2021-05-01

Refined Moderation Analysis with Binary Outcomes

Eric Anto
University of Texas at El Paso

REFINED MODERATION ANALYSIS WITH BINARY OUTCOMES

ERIC ANTO

Master's Program in Statistics

APPROVED:

Xiaogang Su, Ph.D, Chair

Naijun Sha, Ph.D

Oralia Loza, Ph.D

Stephen L. Crites, Jr., Ph.D.
Dean of the Graduate School

To my

Brother BENJAMIN, Mother MARY, Uncle BENJAMIN and my Advisor Dr SU

with love

REFINED MODERATION ANALYSIS WITH BINARY OUTCOMES

by

ERIC ANTO

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

DEPARTMENT OF MATHEMATICAL SCIENCES

THE UNIVERSITY OF TEXAS AT EL PASO

May 2021

Acknowledgements

I would like to express my sincerest gratitude to my advisor and mentor, Professor Xiaogang Su of the Mathematical Science Department at The University of Texas at El Paso, for his invaluable support, expertise, unflinching love and guidance for making this work a great success. In fact, his readiness to provide clear explanations to some key concepts whenever the need arose was quite amazing and heroic. Despite his busy schedules, he was very responsive and easy to contact, hence there were never depressed moments working with him. Again, I also wish to thank the other members of my committee, Professor Naijun Sha of the Mathematical Science Department, and Professor Oralia Loza of the Department of Public Health Sciences both at The University of Texas at El Paso, for their commitments, suggestions, comments and additional guidance which were indispensable to the completion of this work. Finally, I wish to express great appreciation to my family, colleagues and friends for being there for me with their supports and encouragements through thick and thin.

Abstract

With the growing interest in personalized or precision medicine, it is indispensable that moderation analysis which is primarily related to the study of differential treatment effects among patients with different characteristics, also serves as the bedrock for precision medicine is taken more seriously. Concerning moderation analysis with binary outcomes, we start with an interesting observation, which shows that heterogeneous treatment effects could be equivalently estimated via a role exchange between the outcome and the treatment variable. The result holds for both experimental data and observational data, yet with an important difference in interpretation. Two estimators of moderating effects corresponding to two GLM models can be obtained. We combine the two models into a single model and employ the GEE approach to simultaneously obtain parameter estimates associated with the moderating effects (interaction terms), on which basis of refined inference can be made. The improved efficiency is helpful in addressing the lack-of-power problem that is common in the search for moderators. We investigate the proposed method by simulation and provide an illustration with data from a randomized trial concerning wart treatment. Essentially, the new revelation about the 'role swapping' technique can be useful by offering more flexible and computational ease in scenarios where direct modeling of interactions encounters difficulties and becomes inconvenient, owing to modeling complexity, numerical difficulty, or unavailability of implementation.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and Statement of Problem

Moderation analysis has gained a lot of referrals in different fields of study owning to the enormous advantages it offers in recent times. However, these advantages have never been completely or fully archived or perceived by analysts, clinicians, and students in training. The concept of moderation analysis is one of the theories employed to better refine and comprehend a causal relationship empirically. In fact, moderation analysis is fundamental to precision medicine or assessment of differential causal effects. It is highly needed in clinical examination settings today and a similar origination has been stretched out to investigations in different fields.

Customizing healthcare, treatments, medical decisions, or products tailored to a subgroup of patients is the new quest in the health industry which has gained a lot of traction in political domain. With this growing interest in personalized or precision medicine, researchers and clinicians look for best practices that are aimed at advancing and proposing models that seek to alleviate the "one-drug-fits-for-all" approach. Moderation analysis is basically concerned with differential treatment effects among patients with different characteristics, supplying the foundation for precision medicine. Best practices aimed at advancing precision in the medical field for instance, requires that a distinction is made to better identify significant moderators or mediators in order to understand their impacts on the treatment effects. In reviewing the concept of moderation analysis, some basic references are made to related ideas which include but not limited to subgroup investigation, prognostic or prescriptive components, stratified/individualized treatment effects, and

treatment-by-covariates interactions. Moderation analysis is plagued with the challenges of multicollinearity and lack of power.

1.2 Motivation of the Study

The primary goal of this study is aimed at investigating moderation analysis with a binary endpoint, where the effect of a binary treatment is estimated by relative risk (RR) or odds ratio (OR). The evaluation of the treatment effect relies on whether the information gathered is from a randomized trial or an observational study. We present an intriguing result related to moderation analysis, which shows that heterogeneous treatment effect can be studied by an interaction model that regresses the treatment variable on the response variable and covariates. In simple terms, the 'role swap' between the response and the treatment is harnessed to this effect. Meaningful interpretation can be achieved via the resultant interaction model using RR or OR, and this conclusion varies with the source of data and thus has several useful implications. Essentially, estimates of parameters with smaller standard errors could be obtained using this new approach or method of conducting moderation analysis. In situations where direct modeling of interactions encounters difficulties, the role-swapping technique offers a more flexible and computational ease.

1.3 Outline of Thesis

The remaining parts of the thesis are organized in this manner. Chapter 2 provides a literature review on moderation analysis. In Chapter 3, our proposed method, the role swapping technique and the refined moderation analysis via the generalized estimating equations (GEE) approach is presented and explained in detail. A simulation study that verifies the performance of the proposed method is analyzed and compared in Chapter 4. Chapter 5 provides a report on validation of our findings with an illustrative example from a randomized wart treatment trial (real-world data). Finally, a summary of our study and

future work directions are discussed in Chapter 6.

Chapter 2

Literature Review

In this chapter, we present a literature review on moderation analysis. We explore briefly literature related to moderation analysis which includes subgroup analysis, predictive or prescriptive factors, effect modification, treatment-by-covariates interactions and precision medicine as well as generalised and independent estimating equations.

2.1 Moderation Analysis

Three key concepts arise in research which are related to associations or relationships between variables- and these are confounding, mediation, and moderation (or effect modification, interaction). The study of mediation, moderation, and confounding analysis are methodologies that are employed to identify how a third variable may be incorporated in statistical analyses to explore the underlying mechanisms leading to the alteration of an effect on a subgroup of the population or specifically to ascertain the instances where an effect may be strong or weak. In the area of research, determining not only whether predictors affect the response, but then, how and when that association exists (whether strong or weak), is more crucial to our understanding of the outcome of an empirical study. The concept of moderator variables is very essential according to most researchers, however, confusion arises as to what a moderator variable particularly is and how it operates specifically to affect the classical validation model. This confusion in particular renders the comparison of studies of results relatively difficult. For instance, Bennett and Harrell (1975), Ghiselli (1963), and Hobert and Dunnette (1967) suggested a way that tries to ignore the interaction controversy by using an analytic procedure to examine differences

between individuals grouped on the basis of some hypothesized moderator variable.

Two types of moderator variables are known in literature: A homologizer variable, a termed coined by Zedeck (1971), is one that affects or influences the strength rather than the form of the relationship between predictors and outcome (criterion) variable. Pure and Quasi Moderator Variable on the other hand is one which modifies essentially the form of the relationship between the criterion and predictor variables (Sharma et al., 1981).
They proposed a framework for identifying the two types of moderator variables which incorporates moderated regression analysis(MRA) and subgroup analysis with a validation of results via Monte Carlo simulation.

Moderation analysis is a statistical and analytical technique used to ascertain whether the relationship (strength and size) between an independent variables (covariate) and the response or outcome depends a third variable called the moderator. A relationship may apply to some subgroup of a population but not to others and in such scenario, the relationship is said to be moderated. M is said to be a moderator, if the size, sign, or strength of the effect of X on some variable Y depends on or can be predicted by M. Thus M and X interact to influence Y. The primary goal in moderation analysis is to measure and test the differential effect of the independent variable (covariate) on the response (outcome) variable as a function of the moderator. Within the framework of correlational analysis, a third variable that has an effect on the zero-order association or correlation between two different other variables or better still the magnitude of the slope parameter of the response variable on the covariate (independent variable) is a moderator (Baron and Kenny, 1986). The interaction between a covariate (independent variable) and a factor that basically identifies the settings for its operation is usually employed as a yardstick to measure the moderator effect within the context of analysis of variance (ANOVA).

2.2 Treatment-by-Covariate Interactions

The key to understanding moderation is simply by modeling interactions between treatment and covariates. According to Royston and Sauerbrei (2008), in developing strategies or procedures for precision or personalized medicine, it is essential to identify the treatment and covariate interactions in the setting of randomized clinical trials. The interest of analysts and clinicians in research is to hypothesize the effect of a moderator, where the value of a moderator variable determines the effect of an independent variable on an outcome. This effect according to (Hayes and Matthes, 2009), reveals itself statistically as an interaction between the independent variable and moderator variables in a model of the outcome variable.

In regression analysis, interactions (i.e., product between predictors, usually between treatment and covariates) can be explored along with the original predictor variables where statistical significance is further be examined via hypothesis testing, e.g., the Wald test. For instance, suppose Paracetamol (X) has more effect on adults than on children in treating headache (Y), then we say that age (M), moderates the effects of the treatment. In this regard, moderation is shown by an interaction between X and M in predicting Y, which would show that the association or relationship between X and Y depends on the level of M. Describing the overall treatment effects without taking into account group membership could be misleading whenever there is interaction. A construction of a unique term associated with the interaction of X and M in moderation analysis is often employed to determine whether the contribution of this term goes beyond the main effects of X and M in predicting Y. In regression analysis, one way of quantifying this interactions mathematically is via the cross-product of X and M, which are then included in the model. In reality, complex interactions may also present themselves in higher order and in other forms.

There are two usual approaches which are the exploratory and analytic approaches that are practically used to categorize the potential treatment-by-covariate interactions. In the exploratory approach, a cohort of patient subgroup analyses is carried out, where a compar-

ison between the treatment and control arms is explored in the different subgroups defined a priori, such as young and adult, asthmatic and non-asthmatic may be performed which usually follows the main comparison. This approach focuses on mainly simple interactions between treatment and binary covariate and more often than not this leads to false-positive findings as a result of the multiple testings as well as failing to solve the problem of unmasking complicated treatment-by-covariates interactions. In the analytic approach, the binary treatment indicator and the baseline covariates are considered as a product and included in the regression model to examine the treatment-by-covariates interactions via multivariate regression analysis (Tian et al., 2014). Owning to the complexity in identifying and modeling directly (via multivariate regression) the interactions between treatment and high dimensional covariates, Tian et al. (2014), proposed to identify moderators with regularization.

2.3 Subgroup Analysis

It is crucial to a have guidelines in the emerging area of precision medicine that covers the issues of subgroup analyses to better understand how the framework of theory and methodology interacts with empirics. Subgroup analysis refers to any comparison of patient outcomes between treatment groups across subsets of patients defined by patient characteristics. Prevention and intervention studies requires that an investigation is launched to determine whether treatment effects vary among subgroups of patients defined by individual characteristics (Wang and Ware, 2013). Subgroup analysis is a technique that is used to provide an information on how to employ a new program related to prevention or intervention studies. Ferreira and Patino (2017) cautions on common mistakes that either lead to false negative or positive findings, especially when they are not pre-specified in the analysis plan. Various limitations regarding the use of the traditional subgroup analysis have been widely explored (Sleight, 2000; Assmann et al., 2000; Lagakos et al., 2006).

Owning to the controversies (the frequency of risky and inappropriate “false-positive” or

"false-negative" conclusions) that underscore decisions or results based on subgroup analyses (encountered frequently and routinely in reports of observational studies and clinical trials), this anomaly presents some level of difficulties for clinicians and physicians to make inferences from such conclusions. These false conclusions mainly stem from naive uses of subgroup analysis. New developments in subgroup analysis address the issues such as Type I and Type II errors in these naive approaches via cross validation. Though subgroup analyses are widely known to be highly misleading with the tendency to overemphasize results, they can in fact be very informative when results are accurate.

In this era where the quest for precision medicine in the healthcare industry is on the rampage and remains a topical issue, it is essential that guidelines regarding decisions based on subgroup analyses be clearly spelt out. Oxman and Guyatt (1992) presented extensive guidelines on when to act on recommendations based on subgroup analyses and when to ignore them particularly when deciding about how believable the results of subgroup analyses are. In their paper, they highlighted seven main guidelines that could reveal the strength of inference regarding a proposed difference in treatment effect among subgroups, which could be used as a yardstick for concluding on treatment decision based on overall results or on the results of a subgroup analysis.

2.4 Predictive versus Prognostic Factors

Definitions of the terms "predictive" and "prognostic" are rarely defined and more often than not used interchangeably, despite their wide usage in a large number of publications in describing relationships or associations that exist between biomarkers and clinical outcomes. According to Clark et al. (2006), a predictive factor is a measurement that is associated with response or lack of response (clinical endpoints commonly used in clinical trials) to a particular therapy while a prognostic factor is a measurement that is associated with clinical outcome in the absence of therapy or with the application of a standard therapy that patients are likely to receive. To clinicians and physicians in the medical world, the

knowledge about prognostic and predictive factors serves as an essential key in making well-informed decisions about a patient on specifically when to start, stop or change a particular therapy (i.e., prognostic factors) or which particular therapy to prescribe for a patient in question (i.e., prescriptive factors)

Prior to an inclusion of a particular factor in guidelines for treatment selection, it is essential to distinguish its prognostic effects from its ability to predict a differential clinical benefit from the specific treatment (Clark, 2008). Prognostic factors are indispensable in the healthcare of patients and research activities of various diseases such as cancer control as well as many clinical trials. Specifically, a prognostic factor in simple terms, measures or tracks the natural and underlying history of a particular disease and in such studies, a control group from a randomized clinical trial setting is ideal for evaluating the prognostic effect of a biomarker. In the framework of statistics and medical research, Clark (2008), a predictive factor is best evaluated in a randomized clinical trial with a control group and particularly constitutes an interaction between biomarker status and treatment benefit.

The assessment that leads to categorizing a baseline factor into prognostic or prescriptive Simms et al. (2013) is done usually in a prospective or a retrospective way, where respectively, a hypothesis testing prespecifying this analysis in the statistical analysis plan or a hypothesis-generating using exploratory analyses is carried out. Essentially, regression methods such as Cox regression (i.e., used for time-to event endpoints such as survival), logistic regression (i.e., used for binary endpoints such as death/survival outcome) and linear regression (i.e., used for continuous endpoints such as change in tumor size) are useful statistical techniques that are employed to determine whether a factor is prognostic and/or predictive.

2.5 Precision Medicine

With the growing interest in personalized or precision medicine, researchers and clinicians look for best practices that are aimed at advancing and proposing models that alleviate

the "one-drug-fits-for-all" approach and instead customize healthcare, treatments, medical decisions, or products tailored to a subgroup of patients.

Precision medicine as a field of medicine has been structured to consider differences in individual's genetic make-up, microbiomes, environments, family history, as well as their lifestyles when employing diagnostic and therapeutic approaches which are tailored specifically to individual patients (Zhang, 2015). Precision medicine is a newer term that has changed the face of the medical world in recent times and has been envisioned to improve healthcare delivery and treatment of diseases.

The term, 'personalized medicine' which is closely related to precision medicine according to Jain (2002) made waves in the public domain in 1999 when the first publication on the subject came into the limelight, with the creation of some of the core concepts related to the field even dating back to the early 1960s. Personalized medicine originates from the concept that selection of a particular treatment should not be geared towards a decision based on 'standards of care' which is only a derivation of averaging responses across large cohorts of individuals in clinical trials, but instead be tailored according to individual patient's specific characteristics, such as age, gender, height, weight, ethnicity, diet, and environment (Jain and Jain, 2009).

The business strategist, Clayton Christensen, of Harvard Business School in Boston, first coined the expression 'precision medicine' in 2008 which has become popular in the scientific dictionary today, to describe how molecular diagnostics allows health professionals to diagnose the cause of a disease in an unambiguous way without necessarily relying on intuition. In particular, the name gained traction and became more popular in 2011 after a blueprint aimed at modernizing the taxonomy of disease based on molecular information rather than a symptom-based classification system was laid out a by committee of the US National Research Council (Katsnelson, 2013).

Zhang (2011) noted that, following the report of the US National Research Council published in 2011 on the topic, "Toward Precision Medicine: Building a Knowledge Network

for Biomedical Research and a New Taxonomy of Disease", some scientists and researchers began to retire the phrase "personalized medicine" and replace it with "precision medicine", alluding to the fact that molecular information improves the precision with which patients are categorized (based on certain characteristics) and treated (Katsnelson, 2013).

The Obama administration on January 30, 2015, unveiled details about the Precision Medicine Initiative. With a $215 investment in the US President's 2016 budget, the Precision Medicine Initiative is aimed to champion a new model of patient-powered research that will essentially help deliver the required treatment to the right patient at the right time (The White House: Office of the Press Secretary, 2015). This is a big feat and very innovative idea on the part of government in ensuring a very robust healthcare customization and effective control and treatment of diseases by taking into account the variation that exists in individual's environments, genetic make-up, and lifestyles. According to Jameson and Longo (2015), the most daunting challenge that is associated with precision medicine is how to manage the complexity that underscores the progressively refined classification of disease and thus this complexity associated with data supporting precision medicine will require that provisions are made within the health systems to provide diagnostics, informatics, and decision to buttress healthcare providers.

2.6 Independent Estimating Equations versus Generalized Estimating Equations

As a matter of concern, researchers sometimes come face-to-face with the challenge in statistical modeling of longitudinal or clustered data which normally requires that successive measurements on the same or related subjects enrolled in clinical studies.
In statistical modeling, failing to take into account the correlation between repeated measures often leads to making invalid inferences as estimates of parameters may not be con-

sistent and precise (Dobson and Barnett, 2018).

The Generalized Estimating Equations (GEE) and Independence Estimating Equations (IEE) have gained a lot referrals over the past two decades owning to their wide usage and application in longitudinal and clustered data (where repeated observations for a subject are known to be correlated) analysis. With the growing interest of research in the biomedical and health sciences where mostly clustered binary data occur, generalized estimating equations (GEE) as a technique to analysing multivariate binary responses (Liang and Zeger, 1986; Zeger and Liang, 1986) is employed, where the complete specification of the joint distribution of the responses is not required. Like the GEE, the independence estimating equations (IEE) estimator is based on the assumption that the responses are independent, which has be shown by Fitzmaurice (1995) to be nearly efficient relative to the maximum likelihood estimation in a varied number of settings and also outlined an instance where ignoring the independence assumption can cause substantial losses of efficiency of parameter estimates. Particularly, in situations where the correlation between responses is moderate to weak, Zeger (1988) suggests that this estimator should be nearly efficient.

For the regression analysis of correlated observations, Liang and Zeger (1986) introduced the GEE approach which is an extension of the generalised linear model (GLMs) and in spite of the fact that the responses on units within the same cluster are usually (positively) correlated, ordinary logistic regression employs maximum likelihood estimation (which assumes the within- cluster responses are independent) yields estimates which are consistent and asymptotically normal. An interesting fact about the estimator of the GEE approach is that, it is statistically consistent even if the working correlation structure is misspecified. Just like the GEE approach, the penalized GEE procedure requires a specification of the first couple of marginal moments as well as a working correlation matrix. This approach generalized the estimation method of quasi-likelihood of Wedderburn (1974) to correlated data. The quasi-likelihood as a function, has similar properties like that of the log-likelihood function, however, it does not conform to any actual or exact probability distribution. The GEE is not bounded by the exponential family assumption

that underscores the use of GLM; here, the specification of only the first couple of moments is required. In particular, with a quasi-likelihood function, a relation between the mean and variance of the observations (where the variance is given as a function of the mean) is usually specified, and used for estimation. Ziegler et al. (1998) gave an extensive review of the development of the GEE approach. The quasi-likelihood methods provide relatively more computational ease, robustness as well as speed, because the methods use more direct and developed algorithms to fit GLMs. For high-dimensional correlated data, specifying the full joint likelihood is not required, making it more appealing for modeling correlated discrete responses (Wang et al., 2012).

Chapter 3

Proposed Method (Refined Moderation Analysis with Binary Outcomes)

3.1 Role-Swapping

Let $\left\{(y_i, t_i, x_i)\right\}_{i=1}^{n}$ be an n-independent and identically distributed copies of (y, t, x), where both y and t are binary variables and $\mathbf{x} \in \mathbb{R}^p$ be a p-dimensional covariate vector. The variable y is measured on 0/1 as the response or outcome with value 1 indicating the occurrence of an event of interest (such as death) and 0 indicating the absence of the event (e.g., survival) and t, the 0/1 treatment assignment variable with value 1 for the treated and 0 for the untreated. In the domain of clinicians, participants with the event are termed as *cases* and those without the condition/event are otherwise referred to as *controls*. The focus is to evaluate the efficacy or effectiveness of t on y. The two commonly used scales for measuring the effectiveness of treatment on an outcome are the relative risks (RR) scale and the odds ratio (OR) scale.

For a fixed covariates at $\mathbf{X} = \mathbf{x}$, two measures associated with RR, depending on whether the death rate or the survival rate is considered and could be defined as,

$$RR_{1,\mathbf{x}} = \frac{\Pr(Y = 1|T = 1, \mathbf{X} = \mathbf{x})}{\Pr(Y = 1|T = 0, \mathbf{X} = \mathbf{x})}$$

and

$$RR_{0,\mathbf{x}} = \frac{\Pr(Y = 0|T = 1, \mathbf{X} = \mathbf{x})}{\Pr(Y = 0|T = 0, \mathbf{X} = \mathbf{x})},$$

In the same vein, a measure of OR is given by,

$$OR_{\mathbf{x}}^{(Y)} = \frac{\Pr(Y=1|T=1,\mathbf{X}=\mathbf{x})/\Pr(Y=0|T=1,\mathbf{X}=\mathbf{x})}{\Pr(Y=1|Y=0,\mathbf{X}=\mathbf{x})/\Pr(Y=0|Y=0,\mathbf{X}=\mathbf{x})} = \frac{RR_{1,\mathbf{x}}}{RR_{0,\mathbf{x}}}. \tag{3.1}$$

It could be seen from the above results that $OR_{\mathbf{x}}^{(Y)} \approx RR_{1,\mathbf{x}}$ when cases (with $Y=1$) are rare or $\Pr(Y=0|T,\mathbf{X}) \approx 1$ and $OR_{\mathbf{x}}^{(Y)} \approx 1/\mathrm{RR}_{0,\mathbf{x}}$ when controls (with $Y=0$) are rare or $\Pr(Y=1|T,\mathbf{X}) \approx 1$.

It is worth noting that, a measure of the differential treatment effects among patients with heterogeneous characteristics is the primary objective in the studies related to moderation analysis. In evaluating the differential treatment effects, the ratio of relative risks (RRR) and the ratio of odds ratio (ROR) are commonly used as the natural measures. Let examine two sets of patients or individuals; one with covariates $\mathbf{x}$ and the other with covariates $\mathbf{x}'$. By defining the RRR for death or case ($Y=1$) between individuals with covariates $\mathbf{x}$ and individuals with covariates $\mathbf{x}'$,we have,

$$RRR^{(Y=1)}(\mathbf{x}:\mathbf{x}') = \frac{RR_{1,\mathbf{x}}}{RR_{1,\mathbf{x}'}} = \frac{\Pr(Y=1|T=1,\mathbf{X}=\mathbf{x})/\Pr(Y=1|T=0,\mathbf{X}=\mathbf{x})}{\Pr(Y=1|T=1,\mathbf{X}=\mathbf{x}')/\Pr(Y=1|Y=0,\mathbf{X}=\mathbf{x}')} \tag{3.2}$$

In a similar scenario for the survival or control ($Y=0$), we have,

$$RRR^{(Y=0)}(\mathbf{x}:\mathbf{x}') = \frac{RR_{0,\mathbf{x}}}{RR_{0,\mathbf{x}'}} = \frac{\Pr(Y=0|T=1,\mathbf{X}=\mathbf{x})/\Pr(Y=0|T=0,\mathbf{X}=\mathbf{x})}{\Pr(Y=0|T=1,\mathbf{X}=\mathbf{x}')/\Pr(Y=0|Y=0,\mathbf{X}=\mathbf{x}')} \tag{3.3}$$

A key note to observe is that, if $RR_{1,\mathbf{x}} \geq 1$, then we must have $RR_{0,\mathbf{x}} \leq 1$ related to assessing the treatment effect; however, if $RRR_1(\mathbf{x}:\mathbf{x}') \geq 1$ does not necessarily mean a one-to-one correspondence such that $RRR_0(\mathbf{x}:\mathbf{x}') \leq 1$ concerning moderation.

On the other hand, by comparing the OR of individuals with covariates $\mathbf{x}$ against those with covariates $\mathbf{x}'$, which is a measure of the ratio of odds ratio (ROR) can be defined as,

$$ROR^{(Y)}(\mathbf{x}:\mathbf{x}') = \frac{OR_{\mathbf{x}}^{(Y)}}{OR_{\mathbf{x}'}^{(Y)}} = \frac{RR_{1,\mathbf{x}}/RR_{0,\mathbf{x}}}{RR_{1,\mathbf{x}'}/RR_{0,\mathbf{x}'}} = \frac{RRR^{(Y=1)}(\mathbf{x}:\mathbf{x}')}{RRR^{(Y=0)}(\mathbf{x}:\mathbf{x}')}. \tag{3.4}$$

It is clear to note that defining separate $ROR^{(Y)}(\mathbf{x}:\mathbf{x}')$ for *cases* and for *controls* is not necessary owning to the invariance property of odds ratio. Particularly, these are simply reciprocal of each other.

By employing the role-swap of T and Y, the odds ratios for event $T = 1$ that compares cases ($Y = 1$) with $\mathbf{X} = \mathbf{x}$ and cases with $\mathbf{X} = \mathbf{x}'$ is defined as,

$$OR^{(T)}_{Y=1}(\mathbf{x} : \mathbf{x}') = \frac{\Pr(T = 1|Y = 1, \mathbf{X} = \mathbf{x})/\Pr(T = 0|Y = 1, \mathbf{X} = \mathbf{x})}{\Pr(T = 1|Y = 1, \mathbf{X} = \mathbf{x}')/\Pr(T = 0|Y = 1, \mathbf{X} = \mathbf{x}')} \tag{3.5}$$

In the same manner, the odds ratios for event $T = 1$ that compares controls ($Y = 0$) with $\mathbf{X} = \mathbf{x}$ and controls with $\mathbf{X} = \mathbf{x}'$ is defined as,

$$OR^{(T)}_{Y=0}(\mathbf{x} : \mathbf{x}') = \frac{\Pr(T = 1|Y = 0, \mathbf{X} = \mathbf{x})/\Pr(T = 0|Y = 0, \mathbf{X} = \mathbf{x})}{\Pr(T = 1|Y = 0, \mathbf{X} = \mathbf{x}')/\Pr(T = 0|Y = 0, \mathbf{X} = \mathbf{x}')} \tag{3.6}$$

In reference to Rosenbaum and Rubin (1983), let $\pi(\mathbf{x}) = \Pr(T = 1|\mathbf{X} = \mathbf{x})$ denote the propensity score. We have the following lemma.

Lemma 1. *Concerning moderation analysis on the RRR scale,*

$$RRR^{(Y=1)}(\mathbf{x} : \mathbf{x}') = OR^{(T)}_{Y=1}(\mathbf{x} : \mathbf{x}') \cdot \frac{\pi(\mathbf{x}')/(1 - \pi(\mathbf{x}'))}{\pi(\mathbf{x})/(1 - \pi(\mathbf{x}))}. \tag{3.7}$$

and

$$RRR^{(Y=0)}(\mathbf{x} : \mathbf{x}') = OR^{(T)}_{Y=0}(\mathbf{x} : \mathbf{x}') \cdot \frac{\pi(\mathbf{x}')/(1 - \pi(\mathbf{x}'))}{\pi(\mathbf{x})/(1 - \pi(\mathbf{x}))}. \tag{3.8}$$

Concerning moderation analysis on the ROR scale,

$$ROR^{(Y)}(\mathbf{x} : \mathbf{x}') = ROR^{(T)}(\mathbf{x} : \mathbf{x}'), \tag{3.9}$$

where $ROR^{(T)}(\mathbf{x} : \mathbf{x}')$ is the ratio of odds ratio (ROR) after the role exchange of T and Y.

If we further define the odds ratio on propensity as

$$OR^{(T)}(\mathbf{x} : \mathbf{x}') = \frac{\pi(\mathbf{x})/(1 - \pi(\mathbf{x}))}{\pi(\mathbf{x}')/(1 - \pi(\mathbf{x}'))}, \tag{3.10}$$

equations (3.8) and (3.7) in Lemma 1 can be rewritten as

$$RRR^{(Y=0)}(\mathbf{x} : \mathbf{x}') = \frac{OR^{(T)}_{Y=0}(\mathbf{x} : \mathbf{x}')}{OR^{(T)}(\mathbf{x} : \mathbf{x}')} \quad \text{and} \quad RRR^{(Y=1)}(\mathbf{x} : \mathbf{x}') = \frac{OR^{(T)}_{Y=1}(\mathbf{x} : \mathbf{x}')}{OR^{(T)}(\mathbf{x} : \mathbf{x}')}.$$

The results of Lemma 1 hold no matter whether data are obtained from a randomized experiment or from an observational study. Moreover, the OR-based equation (3.9) is applicable to retrospective matched case control studies; (see, e.g., Hosmer Jr et al., 2013).

This propensity score is a constant in study of randomized experiments. That is for $\pi \in (0,1)$, $\pi(\mathbf{x}) = \pi(\mathbf{x}')$. Note that Equation (3.7), simplifies to $RRR^{(Y=1)}(\mathbf{x} : \mathbf{x}') = OR^{(T)}_{Y=1}(\mathbf{x} : \mathbf{x}')$, and in literature, it is commonly called "the case-only analysis." In the case of rare events, $RRR^{(Y=1)}(\mathbf{x} : \mathbf{x}') \approx ROR^{(Y)}(\mathbf{x} : \mathbf{x}')$, implying that $RRR^{(Y=1)}(\mathbf{x} : \mathbf{x}')$ can be roughly construed as $ROR^{(Y)}(\mathbf{x} : \mathbf{x}')$. This phenomenon is first observed by Piegorsch et al. (1994) and its applications and extensions are further explored by several authors including Vittinghoff and Bauer (2006), Dai et al. (2014), Dai et al. (2018), and Dai and LeBlanc (2019). Clearly, control-only analysis could be proceeded similarly by reducing Equation (3.8) to $RRR^{(Y=0)}(\mathbf{x} : \mathbf{x}') = OR^{(T)}_{Y=0}(\mathbf{x} : \mathbf{x}')$. With rare control events, $1/RRR^{(Y=0)}(\mathbf{x} : \mathbf{x}') \approx ROR^{(Y)}(\mathbf{x} : \mathbf{x}')$.

The above RRR and ROR quantities can be naturally connected to generalized linear models (GLM), as prescribed by the following two propositions.

Proposition 1. *Consider the usual logistic regression model for moderation analysis on the odds ratio (OR) scale, where the conditional distribution of* $Y \mid T, \mathbf{X}$ *is formulated by*

$$\log \frac{\Pr(Y = 1|T, \mathbf{X})}{\Pr(Y = 0|T, \mathbf{X})} = \beta_0 + \beta_1 T + \mathbf{X}^T \boldsymbol{\beta}_2 + T \cdot \mathbf{X}^T \boldsymbol{\beta}_3. \tag{3.11}$$

Suppose that the conditional distribution of $T \mid Y, \mathbf{X}$ *is formulated by the logistic regression model*

$$\log \frac{\Pr(T = 1|Y, \mathbf{X})}{\Pr(T = 0|Y, \mathbf{X})} = \alpha_0 + \alpha_1 Y + \mathbf{X}^T \boldsymbol{\alpha}_2 + Y \cdot \mathbf{X}^T \boldsymbol{\alpha}_3. \tag{3.12}$$

Then we must have

$$\boldsymbol{\beta}_3 = \boldsymbol{\alpha}_3,$$

regardless of whether data come from a randomized experiment or from an observational study.

Proposition 1 indicates that the moderation analysis can be conducted via the interaction model (3.12) that regresses the treatment variable T on the outcome Y and covariates $\mathbf{X}$. This amounts to a role exchange between the treatment variable T and the outcome variable Y. This result generally holds for both experimental data and observational data,

under one necessary condition that that the same set of covariates $\mathbf{X}$ is used in both models of (3.11) and (3.12). Since model (3.12) involves an inverse regression by swapping the roles of T and Y, we refer to its related analysis as the 'inverse method' for simplicity while the approach based on model (3.11) is referred to as the 'direct method'.

Model (3.12) may seem surprising at first glance. One key to understand the inverse method is to note that the distribution of $T|\mathbf{X}$ is different from that of $T|\mathbf{X}, Y$. For example, $T \perp\!\!\!\perp \mathbf{X}$ when T is randomized, but $T \not\perp\!\!\!\perp \mathbf{X} \,|\, Y$. For randomized experiments, additional interpretations concerning moderation analysis on the relative risk scale can be extracted from model (3.12), as stated in Proposition 2.

Proposition 2. *For experiments where the assignment mechanism of treatment T is random, consider either of the following two log-linear regression models for moderation analysis on the relative risk (RR) scale, one for the control event and the other for the case event,*

$$\begin{cases} \log \Pr(Y = 0|T, \mathbf{X}) = \gamma_0^{(0)} + \gamma_1^{(0)} T + \mathbf{X}^T \boldsymbol{\gamma}_2^{(0)} + T \cdot \mathbf{X}^T \boldsymbol{\gamma}_3^{(0)}, \\ \log \Pr(Y = 1|T, \mathbf{X}) = \gamma_0^{(1)} + \gamma_1^{(1)} T + \mathbf{X}^T \boldsymbol{\gamma}_2^{(1)} + T \cdot \mathbf{X}^T \boldsymbol{\gamma}_3^{(1)}. \end{cases} \tag{3.13}$$

Furthermore assume that model (3.12) formulates the conditional distribution of $T \,|\, Y, \mathbf{X}$. Then we must have

$$\boldsymbol{\gamma}_3^{(0)} = \boldsymbol{\alpha}_2 \quad \textit{or} \quad \boldsymbol{\gamma}_3^{(1)} = \boldsymbol{\alpha}_2 + \boldsymbol{\alpha}_3.$$

Proposition 2 essentially breaks model (3.12) into two equations,

$$\log \frac{\Pr(T = 1|Y, \mathbf{X})}{\Pr(T = 0|Y, \mathbf{X})} = \begin{cases} \alpha_0 + \mathbf{X}^T \boldsymbol{\alpha}_2, & \text{if } \ Y = 0; \\ (\alpha_0 + \alpha_1) + \mathbf{X}^T (\boldsymbol{\alpha}_2 + \boldsymbol{\alpha}_3), & \text{if } \ Y = 1, \end{cases}$$

one for controls only and the other for cases only. It requires, either of, not necessarily both, the two models in (3.13) hold. Since $\Pr(Y = 0|T, \mathbf{X}) + \Pr(Y = 1|T, \mathbf{X}) = 1$, one is sufficient. If both models in (3.13) hold, then the parameters can be reduced by introducing

constraints. In this case, model (3.11) must hold with $\boldsymbol{\alpha}_j = \boldsymbol{\gamma}_j^{(1)} - \boldsymbol{\gamma}_j^{(0)}$ for $j = 0, 1, 2, 3$, but not *vice versa*.

The proofs of the above lemma and propositions are outlined in the Appendix and essentially involve applications of Bayes's rule. The similar arguments have been used to derive the case-only analysis (Piegorsch et al., 1994) and show equivalence of the conditional odds ratio obtained from a prospective study and that from a retrospective study (see, e.g., Hosmer Jr et al., 2013). As seen from the proofs, the results are also directly applicable to categorical outcomes and treatments with multiple levels.

In terms of inverse regression with role swapping, Efron (1975) has shown that the logistic regression for $Y|\mathbf{X}$ have the same coefficients as the Gaussian linear discriminant analysis that models $\mathbf{X}|Y$, under the assumption that $\mathbf{X}$ follows a multivariate normal distribution. Our results are derived in the same spirit of Bayes' rule, but are quite different in that the logistic model (3.11) for $Y|T, \mathbf{X}$ has the same set of coefficients only for interaction terms as the logistic model (3.12) for $T|Y, \mathbf{X}$, regardless of the distribution of $\mathbf{X}$.

We have shown that moderation analysis could be equivalently obtained by swapping the roles of the response and the treatment in theory. One should be aware that the specific estimates of $\boldsymbol{\beta}_3$ in Model (3.11) and $\boldsymbol{\alpha}_3$ in Model (3.12) could be different since logistic regression is solved numerically, but this difference is inconsequential. Nevertheless, the standard errors (SE) of these estimates could be systematically different, when switching the roles of response and treatment. The difference in SE is relevant and meaningful. This implies that the role swapping strategy facilitates a new estimation method for the parameters involved in moderation analysis, in which possibly more precise estimates could be obtained.

3.2 Refined moderation analysis with GEE

We have a clear trajectory or path on how moderation analysis could be equivalently done by employing the role-swapping technique between the response and the treatment variable.

This gives an insight on possibly how inference could be made on the moderating effects in a more efficient way. Refined estimators associated with the interaction terms are attainable by combining both Model (3.11) and Model (3.12) and estimating the resulting model equation via GEE the approach. This approach generalized the estimation method of the quasi-likelihood of Wedderburn (1974) to correlated data which has similar properties like that of the log-likelihood function, however, it does not conform to any actual or exact probability distribution and also provides relatively more computational ease, robustness as well as speed, because the methods use more direct and developed algorithms to fit GLMs. In particular, with a quasi-likelihood function, a relation between the mean and variance of the observations (where the variance is given as a function of the mean) is usually specified, and used for estimation.

Table 3.1: Data Layout for the Direct Model

ID (i)	**Response** (Y_i)	**Treatment**(T_i)	**Covariates** (X_{ij})			
1	Y_1	T_1	X_{11}	X_{12}	$\ldots$	X_{1p}
2	Y_2	T_2	X_{21}	X_{22}	$\ldots$	X_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
n	Y_n	T_n	X_{n1}	X_{n2}	$\ldots$	X_{np}

Table 3.2: Data Layout for the Inverse Model

ID (i)	**Response** (T_i)	**Treatment** (Y_i)	**Covariates** (X_{ij})			
1	T_1	Y_1	X_{11}	X_{12}	$\ldots$	X_{1p}
2	T_2	Y_2	X_{21}	X_{22}	$\ldots$	X_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
n	T_n	Y_n	X_{n1}	X_{n2}	$\ldots$	X_{np}

Table 3.3: Data Layout (Set-Up) for the GEE Model

ID (i)	j	**Response** (Y'_{ij})	**Treatment** (T'_{ij})	**Covariates** (X'_{ijl})			
1	1	Y'_{11}	T'_{11}	X_{111}	X_{112}	$\dots$	X_{11p}
	2	Y'_{12}	T'_{12}	X_{121}	X_{122}	$\dots$	X_{12p}
2	1	Y'_{21}	T'_{21}	X_{211}	X_{212}	$\dots$	X_{21p}
	2	Y'_{22}	T'_{22}	X_{221}	X_{222}	$\dots$	X_{22p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$
k	1	Y'_{k1}	T'_{k1}	X_{k11}	X_{k12}	$\dots$	X_{k1p}
	2	Y'_{k2}	T'_{k2}	X_{k21}	X_{k22}	$\dots$	X_{k2p}

To serve the purpose of the study which aims at making inference that is more efficient on the interaction terms, a combined dataset is churned out from the two data structure corresponding to the direct and indirect models in Table (3.1) and Table (3.2) .
Using the fact that $Y'_{ij} = Y_i$ if $j = 1$ and $Y'_{ij} = T_i$ if $j = 2$ and $j = 1, 2$. Similarly $T'_{ij} = T_i$ if $j = 1$ and $T'_{ij} = Y_i$ for $j = 2$. with $i = 1, \dots, n$, we obtain the layout for the data in Table (3.3) to fit for the GEE model. The basic data layout for the GEE model presented in the table is such that we have are repeated measures (2 observations) for K subjects so that we have a total of $2K$ observations.

We consider a longitudinal or clustered data consisting of K subjects or clusters. Now for each subject i, $i = 1, 2, \dots, K$, there are $n_i = 2$ observations and Y_{ij} denotes the jth observation from the ith subject with $j = 1 \dots kn_i$, and let $\mathbf{X}_{ij} = X_{ij1}, X_{ij2}, \dots, X_{ijp}$ denote a $p \times 1$ vector of covariates.
Denote $Y_i = (Y_{i1}, Y_{i2} \dots Y_{i2n_i})^T$ to be the response vector for the ith subject with mean vector given by $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{i2n_i})^T$, where $\boldsymbol{\mu}_{ij}$ is the corresponding jth mean and independence of responses across subjects/clusters but correlated within each subject/cluster

are assumed. Here, the marginal model specifies a relation between $\boldsymbol{\mu}_{ij}$ and the covariates $\mathbf{X}_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijp})^T$ which is given by:

$$\mathbf{g}(\mu_{ij}) = \text{logit}\Pr(Y_{ij} = 1|\mathbf{X}_{ij}, T_i) = \log \frac{\Pr\big(Y'_{ij} = 1|T'_{ij}, \mathbf{X}_i\big)}{\Pr\big(Y'_{ij} = 0|T'_{ij}, \mathbf{X}_i\big)} = \mathbf{X}_{ij}^T\boldsymbol{\beta}, \tag{3.14}$$

where $\mathbf{g}$ denote the logit link function and $\boldsymbol{\beta}$ is an unknown $p \times 1$ vector of regression coefficients with the true value as $\boldsymbol{\beta}_0$. We denote the conditional variance of Y_{ij} given X_{ij} as $\nu(\mu_{ij})\boldsymbol{\phi}$, where ν is a specified variance function of the marginal mean, $\boldsymbol{\mu}_{ij}$ and $\boldsymbol{\phi}$ is a scale parameter to be estimated. Usually, ν and $\boldsymbol{\phi}$ depend on the distributions of outcome variable. For e.g., given that Y_{ij} is continuous, $\nu(\mu_{ij})$ is specified to be 1, while $\boldsymbol{\phi}$ denotes the error variance. Similarly, for a count response, Y_{ij}, we have that $\nu(\mu_{ij}) = \boldsymbol{\mu}_{ij}$, with $\boldsymbol{\phi} = 1$. In our setting, we have a logistic model and binary data, hence $var(Y_i) = \mu_{ij}(1-\mu_{ij})$ with scale parameter, $\boldsymbol{\phi} = 1$.

Let V_i denote the variance-covariance matrix for Y_i, where $V_i = \boldsymbol{\phi} D_i^{1/2} R_i(\boldsymbol{\lambda}) D_i^{1/2}$ with $D_i = \text{diag}\{\nu(\mu_{i2}) \ldots \nu(\mu_{in_i})\}$, and $R_i(\boldsymbol{\lambda})$ represents the working correlation structure that describes the pattern of measures within cluster or subjects which depends on a vector of association parameters, $\boldsymbol{\lambda}$.

With biological data sets, (ÖNDER et al., 2010 & Park and Shin, 1999), revealed independent and exchangeable correlation structures as good candidates. In particular, we specify the independence correlation matrix, $R_i(\boldsymbol{\lambda}) = I$, a $T \times T$ identity matrix with no values of $\boldsymbol{\lambda}$ to be estimated, because the intra-cluster correlation is actually considered to be zero but then yields estimates similar to those from simple "pooled" models (Zorn, 2001). Thus variance-covariance matrix simply becomes $V_i = \text{diag}(\mu_{ij}(1 - \mu_{ij}))$.

Asymptotically, GEE yields a consistent $\widetilde{\boldsymbol{\beta}}$ even in the case where the working correlation structure is misspecified Liang and Zeger (1986) and estimating the value of $\boldsymbol{\beta}$ requires solving the estimating equation below:

$$\mathbf{Q}(\boldsymbol{\beta}) = \sum_{i=1}^{K} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}^T}\right)^T \mathbf{V}_i^{-1}(Y_i - \mu_i) = 0 \tag{3.15}$$

$\widetilde{\boldsymbol{\beta}}$ is asymptotically normally distributed with a mean $\boldsymbol{\beta}_0$ with a covariance matrix estimated

based on the sandwich estimator and this holds under mild regularity conditions.Thus,

$$\widehat{\boldsymbol{\Sigma}_i} = \left[\sum_{i=1}^{K}\left(\frac{\partial\mu_i}{\partial\boldsymbol{\beta}^T}\right)^T V_i^{-1}\left(\frac{\partial\mu_i}{\partial\boldsymbol{\beta}^T}\right)\right]^{-1}\widehat{M}\left[\sum_{i=1}^{K}\left(\frac{\partial\mu_i}{\partial\boldsymbol{\beta}^T}\right)^T V_i^{-1}\left(\frac{\partial\mu_i}{\partial\boldsymbol{\beta}^T}\right)\right]^{-1} \tag{3.16}$$

where,

$$\widehat{M} = \sum_{i=1}^{K}\left(\frac{\partial\mu_i}{\partial\boldsymbol{\beta}^T}\right)^T V_i^{-1} Cov(Y_i) V_i^{-1}\left(\frac{\partial\mu_i}{\partial\boldsymbol{\beta}^T}\right) \tag{3.17}$$

with $cov(Y_i) = (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)^T$, which is an estimator of the variance-covariance matrix of Y_i (Liang and Zeger,1986 and Qu et al., 2000). The robustness of the sandwich estimator is such that it is consistent even if the correlation structure, V_i is misspecified. It is however worth noting that, if V_i is correctly specified, then $\widehat{\boldsymbol{\Sigma}_i}$ reduces to:

$$\left[\sum_{i=1}^{K}\left(\frac{\partial\mu_i}{\partial\boldsymbol{\beta}^T}\right)^T V_i^{-1}\left(\frac{\partial\mu_i}{\partial\boldsymbol{\beta}^T}\right)\right]^{-1}$$

, which is called the model-based variance estimator (Kauermann and Carroll, 2001).

3.3 Computation using available GEE packages

The goal is to make a more efficient inference on the moderating effects. To this effect, we consider and use the R package *"geepak"*, which provides a flexible approach to estimating equations and also estimating the covariates depending mean, scale and correlations parameters of correlated observations (Halekoh et al., 2006).

3.4 Computing Trick

More generally, since $\boldsymbol{\alpha}_3 = \boldsymbol{\beta}_3$, we rewrite Model (3.11) and Model (3.12) as below;

$$\begin{cases} \log\dfrac{\Pr(Y_i=1|T_i,\mathbf{X}_i)}{\Pr(Y_i=0|T_i,\mathbf{X}_i)} = \beta_0 + \beta_1 T_i + \mathbf{X}_i^T\boldsymbol{\beta}_2 + T_i\cdot\mathbf{X}_i^T\boldsymbol{\beta}_3 \\ \log\dfrac{\Pr(T_i=1|Y_i,\mathbf{X}_i)}{\Pr(T_i=0|Y_i,\mathbf{X}_i)} = \alpha_0 + \alpha_1 Y_i + \mathbf{X}_i^T\boldsymbol{\alpha}_2 + Y_i\cdot\mathbf{X}_i^T\boldsymbol{\beta}_3. \end{cases} \tag{3.18}$$

To estimate simultaneously the interaction coefficients, with $\boldsymbol{\alpha}_3 = \boldsymbol{\beta}_3$ via the GEE approach, we formulate a model that combines Model (3.11) and Model (3.12) into model equation.

Let $Y'_{ij} = Y_i$ if $j = 1$ and $Y'_{ij} = T_i$ if $j = 2$ for $i = 1, \ldots, n$ and $j = 1, 2$. Similarly, let $T'_{ij} = T_i$ if $j = 1$ and $T'_{ij} = Y_i$ for $j = 2$. And define $Z_{ij} = 0$ if $j = 1$ and $Z_{ij} = 1$ if $j = 2$. Let $\mathbf{X}_{ij} = \mathbf{X}_i$. Consider the model below:

$$\begin{aligned} \log \frac{\Pr\big(Y'_{ij} = 1 | T'_{ij}, \mathbf{X}_i\big)}{\Pr\big(Y'_{ij} = 0 | T'_{ij}, \mathbf{X}_i\big)} &= \gamma_0 + \gamma_1 Z_{ij} + \gamma_2 T'_{ij} + \mathbf{X}_i^T \boldsymbol{\gamma}_3 + \\ &\quad + \gamma_4 T'_{ij} \cdot Z_{ij} + \mathbf{X}_i^T \boldsymbol{\gamma}_5 \cdot Z_{ij} + T'_{ij} \cdot \mathbf{X}_i^T \boldsymbol{\gamma}_6 \end{aligned} \tag{3.19}$$

where $\boldsymbol{\gamma}_6$ denotes the coefficients associated with the interaction term in Model (3.19) is equivalent to $\boldsymbol{\beta}_3$ in Model (3.18).

It can further be observed that, with $j = 1$ and $Z_{ij} = 0$, we have a result from Model (3.19), which is equivalent to Model (3.11), where $\boldsymbol{\gamma}_0 = \boldsymbol{\beta}_0$, $\boldsymbol{\gamma}_2 = \boldsymbol{\beta}_1$, $\boldsymbol{\gamma}_3 = \boldsymbol{\beta}_2$ and $\boldsymbol{\gamma}_6 = \boldsymbol{\beta}_3$. Similarly, with $j = 2$ and $Z_{ij} = 1$, a similar result is obtained from (3.19) which is equivalent to the Model (3.12). In particular, we have that $(\boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1) = \boldsymbol{\alpha}_0$, $(\boldsymbol{\gamma}_2 + \boldsymbol{\gamma}_4) = \boldsymbol{\alpha}_1$, $(\boldsymbol{\gamma}_3 + \boldsymbol{\gamma}_5) = \boldsymbol{\alpha}_2$ and $\boldsymbol{\gamma}_6 = \boldsymbol{\alpha}_3$, with an interesting observation showing that $\boldsymbol{\gamma}_6 = \boldsymbol{\alpha}_3 = \boldsymbol{\beta}_3$.

The above deductions clearly shows how we can combine the two Models in (3.11) and (3.12) into Model (3.19) and employ the GEE approach to simultaneously obtain a more refined parameter estimates associated with the interaction terms.

Chapter 4

Simulation Studies

4.1 Numerical Results

We validate the finding that the role-swapping yields estimates of the same parameters and illustrate the difference in standard errors by simulation.

4.1.1 Simulation Studies

We generate data from model (3.11), with $\mathbf{X} = (X_1, X_2, X_3)^T$, $\beta_0 = -1$, $\beta_1 = 0.5$, $\boldsymbol{\beta}_2 = (0.5, -1, 1)^T$, and $\boldsymbol{\beta}_3 = (-0.5, 1, 0)^T$. Three covariates generated from multivariate uniform[0, 1] distribution with correlation matrix $\left(\rho^{|j-j'|}\right)$ for $j, j' = 1, 2, 3$ with different choices of $\rho = \{0, 0.2, 0.5, 0.8\}$; see Falk (1999) and implementation in R (Team et al., 2013) package **MultiRNG**. To mimic observational studies, the binary treatment T is generated from the following logistic model

$$\log \frac{\Pr(T = 1|\mathbf{X})}{\Pr(T = 0|\mathbf{X})} = \theta_0 + \mathbf{X}^T \boldsymbol{\theta}_1, \tag{4.1}$$

where $\theta_0 = -0.5$ and $\boldsymbol{\theta}_1 = (1, -0.5, 1)^T$. To have data from randomized experiments, we set $\boldsymbol{\theta}_1 = \mathbf{0}$, as well as $\theta_0 = 0$.

For each model configuration, a number of sample sizes with $n \in (100, 200, \ldots 1000)$ are examined and for each setting, a total of 1,000 simulation runs are considered. For each simulated data set, we fit model (3.11), model (3.12) together with model (3.19) and extract the estimates $\widehat{\boldsymbol{\beta}}_3$, $\widehat{\boldsymbol{\alpha}}_3$ and $\widetilde{\boldsymbol{\beta}}_3$ of coefficients associated with the interaction terms, as well as some performance measures such as the standard errors (SE). In particular, $\widehat{\boldsymbol{\beta}}_3$ and $\widehat{\boldsymbol{\alpha}}_3$ and $\widetilde{\boldsymbol{\beta}}_3$ are the empirical estimates of $\boldsymbol{\beta}_3 = (-0.5, 1, 0)^T$. Essentially, setting $\boldsymbol{\beta}_{33} = 0$ helps

in the evaluation of the 'size' issue in hypothesis testing, while the two nonzero coefficients ($\boldsymbol{\beta}_{31}$ and $\boldsymbol{\beta}_{32}$) help address the 'power' issue. In this regard, the p-values from the Wald z test for each coefficient is also computed. The empirical power and size are calculated as the proportion of p-values smaller than the set significance level, $\alpha = 0.05$.

4.2 Verifying the Equivalence Between Direct and Inverse Estimators and Comparing Results to the GEE Estimator

The performance measures employed in evaluating the effectiveness and precision of the estimated means of the model parameters include the standard deviation (SD) of the estimates, the averaged standard error value as well as the empirical size and power associated with interaction terms. Respectively, Tables 4.1 and 4.2 present the aggregated results corresponding to randomized experimental data and observational data over 1000 simulation runs for each setting with $\rho \in \{0, 0.5, 0.8\}$ and $n \in \{100, 500, 800\}$. It can be seen that the direct estimates $\widehat{\boldsymbol{\beta}}_3$ and the inverse estimates $\widehat{\boldsymbol{\alpha}}_3$ are generally close to each other. The estimation performance improves with larger sample sizes. In terms of precision of these estimates, the inverse method gives smaller averaged standard errors than the direct approach most of the time. More generally, $\widehat{\boldsymbol{\alpha}}_3$ and $\widehat{\boldsymbol{\beta}}_3$ match each other by and large. In fact, it is worth noting that, there exist some level of bias in estimating moderation parameters (mean value of estimates) in both the direct and inverse methods in some scenarios especially in the case of observational data. This can be seen for e.g., when $n = 100$, $n = 500$ and $n = 1000$ at different levels of the correlation ρ among covariates.

However, by analyzing and comparing estimates of both the direct and indirect with those obtained by the GEE approach (i.e, $\widetilde{\boldsymbol{\beta}}_3$), it is clear that the performance measures (the mean and standard deviation (SD) of the estimates together with the averaged SE value) corresponding to the GEE approach produces relatively precise and better estimates

to a large extent.This revelation for the most part is true for randomized experimental data. Unfortunately, there exist some estimation gaps using the GEE approach for observational data in the same setting as shown in Table (4.2). This may partly be due to the issue of confounders that may be present since observational studies are not randomized to eliminate imbalances due to chance. An attempt via the propensity scores approach was employed to adjust or account for any issues of confounding that may be present with the study, however, the method could not yield any much promising results.

The evaluation of the size and power issue associated with moderation analysis has always been a major issue in hypothesis testing. It is well known in any statistical analysis or inference that, when assessing the issue associated with 'power', higher values constitute a high probability to reject a false null hypothesis and particularly in our case, that is the hypothesized values of $\boldsymbol{\beta}_{31}$ and $\boldsymbol{\beta}_{32}$ (i.e., reducing type II error). Though these values are generally low, the empirical power estimates associated with the GEE, for the most part are higher. Similarly, with the 'size' issue, it is observed that most of the estimates are within the neighborhood of (or close to) the nominal level (the set significance level, i.e., 0.05 in our setting) which is mostly desired in hypothesis testing (i.e., with $\boldsymbol{\beta}_{33} = 0$ in our scenario) which in practice offers the maximum probability of committing a Type I error.

Table 4.1: Simulation results with randomized experimental data based on 1,000 simulation runs. The regression coefficient $\boldsymbol{\beta}_3$ associated with moderation analysis were estimated with three methods: (I) the direct estimator $\widehat{\boldsymbol{\beta}}_3$ with model (3.11); (II) the inverse estimator $\widehat{\boldsymbol{\alpha}}_3$ with model (3.12) and (III) the GEE estimator $\widetilde{\boldsymbol{\beta}}_3$ with Model (3.19).

			Estimate (Mean)			SD			ASE (Mean)			Size & Power		
ρ	n		D	I	GEE	D	I	GEE	D	I	GEE	D	I	GEE
		β_{31}	−0.585	−0.604	−0.590	1.775	1.776	1.732	1.630	1.624	1.572	0.072	0.068	0.079
	100	β_{32}	1.050	1.024	1.028	1.779	1.794	1.741	1.640	1.630	1.577	0.099	0.098	0.118
		β_{33}	−0.087	0.114	0.107	1.810	1.766	1.746	1.647	1.636	1.581	0.053	0.052	0.057
		β_{31}	−0.518	−0.511	−0.514	0.673	0.664	0.665	0.664	0.657	0.656	0.119	0.108	0.116
0	500	β_{32}	1.031	1.010	1.019	0.680	0.673	0.673	0.668	0.660	0.659	0.333	0.324	0.332
		β_{33}	0.002	0.018	0.011	0.653	0.647	0.647	0.669	0.663	0.661	0.043	0.044	0.043
		β_{31}	−0.499	−0.486	−0.493	0.515	0.512	0.511	0.521	0.515	0.515	0.159	0.161	0.164
	800	β_{32}	0.982	0.959	0.970	0.536	0.531	0.531	0.523	0.517	0.517	0.475	0.462	0.470
		β_{33}	0.022	0.034	0.028	0.535	0.530	0.530	0.525	0.520	0.519	0.057	0.056	0.058
		β_{31}	−0.565	−0.541	−0.550	1.979	1.975	1.925	1.865	1.861	1.792	0.062	0.054	0.064
	100	β_{32}	1.078	1.069	1.066	2.308	2.311	2.246	2.086	2.081	2.001	0.089	0.093	0.109
		β_{33}	−0.017	−0.024	−0.015	2.023	2.015	1.967	1.881	1.874	1.811	0.065	0.061	0.076
		β_{31}	−0.528	−0.528	−0.528	0.789	0.783	0.782	0.757	0.752	0.749	0.122	0.119	0.120
0.5	500	β_{32}	1.016	0.997	1.005	0.851	0.845	0.843	0.843	0.835	0.832	0.229	0.223	0.232
		β_{33}	0.015	0.020	0.018	0.768	0.762	0.761	0.762	0.756	0.753	0.043	0.049	0.051
		β_{31}	−0.514	−0.506	−0.510	0.590	0.586	0.586	0.593	0.589	0.588	0.133	0.132	0.135
	800	β_{32}	1.048	1.034	1.040	0.654	0.647	0.647	0.660	0.654	0.653	0.349	0.349	0.354
		β_{33}	−0.021	−0.016	−0.018	0.593	0.586	0.587	0.597	0.592	0.591	0.047	0.045	0.048
		β_{31}	−0.662	−0.640	−0.647	2.845	2.997	2.812	2.672	2.685	2.566	0.054	0.054	0.064
	100	β_{32}	1.188	1.167	1.166	3.549	3.579	3.451	3.392	3.388	3.238	0.056	0.060	0.071
		β_{33}	−0.060	−0.062	−0.053	2.921	2.908	2.829	2.682	2.669	2.552	0.053	0.060	0.064
		β_{31}	−0.518	−0.511	−0.514	1.118	1.122	1.116	1.069	1.066	1.060	0.080	0.085	0.085
0.8	500	β_{32}	1.025	1.017	1.020	1.419	1.433	1.420	1.358	1.352	1.344	0.125	0.128	0.131
		β_{33}	−0.014	−0.015	−0.013	1.074	1.090	1.078	1.073	1.069	1.064	0.050	0.056	0.054
		β_{31}	−0.515	−0.513	−0.514	0.821	0.818	0.817	0.840	0.837	0.834	0.083	0.077	0.081
	800	β_{32}	0.993	0.984	0.988	1.094	1.083	1.085	1.064	1.058	1.055	0.157	0.161	0.159
		β_{33}	0.032	0.034	0.034	0.883	0.875	0.876	0.842	0.837	0.835	0.058	0.058	0.058

Table 4.2: Simulation results with observational data based on 1,000 simulation runs. The regression coefficient $\boldsymbol{\beta}_3$ associated with moderation analysis were estimated with three methods: (I) the direct estimator $\widehat{\boldsymbol{\beta}}_3$ with model (3.11); (II) the inverse estimator $\widehat{\boldsymbol{\alpha}}_3$ with model (3.12) and (III) the GEE estimator $\widetilde{\boldsymbol{\beta}}_3$ with Model (3.19).

			Estimate (Mean)			SD			ASE (Mean)			Size & Power		
ρ	n		D	I	GEE	D	I	GEE	D	I	GEE	D	I	GEE
		β_{31}	−0.508	−0.408	−0.481	1.717	1.922	1.705	1.683	1.849	1.646	0.064	0.051	0.065
	100	β_{32}	1.072	1.201	1.094	1.837	2.124	1.846	1.711	1.899	1.675	0.103	0.091	0.113
		β_{33}	−0.064	0.013	−0.027	1.782	1.975	1.777	1.641	1.786	1.604	0.058	0.063	0.064
		β_{31}	−0.487	−0.475	−0.483	0.699	0.748	0.703	0.684	0.731	0.686	0.114	0.101	0.116
0	500	β_{32}	1.053	1.041	1.043	0.710	0.755	0.712	0.695	0.742	0.697	0.338	0.284	0.315
		β_{33}	0.016	0.032	0.024	0.675	0.731	0.685	0.667	0.713	0.671	0.058	0.062	0.062
		β_{31}	−0.505	−0.476	−0.493	0.539	0.575	0.542	0.536	0.572	0.539	0.156	0.135	0.150
	800	β_{32}	1.004	1.008	1.003	0.548	0.581	0.550	0.545	0.580	0.547	0.445	0.390	0.444
		β_{33}	0.006	0.012	0.009	0.536	0.577	0.543	0.523	0.558	0.526	0.058	0.061	0.062
		β_{31}	−0.567	−0.440	−0.542	2.079	2.403	2.077	1.923	2.135	1.873	0.068	0.063	0.080
	100	β_{32}	1.222	1.358	1.230	2.250	2.690	2.279	2.147	2.384	2.084	0.075	0.079	0.102
		β_{33}	−0.049	−0.084	−0.053	1.940	2.290	1.949	1.867	2.117	1.831	0.051	0.050	0.059
		β_{31}	−0.507	−0.502	−0.507	0.784	0.854	0.799	0.776	0.828	0.780	0.106	0.104	0.114
0.5	500	β_{32}	1.038	1.017	1.025	0.879	0.928	0.880	0.866	0.924	0.868	0.214	0.195	0.211
		β_{33}	−0.006	0.003	−0.001	0.742	0.800	0.744	0.758	0.832	0.766	0.046	0.045	0.041
		β_{31}	−0.470	−0.443	−0.458	0.602	0.639	0.607	0.608	0.648	0.613	0.116	0.102	0.117
	800	β_{32}	1.011	0.999	1.003	0.657	0.721	0.670	0.678	0.721	0.681	0.322	0.292	0.323
		β_{33}	−0.011	0.003	−0.004	0.573	0.632	0.583	0.595	0.652	0.602	0.046	0.039	0.040
		β_{31}	−0.388	−0.202	−0.342	2.968	3.382	2.950	2.724	3.010	2.647	0.054	0.048	0.062
	100	β_{32}	1.051	1.230	1.069	3.909	4.426	3.886	3.459	3.833	3.341	0.065	0.059	0.083
		β_{33}	−0.037	−0.142	−0.055	2.918	3.381	2.904	2.691	3.025	2.616	0.057	0.052	0.059
		β_{31}	−0.496	−0.499	−0.500	1.102	1.167	1.109	1.086	1.158	1.093	0.076	0.077	0.086
0.8	500	β_{32}	1.036	1.034	1.030	1.407	1.500	1.416	1.375	1.471	1.383	0.125	0.106	0.119
		β_{33}	−0.013	−0.002	−0.006	1.101	1.208	1.119	1.073	1.176	1.085	0.045	0.044	0.049
		β_{31}	−0.569	−0.548	−0.561	0.860	0.901	0.862	0.851	0.906	0.858	0.108	0.098	0.100
	800	β_{32}	1.088	1.092	1.087	1.085	1.176	1.104	1.077	1.150	1.086	0.170	0.152	0.165
		β_{33}	−0.030	−0.032	−0.030	0.854	0.932	0.867	0.841	0.919	0.852	0.049	0.049	0.049

4.2.1 Assessment of Sample Size Effect on Simulated Results via Graphical Display

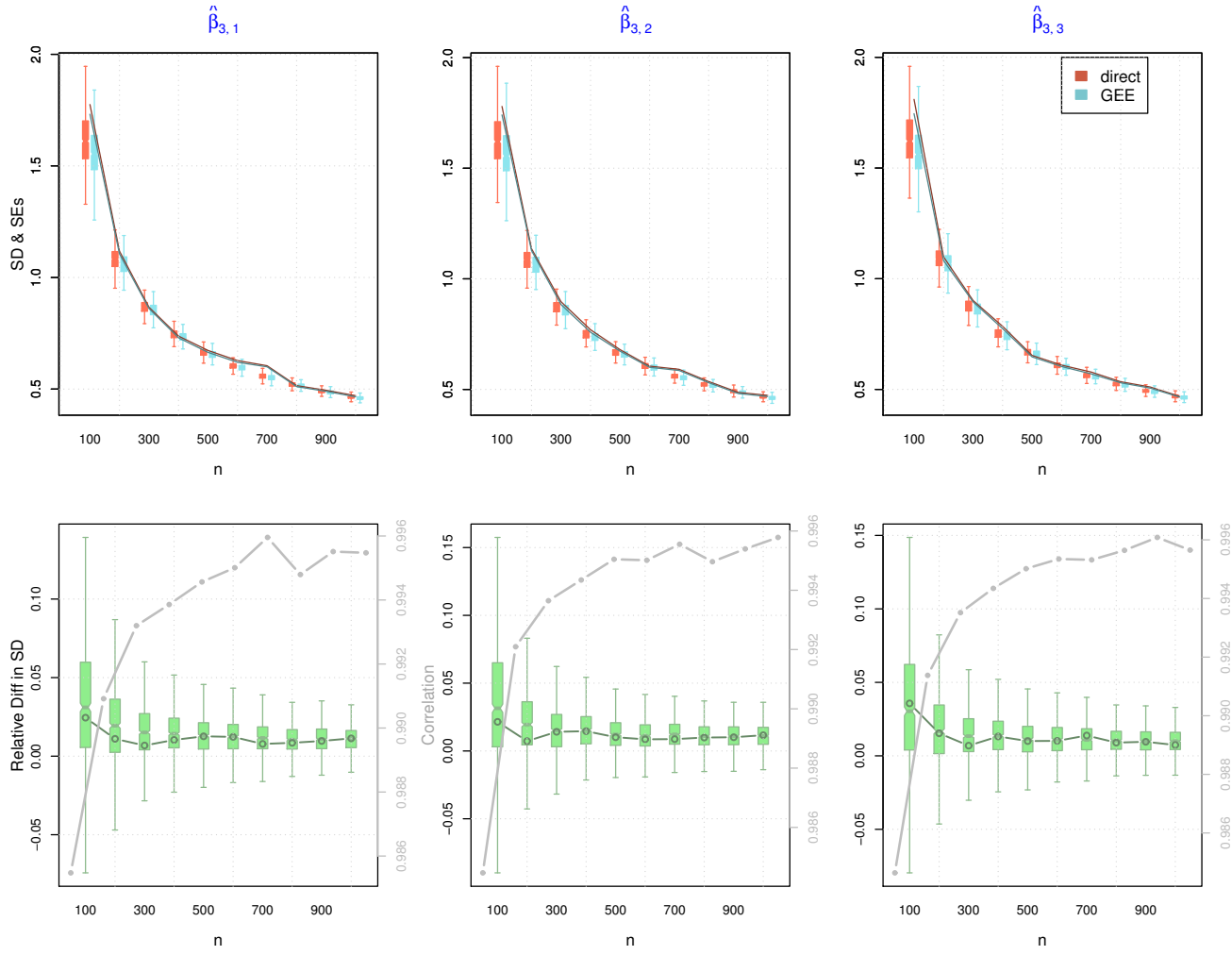

Figure 4.1: Simulation results for assessing the effect of sample size of randomized experimental data with $\rho = 0$.

The box plots in the first case show the standard error estimates associated with each of the 1000 simulation runs, with the lines showing the corresponding standard deviation in each scenario.

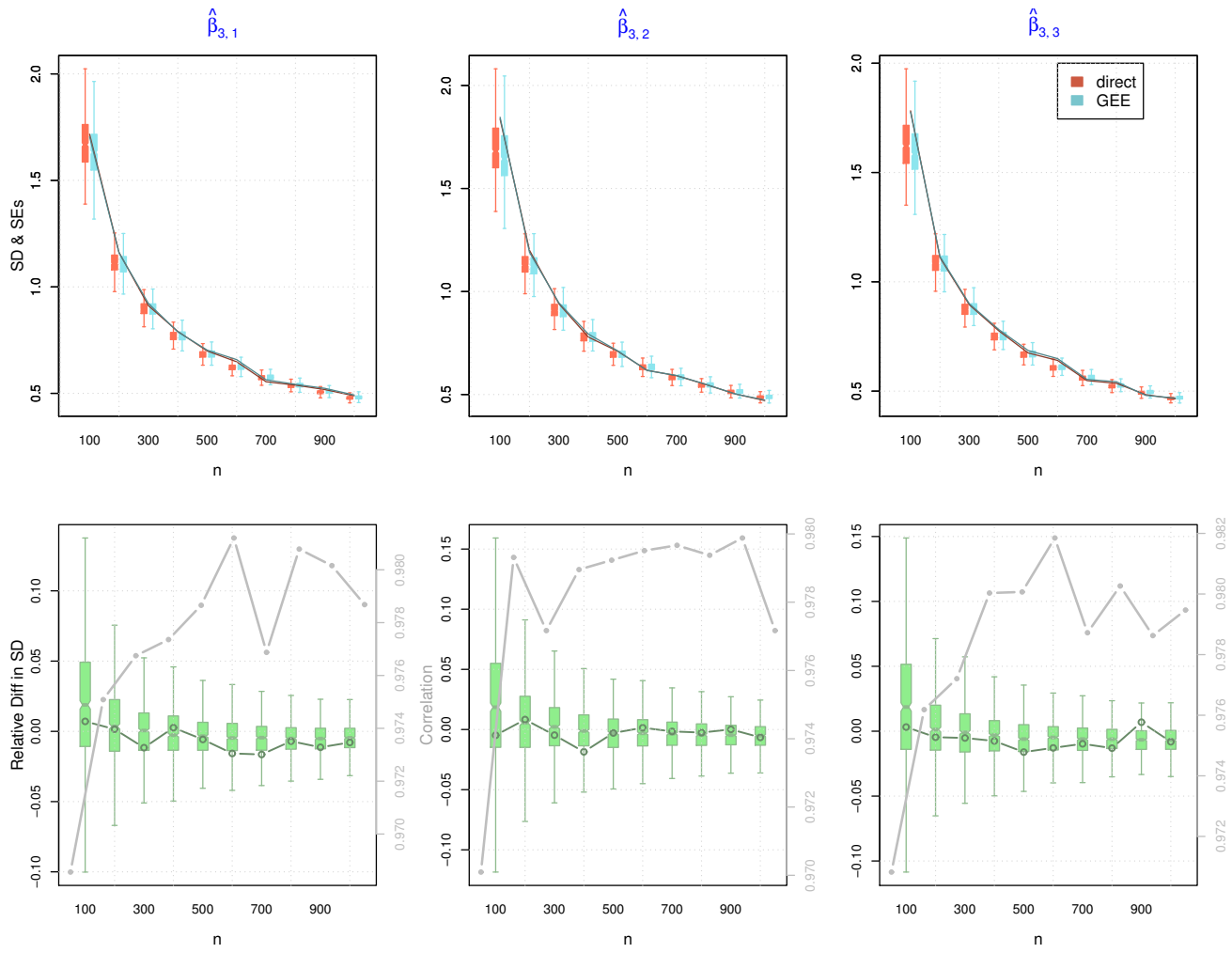

Figure 4.2: Simulation results for assessing the effect of sample size of observational data with $\rho = 0$.

The box plots in the first case show the standard error estimates associated with each of the 1000 simulation runs, with the lines showing the associated standard deviation in each scenario.

Figures (4.1) and (4.2) above represent the visual display of simulation results for assessing the effect of sample size $n = 100, 200, \ldots, 1000$ for two model configuration, (3.11) and (3.19) corresponding to the randomized experimental and observational data respectively. For each setting, 1,000 simulation runs were considered.

To this effect, the average standard errors (SEs) and standard deviations (SDs) (performance measures) of the mean estimates (interaction terms) corresponding to Model (3.11) and Model (3.19) were pooled together and compared with various sample sizes (with the boxplot in the first plot in each figure showing the standard errors obtained for each of the 1000 simulation runs and the lines corresponding to standard deviation in each scenario) . The overlaid plots above reveals clearly how the performance measures improve significantly with increased sample size with no particular respect to data source (whether randomized experimental trials or observational studies are considered). More particularly, the estimator of Model (3.19) produces estimates with relatively and consistently smaller errors (which suggest possibly more precision for the GEE estimator) with our randomized experimental trials as shown by the box plots and the curved lines in Figures (4.1).

However, the same conclusion cannot be made concerning observational data as revealed by the performance measures (SEs and SDs) that characterize the mean estimates (interaction terms), as the GEE estimator could not consistently produce smaller errors across the varying sample sizes that are considered. A confirmation of this result is indicated by the overlaid graph of the relative difference in the SDs (that is computed as the arithmetic difference between SDs of the estimates of direct and the GEE, divided by the SDs of the estimates of direct) versus the sample size shown in the above figures. In our convention, the GEE estimator of Model (3.19) is expected to produce more precise estimates compared to that of the direct estimator of Model (3.11), the relative difference in SD will be positive. The case for the randomized experimental data in Figure (4.1) validates in a way, our approach of simultaneously finding a more refined parameter estimates associated with the moderating effects on which basis of refined inference could be made.

Chapter 5

An Illustrative Example using the Wart dataset

To validate our results, we illustrate with a real-world data collected from a randomized wart treatment trial, as reported in Khozeimeh et al. (2017). In this study, $n = 180$ patients were randomized to be treated with either the cryotherapy (`cryo` = 1) or immunotherapy (`cryo` = 0) method, with 90 patients in either treatment group. These are two known and popular therapeutic approaches for the treatment of wart. Even though they lack the ability to treating all category of patients, they offer to a large extent higher prospect in dealing with wart in about 80% of people who suffer from the disease. As noted about the deficiencies about these therapies, with each lacking the ability to heal all patients, moderation analysis in the face of precision medicine is therefore crucial in designing more effective and customized treatments to help solve this problem. In our data set, the outcome variable `response` is binary with 1 indicating a positive response to the treatment and 0 otherwise. Also included are six covariates: patient gender (`sex`), age (`age`), self-reported time in months before treatment (`time`), the number of warts (`nwarts`), an indicator of whether or not patient has mixed types of warts (`type`), and surface area in mm^2 of the warts (`area`). Prior to experimenting with the wart data, a variable selection procedure via regularization is harnessed to come up with two important covariates; `age` and `type` for moderation analysis.

Table 5.1: Analysis of Wart Trial Data. Model I is obtained with the direct method; Model II from the inverse method; Model III is obtained by the GEE approach.

Model	Term	Estimate	SE	z	p-value	OR / RR
	`intercept`	2.383	0.802	2.969	0.003	10.833
	`cryo`	1.684	1.230	1.369	0.171	5.389
I	`age`	−0.035	0.021	−1.651	0.099	0.965
	`type`	0.458	0.699	0.655	0.513	1.580
	`cryo` × `age`	−0.087	0.039	−2.217	0.027	0.917
	`cryo` × `type`	−2.635	0.976	−2.701	0.007	0.072
	`intercept`	0.345	0.849	0.407	0.684	1.413
	`response`	1.594	1.038	1.535	0.125	4.922
II	`age`	−0.005	0.023	−0.223	0.823	0.995
	`type`	1.887	0.710	2.660	0.008	6.600
	`response` × `age`	−0.078	0.032	−2.415	0.016	0.925
	`response` × `type`	−3.423	0.942	−3.635	0.000	0.033
	`intercept0`	2.387	0.703	3.396	0.001	10.885
	`cryo0`	1.618	0.983	1.647	0.100	5.044
	`age0`	−0.037	0.019	−1.927	0.054	0.964
	`type0`	0.673	0.743	0.906	0.365	1.960
III	`intercept1`	−2.098	0.728	−2.881	0.004	0.123
	`cryo1`	0.043	0.056	0.781	0.435	1.044
	`age1`	0.035	0.019	1.781	0.075	1.035
	`type1`	0.996	0.438	2.276	0.023	2.708
	`cryo` × `age`	−0.082	0.031	−2.674	0.008	0.921
	`cryo` × `type`	−3.052	0.914	−3.340	0.001	0.047

Table 5.1 presents the fitting results from three models, including the parameter estimates, standard errors, the Wald z test, the resultant p-value, and the exponential of the estimates (as shown in the last column).

In Model I, a logistic regression model was fit on `response`, examining the interactions between the treatment variable `cryo` and the covariates. In Model II, the roles of `response` and `cryo` are swapped. By viewing the parameter estimates associated with the interaction terms in these two models, we note that they are within the ballpark of each other despite certain differences. It is interesting to see that the SEs for these estimates in Model II are generally lower than those in Model I. As a result, the interaction terms become more significant in Model II.

It is also worth noting that the main cardinal reason for employing the GEE approach in Model III is to find a way that seeks to combine the two Models in (3.11) and (3.12) into one, Model (3.19) (Model III) and simultaneously find more refined parameter estimates associated with the moderating effects. Owning to this goal, we skip discussion on details associated with the terms of main effects and only focus on interactions in Model III. It is however important to note that, in model III, we have two slope estimates for each of the four main effect terms (i.e., intercept, cryo, age, and type), where the first and second sets of these terms essentially corresponds with the direct and inverse models respectively. Looking at the result from a broad spectrum, it is very important to note the GEE approach in Model III generally produced estimates of the interaction terms (with relatively smaller SE), which is quite an interesting observation we sought to establish with our findings in this study. More importantly, it is revealed that both age and type (whether a patient has a mixed type of warts) have significant moderating effect on the treatment of wart. The odds of having a positive response decreases by a factor of 0.921 per every year of the patient's age if the patient do not take cryotherapy (i.e., take immunotherapy), and increase by a factor of $0.921 \times 5.044 = 4.646$ for every year they take cryotherapy. Similarly, patients with type 1 form of warts have a 46% less odds of having a positive response than patients with type 2 form of warts.

Since the treatment assignment is randomized in this study, Proposition 2 holds. By plugging `response` $= 0$ and `response` $= 1$ into Model II, we break it down into the control-only analysis and the case-only analysis respectively. However, we have skipped this related studies entirely, because the setting of Proposition 2, applying only to data from randomized experiment, has been partly investigated elsewhere; see, e.g., Dai et al. (2014), Dai et al. (2018), and Dai and LeBlanc (2019) for studies that are designed to verify the case-only analysis of which the control-only analysis, albeit new, holds naturally by symmetry.

Chapter 6

Discussion

6.1 Summary of Results

An interesting observation is made concerning moderation analysis with binary outcomes. We have shown that, the role swap between the outcome variable and the treatment variable does not alter the regression coefficients associated with the interaction terms in logistic regression models. The resultant model (3.12) is highly informative of treatment effect moderation and meaningful interpretations can be extracted depending on whether data are experimental or observational. Furthermore, it offers a new way of estimating the moderation-related parameters, relatively with more precision. Since both $\widehat{\boldsymbol{\beta}}_3$ and $\widehat{\boldsymbol{\alpha}}_3$ estimate the same parameters, we have further shown that this can easily be estimated by combining the two models and employ the generalized estimating equation (GEE) approach as an estimation technique to find a more refined parameter estimates corresponding to the interaction terms.

The trick of 'role swapping' can be useful in other scenarios where a direct study of moderation is inconvenient owing to modeling complexity, numerical difficulty, or unavailability of implementation. To elaborate, two immediate applications are outlined for future research avenues. First, consider a study that involves a categorical outcome and a binary treatment. Since the implementation of multinomial logistic regression related methods is not widely available, switching the roles allows us to conduct moderation analysis with binary logistic regression, where regularization (Tibshirani, 1996) can be used to select moderators (Lim and Hastie, 2015). Along the same lines, tree-based modeling (Su et al., 2009, 2012) can be developed for subgroup analysis as well. Secondly, the

study of gene-environment interactions (Ottman, 1996) presents a scenario with multiple treatments which are important genetic biomarkers. In common approaches on basis of hypothesis testings, the main challenge stems from multiplicity of inferences. If we swap the roles and treat multiple genetic variables as clustered binary outcomes, a generalized linear mixed model (GLMM) may facilitate an overall interaction test for environmental variables conveniently.

6.2 Future work

Future work could involve employing the role-swap technique which allows the estimation of regression coefficients associated with the interaction terms to revisit observational data (by accounting for possible confounders) with binary outcome/treatment via the GEE approach. Similar work in the same domain could be extended to both experimental and observational study with nominal outcomes or treatment as well as mediation and confounding analysis, where perhaps a national health data could be be considered.

References

Assmann, S. F., Pocock, S. J., Enos, L. E., and Kasten, L. E. (2000). Subgroup analysis and other (mis) uses of baseline data in clinical trials. *The Lancet*, 355(9209):1064–1069.

Baron, R. M. and Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173.

Bennett, P. D. and Harrell, G. D. (1975). The role of confidence in understanding and predicting buyers' attitudes and purchase intentions. *Journal of Consumer Research*, 2(2):110–117.

Clark, G. M. (2008). Prognostic factors versus predictive factors: examples from a clinical trial of erlotinib. *Molecular Oncology*, 1(4):406–412.

Clark, G. M., Zborowski, D. M., Culbertson, J. L., Whitehead, M., Savoie, M., Seymour, L., and Shepherd, F. A. (2006). Clinical utility of epidermal growth factor receptor expression for selecting patients with advanced non-small cell lung cancer for treatment with erlotinib. *Journal of Thoracic Oncology*, 1(8):837–846.

Dai, J. Y. and LeBlanc, M. (2019). Case-only trees and random forests for exploring genotype-specific treatment effects in randomized clinical trials with dichotomous end points. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(5):1371–1391.

Dai, J. Y., Li, S. S., and Gilbert, P. B. (2014). Case-only method for cause-specific hazards models with application to assessing differential vaccine efficacy by viral and host genetics. *Biostatistics*, 15(1):196–203.

Dai, J. Y., Liang, C. J., LeBlanc, M., Prentice, R. L., and Janes, H. (2018). Case-only approach to identifying markers predicting treatment effects on the relative risk scale. *Biometrics*, 74(2):753–763.

Dobson, A. J. and Barnett, A. G. (2018). *An introduction to generalized linear models*. CRC Press.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352):892–898.

Falk, M. (1999). A simple approach to the generation of uniformly distributed random variables with prescribed correlations. *Communications in Statistics-Simulation and Computation*, 28(3):785–791.

Ferreira, J. C. and Patino, C. M. (2017). Subgroup analysis and interaction tests: why they are important and how to avoid common mistakes. *Jornal Brasileiro de Pneumologia*, 43(3):162–162.

Fitzmaurice, G. M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics*, pages 309–317.

Ghiselli, E. E. (1963). Moderating effects and differential reliability and validity. *Journal of Applied Psychology*, 47(2):81.

Halekoh, U., Højsgaard, S., Yan, J., et al. (2006). The r package geepack for generalized estimating equations. *Journal of Statistical Software*, 15(2):1–11.

Hayes, A. F. and Matthes, J. (2009). Computational procedures for probing interactions in ols and logistic regression: Spss and sas implementations. *Behavior Research Methods*, 41(3):924–936.

Hobert, R. and Dunnette, M. D. (1967). Development of moderator variables to enhance the prediction of managerial effectiveness. *Journal of Applied Psychology*, 51(1):50.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Jain, K. K. (2002). Personalized medicine. *Current opinion in molecular therapeutics*, 4(6):548–558.

Jain, K. K. and Jain, K. (2009). *Textbook of personalized medicine*. Springer.

Jameson, J. L. and Longo, D. L. (2015). Precision medicine—personalized, problematic, and promising. *Obstetrical & gynecological survey*, 70(10):612–614.

Katsnelson, A. (2013). Momentum grows to make'personalized'medicine more'precise'.

Kauermann, G. and Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96(456):1387–1396.

Khozeimeh, F., Alizadehsani, R., Roshanzamir, M., Khosravi, A., Layegh, P., and Nahavandi, S. (2017). An expert system for selecting wart treatment method. *Computers in Biology and Medicine*, 81:167–175.

Lagakos, S. W. et al. (2006). The challenge of subgroup analyses-reporting without distorting. *New England Journal of Medicine*, 354(16):1667.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Lim, M. and Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654.

ÖNDER, H., OLFAZ, M., and SOYDAN, E. (2010). Comparison of working correlation matrices in generalized estimating equations for animal data. *Anadolu Tarım Bilimleri Dergisi*, 25(3):197–201.

Ottman, R. (1996). Gene–environment interaction: definitions and study design. *Preventive medicine*, 25(6):764–770.

Oxman, A. D. and Guyatt, G. H. (1992). A consumer's guide to subgroup analyses. *Annals of internal medicine*, 116(1):78–84.

Park, T. and Shin, D.-Y. (1999). On the use of working correlation matrices in the gee approach for longitudinal data. *Communications in Statistics-Simulation and Computation*, 28(4):1011–1029.

Piegorsch, W. W., Weinberg, C. R., and Taylor, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine*, 13(2):153–162.

Qu, A., Lindsay, B. G., and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika*, 87(4):823–836.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Royston, P. and Sauerbrei, W. (2008). Interactions between treatment and continuous covariates: a step toward individualizing therapy. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 26(9):1397–1399.

Sharma, S., Durand, R. M., and Gur-Arie, O. (1981). Identification and analysis of moderator variables. *Journal of Marketing Research*, 18(3):291–300.

Simms, L., Barraclough, H., and Govindan, R. (2013). Biostatistics primer: what a clinician ought to know—prognostic and predictive factors. *Journal of Thoracic Oncology*, 8(6):808–813.

Sleight, P. (2000). Debate: Subgroup analyses in clinical trials: fun to look at-but don't believe them! *Trials*, 1(1):1–3.

Su, X., Kang, J., Fan, J., Levine, R. A., and Yan, X. (2012). Facilitating score and causal inference trees for large observational studies. *Journal of Machine Learning Research*, 13:2955.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(2).

Team, R. C. et al. (2013). R: A language and environment for statistical computing.

The White House: Office of the Press Secretary (2015). Fact sheet: President obama's precision medicine initiative. `https://obamawhitehouse.archives.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative`.

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532.

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

Vittinghoff, E. and Bauer, D. (2006). Case-only analysis of treatment–covariate interactions in clinical trials. *Biometrics*, 62(3):769–776.

Wang, L., Zhou, J., and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, 68(2):353–360.

Wang, R. and Ware, J. H. (2013). Detecting moderator effects using subgroup analyses. *Prevention Science*, 14(2):111–120.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika*, 61(3):439–447.

Zedeck, S. (1971). Problems with the use of" moderator" variables. *Psychological Bulletin*, 76(4):295.

Zeger, S. (1988). The analysis of discrete longitudinal data: Commentary. *Statistics in Medicine*, 7(1-2):161–8.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, pages 121–130.

Zhang, X. (2015). Precision medicine. *Personalized Medicine, Omics and Big Data: Concepts and Relationships. J Pharmacogenomics Pharmacoproteomics*, 6(1):1000e144.

Zhang, X. D. (2011). *Optimal high-throughput screening: practical experimental design and data analysis for genome-scale RNAi research*. Cambridge University Press.

Ziegler, A., Kastner, C., and Blettner, M. (1998). The generalised estimating equations: an annotated bibliography. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 40(2):115–139.

Zorn, C. J. (2001). Generalized estimating equation models for correlated data: A review with applications. *American Journal of Political Science*, pages 470–490.

APPENDIX

Proofs

To prove Lemma 1, apply Bayes's rule to rewrite the relative risk (RR) terms as

$$\begin{aligned} RR_{1,\mathbf{x}} &= \frac{\Pr(Y=1|T=1,\mathbf{X}=\mathbf{x})}{\Pr(Y=1|T=0,\mathbf{X}=\mathbf{x})} \\ &= \frac{\Pr(Y=1,T=1,\mathbf{X}=\mathbf{x})/\Pr(T=1,\mathbf{X}=\mathbf{x})}{\Pr(Y=1,T=0,\mathbf{X}=\mathbf{x})/\Pr(T=0,\mathbf{X}=\mathbf{x})} \\ &= \frac{\Pr(T=1|Y=1,\mathbf{X}=\mathbf{x})\cdot\Pr(Y=1,\mathbf{X}=\mathbf{x})/\Pr(T=1|\mathbf{X}=\mathbf{x})}{\Pr(T=0|Y=1,\mathbf{X}=\mathbf{x})\cdot\Pr(Y=1,\mathbf{X}=\mathbf{x})/\Pr(T=0|\mathbf{X}=\mathbf{x})} \\ &= \frac{\Pr(T=1|Y=1,\mathbf{X}=\mathbf{x})}{\Pr(T=0|Y=1,\mathbf{X}=\mathbf{x})}\cdot\frac{1-\pi(\mathbf{x})}{\pi(\mathbf{x})} \end{aligned}$$

Similarly,

$$RR_{1,\mathbf{x}'} = \frac{\Pr(T=1|Y=1,\mathbf{X}=\mathbf{x}')}{\Pr(T=0|Y=1,\mathbf{X}=\mathbf{x}')}\cdot\frac{1-\pi(\mathbf{x}')}{\pi(\mathbf{x}')}.$$

Therefore,

$$RRR^{(Y=1)}(\mathbf{x}:\mathbf{x}') = \frac{RR_{1,\mathbf{x}}}{RR_{1,\mathbf{x}'}} = \frac{OR^{(T)}_{Y=1}(\mathbf{x}:\mathbf{x}')}{OR^{(T)}(\mathbf{x}:\mathbf{x}')},$$

where $OR^{(T)}(\mathbf{x}:\mathbf{x}')$ as given in (3.10) is the odds ratio on basis of the propensity scores. This establishes equation (3.7). Similar arguments can be used to establish $RRR^{(Y=1)}(\mathbf{x}:\mathbf{x}')$ in equation (3.8). Equation (3.9) follows since

$$ROR^{(Y)}(\mathbf{x}:\mathbf{x}') = \frac{RRR^{(Y=1)}(\mathbf{x}:\mathbf{x}')}{RRR^{(Y=0)}(\mathbf{x}:\mathbf{x}')} = \frac{OR^{(T)}_{Y=1}(\mathbf{x}:\mathbf{x}')}{OR^{(T)}_{Y=0}(\mathbf{x}:\mathbf{x}')} = ROR^{(T)}(\mathbf{x}:\mathbf{x}')$$

.

Denote $\boldsymbol{\beta}_3 = (\beta_{31},\ldots,\beta_{3j},\ldots,\beta_{3p})^T \in \mathbb{R}^p$ in model (3.11) and $\boldsymbol{\alpha}_3 = (\alpha_{31},\ldots,\alpha_{3j},\ldots,\alpha_{3p})^T$ in model (3.12). Let $\mathbf{x} = (x_1,\ldots,x_j,\ldots,x_p)^T$ and set $\mathbf{x}' = (x_1,\ldots,x_{j-1},x_j{+}1,x_{j+1},\ldots,x_p)^T$. Then

$$\beta_{3j} = ROR^{(Y)}(\mathbf{x}:\mathbf{x}') = ROR^{(T)}(\mathbf{x}:\mathbf{x}') = \alpha_{3j}.$$

Since this holds for every $j = 1, \ldots, p$, conclude that $\boldsymbol{\beta}_3 = \boldsymbol{\alpha}_3$, which completes the proof of Proposition 1.

Proposition 2 follows equations (3.7) and (3.8) in Lemma 1. With data collected from a randomized experiment, the term $OR^{(T)}(\mathbf{x} : \mathbf{x}')$ in (3.10) equals 1 and can be removed. It follows

$$RRR^{(Y=0)}(\mathbf{x} : \mathbf{x}') = OR^{(T)}_{Y=0}(\mathbf{x} : \mathbf{x}') \quad \text{and} \quad RRR^{(Y=1)}(\mathbf{x} : \mathbf{x}') = OR^{(T)}_{Y=1}(\mathbf{x} : \mathbf{x}').$$

Setting $\mathbf{x}$ and $\mathbf{x}'$ equal except that the j-th component of $\mathbf{x}'$ is one unit larger than that of $\mathbf{x}$ and comparing the loglinear model (3.13) with the logistic model (3.12) yield

$$\gamma^{(0)}_{3j} = RRR^{(Y=0)}(\mathbf{x} : \mathbf{x}') = OR^{(T)}_{Y=0}(\mathbf{x} : \mathbf{x}') = \alpha_{2j}$$

and

$$\gamma^{(1)}_{3j} = RRR^{(Y=1)}(\mathbf{x} : \mathbf{x}') = OR^{(T)}_{Y=1}(\mathbf{x} : \mathbf{x}') = \alpha_{2j} + \alpha_{3j},$$

for every $j = 1, \ldots, p$, where $\boldsymbol{\gamma}^{(0)}_3 = \left(\gamma^{(0)}_{3j}\right)$, $\boldsymbol{\gamma}^{(1)}_3 = \left(\gamma^{(1)}_{3j}\right)$, and $\boldsymbol{\alpha}_2 = (\alpha_{2j})$. This completes the proof of Proposition 2. □

R Codes

```
# install.packages("MultiRNG")
# require(MASS) # TO USE FUNCTION mvrnorm()
library(MultiRNG) # MULTIVARIATE UNIFORM DISTRIBUTION WITH GIVEN
 CORRELATION MATRIX

# ====================================
# FUNCTION rdat
# ====================================

is.even <- function(x) x %% 2 == 0

expit <- function(x) (tanh(x/2)+1)/2    # CORRECT & VERIFIED
# expit0 <- function(x) exp(x)/(1+exp(x))  # LOGISTIC FUNCTION, INVERSE
  OF LOGIT FUNCTION

rdat <- function(n=100,
b0=c(-1, 0.5, 0.5, -1, 1,  -0.5, 1, 0),
link.function="logistic",
rho=0.2, # UNIFORM COVARIATES
observational=FALSE, trt.p=0.5, a0=c(-2, 2, 2, 0), details=FALSE)   #
 TRT ASSIGNMENT
{
  # GENERATE X
  if (!is.even(length(b0)) || length(b0)<2) stop("Length of b0 should
 be an even number!")
  p <- (length(b0)-2)/2  # NUMBER OF COVARIATES
  S <- matrix(1, p, p)
  for (i in 1:p){
    for (j in 1:p){
      S[i, j] <- rho^(abs(i-j))
    }
  }
```

```
  X <- draw.d.variate.uniform(no.row=n,d=p,cov.mat=S)

  # TREATMENT (RANDOM)
  if (observational) {
    if (length(a0)!=(p+1)) stop("Lenght of a0 must be consistent with
 b0.")
    eta.trt <- as.vector(cbind(1, X)%*%a0);
    pi.trt <- expit(eta.trt)
    trt <- rbinom(n=n, size=1, prob=pi.trt)
    if (details) print(cbind(p.treatment=mean(pi.trt), p.trt=mean(trt))
 )
  } else {trt <- rbinom(n=n, size=1, prob=trt.p)}

  # MODEL
  X0 <- cbind(1, trt, X, trt*X)
  eta <- as.vector(X0%*%b0);
  if (details) print(cbind(eta.mean=mean(eta), eta.min=max(eta), prop.
 positive=sum(eta>0)/length(eta)))
  if (link.function=="exp") pi0 <- exp(eta*(eta<=0))  # LOGLINEAR
  else pi0 <- exp(eta)/(1+exp(eta))  # LOGISTIC
  y <- rbinom(n, size=1, prob=as.vector(pi0));
  if (details) print(cbind(p.respone=mean(pi0), p.y=mean(y)))

  dat <- data.frame(cbind(y, trt, X))
  colnames(dat) <- c("y", "trt", paste("x", 1:NCOL(X), sep=""))
  dat
}




# ------------------------------------
# FUNCTION swap() SWAPS TWO COLUMNS
# ------------------------------------
```

```
  swap <- function(DF, n, m)
  {
    n <- if (class(n)=="character" & is.na(suppressWarnings(as.integer(n)
   ))) which(colnames(DF)==n) else as.integer(n)
    m <- if (class(m)=="character" & is.na(suppressWarnings(as.integer(m)
   ))) which(colnames(DF)==m) else as.integer(m)

    if (!(1<=n & n<=length(DF))) stop( "`n` represents invalid index!" )
    if (!(1<=m & m<=length(DF))) stop( "`m` represents invalid index!" )

    return (DF[ if (n==m) 1:length(DF) else c( (if (min(n,m)==1) c() else
    1:(min(n,m)-1) ), (if (min(n,m)+1 == max(n,m)) (min(n,m)+1):(max(n,m)
   -1) else c( max(n,m), (min(n,m)+1):(max(n,m)-1), min(n,m))), (if (max(
   n,m)==length(DF)) c() else (max(n,m)+1):length(DF) ) ) ])
  }


  arrange.data <- function(dat, y="y", trt="trt") {
    n <- NROW(dat)
    dat1 <- swap(dat, y, trt)
    names(dat1) <- names(dat)
    dat <- data.frame(rbind(dat, dat1))
    dat$group <- rep(c(0,1), c(n, n))
    dat$ID <- rep(1:n, 2)
    return(dat)
  }

# ---------------------
# GENERATE SOME DATA
# ---------------------
set.seed(777)
b0 <- c(-1, 0.5, 0.5, -1, 1, -0.5, 1, 0)
a0 <- c(-2, 2, 2, 0)
n <- 500
```

```
dat <- rdat(n=n, b0=b0, link.function="logistic",
rho=0.2, # UNIFORM COVARIATES
observational=FALSE, trt.p=0.5, a0=a0, details=TRUE)
head(dat); dim(dat)


# -------------------------
# DIRECT/INVERSE REGRESS
# -------------------------

fit.direct <- glm(y~trt + x1 + x2 + x3 + trt:x1 + trt:x2 + trt:x3, data=
    dat,
family=binomial(link = "logit"))
summary(fit.direct)

fit.inverse <- glm(trt~y + x1 + x2 + x3 + y:x1 + y:x2 + y:x3, data=dat,
family=binomial(link = "logit"))
summary(fit.inverse)



# ----------------------------------
# COMMON ESTIMATORS FOR MODERATION
# ----------------------------------

library(geepack)
dat0 <- arrange.data(dat)
dim(dat0); dim(dat)
head(dat0)

fit <- glm(y~ (trt + x1 + x2 + x3)*group + trt:x1 + trt:x2 + trt:x3, data
    =dat0,
family=binomial(link = "logit"))
summary(fit)
```

```
fit.ind <- geeglm(y~ (trt + x1 + x2 + x3)*group + trt:x1 + trt:x2 + trt:
   x3,
id=ID, data=dat0,  corstr="independence",
family=binomial(link = "logit"))
summary(fit.ind)

# USER-DEFINED WORKING CORRELATION
rho <- rep(0.65, n)
fit.gee <- geeglm(y~ (trt + x1 + x2 + x3)*group + trt:x1 + trt:x2 + trt:
   x3,
id=ID, data=dat0,
corstr="fixed", zcor=rho,
family=binomial(link = "logit"))
summary(fit.gee)


betas <- cbind(coef(fit.direct)[6:8], coef(fit.inverse)[6:8],
coef(fit)[11:13],  coef(fit.ind)[11:13], coef(fit.gee)[11:13])

SE <- cbind(summary(fit.direct)$coefficients[6:8, 2],
summary(fit.inverse)$coefficients[6:8, 2],
summary(fit)$coefficients[11:13, 2],
summary(fit.ind)$coefficients[11:13, 2],
summary(fit.gee)$coefficients[11:13, 2])
out <- data.frame(betas, SE)
colnames(out) <- c(paste("beta", c("direct", "inverse", "common", "ind",
   "gee"), sep="."),
paste("se", c("direct", "inverse", "common", "ind", "gee"), sep="."))
out


# ------------------------------
# EXTRACT VCOV FROM gee FITTING
```

```r
# ------------------------------

summary(fit.gee)
V <- summary(fit.gee)$cov.scaled
sqrt(diag(V))
V0 <- V[11:13, 11:13]
sqrt(diag(V0))




# ============================
# A SIMULATION STUDY
# ============================


source("Functions-Moderation.R")

set.seed(777)
b0 <- c(-1, 0.5, 0.5, -1, 1, -0.5, 1, 0)
a0 <- c(-2, 2, 2, 0)
RHO <- c(0, 0.2, 0.5, 0.8)
nrun <- 1000
STUDY.obs <- c(FALSE, TRUE)
N <- (1:10)*100
BETA <- SE <- array(0, dim=c(nrun, 12, length(N), length(RHO), length(
    STUDY.obs)))
VCOV <- array(0, dim=c(3, 3, nrun, length(N), length(RHO), length(STUDY.
    obs)))
OUT <- NULL
for (i in 1:length(STUDY.obs)) {
  study <- STUDY.obs[i]
  for (m in 1:length(RHO)){
    rho <- RHO[m]
```

```
 for (j in 1:length(N)){
   n <- N[j]
   Beta <- Se <- matrix(0, nrun, 9)
   for (k in 1:nrun) {
     print(cbind(study=ifelse(study, "Observational", "Experimental"),
 rho=rho, n=n, run=k))
     dat <- rdat(n=n, b0=b0, link.function="logistic", rho=rho,
observational=study,
     trt.p=0.5, a0=a0, details=TRUE)
     # DIRECT
     fit.direct <- glm(y~trt + x1 + x2 + x3 + trt:x1 + trt:x2 + trt:x3
, data=dat,
     family=binomial(link = "logit"))
     beta.direct <- coef(fit.direct)[6:8]
     se.direct <- summary(fit.direct)$coefficients[6:8, 2]
     # INVERSE
     fit.inverse <- glm(trt~y + x1 + x2 + x3 + y:x1 + y:x2 + y:x3,
data=dat,
     family=binomial(link = "logit"))
     beta.inverse <- coef(fit.inverse)[6:8]
     se.inverse <- summary(fit.inverse)$coefficients[6:8, 2]
     # GEE - IND
     dat0 <- arrange.data(dat)
     fit.gee <- geeglm(y~ (trt + x1 + x2 + x3)*group + trt:x1 + trt:x2
 + trt:x3,
     id=ID, data=dat0,  corstr="independence",
     family=binomial(link = "logit"))
     beta.gee <- coef(fit.gee)[11:13]
     se.gee <- summary(fit.gee)$coefficients[11:13, 2]
     vcov <- (summary(fit.gee)$cov.scaled)[11:13, 11:13]
     VCOV[, ,k,j,m,i] <- vcov

     # UPDATE RESULT
     Beta[k,] <- c(beta.direct, beta.inverse, beta.gee)
```

```
        Se[k,] <- c(se.direct, se.inverse, se.gee)
      }
      BETA[, ,j,m,i] <- cbind(study=ifelse(study, "Observational", "
   Experimental"), rho=rho, n=n, Beta)
      SE[, ,j,m,i] <- cbind(study=ifelse(study, "Observational", "
   Experimental"), rho=rho, n=n, Se);
      out <- c(study=ifelse(study, "Observational", "Experimental"), rho=
   rho, n=n, apply(Beta, 2, mean),
      apply(Beta, 2, sd),  apply(Se, 2, mean), apply(Se, 2, median))
      OUT <- rbind(OUT, out)
    }
  }
}
OUT <- as.data.frame(OUT)
colnames(OUT) <- c("study", "rho", "n",
paste("b", 1:3, "-direct", sep=""), paste("b", 1:3, "-inverse", sep=""),
   paste("b", 1:3, "-gee", sep=""),
paste("sd", 1:3, "-direct", sep=""), paste("sd", 1:3, "-inverse", sep="")
   , paste("sd", 1:3, "-gee", sep=""),
paste("ase", 1:3, "-direct", sep=""), paste("ase", 1:3, "-inverse", sep="
   "), paste("ase", 1:3, "-gee", sep=""),
paste("median-se", 1:3, "-direct", sep=""), paste("median-se", 1:3, "-
   inverse", sep=""), paste("median-se", 1:3, "-gee", sep=""))
OUT

save(BETA, SE, VCOV, OUT, file="result.Rdat")

#####################################################
# SUMMARY OF THE SIMULATION RESULTS
#####################################################


# rm(list=ls(all=TRUE))
```

```
load("result.Rdat")
ls()

dim(BETA)

# ==============
# NUMERICAL
# ==============

dim(OUT)
write.csv(OUT, file="resul1.csv", row.names=FALSE)

n.study <- dim(BETA)[[5]]
n.rho <- dim(BETA)[[4]]
ns <- dim(BETA)[[3]]
RHO <- c(0, 0.2, 0.5, 0.8)
N <- 1:10*100
alpha <- 0.05
OUT <- NULL
for (i in 1:n.study) {
  study <- ifelse(i==1, "Experimental", "Observational")
  for (j in 1:n.rho) {
    rho <- RHO[j]
    for (k in 1:ns){
      n <- N[k]
      print(cbind(i, j, k, study=study, rho=rho, n=n))
      Beta <- BETA[, 4:12, k, j, i]
      Se <- SE[, 4:12, k, j, i]
      Vcov <- VCOV[, , , k, j, i]
      storage.mode(Beta) <- storage.mode(Se) <- "double"

      beta <- matrix(apply(Beta, 2, mean), nrow=3, byrow=FALSE)
      sd <- matrix(apply(Beta, 2, sd), nrow=3, byrow=FALSE)
      se.mean <- matrix(apply(Se, 2, mean), nrow=3, byrow=FALSE)
```

```
      se.median <- matrix(apply(Se, 2, median), nrow=3, byrow=FALSE)

      storage.mode(Vcov) <- "double"
      vcov.avg <- apply(Vcov, c(1, 2), mean)
      vcov.median <- apply(Vcov, c(1, 2), median)
      vcov <- cov(Beta[, 7:9])

      # WALD Z TEST
      Z <- (Beta/Se)^2
      pvalue <- pchisq(Z, df=1, lower.tail =FALSE)
      power.size <- matrix(apply(pvalue < alpha, 2, mean), nrow=3, byrow=
   FALSE)

      # COLLECT RESULTS
      out <- cbind(study, rho, n, beta, sd, se.mean, se.median, power.
   size,
      vcov, vcov.avg, vcov.median)
      OUT <- rbind(OUT, out)
    }
  }
}
OUT <- as.data.frame(OUT)
methods <- c("direct", "inverse", "gee")
colnames(OUT) <- c("study", "rho", "n", paste("beta.", methods, sep=""),
paste("sd.", methods, sep=""), paste("se.mean.", methods, sep=""),
paste("se.median.", methods, sep=""), paste("power.size.", methods, sep="
   "),
paste("vcov.", 1:3, sep=""), paste("vcov.avg.", 1:3, sep=""),
paste("vcov.median.", 1:3, sep=""))
head(OUT)

write.csv(OUT, file="results.csv", row.names=FALSE)
```

```
# ==============
# GRAPHICAL
# ==============

dim(OUT)
head(OUT)
# ============================
# PLOTTING THE RAW RESULTS
# ============================
load("result.Rdat")
ls()

n.study <- dim(BETA)[[5]]
n.rho <-  dim(BETA)[[4]]
RHO <- c(0, 0.2, 0.5, 0.8)
for (k in 1:n.study) {
  study <- ifelse(k==1, "exp", "obs")
  for (m in 1:n.rho){
    rho <- RHO[m]
    print(cbind(k=k, m=m, study=study, rho=rho))
    filename <- paste("fig-", study, "-", rho, ".eps", sep="")
    BETA0 <- BETA[,, , m, k]; SE0 <- SE[,, , m, k]

    # SD & CORR
    ns <- dim(BETA0)[[3]]
    N <- (1:ns)*100
    SD <- matrix(0, ns, 9)
    CORR <- matrix(0, ns, 3)
    for (i in 1:ns){
      Beta <- as.matrix(BETA0[, 4:12,i])
      storage.mode(Beta) <- "numeric"
      SD[i, ] <- apply(Beta, 2, sd)
```

```
   CORR[i, 1] <- cor(Beta[,1], Beta[,7], method="pearson")
   CORR[i, 2] <- cor(Beta[,2], Beta[,8], method="pearson")
   CORR[i, 3] <- cor(Beta[,3], Beta[,9], method="pearson")
 }
 # AVERAGED SE
 AvgSe <- matrix(0, ns, 9)
 for (i in 1:ns){
   Se <- as.data.frame(SE0[, 4:12,i])
   Se <- as.data.frame(sapply(Se, function(x) as.numeric(as.character(
x))))
   AvgSe[i, ] <- apply(Se, 2, mean)
 }
 # PREPARE BOXPLOT DATA WITH SE
 dat0 <- NULL
 for (i in 1:ns) dat0 <- rbind(dat0, as.matrix(SE0[, ,i]))
 dat0 <- as.data.frame(dat0)
 dat0[, 4:12] <- as.data.frame(sapply(dat0[, 4:12], function(x) as.
numeric(as.character(x))))
 names(dat0) <- c("study", "rho", "n", paste("x", 1:9, sep=""))
 # dat0$n <- ordered(dat0$n, levels =as.character((1:10)*100))
 dat0$n <- as.numeric(as.character(dat0$n))
 dat0$rd1 <- (dat0$x1 - dat0$x7)/dat0$x1
 dat0$rd2 <- (dat0$x2 - dat0$x8)/dat0$x2
 dat0$rd3 <- (dat0$x3 - dat0$x9)/dat0$x3

 # -------------
 # 2 x 3 PLOT
 # -------------
 postscript(file=filename, horizontal=TRUE)
 par(mfrow=c(2, 3), mar=c(4, 4, 3, 2))
 # SD
 for (j in 1:3){
   # BOXPLOTS OF SE
   form <- as.formula(paste(paste("x", j, sep=""), " ~ ", "n"))
```

```
   boxplot(form, data=dat0, boxwex=0.2, border="coral1", col="coral1",
 xlab="n", ylab="", notch=TRUE,
   xaxt="n", outline=FALSE, at=(1:ns)-0.15, add=FALSE, cex.lab=1.2,
lty=1)
   if(j==1) mtext(text="SD & SEs", side=2, line=2.1, col="black", cex
=0.8)
   if (j==3) legend(6, 2.0, fill=c("coral3", "cadetblue3"), border=c("
coral3", "cadetblue3"),
   legend=c("direct", "GEE"), cex=1.2)
   beta.j <- substitute(list(hat(beta))[list(3,j0)], list(j0=j))
   text(5, par("usr")[4] + 0.15, srt=0, adj = 0, labels=beta.j, xpd =
TRUE, col="blue", cex=1.5)
   # mtext(text=beta.j, side=4, line=1, col="blue", cex=2)
   axis(1, at=1:10, labels=N, tick=FALSE , cex=0.3)
   form0 <- as.formula(paste(paste("x", j+6, sep=""), " ~ ", "n"))
   boxplot(form0, data=dat0, boxwex=0.2, border="cadetblue2", col="
cadetblue2", xaxt="n", notch=TRUE,
   add=TRUE, at=(1:ns)+0.15, outline=FALSE, lty=1)
   # Averaged SE
   # lines(1:ns, AvgSe[1:ns,j], col="tomato", lty=1, lwd=0.5)
   # lines(N, AvgSe[1:ns, j+3], col="green4", lwd=0.5, lty=2)
   # lines(1:ns, AvgSe[1:ns, j+6], col="skyblue2", lwd=0.5, lty=1)
   # SD
   lines(1:ns, SD[1:ns, j], col="coral4", type="l", lwd=1, ylab="SD",
xlab="n", cex=0.5)
   grid()
   # lines(N, SD[1:ns, j+3], col="green4", lwd=0.5)
   lines(1:ns, SD[1:ns, j+6], col="cadetblue4", lwd=1, type="l", pch
=20, cex=0.5)
 }

 for (j in 1:3){
   # BOXPLOT OF RELATIVE DIFFERENCE IN SE
   form <- as.formula(paste(paste("rd", j, sep=""), " ~ ", "n"))
```

```
      boxplot(form, data=dat0, boxwex=0.35, border="darkseagreen", col="
    palegreen2",
      xlab="n", ylab="", notch=TRUE, cex.lab=1.2, lty=1,
      xaxt="n", outline=FALSE, at=(1:ns), add=FALSE)
      axis(1, at=1:10, labels=N, tick=FALSE , cex=0.3)
      if(j==1) mtext(text="Relative Diff in SD", side=2, line=2.1, col="
    black", cex=0.8)
      if(j==2) mtext(text="Correlation", side=2, line=2.2, col="gray65",
    cex=0.8)

      # RELATIVE DIFFERENCE
      rd.sd <- (SD[1:ns, j]-SD[1:ns, j+6])/SD[1:ns, j]
      lines(1:ns, rd.sd, col="darkseagreen4", type="b", lwd=1.5);
      grid()
      # rd.ase <- (AvgSe[1:ns, j]-AvgSe[1:ns, j+6])/AvgSe[1:ns, j]
      # lines(1:ns, rd.ase, col="blue", type="b", lwd=1.5)
      par(new=T)
      plot(1:ns, CORR[ , j], lwd=2, col="gray75", xlab="", ylab="", type=
    "b",  # ylim=c(0.90, 1.00),
      pch=20, cex=0.8, cex.lab=1.2, lty=1, axes=F)
      axis(4,col="gray75", col.ticks="gray75", col.axis="gray75")

    }
    dev.off()
  }
}


#####################################################################
# REAL DATA EXPLORATION - MODERATION ANALYSIS OF WART DATA
#####################################################################
library(MASS)
library(geepack)
# rm(list=ls(all=TRUE))
```

```
getwd()
dat <- read.csv(file="wart.csv", header=TRUE)

dim(dat); head(dat)
table(dat$response)/NROW(dat)
table(dat$cryo)/NROW(dat)

table(dat$response, dat$cryo, dat$type)
table(dat$response, dat$type)

# MERGE TYPE 1&2 TOGETHER; type=type3
dat$type <- ifelse(dat$type==3, 1, 0)

dat$agegrp <- cut(dat$age, breaks=quantile(dat$age, probs = seq(0, 1,
   0.25)),
labels = 1:4)

table(dat$response, dat$cryo, dat$agegrp)


# ======================================================================
# GENERAL DIRECT LOGISTIC REGRESSION
# ======================================================================

fit0 <- glm(response ~ cryo + sex + age + time + nwarts + type + area,
   data=dat,  family=binomial(link = "logit"))
summary(fit0)

======================================================================
#VARIABLE SELECTION METHODS
======================================================================
# BEST SUBSET SELECTION
install.packages("glmulti")
library(glmulti)
```

```
out.BSS <- glmulti(response ~ cryo + sex + age + time + nwarts + type +
    area, data=dat,
fitfunc = glm, family=binomial, intercept = TRUE,
crit = bic, level = 1, method="g", plotty=FALSE,
confsetsize=1)  # SELECT ONLY ONE BEST MODEL
fit.BSS <- attributes(out.BSS)$objects[[1]]
fit.BSS$coef
summary(fit.BSS)



fit.direct <- glm(response ~ cryo + sex + age + time + type + cryo:age +
    cryo:time + cryo:type, data=dat,
family=binomial(link = "logit"))
summary(fit.direct)


fit.direct <- glm(response ~ cryo + age + time + factor(type) + cryo:time
     + cryo:age, data=dat,
family=binomial(link = "logit"))
summary(fit.direct)

=============================================================================
#STEPWISE SELECTION
=============================================================================
fit.stepwise <- stepAIC(fit.direct, direction = "both",  k=log(nrow(dat))
    )
# names(fit.stepwise)
fit.stepwise$anova
summary(fit.stepwise)


fit.direct <- glm(response ~ cryo + age + time + type + cryo:age + cryo:
    time + cryo:type, data=dat,
```

```
family=binomial(link = "logit"))
summary(fit.direct)


control0.glm <- glm.control(epsilon = 1e-8, maxit=100, trace = FALSE)
fit.direct <- glm(response ~ cryo  + age + type + cryo:age + cryo:type,
   data=dat,
family=binomial(link = "logit"), control =control0.glm)
summary(fit.direct)$coefficients


## LASSO

library(glmnet)
formula0 <- response~.
X <- model.matrix (as.formula(formula0), data = dat)
y <- dat$response
#response ~ cryo + sex + age + time + nwarts + factor(type) + area, data=
   dat,
fit.lasso <- glmnet(x=X, y=y, family= binomial(link = "logit"), alpha=1,
lambda.min = 1e-6, nlambda = 100, standardize=T, thresh =
1e-07, maxit=1000)
plot(fit.lasso)


#We then determine the optimal tuning parameter.
CV <- cv.glmnet(x=X, y=y, family=binomial(link = "logit"), alpha = 1,
lambda.min = 1e-4, nlambda = 200, standardize = T, thresh = 1e-07,
maxit=100)
plot(CV)

#I select the best tuning parameter ($\lambda$) and use it in the final
   model.
```

```
b.lambda <- CV$lambda.1se; b.lambda
fit.best <- glmnet(x=X, y=train$outcome, family="binomial", alpha = 1,
lambda=b.lambda, standardize = T, thresh = 1e-07,
maxit=1000)
fit.best$beta
================================================================================



# FITTING DIRECT AND INVERSE MODEL WITH
control0.glm <- glm.control(epsilon = 1e-8, maxit=100, trace = FALSE)
fit.direct <- glm(response ~ cryo  + age + type + cryo:age + cryo:type,
   data=dat,
family=binomial(link = "logit"), control =control0.glm)
summary(fit.direct)$coefficients

form.inverse <- cryo ~ response + age  + type + response:age + response:
   type
fit.inverse <- glm(form.inverse, data=dat,
family=binomial(link = "logit"), control =control0.glm)
summary(fit.inverse)$coefficients

fit0.inverse <- glm(cryo ~ age + type, data=dat, subset=(response==0),
family=binomial(link = "logit"), control =control0.glm)
summary(fit0.inverse)$coefficients

fit1.inverse <- glm(cryo ~ age + type, data=dat, subset=(response==1),
family=binomial(link = "logit"), control =control0.glm)
summary(fit1.inverse)$coefficients


#================================GEE Model
   ================================
```

```
arrange.data <- function(dat, y="response", trt="cryo") {
  n <- NROW(dat)
  dat1 <- swap(dat, y, trt)
  names(dat1) <- names(dat)
  dat <- data.frame(rbind(dat, dat1))
  dat$group <- rep(c(0,1), c(n, n))      #Rearranging data
  dat$ID <- rep(1:n, 2)
  dat <- dat[order(dat$ID), ]
  return(dat)
}

dat0 <- arrange.data(dat)
dim(dat0); dim(dat)
head(dat0)

# Fitting the GEE model
fit.gee <- geeglm(response ~ (cryo  + age + type)*group + cryo:age + cryo
   :type,
id=ID, data=dat0,  corstr="independence",
family=binomial(link = "logit"))
summary(fit.gee)
result.gee<-(summary(fit.gee)$coefficients)
names(result.gee)
result.gee$Wald<-(result.gee$Estimate)/(result.gee$Std.err)
result.gee$`Pr(>|W|)`<-2*pnorm(-abs(result.gee$Wald))

#result.gee$OR <- exp(as.numeric(result.gee[, 1]))
#result.gee$"z value" <- result.gee[, 3]
names(result.gee)[names(result.gee) == "Wald"] <- "z value"
names(result.gee)[names(result.gee) == "Pr(>|W|)"] <- "Pr(>|z|)"
names(result.gee)[names(result.gee) == "Std.err"] <- "Std. Error"
#result.gee
```

```
# RESULTS
OUT <- rbind(summary(fit.direct)$coefficients, NA,
summary(fit.inverse)$coefficients, NA,
summary(fit0.inverse)$coefficients, NA,
summary(fit1.inverse)$coefficients, NA,
result.gee)
OUT <- as.data.frame(OUT)

OUT$OR <- exp(as.numeric(OUT[, 1]))

Wart_data_results<-write.csv(OUT, file="result2-wart.csv", row.names=TRUE
    )
```

Curriculum Vitae

Eric Anto was born on June 23, 1990, the only child of Mary Ntsiful to have made it beyond high school. He had all his elementary through college education in Ghana. He graduated from Agona Swedru Senior High, in the central part of Ghana in 2011, where he won a Ghana COCOBOD scholarship Trust Award during his four year term of studies as well as other best student academic wards. He had his Bachelor of Arts (BA) degree in Mathematics and Statistics from the University of Ghana, Legon. While pursuing his bachelors degree, Eric received the Tertiary Education Scholarship Trust award from 2012 to 2016 and he also won the Yi-Boa Scholarship Scheme award in 2014 to 2016 and a subsequent award from the Sadhu T.L. Vaswani/Indian Association of Ghana Endowment Fund Awards for Mathematics Students in 2015 at the University of Ghana awards ceremony. Owning to his good academic performance, he was appointed a Teaching Assistant at the Department of Statistics & Actuarial Science of the University of Ghana for his national service from 2016 to 2017. He was actively engaged in activities like data collection, data entry, data analysis, report writing and teaching which helped broadened his knowledge in statistics as well as improved his interpersonal skills which also challenged his critical thinking and analytical skills. His love for teaching as well as his good track record earned him an employment opportunity from the Ghana Education Service where he was posted to Odorgonno Senior School in Accra to teach high school mathematics from 2018 to 2019. In Fall 2019, he enrolled at the Graduate School of The University of Texas at El Paso, where he was awarded a Teaching Assistantship position while pursuing his master's degree in Statistics at the Mathematical Sciences Department. Eric started his thesis work titled, "Refined Moderation Analysis with Binary Outcomes" which was supervised by his mentor, Professor Xiaogang Su. Eric will pursue his doctoral degree in the Population Health Sciences Program (Biostatistics) beginning Fall 2021 at the University of Utah.

Email: `eanto@miners.utep.edu`