\left(p^{11}\right)^2}p_{110}\tilde{\tilde{\mu}}(1,1,0)N_0n_1}{n_1n_0} \\
&\quad + \frac{\frac{2FI}{\left(p^{01}\right)^2}p_{011}\tilde{\tilde{\mu}}(0,1,1)N_1n_0}{n_1n_0} + \frac{\frac{2FI}{\left(p^{01}\right)^2}p_{010}\tilde{\tilde{\mu}}(0,1,0)N_0n_1}{n_1n_0}
\end{aligned}
$$

$$
\begin{aligned}
H^2g &= \frac{H^2p_{111}\tilde{\tilde{\mu}}(1,1,1)}{\left(p^{11}\right)^2\pi(1)} + \frac{H^2p_{110}\tilde{\tilde{\mu}}(1,1,0)}{\left(p^{11}\right)^2\pi(0)} \\
&\quad + \frac{H^2p_{101}\tilde{\tilde{\mu}}(1,0,1)}{\left(p^{10}\right)^2\pi(1)} + \frac{H^2p_{100}\tilde{\tilde{\mu}}(1,0,0)}{\left(p^{10}\right)^2\pi(0)} \\
&= \frac{\frac{H^2}{\left(p^{11}\right)^2}p_{111}\tilde{\tilde{\mu}}(1,1,1)N_1n_0}{n_1n_0} + \frac{\frac{H^2}{\left(p^{11}\right)^2}p_{110}\tilde{\tilde{\mu}}(1,1,0)N_0n_1}{n_1n_0} \\
&\quad + \frac{\frac{H^2}{\left(p^{10}\right)^2}p_{101}\tilde{\tilde{\mu}}(1,0,1)N_1n_0}{n_1n_0} + \frac{\frac{H^2}{\left(p^{10}\right)^2}p_{100}\tilde{\tilde{\mu}}(1,0,0)N_0n_1}{n_1n_0}
\end{aligned}
$$

$$
\begin{aligned}
2HIj &= \frac{\frac{2HI}{\left(p^{11}\right)^2} p_{111}\tilde{\tilde{\mu}}(1,1,1)}{\pi(1)} + \frac{\frac{2HI}{\left(p^{11}\right)^2} p_{110}\tilde{\tilde{\mu}}(1,1,0)}{\pi(0)} \\
&= \frac{\frac{2HI}{\left(p^{11}\right)^2} p_{111}\tilde{\tilde{\mu}}(1,1,1) N_1 n_0}{n_1 n_0} + \frac{\frac{2HI}{\left(p^{11}\right)^2} p_{110}\tilde{\tilde{\mu}}(1,1,0) N_0 n_1}{n_1 n_0}
\end{aligned}
$$

$$
\begin{aligned}
I^2 h &= \frac{I^2 p_{111}\tilde{\tilde{\mu}}(1,1,1)}{\left(p^{11}\right)^2 \pi(1)} + \frac{I^2 p_{110}\tilde{\tilde{\mu}}(1,1,0)}{\left(p^{11}\right)^2 \pi(0)} \\
&\quad + \frac{I^2 p_{011}\tilde{\tilde{\mu}}(0,1,1)}{\left(p^{01}\right)^2 \pi(1)} + \frac{I^2 p_{010}\tilde{\tilde{\mu}}(0,1,0)}{\left(p^{01}\right)^2 \pi(0)} \\
&= \frac{\frac{I^2}{\left(p^{11}\right)^2} p_{111}\tilde{\tilde{\mu}}(1,1,1) N_1 n_0}{n_1 n_0} + \frac{\frac{I^2}{\left(p^{11}\right)^2} p_{110}\tilde{\tilde{\mu}}(1,1,0) N_0 n_1}{n_1 n_0} \\
&\quad + \frac{\frac{I^2}{\left(p^{01}\right)^2} p_{011}\tilde{\tilde{\mu}}(0,1,1) N_1 n_0}{n_1 n_0} + \frac{\frac{I^2}{\left(p^{01}\right)^2} p_{010}\tilde{\tilde{\mu}}(0,1,0) N_0 n_1}{n_1 n_0}
\end{aligned}
$$

After simplification, we can rewrite,

$$
var(\hat{\beta}_1) = \frac{\zeta_0 n_0 + \zeta_1 n_1}{n_0 n_1 N},
$$

where

$$\begin{aligned}\zeta_0 = &\frac{F^2}{\left(p^{11}\right)^2} p_{111}\tilde{\tilde{\mu}}\left(1,1,1\right) N_1 + \frac{F^2}{\left(p^{10}\right)^2} p_{101}\tilde{\tilde{\mu}}\left(1,0,1\right) N_1 \\ &+ \frac{F^2}{\left(p^{01}\right)^2} p_{011}\tilde{\tilde{\mu}}\left(0,1,1\right) N_1 + \frac{F^2}{\left(p^{00}\right)^2} p_{001}\tilde{\tilde{\mu}}\left(0,0,1\right) N_1 \\ &+ \frac{2FH}{\left(p^{11}\right)^2} p_{111}\tilde{\tilde{\mu}}\left(1,1,1\right) N_1 + \frac{2FH}{\left(p^{10}\right)^2} p_{101}\tilde{\tilde{\mu}}\left(1,0,1\right) N_1 \\ &+ \frac{2FI}{\left(p^{11}\right)^2} p_{111}\tilde{\tilde{\mu}}\left(1,1,1\right) N_1 + \frac{2FI}{\left(p^{01}\right)^2} p_{011}\tilde{\tilde{\mu}}\left(0,1,1\right) N_1 \\ &+ \frac{H^2}{\left(p^{11}\right)^2} p_{111}\tilde{\tilde{\mu}}\left(1,1,1\right) N_1 + \frac{H^2}{\left(p^{10}\right)^2} p_{101}\tilde{\tilde{\mu}}\left(1,0,1\right) N_1 \\ &+ \frac{2HI}{\left(p^{11}\right)^2} p_{111}\tilde{\tilde{\mu}}\left(1,1,1\right) N_1 \\ &+ \frac{I^2}{\left(p^{11}\right)^2} p_{111}\tilde{\tilde{\mu}}\left(1,1,1\right) N_1 + \frac{I^2}{\left(p^{01}\right)^2} p_{011}\tilde{\tilde{\mu}}\left(0,1,1\right) N_1,\end{aligned}$$

$$\begin{aligned}\zeta_1 = &\frac{F^2}{\left(p^{11}\right)^2} p_{110}\tilde{\tilde{\mu}}\left(1,1,0\right) N_0 + \frac{F^2}{\left(p^{10}\right)^2} p_{100}\tilde{\tilde{\mu}}\left(1,0,0\right) N_0 \\ &+ \frac{F^2}{\left(p^{01}\right)^2} p_{010}\tilde{\tilde{\mu}}\left(0,1,0\right) N_0 + \frac{F^2}{\left(p^{00}\right)^2} p_{000}\tilde{\tilde{\mu}}\left(0,0,0\right) N_0 \\ &+ \frac{2FH}{\left(p^{11}\right)^2} p_{110}\tilde{\tilde{\mu}}\left(1,1,0\right) N_0 + \frac{2FH}{\left(p^{10}\right)^2} p_{100}\tilde{\tilde{\mu}}\left(1,0,0\right) N_0 \\ &+ \frac{2FI}{\left(p^{11}\right)^2} p_{110}\tilde{\tilde{\mu}}\left(1,1,0\right) N_0 + \frac{2FI}{\left(p^{01}\right)^2} p_{010}\tilde{\tilde{\mu}}\left(0,1,0\right) N_0 \\ &+ \frac{H^2}{\left(p^{11}\right)^2} p_{110}\tilde{\tilde{\mu}}\left(1,1,0\right) N_0 + \frac{H^2}{\left(p^{10}\right)^2} p_{100}\tilde{\tilde{\mu}}\left(1,0,0\right) N_0 \\ &+ \frac{2HI}{\left(p^{11}\right)^2} p_{110}\tilde{\tilde{\mu}}\left(1,1,0\right) N_0 \\ &+ \frac{I^2}{\left(p^{11}\right)^2} p_{110}\tilde{\tilde{\mu}}\left(1,1,0\right) N_0 + \frac{I^2}{\left(p^{01}\right)^2} p_{010}\tilde{\tilde{\mu}}\left(0,1,0\right) N_0.\end{aligned}$$

According to the Proposition 1 in Chapter 2, the optimal sampling ratio of n_0 to n_1 is given by $R_{OD} = \sqrt{\frac{\zeta_1}{\zeta_0}}$.