Multiplicative Effect Modeling

Jiaqi Yin

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Thomas S. Richardson, Chair

Linbo Wang, Chair

Jon Wakefield

Program Authorized to Offer Degree:
Biostatistics – Public Health

University of Washington

Abstract

Multiplicative Effect Modeling

Jiaqi Yin

Co-Chairs of the Supervisory Committee:
Thomas S. Richardson
Department of Statistics

Linbo Wang
Department of Statistics

Generalized linear models, such as logistic regression, are widely used to model the association between a treatment and a binary outcome as a function of baseline covariates. However, the coefficients of a logistic regression model correspond to log odds ratios, while subject-matter scientists are often interested in relative risks. Although odds ratios are sometimes used to approximate relative risks, this approximation is appropriate only when the outcome of interest is rare for all levels of the covariates. Poisson regressions do measure multiplicative treatment effects including relative risks, but with a binary outcome not all combinations of parameters lead to fitted means that are between zero and one. Enforcing this constraint makes the parameters variation dependent, which is undesirable for modeling, estimation and computation. Focusing on the special case where the treatment is also binary, Richardson et al. (2017) propose a novel binomial regression model, that allows direct modeling of the relative risk. The model uses a log odds-product nuisance model leading to variation independent parameter spaces. However, their method is restricted to binary treatments.

My research presents general approaches to modeling the multiplicative effect of a continuous or categorical treatment on a binary outcome. We also attempt to develop a method to estimate relative risks in a meta-analysis including cohort and case-control studies.

In Chapter 1, we introduce the relative risk modeling of the binary treatment case in Richardson et al. (2017). In their work, they have proposed a nuisance model for the log of the odds product, which is the product of two odds of the treatments. The odds product is variation independent from the primary of interest, relative risk, thus this leads to a valid probability distribution.

In Chapter 2, we introduce a new approach which imposes an assumption that the relative risk is a monotone function of an ordinal treatment. This assumption is reasonable in many real-life situations, such as the recovery probability in the arm receiving full-dosage is often higher than it in the small-dosage arm. Furthermore, having the relative risks and the odds product of the lowest and highest level of treatment, we are able to have a valid probability distribution too. We also provide simulations to demonstrate our proposed method and compare the performance with other methods.

In Chapter 3, we propose another new method for the case where the relative risk is not monotonic in treatment. We introduce the generalized odds product as the nuisance model, which is the product of odds for all the levels of the treatments. Simulation and model comparison are also provided.

In Chapter 4, we illustrate the use of our proposed methods in Chapter 2 and 3 by studying the association between the passenger class and death/survival status in the Titanic data. We also compare the results from our proposed models with those obtained from the generalized linear models: Poisson and logistic regression.

In Chapter 5, we propose a novel method to estimate the relative risk in a meta-analysis including cohort and case-control studies. A case-control is often conducted for rare diseases, however it cannot determine a relative risk because the prevalence of disease is set by the study design. Cohort studies can provide information on prevalence, while for rare events they cannot collect sufficient cases. In this chapter, we combine cohort and case-control studies to have a better estimation of the relative risk. Monte Carlo simulations demonstrate

the superior performance of our proposed method. We also apply the method to the National Longitudinal Survey of Youth Data to study the relative risk of smoking on people's heath.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my advisors, Linbo Wang and Thomas Richardson. They have introduced me to causal inference, guided me as well as supported me along the way.

Back to the Spring of 2017, I just ended an independent study and I was not happy and under a lot of pressure during that time. As a second-year PhD or almost third year, I was standing at a four way crosswalk and didn't know who I wanted to be and where I wanted to go after the PhD training. Even today, when I think back to the old days, I can still feel the anxiety, loneliness, and helplessness, plus the rainy days. Linbo went back to Seattle for some conference one year after he graduated from our department. As an old neighbor of Linbo, we had lunch to catch up a bit. Learning that he had a great time at Harvard as a postdoc, I also shared my situations. He suggested that I talked with Thomas who was his advisor back in graduate school. Can I? I didn't know anything about causal. That was the first question I asked Linbo. He recommended me a book called Causality from Judea Pearl. I did not contact Thomas until I finished reading the book. I can imagine he would laugh at me after knowing I was afraid of meeting him at the very beginning. He is a nice gentleman, always encouraging people, and will never make them feel embarrassed. Thomas is also a great researcher. He is passionate about unknowns and has a systemic way to solve them. After working with Thomas, I started to enjoy my life as a PhD student. I sometimes even picture myself as a professor, helping, teaching and inspiring students. I guess the luckiest things in graduate school are that I could be Linbo's neighbor and also work with both of them.

I would like to extend my sincere thanks to Jon Wakefield, Patrick J. Heagerty, and

Yanqin Fan for their constructive comments and support during my dissertation research. Special thanks to Jon for discussions on my work of meta-analysis.

I would like to thank the Department of Biostatistics, National Alzheimer's Coordinating Center, and Department of Epidemiology for providing the financial support during five years of studying.

I cannot begin to express my thanks to my parents, especially my dad. He is a hard-working man. He has proved to me that people can live the life they want by chasing their dreams and working hard. Finally, the completion of my dissertation would not have been possible without the support of my partner Xiaoxiao Wu, and cat kids, Goudan and Ludan under quarantine. The outside has become chaotic because of the pandemic, however once I have you, I feel happy and peaceful inside.

DEDICATION

to my family

Chapter 1

INTRODUCTION

1.1 Overview

Binary outcomes, such as alive versus dead, yes versus no, success versus failure, and so on, are widely seen in epidemiological and biomedical studies. My dissertation mainly studies the multiplicative effect modeling of a binary outcome.

Researchers are primarily interested in estimating the effect of a treatment Z on a binary outcome Y on the multiplicative scale. Relative risks and odds ratios are two important measures of such association. Relative risks (RR) are ratios contrasting the probability of $Y = 1$ in treatment group $Z = z$ versus the probability of $Y = 1$ in a baseline group $Z = z_0$:

$$\text{RR}(z_0, z) = \frac{\text{pr}(Y = 1 \mid Z = z)}{\text{pr}(Y = 1 \mid Z = z_0)}.$$

Odds is simply the ratio between the probability of $Y = 1$ and the probability of $Y = 0$, and an odds ratio (OR) is the ratio between the odds for two different levels of treatment,

$$\text{OR}(z_0, z) = \frac{\text{pr}(Y = 1 \mid Z = z)/\text{pr}(Y = 0 \mid Z = z)}{\text{pr}(Y = 1 \mid Z = z_0)/\text{pr}(Y = 0 \mid Z = z_0)}.$$

In this work, we consider a continuous or categorical treatment Z.

The logistic model is widely used to estimate odds ratio. In a logistic model, the probability of the outcome Y is modeled as a function of covariates using a logit function. The coefficient associated with a particular binary covariate, which we will refer to as treatment, is a log-odds ratio. Since the resulting likelihood is concave, it is feasible to compute maximum likelihood estimates for large data sets.

However, in many epidemiological and medical studies, researchers are primarily interested in estimating relative risks (Lumley et al., 2006). In practice odds ratios are sometimes used to approximate relative risks. Mathematically, RR can be written as the function of OR and the baseline probability $\text{pr}(Y = 1 \mid Z = z_0)$:

$$\text{RR} = \frac{\text{OR}}{\{1 - \text{pr}(Y = 1 \mid Z = z_0)\} + \{\text{pr}(Y = 1 \mid Z = z_0) \times \text{OR}\}}.$$

Such relationship is showed in Figure 1.1. Any points in the same curve have the same odds ratio. It shows that when the disease or interest of the outcome being studied is rare in the

population, the odds ratio is close to the relative risk Zhang and Yu (1998). However, when the outcome is prevalent, odds ratios and relative risks may be very different. Consequently,

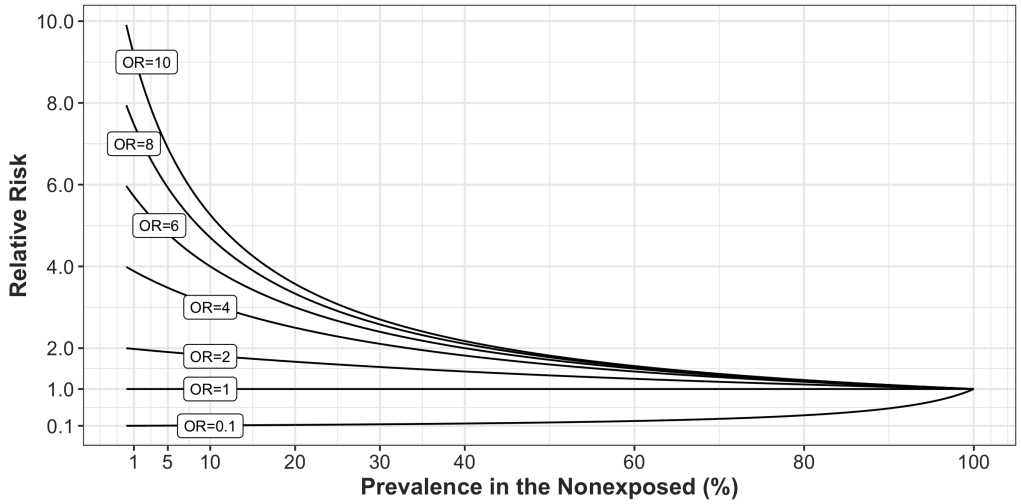


Figure 1.1: The relationship between relative risk and odds ratio by prevalence. Any point in the same curve has the same odds ratio.

it is not usually appropriate to use odds ratios as an approximation for relative risks when the prevalence is large. There are also qualitative differences between these measures: whereas relative risks are collapsible, so that the marginal relative risk will lie in the convex hull of stratum-specific relative risks (Greenland et al., 1999), the same does not hold for odds ratios.

Even though relative risks are often the primary of interesting, in some circumstance, they cannot be obtained such as in a case-control study. It is because the prevalence of the disease is set by the study design. Instead, people have to use odds ratios to quantify the association between the exposure and the outcome in a case control study. This can be

explained by the following qualities:

$$\begin{aligned}\text{OR}(z_0, z) &= \frac{\text{pr}(Y = 1 \mid Z = z)/\text{pr}(Y = 0 \mid Z = z)}{\text{pr}(Y = 1 \mid Z = z_0)/\text{pr}(Y = 0 \mid Z = z_0)} \\ &= \frac{\text{pr}(Z = z \mid Y = 1) \cdot \text{pr}(Z = z_0 \mid Y = 0)}{\text{pr}(Z = z \mid Y = 0) \cdot \text{pr}(Z = z_0 \mid Y = 1)}.\end{aligned}$$

The probability $\text{pr}(Z = z \mid Y = y)$ is identifiable from a case-control study and therefore is used to calculate the odds ratio instead of using the prevalence.

1.2 Multiplicative Effect Modeling

Within the generalized linear model framework, multiplicative treatment effects are typically modeled via a Poisson regression which imposes a linear association between the log of the probability of $Y = 1$ and covariates V,

$$\log\{\text{pr}(Y = 1 \mid Z, V)\} = Z\alpha^{\text{T}}V + \beta^{\text{T}}V.$$

Equivalently, for a binary treatment $Z \in \{0, 1\}$, Poisson model can be written as

$$\begin{aligned}\log\{\text{RR}(0, 1; V)\} &= \alpha^{\text{T}}V, \\ \log\{\text{pr}(Y = 1 \mid Z = 0, V)\} &= \beta^{\text{T}}V.\end{aligned}$$

However, Poisson regression with a binary outcome has drawbacks in terms of modeling, prediction and computation. This is because $\text{RR}(z_0, z)$ is variation dependent on the baseline probability $\text{pr}(Y = 1 \mid Z = z_0)$. This can be directly seen from Figure 1.2. Each gray line represents a constant relative risk. The baseline probability showing as the vertical red line does not intersect with all the gray lines, which indicates that $\text{RR}(z_0, z)$ and the baseline probability $\text{pr}(Y = 1 \mid Z = z_0)$ are not variation independent. For example, if $\text{RR}(z_0, z) = 2$, then $\text{pr}(Y = 1 \mid Z = z) = 2 \times \text{pr}(Y = 1 \mid Z = z_0)$, so that $\text{pr}(Y = 1 \mid Z = z_0) \leq 0.5$. Therefore there is a restricted domain over which the quantities $[\{\text{RR}(z_0, z); z\}, \text{pr}(Y = 1 \mid Z = z_0)]$ are compatible with a valid probability distribution. This may lead to misspecification when modeling. Also the fitted probability for any treatment given covariates can go outside of the range $[0, 1]$.

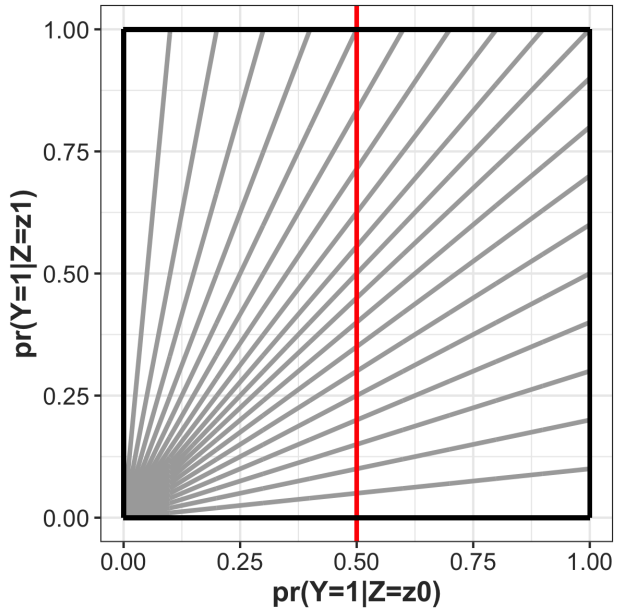


Figure 1.2: Lines of constant relative risks.

Richardson et al. (2017) provide a simple alternative to modeling the relative risk for a binary treatment. In addition to specifying a model for log(RR), they propose a nuisance model for the log of the odds product (OP):

$$\text{OP}(z_0, z) = \frac{\text{pr}(Y = 1 \mid Z = z)\text{pr}(Y = 1 \mid Z = z_0)}{\{1 - \text{pr}(Y = 1 \mid Z = z)\}\{1 - \text{pr}(Y = 1 \mid Z = z_0)\}}.$$

In Figure 1.3, any points in the same gray line have the same odds product value. The red line represents a constant relative risk, which intersects with all the grey lines. This leads to an unrestricted domain for which the quantities $[\{\text{RR}(z_0, z); z\}, \text{OP}(z_0, z)]$ are compatible with a valid probability distribution. Hence the parameters of interest in the two models are variation independent. However, their method is restricted to binary treatments.

Alternatively, Tchetgen Tchetgen (2013) and Dukes and Vansteelandt (2018) propose semi-parametric g-estimation methods for the relative risk of treatment. Their approaches do not employ all the information in the observed data likelihood, and as we illustrate later in the simulations, can be less efficient under correct model specifications.

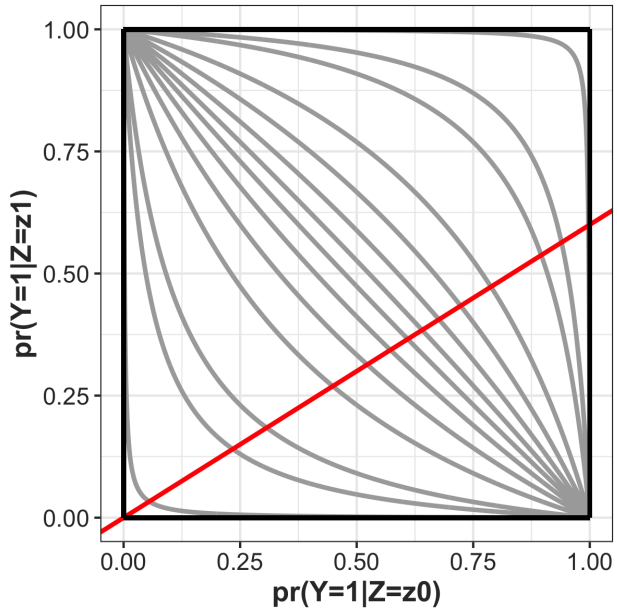


Figure 1.3: Lines of constant odds products.

Building on Richardson et al. (2017), we present two new approaches that model multiplicative effects with continuous or categorical treatments. The first imposes an assumption that the relative risk is a monotone function of an ordinal treatment. The second introduces a new nuisance model, using a so-called generalized odds product (Wang et al., 2017), that is variation independent of the relative risks. These two methods are designed for cohort studies. We further extend the work to modeling relative risk for a meta-analysis including cohort and case-control studies.

This dissertation is organized as follows: In Chapter 2 and Chapter 3, we introduce the two methods, and demonstrate the superior performance of our proposed methods via Monte Carlo simulation; In Chapter 4, a data analysis demonstrate the use of our methods; In Chapter 5, we further extend our models for meta analyses. Simulations and applications are also provided.

Chapter 2

PARAMETERIZATION WITH MONOTONIC TREATMENT EFFECTS

2.1 Methodology

Denote the relative risk between a treatment z and the baseline treatment z_0 given a random vector of covariates v as

$$\text{RR}(z_0, z; v) = \frac{\text{pr}(Y = 1 \mid V = v, Z = z)}{\text{pr}(Y = 1 \mid V = v, Z = z_0)}, \tag{2.1}$$

where Z can be a continuous or categorical treatment. For notational simplicity, we denote $\text{pr}(Y = 1 \mid Z = z, V = v)$ as $p_z(v)$. Similarly, the odds product of treatment z and baseline treatment z_0 is

$$\text{OP}(z_0, z; v) = \frac{p_0(v)p_z(v)}{\{1 - p_0(v)\}\{1 - p_z(v)\}}. \tag{2.2}$$

To fix ideas, first consider the special case where Z is a continuous treatment taking values in a bounded interval, say $[0, 1]$. Our goal is to find $\phi(v)$ so that for any v, the mapping given by

$$(\log\{\text{RR}(0, z; v)\}, z \in [0, 1]; \phi(v)) \to (p_z(v), z \in [0, 1])$$

is a diffeomorphism between the interior of their domains. Recall that Richardson et al. (2017) show that if we let $\phi(v) = \log\{\text{OP}(0, 1; v)\}$, then any possible value of $(\log\{\text{RR}(0, 1; v)\}, \phi(v))$ implies that $(p_0(v), p_1(v)) \in (0, 1)^2$. The key insight for our development is that if $\log\{\text{RR}(0, z; v)\}$ is monotonic in z, or equivalently, the treatment effect is monotonic for all covariate values v, then $p_z(v)$ is also monotonic in z. Consequently,

$$0 < \min\{p_0(v), p_1(v)\} \leq p_z(v) \leq \max\{p_0(v), p_1(v)\} < 1 \quad (z \in [0, 1]).$$

Therefore, any possible value of $(\log \text{RR}(0, z; v), \phi(v))$ such that $\log(\text{RR}(0, z; v))$ is monotone in z implies that $p_z(v) \in (0, 1)$ for all $z \in [0, 1]$.

The monotonic treatment effect assumption we have invoked may be considered reasonable in many real-life situations. For example, the recovery probability in the arm receiving full-dosage is usually at least as high as in the small-dosage arm (Yang et al., 2003; Al-Mamgani et al., 2008; Lang and Birkenmeier, 2015), and greater income may be associated with a higher probability of satisfaction (Easterlin, 2001; Ball and Chernova, 2008).

This idea above can be generalized to accommodate more types of variables for the treatment Z, such as ordinal and unbounded continuous variables.

Theorem 1 (Variation independence with monotonic treatment effects)**.** *Let $\mathcal{Z} \subseteq \mathbb{R}$ and $\mathcal{V}$ be the support of Z and V, respectively. Let $h(z,v)$ and $g(v)$ be real-valued functions with support $\mathcal{Z} \times \mathcal{V}$ and $\mathcal{V}$, respectively. If $h(z,v)$ is bounded in z and monotonic in z, then there exists a unique set of proper probability distributions $\{p_z(v); z \in \mathcal{Z}, v \in \mathcal{V}\}$ such that $\log\{\text{RR}(z_0, z; v)\} = h(z,v)$ and $\log\{\text{OP}(z_{\text{inf}}, z_{\text{sup}}; v)\} = g(v)$, where $z_{\text{inf}} = \inf\{z : z \in \mathcal{Z}\}$, $z_{\text{sup}} = \sup\{z : z \in \mathcal{Z}\}$ and*

$$\text{OP}(z_{\text{inf}}, z_{\text{sup}}; v) = \lim_{z_1 \to z_{\text{inf}}} \lim_{z_2 \to z_{\text{sup}}} \frac{p_{z_1}(v) p_{z_2}(v)}{\{1 - p_{z_1}(v)\}\{1 - p_{z_2}(v)\}}.$$

Remark 1. *The boundedness condition on $h(v,z)$ guarantees that the implied probabilities $p_z(v)$ are bounded away from 0.*

Proof. To prove the existence of a unique set of proper probability distributions $\{p_z(v); z \in \mathcal{Z}, v \in \mathcal{V}\}$, it is sufficient to show that $p_z(v)$ can be written as a function of $h(z,v)$ and $g(v)$. Because for any $v \in \mathcal{V}$, $h(z,v)$ is bounded and monotonic in z, $\lim_{z \to z_{\text{inf}}} h(z,v)$ and $\lim_{z \to z_{\text{sup}}} h(z,v)$ exist, denoted as $h_1(v)$ and $h_2(v)$. Without loss of generality, we assume $h(z,v)$ is monotonically non-decreasing in z. For simplicity, we denote these by $\lim_{z \to z_{\text{inf}}} p_z(v)$ and $\lim_{z \to z_{\text{sup}}} p_z(v)$ as $p_{z_{\text{inf}}}(v)$ and $p_{z_{\text{sup}}}(v)$, respectively; Let $\Delta(v) = e^{2g(v)}\left(e^{h_1(v)-h_2(v)} + 1\right)^2 + 4e^{h_1(v)-h_2(v)+g(v)}\left(1 - e^{g(v)}\right) > 0$.

For any fixed $v \in \mathcal{V}$, $p_{z_{\text{sup}}}(v), p_{z_{\text{inf}}}(v), p_{z_0}(v)$ and $p_z(v)$ via

$$p_{z_{\text{sup}}}(v) = \begin{cases} \frac{e^{g(v)}\{1+e^{h_1(v)-h_2(v)}\} - \sqrt{\Delta(v)}}{2e^{h_1(v)-h_2(v)}\{e^{g(v)}-1\}} & g(v) \neq 0 \\ \frac{1}{1+e^{h_1(v)-h_2(v)}} & g(v) = 0 \end{cases}, \tag{2.3}$$

$$p_{z_{\text{inf}}}(v) = p_{z_{\text{sup}}}(v) e^{h_1(v)-h_2(v)}, \tag{2.4}$$

$$p_{z_0}(v) = p_{z_{\text{sup}}}(v) e^{-h_2(v)}, \tag{2.5}$$

$$p_z(v) = p_{z_{\text{sup}}}(v) e^{h(z,v)-h_2(v)} \quad (z \in \mathcal{Z}). \tag{2.6}$$

We now show

$$\log\{\textsc{rr}(z_0, z; v)\} = h(v, z), \tag{2.7}$$

$$\log\{\textsc{op}(z_{\inf}, z_{\sup}; v\} = g(v). \tag{2.8}$$

In the case where $g(v) = 0$, it is easy to see that (2.7) and (2.8) hold. If $g(v) \neq 0$, for any $v \in \mathcal{V}$, one may divide (2.6) by (2.5) and take the logarithm of both sides. The resulting expression satisfies (2.7). Next we prove that $p_{z_{\sup}}(v) \in (0, 1)$, which is equivalent to showing that $p_{z_{\sup}}(v)\{p_{z_{\sup}}(v) - 1\} < 0$ for any fixed v.

$$\begin{aligned} & p_{z_{\sup}}(v)\{p_{z_{\sup}}(v) - 1\} \\ &= \frac{\left[e^{g(v)}\{1 + e^{h_1(v)-h_2(v)}\} - \sqrt{\Delta(v)}\right]\left[e^{g(v)} - e^{h_1(v)-h_2(v)+g(v)} + 2e^{h_1(v)-h_2(v)} - \sqrt{\Delta(v)}\right]}{\left[2e^{h_1(v)-h_2(v)}\{e^{g(v)} - 1\}\right]^2}. \end{aligned}$$

It is enough to prove that the numerator of the above equation is smaller than 0, which can be directly computed. Further $\textsc{op}(z_{\inf}, z_{\sup}; v)$ maybe obtained explicitly as:

$$\begin{aligned} & \frac{p_{z_{\sup}}(v)p_{z_{\inf}}(v)}{\{1 - p_{z_{\sup}}(v)\}\{1 - p_{z_{\inf}}(v)\}} \\ &= \frac{\{e^{g(v)}(1 + e^{h_1(v)-h_2(v)}) - \sqrt{\Delta(v)}\}^2}{\left(e^{h_1(v)-h_2(v)+g(v)} - 2e^{h_1(v)-h_2(v)} - e^{g(v)} + \sqrt{\Delta(V)}\right)\left(e^{g(v)} - e^{h_1(v)-h_2(v)+g(v)} - 2 + \sqrt{\Delta(V)}\right)} \\ &= \frac{e^{g(v)}\left[2e^{g(v)}\left\{e^{h_1(v)-h_2(v)} + 1\right\}^2 - 4e^{h_1(v)-h_2(v)}\{e^{g(v)} - 1\} - 2\{1 + e^{h_1(v)-h_2(v)}\}\sqrt{\Delta(v)}\right]}{2e^{g(v)}\left\{e^{h_1(v)-h_2(v)} + 1\right\}^2 - 4e^{h_1(v)-h_2(v)}\{e^{g(v)} - 1\} - 2\{1 + e^{h_1(v)-h_2(v)}\}\sqrt{\Delta(v)}} \\ &= e^{g(v)}. \end{aligned}$$

Thus (2.8) is satisfied. This completes our proof.

□

In our simulations and data analysis, we consider a bounded treatment Z and the following models for $\log\{\textsc{rr}(z_0, z; v)\}$ and $\log\{\textsc{op}(z_{\min}, z_{\max}; v)\}$:

$$\log\{\textsc{rr}(z_0, z; V, \gamma)\} = \gamma^{\mathrm{T}} V(z - z_0) \quad z \in \mathcal{Z}, \tag{2.9}$$

$$\log\{\textsc{op}(z_{\min}, z_{\max}; V, \beta)\} = \beta^{\mathrm{T}} V, \tag{2.10}$$

where $z_{\min} = \min\{z : z \in \mathcal{Z}\}$, $z_{\max} = \max\{z : z \in \mathcal{Z}\}$. In light of the boundedness condition on $h(v, z)$, when the treatment is unbounded, researchers should avoid specifying a linear model such as the one on the right hand side of (2.9).

The log-likelihood for a unit i can be written as

$$l(\gamma, \beta | z_i, v_i, y_i) = y_i \log\{p_{z_i}(v_i; \gamma, \beta)\} + (1 - y_i) \log\{1 - p_{z_i}(v_i; \gamma, \beta)\}. \tag{2.11}$$

Inference on γ and β can be obtained in standard fashion. In general, the likelihood is not concave. In practice, we use a simple iterative procedure for finding a solution to the score equation. To be more specific: for the method which assumes monotonicity, we assign a starting value for γ and β. At each step t, we first find $\gamma^{(t)}$ via maximizing the (profile) log-likelihood while holding β fixed at $\beta^{(t-1)}$ and; we then find the optimal $\beta^{(t)}$ via maximizing the log-likelihood holding γ fixed at $\gamma^{(t)}$. The iterations stop when the differences between the parameters at successive iterations are smaller than a pre-defined threshold.

2.2 Variance Formula

In this section, we provide explicit formula for Wald-type variance.

The log-likelihood for a unit can be written as

$$l(\gamma, \beta | z, v, y) = y \log\{p_z(v; \gamma, \beta)\} + (1 - y) \log\{1 - p_z(v; \gamma, \beta)\}. \tag{2.12}$$

Without loss of generality, let both the treatment $z_{\min}$ and the baseline treatment be zero. Denote $\theta(v) = \gamma^{\mathrm{T}} v$, $g(v) = \beta^{\mathrm{T}} v$, $\psi(v) = \log p_0(v)$, and $p_z(v) = e^{z\theta(v) + \psi(v)}$ $(z \in \mathcal{Z})$. For simplicity, we write $l, \theta, g, \psi, p_z, p_0$ referring to $l(\gamma, \beta | z, v, y), \theta(v), g(v), \psi(v), p_z(v), p_0(v)$, respectively. The functional dependence structure of the variables is shown in Figure 2.1. Further we have the derivatives of $l(\gamma, \beta | z_i, v_i, y_i)$ with respect to γ and β:

$$\frac{\partial l}{\partial \gamma} = \frac{\partial l}{\partial p_z} \left(\frac{\partial p_z}{\partial \theta} \frac{\partial \theta}{\partial \gamma} + \frac{\partial p_z}{\partial \psi} \frac{\partial \psi}{\partial \theta} \frac{\partial \theta}{\partial \gamma} \right), \tag{2.13}$$

$$\frac{\partial l}{\partial \beta} = \frac{\partial l}{\partial p_z} \frac{\partial p_z}{\partial \psi} \frac{\partial \psi}{\partial g} \frac{\partial g}{\partial \beta}. \tag{2.14}$$

In the following, we calculate the terms in (2.13) and (2.14).

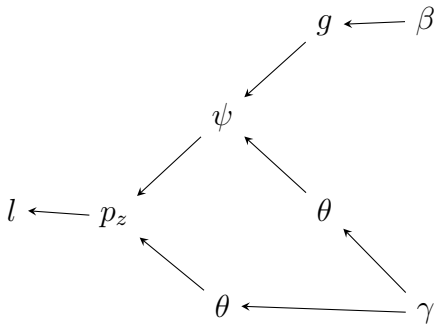


Figure 2.1: Variable structure of the proposed method under the monotonic treatment effects assumption.

$$\frac{\partial l}{\partial p_z} = \frac{y - p_z}{p_z(1 - p_z)},$$

$$\frac{\partial p_z}{\partial \theta} = z p_z, \quad \frac{\partial p_z}{\partial \psi} = p_z.$$

To get $\frac{\partial \psi}{\partial g}, \frac{\partial \psi}{\partial \theta}$, we start from $g(v) = \log\{\text{OP}(0, z_{\max})\}$.

$$\begin{aligned} g &= \log \frac{p_0 p_{z_{\max}}}{\{1 - p_0\}\{1 - p_{z_{\max}}\}} \\ &= \log \frac{p_0^2 e^{z_{\max}\theta}}{(1 - p_0)(1 - p_0 e^{z_{\max}\theta})} \\ &= 2 \log p_0 + k\theta - \log(1 - p_0) - \log(1 - p_0 e^{z_{\max}\theta}) \\ &= 2\psi + z_{\max}\theta - \log(1 - e^{\psi}) - \log(1 - e^{\psi + z_{\max}\theta}). \end{aligned}$$

Because $\frac{\partial g}{\partial \theta} = 0$, we further have

$$\begin{aligned} \frac{\partial g}{\partial \theta} &= 2\frac{\partial \psi}{\partial \theta} + z_{\max} + \frac{e^{\psi}\frac{\partial \psi}{\partial \theta}}{1 - e^{\psi}} + \frac{e^{\psi + z_{\max}\theta}(\frac{\partial \psi}{\partial \theta} + z_{\max})}{1 - e^{\psi + z_{\max}\theta}} \\ &= 2\frac{\partial \psi}{\partial \theta} + z_{\max} + \frac{p_0\frac{\partial \psi}{\partial \theta}}{1 - p_0} + \frac{p_{z_{\max}}(\frac{\partial \psi}{\partial \theta} + z_{\max})}{1 - p_{z_{\max}}} \\ &= 0. \end{aligned}$$

Therefore, we have

$$\frac{\partial \psi}{\partial \theta} = -\frac{z_{\max}(1-p_0)}{1-p_0+1-p_{z_{\max}}};$$

$$\begin{aligned}\frac{\partial g}{\partial \psi} &= 2 + \frac{e^{\psi}}{1-e^{\psi}} + \frac{e^{\psi+z_{\max}\theta}}{1-e^{\psi+z_{\max}\theta}} \\ &= 2 + \frac{p_0}{1-p_0} + \frac{p_{z_{\max}}}{1-p_{z_{\max}}}.\end{aligned}$$

Then

$$\frac{\partial \psi}{\partial g} = \frac{(1-p_{z_{\max}})(1-p_0)}{(1-p_{z_{\max}})+(1-p_0)}.$$

We also have

$$\frac{\partial \theta}{\partial \gamma} = v, \quad \frac{\partial g}{\partial \beta} = v.$$

With the above building blocks, we finally have the derivatives:

$$\frac{\partial l}{\partial \gamma} = \frac{y-p_z}{1-p_z} \cdot \left\{ z - \frac{z_{\max}(1-p_0)}{(1-p_0)+(1-p_{z_{\max}})} \right\} \cdot v, \tag{2.15}$$

$$\frac{\partial l}{\partial \beta} = \frac{y-p_z}{1-p_z} \cdot \frac{(1-p_0)(1-p_{z_{\max}})}{(1-p_0)+(1-p_{z_{\max}})} \cdot v. \tag{2.16}$$

The Fisher Information matrix $\mathcal{I}(\alpha, \beta)$ may be calculated to be

$$\mathcal{I}(\alpha,\beta) = \mathbb{E}\left[\left\{ \left(\frac{\partial l}{\partial \gamma}\right)^{\mathrm{T}}, \left(\frac{\partial l}{\partial \beta}\right)^{\mathrm{T}} \right\}^{\mathrm{T}} \left\{ \left(\frac{\partial l}{\partial \alpha}\right)^{\mathrm{T}}, \left(\frac{\partial l}{\partial \beta}\right)^{\mathrm{T}} \right\} \right] = \mathbb{E}\left[\begin{matrix} \left(\frac{\partial l}{\partial \gamma}\right)\left(\frac{\partial l}{\partial \gamma}\right)^{\mathrm{T}} & \left(\frac{\partial l}{\partial \gamma}\right)\left(\frac{\partial l}{\partial \beta}\right)^{\mathrm{T}} \\ \left(\frac{\partial l}{\partial \beta}\right)\left(\frac{\partial l}{\partial \gamma}\right)^{\mathrm{T}} & \left(\frac{\partial l}{\partial \beta}\right)\left(\frac{\partial l}{\partial \beta}\right)^{\mathrm{T}} \end{matrix} \right].$$

Then variance covariance matrix for $(\gamma^{\mathrm{T}}, \beta^{\mathrm{T}})^{\mathrm{T}}$ is $\{n\mathcal{I}(\gamma^{\mathrm{T}}, \beta^{\mathrm{T}})\}^{-1}$, where n is the sample size.

2.3 Simulation Studies

We evaluate the finite sample performance of our proposed methods. We generate the treatment Z from a uniform distribution on $\{0, 1, 2\}$. The covariates V include an intercept and a random variable generated from a uniform distribution on the interval $[-2, 2]$. We generate Y from models (2.9) and (2.10), where $\gamma = (0, 1)^{\mathrm{T}}$, $\beta = (-0.5, 1)^{\mathrm{T}}$. All simulation results are based on 1000 Monte-Carlo runs.

Table 2.1 summarizes the simulation results. The standard deviation accuracy is defined as the ratio of estimated standard deviation and Monte Carlo standard deviation. Even though, in theory, our estimate is consistent, as shown in Table 2.1, with a small sample of 100, the γ_1's bias can be very large relative to the standard error. In this case, the model-based standard deviation estimate is also much smaller than the Monte Carlo standard deviation. However, as sample size increases, the bias and standard error of our estimator further decreases and the standard deviation accuracy is close to 1; see more simulation in Appendix §B.1. We also display the histogram of our estimates distribution under sample size 100 and 1000 in Figure 2.2. It clearly shows that when sample size is 100, the distribution of estimates of γ_1 is right skewed, which further leads to bias in our estimates. When sample size gets larger, the estimates obey central limit theorem and show a symmetric normal distribution. Coverage indicates the empirical coverage of true parameters in nominal 95%

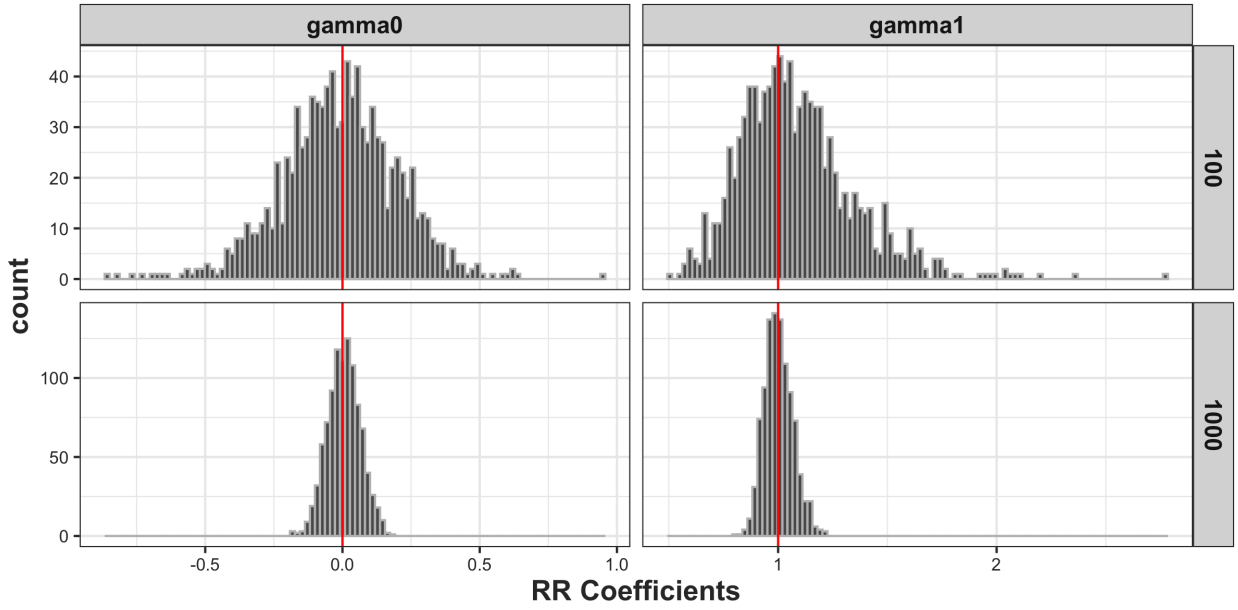


Figure 2.2: Empirical distribution of parameter estimates obtained from 1000 Monte-Carlo runs with sample size 100 and 1000 respectively. The top row is from sample size 100, and the bottom is from sample size 1000. The red vertical line shows the group truth.

Wald-type confidence interval based on point estimators and estimated standard deviation. The coverage probability of the proposed Wald-type confidence intervals also achieve the nominal 95% coverage-rate.

Table 2.1: Monte Carlo simulation results based on 1000 runs for the proposed estimator which assumes monotonic treatment effects. The true values of γ_0 and γ_1 are 0 and 1, respectively

Sample Size		100	500	1000
$\text{Bias}_{\times 10^2}(\text{SE}_{\times 10^2})$				
	γ_0	-0.945(0.671)	-0.005(0.27)	0.533(0.188)
	γ_1	8.728(0.838)	1.200(0.298)	0.232(0.204)
SD Accuracy				
	γ_0	0.948	0.988	0.996
	γ_1	0.881	1.001	1.024
Coverage				
	γ_0	0.962	0.948	0.947
	γ_1	0.951	0.959	0.956

SE, standard error.

SD Accuracy = Average estimated standard deviation/Monte Carlo standard deviation / Monte Carlo standard deviation.

Nominal level = 95%.

We also consider the scenario when the nuisance model is misspecified. An analyst takes variable V^* instead of V in the nuisance model to estimate parameters of interest. V^* includes an intercept and an irrelevant covariate which is distributed as $\text{Unif}[-2, 2]$. We display the simulation results in Table 2.2. The estimates in the relative risk model is bias under all the scenarios, and the 95% coverage rate is also bad.

Table 2.2: Monte Carlo simulation results based on 1000 runs for the proposed estimator which assumes monotonic treatment effects. The nuisance model is **misspecified**. The true values of γ_0 and γ_1 are 0 and 1, respectively

Sample Size		100	500	1000
Bias(SE)				
	γ_0	0.112(0.019)	0.113(0.004)	0.110(0.002)
	γ_1	0.011(0.022)	-0.033(0.004)	-0.037(0.002)
SD Accuracy				
	γ_0	0.963	0.977	1.027
	γ_1	0.964	1.009	0.970
Coverage				
	γ_0	0.907	0.690	0.492
	γ_1	0.945	0.928	0.885

SE, standard error.

SD Accuracy = Average estimated standard deviation/Monte Carlo standard deviation.

Nominal level = 95%.

2.4 Model Comparison

For a binary treatment, Dukes and Vansteelandt (2018) have showed a simple trick for having the doubly robust g-estimator for relative risks using existing generalized estimating equations software. In their method, first fit a model for exposure given confounders, e(V); second fit a gamma generalized linear model for Y with a log-link, $E(Y \mid V, e(V), Z) = \exp(\xi_0^{\mathrm{T}} V + \xi_1^{\mathrm{T}} V e(V) + \xi_2^{\mathrm{T}} V Z)$. ξ_2 is a doubly robust estimator even the Y-on-$e(V)$ is incorrectly specified.

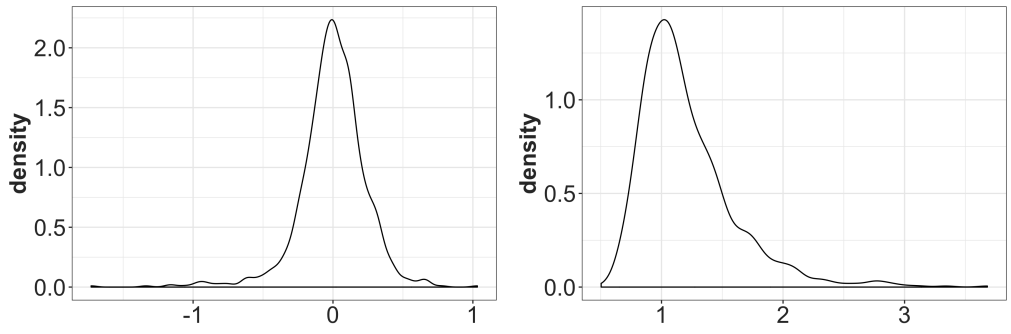
To compare it with our methods, we extend the framework of Dukes and Vansteelandt (2018) to a ordinal or continuous treatment. We use the following generating model: Continuous treatment Z is generated according to a multinomial logistic regression model such that

$$\log\left\{\frac{\mathrm{pr}(Z=1 \mid V)}{\mathrm{pr}(Z=0 \mid V)}\right\} = \eta_1^{\mathrm{T}} V, \quad \text{and} \quad \log\left\{\frac{\mathrm{pr}(Z=2 \mid V)}{\mathrm{pr}(Z=0 \mid V)}\right\} = \eta_2^{\mathrm{T}} V, \tag{2.17}$$

where $\eta_1 = (1, -1)^{\mathrm{T}}$, $\eta_2 = (1, -2)^{\mathrm{T}}$. The covariate vector V includes an intercept and a draw from a uniform distribution on $[-2, 2]$. Y is generated according to models (2.9) and (2.10), where $\gamma = (0, 1)^{\mathrm{T}}$, $\beta = (1, -0.5)^{\mathrm{T}}$, so that the relative risk is linear in z. We apply the method of §2.1 to estimate the relative risk in this setting, and compare it to the doubly robust g-estimator by Dukes and Vansteelandt (2018).

Table 2.3 summarizes the simulation results for sample size 500. The bias of our proposed estimators is small when the sample size is 500, and further decreases as the sample size increases; see the table in the Appendix B.2. The standard deviation accuracy is close to 1 for our proposed estimators. The coverage probability of the proposed Wald-type confidence intervals also achieve the nominal 95% coverage-rate. Even though, in theory, the doubly robust g-estimator is consistent in this setting as the propensity score model is correctly specified, as shown in Table 2.3, with a small sample of 500 the bias can be very large relative to the standard error. In this case, the model-based standard deviation estimate is also much smaller than the Monte Carlo standard deviation. Figure 2.3 displays a density estimate for the doubly robust g-estimator $\gamma = (\gamma_0, \gamma_1)^{\mathrm{T}}$. The estimator of γ_1 appears to not

be normally distributed. However, the bias of the doubly robust decreases as sample size increases; see Table B.2. One can also see that in this simulation, the proposed estimator is much more efficient than the g-estimator.



(a) γ_0 (b) γ_1

Figure 2.3: Probability density of doubly robust estimator by Dukes and Vansteelandt (2018) based on 500 samples and 1000 Monte Carlo runs.

2.5 Discussion

In this chapter, we have introduced a novel method to modeling the relative risks directly. The method has provided researchers an intuitive and straightforward way to interpret the relative risk between treatments/exposures varying with covariates. Our method assumes monotonic treatment effect and allows treatment/exposure to be ordinally categorical or continuous. We have proved that the primary of interest relative risk and the nuisance model odds product is variation independent. Such variation independence further leads to a smooth one-to-one mapping to a series proper probabilities $\{p_z(v), z \in \mathcal{Z}\}$ for all covariates.

When modeling the relative risks between exposures, researchers are supposed to decide a baseline treatment/exposure first. There are no strict rules to determine the baseline. It depends on researchers' study purpose. It is necessary to choose a function $h(v, z)$ for relative risks, which is monotonic in treatment z for all covariates z. In our simulation, we

Table 2.3: Simulation results for three different methods based on 500 samples and 1000 Monte Carlo runs. The true values for γ are $(0, 1)^{\mathrm{T}}$.

	$\text{Bias}_{\times 10^2}(\text{Standard Error}_{\times 10^2})$	SD Accuracy	Coverage (Nominal = 95%)
	γ	γ	γ
Monotone	0.232(0.214)	0.996	0.950
	0.442(0.256)	1.011	0.958
DR-G	−0.267(0.763)	0.659	0.890
	20.93(1.250)	0.558	0.816

Monotone, using models (2.9) and (2.10); DR-G, doubly robust estimator by Dukes and Vansteelandt (2018).

SD Accuracy = Average estimated standard deviation/Monte Carlo standard deviation.

have simply used a linear model. However, researchers can select their own model according to their background knowledge and data exploration.

The proposed assumption of monotonic treatment effects may be falsified from the observed data. In practice, analysts may use descriptive plots to examine the relationship between the treatment and outcome, and use them to assess plausibility of the monotonic treatment effect assumption. See §4 for an illustration.

Further work can be done on exploring the interaction between treatments/exposures. Our original work only focuses on modeling one treatment/exposures. For example, researchers are interested in studying the relative risk of having diabetes between the younger people with higher BMI and slightly older people with lower BMI. Therefore, one may find it important to model relative risks of treatment X and Z conditional on covariates V:

$$\text{RR}(z_0, z; x_0, x; v) = \frac{\text{pr}(Y = 1 \mid Z = z, X = x, V = v)}{\text{pr}(Y = 1 \mid Z = z_0, X = x_0, V = v)},$$

where X and Z are continuous with baseline x_0 and z_0. respectively. Building from the proposed method in §2.1, one simple way is to assume that $\text{RR}(z_0, z; x_0, x; v)$ is bounded in

z and x and monotonic in z and x respectively and to extend the odds product to

$$\textsc{op}(z_0, z; x_0, x; v) = \frac{p_{z_0,x_0}(v)}{1 - p_{z_0,x_0}(v)} \cdot \frac{p_{z,x}(v)}{1 - p_{z,x}(v)},$$

where $p_{z,x}(v) = \text{pr}(Y = 1 \mid Z = z, X = x, V = v)$. One may consider simple linear models:

$$\log\{\textsc{rr}(z_0, z; x_0, x; v)\} = \gamma_1^{\mathrm{T}} W(z - z_0) + \gamma_2^{\mathrm{T}} Q(x - x_0),$$
$$\log\{\textsc{op}(z_{\text{inf}}, z_{\text{sup}}; x_{\text{inf}}, x_{\text{sup}}; V)\} = \beta^{\mathrm{T}} H,$$

where $W = W(v)$, $Q = Q(v)$, $H = H(v)$, $z_{\text{sup}} = \sup\{z : z \in \mathcal{Z}\}$, $z_{\text{inf}} = \inf\{z : z \in \mathcal{Z}\}$, $x_{\text{sup}} = \sup\{x : x \in \mathcal{X}\}$, and $x_{\text{inf}} = \inf\{x : x \in \mathcal{X}\}$. We have to be careful when extending the method because this is an even stronger assumption which indicates the extreme points for both treatments/exposures are taken at their domains' boundary respectively.

Chapter 3

PARAMETERIZATION WITH A CATEGORICAL TREATMENT

3.1 Methodology

The approach introduced in Chapter 2 is not directly applicable if the relative risk is not monotonic in z. We now consider a categorical treatment whose effect on the outcome is not necessarily monotonic. Suppose that the treatment Z takes values in $\{z_0, \ldots, z_K\}$, where z_0 is chosen as the baseline treatment. The quantities of interest are relative risks $\text{RR}(z_0, z_k; v)$ $(k = 1, \ldots, K)$. For notational simplicity, we denote $\text{pr}(Y = 1 \mid Z = z_k, V = v)$ as $p_k(v)$, and $\text{RR}(z_0, z_k; v)$ as $\text{RR}(0, k; v)$. Following Wang et al. (2017), we introduce a nuisance model on the generalized odds product

$$\text{GOP}(v) = \prod_{k=0}^{K} \frac{p_k(v)}{1 - p_k(v)}. \tag{3.1}$$

The following theorem states that the generalized odds product is variation independent of the set of relative risks.

Theorem 2 (Variation independence with a categorical treatment)**.** *Let $\mathcal{M}$ denote a $(K+1)$-dimensional model on*

$$\text{RR}(0, k; v) = \frac{p_k(v)}{p_0(v)} \quad (k = 1, \ldots, K),$$
$$\text{GOP}(v) = \prod_{k=0}^{K} \frac{p_k(v)}{1 - p_k(v)}.$$

For any v, the map given by

$$(p_0(v), \ldots, p_K(v)) \to (\log \text{RR}(0, 1; v), \ldots, \log \text{RR}(0, K; v), \log \text{GOP}(v)) \tag{3.2}$$

is a diffeomorphism from $(0, 1)^{K+1}$ to $(\mathbb{R})^{K+1}$. Furthermore, the models in $\mathcal{M}$ are variation independent of each other.

Proof. In order to prove the map given by

$$(p_0(v), \ldots, p_K(v)) \to (\log \text{RR}(v; 0, 1), \ldots, \log \text{RR}(v; 0, K), \log \text{GOP}(v))$$

is a diffeomorphism, we need to prove that for any fixed v and any vector

$$(\text{RR}(0,1;v),\ldots,\text{RR}(0,K;v),\text{GOP}(v)) \in (\mathbb{R}^+)^{K+1},$$

there is one and only one vector $(p_0(v),\ldots,p_K(v)) \in (0,1)^{K+1}$. Let $\text{RR}(0,k;v) = c_k(v) \in \mathbb{R}^+$ where $k = 1,\ldots,K$, and $\text{GOP}(v) = c_{K+1}(v) \in \mathbb{R}^+$. By definition of $\text{GOP}(v)$, we further have

$$\log\{c_{K+1}(v)\} = (K+1)\log\{p_0(v)\} + \sum_{k=1}^{K}\log\{c_k(v)\} - \log\{1-p_0(v)\} - \sum_{k=1}^{K}\log\{1-p_0(v)c_k(v)\}. \tag{3.3}$$

In the following, we show that there is one and only one solution of Equation (3.3) for $p_0(v) \in (0,1)$. For notational simplicity, write $p_k(v)$ as p_k, and $c_k(v)$ as c_k, $k = 0,1,\ldots,K+1$. Let $f(p_0) = (K+1)\log p_0 + \sum_{k=1}^{K}\log(c_k) - \log(1-p_0) - \sum_{k=1}^{K}\log(1-p_0c_k) - \log(c_{K+1})$. Now

$$\begin{aligned}\frac{df(p_0)}{dp_0} &= \frac{K+1}{p_0} + \frac{1}{1-p_0} + \sum_{k=1}^{K}\frac{c_k}{1-p_0c_k} \\ &= \frac{K+1}{p_0} + \frac{1}{1-p_0} + \sum_{k=1}^{K}\frac{c_k}{1-p_k} > 0.\end{aligned}$$

Therefore $f(p_0)$ is monotonically increasing on $(0,1)$. Because of $\lim_{p_0\to 0} f(p_0) = -\infty$ and $\lim_{p_0\to 1} f(p_0) = +\infty$, there is one and only one root for $f(p_0) = 0$ on $(0,1)$. Since the domain of $\mathcal{M}$, $(\mathbb{R}^+)^{K+1}$, is the Cartesian product of the marginal domains of the Relative Risk and Generalized Odds Product models, the models in $\mathcal{M}$ are variation independent. □

In our simulations and data analysis, we consider the following specifications of $\mathcal{M}$:

$$\log\{\text{RR}(0,k;v)\} = \alpha_k^{\mathrm{T}}X \quad (k = 1,\ldots,K), \tag{3.4}$$

$$\log\{\text{GOP}(v)\} = \beta^{\mathrm{T}}W, \tag{3.5}$$

where $X = X(v)$, $W = W(v)$. Theorem 2 shows that the parameters $\alpha_1,\ldots,\alpha_K$, and β are variation independent so that their domains are unconstrained. Maximum likelihood estimates and associated inference for parameters $\alpha_1,\ldots,\alpha_K$, and β can then be obtained in standard fashion. The relative risk model in this approach is more flexible than the

corresponding model (2.9) in Chapter 2, which assumes monotonicity, thus (3.4) has K-times as many parameters.

3.2 Variance Formula

Suppose we observe a unit in treatment arm z_k. Denote $\theta_k = \alpha_k^{\mathrm{T}} v$, , $g = \beta^{\mathrm{T}} v$. Then the first derivatives of $l(\alpha_1, \ldots, \alpha_K, \beta \mid z, v, y)$ with respect to $\alpha_1, \ldots, \alpha_K, \beta$ are

$$\frac{\partial l}{\partial \alpha_j} = \frac{y}{p_k}\frac{\partial p_k}{\partial \alpha_j} - \frac{1-y}{1-p_k}\frac{\partial p_k}{\partial \alpha_j} = \frac{y-p_k}{p_k(1-p_k)}\frac{\partial p_k}{\partial \alpha_j} \quad (k = 0, 1, \ldots, K; j = 1, \ldots, K), \tag{3.6}$$

$$\frac{\partial l}{\partial \beta} = \frac{y-p_k}{p_k(1-p_k)}\frac{\partial p_k}{\partial \beta}. \tag{3.7}$$

Since $\partial p_k / \partial \alpha_j = \partial (p_0 e^{\theta_k}) / \partial \alpha_j$, we further have

$$\frac{\partial p_k}{\partial \alpha_j} = \begin{cases} \frac{\partial p_0}{\partial \alpha_j} e^{\theta_k} & k \neq 0, k \neq j \\ \frac{\partial p_0}{\partial \alpha_j} e^{\theta_j} + p_j v & k \neq 0, k = j \\ \frac{\partial p_0}{\partial \alpha_j} & k = 0 \end{cases}; \tag{3.8}$$

$$\frac{\partial p_k}{\partial \beta} = \frac{\partial p_0}{\partial \beta} e^{\theta_k}. \tag{3.9}$$

In order to calculate Eq. (3.6) and (3.7), we need to have $\frac{\partial p_0}{\partial \alpha_j}$ and $\frac{\partial p_0}{\partial \beta}$. By definition we have

$$e^{\phi} = \frac{\prod_{k=0}^{K} p_k}{\prod_{k=0}^{K}(1-p_k)}.$$

Taking the logarithm of the both sides gives

$$\phi = \sum_{k=0}^{K} \log p_k - \sum_{k=0}^{K} \log(1-p_k). \tag{3.10}$$

The derivatives of both sides of (3.10) with respect to α_j and β, respectively, are:

$$0 = \frac{1}{p_0}\frac{\partial p_0}{\partial \alpha_j}\left(\sum_{k=0}^{K}\frac{1}{1-p_j}\right) + \frac{1}{1-p_j} v, \tag{3.11}$$

$$v = \frac{1}{p_0}\frac{\partial p_0}{\partial \beta}\left(\sum_{k=0}^{K}\frac{1}{1-p_j}\right). \tag{3.12}$$

By (3.11) and (3.12), we further have

$$\frac{\partial p_0}{\partial \alpha_j} = -\frac{v \cdot \frac{p_0}{1-p_j}}{\sum_{k=0}^{K} \frac{1}{1-p_j}}, \tag{3.13}$$

$$\frac{\partial p_0}{\partial \beta} = \frac{p_0 v}{\sum_{k=0}^{K} \frac{1}{1-p_j}}. \tag{3.14}$$

Substituting (3.13) and (3.14) into (3.6) to (3.9), we have

$$\frac{\partial l}{\partial \alpha_j} = \begin{cases} \frac{v(y-p_k)}{1-p_k} \frac{-\frac{1}{1-p_j}}{\sum_{l=0}^{K} \frac{1}{1-p_l}} & k \neq 0, k \neq j \\ \frac{v(y-p_k)}{1-p_k} \left(1 - \frac{\frac{1}{1-p_j}}{\sum_{l=0}^{K} \frac{1}{1-p_l}}\right) & k \neq 0, k = j \\ \frac{v(y-p_k)}{1-p_k} \frac{-\frac{1}{1-p_j}}{\sum_{l=0}^{K} \frac{1}{1-p_l}} & k = 0 \end{cases},$$

$$\frac{\partial l}{\partial \beta} = \frac{(y-p_k)v}{1-p_k} \frac{1}{\sum_{l=0}^{K} \frac{1}{1-p_l}}.$$

The variance-covariance matrix for $(\alpha_1, \ldots, \alpha_K, \beta)$ can be calculated as the inverse of the Fisher Information matrix.

3.3 Simulation Studies

We conduct a similar set of simulations as in §2.3. Categorical treatment Z is randomly generated from $\{0, 1, 2\}$ with an equal probability. Outcome Y is generated following models (3.4) and (3.5), where $\alpha_1 = (-0.5, 1)^{\mathrm{T}}$, $\alpha_2 = (0.5, 1.5)^{\mathrm{T}}$, $\beta = (1, -0.5)^{\mathrm{T}}$. The simulation results are based on 1000 Monte Carlo runs and shown in Table 3.2. We conduct the simulations under different sample size ($N = 100, 500, 1000, 5000$); see more in §C.1. As theory predicts, the bias of our point and variance estimators goes to zero as sample size increases. Although at sample size 100, the estimated standard deviation and the coverage rate of α_1 are biased upwards, the bias decreases as sample size becomes larger; see simulation results in Appendix §C.1. To understand the bias in estimates and standard deviations under sample size 100, we display the empirical distribution of parameter estimates via histograms in Figure 3.1 and the descriptive statistics in Table 3.1. Distribution histograms in Figure 3.1 show that when sample size is 100, the estimates are skewed especially for α_{22}. It explains

why the bias is relatively large for α_{22} in Table 3.2. We also notice that the standard deviation accuracy of $\alpha_1 = (\alpha_{11}, \alpha_{12})^{\mathrm{T}}$ is much larger than 1 for sample size 100. The Monte Carlo standard deviations for α_{11} and α_{12} are 0.574 and 0.561 respectively. However, there are extremely large model-based estimates of the standard deviation appearing in our simulations; see Table 3.1. That is likely due to numerical instability at this comparatively small sample size.

Table 3.1: Quantiles of model-based estimates of the standard deviation of $\alpha_1 = (\alpha_{11}, \alpha_{12})$

Quantiles	0%	25%	50%	75%	100%
SD(α_{11})	0.237	0.402	0.464	0.559	477.364
SD(α_{12})	0.258	0.429	0.512	0.644	397.217

We also consider the scenario when the nuisance model is misspecified. An analyst takes variable V^* instead of V in the nuisance model to estimate parameters of interest. V^* includes an intercept and an irrelevant covariate which is distributed as Unif$[-2, 2]$. We display the simulation results in Appendix C.2. Note that if the quantity of nuisance model is small, it barely has an impact on the estimates even under misspecification; see Table C.2 and Table C.3 for comparison where β varies.

3.4 Model Comparison

With a categorical treatment taking $K + 1$ levels, a naive alternative is to apply a method designed for modeling the relative risk for a binary treatment K times. However, the resulting relative risk models will not necessarily be compatible.

In this way we compare our proposed generalized odds product method to two previously proposed relative risk models for binary treatment: the likelihood method proposed by Richardson et al. (2017) and the doubly robust g-estimator of Dukes and Vansteelandt (2018).

Similarly as in §2.4, covariates V include an intercept and a random variable uniformly

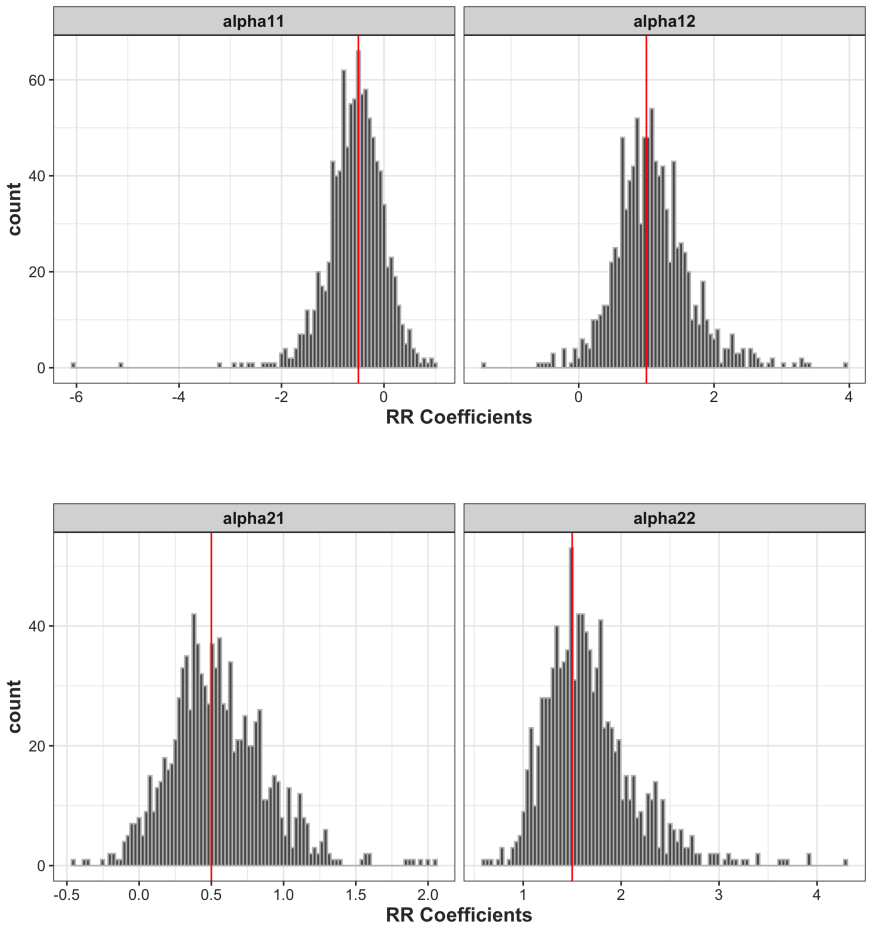


Figure 3.1: Empirical distribution of parameter estimates obtained from 1000 Monte-Carlo runs with sample size 100. The top row shows the histograms of estimates for $\alpha_1 = (\alpha_{11}, \alpha_{12})$; The bottom are estimates of $\alpha_2 = (\alpha_{21}, \alpha_{22})$. The red vertical lines represent the ground truth.

Table 3.2: Monte Carlo simulation results based on 1000 runs for the relative risk model with a generalized odds product nuisance model. The true values for vectors α_1 and α_2 are $(-0.5, 1)^{\mathrm{T}}$ and $(0.5, 1.5)^{\mathrm{T}}$ respectively

Sample Size	100		1000	
	α_1	α_2	α_1	α_2
$\mathrm{Bias}_{\times 10^2}(\mathrm{SE}_{\times 10^2})$				
	-6.077(1.816)	4.529(1.056)	0.367(0.41)	0.721(0.284)
	9.054(1.775)	17.239(1.468)	1.571(0.435)	1.767(0.361)
SD Accuracy				
	1.759	1.004	1.026	1.007
	1.797	1.003	1.017	1.024
Coverage				
	0.975	0.954	0.956	0.954
	0.971	0.973	0.953	0.955

SE, standard error.

SD Accuracy = Average estimated standard deviation /Monte Carlo standard deviation.

Nominal level = 95%.

generated from $[-2, 2]$, categorical treatment $Z \in \{0, 1, 2\}$ is generated according to a multinomial logistic regression model:

$$\log\left\{\frac{\text{pr}(Z=1 \mid V)}{\text{pr}(Z=0 \mid V)}\right\} = \eta_1^{\mathrm{T}} V, \quad \text{and} \quad \log\left\{\frac{\text{pr}(Z=2 \mid V)}{\text{pr}(Z=0 \mid V)}\right\} = \eta_2^{\mathrm{T}} V, \tag{3.15}$$

where $\eta_1 = (1, -1)^{\mathrm{T}}$, $\eta_2 = (1, -2)^{\mathrm{T}}$. Outcome Y is generated from models (3.4) and (3.5), where vectors α_1, α_2, β are $(-0.5, 1)^{\mathrm{T}}$, $(0.5, 1.5)^{\mathrm{T}}$, and $(1, -0.5)^{\mathrm{T}}$.

To apply the methods in Richardson et al. (2017) and Dukes and Vansteelandt (2018), we use the subset of units with $Z \in \{0, 1\}$ to estimate α_1 and the subset with $Z \in \{0, 2\}$ to estimate α_2.

For Richardson et al. (2017)'s method, we assume that

$$\log\{\text{OP}(0, 1; V)\} = \beta_1^{\mathrm{T}} V,$$
$$\log\{\text{OP}(0, 2; V)\} = \beta_2^{\mathrm{T}} V,$$

where $\text{OP}(0, 1; V)$ is the odds products of treatment 0 and 1 conditional on covariates V and so is $\text{OP}(0, 2; V)$. In general, these odds product models will be incompatible with the models for $\text{RR}(0, 1; v)$, $\text{RR}(0, 2; v)$ as they are variation dependent. For the method of Dukes and Vansteelandt (2018), we assume the propensity score model as (3.15) and and a baseline model for outcome Y, $E(Y \mid Z = 0, V) = \exp(\xi^{\mathrm{T}} V)$.

Table 3.3 shows the simulation results for sample sizes 500 and 1000; see more simulations for larger sample size in Appendix Table C.4. The biases of our point and variance estimators are small and, in addition, go to zero as sample size increases. Although the bias of the doubly robust g-estimator is very large, the bias decreases as sample size increases. As expected, two applications of the likelihood method of Richardson et al. (2017) yields biased estimates as the odds product models are misspecified. Similar to the performance reported in Table 2.3, two applications of the doubly-robust g-estimator by Dukes and Vansteelandt (2018) yield results that are consistent but not efficient.

Table 3.3: Simulation results for three different methods based on 500, and 1000 samples and 1000 Monte Carlo runs. The true values for α_1 and α_2 are $(-0.5, 1)^{\mathrm{T}}$ and $(0.5, 1.5)^{\mathrm{T}}$ respectively

Sample Size	Bias$_{\times 10^2}$(Standard Error$_{\times 10^2}$)		SD Accuracy		Coverage	
	α_1	α_2	α_1	α_2	α_1	α_2
500	GOP					
	0.612(0.626)	1.735(0.451)	1.007	1.008	0.957	0.960
	0.080(0.669)	−0.490(0.463)	1.011	0.988	0.961	0.951
	DR-G (applied twice)					
	−8.251(0.988)	3.030(0.565)	0.828	0.939	0.922	0.947
	20.80(1.943)	14.98(1.224)	0.653	0.695	0.885	0.933
	OP (applied twice)					
	0.146(0.638)	2.129(0.457)	0.998	1.011	0.961	0.959
	1.913(0.764)	−1.618(0.47)	0.971	1.008	0.956	0.955
1000	GOP					
	−0.433(0.434)	0.626(0.314)	1.005	0.996	0.964	0.946
	0.519(0.456)	−0.124(0.314)	1.026	0.989	0.963	0.952
	DR-G (applied twice)					
	−4.705(0.605)	1.020(0.363)	0.878	0.962	0.936	0.947
	9.834(1.143)	6.266(0.715)	0.733	0.722	0.911	0.947
	OP (applied twice)					
	−0.531(0.443)	1.106(0.320)	1.002	0.994	0.963	0.952
	1.112(0.523)	−1.251(0.321)	0.991	1.002	0.956	0.947

GOP: Using models (9) and (10); DR-G, doubly robust estimator by Dukes and Vansteelandt (2018); OP: Using nuisance model proposed by Richardson et al. (2017).

SD Accuracy = Average estimated standard deviation / Monte Carlo standard deviation.

95% nominal coverage.

3.5 A note on Bayesian analysis

In this section, we provide a simple Bayesian framework to estimate the parameters of interest of the proposed model in §3.1,

$$\log\{\text{RR}(0,k;V)\} = \alpha_k^{\mathrm{T}} V \quad (k=1,\ldots,K),$$
$$\log\{\text{GOP}(V)\} = \beta^{\mathrm{T}} V.$$

We put non-informative priors on $\alpha_1,\ldots,\alpha_K,\beta$. Let $\alpha_1,\ldots,\alpha_K,\beta$ be multinormal distributed with a large standard deviation, respectively, e.g.

$$\alpha_k \sim \mathcal{N}_p\left(0, D_p(10^2,\ldots,10^2)\right) \ (k=1,\ldots,K),$$
$$\beta \sim \mathcal{N}_p\left(0, D_p(10^2,\ldots,10^2)\right),$$

where D_p represents a $p\times p$ diagonal matrix. The prior probability of the parameters is

$$p(\alpha_1,\ldots,\alpha_K,\beta) = \prod_{k=1}^{K} p(\alpha_k)\cdot p(\beta). \tag{3.16}$$

The observed data $Y_i,\ldots,Y_n$ are independent and have a binomial distribution

$$Y_i \mid \alpha_1,\ldots,\alpha_K,\beta \sim B\left(p_{z_i}(v_i;\alpha_1,\ldots,\alpha_K,\beta)\right).$$

Therefore, the probability of observations, or the likelihood, is

$$p(y_1,\ldots,y_n \mid \alpha_1,\ldots,\alpha_K,\beta) = \prod_{i=1}^{n} y_i^{p_{z_i}(v_i;\alpha_1,\ldots,\alpha_K,\beta)}\cdot(1-y_i)^{1-p_{z_i}(v_i;\alpha_1,\ldots,\alpha_K,\beta)}. \tag{3.17}$$

By Bayes' theorem we further have the posterior distribution of parameters by multiplying Eq.(3.16) and Eq.(3.17),

$$p(\alpha_1,\ldots,\alpha_K,\beta \mid y_1,\ldots,y_n) \propto p(\alpha_1,\ldots,\alpha_K,\beta)\cdot p(y_1,\ldots,y_n \mid \alpha_1,\ldots,\alpha_K,\beta). \tag{3.18}$$

Simulation Studies

We conduct a finite sample simulation ($n=500$) to evaluate our proposed Bayesian method. We generate treatment Z from a uniform distribution on $\{0,1,2\}$ and $Z=0$ is the baseline

treatment. The covariates V include an intercept and a random variable generated from a uniform distribution on the interval $[-2, 2]$. We generate Y from above models, where $\alpha_1 = (-0.5, 1)^{\mathrm{T}}$, $\alpha_2 = (0.5, 1.5)^{\mathrm{T}}$, $\beta = (1, -0.5)^{\mathrm{T}}$. The algorithm of Metropolis hasting is used; see Algorithm 1.

Algorithm 1 Parameter estimation via Metropolis Hastings

Initiate $\xi^{(0)\mathrm{T}} = (\alpha_1^{(0)}, \alpha_2^{(0)}, \beta^{(0)})^{\mathrm{T}} \leftarrow (0, 0, 0, 0, 0, 0)$

Set $t \leftarrow 0$

while $t < T$ **do**

  for $j \leftarrow 1$ to 6 **do**

    Generate $\xi_j' \sim \mathcal{N}\left(\xi_j^{(t)}, 0.1\right)$

    Calculate $R(\xi_j', \xi_j^{(t)}) \leftarrow \frac{p(\xi_1^{(t)},\ldots,\xi_j',\ldots,\xi_6^{(t)}) \cdot p(y_1,\ldots,y_n|\xi_1^{(t)},\ldots,\xi_j',\ldots,\xi_6^{(t)})}{p(\xi_1^{(t)},\ldots,\xi_j^{(t)},\ldots,\xi_6^{(t)}) \cdot p(y_1,\ldots,y_n|\xi_1^{(t)},\ldots,\xi_j^{(t)},\ldots,\xi_6^{(t)})}$

    Generate $U \sim \mathrm{Unif}(0, 1)$

    if $U < R(\xi_j', \xi_j^{(t)})$ **then**

      $\xi_j^{(t+1)} \leftarrow \xi_j'$

    else

      $\xi_j^{(t+1)} \leftarrow \xi_j^{(t)}$

    end if

  end for

  $t \leftarrow t + 1$

end while

We have 11000 iterations in total and discard the first 1000 as the burn-in. The following figures display the posterior distributions. Figure 3.2 displays the trace plot. It is relatively constant and shows that our MCMC has reached stationarity. Figure 3.3 shows the posterior distributions for α_1, α_2, and β respectively. The posterior distributions of each entry of α_1 and α_2 are normally distributed and centered around the ground truth. The posterior mean and 95% credible interval are showed in Table 3.4. The point estimates are close to the

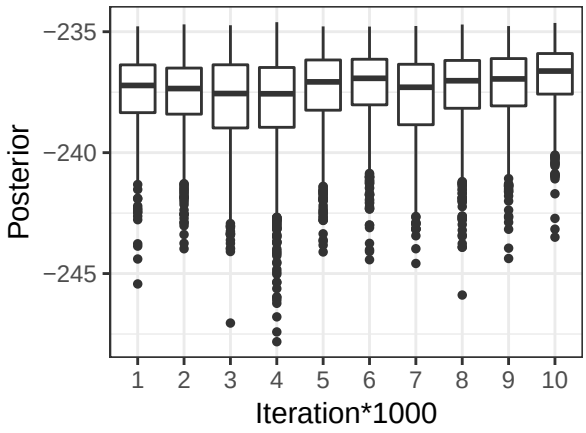


Figure 3.2: Trace Plot. Every 1000 iterations are grouped and showed by boxplots.

ground truth and their 95% credible intervals cover the ground truth too, which suggest that the Bayesian estimation works well. The estimated α_1, α_2 from the frequentist view in §3.3 are $(-0.425, 1.007)^{\mathrm{T}}$ and $(0.475, 1.392)^{\mathrm{T}}$ respectively, which are also close to the posterior mean from Bayesian analysis. The standard deviation from §3.3 and the Bayesian method are closed too.

Table 3.4: Posterior estimation results based on 500 samples via MH samplings. The true values for α_1 and α_2 are $(-0.5, 1)^{\mathrm{T}}$ and $(0.5, 1.5)^{\mathrm{T}}$ respectively.

	α_{11}	α_{12}	α_{21}	α_{22}
Estimate	−0.414	1.041	0.491	1.437
95% CI	(−0.783, −0.083)	(0.715, 1.391)	(0.243, 0.756)	(1.154, 1.740)

Estimates are the posterior mean.

CI is credible interval.

One may also implement more efficient MCMC procedures, such as block Metropolis Hastings. In more detail, we would initialize $\alpha_1^{(0)}$, $\alpha_2^{(0)}$, and $\beta^{(0)}$ at the MLEs $\hat{\alpha}_1$, $\hat{\alpha}_2$, and $\hat{\beta}$.

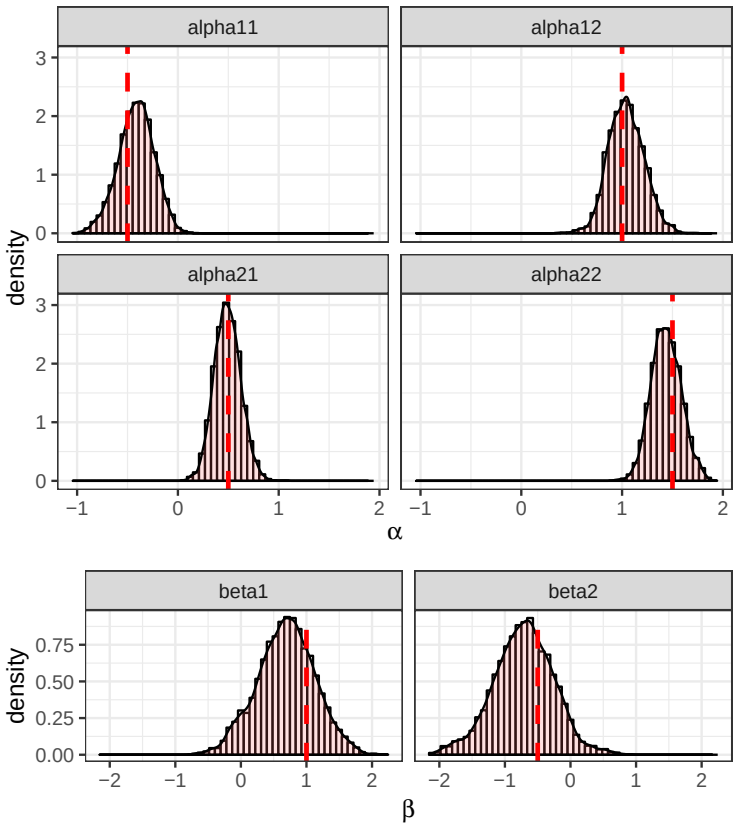


Figure 3.3: Posterior distributions for α_1, α_2, and β. The red dashed lines show where the true values are.

We then iterate, at iteration t, between:

1. Generate $\alpha_1' \sim \mathcal{N}(\alpha_1^{(t)}, c_1\hat{V}_1)$, where $\hat{V}_1$ is the estimated variance of $\hat{\alpha}_1$, from the inverse of the Hessian matrix. Calculate the acceptance probability:

$$R(\alpha_1', \alpha_1^{(t)}) = \min\left\{\frac{\text{pr}(\alpha_1', \alpha_2^{(t)}, \beta^{(t)} \mid \mathbf{y})}{\text{pr}(\alpha_1^{(t)}, \alpha_2^{(t)}, \beta^{(t)} \mid \mathbf{y})}, 1\right\},$$

and set

$$\alpha_1^{(t+1)} = \begin{cases} \alpha_1' & \text{with probability } R(\alpha_1', \alpha_1^{(t)}) \\ \alpha_1^{(t)} & o.w. \end{cases}$$

2. Generate $\alpha_2' \sim \mathcal{N}(\alpha_2^{(t)}, c_2\hat{V}_2)$, where $\hat{V}_2$ is the estimated variance of $\hat{\alpha}_2$. Calculate the acceptance probability:

$$R(\alpha_2', \alpha_2^{(t)}) = \min\left\{\frac{\text{pr}(\alpha_1^{(t+1)}, \alpha_2', \beta^{(t)} \mid \mathbf{y})}{\text{pr}(\alpha_1^{(t+1)}, \alpha_2^{(t)}, \beta^{(t)} \mid \mathbf{y})}, 1\right\},$$

and set

$$\alpha_2^{(t+1)} = \begin{cases} \alpha_2' & \text{with probability } R(\alpha_2', \alpha_2^{(t)}) \\ \alpha_2^{(t)} & o.w. \end{cases}$$

3. Generate $\beta' \sim \mathcal{N}(\beta^{(t)}, c_3\hat{V}_3)$, where $\hat{V}_3$ is the estimated variance of $\hat{\beta}$. Calculate the acceptance probability:

$$R(\beta', \beta^{(t)}) = \min\left\{\frac{\text{pr}(\alpha_1^{(t+1)}, \alpha_2^{(t+1)}, \beta' \mid \mathbf{y})}{\text{pr}(\alpha_1^{(t+1)}, \alpha_2^{(t+1)}, \beta^{(t)} \mid \mathbf{y})}, 1\right\},$$

and set

$$\beta^{(t+1)} = \begin{cases} \beta' & \text{with probability } R(\beta', \beta^{(t)}) \\ \beta^{(t)} & o.w. \end{cases}$$

The constants c_1, c_2, and c_3 are chosen to provide a trade-off between gaining a high proportion of acceptances, and moving around the support of the parameter space.

3.6 Discussion

For circumstances where this monotonicity assumption is not appropriate, we propose an alternative approach that involves a novel generalized odds product model. This approach

applies to a categorical treatment variable, and, in general, requires estimation of a larger number of parameters.

With categorical treatments, it is more straightforward to extend this method to modeling the combination relative effect. Assume multi-level treatments X and Z taking value in $\{x_0, \ldots, x_L\}, L \in \mathbb{N}^+$ and $\{z_0, \ldots, z_K\}, K \in \mathbb{N}^+$, respectively. Recall the relative risks of combination treatments X and Z in §2.5:

$$\textsc{rr}(z_0, z; x_0, x; v) = \frac{\text{pr}(Y = 1 \mid Z = z, X = x, V = v)}{\text{pr}(Y = 1 \mid Z = z_0, X = x_0, V = v)},$$

where without loss of generality, we take $Z = z_0$ and $X = x_0$ as the baseline, respectively. One may simply treat (Z, X) as a new grouped treatment, and there are $K \times L$ different levels of treatments in total. The corresponding nuisance model becomes:

$$\textsc{gop}(v) = \prod_{k=0}^{K} \prod_{l=0}^{L} \frac{\text{pr}(Y = 1 \mid Z = z_k, X = x_l, V = v)}{1 - \text{pr}(Y = 1 \mid Z = z_k, X = x_l, V = v)}. \tag{3.19}$$

For notation simplicity, denote $\textsc{rr}(z_0, z_k; x_0, x_l; v)$ as $\textsc{rr}(0, k; 0, l; v)$, $\text{pr}(Y = 1 \mid Z = z_k, X = x_l, V = v)$ as $p_{kl}(v)$. Theorem 2 can also be extended to the grouped treatment:

Theorem 3. *Let $\mathcal{M}$ denote the $(K+1)(L+1)$ dimensional models consisting of*

$$\textsc{rr}(0, k; 0, l; v) = \frac{p_{kl}(v)}{p_{00}(v)} \quad (k = 0, \ldots, K; l = 0, \ldots, L),$$
$$\textsc{gop}(v) = \prod_{k=0}^{K} \prod_{l=0}^{L} \frac{p_{kl}(v)}{1 - p_{kl}(v)}.$$

Then for any v, the map

$$(p_{00}(v), \ldots, p_{km}(v)) \to (\textsc{rr}(v; 0, 1; 0, 0), \ldots, \textsc{rr}(0, K; 0, L; v), \textsc{gop}(v)) \tag{3.20}$$

is a bijection from $(0,1)^{(K+1)(L+1)}$ to $(\mathbb{R}^+)^{(K+1)(L+1)}$. Models in $\mathcal{M}$ are variation independent of each other.

One can specify parametric forms for relative risks and the nuisance model:

$$\log\{\textsc{rr}(z_0, z_k; x_0, x_l; V)\} = \alpha_{kl}^{\mathrm{T}} W \quad (k = 0, \ldots, K; l = 0, \ldots, L), \tag{3.21}$$
$$\log\{\textsc{gop}(V)\} = \beta^{\mathrm{T}} X, \tag{3.22}$$

where $W = W(V), X = X(V)$, and $\alpha_{00} = 0$. With Model (3.21) and (3.22), the log-likelihood for observations can be written. Then the estimates and their standard deviation can obtained as we have showed before.

In this Chapter, we also provide a Bayesian framework for estimating the parameters under the GOP model. Researchers are able to incorporate their substantial background knowledge to the prior distribution under the Bayesian framework. With the posterior distributions of all the parameters, one can also easily understand how the estimates are distributed under a finite population and make inference based on their own scientific interest. A similar Bayesian analysis can also be done under the setting in §2.1 where the monotonic treatment effect is assumed.

Chapter 4

APPLICATION: TITANIC DATA

4.1 Introduction

In 1912, a British passenger liner, Titanic, sank in the North Atlantic Ocean after it struck an iceberg. Of the estimated 2,224 passengers and crew abroad, more than 1,500 died (Wikipedia contributors, 2020). There were three different passenger classes, First, Second, and Third, locating from the top deck to the bottom deck. The lifeboats were housed in the top deck. Many newspapers reported the tragedy back then, and the majority of them empathised that women and children had survived.

We have accessed Titanic data set. The data set consists of 1,309 passengers from three passenger classes, of whom 809 (61.803%) lost their lives during the event. We find that the death rate among the passenger classes shows a clear pattern. The third Class passengers more likely to die compared with the First Class passengers.

In this Chapter, we illustrate the use of our proposed methods in Chapter 2 and Chapter 3 by studying the association between the passenger class and death/survival status in the tragic sinking of the Titanic. We compare the results from our proposed models with those obtained from generalized linear models.

Data Exploration

Among 1,309 passengers, 263 (20.1%) miss age. For illustration we removed the 263 passengers, resulting in a sample size of 1,046, including 284 (27.1%) passengers in the First Class, 261 (25.0%) in the Second Class, and 501 (47.9%) in the Third Class. Table 4.1 shows that the empirical probability of death is lowest in the First Class at 36.27%, increasing to 55.94% in the Second Class, and 73.85% in the Third Class. Given this, we initially consider modeling the relative risk of death as a monotone function of passenger class, using the First Class as the baseline. On the other hand, the First Class has the highest chance to survive, then the Second Class, and finally the Third. We can also have a model for relative survival, using the Third Class as the baseline. Figure 4.1 shows the survival statuses of passengers by their passenger class, age and sex. Female passengers tend to have lower probability of

death compared to males, and children tend to have lower probability of death compared to adults. These observations suggest that the relative risk of death with respect to passenger class may vary with sex and age.

Table 4.1: Descriptive statistics of passengers' class and survival status

Class	Death	N(%)
First	No	181(63.73%)
First	Yes	103(36.27%)
Second	No	115(44.06%)
Second	Yes	146(55.94%)
Third	No	131(26.15%)
Third	Yes	370(73.85%)

4.2 Relative Risk of Death

We have attempted to show the empirical probability of death after stratifying by passenger class, sex, and age. However, considering the continuity of age, we divide it into three categories: < 20 years old, $20 \sim 45$ years old, and > 45 years old. Figure 4.2 displays the empirical probability of death. For females, the relative risk of death for the Third Class (against the First Class) is highest in the $20 \sim 45$ age group; For males, the highest relative risk of death for the Third Class is in the < 20 age group. Such observations suggests an interaction between age and sex when modeling relative risks. We let the outcome Y be 1 if the passenger died and 0, otherwise; let the covariates X and W be identical, which include age, sex, age squared, and the interaction between age and sex.

We apply four different models to estimate the variation in the relative risk of death stratifying on age and sex: 1) Poisson regression; 2) Logistic regression; 3) Monotone: the model given by (2.9) and (2.10); 4) GOP: the model given by (3.4) and (3.5). Results for

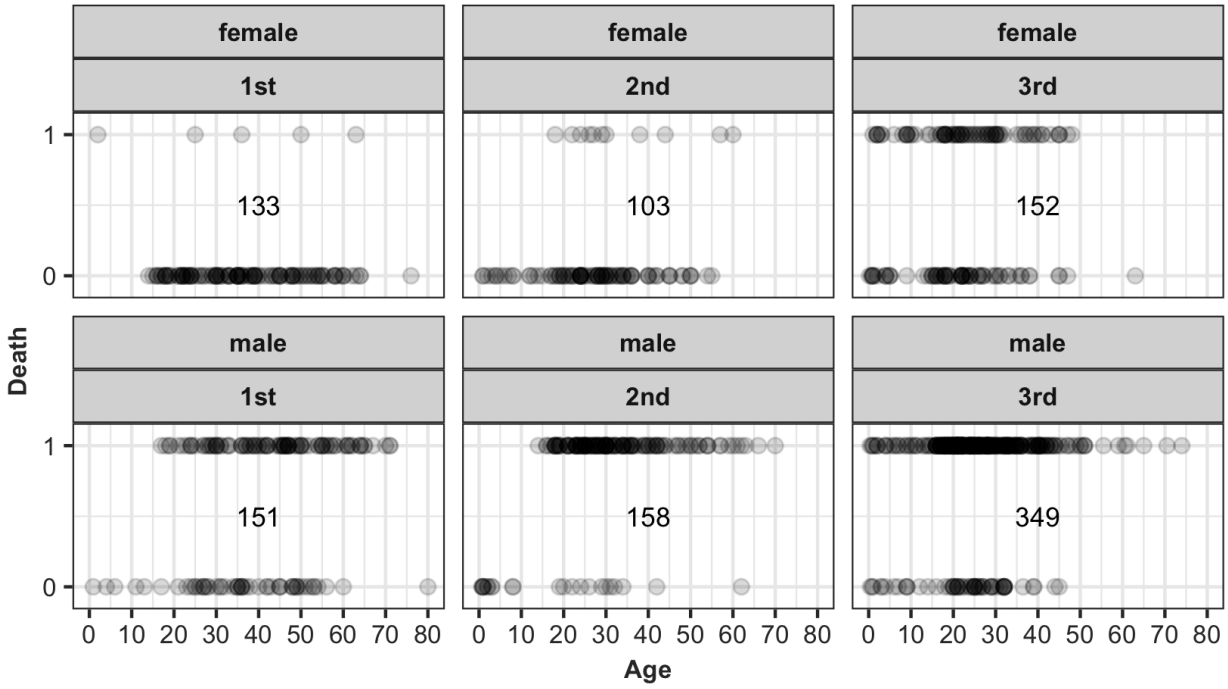


Figure 4.1: Passengers' survival statuses by passenger class, age, and sex. The number of passengers in each group is shown in the center of the corresponding plot.

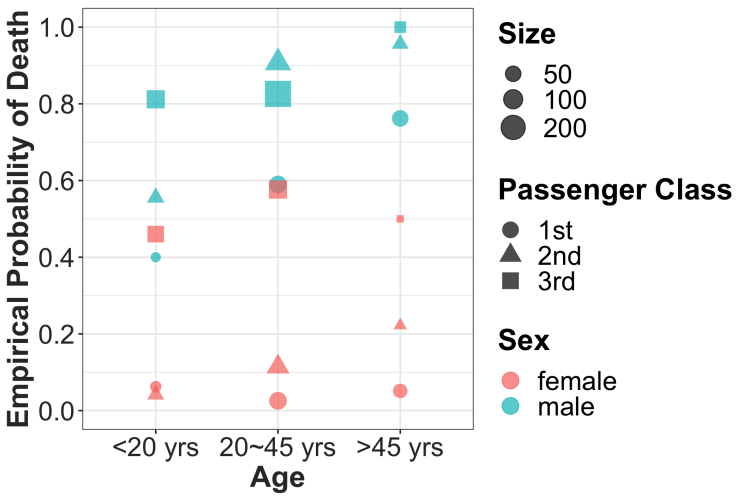


Figure 4.2: Empirical probability of death after stratifying by passenger class, sex, and age. Different color represents different sex; Different shape represents different passenger class; The size of the dots represents the number of people in the certain group.

models 1) and 2) were obtained using the `glm` function in `R` via maximum likelihood with robust standard errors. Table 4.2 shows regression coefficient estimates from our proposed methods and Poisson regression. Coefficient estimates for logistic regression are not shown as they do not directly describe the dependence of the relative risk of death on age and sex. The point estimates from our GOP model are close to those from the Poisson model, while the standard errors are smaller. On the other hand, point estimates for our Monotone model are different from those given by the other two models. Although it appears reasonable from the marginal death rates in each passenger class, the monotonic treatment effects assumption is probably violated after stratifying by age and sex. For example, for males from 25 to 57 years old, the empirical probability of death is 62.8% for the first class , 93.0% for the second class, and 82.9% for the third class.

Figure 4.3 displays the fitted probabilities of death from different models. For male passengers in the second class aged between 30 and 50, the fitted probability of death using the Poisson model does not lie in the valid range $[0, 1]$. Under the logistic regression model the

Table 4.2: Coefficient estimates via different models.

	2nd	2nd* male	2nd* age/10	2nd* age^2/ 100	2nd* male* age/10	3rd	3rd* male	3rd* age/10	3rd* age^2/ 100	3rd* male* age/10
Point Estimate										
Monotone	1.891	−1.543	−0.165	0.011	0.058	3.782	−3.086	−0.329	0.022	0.116
GOP	−1.134	1.439	0.780	−0.033	−0.617	2.204	−1.212	0.053	0.020	−0.309
Poisson	−1.211	0.938	0.969	−0.072	−0.487	2.232	−1.444	0.120	0.005	−0.254
Standard Deviation										
Monotone	0.396	0.407	0.124	0.010	0.107	0.792	0.813	0.247	0.020	0.214
GOP	1.230	1.251	0.369	0.029	0.314	0.888	0.957	0.260	0.021	0.236
Poisson	2.077	1.967	0.620	0.033	0.542	1.874	1.739	0.570	0.030	0.482

1st, 2nd, 3rd: the first passenger class, the second passenger class, and the third passenger class. The first class is chosen as the baseline.

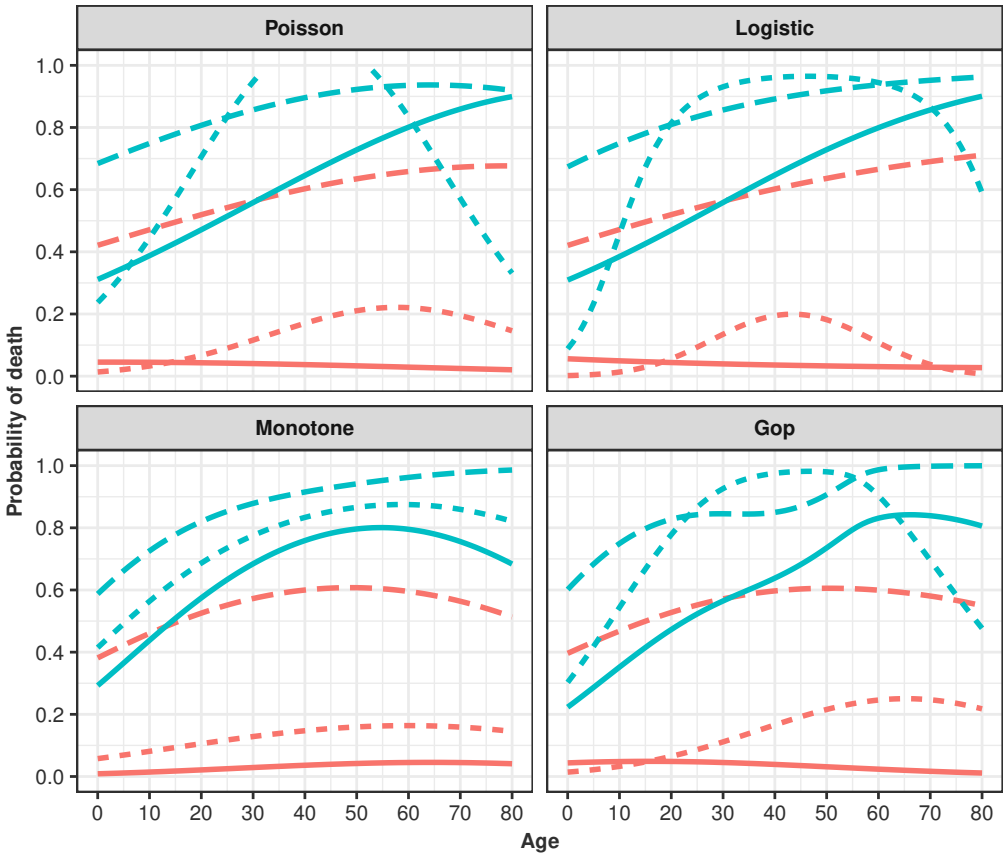


Figure 4.3: Predicted probability of death of the first passenger class (solid line), the second class (dotted line), and the third class (dashed line) with respect to different models. Red represents female, and blue represents male.

fitted probability for second class females decreases to zero as age approaches 80, whereas this does not happen with the Generalized Odds Product model. However, in the data set, there were only two females in the second class who were older than 55 and both of them died. This suggests that our Generalized Odds Product model may fit the data better. Unlike the other three plots, the fitted lines from the Monotone model do not cross each other. This is due to the assumption of monotonic treatment effects. As we discussed earlier, this assumption may not be plausible for the Titanic data set.

4.3 Relative Survival

One may be more interested in knowing which passenger class had a better chance to survive instead of a higher risk of death. A similar analysis can be conducted as in §4.2. The main difference between modeling relative survival and relative risk of death is how the outcome is defined. When modeling relative survival, we let the outcome Y be 1 if the passenger survived, otherwise 0. Figure 4.4 displays the empirical probability of survival. For females, the relative survival for the First Class (against the Third Class) is highest in the $20 \sim 45$ age group; For males, the highest relative survival for the First Class is the <20 age group. Same as in modeling the relative risk of death, we let the covariates X and W be identical, which include age, sex, age squared, and the interaction between age and sex. We consider the Third Class as the baseline.

We also apply the four different models to estimate the relative survival stratifying on age and sex: 1) Poisson regression; 2) Logistic regression; 3) Monotone; 4) GOP.

Table 4.3 and Figure 4.5 are the estimated regression coefficients and the fitted probability of survival, respectively. Similarly as relative risk models, the point estimates from the GOP model are close to those from Poisson model but different from Monotone Model. Figure 4.4 already shows that survival probability is not monotonic in passenger class after stratifying by age and sex. The fitted probability of survival from Poisson regression is larger than 1 when Second Class female passengers is younger than 10 or older than 60. The fitted probability from logistic regression is similar to it from our proposed method with a generalized odds

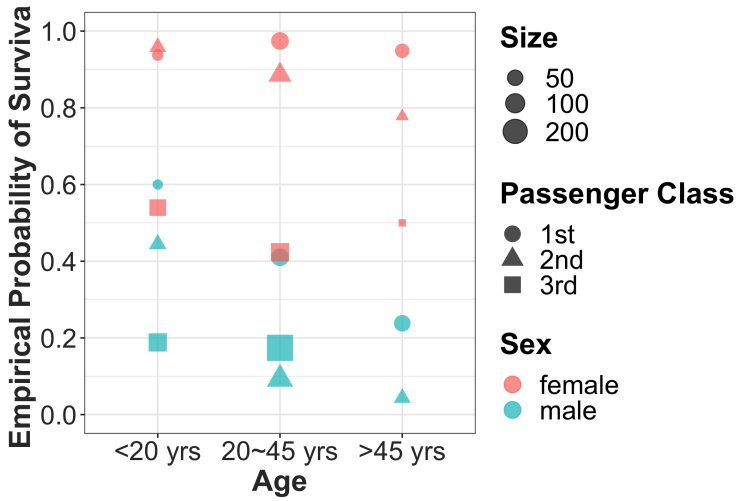


Figure 4.4: Empirical probability of survival after stratifying by passenger class, sex, and age. Different color represents different sex; Different shape represents different passenger class; The size of the dots represents the number of people in the certain group.

product nuisance model. The main differences are fitted probabilities of the Second Class males, and the First Class Females. For the First Class females, their probability of survival first increases until age 16 and than decreases slowly in the GOP method, while it keeps approximately unchanged in the logistic regression. There were six girls whose age was under 16 in the First Class: one 2 years old, one 14 years old, one 15 years old, and three 16 years old. Only the 2-year old baby did not survived; also can be seen in Figure 4.1. For the Second Class males, 10 passengers were older than 55 years ago and only one 62-year old male survived. Both evidence suggested that the GOP model has a better fitting.

In the appendix D, we provide a sensitivity analysis which imputes the missing ages. It also provides similar results.

4.4 Discussion

In this Chapter, we illustrate the use of our proposed methods on the Titanic data by modeling the relative risk of death and relative survival, respectively. Whether to model the

Table 4.3: Coefficient estimates via different models.

	2nd	2nd* male	2nd* age/10	2nd* age^2/ 100	2nd* male* age/10	1st	1st* male	1st* age/10	1st* age^2/ 100	1st* male* age/10
Point Estimate										
Monotone	0.178	0.096	0.071	-0.010	0.082	0.356	0.192	0.143	-0.020	0.164
GOP	0.584	0.565	0.036	-0.003	-0.563	0.311	0.542	0.246	-0.027	-0.072
Poisson	0.736	0.315	-0.117	0.030	-0.475	0.396	0.460	0.155	-0.007	-0.051
Standard Deviation										
Monotone	0.076	0.202	0.063	0.012	0.072	0.153	0.403	0.126	0.023	0.144
GOP	0.208	0.438	0.179	0.037	0.184	0.281	0.456	0.208	0.038	0.168
Poisson	0.433	0.584	0.320	0.059	0.242	0.546	0.590	0.328	0.053	0.193

1st, 2nd, 3rd: the first passenger class, the second passenger class, and the third passenger class. The first class is chosen as the baseline.

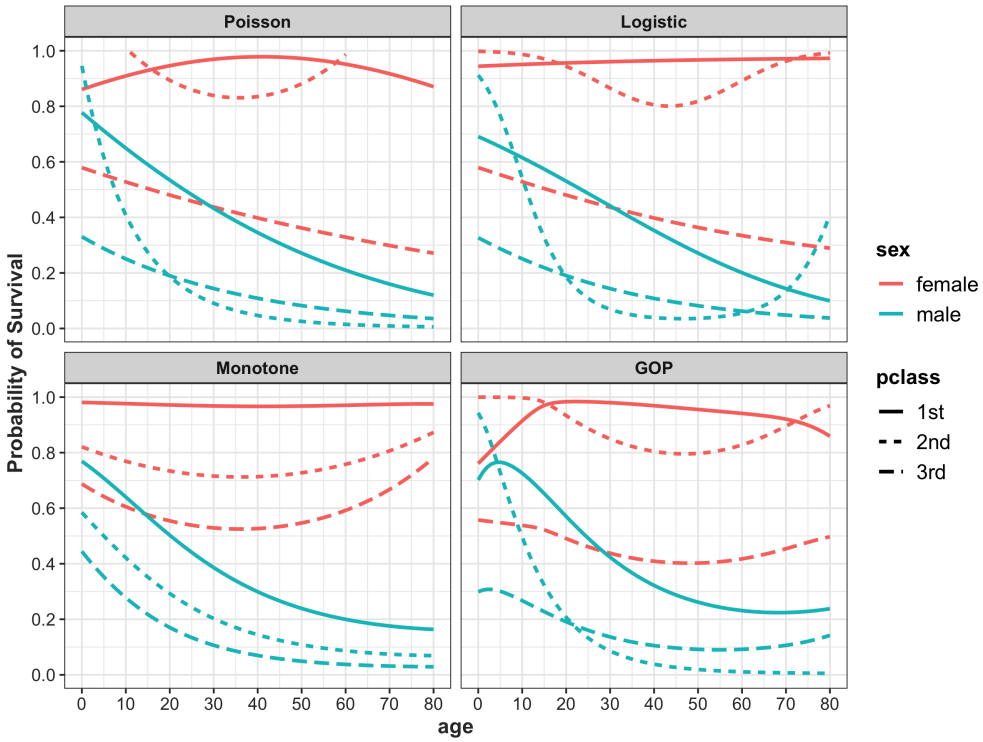


Figure 4.5: Predicted probability of survival of the first passenger class (solid line), the second class (dotted line), and the third class (dashed line) with respect to different models. Red represents female, and blue represents male.

relative risk or relative survival, it depends on researchers' primary of interest. Once that is decided, one can simply change the outcome $Y = 1$ as the case of interest.

In practice, We suggest to start from the data exploration. Analysts may use descriptive plots to examine the relationship between the treatment and outcome, and use them to assess plausibility of the monotonic treatment effect assumption first. If the relative risk/survival is not monotonic in exposure/treatment, one may consider to use our proposed method with a generalized odds product.

Chapter 5

COMBINING CASE-CONTROL AND COHORT DATA

5.1 Introduction

In this chapter, we focus on estimating relative risk by combining cohort and case-control studies. Considered in isolation for a rare event, cohort studies do not have sufficient cases, making it difficult to inform clinical decisions, however cohort studies can provide information about prevalence of the disease; For case-control studies, the prevalence is unidentifiable but there are many more cases compared with cohort studies. By combining those studies, we can obtain a greater statistical power and robust point estimates thus resulting in more trustworthy conclusions.

Case-control Studies

A case-control study is a retrospective observational study to help determine if an exposure is associated with an outcome/disease. In a case-control study, participants are selected based on their outcome/disease and further divided into a case group with the disease and a control group without the disease (Prentice and Pyke, 1979). Then a suspected risk factor is examined by comparing its prevalence in these two groups. A case-control study is often conducted for rare diseases. For example, female breast cancer is the most common cancer in the world, with 124.2 cases per 100,000 people in the United States in 2016. On the other hand, thyroid cancer is only the tenth most common cancer, with just 14 cases per 100,000 people[1].

There are several different types of case-control study.

- A *Matched case-control study* identifies one (or more) control(s) who have the same pre-specified attributes as each case (Breslow and Day, 1980). The cases and the controls are analyzed as a match set. To be more specific, the group of controls is selected in such a way that the distribution of a given factor in the controls is equal to the distribution of that factor in the case group. For example, if 10% people are aged 20-25 years, 45% are females, and 20% are current smokers, then in the control group, we should also enroll 10% of people aged 20-25 years, 45% females, and 20%

[1]In the website for Centers for Disease Control and Prevention.

current smokers. The advantages of a matched case-control study is that it allows the researcher to control a few well-know, and strong confounders. It also increases the statistical efficiency of a case-control study. However, this study sometimes can be time-consuming and expensive. Matching controls can't always be found. Furthermore there is a risk of over-matching, such as occurs when matching on a factor that is not a confounder of the exposure-outcome association.

- A *Nested case-control study* is 'nested' within a defined cohort. First cases are selected, and then controls are randomly sampled from cohort members remaining in cohort at the time each case occurs, or at the end of the study period (Ernster, 1994; Zhou et al., 2007). Nested case-control studies also allow for matching of controls to cases on (measured) confounders. They are often less expensive than a prospective cohort study since it is not necessary to ascertain exposure and all confounders for the whole cohort.
- In a *Case-cohort study*, controls are sampled from among those at risk at the beginning of the study period. Those controls are also called the subcohort. The study population includes all members of the subcohort and all incident cases that occur over the study period (Prentice, 1986). Compared with the nested case-control study, the advantage of the case-cohort study is the ability to study the prevalence of the disease in the subcohort.
- The *Two-phase study* is an extension of the traditional case-control study to contexts where there are rare events. White (1982) first proposed the method to improve estimating efficiency in the binary exposure setting. In the first phase, N individuals from the population are observed and their outcome Y and categorical covariates V are recorded. These data can be drawn cross-sectionally, prospectively, or retrospectively. In Phase I, $\text{pr}(Y \mid V)$ is identifiable. In the second phase, n out of N observations contribute more information. Specifically, looking retrospectively, the exposure Z of those n observations is obtained stratifying on the covariate V. Therefore, in Phase II, $\text{pr}(Z \mid Y, V)$ is identifiable Zelnick et al. (2018).

Cohort Studies

The "cohort" is a group of people. A cohort study is an observational study, which can be forward-looking (prospective) or backward-looking (retrospective). Both types of studies first recruit eligible and willing units, classify participants based on the exposure status (exposed and unexposed group), and finally compare outcome incidences over time between the exposed and unexposed groups. The strengths of this design are that the researcher can directly estimate outcome incidence. However it is inefficient for rare or delayed outcomes, especially for the prospective cohort study, which may require a long time to follow outcomes and secondly, to observe enough cases to be statistically significant, one would need to gather a large population.

Combining Case-Control and Cohort Studies

In contrast, a case-control study is comparatively fast, easy and inexpensive, because sampling is conditional on the outcome. Consequently, it is possible to collect sufficient cases of the rare disease (Wallen and Courtright, 1998). However, traditional case-control studies have limitations. They cannot provide the same evidence as a cohort study. Case-control studies cannot determine a relative risk, because the prevalence of the disease is set by the study design. Instead, odds ratios are usually reported from case-control studies, since

$$\frac{\text{pr}(Y=1 \mid Z=1)/\text{pr}(Y=0 \mid Z=1)}{\text{pr}(Y=1 \mid Z=0)/\text{pr}(Y=0 \mid Z=0)} = \frac{\text{pr}(Z=1 \mid Y=1)/\text{pr}(Z=1 \mid Y=0)}{\text{pr}(Z=0 \mid Y=0)/\text{pr}(Z=0 \mid Y=1)},$$

and $\text{pr}(Z=z \mid Y=y)$ is identified from case-control data. Even though an odds ratio approximates relative risk in rare diseases, it is unclear when the approximation is appropriate.

Cohort studies do provide information on prevalence which case-control studies cannot provide. However, for rare events case-control studies collect sufficient cases but cohort studies can not. It is thus natural to ask: can we combine cohort and case-control studies to have a better estimation of the relative risk?

Currently, to my knowledge, there are no studies modeling relative risk directly from a combination of case-control studies and cohort studies. Studies typically approximate the

relative risk by adjusting the odds ratio under an assumption regarding the prevalence of the disease or summarize the relative risk and the odds ratio separately from cohort and case-control studies respectively (Huxley et al., 2011; Chen et al., 2016). This chapter is organized as follows: In §5.2, we introduce the statistical model; In §5.4, simulations are provided to illustrate the performance of the method; In §5.5, an analysis further demonstrates the use of our method.

5.2 Methodology

Suppose the outcome Y is binary and takes value 0 or 1; treatment Z is ordinal, where $Z \in \{z_0, \ldots, z_K\}$, $K \geq 1$; and covariates $V \in \mathbb{R}^d$. The modeling of relative risk from a cohort study under monotonicity assumption has been described in §2. Under that setting, data is sampled from $\text{pr}(Y \mid Z = z, V = v)$, and the likelihood is further computed. However, a case-control study is retrospective, which means each individual is directly sampled from $\text{pr}(Z \mid Y, V)$ instead. The full prospective model is unidentifiable solely from a case-control study, because $\text{pr}(Y \mid Z = z, V = v)$ cannot be completely determined by $\text{pr}(Z = z \mid Y, V = v)$; see in Prentice and Pyke (1979). If we still want to rely on the framework of relative risk modeling, the key is to connect the prospective probability $\text{pr}(Y \mid Z = z, V = v)$ and the retrospective probability $\text{pr}(Z = z \mid Y, V = v)$ together. Using Bayes' theorem

$$\text{pr}(Z = z \mid Y, V = v) = \frac{\text{pr}(Y \mid Z = z, V = v)\text{pr}(Z = z \mid V = v)}{\sum_{z=0}^{K} \text{pr}(Y \mid Z = z, V = v)\text{pr}(Z = z \mid V = v)}, \tag{5.1}$$

we can simply connect them by building another model for the probability of exposure, which is also known as propensity score $\text{pr}(Z \mid V = v)$. For a binary treatment, the propensity score can be modeled via logistic regression. For a multi-level treatment, multinomial logistic regression can be used.

We consider a study including m studies in total. For notational simplicity, assume first m_1 studies are cohort studies, and the remaining m_2 are case-control studies ($m = m_1 + m_2$). For each study i, we have observations $\{(y_{ij}, z_{ij}, v_{ij})\}_{j=1}^{n_i}$. Our primary interest is to model the relative risks of any exposures against the baseline exposure. We formulate a combined

likelihood for all the studies. The likelihood of one observation in a case-control study can be obtained from Eq.(5.1), where $\text{pr}(Y \mid Z = z, V = v)$, $z \in \mathcal{Z}$ can be calculated from cohort studies under the framework discussed in §2, and $\text{pr}(Z \mid V = v)$ is from the propensity score model. For notational simplicity, we denote $\text{pr}(Z = z \mid V = v)$ as $\pi_z(v)$, and $\text{pr}(Z = 1 \mid Y = y, Z = z)$ as $q_y(z)$.

We specify the following models:

$$\log\{\text{RR}(z_0, z; v, \gamma)\} = \gamma^{\text{T}} X(z - z_0) \quad z \in \mathcal{Z}, \tag{5.2}$$

$$\log\{\text{OP}(z_{\min}, z_{\max}; v, \beta)\} = \beta^{\text{T}} W, \tag{5.3}$$

$$\log\left\{\frac{\pi_k(v;\eta)}{\pi_0(v;\eta)}\right\} = \eta_k^{\text{T}} Q_k \quad k \in \{1, \ldots, K\}, \tag{5.4}$$

where $X = X(v)$, $W = W(v)$, $Q_k = Q_k(v)$ $k \in \{1, \ldots, K\}$, $z_{\min} = \min\{z : z \in \mathcal{Z}\}$, and $z_{\max} = \max\{z : z \in \mathcal{Z}\}$.

The likelihood for the combination study consists of two sets of contributions, one from cohort studies and the other from case-control studies:

$$\prod_{i=1}^{m_1}\prod_{j=1}^{n_i} \text{pr}(Y = y_{ij}, Z = z_{ij} \mid V = v_{ij}) \times \prod_{i=m_1+1}^{m}\prod_{j=1}^{n_i} \text{pr}(Z = z_{ij} \mid Y = y_{ij}, V = v_{ij}). \tag{5.5}$$

Both $\text{pr}(Y = y_{ij}, Z = z_{ij} \mid V = v_{ij})$ and $\text{pr}(Z = z_{ij} \mid Y = y_{ij}, V = v_{ij})$ in Eq.(5.5) are a function of $\text{pr}(Y = y_{ij} \mid Z = z_{ij}, V = v_{ij})$ and $\text{pr}(Z = z_{ij} \mid V = v_{ij})$ $i \in \{1, \ldots, m\}, j \in \{1, \ldots, n_i\}$, where:

$$\text{pr}(Y = y_{ij}, Z = z_{ij} \mid V = v_{ij}) = \text{pr}(Y = y_{ij} \mid Z = z_{ij}, V = v_{ij}) \cdot \text{pr}(Z = z_{ij} \mid V = v_{ij}), \tag{5.6}$$

$$\text{pr}(Z = z_{ij} \mid Y = y_{ij}, V = v_{ij}) = \frac{\text{pr}(Y = y_{ij} \mid Z = z_{ij}, V = v_{ij})\text{pr}(Z = z_{ij} \mid V = v_{ij})}{\sum_{z=z_0}^{z_K} \text{pr}(Y = y_{ij} \mid Z = z, V = v_{ij})\text{pr}(Z = z \mid V = v_{ij})}. \tag{5.7}$$

$\text{pr}(Y = y_{ij} \mid Z = z_{ij}, V = v_{ij})$ and $\text{pr}(Z = z_{ij} \mid V = v_{ij})$ can be further represented as a function of $\gamma, \beta, \eta_k, k \in \{1, \ldots, K\}$ via the models specified in (5.2)-(5.4). Finally, by inserting Eq.(5.6) and (5.7) into Eq.(5.5) and taking the logarithm, we have the log-likelihood

of all the data from cohort and case-control studies:

$$
\begin{aligned}
& l\left(\gamma, \beta, \eta_1, \ldots, \eta_K \mid \{y_{1j}, z_{1j}, v_{1j}\}_{j=1}^{n_i}, \ldots, \{y_{mj}, z_{mj}, v_{mj}\}_{j=1}^{n_m}\right) \\
= & \sum_{i=1}^{m_1} \sum_{j=1}^{n_i} \log\{\mathrm{pr}(y_{ij} \mid z_{ij}, v_{ij})\} + \log\{\mathrm{pr}(z_{ij} \mid v_{ij})\} \\
& + \sum_{i=m_1+1}^{m} \sum_{j=1}^{n_i} \log\left\{\frac{\mathrm{pr}(y_{ij} \mid z_{ij}, v_{ij})\mathrm{pr}(z_{ij} \mid v_{ij})}{\sum_{z=z_0}^{z_K} \mathrm{pr}(y_{ij} \mid z, v_{ij})\mathrm{pr}(z \mid v_{ij})}\right\}. \qquad (5.8)
\end{aligned}
$$

Using this, estimates of γ, β, η_k $\{1, \ldots, K\}$ and their standard error can be obtained by maximizing the log-likelihood.

Remark 2. *If the relative risk is not monotonic in z, one can simply substitute Model (5.3) into a generalized odds product model described in §3.*

Case-control studies are able to assist cohort studies in estimating the relative risk, especially for a rare disease. We assume that participants in case-control studies and cohort studies are from the same underlying population, which means they have a common relative risk, propensity score and odds product. By combining of cohort studies and case-control studies, the resulting estimates of the relative risk are more efficient.

5.3 Analysis of the Simplest Case

To illustrate, we consider a special case where the study includes one cohort study and one case-control study, and the exposure Z is binary with a value of 0 or 1. Then Model (5.2)-(5.4) are simplified to

$$\log\{\mathrm{RR}(v; \gamma)\} = \gamma^{\mathrm{T}} X, \qquad (5.9)$$

$$\log\{\mathrm{OP}(v; \beta)\} = \beta^{\mathrm{T}} W, \qquad (5.10)$$

$$\mathrm{logit}\{\pi(v; \eta)\} = \eta^{\mathrm{T}} Q, \qquad (5.11)$$

where we denote $\text{pr}(Z = 1 \mid V = v)$ as $\pi(v)$, and $X = X(v)$, $W = W(v)$, $Q = Q(v)$. The log-likelihood (5.8) is further formulated as

$$
\begin{aligned}
& l\left(\gamma, \beta, \eta \mid \{y_{1j}, z_{1j}, v_{1j}\}_{j=1}^{n_i}, \{y_{2j}, z_{2j}, v_{2j}\}_{j=1}^{n_2}\right) \\
= & \sum_{j=1}^{n_1} y_{1j} \log\{p_{z_{1j}}(v_{1j})\} + (1 - y_{1j}) \log\{1 - p_{z_{1j}}(v_{1j})\} + z_{1j} \log\{\pi_{1j}(v)\} + (1 - z_{1j} \log\{1 - \pi_{1j}(v)\} \\
+ & \sum_{j=1}^{n_2} z_{2j} \log\{q_{y_{2j}}(v)\} + (1 - z_{2j}) \log\{1 - q_{y_{2j}}(v)\},
\end{aligned}
$$

where the subscript in the models above indicates the j-th individual in different studies (1_j: cohort; 2_j: case-control); see notation in §2. We provide explicit Wald-type confidence intervals in Appendix §E.1.

5.4 Simulation Studies

Continuing with the analysis of the special case where $m_1 = 1$, $m_2 = 1$, suppose further that exposure is binary. In the case-control study, the sample-size ratio between case and control is $1 : 2$. In the following, we simulate a series of hypothetical cohorts based on Model (5.9)-(5.11).

Toy example: no covariates

First, we start with a sanity test where the relative risk, odds product, and propensity score are constant, and then covariates V are added into the model. In the following simulations, we compare estimation bias and its standard error, standard deviation accuracy, 95% nominal coverage rate, and the power to find a non-zero effect when one is present.

We let $\gamma = 0.5$, $\beta = -7$, and $\eta = 0.2$,

$$
\begin{aligned}
\log(\text{RR}) &= 0.5 \\
\log(\text{OP}) &= -7, \\
\text{logit}\{\text{pr}(Z = 1)\} &= 0.2.
\end{aligned}
$$

The resulting prevalence $\text{pr}(Y = 1)$ is 0.035. Simulation results are based on 1000 Monte-Carlo runs. Having simulated samples for the cohort study, we obtain samples from case-control studies ranging in size from 200 to 2000. Table 5.1 summaries the simulation results under different numbers of population.

For all the scenarios, the bias is small and further decreases to 0 as the sample size increases. The standard deviation accuracy is close to 1, and the coverage probability of the proposed Wald-type confidence intervals also achieve the nominal 95% coverage-rate. The statistical power, where the null hypothesis is $\gamma = 0$, is much larger in the combination study. For a fixed combined population size, the higher the percent of people from a case-control study in a combination analysis, the lower the standard deviation of the estimates and the better the power is. When 200 subjects are added to the original 1000 subjects from the cohort study, the bias and standard deviation of the estimates obtained from the combination analysis are even smaller than the study with 2000 subjects solely from the cohort study. If we set the total population to be 2000, the estimate from a combination study, with 1000 people in the cohort study and 1000 in the case-control study, performs better than the cohort study with 2000 people; both the bias and standard error are much smaller, and the power is larger.

Given solely case-control data, people may simply use the fitted odds ratio model to approximate the relative risk model if, as in this case, the prevalence is low (0.035). Therefore, we also conduct a similar simulation of using a logistic regression to see if the approximation is appropriate. Table 5.2 summaries simulation results based on 1000 Monte Carlo runs. The bias of the estimates via logistic regression is much larger compared to our proposed method; see Table 5.1. It shows that using odds ratio to approximate is not as good as using our method to model relative risks directly. Such an approximation can be much worse with a non-rare disease.

Table 5.1: Monte Carlo simulation results based on 1000 runs for sanity test. The true values of γ is 0.5.

Case-control	0	200	500	1000	2000
Cohort 1000					
$\text{Bias}_{\times 10^2}(\text{SE}_{\times 10^2})$	2.736(1.246)	0.688(0.675)	1.219(0.499)	-0.242(0.377)	-0.308(0.274)
SD Accuracy	0.995	1.024	0.996	0.996	1.019
Coverage	0.965	0.961	0.948	0.95	0.947
Power	0.239	0.663	0.917	0.992	1
CI Width	1.536	0.857	0.616	0.465	0.346
Cohort 2000					
$\text{Bias}_{\times 10^2}(\text{SE}_{\times 10^2})$	1.179(0.852)	-0.354(0.583)	0.372(0.434)	-0.242(0.352)	-0.012(0.273)
SD Accuracy	1.006	1.020	1.041	0.995	0.972
Coverage	0.963	0.957	0.956	0.954	0.947
Power	0.480	0.760	0.966	0.997	1
CI Width	1.063	0.736	0.561	0.435	0.329
Cohort 4000					
$\text{Bias}_{\times 10^2}(\text{SE}_{\times 10^2})$	-0.304(0.583)	0.042(0.494)	-0.066(0.408)	-0.409(0.311)	-0.277(0.248)
SD Accuracy	1.027	0.98	0.968	1.026	0.999
Coverage	0.961	0.951	0.947	0.962	0.957
Power	0.762	0.913	0.981	0.997	1
CI Width	0.742	0.600	0.490	0.395	0.307

SE, standard error.

SD Accuracy = Average estimated standard deviation/Monte Carlo standard deviation.

Coverage nominal level = 95%.

Power, the null hypothesis: $\gamma = 0$.

CI width, confident interval width.

Table 5.2: Monte Carlo simulation results based on 1000 runs for sanity test. Here we fit a logistic regression model to case-control data and use the resulting estimate as an approximation to the relative risk. The true values of γ is 0.5.

N	200	500	1000	2000
$\text{Bias}_{\times 10^2}(\text{SE}_{\times 10^2})$	3.123(1.011)	2.093(0.628)	1.271(0.454)	1.823(0.315)
SD Accuracy	0.987	1.000	0.977	0.995
Coverage	0.947	0.943	0.950	0.939
Power	0.389	0.764	0.959	1.000
CI Width	1.237	0.778	0.549	0.388

SE, standard error.

SD Accuracy = Average estimated standard deviation/Monte Carlo standard deviation.

Coverage nominal level = 95%.

Power, the null hypothesis: $\gamma = 0$.

CI width, confident interval width.

Simulation with covariates

We now consider the more realistic case in which the models depend on covariates V. Specifically let V include an intercept and a random variable V_1 generated from a uniform distribution on the interval $[-2, 2]$. We generate Z from Model (5.4), and Y from models Model (5.2), (5.3), where $\gamma = (0.5, 1)^{\mathrm{T}}$, $\beta = (-7, 0.5)^{\mathrm{T}}$, $\eta = (0.2, -1)^{\mathrm{T}}$. For generating a case-control study, similarly as in the toy example, the ratio between cases and controls is 1:2. Cases and controls are randomly selected from the group with $Y = 1$ and $Y = 0$ in a bigger cohort study, respectively. Figure 5.1 displays how the probabilities and relative risk, and odds product change with covariates in an observational cohort study. For any $V_1 \in [-2.2]$, the stratified prevalence $\mathrm{pr}(Y = 1 \mid V = v_1)$ is smaller than 0.04. Figure 5.2 shows that in a case-control study, V_1 is fairly uniformly distributed in the control group, while in the case group more samples have larger v_1. The plots in Figure 5.2 show that the distribution of the covariate V differs for cases $(Y = 1)$ and controls $(Y = 0)$.

Table 5.3 summarizes the simulation results. We compare the performance of estimating the relative risk from a cohort of size $(2000 + x,)$ versus using a combination study with a cohort study of size 2000 and a case-control study of size x, for $x \in \{0, 500, 1500, 2000\}$. The performance is qualitatively similar to that observed in the toy example. As one would expect, for both estimators the bias and standard deviation become smaller as the sample size increases. In addition, the standard deviation accuracy is close to 1 and coverage rate reaches 95%. The power of testing a non-zero effect approaches 1 as the sample size increases. Furthermore, for a fixed combined population size, the estimates from a combination study always perform better than those obtained from a cohort study; in particular the combination study estimates have smaller standard deviation and the corresponding test has better power. More simulations for different sample sizes are shown in Appendix E.2.

We also illustrate that using logistic regression to approximate a relative risk model is not valid. Table 5.4 is similar to Table 5.2. We use logistic regression on the case-control data with different sample sizes. Even though the upper right plot in Figure 5.1 shows that the

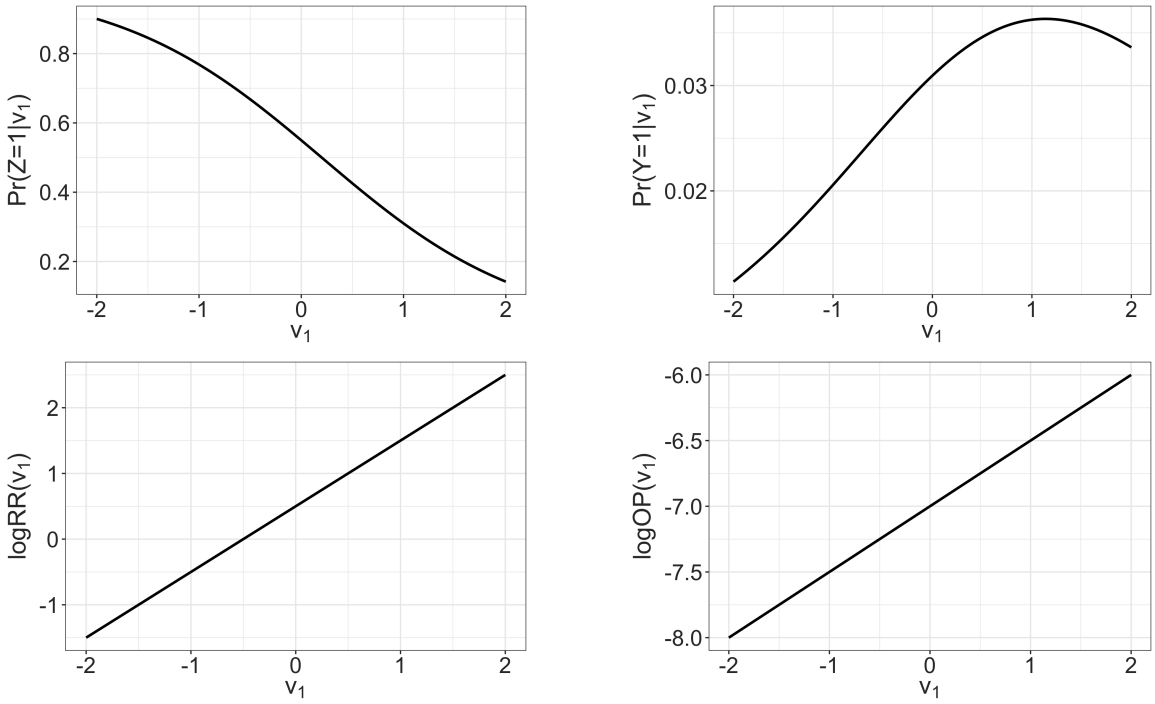


Figure 5.1: Propensity score, prevalence, logarithm of relative risk, and logarithm of odds product with respective to covariate V_1.

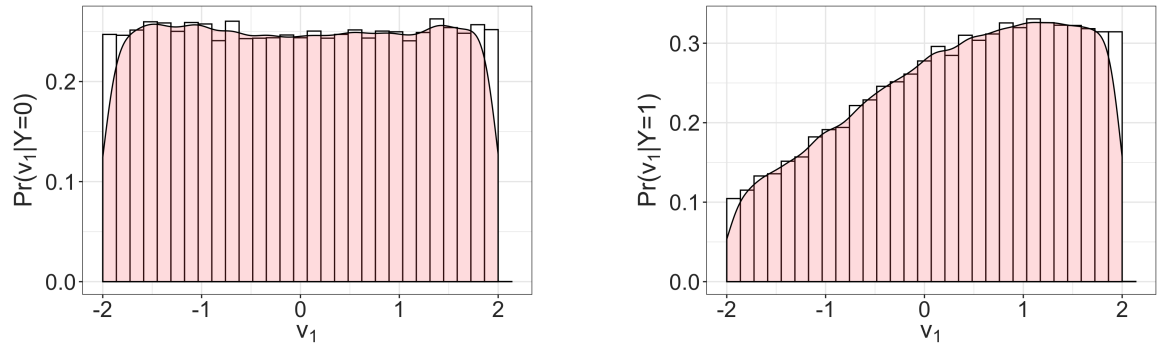


Figure 5.2: In the case-control study with sample size 100,000, the empirical distribution of V_1 conditional on $Y = 1$ and $Y = 0$, respectively.

Table 5.3: Monte Carlo simulation results based on 1000 runs. The true values of $\gamma = (\gamma_0, \gamma_1)^{\mathrm{T}}$ is $(0.5, 1)^{\mathrm{T}}$. The first two columns represent the number of subjects in each study.

		γ	γ	γ	γ	γ
Cohort	CC	$\text{Bias}_{\times 10^2}$(Standard $\text{Error}_{\times 10^2}$)	SD Accuracy	Coverage	Power	CI Width
2000	0	4.074(0.997)	0.978	0.955	0.409	1.208
		-0.201(0.880)	0.985	0.951	0.931	1.075
2000	500	0.923(0.488)	0.993	0.947	0.929	0.600
		0.292(0.435)	1.010	0.957	1.000	0.544
2500	0	3.521(0.871)	0.987	0.953	0.513	1.067
		0.629(0.750)	1.025	0.962	0.971	0.952
2000	1000	0.356(0.365)	1.025	0.953	0.994	0.464
		0.429(0.347)	0.983	0.954	1.000	0.423
3000		1.362(0.774)	1.004	0.956	0.541	0.963
		-0.182(0.716)	0.971	0.947	0.989	0.862
2000	1500	-0.285(0.310)	1.026	0.954	0.999	0.394
		0.679(0.283)	1.028	0.960	1.000	0.360
3500	0	1.416(0.723)	0.995	0.950	0.629	0.892
		0.668(0.636)	1.014	0.962	0.996	0.799
2000	2000	0.324(0.281)	1.008	0.951	1.000	0.351
		0.361(0.259)	1.002	0.951	1.000	0.321
4000	0	0.79(0.667)	1.004	0.954	0.692	0.831
		-0.655(0.594)	1.012	0.949	0.999	0.745

CC, case-control study

SD Accuracy = Average estimated standard deviation / Monte Carlo standard deviation.

Nominal level = 95%.

Power, null hypothesis: $\gamma_0 = 0$, $\gamma_1 = 0$, separately.

CI Width, confident interval width.

prevalence is fairly small across the range of the covariate, the bias is much larger compared with the estimation bias from using our method, especially for γ_1.

Table 5.4: Monte Carlo simulation results based on 1000 runs. The true values of $\gamma = (\gamma_0, \gamma_1)^{\mathrm{T}}$ is $(0.5, 1)^{\mathrm{T}}$. Here we fit a logistic regression model to case-control data and use the resulting estimate as an approximation to the relative risk. The first two columns represent the number of subjects in the case-control study.

	γ	γ	γ	γ	γ
N	Bias$_{\times 10^2}$(Standard Error$_{\times 10^2}$)	SD Accuracy	Coverage	Power	CI Width
500	2.933(0.718)	0.954	0.937	0.679	0.848
	4.743(0.643)	0.981	0.940	0.999	0.782
1000	3.103(0.481)	1.001	0.951	0.941	0.597
	5.439(0.442)	1.004	0.939	1.000	0.550
1500	2.841(0.387)	1.014	0.954	0.993	0.487
	5.253(0.355)	1.019	0.935	1.000	0.448
2000	2.402(0.348)	0.975	0.946	0.998	0.420
	4.128(0.308)	1.011	0.924	1.000	0.387

SD Accuracy = Average estimated standard deviation / Monte Carlo standard deviation.

Nominal level = 95%.

Power, null hypothesis: $\gamma_0 = 0$, $\gamma_1 = 0$, separately.

CI Width, confident interval width.

5.5 Application

The National Longitudinal Survey of Youth (NLSY) has been following the same participants from a specific birth cohort and collecting data on demography, labor market activity, education, health condition, and so on. The number of years of participation in the program varies for each person. Each year new participants have joined the study though others have

dropped out of. The study started in 1979 and continues to the present. Overall, 32,208 participants have joined over the past four decades. In our work, we specifically consider data from participants who responded to the survey in 2017. We treat the samples in 2017 as the target population.

In this section, we demonstrate the use of our proposed methods in Section 5.2 by studying the association between smoking and self-rated health (srh) in the NLSY of 2017. The point estimates based on the whole cohort population via Model (5.2)-(5.4) are treated as the ground truth. Then we draw smaller cohort and case-control samples from this population to illustrate the performance of our method.

5.5.1 Data Exploration

A large collection of work has showed that smoking affects people's health (Chapman et al., 1993; Escobedo et al., 1997; Akl et al., 2010). The association between smoking and health is affected by age and gender (Haber and Kent, 1992; Escobedo et al., 1993; Boezen et al., 1994; Allen et al., 2015). Therefore, in our analysis, we study the relative 'risk' of self-related health between the smoking group and the non-smoking group after adjusting by age and sex. In 2017, in total 14,570 people joined the survey. Fifty-five (0.36%) of them do not have self-rated health information, forty-four (0.30%) miss smoking status, and nine (0.062%) miss age. After removing the missing data, we have 14,506 participants left.

There are five different levels of srh: poor, fair, good, very good, and excellent. We treat srh as a binary outcome by combining fair, good, very good and excellent into one category called non-poor. Smoking status is the binary exposure indicating if the participant is currently smoking or not. Table 5.5 reports the number and proportion of people's srh after stratifying by the smoking status. We can see that people who currently smoke have a higher probability of poor srh compared with the non-smoking group. Figure 5.3 shows the srh of participants by their smoking status, age, and sex. Females tend to have a poorer self-rated health compared with males, and the older tend to have poorer srh compared to the younger. As seen by Figure 5.3, few males are surviving after age 70 in the smoking group.

Consequently, the observed association between smoking and self-rated health is attenuated due to censoring by death. For this reason we restrict our analysis to people under 70, which reduces the size of the population to 13,522. Among the 13,522 participants, 472 (3.49%) have rated themselves as having poor health. We see that a fairly small proportion of people in each group rate their health as poor.

Table 5.5: Descriptive statistics of people's self-rated health and smoking status.

Smoking Status	self-rated health	N(%)
No	Non-poor	11733(96.35%)
No	Poor	444(3.65%)
Yes	Non-poor	2181(93.65%)
Yes	Poor	148(6.35%)

We code poor self-rated health as 1, otherwise 0. The non-smoking group is the baseline. In order to have an intuition of the association between the outcome srh and the exposure smoking status, we display the LOWESS (locally weighted scatterplot smoothing) plot in Figure 5.4. The LOWESS fit is accessed by using the `histSpikeg` function in `R`. Figure 5.4 shows that the relative health between the smoking and the non-smoking group is different among females and males at the same age, and further, the difference varies as age changes. This suggests that we should include an interaction between age and gender when modeling the relative health.

5.5.2 Estimation

We model the relative risk model, the odds product, and the propensity score as a function of the same set of covariates, which include age, sex, and the interaction between age and sex.

The binary relative risk model (brm) first described in Richardson et al. (2017), is also

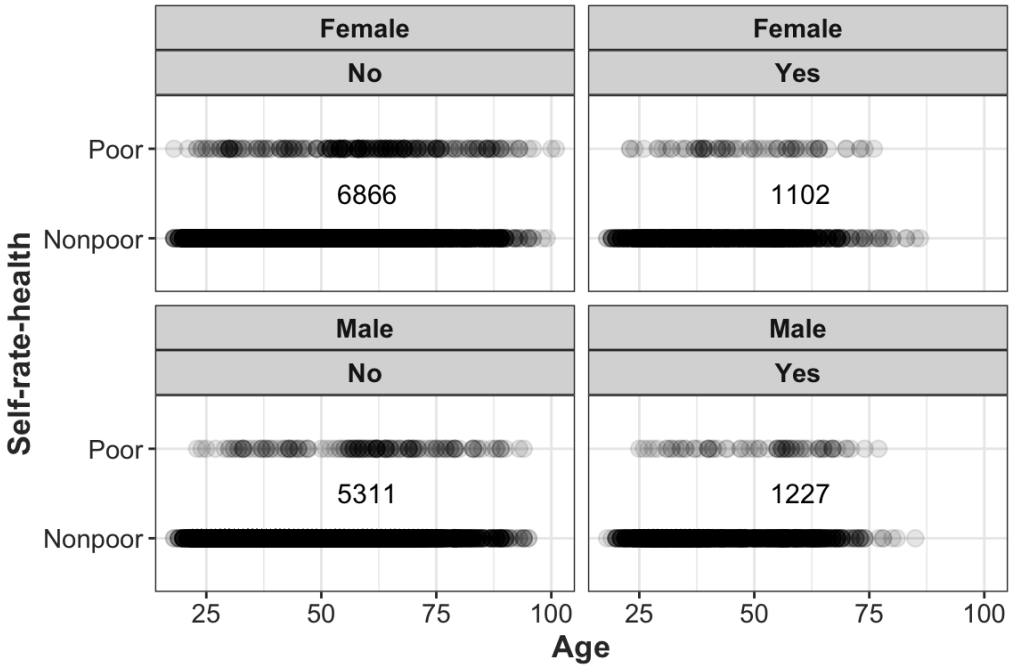


Figure 5.3: Participants' self-rated health by smoking status, age, and sex. The number of participants in each group is shown in the center of the corresponding plot.

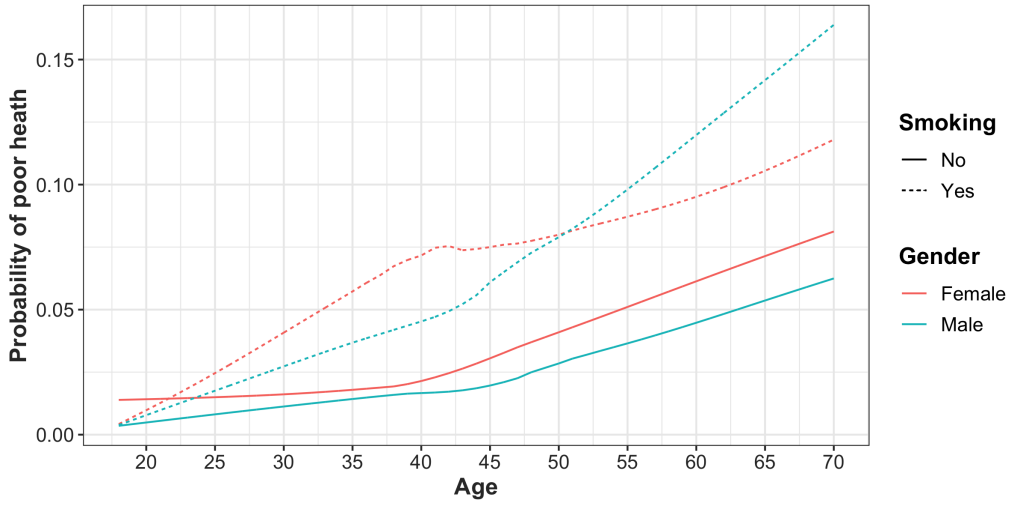


Figure 5.4: The LOWESS fit of probability of poor health against smoking status, age, and gender based on the whole 2017 cohort population. The dashed line is the smoking group, solid line is the non-smoking group. Red represents female, and blue represents male.

a special case of our proposed models in Chapter 2 and Chapter 3. First we apply the `brm` function in `R` to obtain the point estimates of Model (5.9)-(5.10). Then we use the `glm` function with a logit link to obtain the estimates for Model (5.11). Table 5.6 shows the regression coefficients for the suggesting models.

Table 5.6: Coefficients estimates based on the cohort population from 2017.

Est.	Intercept	Age	Male	Age * Male
RR (γ)	1.529	-0.016	-0.682	0.019
OP (β)	-9.537	0.076	-1.625	0.024
PS (η)	-1.549	-0.005	0.754	-0.009

RR, OP, PS: relative risk, odds product, and propensity score.

Figure 5.5 displays the fitted probabilities of poor health from the brm model. Figure 5.4 and Figure 5.5 are quite similar, suggesting that the brm model fitted to the full cohort describes the population well. We see the same qualitative features: as people get older, their health condition becomes worse. The condition is even worse for males who smoke after age 50.

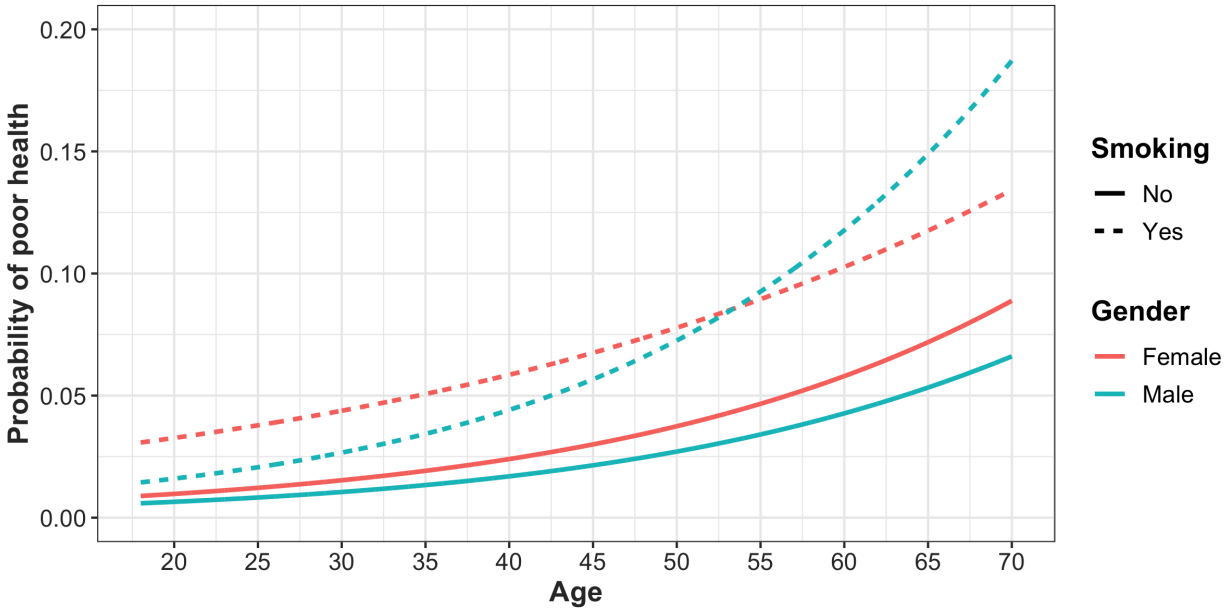


Figure 5.5: Predicted probability of poor health of the smoking group (dash line), and the non-smoking group (solid line) based on the whole 2017 cohort population via brm model. Red represents females, and blue represents males.

We now treat the 13,522 people as our population and treat the coefficients in Table 5.6 as the ground truth. We develop a similar strategy as in §5.4 to illustrate our proposed method for a combination study. Each time we draw a cohort and a case-control sample from the population and treat them as one instance of a combination study. We apply the brm model to the cohort sample, and our proposed method to the meta-study consisting of the cohort sample and the case-control sample together. Specifically, we obtain the combination

study samples as follows:

(a) Cohort sample: Randomly draw a sample of size n_1 without replacement from the population. Run the `brm` function on this sample.

(b) Case-control sample: Randomly draw $\frac{1}{3}n_2$ cases and $\frac{2}{3}n_2$ controls without replacement from the rest of the population in (a). Apply our proposed model to the cohort sample and case-control sample.

(c) To make a comparison, further randomly draw a sample of n_2 without replacement from the rest of the population after drawing the sample in (a). Combine the cohort in (a) with the cohort in (c), and run the `brm` on this cohort study of size $(n_1 + n_2)$.

For example, taking $n_1 = 300$ and $n_2 = 150$, we have three studies: a cohort study of size 300 denoted *cohort 300*; a cohort study of size 450 denoted *cohort 450*; and a combination study of size 450 denoted *meta 450*. Note that cohort 450 and meta 450 include all the participants from cohort 300. For the cohort 450, there are around 17 cases on average, and for the same size meta 450, there are around 60 cases on average.

The number of Monte Carlo runs is 1,000. Table 5.7 summarizes the results for $n_1 = 300$ and $n_2 = 150$. We also display the histogram of the estimates from 1000 Monte Carlo runs in Figure 5.6. The bias of our proposed method for the combination study is small when the sample size is 450. It further decreases as the sample size increases; see Appendix E.3. The standard deviation for the meta approach is also the smallest compared with the other two cohort studies, which can also be seen from Figure 5.6. The standard deviation accuracy is much larger than 1 for *cohort 300* and *cohort 450*. This is because for some instances of cohort study there are few cases. Such data sets will cause a large uncertainty in the estimates (see the histogram of standard deviation plot in Figure 5.6). With the larger standard deviation, the coverage probability is also larger than 95% for cohort studies of size 300 and 450. The estimates from the combination study also have better power (null hypothesis: each coefficient is 0 respectively) when compared with the same size cohort study. The distribution of the coefficients in the odds product model and the propensity score model are displayed in figures in §E.2.

In conclusion, the example shows that when the cases are rare, the estimates from our proposed model for meta data always have better performance when compared with the same size cohort study. However, drawing samples without replacement from a finite population in this way has some limitations. Specifically, the population size here is 13,533 and the number of cases is 472. The larger the sample size of each draw, the more likely the same samples keep appearing. The same issue happens for drawing the case-control data. With highly similar sub-population, we will end up with similar point estimates, which further leads to smaller Monte Carlo standard deviations. Consequently the model-based standard error, and hence our confidence intervals, have coverage that is higher than the nominal level. In Appendix §E.3, we show the results for a larger size cohort and case-control study.

5.6 Discussion

A case-control study cannot directly provide relative risks due to its design, unless the disease or outcome being studied is rare. However, we have developed a novel method of modeling relative risks for the meta analysis data which includes observations from cohort studies and case-control studies. Our method works especially well under the rare disease setting in terms of estimation accuracy and power.

In a situation where N participants' outcome Y is known as well as their covariates V, prospectively or retrospectively, but without knowing their exposure Z. Then n samples are randomly selected from the N participants, and the detailed exposure of each sampled individual is obtained. This contains two steps. In the first step, data arise via a sample from the underlying population; and in the second step, data is collected as the scheme in a case-control study. Therefore, the likelihood will include the information from the two steps:

$$\prod_{i=1}^{N} \text{pr}(Y = y_i \mid V = v_i) \times \prod_{j \in \mathcal{S}} \text{pr}(Z = z_j \mid Y = y_j, V = v_j), \tag{5.12}$$

where $\mathcal{S}$ is the index set of these n samples in step 2. The probability $\text{pr}(Y = y \mid V = v)$ in Eq.(5.12) can be computed as $\sum_{z=z_0}^{z_K} \text{pr}(Y = y, Z = z \mid V)$, where $\text{pr}(Y = y, Z = z \mid V)$ is calculated from Eq.(5.6) in §3.1. One may have multiple studies including cohort studies,

Table 5.7: Coefficient estimates for three different studies based on 1000 Monte Carlo runs.

	Intercept	Age/100	Gender	Age/100 * Gender
Bias(Standard Error)				
Chrt 300	−0.271(0.135)	−0.690(0.264)	0.255(0.206)	0.069(0.412)
Chrt 450	−0.229(0.110)	−0.402(0.216)	0.245(0.167)	−0.187(0.331)
Chrt 300 & CC 150	0.179(0.051)	−0.482(0.101)	−0.036(0.078)	0.133(0.153)
SD Accuracy				
Chrt 300	1.841	2.004	2.037	2.195
Chrt 450	1.38	1.469	1.541	1.623
Chrt 300 & CC 150	1.003	1.017	1.015	1.018
Coverage (Nominal = 95%)				
Chrt 300	0.989	0.988	0.994	0.992
Chrt 450	0.983	0.979	0.985	0.982
Chrt 300 & CC 150	0.968	0.965	0.968	0.965
Power				
Chrt 300	0.034	0.013	0.010	0.014
Chrt 450	0.070	0.025	0.014	0.022
Chrt 300 & CC 150	0.180	0.075	0.041	0.053

Chrt, cohort study; CC, case-control study.

SD Accuracy = estimated standard deviation / Monte Carlo standard deviation.

Power, the null hypothesis: the coefficients are 0.

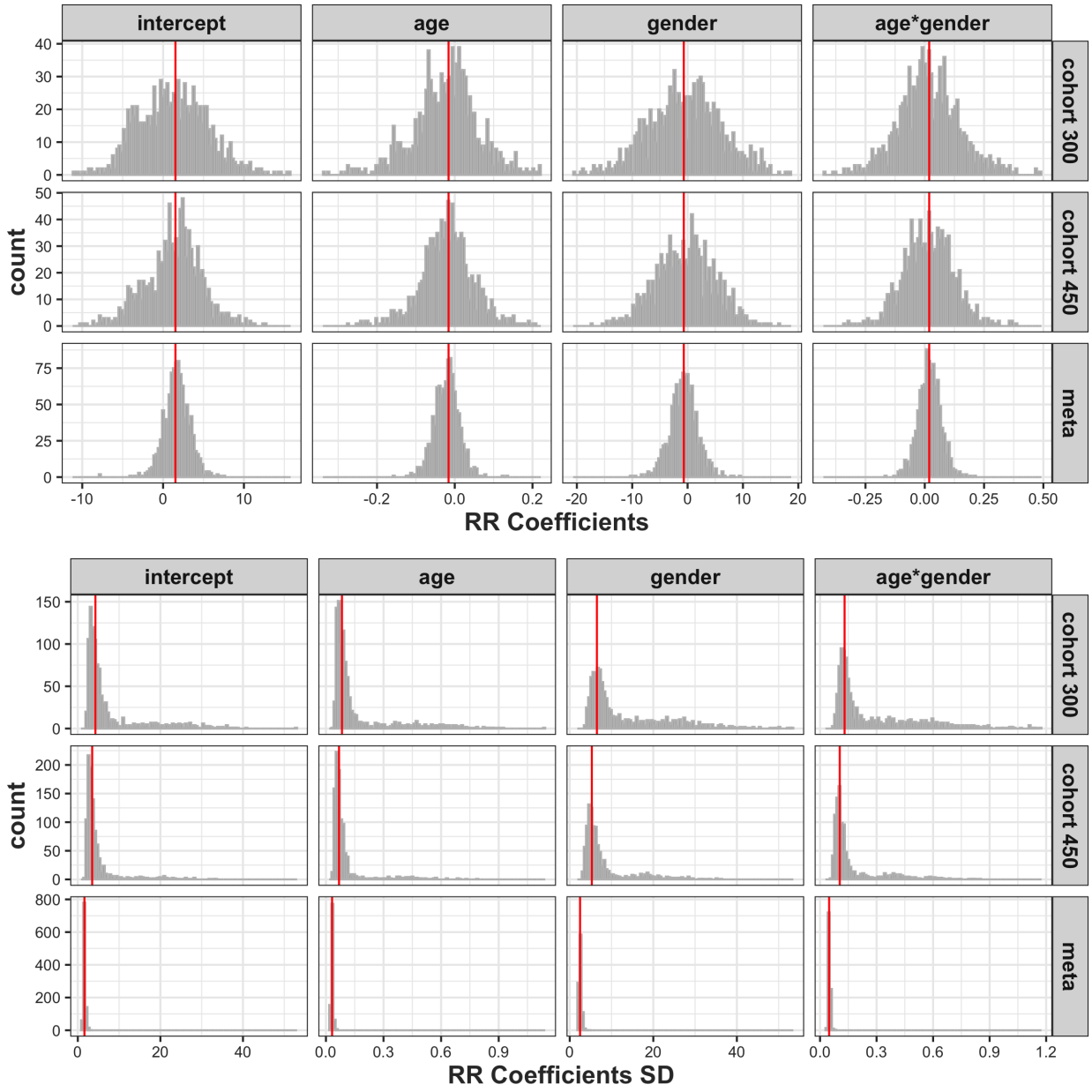


Figure 5.6: Empirical distributions of parameters estimated from 1000 Monte Carlo runs. The top figure shows the estimated coefficients in Model 5.9. The bottom figure is the model based standard deviation estimate. Cohort1 is the cohort study of size 300, cohort2 is size 450, and meta is the combination study of combining the cohort of size 300 and case-control study of size 150. The red vertical lines are the ground truth of the estimates and Monte Carlo standard deviation, respectively.

case cohort studies in which data are available on those for whom exposure status is unknown, and case control studies where there is no information regarding those who are not selected cases or controls. Estimates may be obtained as standard fashion via maximum likelihood estimation.

Random Effect

We have considered a fixed-effect relative risk model. That is, the effect expected from each study (cohort studies as well as case-control studies) is the same. However, studies sometimes differ in terms of the population and context as well as the study design. For example, different studies are conducted by different researchers in different countries where differences in regulation may lead to different eligibility criteria. In such settings, it would be natural to consider a random effects model. Such an approach would be a natural extension of the methods presented here. We may add a hierarchical model to the intercept of the relative risk model (5.2). By doing this, we allow the relative risks to differ across studies. A natural parameter of interest is the expected average of the relative risks (between different exposures/treatments) as a function of baseline covariates.

Alternatively, one may consider the relative risk fixed (given baseline covariates) and instead allow a random intercept in the odds product model (5.3). The advantage of the this latter approach is that the relative risk model may be easier to interpret as it is still a single value (or function) rather than the mean of a distribution of relative risks.

BIBLIOGRAPHY

Akl, E. A., Gaddam, S., Gunukula, S. K., Honeine, R., Jaoude, P. A., and Irani, J. (2010). The effects of waterpipe tobacco smoking on health outcomes: a systematic review. *International Journal of Epidemiology*, 39(3):834–857.

Al-Mamgani, A., van Putten, W. L., Heemsbergen, W. D., van Leenders, G. J., Slot, A., Dielwart, M. F., Incrocci, L., and Lebesque, J. V. (2008). Update of dutch multicenter dose-escalation trial of radiotherapy for localized prostate cancer. *International Journal of Radiation Oncology Biology Physics*, 72(4):980 – 988.

Allen, A. M., Scheuermann, T. S., Nollen, N., Hatsukami, D., and Ahluwalia, J. S. (2015). Gender Differences in Smoking Behavior and Dependence Motives Among Daily and Nondaily Smokers. *Nicotine and Tobacco Research*, 18(6):1408–1413.

Ball, R. and Chernova, K. (2008). Absolute income, relative income, and happiness. *Social Indicators Research*, 88(3):497–529.

Boezen, H., Schouten, J., Postma, D., and Rijcken, B. (1994). Distribution of peak expiratory flow variability by age, gender and smoking habits in a random population sample aged 20-70 yrs. *European Respiratory Journal*, 7(10):1814–1820.

Breslow, N. and Day, N. (1980). Statistical methods in cancer research. volume i - the analysis of case-control studies. *IARC scientific publications*, (32):5—338.

Chapman, S., Wong, W. L., and Smith, W. (1993). Self-exempting beliefs about smoking and health: differences between smokers and ex-smokers. *American Journal of Public Health*, 83(2):215–219.

Chen, C., Xun, P., Nishijo, M., Carter, S., and He, K. (2016). Cadmium exposure and risk of prostate cancer: a meta-analysis of cohort and case-control studies among the general and occupational populations. *Scientific reports*, 6(25814).

Dukes, O. and Vansteelandt, S. (2018). A Note on G-Estimation of Causal Risk Ratios. *American Journal of Epidemiology*, 187(5):1079–1084.

Easterlin, R. A. (2001). Income and happiness: Towards a unified theory. *The Economic Journal*, 111(473):465–484.

Ernster, V. (1994). Nested case-control studies. *Preventive Medicine*, 23(5):587 – 590.

Escobedo, L. G., Marcus, S. E., Holtzman, D., and Giovino, G. A. (1993). Sports Participation, Age at Smoking Initiation, and the Risk of Smoking Among US High School Students. *JAMA*, 269(11):1391–1395.

Escobedo, L. G., Reddy, M., and DuRant, R. H. (1997). Relationship Between Cigarette Smoking and Health Risk and Problem Behaviors Among US Adolescents. *Archives of Pediatrics and Adolescent Medicine*, 151(1):66–71.

Greenland, S., Robins, J. M., and Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46.

Haber, J. and Kent, R. L. (1992). Cigarette smoking in a periodontal practice. *Journal of Periodontology*, 63(2):100–106.

Huxley, R. R., Filion, K. B., Konety, S., and Alonso, A. (2011). Meta-analysis of cohort and case–control studies of type 2 diabetes mellitus and risk of atrial fibrillation. *The American Journal of Cardiology*, 108(1):56 – 62.

Lang, C. E., L. K. R. and Birkenmeier, R. L. (2015). Dose and timing in neurorehabilitation: prescribing motor therapy after stroke. *Current Opinion in Neurology*, 28(6):549–555.

Lumley, T., Kronmal, R., and Ma, S. (2006). Relative risk regression in medical research: Models, contrasts, estimators, and algorithms. *UW Biostatistics Working Paper Series*, page 293.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73(1):1–11.

Prentice, R. L. and Pyke, R. (1979). Logistic Disease Incidence Models and Case-Control Studies. *Biometrika*, 66(3):403.

Richardson, T. S., Robins, J. M., and Wang, L. (2017). On modeling and estimation for the relative risk and risk difference. *Journal of the American Statistical Association*, 112(519):1121–1130.

Tchetgen Tchetgen, E. (2013). Estimation of risk ratios in cohort studies with a common outcome: A simple and efficient two-stage approach. *International Journal of Biostatistics*, 9(2):251–264.

van der Laan, M. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Series in Statistics. Springer New York.

Wallen, S. and Courtright, P. (1998). Epidemiology in practice: case-control studies. *Community eye health*, 11(28):57–58.

Wang, L., Richardson, T. S., and Robins, J. M. (2017). Congenial causal inference with binary structural nested mean models. *arXiv preprint arXiv:1709.08281*.

White, J. E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, 115(1):119–128.

Wikipedia contributors (2020). RMS Titanic — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=RMS_Titanic&oldid=951749670. [Online; accessed 27-April-2020].

Yang, J. C., Sherry, R. M., Steinberg, S. M., Topalian, S. L., Schwartzentruber, D. J., Hwu, P., and Rosenberg, S. A. (2003). Randomized study of high-dose and low-dose interleukin-2 in patients with metastatic renal cancer. *Journal of clinical oncology: Official journal of the American Society of Clinical Oncology*, 21(16).

Zelnick, L. R., Schildcrout, J. S., and Heagerty, P. J. (2018). Likelihood-based analysis of outcome-dependent sampling designs with longitudinal data. *Statistics in Medicine*, 37(13):2120–2133.

Zhang, J. and Yu, K. F. (1998). What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *Journal of the American Medical Association*, 280(19):1690–1691.

Zhou, H., Chen, J., Rissanen, T. H., Korrick, S. A., Hu, H., Salonen, J. T., and Longnecker, M. P. (2007). Outcome-dependent sampling: an efficient sampling and inference procedure for studies with a continuous outcome. *Epidemiology (Cambridge, Mass.)*, 18(4):461—468.

Appendix A

DOUBLY ROBUST ESTIMATOR

A.1 Doubly robust estimator

van der Laan and Rose (2011, §A.15) have provided the efficient scores for the parameters of interest in relative risk models which allow the treatment to be categorical or continuous, and also allow interactions between treatment and baseline covariates. In the following, we separately show the score functions for our two proposed methods.

- Parameterization assuming a monotonic treatment relative risk. Our model of interest is $\log\{\text{RR}(0, Z; V, \gamma)\} = \gamma^{\text{T}} VZ$. The efficient score function is

$$S(\gamma) = \{Y \exp(-\gamma^{\text{T}} VZ) - p_0(V)\}\, h(Z \mid V), \tag{A.1}$$

  where

$$h(Z \mid V) = \frac{V p_Z(V)}{p_0(V)\{1 - p_Z(V)\}} \left[Z - \frac{E\left\{ \frac{Z p_Z(V)}{1-p_Z(V)} \mid V \right\}}{E\left\{ \frac{p_Z(V)}{1-p_Z(V)} \mid V \right\}} \right]. \tag{A.2}$$

  This representation shows $ES(\gamma) = 0$ if either the model for the baseline $p_0(V)$ is correctly specified or the conditional distribution of treatment Z given covariates V is correctly specified. This yields a doubly robust estimator for γ.

- Parameterization with a categorical treatment (without a monotonicity assumption). The model of interest is $\log\{\text{RR}(0, Z; V)\} = \sum_{k=1}^{K} \mathbb{1}\{Z = k\} \alpha_k^{\text{T}} V$. $S(\alpha) = (S(\alpha_1)^{\text{T}}, \ldots, S(\alpha_K)^{\text{T}})$ be the score function for $(\alpha_1, \ldots, \alpha_K)$. Similarly to the monotonic treatment effect model,

$$S(\alpha_i) = \left[Y \exp\left\{ -\sum_{k=1}^{K} \mathbb{1}\{Z = k\} \alpha_k^{\text{T}} V \right\} - p_0(V) \right] h_i(Z \mid V) \quad i \in \{1, \ldots, K\}, \tag{A.3}$$

  where

$$h_i(Z \mid V) = \frac{V p_Z(V)}{p_0(V)\{1 - p_Z(V)\}} \left[\mathbb{1}\{Z = i\} - \frac{E\left\{ \frac{\mathbb{1}\{Z=i\} p_Z(V)}{1-p_Z(V)} \mid V \right\}}{E\left\{ \frac{p_Z(V)}{1-p_Z(V)} \mid V \right\}} \right]. \tag{A.4}$$

As in our first method, the doubly robust estimator of $(\alpha_1, \ldots, \alpha_K)$ can be shown to be consistent if either the baseline risk model or the conditional probability distribution $\text{pr}(Z \mid V)$ are correctly specified.

Appendix B

PARAMETERIZATION WITH MONOTONIC TREATMENT EFFECTS

B.1 Simulations with larger population

Table B.1: Monte Carlo simulation results based on 1000 runs for the proposed estimator which assumes monotonic treatment effects. The true values of γ_0 and γ_1 are 0 and 1, respectively

Sample Size		5000	10000
$\text{Bias}_{\times 10^2}(\text{SE}_{\times 10^2})$			
	γ_0	0.051(0.081)	-0.053(0.058)
	γ_1	-0.04(0.092)	-0.012(0.065)
SD Accuracy			
	γ_0	1.037	1.026
	γ_1	1.01	1.013
Coverage			
	γ_0	0.963	0.956
	γ_1	0.95	0.948

SE, standard error.

SD Accuracy = estimated standard deviation / Monte Carlo standard deviation.

Nominal level = 95%.

B.2 More simulations for Model Comparisons

Table B.2: Simulation results for three different methods based on 1000, 5000 samples and 1000 Monte Carlo runs. The true values for $\gamma$are $(0,1)^{\mathrm{T}}$, $(0,1)^{\mathrm{T}}$

Sample Size	Bias$_{\times 10^2}$(Standard Error$_{\times 10^2}$)	SD Accuracy	Coverage (Nominal = 95%)
1000	γ	γ	γ
	Monotone		
	$-0.122(0.145)$	1.020	0.957
	0.011(0.174)	1.025	0.954
	DR-G		
	$-0.363(0.391)$	0.827	0.924
	9.178(0.653)	0.713	0.883
5000	γ	γ	γ
	Monotone		
	0.002(0.064)	1.032	0.956
	0.068(0.079)	1.001	0.956
	DR-G		
	$-0.026(0.138)$	0.959	0.937
	1.290(0.212)	0.914	0.927

Monotone, using models (2.9) and (2.10); DR-G, doubly robust estimator by Dukes and Vansteelandt (2018). SD Accuracy = estimated standard deviation / Monte Carlo standard deviation.

Appendix C

PARAMETERIZATION WITH A CATEGORICAL TREATMENT

C.1 Simulations with different population size

Table C.1: Monte Carlo simulation results based on 1000 runs for the relative risk model with a generalized odds product nuisance model. The true values for vectors α_1 and α_2 are $(-0.5, 1)^{\mathrm{T}}$ and $(0.5, 1.5)^{\mathrm{T}}$ respectively

Sample Size	500		5000	
	α_1	α_2	α_1	α_2
$\text{Bias}_{\times 10^2}(\text{SE}_{\times 10^2})$				
	-0.913(0.612)	0.587(0.409)	-0.202(0.188)	-0.168(0.122)
	1.739(0.627)	2.741(0.514)	-0.057(0.189)	0.051(0.159)
SD Accuracy				
	0.983	0.991	0.989	1.031
	1.02	1.028	1.027	1.015
Coverage				
	0.946	0.948	0.947	0.955
	0.955	0.96	0.961	0.948

SE, standard error.

SD Accuracy = estimated standard deviation / Monte Carlo standard deviation.

Nominal level = 95%.

C.2 Simulation results if misspecifying the nuisance model

An analyst takes variable V^* instead of V in the nuisance model to estimate parameters of interest. V^* includes an intercept and an irrelevant covariate which is distributed as Unif$[-2, 2]$. In Table C.2, the ground truth $\beta = (1, -0.5)^{\mathrm{T}}$ and $\log\{\textsc{gop}(v)\}$ takes value in $[0, 2]$. Note that if the quantity of nuisance model is small, it barely has an impact on the estimates even under misspecification; see Table C.2. However, if β is changed to $(2, -3)^{\mathrm{T}}$

and then $\log\{\text{GOP}(v)\}$ takes value in $[-4, 8]$, we could see the estimates are biased and inconsistent; see in Table C.3.

Table C.2: Monte Carlo simulation results based on 1000 runs for the relative risk model with a generalized odds product nuisance model. The nuisance model is **misspecified**. The true values for vectors α_1 and α_2, β are $(-0.5, 1)^{\text{T}}$, $(0.5, 1.5)^{\text{T}}$ and $(1, -0.5)^{\text{T}}$ respectively

Sample Size	500		1000	
	α_1	α_2	α_1	α_2
$\text{Bias}_{\times 10^2}(\text{SE}_{\times 10^2})$				
	-4.02(0.589)	-1.586(0.377)	-2.472(0.415)	-1.446(0.267)
	2.715(0.612)	1.244(0.518)	1.42(0.432)	-0.404(0.351)
SD Accuracy				
	1.02	1.059	1.006	1.048
	1.04	1.013	1.019	1.032
Coverage				
	0.963	0.958	0.946	0.958
	0.962	0.957	0.96	0.962

SE, standard error.

SD Accuracy = estimated standard deviation / Monte Carlo standard deviation.

Nominal level = 95%.

Table C.3: Monte Carlo simulation results based on 1000 runs for the relative risk model with a generalized odds product nuisance model. The nuisance model is **misspecified**. The true values for vectors α_1 and α_2, β are $(-0.5, 1)^{\mathrm{T}}$, $(0.5, 1.5)^{\mathrm{T}}$ and $(2, -3)^{\mathrm{T}}$ respectively

Sample Size	500		1000	
	α_1	α_2	α_1	α_2
$\mathrm{Bias}_{\times 10^2}(\mathrm{SE}_{\times 10^2})$				
	-19.883(0.615)	-13.062(0.332)	-18.455(0.414)	-13.428(0.232)
	5.551(0.578)	-19.416(0.454)	4.765(0.404)	-20.75(0.313)
SD Accuracy				
	0.959	1.207	0.991	1.218
	1.09	1.035	1.082	1.039
Coverage				
	0.84	0.861	0.723	0.702
	0.966	0.686	0.963	0.447

SE, standard error.

SD Accuracy = estimated standard deviation / Monte Carlo standard deviation.

Nominal level = 95%.

C.3 More simulations for model comparison

Table C.4: Simulation results for three different methods based on 5000 samples and 1000 Monte Carlo runs. The true values for α_1 and α_2 are $(-0.5, 1)^{\mathrm{T}}$ and $(0.5, 1.5)^{\mathrm{T}}$ respectively

Sample Size	Bias$_{\times 10^2}$(Standard Error$_{\times 10^2}$)		SD Accuracy		Coverage	
	α_1	α_2	α_1	α_2	α_1	α_2
5000	GOP					
	0.006(0.189)	0.122(0.134)	1.014	1.027	0.959	0.956
	−0.261(0.203)	0.099(0.136)	1.016	1.00	0.953	0.964
	DR-G (applied twice)					
	−0.566(0.216)	0.091(0.144)	1.027	1.026	0.958	0.948
	1.120(0.357)	1.219(0.210)	0.972	0.963	0.949	0.942
	OP (applied twice)					
	−0.002(0.194)	0.462(0.136)	1.009	1.027	0.953	0.961
	−0.203(0.230)	−0.798(0.142)	0.992	0.984	0.949	0.956

GOP: Using models (9) and (10); DR-G, doubly robust estimator by Dukes and Vansteelandt (2018); OP: Using nuisance model proposed by Richardson et al. (2017).
SD Accuracy = estimated standard deviation / Monte Carlo standard deviation.
95% nominal coverage.

Appendix D

A SENSITIVE ANALYSIS OF THE TITANIC DATA SET

We impute the missing via function `transcan` in `R`, which automatically transforms continuous variables to have maximum correlation with the best linear combination of the other variables; see more details in `transcan{Hmisc}`. After imputing the missing age, we repeat the analysis in Chapter 4.

Table D.1: Coefficient estimates via different models.

	2nd	2nd* male	2nd* age/10	2nd* age^2/ 100	2nd* male* age/10	1st	1st* male	1st* age/10	1st* age^2/ 100	1st* male* age/10
Point Estimate										
Monotone	0.236	0.050	0.038	-0.006	0.107	0.472	0.100	0.076	-0.013	0.213
GOP	0.787	0.451	-0.144	0.025	-0.439	0.494	0.502	0.080	-0.001	-0.013
Poisson	0.876	0.199	-0.228	0.039	-0.339	0.543	0.368	0.046	0.002	0.041
Standard Deviation										
Monotone	0.084	0.203	0.063	0.011	0.071	0.169	0.405	0.127	0.023	0.142
GOP	0.224	0.447	0.179	0.036	0.181	0.291	0.457	0.208	0.038	0.166
Poisson	0.435	0.579	0.315	0.058	0.234	0.542	0.590	0.321	0.052	0.191

1st, 2nd, 3rd: the first passenger class, the second passenger class, and the third passenger class. The first class is chosen as the baseline.

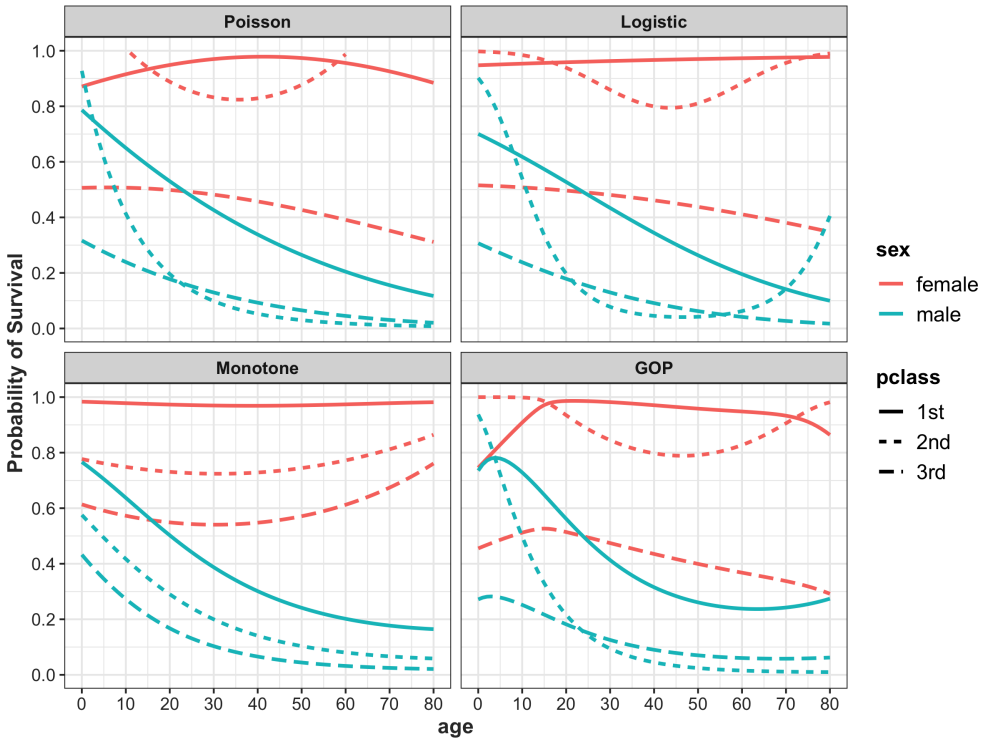


Figure D.1: Predicted probability of survival of the first passenger class (solid line), the second class (dotted line), and the third class (dashed line) with respect to different models. Red represents female, and blue represents male.

Appendix E

COMBINING CASE-CONTROL AND COHORT DATA

E.1 Variance Formula for Special Case

We consider the case of including one cohort study and one case-control study. The treatment Z is binary taking value in $\{0,1\}$. For notational simplification, we denote $\text{pr}(Z_{ij}=1 \mid v_{ij})$ as π_{ij}, $\text{pr}(Z_{ij}=1 \mid y_{ij}, v_{ij})$ as $q_{y_{ij}}$. Then Equation (5.8) can be rewritten as

$$l(\gamma,\beta,\eta) = \sum_{j=1}^{n_1} y_{1j}\log\{p_{z_{1j}}(v_{1j};\gamma,\beta)\} + (1-y_{1j})\log\{1-p_{z_{1j}}(v_{1j};\gamma,\beta)\} \tag{E.1}$$

$$+ \sum_{j=1}^{n_1} z_{1j}\log\pi_{1j} + (1-z_{1j})\log(1-\pi_{1j}) \tag{E.2}$$

$$+ \sum_{j=1}^{n_2} z_{2j}\log q_{y_{2j}} + (1-z_{2j})\log(1-q_{y_{2j}}). \tag{E.3}$$

We also let Equation (E.1), (E.2), and (E.3) as l_1, l_2, and l_3, respectively. To have the derivatives of $l(\gamma,\beta,\eta)$ with respect to γ, β, and η, we can calculate the derivatives of l_1, l_2, and l_3 respectively.

Via §2.2, or §3.2, we have

$$\frac{\partial l_1}{\partial\gamma} = \sum_{j=1}^{n_1} \frac{y_{1j}-p_{z_{1j}}}{1-p_{z_{1j}}} \cdot \left\{ z_{1j} - \frac{1-p_{0,1j}}{(1-p_{0,1j})+(1-p_{1,1j})} \right\} \cdot v_{1j}, \tag{E.4}$$

$$\frac{\partial l_1}{\partial\beta} = \sum_{j=1}^{n_1} \frac{y_{1j}-p_{z_{1j}}}{1-p_{z_{1j}}} \cdot \frac{(1-p_{0,1j})(1-p_{1,1j})}{(1-p_{0,1j})+(1-p_{1,1j})} \cdot v_{1j}, \tag{E.5}$$

$$\frac{\partial l_1}{\partial\eta} = 0. \tag{E.6}$$

The derivatives of l_2 with respect to γ,β,η are

$$\frac{\partial l_2}{\partial\gamma} = 0, \tag{E.7}$$

$$\frac{\partial l_2}{\partial\beta} = 0, \tag{E.8}$$

$$\frac{\partial l_2}{\partial\eta} = \sum_{j=1}^{n_1} \frac{z_{1j}-\pi_{1j}}{\pi_{1j}(1-\pi_{1j})} \cdot \frac{\partial\pi_{1j}}{\partial\eta}. \tag{E.9}$$

With propensity score model, we have $\frac{\partial\pi_{ij}}{\partial\gamma}=0$, $\frac{\partial\pi_{ij}}{\partial\beta}=0$, and

$$\frac{\partial\pi_{ij}}{\partial\eta} = \frac{e^{\eta^{\mathrm{T}}v_{ij}}v_{ij}}{(1+e^{\eta^{\mathrm{T}}v_{ij}})^2} = \pi_{ij}(1-\pi_{ij})v_{ij} \quad (j=1,\ldots,n_i; i=1,2). \tag{E.10}$$

Plug Equation (E.10) to Equation (E.9), we further have

$$\frac{\partial l_2}{\partial \eta} = \sum_{j=1}^{n_1} (z_{1j} - \pi_{1j}) v_{1j}. \tag{E.11}$$

The derivatives of l_3 with respect to γ, β, η are

$$\frac{\partial l_3}{\partial \gamma} = \sum_{j=1}^{n_2} \frac{z_{2j} - q_{y_{2j}}}{q_{y_{2j}}(1 - q_{y_{2j}})} \cdot \frac{\partial q_{y_{2j}}}{\partial \gamma} \tag{E.12}$$

$$\frac{\partial l_3}{\partial \beta} = \sum_{j=1}^{n_2} \frac{z_{2j} - q_{y_{2j}}}{q_{y_{2j}}(1 - q_{y_{2j}})} \cdot \frac{\partial q_{y_{2j}}}{\partial \beta} \tag{E.13}$$

$$\frac{\partial l_3}{\partial \eta} = \sum_{j=1}^{n_2} \frac{z_{2j} - q_{y_{2j}}}{q_{y_{2j}}(1 - q_{y_{2j}})} \cdot \frac{\partial q_{y_{2j}}}{\partial \eta}. \tag{E.14}$$

To have $\frac{\partial l_3}{\partial \gamma}$, $\frac{\partial l_3}{\partial \beta}$, and $\frac{\partial l_3}{\partial \eta}$, we need to calculate $\frac{\partial q_{y_{2j}}}{\partial \gamma}$, $\frac{\partial q_{y_{2j}}}{\partial \beta}$, and $\frac{\partial q_{y_{2j}}}{\partial \eta}$.

$$q_{y_{2j}} = \frac{\{y_{2j}p_{1,2j} + (1 - y_{2j})(1 - p_{1,2j})\}\pi_{2j}}{\{y_{2j}p_{1,2j} + (1 - y_{2j})(1 - p_{1,2j})\}\pi_{2j} + \{y_{2j}p_{0,2j} + (1 - y_{2j})(1 - p_{0,2j})\}(1 - \pi_{2j})}. \tag{E.15}$$

Let the denominator of the right hand side of Equation (E.15) as A, and the numerator as B. Then

$$\frac{\partial (A/B)}{\partial \cdot} = \frac{B \cdot \partial A/\partial \cdot - A \cdot \partial B/\partial \cdot}{B^2}.$$

We further have the derivatives to γ, β as

$$\frac{\partial A}{\partial \cdot} = (2y_{2j} - 1)\pi_{2j} \frac{\partial p_{1,2j}}{\partial \cdot} \tag{E.16}$$

$$\frac{\partial B}{\partial \cdot} = \frac{\partial A}{\partial \cdot} + (2y_{2j} - 1)\frac{\partial p_{0,2j}}{\partial \cdot}(1 - \pi_{2j}), \tag{E.17}$$

and

$$\frac{\partial A}{\partial \eta} = \{y_{2j}p_{1,2j} + (1 - y_{2j})(1 - p_{1,2j})\}\pi_{2j}(1 - \pi_{2j})v_{2j} \tag{E.18}$$

$$\frac{\partial B}{\partial \eta} = (2y_{2j} - 1)(p_{1,2j} - p_{0,2j})\pi_{2j}(1 - \pi_{2j})v_{2j} \tag{E.19}$$

By Equation (3.8), (3.9), (3.13), and, (3.14) in §3.2,

$$\frac{\partial p_z}{\partial \gamma} = p_z v \left(z - \frac{1 - p_0}{(1 - p_0) + (1 - p_1)} \right), \quad \frac{\partial p_z}{\partial \beta} = p_z v \frac{(1 - p_0)(1 - p_1)}{(1 - p_0) + (1 - p_1)}.$$

Plug above equations into (E.16) and (E.17) to have $\frac{\partial A}{\partial \gamma}$, $\frac{\partial A}{\partial \beta}$, $\frac{\partial B}{\partial \gamma}$, and $\frac{\partial B}{\partial \beta}$. Once we have each building blocks, we can calculate (E.12), (E.13), and (E.14). Then $\frac{\partial l}{\partial \cdot} = \sum_{i=1}^{3} \frac{\partial l_i}{\partial \cdot}$.

E.2 More simulations

Table E.1: Monte Carlo simulation results based on 1000 runs for combining cohort and case-control study. The true values of γ is $(0.5, -1)^{\mathrm{T}}$. The first two columns represent the number of subjects in each certain study.

		$\text{Bias}_{\times 10^2}(\text{Standard Error}_{\times 10^2})$	SD Accuracy	Coverage	Power	CI Width
Cohort	CC	γ	γ	γ	γ	γ
1000	0	5.246(1.592)	0.924	0.968	0.177	1.824
		-0.101(1.378)	0.934	0.954	0.689	1.594
1000	500	1.346(0.533)	0.998	0.953	0.869	0.659
		-0.091(0.494)	0.979	0.954	1.000	0.600
1500	0	5.322(1.237)	0.931	0.960	0.305	1.427
		-1.378(1.084)	0.939	0.944	0.837	1.262
1000	1000	0.374(0.403)	0.994	0.951	0.979	0.497
		0.601(0.363)	1.011	0.951	1.000	0.455
2000	0	3.056(0.977)	0.996	0.955	0.391	1.207
		-0.26(0.899)	0.963	0.950	0.929	1.074

CC, case-control study

SD Accuracy = estimated standard deviation / Monte Carlo standard deviation.

Nominal level = 95%.

Power, null hypothesis: $\gamma_0 = 0$, $\gamma_1 = 0$, separately.

CI width, confident interval width.

Table E.2: Monte Carlo simulation results based on 1000 runs. The true values of γ is $(0.5, 1)^{\mathrm{T}}$. The first two columns represent the number of subjects in each certain study.

		$\text{Bias}_{\times 10^2}(\text{Standard Error}_{\times 10^2})$	SD Accuracy	Coverage	Power	
Cohort	CC	γ	γ	γ	γ	γ
3000	0	2.208(0.79)	0.990	0.946	0.557	0.970
		-0.339(0.72)	0.972	0.951	0.992	0.867
3000	1000	0.574(0.368)	0.966	0.940	0.994	0.441
		0.346(0.332)	0.975	0.948	1.000	0.401
4000	0	1.785(0.679)	0.988	0.955	0.703	0.831
		0.036(0.593)	1.014	0.956	0.998	0.745
3000	2000	0.425(0.274)	0.994	0.961	1.000	0.337
		0(0.253)	0.982	0.952	1.000	0.308
5000	0	1.083(0.6)	0.996	0.950	0.793	0.740
		-0.225(0.547)	0.979	0.948	0.999	0.664
3000	3000	-0.13(0.232)	0.995	0.948	1.000	0.286
		0.188(0.209)	1.011	0.957	1.000	0.262
6000	0	1.562(0.565)	0.964	0.952	0.857	0.675
		-0.354(0.477)	1.024	0.963	1.000	0.606

CC, case-control study

SD Accuracy = estimated standard deviation / Monte Carlo standard deviation.

Nominal level = 95%.

Power, null hypothesis: $\gamma_0 = 0$, $\gamma_1 = 0$, separately.

CI width, confident interval width.

Table E.3: Monte Carlo simulation results based on 1000 runs. The true values of γ is $(0.5, 1)^{\mathrm{T}}$. The first two columns represent the number of subjects in each certain study.

		$\text{Bias}_{\times 10^2}(\text{Standard Error}_{\times 10^2})$	SD Accuracy	Coverage	Power	CI Width
Cohort	CC	γ	γ	γ	γ	
4000	0	2.02(0.703)	0.956	0.955	0.684	0.834
		-1.49(0.612)	0.984	0.954	0.997	0.746
4000	1000	0.151(0.342)	0.996	0.940	0.997	0.423
		0.052(0.321)	0.963	0.945	1.000	0.384
5000	0	1.119(0.617)	0.968	0.949	0.783	0.740
		0.076(0.528)	1.016	0.946	1.000	0.665
4000	2000	0.237(0.269)	0.979	0.949	1.000	0.327
		0.194(0.228)	1.052	0.959	1.000	0.298
6000	0	1.347(0.546)	0.995	0.954	0.872	0.673
		0.432(0.476)	1.025	0.950	1.000	0.605
4000	3000	0.024(0.223)	1.009	0.953	1.000	0.278
		0.047(0.204)	1.004	0.953	1.000	0.254
7000	0	1.22(0.497)	1.009	0.948	0.925	0.621
		0.43(0.46)	0.980	0.955	1.000	0.558
4000	4000	0.239(0.203)	0.984	0.944	1.000	0.248
		0.145(0.187)	0.977	0.944	1.000	0.227
8000	0	1.039(0.466)	1.007	0.948	0.944	0.582
		0.22(0.427)	0.987	0.953	1.000	0.523

CC, case-control study

SD Accuracy = estimated standard deviation / Monte Carlo standard deviation.

Nominal level = 95%.

Power, null hypothesis: $\gamma_0 = 0$, $\gamma_1 = 0$, separately.

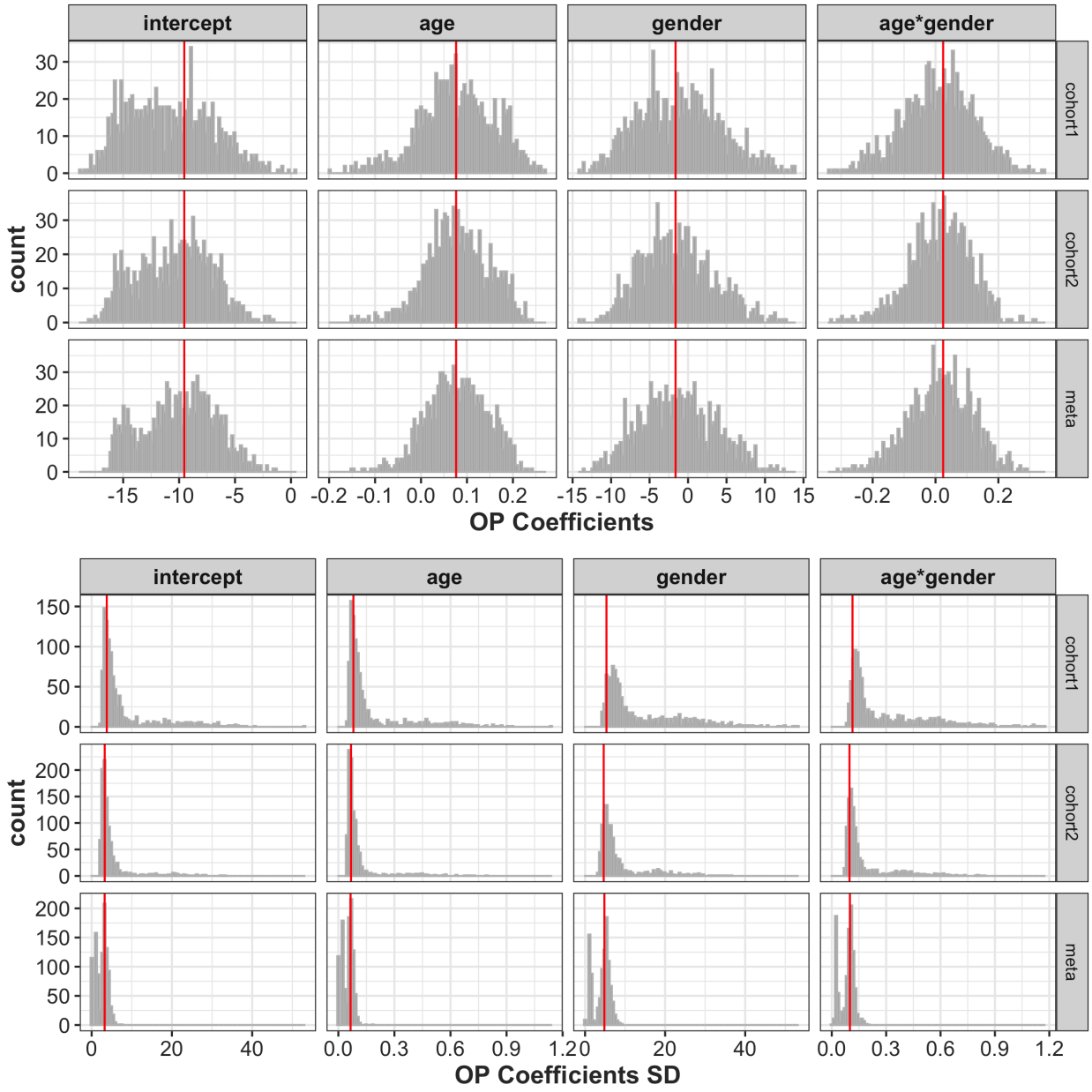


Figure E.1: The histogram of estimates from 1000 Monto Carlo runs. The above figure shows the estimated coefficients in Model 5.10. The bottom figure is the estimated standard deviation of each estimate. Cohort1 is the cohort study of size 300, cohort2 is size 450, and meta is the meta study of combining the cohort of size 300 and case-control study of size 150. The red vertical lines are the ground truth of the estimates, and Monte Carlo standard deviation respectively.

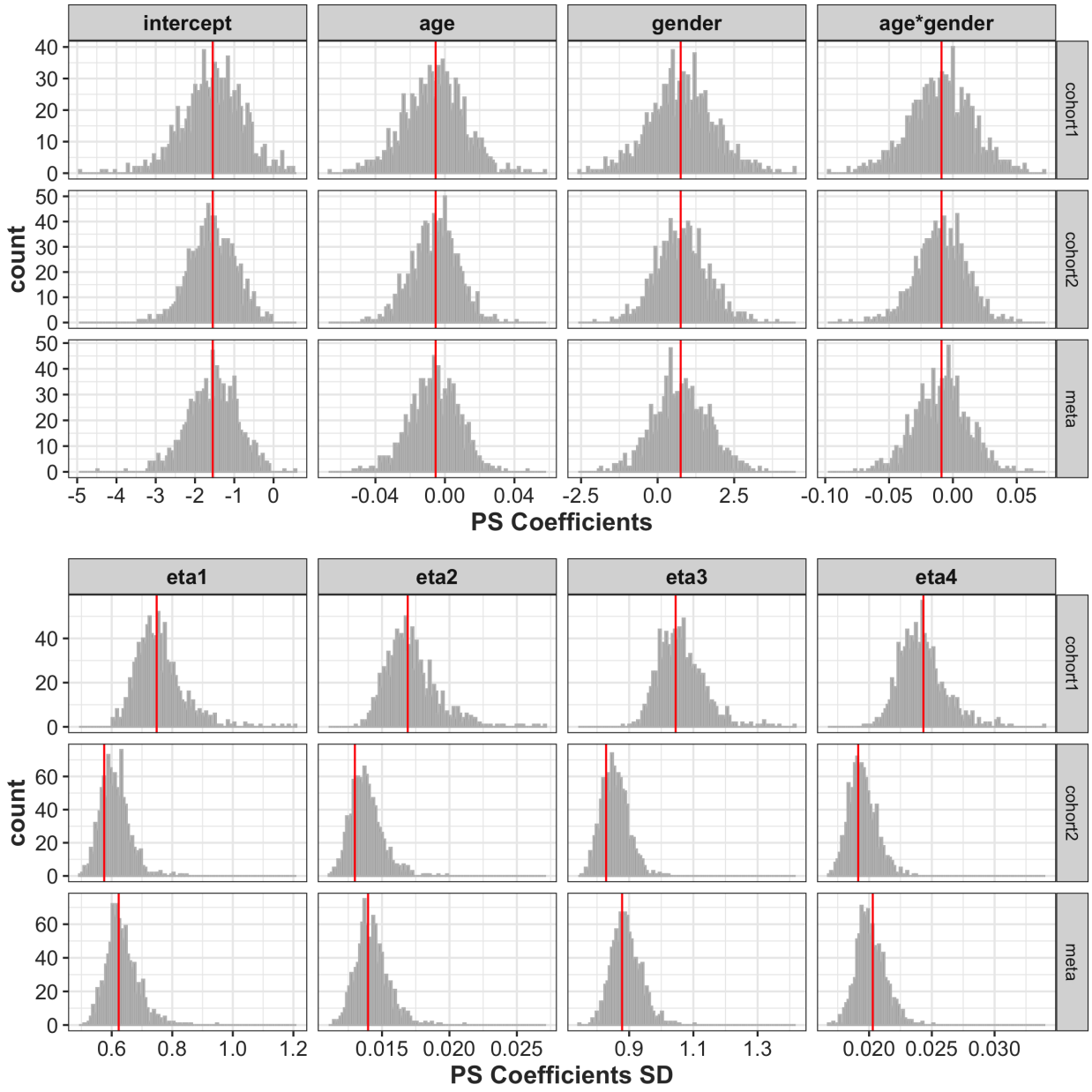


Figure E.2: The histogram of estimates from 1000 Monto Carlo runs. The above figure shows the estimated coefficients in Model 5.11. The bottom figure is the estimated standard deviation of each estimate. Cohort1 is the cohort study of size 300, cohort2 is size 450, and meta is the meta study of combining the cohort of size 300 and case-control study of size 150. The red vertical lines are the ground truth of the estimates, and Monto Carlo standard deviation respectively.

E.3 More simulations for application

Table E.4: Coefficient estimates for three different studies based on 1000 Monte Carlo runs. $n_1 = 200$, $n_2 = 200$.

	Intercept	Age/100	Gender	Age/100 * Gender
Bias(Standard Error)				
Chrt 200	-0.434(0.154)	-0.651(0.311)	0.547(0.24)	-0.407(0.496)
Chrt 4000	-0.323(0.117)	-0.349(0.229)	0.339(0.181)	-0.284(0.363)
Chrt 200 & CC 2000	0.216(0.045)	-0.463(0.091)	-0.305(0.072)	0.558(0.141)
SD Accuracy				
Chrt 200	2.534	2.759	2.767	3.004
Chrt 400	1.55	1.665	1.624	1.703
Chrt 200 & CC 200	1.038	1.05	1.011	1.021
Coverage (Nominal = 95%)				
Chrt 200	0.998	0.996	0.996	0.997
Chrt 400	0.991	0.994	0.991	0.984
Chrt 200 & CC 200	0.955	0.968	0.956	0.961
Power				
Chrt 200	0.017	0.002	0.005	0.004
Chrt 400	0.053	0.013	0.011	0.017
Chrt 200 & CC 200	0.203	0.079	0.05	0.065

Chrt, cohort study; CC, case-control study.

SD Accuracy = estimated standard deviation / Monte Carlo standard deviation.

Power, the null hypothesis: the coefficients are 0.

Table E.5: Coefficient estimates for three different studies based on 1000 Monte Carlo runs. $n_1 = 400$, $n_2 = 200$.

	Intercept	Age/100	Gender	Age/100 * Gender
Bias(Standard Error)				
Chrt 400	-0.195(0.12)	-0.432(0.236)	-0.034(0.179)	0.053(0.36)
Chrt 600	-0.164(0.096)	-0.152(0.194)	-0.2(0.146)	0.389(0.288)
Chrt 400 & CC 200	0.275(0.04)	-0.56(0.081)	-0.18(0.066)	0.368(0.128)
SD Accuracy				
Chrt 400	1.405	1.475	1.608	1.669
Chrt 600	1.116	1.126	1.234	1.242
Chrt 400 & CC 200	1.06	1.065	1.007	1.021
Coverage (Nominal = 95%)				
Chrt 400	0.987	0.987	0.994	0.99
Chrt 600	0.983	0.985	0.986	0.979
Chrt 400 & CC 200	0.965	0.963	0.96	0.962
Power				
Chrt 400	0.059	0.021	0.008	0.015
Chrt 600	0.078	0.023	0.018	0.027
Chrt 400 & CC 200	0.263	0.093	0.055	0.077

Chrt, cohort study; CC, case-control study.

SD Accuracy = estimated standard deviation / Monte Carlo standard deviation.

Power, the null hypothesis: the coefficients are 0.

Table E.6: Coefficient estimates for three different studies based on 1000 Monte Carlo runs. $n_1 = 600$, $n_2 = 200$.

	Intercept	Age/100	Gender	Age/100 * Gender
Bias(Standard Error)				
Chrt 600	-0.21(0.099)	-0.167(0.195)	0.094(0.151)	-0.074(0.302)
Chrt 800	-0.17(0.081)	-0.034(0.159)	0.056(0.124)	-0.003(0.243)
Chrt 600 & CC 200	0.151(0.037)	-0.336(0.075)	-0.018(0.061)	0.057(0.121)
SD Accuracy				
Chrt 600	1.136	1.176	1.173	1.181
Chrt 800	1.042	1.078	1.084	1.087
Chrt 600 & CC 200	1.085	1.09	1.023	1.015
Coverage (Nominal = 95%)				
Chrt 600	0.981	0.989	0.981	0.976
Chrt 800	0.983	0.986	0.983	0.977
Chrt 600 & CC 200	0.972	0.974	0.963	0.957
Power				
Chrt 600	0.095	0.028	0.023	0.03
Chrt 800	0.113	0.037	0.024	0.029
Chrt 600 & CC 200	0.247	0.088	0.052	0.07

Chrt, cohort study; CC, case-control study.

SD Accuracy = estimated standard deviation / Monte Carlo standard deviation.

Power, the null hypothesis: the coefficients are 0.

VITA

Jiaqi Yin was born in Xi'an, China, where she spent the first 18 years of her life. She graduated from Xi'an Gaoxin No.1 High School and attended Zhiyuan College, Shanghai Jiaotong University afterwards. After receiving her bachelor's degree in Math and Applied Math, she continued her education in the Department of Biostatistics at University of Washington. She completed her Ph.D in Biostatistics in June 2020 under the supervision of Professor Thomas S. Richardson and Assistant Professor Linbo Wang.