Modifications of Q-learning to Optimize Dynamic Treatment Regimes

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Yuan Zhang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advised by Thomas A. Murray, Ph.D.
Co-advised by David M. Vock, Ph.D.

August, 2021

Acknowledgements

First and foremost it is a genuine pleasure to express my sincere gratitude to my advisers Dr. Tom Murray and Dr. David Vock for their invaluable advice, support, and encouragement during my PhD study. I am extremely fortunate to work with Tom who always has great ideas and provides continuous guidance and patience throughout my research. I have also received ingenious suggestions from David not only for the completion of this dissertation, but also on the insights into academia, methodological research and interdisciplinary collaboration as a whole. Many thanks to Dr. Megan Patrick for providing the M-bridge data to illustrate our proposed methods, and Grace Lyden for her generous assistance in processing the raw data. I would also like to acknowledge Dr. Lizbeth Finestack for serving on my committee and initiating the first project of the thesis.

I am deeply grateful to Dr. Lynn Eberly for her mentorship on my work as a research assistant in the diabetes projects. Lynn, together with the principal investigators Dr. Elizabeth Seaquist and Dr. Gülin Öz, serving as excellent role models for female researchers, has inspired and motivated me in the pursuit of an academic career. It is indeed my privilege to spend more than three years in the Seaquist Lab and collaborate with the wonderful team. I have received tremendous help from Evan Olawsky, Anjali Kumar, Michelle Snyder, Dr. Amir Moheet and Dr. Lisa Chow. I would also like to extend my appreciation to Dr. Silvia Mangia and Dr. Antonietta Canna for their trust and compliment on my independent analyses of neuroimaging data.

Furthermore, I wish to offer my special thanks to my internship mentor at the Merck Research Laboratories, Dr. John Kang, who has expanded my horizons by exemplifying biometrics research in drug development. I vastly enjoyed the respectful and innovative working environment and the interesting project of adjusting for treatment switching in oncology studies. It was really a great step forward and contributed to my confidence in developing statistical methods within a time constraint.

I am indebted to the Division of Biostatistics, the University of Minnesota. I very much appreciate the extensive help from Sally Olander since the very first day I arrived at the division. Thanks should also go to Susan Wei for giving the most intellectually engaging lectures and providing me with useful tips to survive through the winter in Minnesota. My journey as a PhD student would not be complete without friends who never wavered in their support. I must thank Chuyu Deng and Shannon McKearnan for their kindness since I joined the cohort, Mengli Xiao and Tianzhong Yang for their company during the pandemic, and Jin Jin for talking me through everything.

Finally, I gratefully thank my dearest parents for always believing in me and respecting every decision I have ever made, and my fiancé for always taking great care of me. Without their trust and support I could not have made it this far. I would also like to thank my lifelong best friends Yilin Qiao, Shengyun Wang and Wenxin Tian who live in different continents but always understand and support me one way or another.

This dissertation is dedicated to my fiancé Tingyang Zhou

and my parents Wei Su and Jianfeng Zhang

for their endless love, trust and support.

Abstract

With an emerging interest in personalized medicine and quality healthcare, the design of clinical trials that incorporates multiple stages of randomization and intervention, for example, a sequential multiple assignment randomized trial (SMART), has become a popular choice for investigators as it facilitates the construction and analysis of dynamic treatment regimes (DTRs). There exists a comprehensive body of literature on various statistical methods to analyze data collected from such trials and estimate the optimal DTR for an individual subject, among which Q-learning with linear regression is widely used due to its simplicity and ease of interpretation. This thesis discusses three important challenges that cause problems in the implementation of Q-learning and proposes multiple modifications of Q-learning to address them.

The first challenge arises from the repeatedly monitored outcome of interest at intermediate stages of randomization and at longer follow-up intervals after the final stage of randomization. Clinical investigators are usually interested in identifying the optimal DTR and estimating the outcome trajectory under the optimal DTR. However, in the presence of stagewise repeated-measures outcomes, standard Q-learning fails to provide point estimates of the optimal trajectory with time-specific heterogeneous causal effects. To address this problem, we propose a modified algorithm of Q-learning with a generalized estimating equation to estimate each Q-function. The second challenge is model misspecification. Model misspecification is a common problem in Q-learning, but little attention has been given to its impact when treatment effects are heterogeneous across subjects. We describe the integrative impact of two possible types of model misspecification related to treatment effect heterogeneity: unexplained early-stage treatment effects in late-stage main effect model, and misspecified linearity between pseudo-outcomes and predictors as a result of the optimization operation. The proposed method, aiming to deal with both types of misspecification concomitantly, builds interactive models into

residual-modified parametric Q-learning. The third challenge is generalizing modified Q-learning to dichotomous outcomes. It is difficult to include informative residuals from estimation of late-stage models into early-stage pseudo-outcomes due to the non-identity link function. We propose a modification based on monotonicity of preferences to address model misspecification in Q-learning with probit regression. The improvement in robustness of the proposed modification is subject to the extent of model misspecification and can be limited. Thus, we take a latent variable approach and propose a novel algorithm using sampled surrogates of the underlying continuous outcome conditional on the binary observations. The methods proposed in this thesis are assessed via simulations and illustrated using the M-bridge study, a SMART with embedded tailoring which develops and evaluates adaptive interventions for preventing binge drinking among college students.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

With an increasing demand for quality healthcare and precision medicine, dynamic treatment regime has become an emerging field for statistical researchers. A dynamic treatment regime (DTR) (Murphy, 2003; Orellana et al., 2010; Chakraborty and Murphy, 2014; Laber et al., 2014), also known as adaptive intervention (Collins et al., 2004; Murphy et al., 2007; Lei et al., 2012; Nahum-Shani et al., 2020), or adaptive treatment strategy (Lavori and Dawson, 2000; Murphy and McKay, 2004; Kosorok and Moodie, 2015), is a sequence of functions, one for each decision, which map from a participant's history of characteristics and actions to a set of possible subsequent actions to take over time. The actions in general can be, depending on the context, treatments or interventions, and will be used interchangeably in this thesis. Such treatment regimes are *dynamic* because individualized treatments vary according to tailoring variables (Collins et al., 2004; Nahum-Shani et al., 2012). An example of a tailoring variable is whether a participant responds to the treatment assigned at the first stage, and the corresponding DTR might prescribe that responders continue the initial treatment at the second stage, whereas non-responders switch to an alternative treatment. This adaptive and

sequential nature of treatment assignments take patient heterogeneity into account and effectuates personalized medicine.

The sequential multiple assignment randomized trial (SMART) (Murphy, 2005a; Lavori and Dawson, 2014; Collins et al., 2007) is an experimental design that allows multiple stages of randomization and collects longitudinal data over these stages to help construct and analyze DTRs. The two main goals of analyzing the data collected from SMART studies are: (1) comparison between embedded DTRs, and (2) estimation of the optimal DTR. The former approach aims to analyze deterministic dynamic regimes, whereas the latter takes advantage of sequentially accrued information to individualize treatment at subsequent stages so that the subject ends up with the best possible result. This thesis focuses mainly on identifying the sequence of decision rules that optimizes the outcome in question and estimating this value at each time point of interest. Various methods to address this problem were proposed and discussed in literature. Chakraborty and Murphy (2014) presented a comprehensive description of direct and indirect methods to identify the optimal DTR, direct methods being marginal structural models (Robins, 2000a,b) and inverse probability of treatment weighting, and indirect methods including Q-learning and dynamic system models.

1.2 Q-learning: A Backward Induction Algorithm

Q-learning (Watkins, 1989) is a dynamic programming (Bellman, 1957) algorithm and has been widely implemented to identify the optimal DTR (Murphy and McKay, 2004; Nahum-Shani et al., 2012). At each stage of randomization, Q-learning specifies a model of expected outcomes conditional on past history, given that the optimal interventions are followed thereafter. The parametric form of the conditional expectation is called a *Q-function* (Sutton and Barto, 1998; Murphy, 2005b) and is usually written as a linear regression. Model parameters are estimated stagewise using backward induction (Aumann, 1995). In classification problems, the Q-functions may be estimated by regression

trees or kernels.

Suppose the longitudinal data collected from a SMART study are

$$(Z_1, A_1, Z_2, A_2, \ldots, Z_k, A_k, \ldots, Z_K, A_K, Y)$$

for stages $k = 1, \ldots, K$, where Z_k denotes the observed covariates and intermediate outcomes which are measured prior to randomization at stage k, $A_k \in \mathcal{A}_k$ denotes the treatment received at stage k, and Y denotes the final outcome of interest with small values preferred. Let $H_k = (Z_1, A_1, \ldots, Z_{k-1}, A_{k-1}, Z_k)$ be the history of covariates and treatments prior to randomization at stage k.

Optimization of DTRs borrows the concept of potential outcomes, and therefore the following causal assumptions are necessary:

- Consistency: The potential outcome under the observed treatment agrees with the observed outcome;
- No unmeasured confounders, also known as sequential ignorability: A_k is independent of all future potential outcomes, conditional on the history H_k.

The no unmeasured confounders assumption is valid with the sequential randomization of a SMART.

The Q-function at stage k, $Q_k(H_k, A_k)$, is a function of the observed history up to stage k and the treatment received at stage k, and usually consists of a main effect model and a treatment effect model.

Starting from the last stage, the linear regression model is specified as

$$Q_K(H_K, A_K) = \mathbb{E}(Y|H_K, A_K) = \boldsymbol{x}_{K0}^T \boldsymbol{\beta}_K + A_K \boldsymbol{x}_{K1}^T \boldsymbol{\psi}_K,$$

where $\boldsymbol{x}_{K0}$ and $\boldsymbol{x}_{K1}$ are realizations of functions of H_K that represent the covariates in the main effect and treatment effect model respectively. The least squares estimators of stage K parameters, $\hat{\boldsymbol{\beta}}_K$ and $\hat{\boldsymbol{\psi}}_K$, are obtained.

For a precedent stage k, $k = K-1, K-2, \ldots, 1$, the Q-function is specified as

$$Q_k(H_k, A_k) = \mathbb{E}\left[\min_{a_{k+1}\in\mathcal{A}_{k+1}} Q_{k+1}(H_{k+1}, A_{k+1} = a_{k+1}) \middle| H_k, A_k\right] = \boldsymbol{x}_{k0}^T\boldsymbol{\beta}_k + A_k\boldsymbol{x}_{k1}^T\boldsymbol{\psi}_k,$$

where $\boldsymbol{x}_{k0}$ and $\boldsymbol{x}_{k1}$ are realizations of functions of H_k. The least squares estimators of stage k parameters, $\hat{\boldsymbol{\beta}}_k$ and $\hat{\boldsymbol{\psi}}_k$ are obtained.

The estimated optimal DTR is $\left(\hat{d}_1^{\text{opt}}, \ldots, \hat{d}_k^{\text{opt}}, \ldots, \hat{d}_K^{\text{opt}}\right)$, with

$$\hat{d}_k^{\text{opt}} = \arg\min_{a_k\in\mathcal{A}_k} Q_k(H_k, A_k = a_k; \hat{\boldsymbol{\beta}}_k, \hat{\boldsymbol{\psi}}_k).$$

We assume a two-stage setting throughout the thesis, where $K = 2$.

1.3 The M-bridge Study

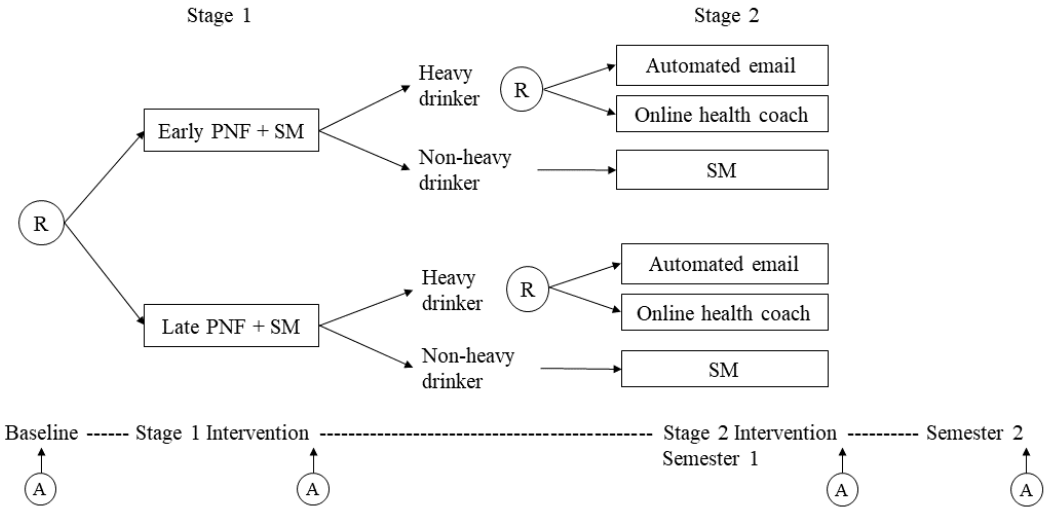

Figure 1.1: The M-bridge study: a sequential, multiple assignment, randomized trial. This figure is adapted from the figure of study design in Patrick et al. (2020). (R) indicates a randomization stage with arrows pointing to available treatment options, and (A) indicates time of assessment where measurements of intermediate or final outcome take place.

The M-bridge study (Patrick et al., 2020) is a SMART conducted at the University of Minnesota Twin Cities. The study develops and evaluates an adaptive preventive intervention for college drinking and related problems among first-year college students. As illustrated in Figure 1.1, enrolled students were randomized to receiving a combined universal preventive intervention, personalized normative feedback and self-monitoring (PNF+SM), either prior to attending college (early intervention) or in the first month of the first semester (late intervention). Two intermediate measures of alcohol use, namely, the frequency of binge drinking (consuming 4/5+ drinks in a row for women/men) and the frequency of high-intensity drinking (consuming 8/10+ drinks in a row for women/men) in the past two weeks, were self-monitored and recorded by students. Prior to re-randomization, students were flagged as a "heavy drinker" if they report two or more occasions of binge drinking, or one or more occasion of high-intensity drinking during a two-week self-monitoring period; otherwise, they were flagged as a "non-heavy drinker". This is a tailoring variable embedded in the SMART design, and the set of available intervention options at stage 2 depended on this tailoring variable. Heavy drinkers were re-randomized to receiving the indicated intervention of an automated email or an invitation to online health coaching, whereas non-heavy drinkers continued self-monitoring for the rest of the first semester. The adaptive design helps to determine whether more resources should be allocated to bridge the students to indicated interventions for alcohol use. The investigators collected a variety of continuous and dichotomous outcomes related to drinking behavior, consequences, and health services utilization.

1.4 Outline

This thesis discusses three challenges and concerns associated with the implementation of Q-learning in various contexts, and develops multiple modifications of Q-learning to address these problems. Chapter 2 is motivated by SMARTs with repeated-measures

outcomes at each stage. We develop Q-learning with generalized estimating equations which aims to generate an optimal trajectory considering all measurement times. One existing modification to Q-learning with Murphy's regret function (Murphy, 2003), that acts as a fundamental thought upon which this thesis is written, was proposed by Huang et al. (2015) and takes care of as much *residuals* from model misspecification as possible. To increase robustness, we then build Murphy's regret function into the proposed method. The susceptibility of the modified algorithm to different correlation structures and model misspecification will be explored through simulation studies. Chapter 3 considers modifications of Q-learning that specifically tackle model misspecifications associated with treatment effect heterogeneity. Some theoretical results will be derived to understand how unmeasured variables in the stage 2 main effect model affect stage 1 estimation, where omission of stage 1 treatment effects is a specific case. We propose to build interactive model (Laber et al., 2014) into residual-modified Q-learning to correct the bias generated as a result of heterogeneous treatment effects at both stages. Finally, Chapter 4 generalizes this approach to dichotomous outcomes. The feasibility to develop a robust algorithm to correctly identify the optimal DTR with dichotomous outcomes will be briefly discussed. It is indeed difficult to incorporate residual remainders from estimation of late-stage models into early-stage pseudo-outcomes. Instead, we develop a simple but nontrivial modification to Q-learning with probit regression by imposing monotonicity of preferences. The advantages and drawbacks of this modified algorithm will be demonstrated using simulation studies. We also propose an alternative approach using latent variable modeling and develop a novel algorithm that incorporates Murphy's regret function to sampled surrogates of the underlying latent variable. All the proposed modifications will be illustrated using the data collected from the M-bridge study. A different structure of the M-bridge data may be used in each chapter, and the data structure and the associated framework will be introduced separately.

Chapter 2

Modified Q-learning with Generalized Estimating Equations for Repeated-measures Outcomes

2.1 Motivation

In some SMARTs, investigators collect repeated-measures outcomes at one or more stages to monitor longitudinal treatment effects, especially at the final stage. This chapter is motivated by two such studies, the M-bridge study and the developmental language disorder (DLD) study. In the M-bridge study (Figure 1.1), the primary outcome is the frequency of binge drinking during a 30-day period and there was a baseline measurement before the randomization stages. The frequency of binge drinking over the past two weeks was recorded at the end of stage 1 intervention via self-monitoring. We multiply this frequency by 2 to give the approximate monthly frequency. The investigators repeatedly monitored the frequency of binge drinking thereafter, with follow-up

assessments immediately at the end of the first semester and at the end of the second semester. The DLD study evaluates a sequence of treatments for children with developmental language disorder (Finestack, 2018), and has a similar design. The investigators have a scientific interest in sustained treatment effects, so the participating children's performance was repeatedly assessed at 1, 6, and 12 months after the end of the treatment period (additional details can be found in Appendix A.2). The data from the DLD study are not available for analysis at the time of writing, so we will use the M-bridge study to illustrate our framework and method throughout the chapter. Recently, SMARTs with repeated-measures outcomes at one or more stages are of growing interest in literature, with the autism study (Kasari et al., 2014), the ENGAGE study (McKay et al., 2015), and the PLUTO study (Fu et al., 2017) as examples. The stagewise repeated measures provide information on the treatment effects from previous stages over time and motivate the development of new methodologies to adequately incorporate these additional follow-ups into statistical analysis.

The two main goals of statistical analysis for SMART studies often are (1) comparing embedded DTRs, and (2) estimating the optimal DTR. The first goal aims to estimate the marginal benefit of each DTR embedded in the SMART. For example, there are four embedded DTRs (or *adaptive preventive interventions*) in the M-bridge study: Early/Email, where first-year college students participated in PNF+SM prior to Semester 1, and were bridged to automated emails if they were flagged as heavy drinkers or continued self-monitoring if they were flagged as non-heavy drinkers; Early/Coach, where students were bridged to online health coach if they were flagged as heavy drinkers; Late/Email, where students participated in PNF+SM during the first month of Semester 1, and were bridged to automated emails if they were flagged as heavy drinkers; Late/Coach, where students were bridged to online health coach if they were flagged as heavy drinkers. The existing method to estimate the longitudinal trajectory of these embedded DTRs is the weighted and replicated GEE (Lu et al.,

2015). The second goal takes advantage of subject heterogeneity to identify the optimal DTR for each subject. The optimal DTR is a sequence of decision rules, one for each intervention stage, that optimize the expectation of an outcome of interest. In the M-bridge study, smaller values of the outcome are preferred. In addition, the investigators are interested in estimating the longitudinal trajectory of the optimal DTR and the heterogeneous causal effects at each stage to understand how the optimal DTR affects the outcome over time. The reasons for doing so might be to simply visualize the marginal trajectory of outcomes under the course of optimal regime, or to address a scientific interest, for example, if any DTR-specific trajectory almost coincides with the optimal trajectory. To the best of our knowledge, there is a lack of methods to tackle potential challenges in the optimization problem with repeated-measures outcomes. In this chapter, we propose some modifications of the Q-learning algorithm to address this issue.

One of the challenges in this context is to incorporate the longitudinal outcomes, whose clinical importance may differ according to the time of measurement. Classic Q-learning algorithm can deal with the repeated measures by simply collapsing them to a composite value, usually through a weighted sum. However, it fails to estimate the trajectory of heterogeneous causal effects or expected outcomes over the entire time period of interest. Thus, we propose to write the Q-functions as marginal models possibly with time-varying coefficients and estimate model parameters using generalized estimating equations (GEEs). Moreover, there is a possibility to misspecify the late-stage model. In the M-bridge study, some important interactions between baseline covariates and stage 1 intervention may be neglected when constructing the stage 2 model. To address this problem, we modify the proposed method with Murphy's regret function (Murphy, 2003; Huang et al., 2015).

We proceed by first formulating the causal framework to identify the optimal DTR and estimate the heterogeneous causal effects in Section 2.2. Furthermore, the details of our proposed method are explained in Section 2.3. Section 2.4 presents a comprehensive

simulation study to explore the susceptibility of the proposed method to different correlation structures between repeated-measures outcomes and model misspecification. We then apply modified Q-learning with GEE in Section 2.5, and compare the optimal trajectory with marginal trajectories of the embedded DTRs, which can also be estimated as a by-product of the method. Finally, we conclude with a brief discussion.

2.2 Statistical Framework for a SMART with Repeated-measures Outcomes

2.2.1 Causal Framework

Suppose that the longitudinal data collected from a SMART with repeated-measures outcomes are represented by a sequence of random variables $(Z_1, A_1, \boldsymbol{Y}_1, Z_2, A_2, \boldsymbol{Y}_2)$, where Z_1 is the set of baseline covariates and outcome measured prior to stage 1 randomization, Z_2 is the set of time-varying covariates and tailoring variables measured after stage 1 and before stage 2 randomization, and $A_k \in \mathcal{A}_k$, $k = 1, 2$, is the treatment that the participant receives at stage k, with $\mathcal{A}_k$ being the set of all possible treatments. In the M-bridge study, $\mathcal{A}_1 = \{-1, 1\}$, where $A_1 = 1$ represents early intervention and $A_1 = -1$ represents late intervention, and $\mathcal{A}_2 = \{-1, 1\}$, where $A_2 = 1$ represents online health coach and $A_2 = -1$ represents automated email. $\boldsymbol{Y}_1$ is the outcome measured between stage 1 and stage 2 treatments, and $\boldsymbol{Y}_2$ is the outcome measured after stage 2 treatment. This framework is generalized to stagewise repeated-measures outcomes, so both $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ can be vectors. In the M-bridge study, Y_1 is a scalar and $\boldsymbol{Y}_2$ is a vector. Our framework and method are flexible in both contexts, as long as we have repeated-measures outcome for at least one stage. In the M-bridge study, the outcome is the frequency of binge drinking, with smaller values preferred. Now let $H_1 = Z_1$ and $H_2 = (Z_1, A_1, \boldsymbol{Y}_1, Z_2)$ denote the covariate and treatment history up to the stage 1 and stage 2 randomization respectively.

Recall that the desired decision rules may not be the same as the observed ones, so solving the problem requires the concept of potential outcomes and naturally falls into the causal framework. For $k = 1, 2$, let $\bar{A}_k = (A_1, \ldots, A_k)$ be a stochastic sequence of treatments up to stage k, with values $\bar{a}_k = (a_1, \ldots, a_k)$. Define $\boldsymbol{Y}_k^{\bar{a}_k}$ to be the potential outcomes at stage k under the treatment sequence $\bar{a}_k$. In the context of repeated-measures outcomes, we rewrite the causal assumptions:

- Consistency: Let $\boldsymbol{Y}_k$ be the observed outcome at stage k under the treatment history $\bar{A}_k$. Then $\boldsymbol{Y}_k$ is equal to the potential outcome under $\bar{A}_k$. Thus, $\boldsymbol{Y}_k = \sum_{\bar{a}_k \in \prod_{j=1}^{k} \mathcal{A}_j} \mathbb{1}\left(\bar{A}_k = \bar{a}_k\right) \boldsymbol{Y}_k^{\bar{a}_k}$, where $\mathbb{1}\left(\bar{A}_k = \bar{a}_k\right) = \begin{cases} 1 & \text{if } \bar{A}_k = \bar{a}_k \\ 0 & \text{if } \bar{A}_k \neq \bar{a}_k \end{cases}$ is an indicator function.

- No unmeasured confounders, also known as sequential ignorability: For any treatment sequence $\bar{a}_{k-1}$ prior to stage k, A_k is independent of all future potential outcomes, conditional on the history H_k. Hence, $\left\{\boldsymbol{Y}_k^{(\bar{a}_{k-1}, a_k)} : a_k \in \mathcal{A}_k\right\} \perp A_k \mid H_k$. This assumption is satisfied in a SMART due to stagewise randomization.

Let $d_k : \mathcal{H}_k \mapsto \mathcal{A}_k$ be the decision rule at stage k with $d_k = d_k(H_k)$ for $H_k \in \mathcal{H}_k$, where $\mathcal{H}_k$ is a set of all possible history information prior to stage k. Applying the causal framework discussed by Chakraborty and Murphy (2014), the expected potential outcome under DTR (d_1, d_2) can be expressed as:

$$\mathbb{E}\begin{pmatrix} \boldsymbol{Y}_1^{(d_1)} \\ \boldsymbol{Y}_2^{(d_1,\ d_2)} \end{pmatrix} = \mathbb{E}_{H_2|H_1,A_1}\left\{ \sum_{a_2 \in \mathcal{A}_2} \mathbb{1}(a_2 = d_2) \underbrace{\begin{pmatrix} \boldsymbol{Y}_1^{(d_1)} \\ \mathbb{E}_{\boldsymbol{Y}_2|H_2,A_2}\left(\boldsymbol{Y}_2^{(d_1,A_2)} \middle| H_2, A_2 = a_2\right) \end{pmatrix}}_{(b)} \right\}$$

$$= \mathbb{E}_{H_1}\left[\sum_{a_1 \in \mathcal{A}_1} \mathbb{1}(a_1 = d_1) \overbrace{\mathbb{E}_{H_2|H_1,A_1}\left\{ \sum_{a_2 \in \mathcal{A}_2} \mathbb{1}(a_2 = d_2) \underbrace{\begin{pmatrix} \boldsymbol{Y}_1 \\ \mathbb{E}_{\boldsymbol{Y}_2|H_2,A_2}(\boldsymbol{Y}_2|H_2, A_2 = a_2) \end{pmatrix}}_{(a)} \middle| H_1, A_1 = a_1 \right\}}^{(b)} \right]. \tag{2.1}$$

This causal framework allows for estimation of expected potential outcomes using observed outcomes, and it is straightforward to generalize it to more than two stages. The

doubly-iterated expectation in Equation (2.1) provides a mathematical guidance for implementing Q-learning. Here, (a) and (b) are the key estimands and will be used to construct estimands to address the scientific problems of interest. We will discuss these estimands in greater detail in Section 2.2.2. The vector nature of $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ inspires the use of GEE techniques (Zeger and Liang, 1986) to estimate them, which will be discussed in Section 2.3.

2.2.2 Optimization Problem

There are four main scientific problems we would like to address in this chapter: (1) to identify the optimal decision rule at each stage as a function of the history prior to that stage, (2) to estimate the optimal trajectory over time, (3) to estimate the heterogeneous causal effects of stage 2 treatment over time, conditional on the history prior to stage 2, and (4) to estimate the heterogeneous causal effects of stage 1 treatment over time, provided that participants follow the optimal treatment at stage 2. Note that the outcome of interest is formed by a sequence of vectors, and it is impossible to minimize a vector without any partial orders. Scientifically, the importance of these outcomes may vary based on the time of measurement, and it is for clinical investigators to decide. Thus, we assign weights $\boldsymbol{w}_k$ to each $\boldsymbol{Y}_k$, $k = 1, 2$, where elements of $\boldsymbol{w}_k$ are nonnegative, and the scalar $\sum_{k=1}^{2} \boldsymbol{w}_k^T \boldsymbol{Y}_k$ is the target of optimization. Since outcomes with smaller magnitude are more desirable in the M-bridge study and elements of $\boldsymbol{w}_k$ are nonnegative, smaller values of $\sum_{k=1}^{2} \boldsymbol{w}_k^T \boldsymbol{Y}_k$ are preferred.

The underlying estimands in the above-mentioned scientific problems can be expressed using (a) and (b). We restrict the problem in the comparison between two available treatments at a specific stage, as described in the M-bridge study. The framework can be generalizable to multiple treatment options, with treatment effects defined through pairwise comparison and contrasts.

Conditional on the history prior to stage 2, the heterogeneous causal effects, also

known as conditional average treatment effects (CATE) (Abrevaya et al., 2015), or moderated causal effects (Wodtke and Almirall, 2017) of stage 2 treatment over time, are defined as

$$\boldsymbol{\tau}_2(H_2) = \mathbb{E}(\boldsymbol{Y}_2|H_2, A_2 = 1) - \mathbb{E}(\boldsymbol{Y}_2|H_2, A_2 = -1). \tag{2.2}$$

Therefore, the optimal treatment rule at stage 2 is $d_2^{\text{opt}}(H_2) = -\text{sgn}\left\{\boldsymbol{w}_2^T\boldsymbol{\tau}_2(H_2)\right\}$, where $\text{sgn}(x) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{otherwise} \end{cases}$.

Suppose the subjects follow the optimal treatment at stage 2, the heterogeneous causal effects of stage 1 treatment over time is defined as

$$\boldsymbol{\tau}_1(H_1) = \mathbb{E}\left\{ \begin{pmatrix} \boldsymbol{Y}_1 \\ \mathbb{E}\left(\boldsymbol{Y}_2 \left| H_2, A_2 = d_2^{\text{opt}} \right.\right) \end{pmatrix} \middle| H_1, A_1 = 1 \right\} - \mathbb{E}\left\{ \begin{pmatrix} \boldsymbol{Y}_1 \\ \mathbb{E}\left(\boldsymbol{Y}_2 \left| H_2, A_2 = d_2^{\text{opt}} \right.\right) \end{pmatrix} \middle| H_1, A_1 = -1 \right\}, \tag{2.3}$$

and hence the optimal treatment rule at stage 1 is $d_1^{\text{opt}}(H_1) = -\text{sgn}\left\{\left(\boldsymbol{w}_1^T\ \boldsymbol{w}_2^T\right)\boldsymbol{\tau}_1(H_1)\right\}$.

The marginal optimal trajectory is obtained by substituting $d_1 = d_1^{\text{opt}}(H_1)$ and $d_2 = d_2^{\text{opt}}(H_2)$ in Equation (2.1).

2.3 Modified Q-learning with Generalized Estimating Equations

Q-learning is a widely implemented algorithm to identify the optimal DTR. It is also discussed in literature (Huang et al., 2015) that the expected trajectory of embedded treatment regimes can be estimated as a by-product of Q-learning. Instead of specifying a simple linear regression for each Q-function, we propose to utilize a marginal model that can generate point estimates at all measurement times. GEE techniques are then used to estimate the model parameters, taking advantage of the robustness to misspecification of working covariance matrix.

2.3.1 Composite Q-learning

Suppose $(Z_{1i}, A_{1i}, \boldsymbol{Y}_{1i}, Z_{2i}, A_{2i}, \boldsymbol{Y}_{2i})$ is the sample data collected for subject i, $i = 1, \ldots, n$. In order to apply the standard Q-learning algorithm, we could *collapse* repeated-measures outcomes using a weighted sum, and the Q-functions are defined as:

$$\begin{aligned} Q_2(H_{2i}, A_{2i}) &= \mathbb{E}\left(\boldsymbol{w}_2^T \boldsymbol{Y}_{2i} \middle| H_{2i}, A_{2i}\right), \\ Q_1(H_{1i}, A_{1i}) &= \mathbb{E}\left(\boldsymbol{w}_1^T \boldsymbol{Y}_{1i} + Q_2\left(H_{2i}, A_{2i} = d_2^{\text{opt}}\right) \middle| H_{1i}, A_{1i}\right), \end{aligned} \tag{2.4}$$

where $\boldsymbol{w}_k$ is a clinically specified weight vector for outcomes at stage k, as described in Section 2.2.2. Consider Q-functions of the parametric form

$$Q_k(H_{ki}, A_{ki}; \boldsymbol{\beta}_k, \boldsymbol{\psi}_k) = \boldsymbol{x}_{k0,i}^T \boldsymbol{\beta}_k + A_{ki} \boldsymbol{x}_{k1,i}^T \boldsymbol{\psi}_k, \ k = 1, 2, \tag{2.5}$$

where $\boldsymbol{x}_{k0,i}$ and $\boldsymbol{x}_{k1,i}$ are realizations of functions of H_{ki}. Starting from stage 2, $\hat{\boldsymbol{\beta}}_2$ and $\hat{\boldsymbol{\psi}}_2$, are obtained using least squares estimation. Stage 1 estimators, $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\psi}}_1$, are estimated by regressing $\boldsymbol{w}_1^T \boldsymbol{Y}_1 + \min_{a_2 \in \{-1,1\}} Q_2\left(H_2, A_2 = a_2; \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\psi}}_1\right)$ on H_1 and A_1. We term this type of Q-learning as *composite Q-learning*.

2.3.2 Q-learning with Generalized Estimating Equations

Our proposed method does not collapse repeated-measures outcomes but treats repeated-measures outcomes as a vector. Let $\boldsymbol{t}_k = (t_{k1} \ \ldots \ t_{kj} \ \ldots \ t_{kJ_k})^T$ be a vector of all measurement times right after the treatment period at stage k, including measurements at future stages. Throughout the chapter, we assume that measurements are taken at the same time across subjects. Therefore, the Q-functions are re-defined as

$$\begin{aligned} \boldsymbol{Q}_2(H_{2i}, A_{2i}, \boldsymbol{t}_2) &= \mathbb{E}\left(\boldsymbol{Y}_{2i} | H_{2i}, A_{2i}\right), \\ \boldsymbol{Q}_1(H_{1i}, A_{1i}, \boldsymbol{t}_1) &= \mathbb{E}\left\{ \begin{pmatrix} \boldsymbol{Y}_{1i} \\ \boldsymbol{Q}_2\left(H_{2i}, A_{2i} = d_{2i}^{\text{opt}}, \boldsymbol{t}_2\right) \end{pmatrix} \middle| H_{1i}, A_{1i} \right\}. \end{aligned} \tag{2.6}$$

Consider a linear model with time-varying coefficients $\boldsymbol{\beta}_{kj}$ and $\boldsymbol{\psi}_{kj}$

$$Q_{kj}\left(H_{ki}, A_{ki}, t_{kj}; \boldsymbol{\beta}_{kj}, \boldsymbol{\psi}_{kj}\right) = \boldsymbol{x}_{k0,i}^T \boldsymbol{\beta}_{kj} + A_{ki} \boldsymbol{x}_{k1,i}^T \boldsymbol{\psi}_{kj},\ j = 1, \ldots, J_k,\ k = 1, 2, \quad (2.7)$$

and $\boldsymbol{Q}_k = (Q_{k1}\ \ldots\ Q_{kj}\ \ldots\ Q_{kJ_k})^T$. Thus, $\boldsymbol{\beta}_k = \left(\boldsymbol{\beta}_{k1}^T\ \ldots\ \boldsymbol{\beta}_{kj}^T\ \ldots\ \boldsymbol{\beta}_{kJ_k}^T\right)^T$ and $\boldsymbol{\psi}_k = \left(\boldsymbol{\psi}_{k1}^T\ \ldots\ \boldsymbol{\psi}_{kj}^T\ \ldots\ \boldsymbol{\psi}_{kJ_k}^T\right)^T$ are the parameters to be estimated at stage k.
The stage 2 estimators, $\hat{\boldsymbol{\beta}}_2$ and $\hat{\boldsymbol{\psi}}_2$, are obtained using generalized estimating equations

$$\sum_{i=1}^{n} \boldsymbol{D}_2(H_{2i}, A_{2i}, \boldsymbol{t}_2)^T V_2^{-1} \left\{\boldsymbol{Y}_{2i} - \boldsymbol{Q}_2\left(H_{2i}, A_{2i}, \boldsymbol{t}_2; \boldsymbol{\beta}_2, \boldsymbol{\psi}_2\right)\right\} = 0, \quad (2.8)$$

where $\boldsymbol{D}_2(H_{2i}, A_{2i}, \boldsymbol{t}_{2i}) = \dfrac{\partial \boldsymbol{Q}_2}{\partial\left(\boldsymbol{\beta}_2^T\ \boldsymbol{\psi}_2^T\right)}$ and V_2 is the working covariance for $\mathrm{Var}(\boldsymbol{Y}_2 | H_2, A_2)$.
Similarly, the stage 1 estimators, $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\psi}}_1$, are calculated using

$$\sum_{i=1}^{n} \boldsymbol{D}_1(H_{1i}, A_{1i}, \boldsymbol{t}_1)^T V_1^{-1} \left\{ \begin{pmatrix} \boldsymbol{Y}_{1i} \\ \hat{\boldsymbol{Y}}_{2i} \end{pmatrix} - \boldsymbol{Q}_1(H_{1i}, A_{1i}, \boldsymbol{t}_1; \boldsymbol{\beta}_1, \boldsymbol{\psi}_1) \right\} = 0, \quad (2.9)$$

where

$$\begin{aligned} \boldsymbol{D}_1(H_{1i}, A_{1i}, \boldsymbol{t}_1) &= \frac{\partial \boldsymbol{Q}_1}{\partial\left(\boldsymbol{\beta}_1^T\ \boldsymbol{\psi}_1^T\right)}, \\ \hat{d}_{2i}^{\mathrm{opt}} &= \underset{a_2 \in \mathcal{A}_2}{\arg\min}\, \boldsymbol{w}_2^T \boldsymbol{Q}_2\left(H_{2i}, A_{2i} = a_2; \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\psi}}_2\right), \\ \hat{\boldsymbol{Y}}_{2i} &= \boldsymbol{Q}_2\left(H_{2i}, A_{2i} = \hat{d}_{2i}^{\mathrm{opt}}, \boldsymbol{t}_2; \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\psi}}_2\right), \end{aligned}$$

and V_1 is the working covariance for $\mathrm{Var}\left\{ \begin{pmatrix} \boldsymbol{Y}_{1i} \\ \hat{\boldsymbol{Y}}_{2i} \end{pmatrix} \middle| H_1, A_1 \right\}$.

The GEE approach allows for missingness in responses over time. Taking the M-bridge study as an example, we have missingness in stage 1 outcome and stage 2 outcome at both follow-ups. Since the outcomes depend only on covariates, we should have a complete set of baseline and intermediate covariates for model estimation at each stage.

An issue with respect to the estimation procedure in Q-learning is that stage 2 outcome depends on stage 1 outcome. Hence, missingness in stage 1 outcome causes problems in stage 2 estimation. The design of the M-bridge study obviates this issue by introducing the tailoring variable. The tailoring variable, whether the students were flagged as heavy drinkers, is defined based on the stage 1 outcome, and those with missing stage 1 outcomes were not flagged as heavy drinkers. Since only heavy drinkers were re-randomized at stage 2, the students we include in stage 2 estimation all have complete covariate data.

In practice, choosing a wrong working structure for V_1 and V_2 can raise issues in estimation. Liang and Zeger (1986) argued that there was little difference in estimation of parameters when the true correlation was moderate. Zhao et al. (1992) claimed that wrong specification of an independent correlation matrix when outcomes were strongly dependent would result in "important losses of efficiency", as compared to the smaller reduction in efficiency when specifying unnecessary high correlations. In the presence of missingness not completely at random, Fitzmaurice et al. (1993) advocated for obtaining a close approximation to the covariance structure in outcomes so that the estimators of time-dependent treatment effects were substantially less biased. SMARTs usually result in a balanced design with time-invariant covariates. Moreover, the M-bridge study has relatively few repeated measures at each stage. Considering all these arguments, we recommend an unstructured working correlation structure.

The estimands discussed in Section 2.2.2 can now be written as functions of parameter estimates. The estimated optimal decision rule at stage k for individual i is identified by minimizing a weighted sum of Q-functions at stage k:

$$\begin{aligned} \hat{d}_{2i}^{\text{opt}} &= \arg\min_{a_2 \in \mathcal{A}_2} \boldsymbol{w}_2^T \boldsymbol{Q}_2 \left(H_{2i}, A_{2i} = a_2; \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\psi}}_2 \right), \\ \hat{d}_{1i}^{\text{opt}} &= \arg\min_{a_1 \in \mathcal{A}_1} \left(\boldsymbol{w}_1^T \ \boldsymbol{w}_2^T \right) \boldsymbol{Q}_1 \left(H_{1i}, A_{1i} = a_1; \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\psi}}_1 \right). \end{aligned} \tag{2.10}$$

The estimated heterogeneous causal effects at times of interest following the stage k

treatment period are calculated as:

$$\hat{\boldsymbol{\tau}}_k(H_k) = \boldsymbol{Q}_k\left(H_k, A_k = 1, \boldsymbol{t}_k; \hat{\boldsymbol{\beta}}_k, \hat{\boldsymbol{\psi}}_k\right) - \boldsymbol{Q}_k\left(H_k, A_k = -1, \boldsymbol{t}_k; \hat{\boldsymbol{\beta}}_k, \hat{\boldsymbol{\psi}}_k\right),\ k = 1, 2. \tag{2.11}$$

2.3.3 Modified Q-learning with Murphy's Regret Function

Model misspecification is a common issue in the implementation of Q-learning. Q-learning requires models at all stages to be correctly specified for consistent estimation of the optimal DTR and heterogeneous causal effects. In this thesis, we assume that the treatment effect models are correctly specified. Backward induction features multi-stage analysis, so even without any unmeasured confounders, misspecification of main effect model at a later stage can bias the estimation of treatment effects at earlier stages, thus resulting in an adverse impact on identification of the optimal DTR and the estimation of the optimal trajectory. Because the focus of the stage 2 model is on the estimation of the stage 2 heterogeneous causal effects, the interaction between baseline covariates and stage 1 treatment may be overlooked in the stage 2 model. Not adjusting for some important interactions in the stage 2 model and only including them in the stage 1 model will result in biased stage 1 treatment effect estimators. In the example of M-bridge, it is more difficult to select correct $Z_1 * A_1$ interactions into stage 2 model a priori due to the large number of baseline covariates that the study is considering. Thus, we would like to introduce some robustness to our proposed method by borrowing the concept of modified Q-learning (Huang et al., 2015).

The modified version of composite Q-learning defines the Q-function at stage 1 as

$$Q_1(H_{1i}, A_{1i}) = \mathbb{E}\left(\boldsymbol{w}_1^T \boldsymbol{Y}_{1i} + \boldsymbol{w}_2^T \boldsymbol{Y}_{2i} + \Delta_{2i} \middle| H_{1i}, A_{1i}\right), \tag{2.12}$$

where Δ_{2i} is the Murphy's regret function at stage 2 and

$$\Delta_{2i} = \begin{cases} 0 & \text{if } A_{2i} = d_{2i}^{\text{opt}} \\ -2\left|\boldsymbol{x}_{21,i}^T \boldsymbol{\psi}_2\right| & \text{if } A_{2i} \neq d_{2i}^{\text{opt}} \end{cases}.$$

Thus, the stage 1 estimators in modified composite Q-learning are obtained by regressing $\boldsymbol{w}_1^T \boldsymbol{Y}_1 + \boldsymbol{w}_2^T \boldsymbol{Y}_{2i} - 2\left|\boldsymbol{x}_{21,i}^T \hat{\boldsymbol{\psi}}_2\right| \mathbb{1}\left\{A_{2i} \neq \hat{d}_{2i}^{\text{opt}}\right\}$ on H_1 and A_1.

Applying this technique to Q-learning with GEE, our proposed modification to the algorithm illustrated in Section 2.3.2 updates stage 1 Q-function (2.6) as

$$Q_1(H_{1i}, A_{1i}) = \mathbb{E}\left\{ \begin{pmatrix} \boldsymbol{Y}_{1i} \\ \boldsymbol{Y}_{2i} + \boldsymbol{\Delta}_{2i} \end{pmatrix} \middle| H_{1i}, A_{1i} \right\}, \tag{2.13}$$

where

$$\boldsymbol{\Delta}_{2i} = \begin{cases} \boldsymbol{0} & \text{if } A_{2i} = d_{2i}^{\text{opt}} \\ d_{2i}^{\text{opt}} \boldsymbol{\tau}_2(H_{2i}) & \text{if } A_{2i} \neq d_{2i}^{\text{opt}} \end{cases}.$$

Thus, the stage 1 estimators in modified Q-learning with GEE are obtained using Equation (2.9) with $\hat{\boldsymbol{Y}}_{2i} = \boldsymbol{Y}_{2i} + \hat{d}_{2i}^{\text{opt}} \hat{\boldsymbol{\tau}}_2(H_{2i}) \mathbb{1}\left\{A_{2i} \neq \hat{d}_{2i}^{\text{opt}}\right\}$. This approach makes as much use of observations as possible and is robust to misspecification of stage 2 main effect model.

2.4 Simulation Study

In this section, we present a simulation study to compare the performance of the four methods illustrated in Section 3.5: (1) composite Q-learning (Q), (2) modified composite Q-learning (mQ), (3) Q-learning with GEE (Q-GEE), and (4) modified Q-learning with GEE (mQ-GEE).

2.4.1 Data Generative Mechanism

For simplicity, we omit any intermediate and tailoring variables in the data generative mechanism. Assume a sequence of observations from a simple SMART study with no embedded tailoring variable is $(Z_{1i}, A_{1i}, Y_{1i}, A_{2i}, \boldsymbol{Y}_{2i})$, $i = 1, \ldots, n$, where $Z_{1i} \overset{\text{i.i.d.}}{\sim} N(0,1)$. A_{1i} and A_{2i} both follow an i.i.d. Rademacher distribution, so $\frac{A_{1i}+1}{2} \overset{\text{i.i.d.}}{\sim}$ Bernoulli(0.5) and $\frac{A_{2i}+1}{2} \overset{\text{i.i.d.}}{\sim}$ Bernoulli(0.5). The scalar $Y_{1i} \in \mathbb{R}$ is the observed outcome at the end of stage 1. The vector $\boldsymbol{Y}_{2i} \in \mathbb{R}^2$ is the observed repeated-measures outcome at the end of stage 2. Smaller values are preferred for Y_{1i} and each element of $\boldsymbol{Y}_{2i}$. Suppose that

$$\begin{pmatrix} Y_{1i} \\ \boldsymbol{Y}_{2i} \end{pmatrix} \Bigg| \, Z_{1i}, A_{1i}, A_{2i} \sim MVN \left[\begin{pmatrix} \alpha_1 Z_{1i} A_{1i} \\ \alpha_2 Z_{1i} A_{1i} + \gamma_2 A_{1i} A_{2i} \\ \alpha_3 Z_{1i} A_{1i} + \gamma_3 A_{1i} A_{2i} \end{pmatrix}, \Sigma \right]$$

where $\alpha_1, \alpha_2, \gamma_2, \alpha_3, \gamma_3 < 0$ and

$$\Sigma = \begin{pmatrix} \lambda_1^2\sigma_v^2 + \sigma_e^2 & \lambda_1\lambda_2\sigma_v^2 & \lambda_1\lambda_3\sigma_v^2 \\ \lambda_1\lambda_2\sigma_v^2 & \lambda_2^2\sigma_v^2 + \sigma_e^2 & \lambda_2\lambda_3\sigma_v^2 \\ \lambda_1\lambda_3\sigma_v^2 & \lambda_2\lambda_3\sigma_v^2 & \lambda_3^2\sigma_v^2 + \sigma_e^2 \end{pmatrix}.$$

We are able to control the direction and magnitude of the conditional covariance Σ through parameters $\lambda_1, \lambda_2, \lambda_3, \sigma_e, \sigma_v$. The existence of conditional covariance can be viewed as adding an unmeasured covariate to the expected outcomes (refer to Appendix A.3 for a full derivation). The effects of the unmeasured covariate on Y_1 and $\boldsymbol{Y}_2$ can be similar or opposite. Suppose that the parameter values are $(\alpha_1, \alpha_2, \alpha_3, \gamma_2, \gamma_3) = -(2.0, 1.7, 1.6, 1.2, 0.8)$. The error dispersion is set to $\sigma_e = 2.7$ and $\sigma_v = 3.2$ so that the marginal coefficient of determination R^2 (Zheng, 2000) is around 0.25 to 0.60 for stage 1 model, and around 0.33 to 0.45 for stage 2 model. We will explore the performance of above-mentioned methods when partial correlations vary in the following cases:

(I) Y_1 and $\boldsymbol{Y}_2$ are not conditionally correlated, $\lambda_1 = \lambda_2 = \lambda_3 = 0$.

(II) Y_1 and $\boldsymbol{Y}_2$ are positively conditionally correlated, $\lambda_1 = \lambda_2 = \lambda_3 = 1$.

(III) Y_1 and $\boldsymbol{Y}_2$ are negatively conditionally correlated, $\lambda_1 = 1$ and $\lambda_2 = \lambda_3 = -1$.

2.4.2 Evaluation Criteria

Three metrics are defined to assess the consistency and efficiency of $\hat{d}_k^{\text{opt}}$ and $\hat{\boldsymbol{\tau}}_1$:

(1) $\text{PCI}_k = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{\hat{d}_{ki}^{\text{opt}} = d_{ki}^{\text{opt}}\right\}$: *the probability of correctly identified optimal rules* (PCI) at stage k, where $\hat{d}_{ki}^{\text{opt}}$ is the estimated optimal decision rule at stage k for subject i using Equation (2.10) with elements of $\boldsymbol{Q}_k$ being equally weighted, and d_{ki}^{opt} is the true optimal decision rule at stage k for subject i, minimizing the averaged outcomes following stage k treatment.

(2) $\text{RMSE}_{\tau_{1j}} = \left[\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\tau}_{1j}(Z_{1i}) - \tau_{1j}(Z_{1i})\right\}^2\right]^{1/2}$: *root mean square error* (RMSE) of the estimated heterogeneous causal effects of stage 1 treatment at the jth measurement, $j = 1, 2, 3$, assuming participants follow the optimal treatment at stage 2. $\hat{\tau}_{1j}(Z_{1i})$ is the jth element of $\hat{\boldsymbol{\tau}}_1$ from Equation (2.11).

(3) $\text{Bias}_{\tau_{1j}(Z_1)} = \hat{\tau}_{1j}(Z_1) - \tau_{1j}(Z_1)$ for a grid of values Z_1, where $\tau_{1j}(Z_1)$ is the jth element of $\boldsymbol{\tau}_1$ from Equation (2.3).

RMSE is the standard deviation of prediction errors. $\text{RMSE}_{\tau_{1j}}$ provides information on both accuracy and efficiency of using the stage 1 model to estimate heterogeneous causal effects across samples, and measures how accurate and precise the predicted heterogeneous causal effects are to the true values of the estimands over a sample of participants. $\text{Bias}_{\tau_{1j}(Z_1)}$ is composed of two components: (1) the bias from the estimation of stage 2 optimal rule, and (2) the bias from the estimation of stage 1 parameters. Bias should be heterogeneous across participants and dependent on the value of baseline measurement Z_1.

True values of the estimands are derived analytically as follows.

Let d_{ki}^{opt} be the true optimal decision rule at stage k for subject i, minimizing a weighted average of all outcomes following stage k treatment:

$$\begin{aligned} d_{2i}^{\text{opt}}(Z_{1i}, A_{1i}) &= \underset{a_2 \in \{-1,1\}}{\arg\min}\ \mathbb{E}\left(\boldsymbol{w}_2^T \boldsymbol{Y}_{2i} \middle| Z_{1i}, A_{1i}, Y_{1i}, A_{2i} = a_2\right) \\ &= -\text{sgn}\left\{\boldsymbol{w}_2^T \begin{pmatrix} \gamma_2 A_{1i} \\ \gamma_3 A_{1i} \end{pmatrix}\right\} = A_{1i}, \\ d_{1i}^{\text{opt}}(Z_{1i}) &= \underset{a_1 \in \{-1,1\}}{\arg\min}\ \mathbb{E}\left(w_1 Y_{1i} + \boldsymbol{w}_2^T \boldsymbol{Y}_{2i}\left(d_{2i}^{\text{opt}}\right) \middle| Z_{1i}, A_{1i} = a_1\right) \\ &= -\text{sgn}\left\{w_1 \alpha_1 Z_{1i} + \boldsymbol{w}_2^T \begin{pmatrix} \alpha_2 Z_{1i} \\ \alpha_3 Z_{1i} \end{pmatrix}\right\} = \text{sgn}(Z_{1i}), \end{aligned}$$

Let $\tau_{1j}(Z_1)$ be the expected stage 1 heterogeneous treatment effect at time j conditional on the baseline characteristics Z_1, provided that the true optimal treatment is followed at stage 2:

$$\boldsymbol{\tau}_1(Z_{1i}) = 2 \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} Z_{1i}.$$

2.4.3 Results for Equally Weighted Outcomes

To apply Q-learning with GEE, $H_2 = (Z_1, A_1, Y_1)$ and the interaction between A_2 and H_2 are included as covariates in Q_2, and $H_1 = Z_1$ and the interaction between A_1 and H_1 are included as covariates in Q_1, all with time-varying coefficients. Composite Q-learning uses the same set of covariates, but without time-varying coefficients. We consider two scenarios: (I) Q_2 is correctly specified with $Z_1 * A_1$ in the main effect model, and (II) Q_2 is misspecified by omitting $Z_1 * A_1$ in the main effect model. Here, we show the results under weights $w_1 = 1/3$ and $\boldsymbol{w}_2 = (1/3, 1/3)$ with a sample size $n = 200$. Details of model specifications can be found in Appendix A.4.

2.4.3.1 Scenario I: no model misspecification

As shown in Table 2.1 and Figure 2.1(a), there is a tiny loss in the accuracy of identifying stage 1 optimal rules using Q-learning with GEE as compared to composite Q-learning. The loss is likely due to the additional time-varying coefficients in the marginal model that Q-GEE has to estimate and is negligible considering the high accuracy. Q-lGEE is able to estimate heterogeneous causal effects (Table 2.1 and Figure 2.2(a)), and the modified version of Q-GEE has exactly the same performance as the standard algorithm.

Partial Correlation	Method	PCI_1	PCI_2	$RMSE_{\tau_{11}}$	$RMSE_{\tau_{12}}$	$RMSE_{\tau_{13}}$
Positive	Q	0.950	0.983	-	-	-
	mQ	0.950	0.983	-	-	-
	Q-GEE	0.947	0.983	0.74 (0.38)	0.97 (0.52)	0.94 (0.49)
	mQ-GEE	0.947	0.983	0.74 (0.38)	0.97 (0.52)	0.94 (0.49)
Independent	Q	0.976	0.998	-	-	-
	mQ	0.976	0.998	-	-	-
	Q-GEE	0.974	0.998	0.49 (0.26)	0.69 (0.36)	0.68 (0.35)
	mQ-GEE	0.974	0.998	0.49 (0.26)	0.69 (0.36)	0.68 (0.35)
Negative	Q	0.970	0.984	-	-	-
	mQ	0.970	0.984	-	-	-
	Q-GEE	0.966	0.984	0.74 (0.38)	0.98 (0.51)	0.94 (0.51)
	mQ-GEE	0.966	0.984	0.74 (0.38)	0.98 (0.51)	0.94 (0.51)

Table 2.1: (Scenario I) PCI of optimal rules and RMSE (mean (SD)) of estimated heterogeneous causal effects, based on estimated stage 1 Q-functions from 1000 simulations with sample size $n = 200$.

2.4.3.2 Scenario II: misspecified stage 2 main effect model

Under this circumstance, mQ-GEE universally outperforms Q-GEE in PCI_1 (Table 2.2 and Figure 2.1(b)), the RMSE of predicting heterogeneous causal effects (Table 2.2 and Figure 2.2(b)), and the bias of estimating heterogeneous causal effects based on a grid of Z_1 values (Figure 2.3(b)). This confirms our hypothesis that modified Q-learning avoids bias caused by misspecification of stage 2 main effect model. For more extreme

values of Z_1, there exists a remarkable bias (Figure 2.3) using Q-GEE in the presence of nonzero partial correlations. This bias does not decrease to 0 as sample size increases, suggesting that Q-GEE estimators are not consistent in this case.

Partial Correlation	Method	PCI_1	PCI_2	$RMSE_{\tau_{11}}$	$RMSE_{\tau_{12}}$	$RMSE_{\tau_{13}}$
Positive	Q	0.944	0.990	-	-	-
	mQ	0.951	0.990	-	-	-
	Q-GEE	0.940	0.990	0.73 (0.39)	1.20 (0.56)	1.13 (0.52)
	mQ-GEE	0.949	0.990	0.73 (0.39)	0.96 (0.50)	0.96 (0.48)
Independent	Q	0.957	0.995	-	-	-
	mQ	0.973	0.995	-	-	-
	Q-GEE	0.948	0.995	0.49 (0.25)	2.28 (0.55)	2.13 (0.53)
	mQ-GEE	0.971	0.995	0.49 (0.25)	0.75 (0.40)	0.73 (0.37)
Negative	Q	0.845	0.954	-	-	-
	mQ	0.965	0.954	-	-	-
	Q-GEE	0.702	0.954	0.74 (0.40)	4.70 (0.76)	4.56 (0.65)
	mQ-GEE	0.958	0.954	0.74 (0.40)	1.11 (0.56)	1.05 (0.55)

Table 2.2: (Scenario II) PCI of stage 1 optimal rules and RMSE of estimated heterogeneous causal effects at time 2 and 3, based on estimated stage 1 Q-functions from 1000 simulations.

Combining the results in both scenarios discussed above, we can conclude that partially correlated outcomes can be treated as a specific type of model misspecification, but it cannot be relieved by mQ-GEE for all cases of correlated outcomes. However, mQ-GEE is a more robust algorithm and has a decent and stable performance over all cases and metrics. Therefore, mQ-GEE should be used in analyzing SMART data with repeated-measures outcomes, especially in the case where stage 2 outcomes are believed to be conditionally negatively correlated with stage 1 outcomes. We also explored the relative efficiency of estimators under different stage 1 working correlations (Appendix A.5). With the relatively small data structure (3 repeated measures in total) in this simulation study and the saturated models considered, the potential impact of choosing a wrong working correlation matrix is small.

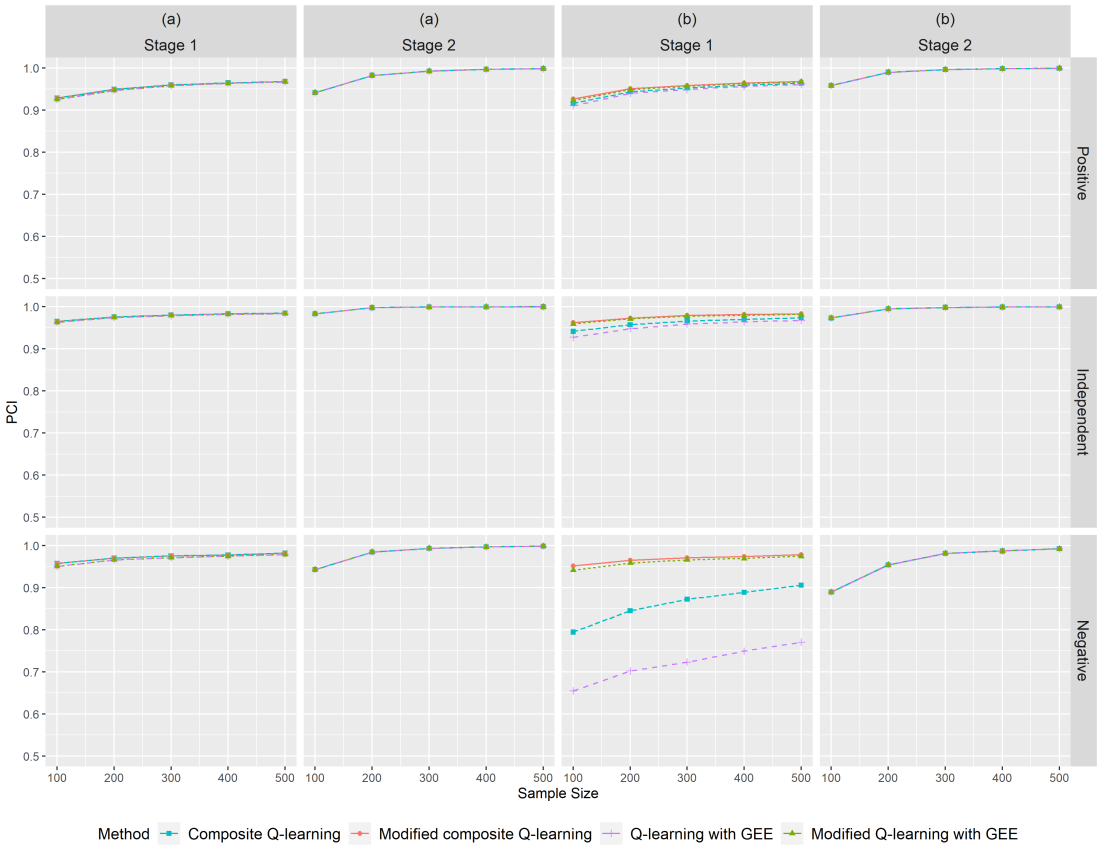

Figure 2.1: The probability of correctly identifying the optimal rules across a grid of sample sizes for scenarios (a) when outcomes are correlated, and (b) when outcomes are correlated and the main effect model in stage 2 Q-function is misspecified, based on 1000 iterations.

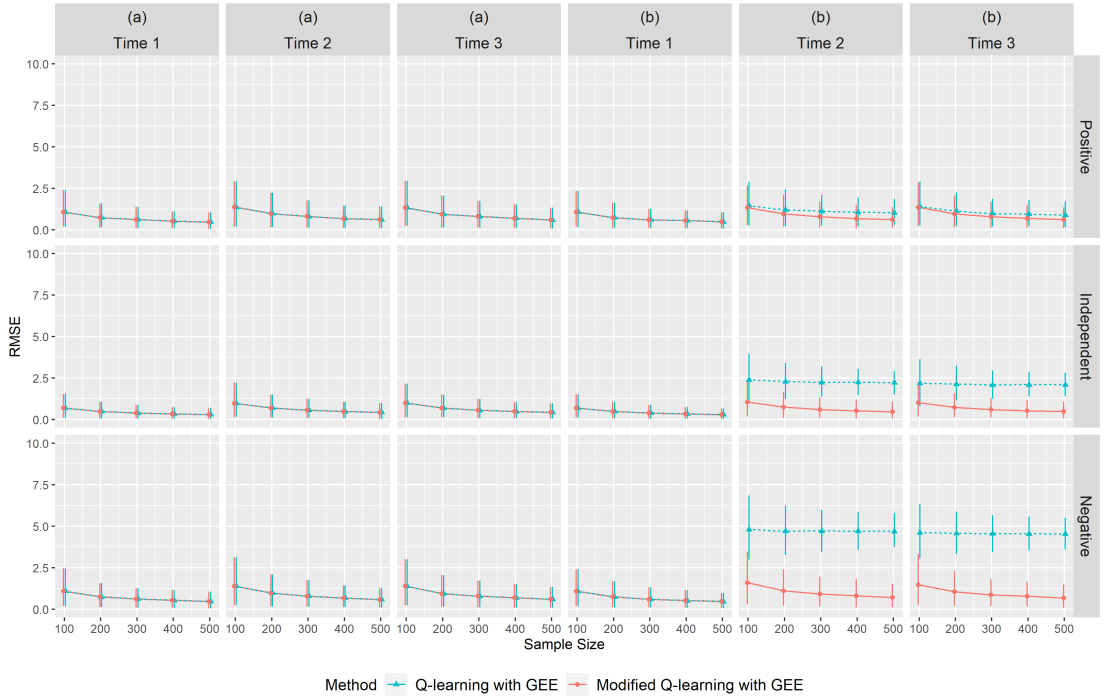

Figure 2.2: Root mean square error (mean and 95% confidence interval) of heterogeneous causal effects across a grid of sample sizes for scenarios (a) when outcomes are correlated, and (b) when outcomes are correlated and the main effect model in stage 2 Q-function is misspecified, based on 1000 iterations.

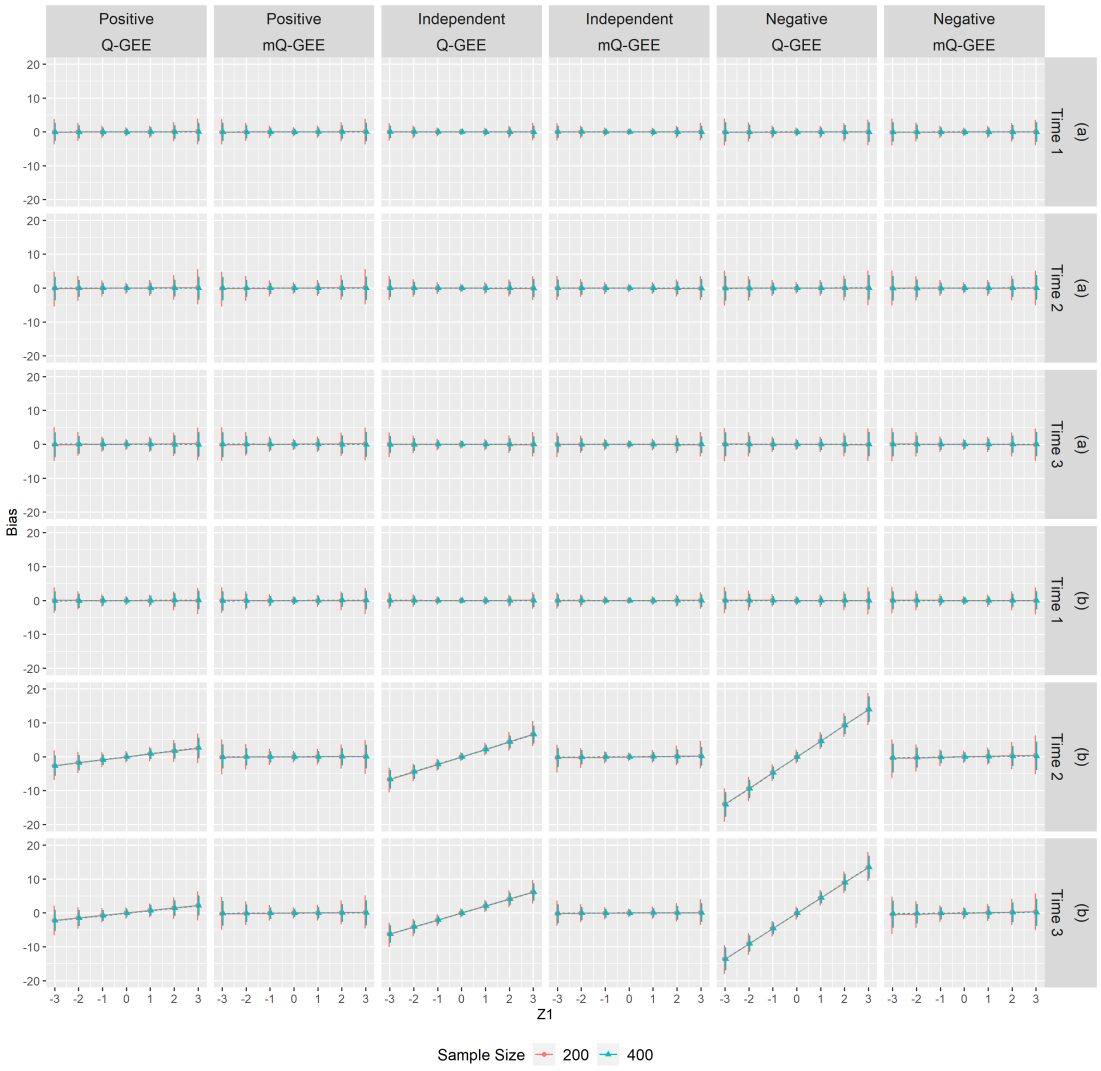

Figure 2.3: Bias (mean and 95% confidence interval) of heterogeneous causal effects across a grid of Z_1 for scenarios (a) when outcomes are correlated, and (b) when outcomes are correlated and the main effect model in the Q-function is misspecified, based on 1000 iterations.

2.4.4 Results for Unequally Weighted Outcomes

We run the simulation under a set of different weights. Figure 2.4 shows the result for the metric PCI_k, $k = 1, 2$. A significant difference in performance between Q-learning and modified Q-learning is observed for the negatively correlated outcomes, especially when more weights are assigned to stage 2 outcomes. Figure 2.5 shows the result for the metric $\text{RMSE}_{\tau_{1j}}$, $j = 1, 2, 3$. The findings are consistent with Section 2.4.3: across all weights considered, mQ-GEE has a universally better performance than Q-GEE for all scenarios. The sample size is set at $n = 200$.

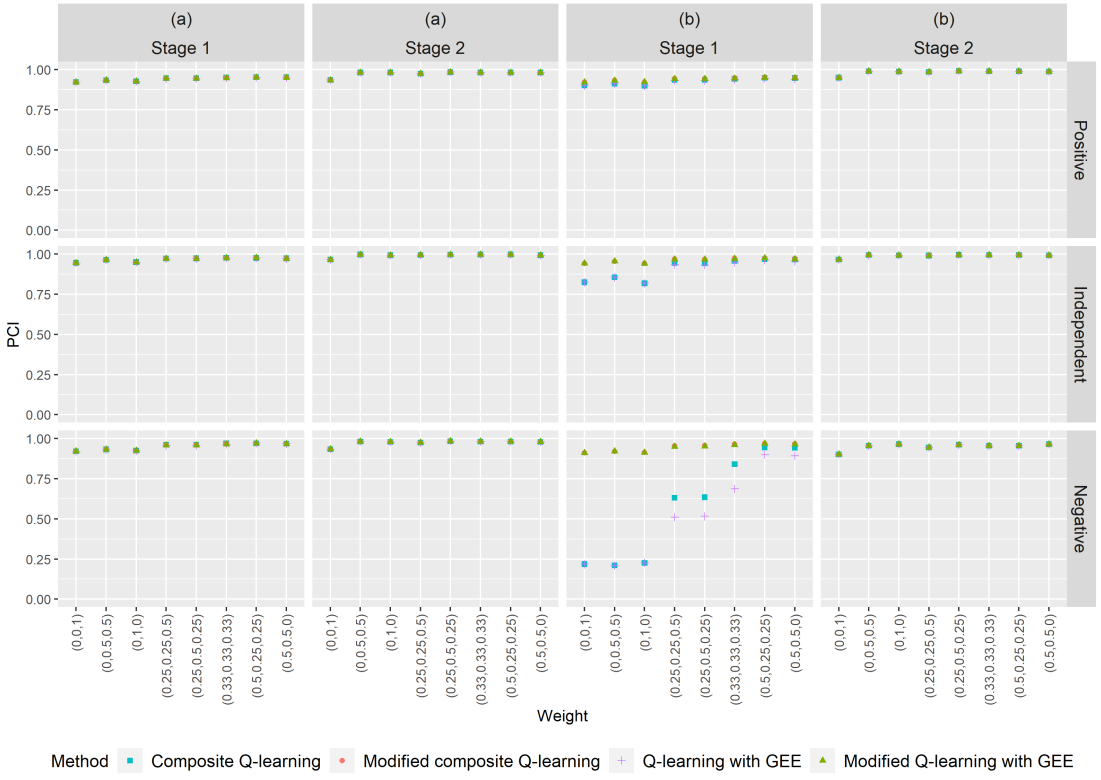

Figure 2.4: The probability of correctly identifying the optimal rules across a set of weights $(w_1, \boldsymbol{w}_2)$ for scenarios (a) when outcomes are correlated, and (b) when outcomes are correlated and the main effect model in stage 2 Q-function is misspecified, based on 1000 iterations.

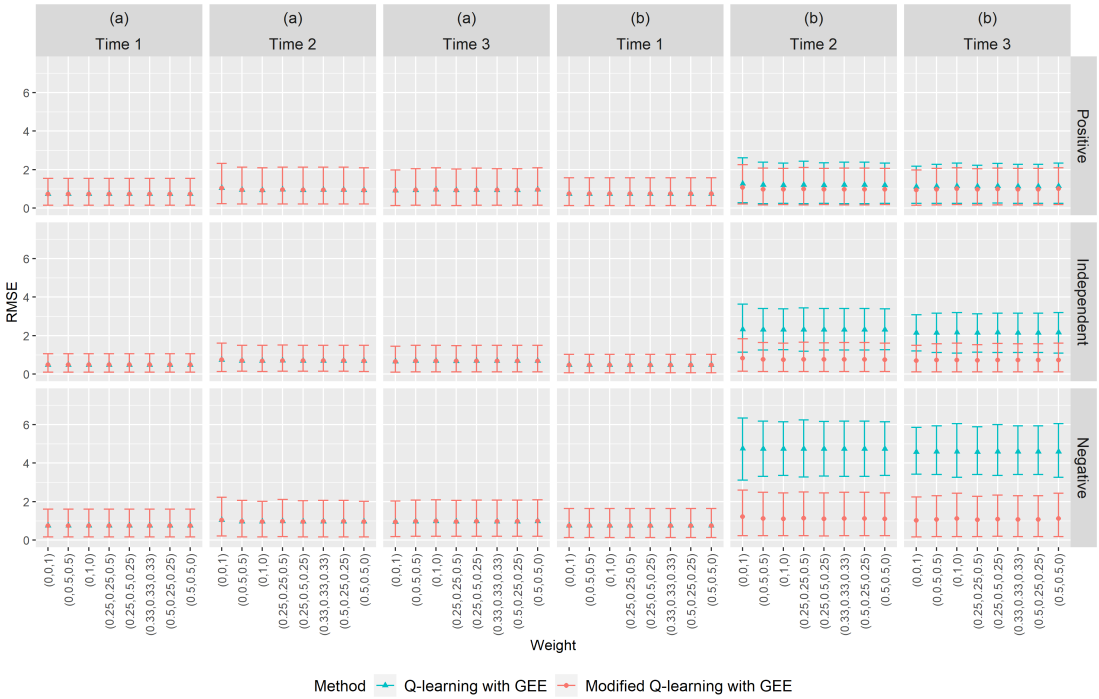

Figure 2.5: Root mean square error (mean and 95% confidence interval) of heterogeneous causal effects across a set of weights $(w_1, \boldsymbol{w}_2)$ for scenarios (a) when outcomes are correlated, and (b) when outcomes are correlated and the main effect model in stage 2 Q-function is misspecified, based on 1000 iterations.

2.5 Application

We use the M-bridge study ($n = 591$) to illustrate the application of mQ-GEE. Our goal is (1) to identify the optimal rules considering different weights for the repeated-measure outcomes, (2) to estimate the optimal trajectory and compare it with estimated trajectories of the four embedded DTRs, and (3) to estimate time-specific heterogeneous causal effects. The baseline variables considered are gender, race/ethnicity, pre-college intention to pledge to a sorority or fraternity [i.e., "Greek"], and baseline drinking habits. We are interested in the frequency of binge drinking during a 30-day period at baseline, self-monitoring (stage 1) and follow-up 1 and 2 (stage 2).

2.5.1 Identification of Personalized Optimal Rules

Applying mQ-GEE and assigning equal weights to each component of Y_1 and $\boldsymbol{Y}_2$, we estimate the individualized optimal decision rules that should be made at each stage for each college student. Variable selection based on QIC (Pan, 2001) is performed in this application to avoid over-fitting. The fitted stage 1 model selects gender and time-varying effects of intention to pledge to Greek and race as tailoring variables, whereas the fitted stage 2 model selects stage 1 outcome, baseline drinking habits, and time-varying effects of race and baseline outcome as tailoring variables (Appendix A.7). In summary, at stage 2, conditional on the covariate and treatment history, 58% of the heavy drinkers in this study would have benefited more from automated email ($A_2 = -1$) and 42% would have benefited more from online health coach ($A_2 = 1$). At stage 1, provided that participants follow the estimated optimal decision rules at stage 2, 22% of the participants would have benefited from late intervention ($A_1 = -1$) and 78% would have benefited from early intervention ($A_1 = 1$).

Table 2.3: Summary of estimated stage 1 and stage 2 optimal rules as a proportion of college students who were eligible for randomization at stage $k, k = 1, 2$, under different weights.

w_1	$\boldsymbol{w}_2$	Stage 1				Stage 2[a]	
		Q-GEE		mQ-GEE			
		$A_1 = -1$	$A_1 = 1$	$A_1 = -1$	$A_1 = 1$	$A_2 = -1$	$A_2 = 1$
0.33	(0.33, 0.33)	0.22	0.78	0.22	0.78	0.58	0.42
0	(1, 0)	0.67	0.33	0.40	0.60	0.61	0.39
0	(0.5, 0.5)	1	0	0.63	0.37	0.58	0.42
0	(0, 1)	1	0	1	0	0.59	0.41
0.25	(0.5, 0.25)	0.22	0.78	0.54	0.46	0.59	0.41
0.25	(0.25, 0.5)	0.24	0.76	0.63	0.37	0.59	0.41
0.5	(0.5, 0)	0.22	0.78	0.14	0.86	0.61	0.39
0.5	(0.25, 0.25)	0.22	0.78	0.22	0.78	0.58	0.42

[a]The percentage of stage 2 optimal rules is calculated based on the total number of heavy drinkers. Q-GEE and mQ-GEE generate the same results because the corresponding stage 2 models are the same.

We also explored the performance of Q-GEE and mQ-GEE under different weights. Table 2.3 shows a summary of the results. The assignment of stage 2 optimal rule is similar for both methods. However, it seems that mQ-GEE is more sensitive to the change in weights.

2.5.2 Point Estimation of the Optimal Trajectory and Heterogeneous Causal Effects

Figure 2.6 shows the estimates of expected repeated-measures outcomes under the optimal decision rule and specific embedded DTRs respectively. The optimal trajectory is estimated using mQ-GEE with equally weighted outcomes. The DTR-specific trajectories are estimated using Q-GEE with stagewise QIC-based variable selection by plugging in a set of deterministic values for $(A_1 = a_1, A_2 = a_2)$.

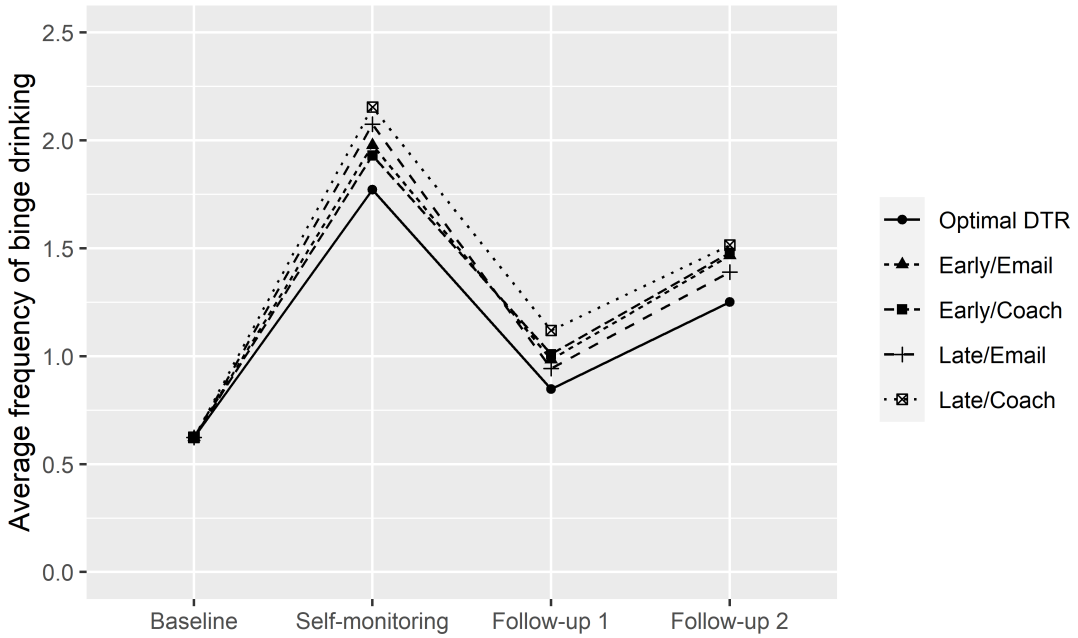

Figure 2.6: Estimated optimal trajectory versus marginal DTR-specific trajectories.

Figure 2.7 shows the empirical distributions of estimated heterogeneous treatment effects across all time points of measurement. For stage 1 intervention, it is not surprising to see that early intervention has more positive effects on early measurements (self-monitoring and follow-up 1) and late intervention has more positive effects on late measurement (follow-up 2). The result for stage 2 intervention is consistent with the investigator's findings (Patrick et al., in press) that the automated email was more effective in reducing frequency of binge drinking for heavy drinkers as compared to online health coach.

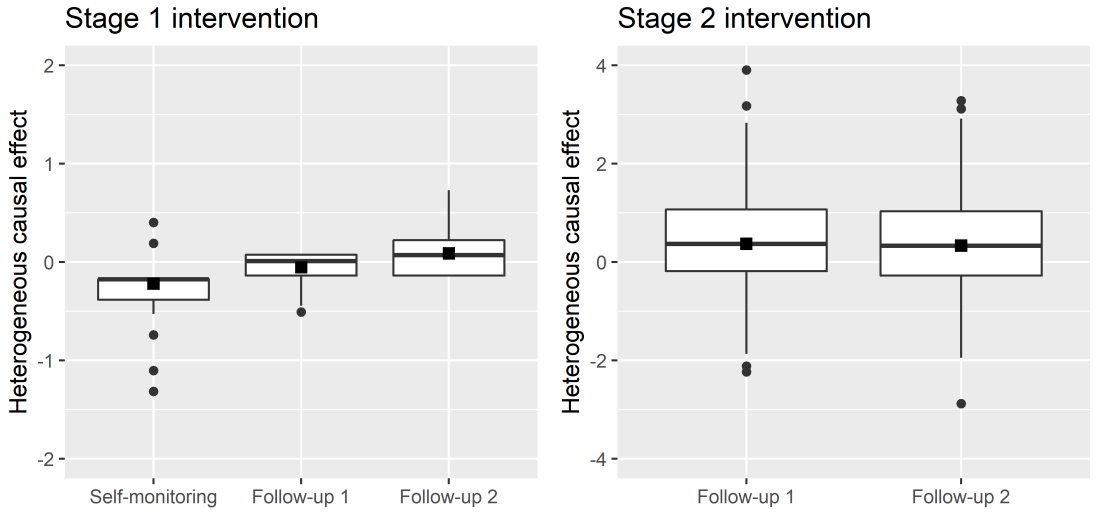

Figure 2.7: Distribution of estimated heterogeneous causal effects using modified Q-learning with GEE. The first panel is estimated based on stage 1 Q-function, with each boxplot representing the distribution of the contrast between Early and Late over time. Since negative treatment effects are preferred, an average treatment effect (indicated by a black square) below 0 shows that Early wins. The second panel is estimated based on stage 2 Q-function, with each boxplot representing the distribution of the contrast between Coach and Email over time. Similarly, an average treatment effect (indicated by a black square) below 0 shows that Coach wins.

2.6 Discussion

We developed an implementation of Q-learning on stagewise repeated-measures outcomes using generalized estimating equations. To apply the proposed method, investigators have the flexibility to choose science-driven weights for the repeated-measures outcomes. For implementing GEE, we recommend specifying unstructured working correlation, because in conventional SMARTs, participants are followed a considerably small number of times before the next stage of re-randomization or the end of study, and sample size is comparatively large. Furthermore, by using a saturated marginal model with time-varying coefficients, the impact of choosing a wrong working correlation is negligible. We also explored the performance of the proposed method in terms of rule identification and the consistency and efficiency of estimating heterogeneous causal effects. The performance of regression-based Q-learning depends heavily on correct specification of models at all stages. The modified version of Q-learning with GEE mitigates the problem by incorporating the residual from observed outcomes at subsequent stages into the pseudo-outcome at previous stages. This thesis focuses only on point estimation of the optimal trajectory. Inference under nonregularity conditions, where the weighted stage 2 treatment effect equals to 0, can be a complex problem that needs further research. Moreover, future work needs to be done to extend the proposed method for discrete outcomes, e.g., dichotomous responses which many investigators care about in clinical trials.

Chapter 3

On the Model Misspecification in Q-learning with Treatment Effect Heterogeneity

3.1 Literature Review

Q-learning with linear regression (Murphy, 2005b; Nahum-Shani et al., 2012) is a widely used backward induction algorithm to identify the optimal DTR due to its ease of implementation. At each stage of randomization, Q-learning specifies a *Q-function*, a parametric model of the expected pseudo-outcomes conditional on past history, where the pseudo-outcomes are the potential outcomes assuming that the optimal interventions are followed thereafter. Typically in practice, one specifies the Q-function as a linear model, which consists of a main effect model and a treatment effect model, and the model parameters are usually estimated using least squares estimation. The main effect component characterizes the variation in the outcome that is explained by pre-treatment covariates, whereas the treatment effect component characterizes the average effect of the observed treatment allowing for variation with pre-treatment covariates.

Decision making at a single stage, for example, using data from a randomized controlled trial, does not depend on the main effect model, as the treatment effect model fully defines the estimated optimal rule. However, this is not the case for backward induction over multiple stages. The performance of Q-learning is susceptible to model misspecification of the main effect model. Heterogeneous treatment effects at an earlier stage on a final outcome are in fact part of the main effects at later stages (either the earlier treatment-covariate interaction or an intermediate measurement which depends on prior treatement). It is reasonable to conjecture that omitting them at later stages would result in biased estimation of the treatment effects at early stages. This is an example of *informative residual bias* in optimizing over multiple stages.

Additionally, heterogeneous treatment effects at a later stage result in biased estimation if linear models are used in earlier stages as the optimization operation necessarily makes the linearity misspecified. Though the treatment effect model of a Q-function is assumed to be correctly specified with no unmeasured confounders, a nonlinear relationship between the predictors and the pseudo-outcome, which usually involves an absolute value function of the estimated treatment effects, arises inevitably when the estimated treatment effects at late stages are heterogeneous. This problem was studied comprehensively by Laber et al. (2014).

Both types of misspecification described above result in a nonnegligible bias in the prediction of stage 1 optimal rules, and have been addressed with carefully constructed methods. Existing methods that deal with residual bias are modified Q-learning (Huang et al., 2015), A-learning (Schulte et al., 2014) and robust Q-learning (Ertefaie et al., 2021). A-learning takes a propensity score approach and allows for flexible modeling of the main effects. Robust Q-learning as well takes a propensity score approach, but obviates the need to specify the main effect model. However, nonparametric methods are usually used to estimate the main effects in A-learning and the expected outcome in robust Q-learning. Nonparametric methods work ideally for nonlinearity between

outcomes and covariates, but are less straightforward when implementation and interpretation come into consideration. Moreover, model checking and residual diagnostics for Q-learning with linear regression can be easily performed using standard approaches (Henderson et al., 2010; Chakraborty and Moodie, 2013). Therefore, we advocate the use of modified Q-learning, a parametric approach that takes account of stage 2 residuals, for dealing with misspecification of the main effect model. For the treatment effect model, Laber et al. (2014) proposed an interactive model building of Q-learning to correct the bias caused by misspecified linearity between pseudo-outcome and predictors.

The above-mentioned methods work well for one specific type of misspecification, but fail to consider the coexistence of misspecifications as a result of heterogeneous treatment effects. This chapter starts with an introduction of the data structure and a further elaboration on the importance of the problem using the M-bridge study as an example. We then discuss the integrative impact of late-stage unadjusted residuals and early-stage nonlinearity on the prediction of optimal rules, with mathematical formulation and proof in Section 3.4 to help understand the statistical aspects of the problem. We then propose to build interactive models into modified parametric Q-learning with Murphy's regret function in Section 3.5. Simulations are performed to show the robustness of our proposed algorithm. Finally, we demonstrate its application on SMARTs with embedded tailoring using the M-bridge data.

3.2 Framework

3.2.1 Data Example

The data example we use to illustrate the problem omits the repeated-measures outcome in the M-bridge design. As shown in Figure 3.1, the outcomes of interest, binge drinking (primary outcome) and negative drinking-related consequences (secondary outcome), were measured at the end of Semester 1 for all students. The target outcomes for analysis in this section are the maximum number of drinks and the total number of

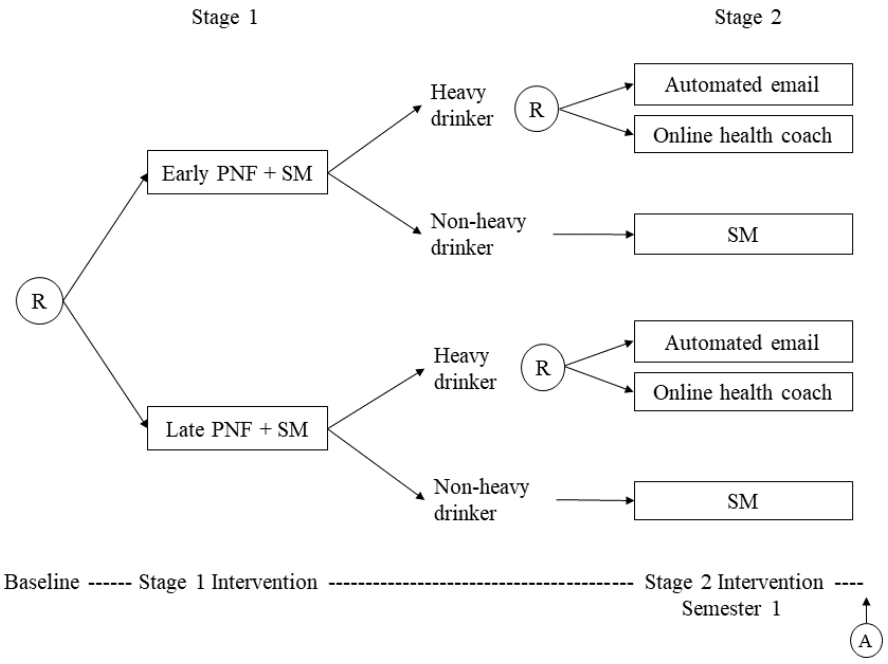

Figure 3.1: The M-bridge study design and data structure. (R) indicates a randomization stage with arrows pointing to available treatment options, and (A) indicates an assessment of outcomes. Note that this design only considers a final outcome measured at the end of the treatment course.

drinking-related consequences, both of which are continuous.

The scientific question is to predict, for each potential subject, the embedded DTR that optimizes two different final outcomes at the end of the treatment course: minimizes the maximum number of drinks consumed within a 24-hour period, or the total number of alcohol-related consequences in the past 30 days, which was measured using 24 items from the Brief Young Adult Alcohol Consequences Questionnaire (B-YAACQ). In other words, our aim is (1) to estimate the stage 1 intervention, i.e., if a student should receive PNF+SM early or late, conditional on his/her baseline characteristics, and then (2) to estimate stage 2 strategy, if the student should receive an automated email or online health coaching if he/she is a heavy drinker but continue self-monitoring if he/she is not a heavy drinker, so that the combined regime optimizes the potential outcome.

In the M-bridge study, the investigators hypothesize that pre-college alcohol use norms and pre-college intentions for college drinking are treatment effect moderators. These baseline variables might moderate both stage 1 and stage 2 treatment effects, and inappropriate adjustment is likely to occur in the Q-functions at both stages as a result of heterogeneous treatment effects. Details will be discussed in Section 3.4.

3.2.2 Data Structure

As in the M-bridge study, we assume a two-stage setting, although the following could be generalized to multiple decision points. Suppose that the data collected from a SMART are represented by a sequence of independently and identically distributed random variables (Z_1, A_1, Z_2, A_2, Y), where Z_1 is the set of baseline covariates and potential moderators measured prior to stage 1 randomization, Z_2 is the set of time-varying covariates and tailoring variables measured after stage 1 and before stage 2 randomization, $A_k \in \mathcal{A}_k$, $k = 1, 2$, is the treatment that the participant receives at stage k, with $\mathcal{A}_k$ being the set of all possible treatments, and Y is the outcome measured after treatment stages, with smaller values preferred.

In the M-bridge study, $\mathcal{A}_1 = \{-1, 1\}$, where $A_1 = 1$ represents early intervention and $A_1 = -1$ represents late intervention, and $\mathcal{A}_2 = \{-1, 1\}$, where $A_2 = 1$ represents online health coach and $A_2 = -1$ represents automated email. Z_1 includes baseline information on subject characteristics (gender, race, intention to pledge to a sorority or fraternity [i.e., "Greek"], and indicator whether parent has a significant drinking problem), pre-college drinking norms and intention for college drinking, and pre-college drinking habits. Z_2 includes solely the embedded tailoring variable, i.e., the flag of heavy drinkers. Y can be either the maximum number of drinks consumed within a 24-hour period or the total number of negative alcohol-related consequences in the past 30 days, and separate analyses should be done for each outcome. Let $H_1 = Z_1$ and $H_2 = (Z_1, A_1, Z_2)$ denote the covariate and treatment history up to the stage 1 and stage 2 randomization respectively.

3.2.3 Q-learning

We describe the algorithm of Q-learning with linear regression to answer the scientific question. Starting from stage 2, the Q-function is specified as

$$Q_2(H_2, A_2) = \mathbb{E}(Y|H_2, A_2) = \beta_{200} + \boldsymbol{X}_{20}^T\boldsymbol{\beta}_{201} + A_2\left(\beta_{210} + \boldsymbol{X}_{21}^T\boldsymbol{\beta}_{211}\right), \tag{3.1}$$

where $\boldsymbol{X}_{20}$ and $\boldsymbol{X}_{21}$ denote the vectors formed by elements in H_2 that represent the predictors in stage 2 main effect model and treatment effect model respectively. The estimators of the parameters, $\left(\hat{\beta}_{200} \quad \hat{\boldsymbol{\beta}}_2^T\right) = \left(\hat{\beta}_{200} \quad \hat{\boldsymbol{\beta}}_{201}^T \quad \hat{\beta}_{210} \quad \hat{\boldsymbol{\beta}}_{211}^T\right)$, are obtained using ordinary least squares.

For the preceding stage (i.e., stage 1), the Q-function is specified as

$$\begin{aligned} Q_1(H_1, A_1) &= \mathbb{E}\left\{ \min_{a_2 \in \mathcal{A}_2} Q_2(H_2, A_2 = a_2) \middle| H_1, A_1 \right\} \\ &= \beta_{100} + \boldsymbol{X}_{10}^T\boldsymbol{\beta}_{101} + A_1\left(\beta_{110} + \boldsymbol{X}_{11}^T\boldsymbol{\beta}_{111}\right), \end{aligned} \tag{3.2}$$

where $\boldsymbol{X}_{10}$ and $\boldsymbol{X}_{11}$ denote the vectors formed by elements in H_1 that represent the predictors in stage 1 main effect model and treatment effect model respectively. The estimators of the parameters, $\left(\hat{\beta}_{100} \quad \hat{\boldsymbol{\beta}}_1^T\right) = \left(\hat{\beta}_{100} \quad \hat{\boldsymbol{\beta}}_{101}^T \quad \hat{\beta}_{110} \quad \hat{\boldsymbol{\beta}}_{111}^T\right)$, are obtained using ordinary least squares.

The predicted optimal DTR is $\left(\hat{d}_1^{\text{opt}}, \hat{d}_2^{\text{opt}}\right)$, with

$$\hat{d}_j^{\text{opt}} = \arg\min_{a_j \in \mathcal{A}_j} Q_j\left(H_j, A_j = a_j; \hat{\beta}_{j00}, \hat{\boldsymbol{\beta}}_j\right) \text{ for } j = 1, 2. \tag{3.3}$$

3.3 Unmeasured Variables

3.3.1 Misspecification of Stage 2 Main Effects

To understand the problem thoroughly, we first focus on model misspecification in the presence of an important variable prior to stage 2 randomization which is not measured by design or omitted in the data analysis. Huang et al. (2015) argues that an unmeasured variable causes bias in the estimation of stage 1 optimal rules and model parameters. We briefly outline the argument here. Let $\boldsymbol{V}_{20}$ denote a vector of unmeasured covariates at stage 2 that are independent of A_2. We use uppercase to denote random variables and lowercase to denote a realization of the corresponding random variable.

Suppose Y conditional on all the covariates, measured or unmeasured, has mean

$$\mathbb{E}(Y|H_2, A_2, \boldsymbol{V}_{20}) = \psi_{200} + \boldsymbol{x}_{20}^T\boldsymbol{\psi}_{201} + \boldsymbol{V}_{20}^T\boldsymbol{\gamma}_{20} + a_2\left(\psi_{210} + \boldsymbol{x}_{21}^T\boldsymbol{\psi}_{211}\right) \tag{3.4}$$

and variance $\mathrm{Var}(Y|H_2, A_2, \boldsymbol{V}_{20}) = \sigma_2^2$.

3.3.2 Estimation Bias

The proof of the theorems can be found in the supplementary materials.

Theorem 3.3.1 (Matrix Version of the Omitted Variable Bias Theorem)**.** *Suppose that the true regression model for Y is $Y = \psi_0 + \boldsymbol{X}^T\boldsymbol{\psi}_1 + \boldsymbol{V}^T\boldsymbol{\gamma} + \varepsilon$, where $\boldsymbol{X}$ is a random vector formed by measured covariates, $\boldsymbol{V}$ is formed by unmeasured covariates, and $\varepsilon \sim N(0, \sigma^2)$. The parameters associated with measured covariates, $\boldsymbol{\psi} \equiv \begin{pmatrix} \psi_0 \\ \boldsymbol{\psi}_1 \end{pmatrix}$, are thus estimated via the misspecified model $y = \beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}_1 + \varepsilon^*$, where y and $\boldsymbol{x}$ are realizations of Y and $\boldsymbol{X}$ respectively. Then*

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) \equiv \mathbb{E}\begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_1 \end{pmatrix} = \boldsymbol{\psi} + \begin{pmatrix} \mathbb{E}(\boldsymbol{V}^T) - \mathbb{E}(\boldsymbol{X}^T)\,Cov(\boldsymbol{X})^{-1}\,Cov(\boldsymbol{X}, \boldsymbol{V}) \\ Cov(\boldsymbol{X})^{-1}\,Cov(\boldsymbol{X}, \boldsymbol{V}) \end{pmatrix}\boldsymbol{\gamma}.$$

Let $\boldsymbol{X}_2 = \begin{pmatrix} \boldsymbol{X}_{20}^T & A_2 & A_2\boldsymbol{X}_{21}^T \end{pmatrix}^T$ denote the full predictor vector and assume that the covariance matrix of $\boldsymbol{X}_2$ is invertible. By Theorem 3.3.1, $\mathbb{E}\begin{pmatrix} \hat{\beta}_{200} \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} = \begin{pmatrix} \psi_{200} \\ \boldsymbol{\psi}_2 \end{pmatrix} + \mathcal{B}\boldsymbol{\gamma}_{20}$, where

$$\mathcal{B} = \begin{pmatrix} \mathbb{E}(\boldsymbol{V}_{20}^T) - \mathbb{E}(\boldsymbol{X}_2^T)\mathrm{Cov}(\boldsymbol{X}_2)^{-1}\mathrm{Cov}(\boldsymbol{X}_2, \boldsymbol{V}_{20}) \\ \mathrm{Cov}(\boldsymbol{X}_2)^{-1}\mathrm{Cov}(\boldsymbol{X}_2, \boldsymbol{V}_{20}) \end{pmatrix}. \tag{3.5}$$

The first element of $\mathcal{B}$ corresponds to the bias associated with the intercept term and the second element of $\mathcal{B}$ corresponds to the bias associated with the covariate effects. The existence of bias in the estimation of covariate effects is then characterized by the term $\mathrm{Cov}(\boldsymbol{X}_2, \boldsymbol{V}_{20})$. The bias of $\hat{\boldsymbol{\beta}}_2$ can be rewritten as

$$\mathrm{Bias}(\hat{\boldsymbol{\beta}}_2) = \mathrm{Cov}(\boldsymbol{X}_2)^{-1} \begin{pmatrix} \mathrm{Cov}(\boldsymbol{X}_{20}, \boldsymbol{V}_{20}) \\ \mathrm{Cov}(A_2, \boldsymbol{V}_{20}) \\ \mathbb{E}(A_2)\mathrm{Cov}(\boldsymbol{X}_{21}, \boldsymbol{V}_{20}) \end{pmatrix} \boldsymbol{\gamma}_{20}, \tag{3.6}$$

and $\mathrm{Cov}(A_2, \boldsymbol{V}_{20}) = \boldsymbol{0}^T$ for SMARTs due to sequential randomization.

Theorem 3.3.2 (Bias of Stage 2 Treatment Effect Estimators)**.** *Assume that $\boldsymbol{V}_{20}$ is a vector of unmeasured covariates that are independent of A_2 and $Cov(\boldsymbol{X}_2)$ is invertible. The estimators of stage 2 heterogeneous treatment effects are unbiased if and only if at least one of the following conditions is satisfied:*

- $\mathbb{E}(A_2) = 0$*;*
- $\boldsymbol{V}_{20}$ *is correlated with neither $\boldsymbol{X}_{20}$ nor $\boldsymbol{X}_{21}$.*

Theorem 3.3.3 (Bias of Stage 2 Main Effect Estimators)**.** *Assume that $\boldsymbol{V}_{20}$ is a vector of unmeasured covariates that are independent of A_2 and $\mathbb{E}(A_2) = 0$. Suppose that $\boldsymbol{V}_{20}$ is correlated with $\boldsymbol{X}_{20}$ and $Cov(\boldsymbol{X}_2)$ is invertible. Then the estimators of stage 2 main*

effects are biased and the bias is $\mathcal{B}'\boldsymbol{\gamma}_{20}$*, where*

$$\mathcal{B}' = \begin{pmatrix} \mathbb{E}(\boldsymbol{V}_{20}^T) - \mathbb{E}(\boldsymbol{X}_2^T)\,Cov(\boldsymbol{X}_2)^{-1}\,Cov(\boldsymbol{X}_2, \boldsymbol{V}_{20}) \\ Cov(\boldsymbol{X}_{20})^{-1}\,Cov(\boldsymbol{X}_{20}, \boldsymbol{V}_{20}) \end{pmatrix}. \tag{3.7}$$

Theorem 3.3.2 shows the importance of balancing sample size in the randomization arms. With unbalanced designs, it is possible to bias the estimation of stage 2 treatment effects. However, this is not the case we consider in this thesis. In the M-bridge study, heavy-drinkers were re-randomized to $A_2 = 1$ and $A_2 = -1$ with equal probabilities, i.e., $\mathbb{E}(A_2) = 0$, so no bias would be induced in the identification of stage 2 optimal rules. Theorem 3.3.3 shows that the estimators of stage 2 main effects, however, can be biased if the unmeasured variable $\boldsymbol{V}_{20}$ is correlated with $\boldsymbol{X}_{20}$.

3.3.3 A Special Case: Omission of Stage 1 Heterogeneous Treatment Effects

Now we understand how omission of an unmeasured variable causes bias in the estimation in the main effect and treatment effect models. As a special case, omitted stage 1 heterogeneous treatment effects in the stage 2 main effect model may similarly cause a loss in the power to correctly predict stage 1 optimal rules. Investigators usually do not care much about the adjustment for stage 1 treatment effects in the stage 2 model as only the (heterogeneous) treatment effects of stage 2 intervention would determine the stage 2 optimal rules. Thus, Q-learning with linear regression is often implemented using a linear predictor function such that the same design matrix is used for both the main effect and treatment effect models. Moreover, three way interactions are rarely included in the stage 2 treatment model, so $\boldsymbol{X}_{21}$ often does not include interactions between A_1 and baseline covariates. However, doing so may result in a bias for stage 1 estimation.

As a special case of Equation (3.4), suppose Y follows an independent and identical

distribution with mean

$$\mathbb{E}(Y|H_2, A_2) = \psi_{200} + \boldsymbol{x}_{20}^T\boldsymbol{\psi}_{201} + a_1\boldsymbol{x}_{11}^T\boldsymbol{\gamma}_{20} + a_2\left(\psi_{210} + \boldsymbol{x}_{21}^T\boldsymbol{\psi}_{211}\right), \tag{3.8}$$

where $\boldsymbol{X}_{20} = \boldsymbol{X}_{21} = \begin{pmatrix}\boldsymbol{X}_{10}^T & A_1 & Z_2\end{pmatrix}^T$ does not include the interaction between baseline covariates and stage 1 treatment. Assume that $\mathbb{E}(A_2) = 0$. Substituting $\boldsymbol{X}_2 = \begin{pmatrix}\boldsymbol{X}_{20}^T & A_2 & \boldsymbol{X}_{21}^T\end{pmatrix}^T$ and the unmeasured variables $A_1\boldsymbol{X}_{11}$ in Equation (3.7), the bias of stage 2 main effect estimators is

$$\mathcal{B}_s' = \begin{pmatrix} \mathbb{E}(A_1\boldsymbol{X}_{11}^T) - \mathbb{E}(\boldsymbol{X}_2^T)\mathrm{Cov}(\boldsymbol{X}_2)^{-1}\mathrm{Cov}(\boldsymbol{X}_2, A_1\boldsymbol{X}_{11}) \\ \mathrm{Cov}(\boldsymbol{X}_{20})^{-1}\begin{pmatrix}\mathbb{E}(A_1)\mathrm{Cov}(\boldsymbol{X}_{10}, \boldsymbol{X}_{11}) \\ \mathrm{Var}(A_1)\mathbb{E}(\boldsymbol{X}_{11}^T) \\ \mathrm{Cov}(Z_2, A_1\boldsymbol{X}_{11})\end{pmatrix} \end{pmatrix}. \tag{3.9}$$

Even if $\mathbb{E}(A_1) = 0$, $\mathcal{B}_s'$ is a nonzero vector.

Thus, we have shown that wrongly omitting stage 1 heterogeneous treatment effects results in biased estimation of stage 2 main effects. This is the case we would like to address in this chapter. Furthermore, investigators should be alert to this issue in the use of Q-learning software. `qLearn` (Xin et al., 2012) allows for different and explicit specifications of the main effect and treatment effect model, but it is less straightforward in `qlaci` (Ertefaie et al., 2014) and `iqLearn` (Linn et al., 2015).

3.4 Model Misspecification with Treatment Effect Heterogeneity

In this section, the integrated impact of the two sources of model misspecification caused by heterogeneous treatment effects is formulated mathematically and discussed in detail. Following the setting in Section 3.3.3, suppose Y follows an independent and identical

distribution with mean

$$\mathbb{E}(Y|H_2, A_2) = \psi_{200} + \boldsymbol{x}_{20}^T\boldsymbol{\psi}_{201} + a_1\boldsymbol{x}_{11}^T\boldsymbol{\gamma}_{20} + a_2\left(\psi_{210} + \boldsymbol{x}_{21}^T\boldsymbol{\psi}_{211}\right),$$

where $\boldsymbol{X}_{20} = \boldsymbol{X}_{21}$ does not include the interaction between baseline covariates and stage 1 treatment. Assume $\mathbb{E}(A_1) = \mathbb{E}(A_2) = 0$. In the backward induction setting, stage 1 optimization is contingent on compliance with the optimal rule at stage 2, so the true optimal pseudo-outcome at stage 1 is

$$Y^{\text{opt}} = \mathbb{E}\left(Y \left| H_2, A_2 = d_2^{\text{opt}}\right.\right) = \psi_{200} + \boldsymbol{x}_{20}^T\boldsymbol{\psi}_{201} + a_1\boldsymbol{x}_{11}^T\boldsymbol{\gamma}_{20} - \left|\psi_{210} + \boldsymbol{x}_{21}^T\boldsymbol{\psi}_{211}\right|,$$

where d_2^{opt} is the optimal decision rule at stage 2 and

$$d_2^{\text{opt}}(H_2) = -\text{sgn}\left\{\psi_{210} + \boldsymbol{x}_{21}^T\boldsymbol{\psi}_{211}\right\}.$$

For the sake of notation convention, we define $\tilde{\boldsymbol{X}}_{20} = \begin{pmatrix}1 & \boldsymbol{X}_{20}^T\end{pmatrix}^T$, $\tilde{\boldsymbol{X}}_{21} = \begin{pmatrix}1 & \boldsymbol{X}_{21}^T\end{pmatrix}^T$, $\boldsymbol{\psi}_{20} = \begin{pmatrix}\psi_{200} & \boldsymbol{\psi}_{201}^T\end{pmatrix}^T$, and $\boldsymbol{\psi}_{21} = \begin{pmatrix}\psi_{210} & \boldsymbol{\psi}_{211}^T\end{pmatrix}^T$, and rewrite Y^{opt} as

$$Y^{\text{opt}} \equiv \tilde{\boldsymbol{x}}_{20}^T\boldsymbol{\psi}_{20} + a_1\boldsymbol{x}_{11}^T\boldsymbol{\gamma}_{20} - \left|\tilde{\boldsymbol{x}}_{21}^T\boldsymbol{\psi}_{21}\right|. \tag{3.10}$$

$-\left|\tilde{\boldsymbol{X}}_{21}^T\boldsymbol{\psi}_{21}\right|$ represents the stage 2 optimal treatment effect, and is a non-smooth function of $\boldsymbol{\psi}_{21}$. Nonregularity of stage 1 parameters due to the non-smooth function $-\left|\tilde{\boldsymbol{X}}_{21}^T\boldsymbol{\psi}_{21}\right|$ has been extensively studied in literature (Robins, 2004; Chakraborty et al., 2010). In order to satisfy the regularity conditions for statistical inference, we assume

$$P\left\{H_2 : \tilde{\boldsymbol{X}}_{21}^T\boldsymbol{\psi}_{21} = 0\right\} = 0. \tag{3.11}$$

The estimator of Y^{opt} is

$$\hat{Y}^{\text{opt}} = \min_{a_2} Q_2\left(\tilde{\boldsymbol{x}}_2, a_2; \hat{\beta}_{200}, \hat{\boldsymbol{\beta}}_2\right) = \tilde{\boldsymbol{x}}_{20}^T\hat{\boldsymbol{\beta}}_{20} - \left|\tilde{\boldsymbol{x}}_{21}^T\hat{\boldsymbol{\beta}}_{21}\right|, \tag{3.12}$$

where $\hat{\boldsymbol{\beta}}_{20} = \left(\hat{\beta}_{200} \quad \hat{\boldsymbol{\beta}}_{201}^T\right)^T$ and $\hat{\boldsymbol{\beta}}_{21} = \left(\hat{\beta}_{210} \quad \hat{\boldsymbol{\beta}}_{211}^T\right)^T$. For large samples, $\hat{\boldsymbol{\beta}}_{21}$ is a consistent estimator of $\boldsymbol{\psi}_{21}$. Under assumption (3.11), $\left|\tilde{\boldsymbol{x}}_{21}^T\hat{\boldsymbol{\beta}}_{21}\right|$ is also a consistent estimator of $\left|\tilde{\boldsymbol{x}}_{21}^T\boldsymbol{\psi}_{21}\right|$ by the continuous mapping theorem. However, $\left|\tilde{\boldsymbol{x}}_{21}^T\hat{\boldsymbol{\beta}}_{21}\right|$ is a biased estimator of $\left|\tilde{\boldsymbol{x}}_{21}^T\boldsymbol{\psi}_{21}\right|$ as $\left|\tilde{\boldsymbol{x}}_{21}^T\boldsymbol{\psi}_{21}\right| \neq 0$.

Normality and linearity are usually assumed for the conditional distribution of stage 2 effects on stage 1 covariates, but in stage 1 estimation, bias can still be induced by the absolute value function $\left|\tilde{\boldsymbol{x}}_{21}^T\boldsymbol{\psi}_{21}\right|$. Q-learning requires the causal assumption of no unmeasured confounders to be satisfied in order to obtain unbiased estimators of treatment effects, i.e., the treatment model at each stage is correctly specified. Thus, the bias discussed here does not result from misspecification of the treatment model, but intrinsically from the misspecified linear relationship between the pseudo-outcome and stage 1 covariates, as a result of the optimization operation. The detailed proof of the nonlinear relationship can be found in the supplementary materials of *Interactive Model Building for Q-learning* (Laber et al., 2014), and the main idea is built upon Theorem 3.4.1.

Theorem 3.4.1. *Suppose* $X \sim N(\mu, \sigma^2)$. *Then*

$$\mathbb{E}(|X|) = \left(\frac{2\sigma^2}{\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \mu\left\{1 - 2\Phi\left(-\frac{\mu}{\sigma}\right)\right\}.$$

Now we derive an expression for stage 1 bias from model misspecification associated with heterogeneous treatment effects at both stages. Rewrite $Q_1(H_1, A_1) = \beta_{100} + \boldsymbol{X}_{10}^T\boldsymbol{\beta}_{101} + a_1\left(\beta_{110} + \boldsymbol{X}_{11}^T\boldsymbol{\beta}_{111}\right) \equiv \tilde{\boldsymbol{X}}_{10}^T\boldsymbol{\beta}_{10} + a_1\tilde{\boldsymbol{X}}_{11}^T\boldsymbol{\beta}_{11}$. The values of the stage 1 parameters can be obtained as

$$(\boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{11}) = \underset{\boldsymbol{\beta}_{10},\boldsymbol{\beta}_{11}}{\arg\min}\, \mathbb{E}\left[\left\{\tilde{\boldsymbol{X}}_{20}^T\boldsymbol{\beta}_{20} - \left|\tilde{\boldsymbol{X}}_{21}^T\boldsymbol{\beta}_{21}\right| - \tilde{\boldsymbol{X}}_{10}^T\boldsymbol{\beta}_{10} - A_1\tilde{\boldsymbol{X}}_{11}^T\boldsymbol{\beta}_{11}\right\}^2\right].$$

Let $\hat{\boldsymbol{\beta}}_{10}$ and $\hat{\boldsymbol{\beta}}_{11}$ be the corresponding ordinary least squares estimators. Suppose

$\tilde{\boldsymbol{X}}_1 = \begin{pmatrix} \tilde{\boldsymbol{X}}_{10}^T & A_1 \tilde{\boldsymbol{X}}_{11}^T \end{pmatrix}^T$ is the vector of predictors, $\tilde{\mathbf{X}}_1 = \begin{pmatrix} \tilde{\mathbf{X}}_{10} & \mathbf{A}_1 \tilde{\mathbf{X}}_{11} \end{pmatrix}$ is the design matrix for stage 1 estimation and is of full column rank, and $\boldsymbol{Y}^{\text{opt}}$ is the outcome vector with Y_i^{opt} as the ith element, $i = 1, \ldots, n$. Then the estimators of stage 1 model parameters are

$$\begin{aligned} \begin{pmatrix} \hat{\boldsymbol{\beta}}_{10} \\ \hat{\boldsymbol{\beta}}_{11} \end{pmatrix} &= \left(\tilde{\mathbf{X}}_1^T \tilde{\mathbf{X}}_1 \right)^{-1} \tilde{\mathbf{X}}_1^T \hat{\boldsymbol{Y}}^{\text{opt}} \\ &= \left(\tilde{\mathbf{X}}_1^T \tilde{\mathbf{X}}_1 \right)^{-1} \tilde{\mathbf{X}}_1^T \left(\tilde{\mathbf{X}}_{20} \hat{\boldsymbol{\beta}}_{20} - \left| \tilde{\mathbf{X}}_{21} \hat{\boldsymbol{\beta}}_{21} \right| \right), \end{aligned}$$

and the bias of stage 1 estimation is

$$\begin{aligned} & \mathbb{E}\left(\tilde{\boldsymbol{x}}_{10}^T \hat{\boldsymbol{\beta}}_{10} + a_1 \tilde{\boldsymbol{x}}_{11}^T \hat{\boldsymbol{\beta}}_{11} \right) - \mathbb{E}(Y^{\text{opt}} \big| \, \tilde{\boldsymbol{x}}_1, a_1) \\ = \; & \tilde{\boldsymbol{x}}_1^T \mathbb{E}\left\{ \left(\tilde{\mathbf{X}}_1^T \tilde{\mathbf{X}}_1 \right)^{-1} \tilde{\mathbf{X}}_1^T \left(\tilde{\mathbf{X}}_{20} \hat{\boldsymbol{\beta}}_{20} - \left| \tilde{\mathbf{X}}_{21} \hat{\boldsymbol{\beta}}_{21} \right| \right) \right\} \\ - \; & \mathbb{E}\left(\tilde{\boldsymbol{x}}_{20}^T \boldsymbol{\psi}_{20} + a_1 \boldsymbol{x}_{11}^T \boldsymbol{\gamma}_{20} - \left| \tilde{\boldsymbol{x}}_{21}^T \boldsymbol{\psi}_{21} \right| \big| \, H_1, A_1 \right) \\ = \; & \left[\tilde{\boldsymbol{x}}_1^T \mathbb{E}\left\{ \left(\tilde{\mathbf{X}}_1^T \tilde{\mathbf{X}}_1 \right)^{-1} \tilde{\mathbf{X}}_1^T \mathcal{B}_s' \right\} - a_1 \boldsymbol{x}_{11}^T \right] \boldsymbol{\gamma}_{20} \qquad \text{(A)} \\ + \; & \left[\mathbb{E}\left(\left| \tilde{\boldsymbol{x}}_{21}^T \boldsymbol{\psi}_{21} \right| \big| \, \tilde{\boldsymbol{x}}_1, a_1 \right) - \tilde{\boldsymbol{x}}_1^T \mathbb{E}\left\{ \left(\tilde{\mathbf{X}}_1^T \tilde{\mathbf{X}}_1 \right)^{-1} \tilde{\mathbf{X}}_1^T \mathbb{E}\left(\left| \tilde{\mathbf{X}}_{21} \hat{\boldsymbol{\beta}}_{21} \right| \Big| \, H_1, A_1 \right) \right\} \right], \qquad \text{(B)} \end{aligned}$$

where $\mathcal{B}_s'$ is the expression in Equation (3.9). Bias (A) is induced by the omission of stage 1 heterogeneous treatment effects in the stage 2 main effect model, and bias (B) is induced by the false assumption of linearity between the absolute value of stage 2 heterogeneous treatment effects and stage 1 predictors.

3.5 The Proposed Method

3.5.1 Modified Interactive Q-learning

Interactive Q-learning (Laber et al., 2014) was proposed to address the misspecified linearity in stage 1 estimation by separately regressing stage 2 main effects on stage 1

predictors and estimating the conditional distribution of stage 2 treatment effects conditional on stage 1 predictors, and then combining the former and the expected absolute value of the latter to get the estimated stage 1 Q-function. To address both types of bias concomitantly, our proposed method follows the virtue of interactive Q-learning and modifies the main effect portion of the algorithm to account for any informative residuals from stage 2 estimation.

The modified interactive Q-learning (mIQ) algorithm comprises the following steps:

(mIQ-1) Regress Y on H_2, A_2 based on stage 2 Q-function

$$Q_2\left(H_2, A_2; \boldsymbol{\beta}_{20}, \boldsymbol{\beta}_{21}\right) = \tilde{\boldsymbol{X}}_{20}^T \boldsymbol{\beta}_{20} + A_2 \tilde{\boldsymbol{X}}_{21}^T \boldsymbol{\beta}_{21}$$

to obtain the ordinary least squares estimators $\hat{\boldsymbol{\beta}}_{20}, \hat{\boldsymbol{\beta}}_{21}$;

(mIQ-2) Predict the stage 2 optimal rule for a subject with characteristics $\tilde{\boldsymbol{x}}_{21}$:

$$\hat{d}_2^{\text{opt}} = -\text{sgn}\left\{\tilde{\boldsymbol{x}}_{21}^T \hat{\boldsymbol{\beta}}_{21}\right\};$$

(mIQ-3) Regress $Y - a_2 \tilde{\boldsymbol{x}}_{21}^T \hat{\boldsymbol{\beta}}_{21}$ on H_1, A_1 to obtain the estimator of

$$\mathbb{E}\left(Y - A_2 \tilde{\boldsymbol{X}}_{21}^T \boldsymbol{\psi}_{21} \middle| H_1, A_1\right),$$

denoted by $\hat{m}(H_1, A_1)$;

(mIQ-4) Estimate the conditional distribution $g\left(\tilde{\boldsymbol{X}}_{21}^T \boldsymbol{\psi}_{21} \middle| H_1, A_1\right)$, denoted by $\hat{g}(\cdot|H_1, A_1)$: If g is a conditional normal density with constant variance, i.e.,

$$\tilde{\boldsymbol{X}}_{21}^T \boldsymbol{\psi}_{21} \middle| H_1, A_1 \sim N(\mu(H_1, A_1), \sigma^2),$$

then regress $\tilde{\boldsymbol{x}}_{21}^T \hat{\boldsymbol{\beta}}_{21}$ on H_1, A_1 to obtain the estimators $\hat{\mu}(H_1, A_1)$ and $\hat{\sigma}$;

(mIQ-5) Obtain the estimator of $\mathbb{E}\left(Y^{\text{opt}}\middle|\, H_1, A_1\right) = \mathbb{E}\left(\tilde{\boldsymbol{X}}_{20}^T\boldsymbol{\psi}_{20} - \left|\tilde{\boldsymbol{X}}_{21}^T\boldsymbol{\psi}_{21}\right|\middle|\, H_1, A_1\right)$ by combining the above estimators:

$$\hat{Q}_1(H_1, A_1) = \hat{m}(H_1, A_1) - \int |z|\hat{g}(z|H_1, A_1)dz,$$

where the integral can be easily calculated for a location-scale distribution g using Theorem 3.4.1;

(mIQ-6) Predict the stage 1 optimal rule for a subject with baseline covariates h_1:

$$\hat{d}_1^{\text{opt}} = \underset{a_1\in\{-1,1\}}{\arg\min}\ \hat{Q}_1(H_1 = h_1, A_1 = a_1).$$

Step **(mIQ-3)** incorporates any stage 2 residual remainder from misspecification of the main effect model. As $\hat{\boldsymbol{\beta}}_{21}$ is a consistent estimator of $\boldsymbol{\psi}_{21}$, $\hat{m}(H_1, A_1)$ is a consistent estimator of $\mathbb{E}\left(Y - A_2\tilde{\boldsymbol{X}}_{21}^T\boldsymbol{\psi}_{21}\middle|\, H_1, A_1\right)$. The assumption of g in Step **(mIQ-4)** can be loosened and empirical methods can be applied.

3.5.2 Small Sample Properties of the Proposed Estimator

The pseudo-outcome in the stage 1 estimation, $Y^{\text{opt}} = \mathbb{E}\left(Y\middle|H_2, A_2 = d_2^{\text{opt}}(H_2)\right)$ can be a counterfactual outcome. If $d_2^{\text{opt}} = a_2$, then the expression represents the expected observed outcome at stage 2; if $d_2^{\text{opt}} \neq a_2$, then it represents the expected counterfactual outcome. With the observation that

$$\begin{aligned} Y^{\text{opt}} &= \mathbb{E}(Y|H_2, A_2 = a_2) - 2\mathbb{1}\left\{d_2^{\text{opt}} \neq a_2\right\}\left|\tilde{\boldsymbol{x}}_{21}^T\boldsymbol{\psi}_{21}\right| \\ &= \mathbb{E}(Y|H_2, A_2 = a_2) + \left(d_2^{\text{opt}} - a_2\right)\tilde{\boldsymbol{x}}_{21}^T\boldsymbol{\psi}_{21}, \end{aligned}$$

the estimator of stage 1 Q-function, $\hat{Q}_1(H_1, A_1)$, has a bias of

$$\begin{aligned}
\text{Bias}\left(\hat{Q}_1\right) &= \mathbb{E}\left\{\tilde{\boldsymbol{X}}_1^T\left(\tilde{\boldsymbol{X}}_1\tilde{\boldsymbol{X}}_1^T\right)^{-1}\tilde{\boldsymbol{X}}_1\left(Y - A_2\tilde{\boldsymbol{X}}_{21}^T\hat{\boldsymbol{\beta}}_{21}\right)\right\} - \mathbb{E}\left\{\mathbb{E}_{\hat{g}}\left(\left|\tilde{\boldsymbol{X}}_{21}^T\boldsymbol{\psi}_{21}\right|\middle| H_1, A_1\right)\right\} \\
&\quad - \mathbb{E}\left\{Y + \left(d_2^{\text{opt}} - A_2\right)\tilde{\boldsymbol{X}}_{21}^T\boldsymbol{\psi}_{21}\middle| H_1, A_1\right\} \\
&= \mathbb{E}\left\{\tilde{\boldsymbol{X}}_1^T\left(\tilde{\boldsymbol{X}}_1\tilde{\boldsymbol{X}}_1^T\right)^{-1}\tilde{\boldsymbol{X}}_1 Y\right\} - \mathbb{E}(Y|H_1, A_1) \qquad (3.13) \\
&\quad - \left[\mathbb{E}\left\{\mathbb{E}_{\hat{g}}\left(\left|\tilde{\boldsymbol{X}}_{21}^T\boldsymbol{\psi}_{21}\right|\middle| H_1, A_1\right)\right\} - \mathbb{E}\left(\left|\tilde{\boldsymbol{X}}_{21}^T\boldsymbol{\psi}_{21}\right|\middle| H_1, A_1\right)\right] \qquad (3.14)
\end{aligned}$$

Therefore, unbiased estimation of stage 1 Q-function requires that (1) stage 1 model is correctly specified so that the linearity between Y and stage 1 predictors is valid (shown by Formula (3.13)), and (2) the assumption of normality of the underlying distribution g is true so that the conditional distribution of stage 2 treatment effects on stage 1 predictors is consistently estimated (shown by Formula (3.14)). Laber et al. (2014) proposed additional nonparametric modeling of g using empirical methods, which helps to increase the modeling flexibility of this algorithm.

3.6 Simulation

3.6.1 Preliminaries

We conduct a simulation study to show the *predictive* performance of the proposed algorithm in the context of small samples. We start the discussion with a description of the data generative mechanism. Assume a sequence of observations from a SMART study is $(Z_{1i}, A_{1i}, Z_{2i}, A_{2i}, Y_i)$, $i = 1, \ldots, n$, where $Z_{1i} \overset{\text{i.i.d.}}{\sim} N(-2, 1)$, $Z_{2i} = Z_{1i} + \phi_i$ and $\phi_i \overset{\text{i.i.d.}}{\sim} N(0, 4)$, A_{1i} and A_{2i} both follow an i.i.d. Rademacher distribution, so $\frac{A_{1i}+1}{2} \overset{\text{i.i.d.}}{\sim}$ Bernoulli(0.5) and $\frac{A_{2i}+1}{2} \overset{\text{i.i.d.}}{\sim}$ Bernoulli(0.5). There is no embedded tailoring variable in this design. Suppose

$$Y_i = \tilde{\boldsymbol{X}}_{20,i}^T\boldsymbol{\psi}_{20} + c_2 A_{2i}\tilde{\boldsymbol{X}}_{21,i}^T\boldsymbol{\psi}_{21} + \varepsilon_i,$$

where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0,1)$ and small values of Y_i are preferred. The preliminary study considers three scenarios for the stage 2 main effects and two scenarios for the stage 2 treatment effects:

Stage 1 treatment effects $\tilde{\boldsymbol{X}}_{20,i}^T\boldsymbol{\psi}_{20} = 3 - Z_{1i} + 0.1A_{1i} - 0.1Z_{2i} + c_1V_i$	**Stage 2 treatment effects** $\tilde{\boldsymbol{X}}_{21,i}^T\boldsymbol{\psi}_{21} = -6 + \alpha_1 Z_{1i} + 5A_{1i} + \alpha_2 Z_{2i}$
Homogeneous	**Homogeneous**
• V_i uncorrelated with H_{2i}: $V_i \sim N(-1,1)$	• $\alpha_1 = 0,\ \alpha_2 = 0$
• V_i correlated with H_{2i}: $V_i \sim N(2Z_{1i}Z_{2i}, 1)$	
Heterogeneous	**Heterogeneous**
• $V_i = Z_{1i}A_{1i}$	• $\alpha_1 = -4,\ \alpha_2 = -0.2$

Table 3.1: Specifications of stage 2 main effect and treatment effect model in the data generative mechanism.

For a data generative mechanism that utilizes one specification of both $\tilde{\boldsymbol{X}}_{20,i}^T\boldsymbol{\psi}_{20}$ and $\tilde{\boldsymbol{X}}_{21,i}^T\boldsymbol{\psi}_{21}$, we generate training datasets of sample size n, and a test dataset of N subjects with potential outcomes under the four treatment regimes for each subject. Hence, the true optimal rules, d_1^{opt} and d_2^{opt}, are known for each subject in the test dataset by minimizing the potential outcomes over all treatment regimes. The training datasets are then used to estimate the linear models specified at each stage, and the estimated models are used to predict the optimal rules $\hat{d}_1^{\text{opt}}$ and $\hat{d}_2^{\text{opt}}\left(d_1^{\text{opt}}\right)$ for each subject in the test dataset. In our framework, Z_2 is defined as the set of stage 2 time-varying covariates and tailoring variables. Note that in this setting, Z_2 does not vary with A_1 in predicting $\hat{d}_2^{\text{opt}}$. Thus, we do not need to worry about the change in Z_2 with respect to $\hat{d}_1^{\text{opt}}$ which in turn affects the prediction of $\hat{d}_2^{\text{opt}}\left(d_1^{\text{opt}}\right)$.

Preliminary results (Table 3.2) show that *omission of stage 1 heterogeneous treatment effects in stage 2 main effect model causes significant bias in stage 1 rule identification.* Omission of a variable that is uncorrelated with other stage 2 main predictors does not cause bias, but if the omitted variable is correlated with other main predictors, then bias is generated. Interactive Q-learning indeed does not tackle this problem and may

have a slightly worse performance than standard Q-learning in this scenario. Onwards, we focus on the performance of our proposed method in the case where both stage 1 and stage 2 treatment effects are heterogeneous, i.e., $V_i = Z_{1i}A_{1i}$ and $\alpha_1 = -4, \alpha_2 = -0.2$.

		$c_1 = 0$	$c_1 = 1$	$c_1 = 2$	$c_1 = 3$	$c_1 = 4$
Q-learning	$V_i \sim N(-1, 1)$	1	1	1	1	1
	$V_i \sim N(2Z_{1i}Z_{2i}, 1)$	1	1	1	0.992	0.981
	$V_i = Z_{1i}A_{1i}$	1	1	0.965	0.707	0.543
Interactive Q-learning	$V_i \sim N(-1, 1)$	1	1	1	1	1
	$V_i \sim N(2Z_{1i}Z_{2i}, 1)$	1	1	0.998	0.990	0.976
	$V_i = Z_{1i}A_{1i}$	1	1	0.965	0.707	0.541

Table 3.2: Percentage of correctly identified stage 1 optimal rules when stage 2 treatment effects are homogeneous across subjects ($\alpha_1 = \alpha_2 = 0$), prediction using standard Q-learning and interactive Q-learning, based on a set of test data ($N = 10000$) and 100 simulations of training data ($n = 250$).

3.6.2 Results

We compare the proposed method (mIQ) with standard Q-learning (sQ), modified Q-learning (mQ; Huang et al. (2015)), and interactive Q-learning (IQ; Laber et al. (2014)).

c_1	c_2	sQ	mQ	IQ	mIQ
0.0	1.0	0.963	0.963	0.969	0.969
	2.0	0.969	0.970	0.976	0.976
	3.0	0.969	0.969	0.973	0.973
2.0	1.0	0.869	0.965	0.910	0.985
	2.0	0.900	0.964	0.947	0.981
	3.0	0.921	0.966	0.963	0.979
4.0	1.0	0.893	0.972	0.892	0.986
	2.0	0.871	0.966	0.910	0.985
	3.0	0.883	0.963	0.934	0.981

Table 3.3: Percentage of correctly identified stage 1 optimal rules when stage 2 treatment effects are heterogeneous across subjects ($\alpha_1 = -4, \alpha_2 = -0.2$), based on a set of test data ($N = 10000$) and 100 simulations of training data ($n = 250$).

Table 3.3 summarizes the probabilities of correctly identifying stage 1 optimal rules based on the four methods considered, and mIQ has the highest accuracy across all scenarios. The results show that *the modified interactive Q-learning algorithm corrects for potential bias generated from both stage 1 and stage 2 heterogeneous treatment effects.*

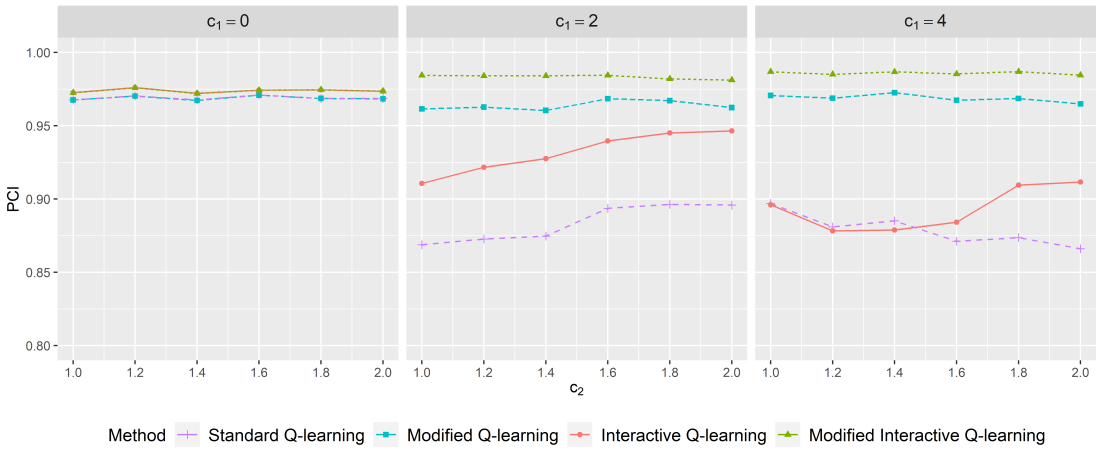

Figure 3.2: Probability of correctly identifying stage 1 optimal rules as a function of c_2 for $c_1 = 0, 2, 4$.

Figure 3.2 plots the trend with varied c_2 at finer intervals for the three values of c_1. If stage 1 treatment effects are homogeneous ($c_1 = 0$), then the four methods work similarly, and modified interactive Q-learning performs exactly the same as interactive Q-learning. If stage 1 treatment effects are highly heterogeneous ($c_1 = 4$), then modified interactive outperforms all other methods, and there is a substantial difference in the probability of correctly identifying stage 1 optimal rules between modified Q-learning and modified interactive Q-learning. It further stresses the importance of building interactive models into modified Q-learning.

3.7 Data Analysis

We use the M-bridge data ($n = 591$) to illustrate the application of the proposed method on SMARTs with embedded tailoring. Figure 3.1 shows the design of M-bridge for this analysis. Our aim is to identify the personalized optimal DTR for each subject in the dataset, which minimizes binge drinking (primary outcome, 490 subjects with complete data on the maximum number of drinks within a 24-hour period at baseline and follow-up), and negative drinking-related consequences (secondary outcome, 496 subjects with complete data on the total number of drinking-related consequences in the past 30 days at baseline and follow-up). We perform the data analysis separately for these two outcomes.

Stage 2 model uses as covariates baseline characteristics, including gender, race, intention to pledge Greek, indicator whether parent has a significant drinking problem, pre-college drinking norms and intention for college drinking, and pre-college drinking habits, and the interaction between stage 2 intervention and baseline characteristics. Stage 1 model uses as covariates baseline characteristics and their interaction with stage 1 intervention. A summary of the covariates and outcomes used in the model is presented in Table 3.4.

Table 3.4: Summary statistics of subject characteristics by initial randomization (stage 1 intervention). Discrete variables are summarized by counts; continuous variables are summarized by mean (SD).

	Early Intervention ($n_1 = 295$)	Late Intervention ($n_2 = 296$)	Overall ($n = 591$)
Demographics			
Gender			
Male	115	104	219
Female	180	192	372

Continued on next page

Table 3.4 – *Continued from previous page*

	Early Intervention ($n_1 = 295$)	Late Intervention ($n_2 = 296$)	Overall ($n = 591$)
Race			
White	229	222	451
Nonwhite	66	74	140
Intention to pledge Greek			
Yes	34	33	67
No or Undecided	261	263	524
Parent drinking problem[1]			
Clearly Yes	46	31	77
Clearly No or Not Sure	247	264	511
Pre-college drinking norms			
Percent of students drink (%)[2]	53.0 (19.9)	52.4 (21.1)	52.7 (20.5)
Number of drinks per week[3]	5.65 (8.35)	5.24 (6.01)	5.45 (7.27)
Max number of drinks in a row[4]	5.62 (3.32)	5.35 (3.55)	5.48 (3.44)
Percent of students binge (%)[5]	22.4 (17.5)	22.9 (17.8)	22.7 (17.7)
Intention for college drinking			
Drinking frequency per month [6]	2.37 (2.70)	2.29 (2.80)	2.33 (2.75)
Number of drinks[7]	1.98 (1.77)	2.14 (1.91)	2.06 (1.84)
Drunk frequency per month[8]	1.39 (1.98)	1.43 (2.28)	1.41 (2.13)
Pre-college drinking habits			
Number of days [9]	2.05 (3.09)	2.03 (3.25)	2.04 (3.17)

Continued on next page

[1]Indicator whether the subject's mother or father has had a significant drinking problem that did or should have led to treatment

[2]Norm on the percentage of UMN first-year students who used alcohol during the last 30 days

[3]Norm on the number of alcoholic drinks a typical UMN first-year student consumed during an average week

[4]Norm on the largest number of drinks a typical college student had in a row during the last two weeks

[5]Norm on the percentage of UMN first-year students had five or more drinks in a sitting during the last two weeks

[6]Intent frequency of drinking alcohol in the next 6 months

[7]Intent number of drinks on a typical occasion

[8]Intent frequency of consuming enough alcohol to feel drunk or intoxicated in the next 6 months

[9]Number of days using alcohol during the last 30 days

Table 3.4 – *Continued from previous page*

	Early Intervention ($n_1 = 295$)	Late Intervention ($n_2 = 296$)	Overall ($n = 591$)
Average number of drinks[10]	1.83 (2.37)	1.75 (2.35)	1.79 (2.36)
Re-randomization/Stage 2			
Heavy drinker	75	83	158
Non-heavy drinker	220	213	433
Stage 2 intervention			
Continued self-monitoring[11]	220	213	433
Online health coach	37	43	80
Automated email	38	40	78

	`max_drinks` ($n = 490$)		`byaacq` ($n = 496$)	
	Baseline	End of Semester 1	Baseline	End of Semester 1
Early/Email	5.83 (3.17)	7.20 (3.22)	2.89 (3.06)	5.06 (4.04)
Early/Coach	4.94 (3.08)	7.15 (2.68)	2.97 (2.26)	5.18 (4.77)
Early/-	1.72 (2.88)	2.37 (2.90)	0.92 (2.15)	1.32 (2.43)
Late/Email	3.78 (2.66)	6.03 (2.86)	2.79 (2.46)	4.15 (4.06)
Late/Coach	6.22 (4.81)	8.03 (3.69)	3.28 (3.43)	5.22 (4.18)
Late/-	1.20 (2.10)	2.34 (2.87)	0.57 (1.52)	1.13 (2.05)
Overall	2.55 (3.38)	3.72 (3.68)	1.39 (2.42)	2.28 (3.40)

Table 3.5: Summary statistics of outcomes by DTR. Continuous variables are summarized by mean (SD). "Early/-" and "Late/-" indicate the subgroup of non-heavy drinkers who were not re-randomized to stage 2 interventions.

Table 3.5 summarizes the outcomes by observed DTRs. For the primary outcome (`max_drinks`), stage 2 model utilizes data on the 140 heavy drinkers with complete data, who were flagged based on the frequency of binge and high-intensity drinking during

[10]Number of drinks had on a typical day when drinking alcohol during the last 30 days

[11]Subjects who were identified as non-heavy drinkers were not randomized at stage 2 and continued self-monitoring

the self-monitoring period at stage 1, whereas stage 1 model utilizes data on the 490 enrolled students with complete data. Implementing modified interactive Q-learning, 181 (37.0%) subjects are predicted to benefit most from receiving late intervention at stage 1 based on their baseline characteristics, and 90 (64.3%) are predicted to benefit most from receiving automated email at stage 2 had they received the predicted stage 1 optimal treatment.

For the secondary outcome (`byaacq`), stage 2 model utilizes data on the 142 heavy drinkers with complete data, whereas stage 1 model utilizes data on the 496 enrolled students with complete data. Implementing modified interactive Q-learning, 229 (46.2%) subjects are predicted to benefit most from receiving late intervention at stage 1 based on their baseline characteristics, and 77 (54.2%) are predicted to benefit most from receiving automated email at stage 2 had they received the predicted stage 1 optimal treatment.

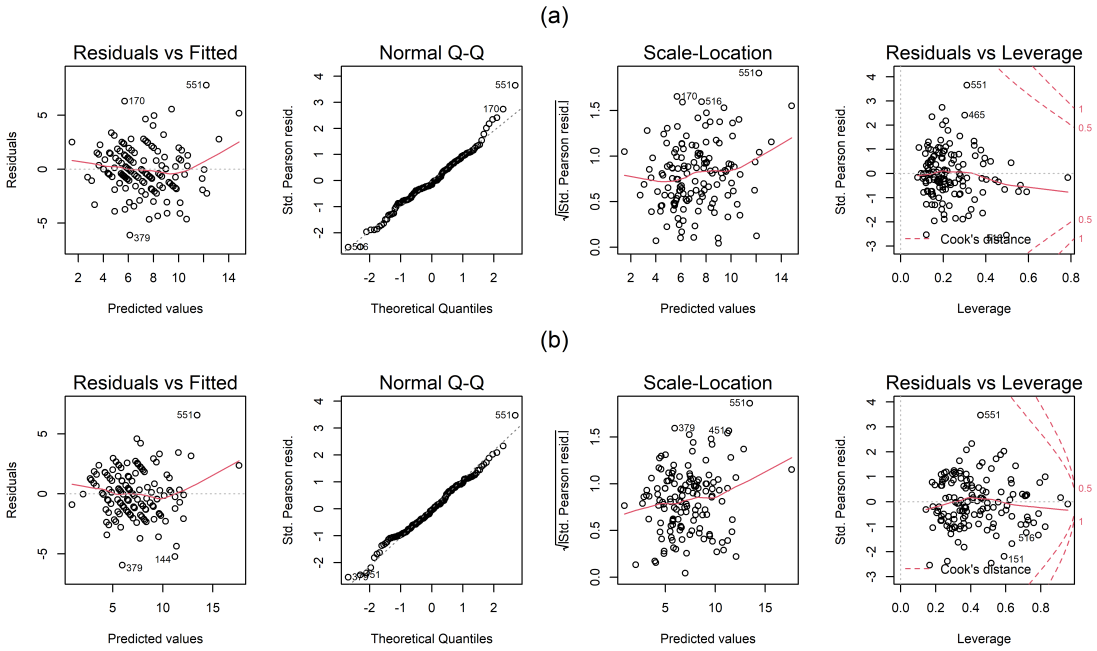

Figure 3.3: Residual diagnostics for (a) the parsimonious model (b) the saturated model: an illustration using the primary outcome `max_drinks`.

The proposed method requires that any missingness in the outcome measurements is not informative. M-bridge might not be a perfect example of informative residuals, because the interaction between A_1 and Z_1 does not have a significantly impact on the outcome. Figure 3.3 shows that, for the primary outcome, residual diagnostics of the parsimonious model do not differ much from those of the saturated model. However, modified interactive Q-learning still helps as the huge amount of baseline covariates might cause the problem of overfitting. Nevertheless, we anticipate that modified interactive Q-learning performs similarly as interactive Q-learning in this case. The comparison is shown in the supplementary materials.

3.8 Discussion

This chapter attempts to understand the bias caused by misspecification of the main effect model in the implementation of Q-learning, specifically, omitting an unmeasured variable, and confirms the existence of bias when the unmeasured variable is correlated with the observed predictors (especially stage 1 treatment). As stage 1 heterogeneous treatment effects are usually neglected in the stage 2 main effect model and might be a significant predictor, we then modify the Q-learning algorithm to account for the unexplained residuals. This modification is then built into the interactive Q-learning, which corrects the bias generated by stage 2 heterogeneous treatment effects and the optimization operation, to improve the overall performance where both stage 1 and 2 treatment effects are heterogeneous. The proposed method and the simulation results assume that the two randomization arms at each stage are equally assigned. If the randomization scheme has an unbalanced allocation ratio, then the estimators of stage 2 treatment effects would be biased in the presence of an unmeasured variable. Future work and exploration of Q-learning in this context would be merited.

Chapter 4

A Generalization to Dichotomous Outcomes

4.1 Background

Clinical investigators are commonly interested in dichotomous endpoints as the outcome can be easily differentiated across subjects, especially when there is a gold standard cut-off point with a clear justification. To optimize DTRs for dichotomous outcomes, one common method is Q-learning with logistic or probit regression. This chapter aims to address the informative residuals due to misspecification of late-stage main effect model for dichotomous outcomes. Methods that posit more flexibility in modeling Q-functions include generalized additive model (Moodie et al., 2014), which allows a nonlinear relationship between outcome and predictors, and Bayesian machine learning (Murray et al., 2018), which obviates the nonregularity problem in Q-learning and utilizes a probit Bayesian additive regression trees model for the dichotomous outcomes. Despite the fact that these methods perform well in many circumstances, they are not developed for the purpose of dealing with informative residuals. For Q-learning with linear regression, a modification with Murphy's regret function (Huang et al., 2015) was

proposed to correct for the informative residuals. However, a difficulty arises if modified Q-learning is used for dichotomous outcomes, as the probability of a success is linked to the linear combination of predictors using a non-identity function.

One solution to this problem would be imposing monotonicity of preferences to early-stage pseudo-outcomes. If a subject was observed to receive an optimal treatment, then the corresponding pseudo-outcome would be the observed outcome. If a subject was observed to receive a suboptimal treatment and responded with the desirable outcome, then they must have responded, had they received the optimal treatment. However, there is no restriction that we could impose on potential pseudo-outcomes for subjects who received the suboptimal treatment and responded with the undesirable outcome. These subjects could possibly respond with desirable or undesirable outcomes had they received the optimal treatment. Therefore, this approach results in a loss of information.

We present Q-learning with probit regression and the proposed modification in Section 4.2. A simulation study is performed in Section 4.3 to investigate the potential improvement of the modification. The data example in Section 3.2.1 with a different set of outcomes is used to illustrate the methods, and the results are summarized in Section 4.4. Moreover, probit regression is equivalent to a latent variable model where the underlying continuous variable is modeled explicitly. In Section 4.5, we discuss the feasibility of using latent variable modeling to correctly estimate the underlying continuous variable, and propose a novel algorithm which samples surrogates from a truncated normal distribution as estimates of the latent variable and adds back the Murphy's regret function to generate early-stage pseudo-outcomes.

4.2 Methods

Suppose that the data collected from a SMART are represented by a sequence of independently and identically distributed random variables (Z_1, A_1, Z_2, A_2, Y), where Z_1 is the set of baseline covariates and moderators measured prior to stage 1 randomization,

Z_2 is the set of time-varying covariates and potential tailoring variables measured between stage 1 and stage 2 randomizations, $A_k \in \mathcal{A}_k$, $k = 1, 2$, is the treatment at stage k, with $\mathcal{A}_k$ being the set of all possible treatments, and $Y \in (0, 1)$ is the dichotomous outcome measured after treatment stages, with $Y = 0$ preferred.

In the M-bridge study, the investigators are primarily interested in preventing binge and high-intensity drinking, so whether a student had binge drinking, i.e., the frequency of binge drinking was greater than 0, would be an outcome of great interest. We hypothesize a scenario where the actual frequency of binge drinking for each college student is hidden in the data cleaning process, and we only have the information whether the frequency of binge drinking is greater than 0. A value of 0 is preferred for the indicator of binge drinking. The secondary outcome considered by the investigators is health services utilization, which is a dichotomous measure of whether students used any health services over the past three months, including through a healthcare clinic, individual counseling, group therapy, support groups, self-help resources, or other services/resources. A value of 1 is preferred for health services utilization.

4.2.1 Q-learning with Probit Regression

One commonly used method to optimize DTRs for binary outcomes in the literature is Q-learning with logistic regression or probit regression. With the probit link, an equivalent formulation is to assume that there is a latent variable that is linearly related to covariates with an error term following a standard normal distribution, and we only observe the indicator for whether the latent variable is greater than a certain threshold. The alternative formulation will be discussed in detail in Section 4.5.

Consider the model for $Q_2(H_2, A_2) = \mathbb{E}(Y|H_2, A_2) = \mathbb{P}(Y = 1|H_2, A_2)$:

$$Q_2(H_{2i}, A_{2i}; \boldsymbol{\beta}_{20}, \boldsymbol{\beta}_{21}) = \Phi\left(\boldsymbol{x}_{20,i}^T \boldsymbol{\beta}_{20} + a_2 \boldsymbol{x}_{21,i}^T \boldsymbol{\beta}_{21}\right), \tag{4.1}$$

where Φ is the cumulative distribution function of a standard normal random variable.

As Φ is a monotonically increasing function, the stage 2 optimal rule is

$$\hat{d}_{2i}^{\text{opt}} = \underset{a_2 \in \{-1,1\}}{\arg\min}\ \Phi\left(\boldsymbol{x}_{20,i}^T \hat{\boldsymbol{\beta}}_{20} + a_2 \boldsymbol{x}_{21,i}^T \hat{\boldsymbol{\beta}}_{21}\right) = -\text{sgn}\left\{\boldsymbol{x}_{21,i}^T \hat{\boldsymbol{\beta}}_{21}\right\}.$$

The pseudo-outcome at stage 1, conditional on stage 2 optimization, is

$$\tilde{Y}_{1i} = \min_{a_2 \in \{-1,1\}} \Phi^{-1} Q_2(H_{2i}, A_{2i} = a_2) = \boldsymbol{x}_{20,i}^T \hat{\boldsymbol{\beta}}_{20} - \left|\boldsymbol{x}_{21,i}^T \hat{\boldsymbol{\beta}}_{21}\right|. \tag{4.2}$$

Consider $Q_1(H_{i1}, A_{1i}; \boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{11}) = \boldsymbol{x}_{10,i}^T \boldsymbol{\beta}_{10} + a_1 \boldsymbol{x}_{11,i}^T \boldsymbol{\beta}_{11}$. Stage 1 least squares estimators are obtained by

$$\left(\hat{\boldsymbol{\beta}}_{10}, \hat{\boldsymbol{\beta}}_{11}\right) = \underset{\boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{11}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left\{\tilde{Y}_{1i} - Q_1(H_{i1}, A_{1i}; \boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{11})\right\}^2.$$

The stage 1 optimal rule is estimated as

$$\hat{d}_{1i}^{\text{opt}} = \underset{a_1 \in \{-1,1\}}{\arg\min}\ Q_1\left(H_{i1}, A_{1i}; \hat{\boldsymbol{\beta}}_{10}, \hat{\boldsymbol{\beta}}_{11}\right) = -\text{sgn}\left\{\boldsymbol{x}_{11,i}^T \hat{\boldsymbol{\beta}}_{11}\right\}.$$

Logit link is sometimes preferred in stage 2 estimation to analyze real data, as it leads to the estimation of a commonly used association measure, odds ratio, which is easily interpretable by clinical investigators. Nevertheless, both link functions produce similar results.

4.2.2 The Proposed Modification

If the main effect model of the stage 2 Q-function in Equation (4.1) is misspecified, i.e., an important predictor is omitted, it is difficult to correct for any informative residual remaining due to the presence of the non-identity link. It is, therefore, tempting to impose some restrictions directly on stage 1 pseudo-outcome $\tilde{Y}_1$ so that the imputed distribution over $\tilde{Y}_1$ is closer to reality. Our proposed modification follows the concept of Murphy's regret function. At stage 2, if a subject is observed to receive the optimal

treatment, then the stage 1 pseudo-outcome would be the same as the observed outcome. On the other hand, if a subject receives the suboptimal treatment, we are not able to simply add back Murphy's regret function to obtain the counterfactual outcome under the optimal treatment. Instead, we impose monotonicity of preferences, i.e., the treatment which yields a more desirable potential outcome is *always preferred*. Thus, the treatment associated with the *most desirable* potential outcome is *optimal*. Specifically, if a subject receives the suboptimal treatment and responds with the desirable outcome $Y = 0$, then they must have responded with $Y = 0$ had they received the optimal treatment. If a subject receives the suboptimal treatment and responds with the undesirable outcome $Y = 1$, then we cannot make any correction to stage 1 pseudo-outcome as it is not guaranteed that the subject would respond desirably enough to reach $Y = 0$ counterfactually. The stage 1 pseudo-outcome in this case should still be drawn from the model in Equation (4.1). If the estimated optimal probability of an event, $\min_{a_2 \in \{-1,1\}} \hat{\mathbb{P}}(Y = 1 | H_2, A_2 = a_2)$, is greater than 0.5, then we assign the stage 1 pseudo-outcome as 1; otherwise we assign the stage 1 pseudo-outcome as 0. Thus, the modified pseudo-outcome at stage 1, conditional on stage 2 optimization, is

$$\begin{aligned}
\tilde{Y}_{1i}^{m} &= \mathbb{1}\left\{\hat{d}_{2i}^{\text{opt}} = A_{2i}\right\} Y_i \\
&\quad + \mathbb{1}\left\{\hat{d}_{2i}^{\text{opt}} \neq A_{2i}\right\} \left[\mathbb{1}\left\{Y_i = 0\right\} Y_i + \mathbb{1}\left\{Y_i = 1\right\} \mathbb{1}\left\{\Phi\left(\tilde{Y}_{1i}\right) > 0.5\right\}\right] \\
&= \mathbb{1}\left\{\hat{d}_{2i}^{\text{opt}} = A_{2i}\right\} Y_i + \mathbb{1}\left\{\hat{d}_{2i}^{\text{opt}} \neq A_{2i}\right\} \mathbb{1}\left\{Y_i = 1\right\} \mathbb{1}\left\{\Phi\left(\tilde{Y}_{1i}\right) > 0.5\right\}. \qquad (4.3)
\end{aligned}$$

where $\tilde{Y}_{1i}$ is defined in Equation (4.2) and $\Phi\left(\tilde{Y}_{1i}\right) = \min_{a_2 \in \{-1,1\}} Q_2\left(H_{2i}, A_{2i} = a_2; \hat{\boldsymbol{\beta}}_{20}, \hat{\boldsymbol{\beta}}_{21}\right)$.

Different from Q-learning with probit regression, where we obtain stage 1 estimators directly using least squares estimation, stage 1 estimators in the modified algorithm are again obtained by a probit regression. Consider $Q_1(H_{i1}, A_{1i}; \boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{11}) =$

$\Phi\left(\boldsymbol{x}_{10,i}^T\boldsymbol{\beta}_{10} + a_1\boldsymbol{x}_{11,i}^T\boldsymbol{\beta}_{11}\right)$. The stage 1 optimal rule is estimated as

$$\hat{d}_{1i}^{\text{opt}} = \underset{a_1\in\{-1,1\}}{\arg\min}\ \Phi\left(\boldsymbol{x}_{10,i}^T\hat{\boldsymbol{\beta}}_{10} + a_1\boldsymbol{x}_{11,i}^T\hat{\boldsymbol{\beta}}_{11}\right) = -\text{sgn}\left\{\boldsymbol{x}_{11,i}^T\hat{\boldsymbol{\beta}}_{11}\right\}.$$

The definition of $\tilde{Y}_{1i}^m$ in Equation (4.3) has a piece of information missing: if a subject receives a suboptimal treatment and responds with the undesirable outcome $Y = 1$, though we are not able to make any assumption on stage 1 pseudo-outcome, the underlying continuous variable should have at most the same value had they received the optimal treatment. Since the proposed modification does not utilize all the information, its performance is expected to be compromised. However, the proposed approach has some potential advantages to release the reliance of Q-learning with probit regression on the correct specification of stage 2 Q-function.

4.3 Simulation Study

We conduct a simulation study to assess the performance of the proposed modification. Assume a sequence of data collected from a two-stage SMART study is

$$(Z_{1i}, A_{1i}, Z_{2i}, A_{2i}, Y_i),\ i = 1, \ldots, n,$$

where $Z_{1i} \overset{\text{i.i.d.}}{\sim} N(-2, 1)$, $Z_{2i} = Z_{1i} + \phi_i$ and $\phi_i \overset{\text{i.i.d.}}{\sim} N(0, 1)$, $\dfrac{A_{1i}+1}{2} \overset{\text{i.i.d.}}{\sim}$ Bernoulli(0.5) and $\dfrac{A_{2i}+1}{2} \overset{\text{i.i.d.}}{\sim}$ Bernoulli(0.5). There is no embedded tailoring variable in this design. Suppose

$$\begin{aligned} Y_i = \mathbb{1}\{&3 - Z_{1i} + 0.1A_{1i} - 0.1Z_{2i} + \gamma Z_{1i}A_{1i} \\ &+ A_{2i}\left(-9 - 6Z_{1i} + 7.5A_{1i} - 0.3Z_{2i}\right) + \varepsilon_i > 10\}, \end{aligned}$$

where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. Suppose $Y_i = 0$ is preferred.

A test dataset of N subjects is generated with potential outcomes under the four

treatment regimes for each subject. The true optimal rules, d_1^{opt} and d_2^{opt}, are known for each subject in the test dataset by minimizing the potential latent outcomes over all treatment regimes. Hence, the test dataset assumes that monotonicity of preferences holds. Training datasets of sample size n are simulated 1000 times. The training datasets are then used to estimate the linear models specified for Q-function at each stage, and the estimated models are used to predict the optimal rules $\hat{d}_1^{\text{opt}}$ and $\hat{d}_2^{\text{opt}}\left(d_1^{\text{opt}}\right)$ for each subject in the test dataset.

The datasets are simulated based on varying values of γ and σ, where γ controls the size of the omitted stage 1 heterogeneous treatment effects, indicating the extent to which the stage 2 main effect model is misspecified, and σ controls the variance of the outcome. The results are summarized below.

γ	PCI_1		PCI_2	
	Q-probit	mQ-probit	Q-probit	mQ-probit
0	0.796 (0.090)	0.591 (0.218)	0.986 (0.000)	0.986 (0.000)
1	0.837 (0.069)	0.760 (0.080)	0.989 (0.000)	0.989 (0.000)
2	0.859 (0.058)	0.814 (0.017)	0.993 (0.003)	0.993 (0.003)
3	0.882 (0.046)	0.856 (0.029)	0.991 (0.015)	0.991 (0.015)
4	0.880 (0.018)	0.908 (0.055)	0.984 (0.036)	0.984 (0.036)
5	0.885 (0.009)	0.932 (0.053)	0.987 (0.047)	0.987 (0.047)

Table 4.1: Percentage of correctly identified optimal rules (mean (SD); $\sigma = 3$), prediction using Q-learning with probit regression and modified Q-learning with probit regression, based on a set of test data ($N = 10000$) and 1000 simulations of training data ($n = 250$).

Table 4.1 summarizes the probability of correctly identifying optimal rules (PCI) at both stages using (1) Q-learning with probit regression (2) modified Q-learning with probit regression. The modified algorithm does not necessarily perform better than the original algorithm for all γ's. As γ increases, i.e., stage 2 model residuals become more informative to stage 1 rule estimation, the modified algorithm starts to outperform the original algorithm. The turnaround, in this simulation, occurs at $\gamma = 4$. An intuitive explanation to the exceptionally bad performance of mQ-probit when there is

no model misspecification is that, hastily using the observed binary outcome as stage 1 pseudo-outcome, when the estimated stage 2 optimal rule is the same as the observed treatment, might cause harm as misspecification of stage 2 main effect model can affect the estimation of stage 2 treatment effect model due to the probit link function.

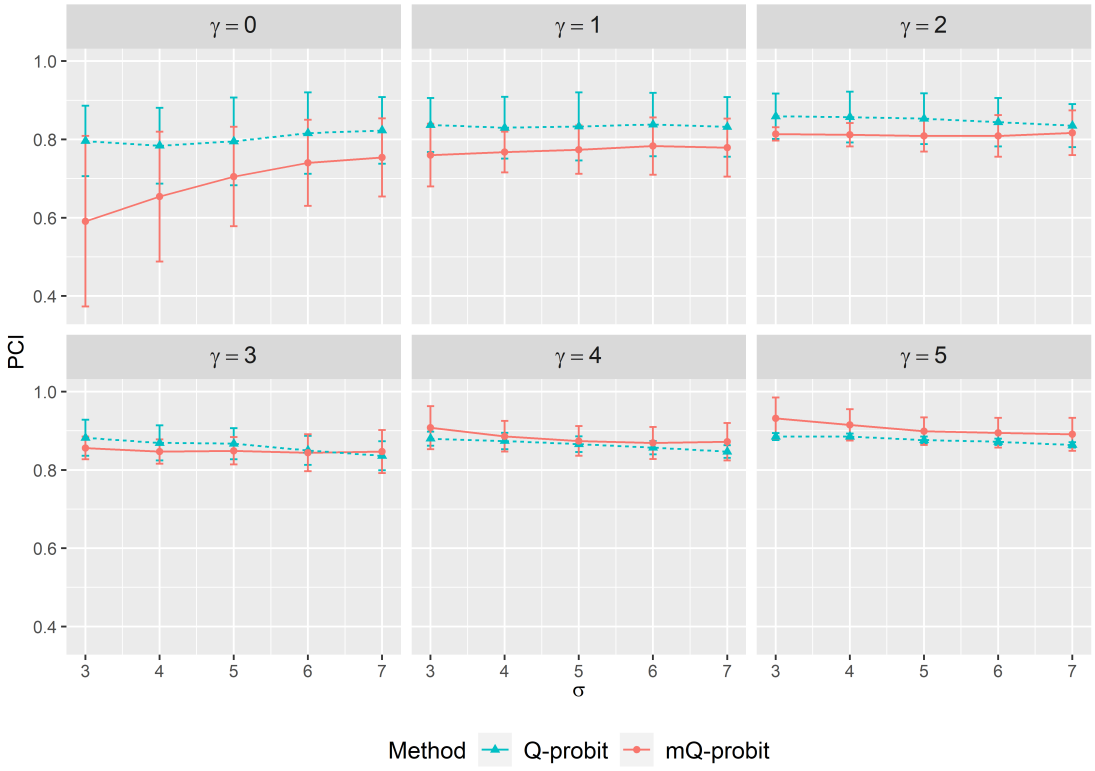

Figure 4.1: Probability of correctly identifying stage 1 optimal rules as a function of σ for different values of γ using (1) Q-learning with probit regression, and (2) modified Q-learning with probit regression.

Figure 4.1 shows that improvement of the proposed modification on correctly identifying the optimal rules is not consistent across the value of σ. For scenarios where the proposed modification improves the algorithm performance, PCI has a general decreasing trend as σ increases. This drawback, however, is not surprising, because the

uncertainty of classifying the outcome as desirable or undesirable increases as the underlying latent variable is largely dispersed.

4.4 Data Analysis

We use the M-bridge data ($n = 591$) to illustrate the application of modified Q-learning with probit regression on SMARTs with embedded tailoring. We follow the design illustrated in Section 3.2.1 for this analysis, but analyze a different set of outcomes. Our aim is to identify the personalized optimal DTR for each subject in the dataset, which minimizes binge drinking (primary outcome, 495 subjects with complete data on the frequency of binge drinking during the last 30 days at baseline and follow-up), or maximizes health services utilization (secondary outcome, 493 subjects with complete data on if the subject used any health resources or services indicated in the questionnaire in the last 3 months at baseline and follow-up). We perform data analysis separately for these two outcomes.

		`ind_binge` (%)			`hs_util` (%)	
	n	Baseline	Semester 1	n	Baseline	Semester 1
Early/Email	36	66.7	88.9	36	72.2	52.8
Early/Coach	33	51.5	93.9	33	60.6	48.5
Early/-	179	16.8	24.6	178	71.3	52.8
Late/Email	33	57.6	75.8	33	81.8	54.5
Late/Coach	40	62.5	92.5	40	70.0	55.0
Late/-	174	12.6	25.9	173	72.8	56.6
Overall	495	27.7	43.2	493	71.8	54.2

Table 4.2: Summary statistics of outcomes by DTR: the percentage of subjects who had binge drinking, and the percentage of subjects who did *not* utilize health services. Higher percentage is undesirable. "Early/-" and "Late/-" indicate the subgroup of non-heavy drinkers who were not re-randomized to stage 2 interventions, and n denotes the sample size in each DTR.

Suppose the hypothetical definition of the indicator for binge drinking (`ind_binge`) is that the frequency of binge drinking is nonzero during the last 30 days at the time of measurement. Thus, `ind_binge` = 0 is desirable. On the other hand, the indicator for health services utilization (`hs_util`) is defined by utilizing one or more type of health sources or services, including a healthcare clinic, individual counseling, group therapy, support groups, self-help resources, over the last 3 months. Thus, `hs_util` = 1 is desirable. Table 4.2 summarizes the percentage of students who experienced the undesirable outcomes marginally, or for each combination of the stage 1/2 interventions.

Data analysis on SMARTs with embedded tailoring using Q-learning with probit regression is slightly different from the algorithm in Section 4.2.1 as a result of subgroup re-randomization. Stage 1 pseudo-outcome needs to be a dichotomous outcome with values 0/1 as opposed to a continuous outcome on the scale of linear predictors. If the estimated optimal probability of having an event is greater than 0.5, then the stage 1 pseudo-outcome is assigned value 1; otherwise, it is assigned value 0.

Method	n_1	Stage 1 Prediction		n_2	Stage 2 Prediction	
		$\hat{d}_1^{\text{opt}} = -1$	$\hat{d}_1^{\text{opt}} = 1$		$\hat{d}_2^{\text{opt}} = -1$	$\hat{d}_2^{\text{opt}} = 1$
			`ind_binge`			
Q-probit	495	294 (59.4%)	201 (40.6%)	142	128 (90.1%)	14 (9.86%)
mQ-probit	495	268 (54.1%)	227 (45.9%)	142	128 (90.1%)	14 (9.86%)
			`hs_util`			
Q-probit	493	205 (41.6%)	288 (58.4%)	142	55 (38.7%)	87 (61.3%)
mQ-probit	493	233 (47.3%)	260 (52.7%)	142	55 (38.7%)	87 (61.3%)

Table 4.3: Data analysis results using Q-probit and mQ-probit.

The results in Table 4.3 show that Q-probit generates a more distinctive difference when predicting if a student would benefit most from receiving stage 1 intervention early ($\hat{d}_1^{\text{opt}} = 1$) or late ($\hat{d}_1^{\text{opt}} = -1$) than mQ-probit for both outcomes `ind_binge` and `hs_util`. However, Table 4.2 does not show a consistent and distinctive trend in either

outcome comparing treatment regimes with early stage 1 intervention and treatment regimes with late stage 1 intervention. Though we cannot validate the accuracy of estimating individualized optimal DTR, the marginal summary indicates that mQ-probit generates a more reasonable result.

4.5 A Latent Variable Approach

The proposed modification in Section 4.2.2 improves Q-learning with probit regression if residuals from stage 2 model is largely informative, but the improvement is limited, and if misspecification of stage 2 main effect model does not deviate from reality much, the proposed modification may be harmful to estimating the optimal rules. To further improve the robustness of our algorithm, it is natural to consider the latent variable approach. We propose a novel algorithm to optimize DTRs for dichotomous outcome and explore the feasibility of using this framework to deal with informative residuals.

4.5.1 Framework

Let Y^* denote the latent continuous variable. The dichotomous outcome Y is defined as

$$Y = \begin{cases} 1 & \text{if } Y^* > \xi \\ 0 & \text{otherwise} \end{cases},$$

where ξ is a known threshold of scientific choice.

In the literature of latent variable modeling (Skrondal and Rabe-Hesketh, 2004), the expectation of the observed variable Y can be expressed as a function of the underlying distribution of the latent variable Y^*. Suppose $Y^* = \mu + \epsilon$ and ϵ is independent and identically distributed with mean 0 and variance σ^2. Then

$$\mathbb{E}(Y|\mu) = \mathbb{P}(Y = 1|\mu) = P(Y^* > \xi|\mu) = P(\mu + \epsilon > \xi) = 1 - F_\epsilon(\xi - \mu), \tag{4.4}$$

where F_ϵ is the cumulative distribution function of ϵ.

4.5.2 The Proposed Algorithm

Suppose $\mu_i = \boldsymbol{x}_{20,i}^T\boldsymbol{\beta}_{20} + a_2\boldsymbol{x}_{21,i}^T\boldsymbol{\beta}_{21}$. Following the formulation in Section 4.5.1, we derive the log-likelihood of model parameters:

$$\boldsymbol{\ell}\left(\boldsymbol{\beta}_{20}, \boldsymbol{\beta}_{21}, \sigma \,|\{y_i\}_{i=1}^n, \{y_i^*\}_{i=1}^n\right) = -n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n \left(y_i^* - \boldsymbol{x}_{20,i}^T\boldsymbol{\beta}_{20} - a_2\boldsymbol{x}_{21,i}^T\boldsymbol{\beta}_{21}\right)^2. \tag{4.5}$$

Now we propose a novel algorithm called *modified latent Q-learning* (mLQ) to optimize DTRs for dichotomous outcomes. Here is a brief outline of the proposed algorithm. First, we utilize an Expectation-Maximization (EM) algorithm to estimate the stage 2 model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}_{20}, \boldsymbol{\beta}_{21}, \sigma)$. Using an idea adapted from Liu and Zhang (2018), a surrogate of Y^* is drawn from the conditional distribution $f(y^*|y)$ so that the surrogate variable follows a distribution of $\sum_y f(y^*|y)f(y)$. Note that the marginal distribution $f(y)$ is observed, but the conditional distribution $f(y^*|y)$ can be misspecified. We then use the surrogate variable and Murphy's regret function to construct stage 1 pseudo-outcomes.

The detailed steps are as follows:

Stage 2 estimation: the EM algorithm

(mLQ-1) Initialize $\boldsymbol{\theta}^{(v)}$ for $v = 0$;

(mLQ-2) *E-step*: Take the expectation of $\boldsymbol{\ell}\left(\boldsymbol{\theta}\,|\{y_i\}_{i=1}^n, \{y_i^*\}_{i=1}^n\right)$ over the conditional distribution $f(Y^*|Y)$ characterized by $\boldsymbol{\theta}^{(v)}$ to obtain

$$\begin{aligned} Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(v)}; \{y_i\}_{i=1}^n\right) = &- n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta}^{(v)}}\left(y_i^{*2}\middle|\, y_i\right) \\ &+ \frac{1}{\sigma^2}\sum_{i=1}^n \mu_i \mathbb{E}_{\boldsymbol{\theta}^{(v)}}\left(y_i^*\middle|\, y_i\right) - \frac{1}{2\sigma^2}\sum_{i=1}^n \mu_i^2, \end{aligned} \tag{4.6}$$

where

$$\mathbb{E}_{\boldsymbol{\theta}^{(v)}}\left(y_i^*\middle|\, y_i\right) = \mu_i^{(v)} + \frac{(2y_i-1)\sigma^{(v)2}\phi\left(\mu_i^{(v)},\sigma^{(v)2};\xi\right)}{y_i\left\{1-\Phi\left(\mu_i^{(v)},\sigma^{(v)2};\xi\right)\right\}+(1-y_i)\Phi\left(\mu_i^{(v)},\sigma^{(v)2};\xi\right)},$$

$$\mathbb{E}_{\boldsymbol{\theta}^{(v)}}\left(y_i^{*2}\middle|\, y_i\right) = \mu_i^{(v)2} + \sigma^{(v)2} + \frac{(2y_i-1)(\mu_i^{(v)}+\xi)\sigma^{(v)2}\phi\left(\mu_i^{(v)},\sigma^{(v)2};\xi\right)}{y_i\left\{1-\Phi\left(\mu_i^{(v)},\sigma^{(v)2};\xi\right)\right\}+(1-y_i)\Phi\left(\mu_i^{(v)},\sigma^{(v)2};\xi\right)},$$

$$\mu_i^{(v)} = \boldsymbol{x}_{20,i}^T\boldsymbol{\beta}_{20}^{(v)} + a_2\boldsymbol{x}_{21,i}^T\boldsymbol{\beta}_{21}^{(v)},$$

and $\phi(\mu,\sigma^2)$ and $\Phi(\mu,\sigma^2)$ denote the probability density function and cumulative distribution function of $N(\mu,\sigma^2)$ respectively;

(mLQ-3) *M-step*: Let $\boldsymbol{x}_{2i} = \begin{pmatrix}\boldsymbol{x}_{20,i}^T & a_2\boldsymbol{x}_{21,i}^T\end{pmatrix}^T$. Maximize $Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(v)};\{y_i\}_{i=1}^n\right)$ to derive $\boldsymbol{\theta}^{(v+1)}$ as

$$\begin{pmatrix}\boldsymbol{\beta}_{20}^{(v+1)}\\ \boldsymbol{\beta}_{21}^{(v+1)}\end{pmatrix} = \left\{\sum_{i=1}^n \boldsymbol{x}_{2i}\boldsymbol{x}_{2i}^T\right\}^{-1}\left\{\sum_{i=1}^n \boldsymbol{x}_{2i}\mathbb{E}_{\boldsymbol{\theta}^{(v)}}\left(y_i^*\middle|\, y_i\right)\right\},$$

$$\sigma^{(v+1)} = \sqrt{\frac{1}{n}\left\{\sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta}^{(v)}}\left(y_i^{*2}\middle|\, y_i\right) - 2\sum_{i=1}^n \mu_i\mathbb{E}_{\boldsymbol{\theta}^{(v)}}\left(y_i^*\middle|\, y_i\right) + \sum_{i=1}^n \mu_i^2\right\}};$$

(mLQ-4) Iterate steps 2 and 3 until convergence, where the difference in log-likelihood is negligible;

(mLQ-5) The stage 2 optimal rule for the ith subject is estimated as

$$\hat{d}_{2i}^{\text{opt}} = -\text{sgn}\left\{\boldsymbol{x}_{21,i}^T\hat{\boldsymbol{\beta}}_{21}\right\}.$$

Stage 1 estimation: the linear model

(mLQ-6) *Sample surrogate variables*: Sample a surrogate of Y_i^*, denoted by s_i, for

each individual from a truncated normal distribution conditional on y_i,

$$f(y_i^*|y_i) = y_i \text{TN}(\hat{\mu}_i, \hat{\sigma}^2, \xi, \infty) + (1 - y_i)\text{TN}(\hat{\mu}_i, \hat{\sigma}^2, -\infty, \xi),$$

where $\hat{\mu}_i = \boldsymbol{x}_{20,i}^T \hat{\boldsymbol{\beta}}_{20} + a_2 \boldsymbol{x}_{21,i}^T \hat{\boldsymbol{\beta}}_{21}$;

(mLQ-7) *Construct stage 1 pseudo-latent outcomes*:

$$\tilde{Y}_{1i}^* = s_i - 2\mathbb{1}\left\{a_{2i} \neq \hat{d}_{2i}^{\text{opt}}\right\}\left|\boldsymbol{x}_{21,i}^T \hat{\boldsymbol{\beta}}_{21}\right|; \tag{4.7}$$

(mLQ-8) Obtain the least squares estimators $\hat{\boldsymbol{\beta}}_{10}$ and $\hat{\boldsymbol{\beta}}_{11}$ through a linear regression

$$\mathbb{E}\left(\tilde{Y}_{1i}^* \middle| H_{1i}, A_{1i}\right) = \boldsymbol{x}_{10,i}^T \boldsymbol{\beta}_{10} + a_1 \boldsymbol{x}_{11,i}^T \boldsymbol{\beta}_{11};$$

(mLQ-9) The stage 1 optimal rule for the ith subject is estimated as

$$\hat{d}_{1i}^{\text{opt}} = -\text{sgn}\left\{\boldsymbol{x}_{11,i}^T \hat{\boldsymbol{\beta}}_{11}\right\}.$$

The EM algorithm is equivalent to probit regression in the prediction of stage 2 optimal rules, but yields model estimates associated with the expectation of the latent variable. The variance, or the standard error of the estimators can be obtained using Bootstrap.

4.5.3 Simulation Study Revisited

Following the simulation setting in Section 4.3, we add the performance of the proposed mLQ. Figure 4.2 shows that mLQ has a consistent and moderately decent performance across all γ's. When a reasonably small amount of model misspecification is present, specifically, for $\gamma = 1, 2, 3$, mLQ outperforms Q-probit and mQ-probit. However, when a sizable amount of model misspecification is present ($\gamma \geq 4$), mLQ starts to generate

a lower accuracy in estimating stage 1 optimal rules than mQ-probit, but mLQ still improves upon the standard Q-probit algorithm.

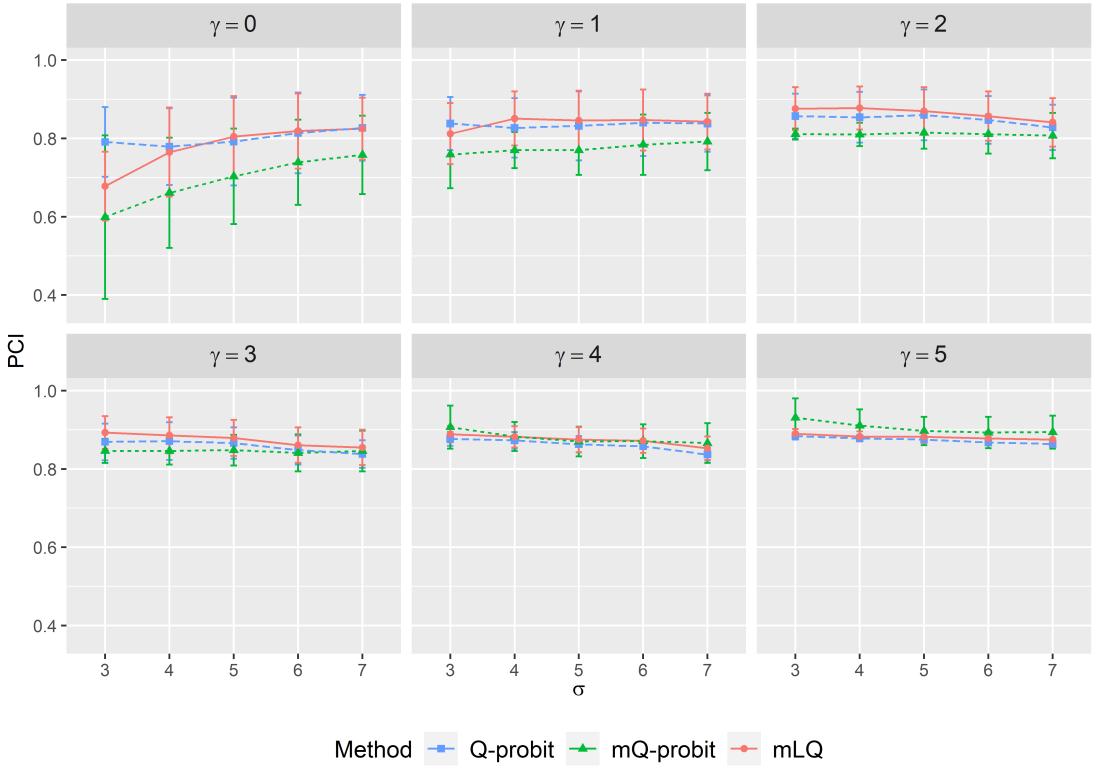

Figure 4.2: Probability of correctly identifying stage 1 optimal rules as a function of σ for different values of γ using (1) Q-learning with probit regression, (2) modified Q-learning with probit regression, and (3) modified latent Q-learning.

The intuition behind the moderate performance when stage 2 main effect model is severely misspecified (or when stage 1 heterogeneous treatment effects of a considerably large size are omitted from stage 2 Q-function), is that the surrogates of Y^* follow a marginal distribution $\sum_y f(y^*|y)f(y)$, where $f(y^*|y)$ has a specified model, and $f(y)$ is observed and consistent with the truth. Thus, as γ increases, the misspecification of the model $f(y^*|y)$ starts to dominate the marginal distribution and reduce the accuracy of estimating stage 1 optimal rules. However, mLQ still has a better performance than

the standard Q-learning algorithm with probit regression due to the incorporation of the observed distribution $f(y)$.

4.5.4 Challenges with the Monotonicity Assumption

The latent variable approach generates estimates for the hidden continuous variable, but we never get to observe them. If the underlying model of the latent variable Y^*, $\mathbb{E}\left(Y^* | H_2, A_2\right)$ is misspecified, then we are not able to consistently estimate $\tilde{Y}_1^*$, as the surrogates of the latent variable, s_i in Equation (4.7), are sampled from a semi-misspecified model. Therefore, it remains a challenge to include the missing piece of information discussed in Section 4.2.2.

Suppose the monotonicity assumption holds. At stage 2, if a subject received a suboptimal treatment and responded with the desirable outcome $Y = 0$, then the pseudo-outcome (conditional on stage 2 optimal rule) must have been $\tilde{Y}_1 = 0$. If a subject received a suboptimal treatment and responded with the undesirable outcome $Y = 1$, then the latent pseudo-outcome $\tilde{Y}_1^*$ should follow a truncated normal distribution with an upper limit Y_i^* for each subject. The list of conditions is written down mathematically below:

- If $d_{2i}^{\text{opt}} = a_{2i}$, then $\tilde{Y}_{1i} = Y_i$;
- If $d_{2i}^{\text{opt}} \neq a_{2i}$ and $Y_i = 0$, then $\tilde{Y}_{1i} = 0$;
- If $d_{2i}^{\text{opt}} \neq a_{2i}$ and $Y_i = 1$, then $\tilde{Y}_{1i}^* < Y_i^*$.

The problem now comes down to the consistent estimation of Y_i^*. Moreover, one direction of future work can be developing an algorithm by sampling directly from the truncated normal distribution with a subject-specific upper limit Y_i^*, and comparing it with mLQ, where the stage 1 latent pseudo-outcome is constructed using surrogates of the unobservable latent variable and Murphy's regret function.

4.6 Discussion

This chapter serves as a first look at main effect model misspecification in Q-learning for dichotomous outcomes, especially in terms of omitting stage 1 heterogeneous treatment effects in the stage 2 model. We proposed modified Q-learning with probit regression that corrects stage 1 pseudo-outcomes by imposing monotonicity-based restrictions. We showed that the proposed approach led to an improved performance in estimating the optimal DTR when model misspecification was considerably large. As the improvement was limited when model misspecification was small due to some loss of information, we explored the feasibility of a latent variable approach to increase sensitivity of the algorithm, but figured that the approach did not solve the problem of systematic model misspecification as the underlying continuous variable was unobservable. Moving forward, we would like to further investigate a way to formalize residuals from the latent variable modeling and understand how the latent variable approach helps with the optimization of DTRs.

Chapter 5

Conclusion

5.1 Summary

As an emerging research area, there exists a rich body of literature discussing frameworks and methods to study DTRs. Q-learning is one of the most popular amongst the reinforcement learning algorithms that were introduced to analyze adaptive interventions. While many researchers dive into more advanced techniques such as machine learning to address certain problems arisen with the implementation of Q-learning, the simpler and more interpretable regression-based Q-learning has somehow been overlooked in recent literature. The assumption that all Q-functions are correctly specified is indeed a strong prerequisite for consistently estimating the optimal DTR, especially for studies where a large number of features is considered. This thesis revisits regression-based Q-learning and develops statistical methods in optimizing DTRs to improve clinical applicability.

This thesis has proposed multiple modifications of Q-learning to address three problems associated with its implementation to optimize DTRs. Here we provide a brief recap. A Q-learning algorithm that posits a marginal model as the Q-function for each stage and estimates model parameters using a generalized estimation equation was proposed to generate individual or marginal trajectory of the optimal DTR that accounts for

all repeated-measures times. One important message conveyed in the simulation study was that, even though the partial correlation between repeated-measures outcomes can be viewed as adding an unmeasured covariate to the expected outcome model, not adjusting for the unmeasured covariate is not harmful to estimation of the optimal rules. This discovery motivated us to further explore the impact of omitted variables on stage 1 rule identification. We found that omission of stage 1 heterogeneous treatment effects in stage 2 Q-function resulted in biased estimation of the stage 1 optimal rules. It is also discussed in literature that stage 2 heterogeneous treatment effects cause bias under misspecified stage 1 linearity due to the optimization operation. To address both sources of bias concomitantly, we proposed a modified interactive Q-learning algorithm, which is robust to both types of model misspecification. It is difficult to extend the implementation of modified Q-learning to optimizing DTRs for dichotomous outcomes due to the non-identity link. Thus, we proposed two alternative approaches to deal with model misspecification in Q-learning with probit/logistic regression. The first approach imposes some restrictions on stage 1 pseudo-outcomes based on monotonicity of preferences, whereas the second approach utilizes a latent variable modeling and samples surrogates of the underlying continuous outcome conditional on the observed dichotomous outcomes. Both approaches cannot obviate specification of stage 2 Q-function, so the proposed methods can only alleviate but not solve the problem.

5.2 Future Work

A major area of future work would be extending modified Q-learning to dichotomous outcomes. In addition to the approaches explored in this thesis, other sampling methods, such as Bayesian techniques, may be applied. However, it might still remain a problem to include late-stage informative residuals into early-stage pseudo-outcomes. It would also be interesting to combine the proposed methods in Chapter 4 and interactive Q-learning for quantiles (Linn et al., 2017) to address the bias associated with

heterogeneous treatment effects for dichotomous outcomes. Moreover, prior to data manipulation, the frequency of binge drinking was originally recorded as an ordinal outcome. The methods proposed in this thesis have the potential to be extended to ordinal outcomes, and more modifications can be explored. One of my takeaways from writing this thesis is there are much to explore in statistics and econometrics literature that can be applied to practical problems in biostatistics. Furthermore, our work in this thesis focuses primarily on point estimation and rule identification. Inference and non-regularity has been a key problem for optimizing DTRs due to the non-smooth absolute value function in the construction of stage 1 pseudo-outcome. The proposed modifications should also be assessed in terms of statistical inference. Finally, a potential research area would be integrating causal inference, optimization of DTRs, and survival analysis to develop methodologies in oncology studies with treatment switching which invalidates randomization and the assumption of no unmeasured confounder. The concept of dynamic treatment regimes and reinforcement learning can be applied to inform the conditions (associated with potential tailoring variables) when patients would benefit most from switching treatment.

Bibliography

Abrevaya, J., Hsu, Y.-C., and Lieli, R. P. (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics* **33,** 485–505.

Aumann, R. J. (1995). Backward induction and common knowledge of rationality. *Games and Economic Behavior* **8,** 6–19.

Bellman, R. (1957). *Dynamic programming.* Princeton University Press, Princeton.

Chakraborty, B. and Moodie, E. E. (2013). *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine.* Springer, New York, NY.

Chakraborty, B. and Murphy, S. A. (2014). Dynamic treatment regimes. *Annual Review of Statistics and Its Application* **1,** 447–464.

Chakraborty, B., Strecher, V., and Murphy, S. (2010). Inference for nonregular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research* **19,** 317–343.

Collins, L. M., Murphy, S. A., and Bierman, K. L. (2004). A conceptual framework for adaptive preventive interventions. *Prevention Science* **5,** 185–196.

Collins, L. M., Murphy, S. A., and Strecher, V. (2007). The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART). *American Journal of Preventive Medicine* **32,** 112–118.

Ertefaie, A., Deng, K., Wagner, A. T., and Murphy, S. A. (2014). qlaci R package for using Q-learning to construct adaptive interventions using data from a SMART

(Version 1.0). *University Park: The Methodology Center, Penn State* .

Ertefaie, A., McKay, J. R., Oslin, D., and Strawderman, R. L. (2021). Robust Q-learning. *Journal of the American Statistical Association* **116,** 368–381.

Finestack, L. H. (2018). Evaluation of an explicit intervention to teach novel grammatical forms to children with developmental language disorder. *Journal of Speech Language and Hearing Research* **61,** 2062-2075.

Fitzmaurice, G. M., Laird, N. M., and Rotnitzky, A. G. (1993). Regression models for discrete longitudinal responses. *Statistical Science* **8,** 284–299.

Fu, S. S., Rothman, A. J., Vock, D. M., Lindgren, B., Almirall, D., Begnaud, A., Melzer, A., Schertz, K., Glaeser, S., Hammett, P., and Joseph, A. M. (2017). Program for lung cancer screening and tobacco cessation: study protocol of a sequential, multiple assignment, randomized trial. *Contemporary Clinical Trials* **60,** 86–95.

Greene, W. H. (2002). Specification analysis and model selection. In *Econometric Analysis*, pages 148–161. Prentice Hall, Upper Saddle River, NJ.

Henderson, R., Ansell, P., and Alshibani, D. (2010). Regret-regression for optimal dynamic treatment regimes. *Biometrics* **66,** 1192–1201.

Huang, X., Choi, S., Wang, L., and Thall, P. F. (2015). Optimization of multi-stage dynamic treatment regimes utilizing accumulated data. *Statistics in Medicine* **34,** 3424–3443.

Kasari, C., Kaiser, A., Goods, K., Nietfeld, J., Mathy, P., Landa, R., Murphy, S. A., and Almirall, D. (2014). Communication interventions for minimally verbal children with autism: a sequential multiple assignment randomized trial. *Journal of the American Academy of Child & Adolescent Psychiatry* **53,** 635–646.

Kosorok, M. R. and Moodie, E. E. (2015). *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Laber, E. B., Linn, K. A., and Stefanski, L. A. (2014). Interactive model building for Q-learning. *Biometrika* **101,** 831–847.

Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E., and Murphy, S. A. (2014). Dynamic treatment regimes: technical challenges and applications. *Electronic Journal of Statistics* **8,** 1225–1272.

Lavori, P. W. and Dawson, R. (2000). A design for testing clinical strategies: biased adaptive within-subject randomization. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **163,** 29–38.

Lavori, P. W. and Dawson, R. (2014). Introduction to dynamic treatment strategies and sequential multiple assignment randomization. *Clinical Trials* **11,** 393–399.

Lei, H., Nahum-Shani, I., Lynch, K., Oslin, D., and Murphy, S. A. (2012). A "SMART" design for building individualized treatment sequences. *Annual Review of Clinical Psychology* **8,** 21–48.

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

Linn, K. A., Laber, E. B., and Stefanski, L. A. (2015). iqLearn: interactive Q-learning in R. *Journal of Statistical Software* **64,** i01.

Linn, K. A., Laber, E. B., and Stefanski, L. A. (2017). Interactive q-learning for quantiles. *Journal of the American Statistical Association* **112,** 638–649.

Liu, D. and Zhang, H. (2018). Residuals and diagnostics for ordinal regression models: A surrogate approach. *Journal of the American Statistical Association* **113,** 845–854.

Lu, X., Nahum-Shani, I., Kasari, C., Lynch, K. G., Oslin, D. W., Pelham, W. E., Fabiano, G., and Almirall, D. (2015). Comparing dynamic treatment regimes using repeated-measures outcomes: modelling considerations in SMART studies. *Statistics in Medicine* **35,** 1595–1615.

McKay, J. R., Drapkin, M. L., Van Horn, D. H. A., Lynch, K. G., Oslin, D. W., DePhilippis, D., Ivey, M., and Cacciola, J. S. (2015). Effect of patient choice in an adaptive sequential randomization trial of treatment for alcohol and cocaine dependence. *Journal of Consulting and Clinical Psychology* **83,** 1021–1032.

Moodie, E. E. M., Dean, N., and Sun, Y. R. (2014). Q-learning: flexible learning about

useful utilities. *Statistics in Biosciences* **6,** 223–243.

Murphy, S. A. (2003). Optimal dynamic treatment regimes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65,** 331–366.

Murphy, S. A. (2005a). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine* **24,** 1455–1481.

Murphy, S. A. (2005b). A generalization error for Q-learning. *Journal of Machine Learning Research* **6,** 1073–1097.

Murphy, S. A., Collins, L. M., and Rush, A. J. (2007). Customizing treatment to the patient: adaptive treatment strategies. *Drug and Alcohol Dependence* **88,** S1–S3.

Murphy, S. A. and McKay, J. R. (2004). Adaptive treatment strategies: an emerging approach for improving treatment effectiveness. *Clinical Science (Newsletter of the American Psychological Association Division 12, Section III: The Society for the Science of Clinical Psychology)* page Winter 2003/Spring 2004.

Murray, T. A., Yuan, Y., and Thall, P. F. (2018). A Bayesian machine learning approach for optimizing dynamic treatment regimes. *Journal of the American Statistical Association* **113,** 1255–1267.

Nahum-Shani, I., Almirall, D., Yap, J. R. T., McKay, J. R., Lynch, K. G., Freiheit, E. A., and Dziak, J. J. (2020). Smart longitudinal analysis: a tutorial for using repeated outcome measures from SMART studies to compare adaptive interventions. *Psychological Methods* **25,** 1–29.

Nahum-Shani, I., Qian, M., Almirall, D., Pelham, W. E., Gnagy, B., Fabiano, G., Waxmonsky, J., Yu, J., and Murphy, S. A. (2012). Q-learning: a data analysis method for constructing adaptive interventions. *Psychological Methods* **17,** 478–494.

Orellana, L., Rotnitzky, A., and Robins, J. M. (2010). Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: Main content. *The International Journal of Biostatistics* **6,** Article 7.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations.

Biometrics **57,** 120–125.

Patrick, M. E., Boatman, J. A., Morrell, N., Wagner, A. C., Lyden, G. R., Nahum-Shani, I., King, C. A., Bonar, E. E., Lee, C. M., Larimer, M. E., Vock, D. M., and Almirall, D. (2020). A sequential multiple assignment randomized trial (SMART) protocol for empirically developing an adaptive preventive intervention for college student drinking reduction. *Contemporary Clinical Trials* **96,** 106089.

Patrick, M. E., Lyden, G. R., Morrell, N., Mehus, C. J., Gunlicks-Stoessel, M., Lee, C. M., King, C. A., Bonar, E. E., Nahum-Shani, I., Almirall, D., Larimer, M. E., and Vock, D. M. (in press). Main outcomes of M-bridge: a sequential multiple assignment randomized trial (SMART) for developing an adaptive preventive intervention for college drinking. *Journal of Consulting and Clinical Psychology* .

Robins, J. M. (2000a). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11,** 550–560.

Robins, J. M. (2000b). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology, the Environment and Clinical Trials*, pages 95–134. Springer-Verlag, New York, NY.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium on Biostatistics*, pages 189–326. Springer, New York, NY.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014). Q- and A-learning methods for estimating optimal dynamic treatment regimes. *Statstical Science* **29,** 640–661.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models*. Interdisciplinary statistics. Chapman & Hall/CRC, Boca Raton.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: an introduction*. MIT Press, Cambridge.

Watkins, C. (1989). *Learning from delayed rewards.* PhD thesis, King's College, University of Cambridge, UK.

Wodtke, G. T. and Almirall, D. (2017). Estimating moderated causal effects with time-varying treatments and time-varying moderators: structural nested mean models and regression with residuals. *Sociological Methodology* **47,** 212–245.

Xin, J., Chakraborty, B., and Laber, E. B. (2012). qLearn: estimation and inference for Q-learning. *R package version* **1,** 87.

Zeger, S. L. and Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42,** 121–130.

Zhao, L. P., Prentice, R. L., and Self, S. G. (1992). Multivariate mean parameter estimation by using a partly exponential model. *Journal of the Royal Statistical Society: Series B (Methodological)* **54,** 805–811.

Zheng, B. (2000). Summarizing the goodness of fit of generalized linear models for longitudinal data. *Statistics in Medicine* **19,** 1265–1275.

Appendix A

Supplementary Materials for mQ-GEE

A.1 Inference for mQ-GEE

The estimators obtained from (modified) Q-learning with GEE are M-estimators, and the asymptotics closely follow the inference for M-estimators under the restriction of $\mathbb{P}(\boldsymbol{w}_2^T X_{21}\boldsymbol{\psi}_2 = 0) = 0$.

The Q-learning with GEE estimators, $\hat{\boldsymbol{\beta}} = \begin{pmatrix}\hat{\boldsymbol{\beta}}_1^T & \hat{\boldsymbol{\beta}}_2^T\end{pmatrix}^T$ and $\hat{\boldsymbol{\psi}} = \begin{pmatrix}\hat{\boldsymbol{\psi}}_1^T & \hat{\boldsymbol{\psi}}_2^T\end{pmatrix}^T$, are solutions to

$$\sum_{i=1}^{n} \boldsymbol{\rho}\left(\{\boldsymbol{Y}_{ki}, X_{ki}, A_{ki}\}_{k=1,2}; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}\right) = \mathbf{0},$$

where $\boldsymbol{\rho}\left(\{\boldsymbol{Y}_{ki}, X_{ki}, A_{ki}\}_{k=1,2}; \boldsymbol{\beta}, \boldsymbol{\psi}\right) = \begin{bmatrix} D_2^T V_2^{-1}\left\{\boldsymbol{Y}_{2i} - (X_{20,i}\boldsymbol{\beta}_2 + A_{2i}X_{21,i}\boldsymbol{\psi}_2)\right\} \\ D_1^T V_1^{-1}\left\{\begin{pmatrix}\boldsymbol{Y}_{1i} \\ \boldsymbol{Y}_{2i}^{\text{opt}}\end{pmatrix} - (X_{10,i}\boldsymbol{\beta}_1 + A_{1i}X_{11,i}\boldsymbol{\psi}_1)\right\} \end{bmatrix}$,

and $\boldsymbol{Y}_{2i}^{\text{opt}} = X_{20,i}\boldsymbol{\beta}_2 - \text{sgn}\{\boldsymbol{w}_2^T X_{21,i}\boldsymbol{\psi}_2\}X_{21,i}\boldsymbol{\psi}_2$.

Since $\text{sgn}(x) = \begin{cases} -1 & x < 0 \\ 1 & \text{otherwise} \end{cases}$ is discontinuous and non-differentiable at $x = 0$,

$\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\psi}}_1$ lose consistency and asymptotic normality under nonregular conditions at $\boldsymbol{w}_2^T X_{21,i} \boldsymbol{\psi}_2 = 0$.

A.2 The DLD Study

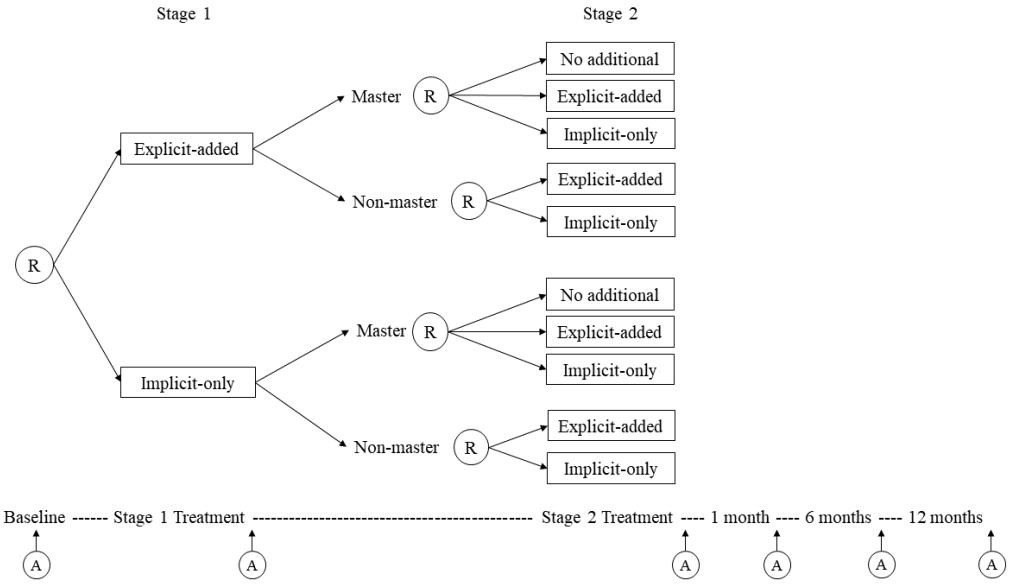

Figure A.1: A sequential, multiple assignment, randomized trial for children with developmental language disorder (DLD). (R) indicates state of randomization with arrows pointing to available treatment options, and (A) indicates time of assessment where measurements of intermediate or final outcome take place.

The DLD study is proposed to evaluate a sequence of treatments for children with developmental language disorder (Finestack, 2018). The study uses an embedded tailoring variable based on a child's performance following an initial stage of treatment. Upon recruitment, children will be randomized between two treatment arms: "Implicit-only," in which children are only provided with models and recasts of problematic forms at a high frequency, and "Explicit-added," where the pattern or pedagogic rule is directly presented. After stage 1 treatment, children are categorized as "Master" or "Non-master"

based on their performance in target forms designed by the investigator. Masters will be re-randomized to stop ("No additional"), boost (i.e., continue stage 1 treatment), or switch to the alternative treatment, whereas non-masters will be randomized to boost or switch. Similar to the M-bridge study, sustained performance after stage 2 is clinically important; hence, the investigators also assess children's performance at 1, 6, and 12 months after the treatment period.

A.3 Marginalization over an Unmeasured Covariate

Proposition 1. *Suppose each element in a sequence of repeated-measures outcomes* (Y_1, Y_2, Y_3) *follows an independent normal distribution,* $Y_{ij} \sim N(\mu_j, \sigma_e^2)$ *for subjects* $i = 1, \ldots, n$ *and* $j = 1, 2, 3$. *Suppose the true mean* μ_j *is associated with a measurable* $\tilde{\mu}_j$ *and an additional unmeasured covariate* $V_i \sim N(0, \sigma_v^2)$ *and* $\mu_j = \tilde{\mu}_j + \lambda_j V_i$. *Then the joint distribution of* Y_1, Y_2, Y_3 *conditional on measurable* $\tilde{\mu}_1, \tilde{\mu}_2, \tilde{\mu}_3$ *is a multivariate normal distribution with* $\tilde{\boldsymbol{\mu}} = \begin{pmatrix} \tilde{\mu}_1 & \tilde{\mu}_2 & \tilde{\mu}_3 \end{pmatrix}^T$ *and*

$$\Sigma = \begin{pmatrix} \lambda_1^2\sigma_v^2 + \sigma_e^2 & \lambda_1\lambda_2\sigma_v^2 & \lambda_1\lambda_3\sigma_v^2 \\ \lambda_1\lambda_2\sigma_v^2 & \lambda_2^2\sigma_v^2 + \sigma_e^2 & \lambda_2\lambda_3\sigma_v^2 \\ \lambda_1\lambda_3\sigma_v^2 & \lambda_2\lambda_3\sigma_v^2 & \lambda_3^2\sigma_v^2 + \sigma_e^2 \end{pmatrix}.$$

Proof. Since $\boldsymbol{Y} = \begin{pmatrix} Y_1 & Y_2 & Y_3 \end{pmatrix}^T \sim MVN\left(\boldsymbol{\mu}, \sigma_e^2 I_3\right)$, where $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 & \mu_2 & \mu_3 \end{pmatrix}^T$ and I_3 is a 3×3 identity matrix, the joint distribution of (Y_1, Y_2, Y_3) conditional on $\tilde{\boldsymbol{\mu}}$ is

$$\begin{aligned} f(Y_1, Y_2, Y_3|\tilde{\boldsymbol{\mu}}) &= \int_{-\infty}^{\infty} f(Y_1, Y_2, Y_3|\tilde{\boldsymbol{\mu}}, V) f(V|\tilde{\boldsymbol{\mu}}) dV \\ &= \int_{-\infty}^{\infty} (2\pi\sigma_e^2)^{-3/2} \exp\left\{-\frac{1}{2\sigma_e^2}(\boldsymbol{Y} - \boldsymbol{\mu})^T(\boldsymbol{Y} - \boldsymbol{\mu})\right\} (2\pi\sigma_v^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma_v^2}V^2\right\} dV \\ &= \int_{-\infty}^{\infty} (2\pi)^{-2}\sigma_e^{-3}\sigma_v^{-1} \exp\left\{-\frac{1}{2\sigma_e^2}(\boldsymbol{Y} - \boldsymbol{\mu})^T(\boldsymbol{Y} - \boldsymbol{\mu}) - \frac{1}{2\sigma_v^2}V^2\right\} dV, \end{aligned}$$

where

$$(\boldsymbol{Y}-\boldsymbol{\mu})^T(\boldsymbol{Y}-\boldsymbol{\mu}) = (\lambda_1^2+\lambda_2^2+\lambda_3^2)V^2 - 2V\left\{\lambda_1(Y_1-\tilde{\mu}_1)+\lambda_2(Y_2-\tilde{\mu}_2)+\lambda_3(Y_3-\tilde{\mu}_3)\right\} + (Y_1-\tilde{\mu}_1)^2+(Y_2-\tilde{\mu}_2)^2+(Y_3-\tilde{\mu}_3)^2.$$

Thus,

$$\begin{aligned} f(Y_1,Y_2,Y_3|\tilde{\boldsymbol{\mu}}) &= \int_{-\infty}^{\infty}(2\pi)^{-2}\sigma_e^{-3}\sigma_v^{-1}\exp\left\{-\frac{1}{2}(aV^2-2bV+c)\right\}dV \\ &= (2\pi)^{-3/2}\sigma_e^{-3}\sigma_v^{-1}a^{-1/2}\exp\left\{-\frac{1}{2}\underbrace{(c-a^{-1}b^2)}_{\zeta}\right\}, \end{aligned} \tag{A.1}$$

where

$$\begin{aligned} a &= \sigma_e^{-2}(\lambda_1^2+\lambda_2^2+\lambda_3^2)+\sigma_v^{-2} \\ b &= \sigma_e^{-2}\left\{\lambda_1(Y_1-\tilde{\mu}_1)+\lambda_2(Y_2-\tilde{\mu}_2)+\lambda_3(Y_3-\tilde{\mu}_3)\right\} \\ c &= \sigma_e^{-2}\left\{(Y_1-\tilde{\mu}_1)^2+(Y_2-\tilde{\mu}_2)^2+(Y_3-\tilde{\mu}_3)^2\right\}. \end{aligned}$$

The function indicated by ζ in Equation (A.1) can be expanded and simplified as

$$\begin{aligned} \zeta &= \sigma_e^{-2}\left\{Y_1^2+Y_2^2+Y_3^2-2\tilde{\mu}_1Y_1-2\tilde{\mu}_2Y_2-2\tilde{\mu}_3Y_3-\frac{\left(\lambda_1Y_1+\lambda_2Y_2+\lambda_3Y_3-\lambda_1\tilde{\mu}_1-\lambda_2\tilde{\mu}_2-\lambda_3\tilde{\mu}_3\right)^2}{\lambda_1^2+\lambda_2^2+\lambda_3^2+\sigma_e^2\sigma_v^{-2}}\right\}+C \\ &= \sigma_e^{-2}\left[\sum_{i=1}^{3}\left\{1-\left(\lambda_1^2+\lambda_2^2+\lambda_3^2+\sigma_e^2\sigma_v^{-2}\right)^{-1}\lambda_i^2\right\}Y_i^2-2\sum_{i=1}^{3}\left\{\tilde{\mu}_i-\left(\lambda_1^2+\lambda_2^2+\lambda_3^2+\sigma_e^2\sigma_v^{-2}\right)^{-1}\lambda_i\sum_{j=1}^{3}\lambda_j\tilde{\mu}_j\right\}Y_i \right. \\ &\left. -\left(\lambda_1^2+\lambda_2^2+\lambda_3^2+\sigma_e^2\sigma_v^{-2}\right)^{-1}\sum_{i=1,i\neq j}^{3}\sum_{j=1}^{3}\lambda_i\lambda_jY_iY_j\right]+C. \end{aligned}$$

Let $\omega = \left(\lambda_1^2+\lambda_2^2+\lambda_3^2+\sigma_e^2\sigma_v^{-2}\right)^{-1}$. Therefore, by the definition of multivariate normal distribution, the distribution of $\boldsymbol{Y}$ conditional on measurable covariates has

mean $\tilde{\boldsymbol{\mu}}$ and precision matrix

$$\Sigma^{-1} = \sigma_e^{-2} \begin{pmatrix} 1-\omega\lambda_1^2 & -\omega\lambda_1\lambda_2 & -\omega\lambda_1\lambda_3 \\ -\omega\lambda_1\lambda_2 & 1-\omega\lambda_2^2 & -\omega\lambda_2\lambda_3 \\ -\omega\lambda_1\lambda_3 & -\omega\lambda_2\lambda_3 & 1-\omega\lambda_3^2 \end{pmatrix}.$$

Finally, we obtain the desired covariance matrix by inverting the precision matrix. □

A.4 Model Misspecifications in the Simulation Study

To apply Q-learning algorithms with GEE, we specify the Q-functions as

for $t_1 \in \{1,2,3\}$ and $t_2 \in \{2,3\}$,

$$\begin{aligned} Q_2^{\text{mis}}(H_2, A_2, t_2; \boldsymbol{\beta}_2, \boldsymbol{\psi}_2) &= \beta_{20,t_2} + \beta_{21,t_2} Z_1 + \beta_{22,t_2} A_1 + \beta_{23,t_2} Y_1 \\ &\quad + A_2 \left(\psi_{20,t_2} + \psi_{21,t_2} Z_1 + \psi_{22,t_2} A_1 + \psi_{23,t_2} Y_1\right), \\ Q_2(H_2, A_2, t_2; \boldsymbol{\beta}_2, \boldsymbol{\psi}_2) &= \beta_{20,t_2} + \beta_{21,t_2} Z_1 + \beta_{22,t_2} A_1 + \beta_{23,t_2} Z_1 A_1 + \beta_{24,t_2} Y_1 \\ &\quad + A_2 \left(\psi_{20,t_2} + \psi_{21,t_2} Z_1 + \psi_{22,t_2} A_1 + \psi_{23,t_2} Y_1\right), \\ Q_1(H_1, A_1, t_1; \boldsymbol{\beta}_1, \boldsymbol{\psi}_1) &= \beta_{10,t_1} + \beta_{11,t_1} Z_1 + A_1 \left(\psi_{10,t_1} + \psi_{11,t_1} Z_1\right); \end{aligned}$$

whereas on the other hand, composite Q-learning fits

$$\begin{aligned} Q_2^{\text{mis}}(H_2, A_2, t_2; \boldsymbol{\beta}_2, \boldsymbol{\psi}_2) &= \beta_{20} + \beta_{21} Z_1 + \beta_{22} A_1 + \beta_{23} Y_1 \\ &\quad + A_2 \left(\psi_{20} + \psi_{21} Z_1 + \psi_{22} A_1 + \psi_{23} Y_1\right), \\ Q_2(H_2, A_2, t_2; \boldsymbol{\beta}_2, \boldsymbol{\psi}_2) &= \beta_{20} + \beta_{21} Z_1 + \beta_{22} A_1 + \beta_{23} Z_1 A_1 + \beta_{24} Y_1 \\ &\quad + A_2 \left(\psi_{20} + \psi_{21} Z_1 + \psi_{22} A_1 + \psi_{23} Y_1\right), \\ Q_1(H_1, A_1, t_1; \boldsymbol{\beta}_1, \boldsymbol{\psi}_1) &= \beta_{10} + \beta_{11} Z_1 + A_1 \left(\psi_{10} + \psi_{11} Z_1\right). \end{aligned}$$

A.5 Relative Efficiency Using Different Working Correlations

Misspecification of working correlation structures in GEE can cause loss in efficiency. In this section, we assess the relative efficiency of using "unstructured", "exchangeable" and "independent" working correlations in stage 1 estimation. Estimates are the same under different correlation structures due to the inclusion of time-varying coefficients, but standard errors are different. So we use variance instead of MSE to quantify efficiency. Table A.1 summarizes the ratio of variances of $\hat{\tau}_{j1}(0)$ at time j, $j = 1, 2, 3$, under different partial correlations. We assume there is not model misspecification in this simulation. The λ's specified in each scenario is summarized below:

1. Low positive partial correlation: $\lambda_1 = 1, \lambda_2 = 0.2, \lambda_3 = 0.2$
2. Medium positive partial correlation: $\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 1$
3. High positive partial correlation: $\lambda_1 = 1, \lambda_2 = 3, \lambda_3 = 1$
4. Low negative partial correlation: $\lambda_1 = 1, \lambda_2 = -0.2, \lambda_3 = -0.2$
5. Medium negative partial correlation: $\lambda_1 = 1, \lambda_2 = -1, \lambda_3 = -1$
6. High negative partial correlation: $\lambda_1 = 1, \lambda_2 = -3, \lambda_3 = -1$
7. Zero partial correlation: $\lambda_1 = 0, \lambda_2 = 0, \lambda_3 = 0$

Note that "exchangeable" working correlation is the best choice for positive correlations; "unstructured" working correlation is the best choice for negative correlations; "independence" working correlation is the best choice for independent correlations.

Table A.1 shows that the relative efficiency does not change much across different working correlations. For high positive partial correlations, "exchangeable" working correlation seems to work slightly better than the other two options. For high negative

partial correlations, "unstructured" working correlation work better than the other two options for Q-GEE, but this pattern is not observed for mQ-GEE. An intuitive explanation of this result is that the sample size $n = 200$ is relatively large compared to the cluster size.

Partial Correlation		Working Correlation	Q-GEE Time 1	Q-GEE Time 2	Q-GEE Time 3	mQ-GEE Time 1	mQ-GEE Time 2	mQ-GEE Time 3
Positive	Low	UN/IN	0.96	0.96	0.96	0.89	1.02	0.95
		EX/IN	1.03	0.94	1.02	0.94	1.00	0.98
	Med	UN/IN	0.95	0.98	1.08	0.88	1.03	0.89
		EX/IN	1.01	0.96	1.04	0.93	1.05	0.92
	High	UN/IN	0.97	1.06	1.04	1.02	0.88	0.89
		EX/IN	0.95	0.96	0.99	0.98	0.91	0.94
Negative	Low	UN/IN	1.01	0.94	0.92	1.13	1.01	1.04
		EX/IN	1.01	0.96	1.05	1.09	0.97	1.05
	Med	UN/IN	0.99	0.95	0.90	0.93	1.05	1.09
		EX/IN	0.95	0.86	0.97	0.99	0.96	1.01
	High	UN/IN	0.84	0.93	0.93	0.98	1.06	1.08
		EX/IN	0.92	1.10	1.04	0.97	1.01	1.10
Independent		UN/IN	0.95	1.04	1.05	0.95	0.99	1.06
		EX/IN	0.91	0.97	1.03	0.91	1.01	1.02

Table A.1: Relative efficiency under correlation structures (1) unstructured (UN) and (2) exchangeable (EX) versus independence (IN) respectively, based on a sample size of $n = 200$ and 1000 simulations.

A.6 Simulation Study: Parsimonious Models

In Section 2.4, the simulation study considers a saturated model with time-varying coefficients. There is a tiny loss of accuracy in the estimation of stage 1 optimal rules using Q-learning with GEE, and the loss pattern is consistent across different partial correlations. One reasonable speculation of this finding is that the existence of additional time-varying coefficients in the marginal model inflates the estimation errors. We

would like to investigate if Q-learning with GEE can possibly outperform standard Q-learning when parsimonious model is considered. Suppose both Q-learning with GEE and composite Q-learning fit parsimonious models with time-invariant coefficients:

$$\begin{aligned}
Q_2^{\text{mis}}(H_2, A_2, t_2; \boldsymbol{\beta}_2, \boldsymbol{\psi}_2) &= \beta_{20} + \beta_{21} Z_1 + \beta_{22} A_1 + \beta_{23} Y_1 \\
&\quad + A_2 \left(\psi_{20} + \psi_{21} Z_1 + \psi_{22} A_1 + \psi_{23} Y_1\right), \\
Q_2(H_2, A_2, t_2; \boldsymbol{\beta}_2, \boldsymbol{\psi}_2) &= \beta_{20} + \beta_{21} Z_1 + \beta_{22} A_1 + \beta_{23} Z_1 A_1 + \beta_{24} Y_1 \\
&\quad + A_2 \left(\psi_{20} + \psi_{21} Z_1 + \psi_{22} A_1 + \psi_{23} Y_1\right), \\
Q_1(H_1, A_1, t_1; \boldsymbol{\beta}_1, \boldsymbol{\psi}_1) &= \beta_{10} + \beta_{11} Z_1 + A_1 \left(\psi_{10} + \psi_{11} Z_1\right).
\end{aligned}$$

Consider $(\alpha_1, \alpha_2, \alpha_3, \gamma_2, \gamma_3) = -(1.6, 1.6, 1.6, 1.2, 1.2)$ and rerun the simulation study with the parsimonious models. In both scenarios, Q-GEE performs slightly better than composite Q-learning for negatively (conditionally) correlated outcomes, but performs much worse for positively correlated outcomes.

Partial Correlation	Method	PCI_1	PCI_2	$\text{RMSE}_{\tau_{11}}$	$\text{RMSE}_{\tau_{12}}$	$\text{RMSE}_{\tau_{13}}$
Positive	Q	0.943	0.992	-	-	-
	mQ	0.943	0.992	-	-	-
	Q-GEE	0.888	0.992	3.18 (25.6)	3.18 (25.6)	3.18 (25.6)
	mQ-GEE	0.941	0.992	0.72 (0.38)	0.72 (0.38)	0.72 (0.38)
Independent	Q	0.973	0.999	-	-	-
	mQ	0.973	0.999	-	-	-
	Q-GEE	0.972	0.999	0.36 (0.19)	0.36 (0.19)	0.36 (0.19)
	mQ-GEE	0.972	0.999	0.36 (0.19)	0.36 (0.19)	0.36 (0.19)
Negative	Q	0.966	0.993	-	-	-
	mQ	0.966	0.993	-	-	-
	Q-GEE	0.967	0.993	0.40 (0.21)	0.40 (0.21)	0.40 (0.21)
	mQ-GEE	0.967	0.993	0.41 (0.21)	0.41 (0.21)	0.41 (0.21)

Table A.2: (Scenario I) PCI of stage 1 optimal rules and RMSE (mean (SD)) of estimated heterogeneous causal effects at time 2 and 3, based on estimated stage 1 parsimonious Q-functions from 1000 simulations with sample size $n = 200$.

Partial Correlation	Method	PCI_1	PCI_2	$RMSE_{\tau_{11}}$	$RMSE_{\tau_{12}}$	$RMSE_{\tau_{13}}$
Positive	Q	0.930	0.996	-	-	-
	mQ	0.945	0.996	-	-	-
	Q-GEE	0.871	0.996	3.47 (26.8)	3.47 (26.8)	3.47 (26.8)
	mQ-GEE	0.943	0.996	0.71 (0.37)	0.71 (0.37)	0.71 (0.37)
Independent	Q	0.943	0.998	-	-	-
	mQ	0.969	0.998	-	-	-
	Q-GEE	0.911	0.998	1.93 (3.33)	1.93 (3.33)	1.93 (3.33)
	mQ-GEE	0.968	0.998	0.42 (0.22)	0.42 (0.22)	0.42 (0.22)
Negative	Q	0.782	0.977	-	-	-
	mQ	0.960	0.977	-	-	-
	Q-GEE	0.824	0.977	2.39 (0.40)	2.39 (0.40)	2.39 (0.40)
	mQ-GEE	0.962	0.977	0.49 (0.25)	0.49 (0.25)	0.49 (0.25)

Table A.3: (Scenario II) PCI of stage 1 optimal rules and RMSE of estimated heterogeneous causal effects at time 2 and 3, based on estimated stage 1 parsimonious Q-functions from 1000 simulations.

A.7 Application: Model Fit and Variable Selection

We performed constrained variable selection based on QIC for both stage 1 and stage 2 saturated models (with time-varying coefficients), where the time-dependent effects of A_1 are forced in the stage 1 model and the time-dependent effects of A_2 and $A_1 * A_2$ are forced in the stage 2 model. Table A.4 shows a summary of the estimated parameters using mQ-GEE, assuming equal weights for all three outcome measurements.

Table A.4: Summary of stage 1 model and stage 2 model with variable selection.

	Estimate	SE	p-value[1]
Stage 1 Model			
Main Effects			

Continued on next page

[1] * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.4 – *Continued from previous page*

	Estimate	SE	p-value
binge_bl[2]	0.298	0.118	0.012 *
race(nonwhite)	-0.415	0.242	0.086
time2	-0.381	0.146	0.009 **
time3	0.038	0.158	0.810
avg_drinks_bl[3]×time1	0.523	0.115	<0.001 ***
avg_drinks_bl×time2	0.185	0.053	<0.001 ***
avg_drinks_bl×time3	0.172	0.072	0.017 *
num_days_bl[4]×time1	0.072	0.066	0.270
num_days_bl×time2	0.007	0.035	0.850
num_days_bl×time3	0.092	0.055	0.092
intent_greek(yes)×time1	0.585	0.438	0.181
intent_greek(yes)×time2	-0.136	0.201	0.499
intent_greek(yes)×time3	-0.382	0.297	0.198
race(nonwhite)×time2	0.407	0.231	0.078
race(nonwhite)×time3	0.070	0.245	0.775
Treatment Effects			
A1×time1	-0.192	0.182	0.290
A1×time2	-0.068	0.124	0.583
A1×time3	-0.069	0.140	0.619
A1×gender(female)	0.105	0.142	0.460
A1×intent_greek(yes)×time1	-0.465	0.411	0.258
A1×intent_greek(yes)×time2	-0.154	0.197	0.433
A1×intent_greek(yes)×time3	0.253	0.287	0.377
A1×race(nonwhite)×time1	0.287	0.239	0.229
A1×race(nonwhite)×time2	-0.032	0.143	0.824
A1×race(nonwhite)×time3	0.076	0.169	0.655

Stage 2 Model

Main Effects

Continued on next page

[2]Frequency of binge drinking during the last 30 days at baseline
[3]Average number of drinks consumed within a 24-hour period during the last 30 days at baseline
[4]Number of days using alcohol during the last 30 days at baseline

Table A.4 – *Continued from previous page*

	Estimate	SE	p-value
binge_bl	0.164	0.1224	0.181
gender(female)	-0.674	0.2647	0.011 *
intent_greek(yes)	-0.661	0.3530	0.061
race(nonwhite)	0.557	0.5154	0.280
avg_drinks_bl	0.075	0.0678	0.272
binge_sm[5]×time2	0.122	0.0594	0.040 *
binge_sm×time3	0.184	0.0646	0.004 **
num_days_bl×time2	0.044	0.0592	0.457
num_days_bl×time3	0.117	0.0702	0.096
Treatment Effects			
A2	-0.693	0.3320	0.037 *
A2×binge_sm	0.116	0.0555	0.036 *
A2×num_days_bl	0.105	0.0513	0.041 *
A2×avg_drinks_bl	0.066	0.0681	0.335
A2×A1	-0.135	0.1614	0.401
A2×time3	-0.041	0.2357	0.862
A2×race(nonwhite)×time2	-0.734	0.5305	0.166
A2×race(nonwhite)×time3	0.825	0.6751	0.222
A2×binge_bl×time2	-0.242	0.1307	0.064
A2×binge_bl×time3	-0.371	0.1402	0.008 **
A2×A1×time3	-0.013	0.2213	0.955

[5]Frequency of binge drinking during the last 30 days at self-monitoring (stage 1)

Appendix B

Supplementary Materials for mIQ

B.1 Proof

Theorem 3.3.1 (Matrix Version of the Omitted Variable Bias Theorem)**.** *Suppose that the true regression model for Y is $Y = \psi_0 + \boldsymbol{X}^T\boldsymbol{\psi}_1 + \boldsymbol{V}^T\boldsymbol{\gamma} + \varepsilon$, where $\boldsymbol{X}$ is a random vector formed by measured covariates, $\boldsymbol{V}$ is formed by unmeasured covariates, and $\varepsilon \sim N(0, \sigma^2)$. The parameters associated with measured covariates, $\boldsymbol{\psi} \equiv \begin{pmatrix} \psi_0 \\ \boldsymbol{\psi}_1 \end{pmatrix}$, are thus estimated via the misspecified model $y = \beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}_1 + \varepsilon^*$, where y and $\boldsymbol{x}$ are realizations of Y and $\boldsymbol{X}$ respectively. Then*

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) \equiv \mathbb{E}\begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_1 \end{pmatrix} = \boldsymbol{\psi} + \begin{pmatrix} \mathbb{E}(\boldsymbol{V}^T) - \mathbb{E}(\boldsymbol{X}^T)\,Cov(\boldsymbol{X})^{-1}\,Cov(\boldsymbol{X}, \boldsymbol{V}) \\ Cov(\boldsymbol{X})^{-1}\,Cov(\boldsymbol{X}, \boldsymbol{V}) \end{pmatrix} \boldsymbol{\gamma}.$$

Theorem 3.3.1 is developed based on the *omitted variable formula* and example in *Econometric Analysis* by Greene (2002).

Proof. Let $\mathbf{X}$ denote a design matrix with $\boldsymbol{x}_i^T$ as the ith row, $i = 1, \ldots, n$ and $\boldsymbol{y}$ denote a vector with y_i as the ith element. Suppose that $\boldsymbol{V}$ is observable and let $\mathbf{V}$ denote

a matrix with $\boldsymbol{v}_i$, a realization of $\boldsymbol{V}$, as the ith row. Define $\tilde{\mathbf{X}} = \begin{pmatrix} \mathbf{1} & \mathbf{X} \end{pmatrix}$. The least squares estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\boldsymbol{y}$.

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\psi} + \mathbb{E}\left[(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathbf{V}\right]\boldsymbol{\gamma}.$$

Consider the multivariate regression $\boldsymbol{V} = \boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}_1\boldsymbol{x} + \boldsymbol{\varepsilon}^{**}$, where $\boldsymbol{\varepsilon}^{**} \sim MVN(\mathbf{0}, \Sigma)$. The least squares estimator of $\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\lambda}_0^T \\ \boldsymbol{\Lambda}_1^T \end{pmatrix}$ is $\hat{\boldsymbol{\Lambda}} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathbf{V}$. Hence, $\mathbb{E}\left[(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathbf{V}\right] = \mathbb{E}(\hat{\boldsymbol{\Lambda}}) = \boldsymbol{\Lambda}$ and $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\psi} + \boldsymbol{\Lambda}\boldsymbol{\gamma}$. Now, it remains to prove that

$$\boldsymbol{\Lambda} = \begin{pmatrix} \mathbb{E}(\boldsymbol{V}^T) - \mathbb{E}(\boldsymbol{X}^T)\mathrm{Cov}(\boldsymbol{X})^{-1}\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{V}) \\ \mathrm{Cov}(\boldsymbol{X})^{-1}\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{V}) \end{pmatrix}.$$

The true values of $\boldsymbol{\Lambda}$ satisfy the equation

$$(\boldsymbol{\lambda}_0, \boldsymbol{\Lambda}_1) = \arg\min_{\boldsymbol{\Lambda}} \mathbb{E}(\boldsymbol{V} - \boldsymbol{\lambda}_0 - \boldsymbol{\Lambda}_1\boldsymbol{X})^T(\boldsymbol{V} - \boldsymbol{\lambda}_0 - \boldsymbol{\Lambda}_1\boldsymbol{X}) = \arg\min_{\boldsymbol{\Lambda}} L.$$

The first derivative of L with respect to $(\boldsymbol{\lambda}_0, \boldsymbol{\Lambda}_1)$ is

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{\lambda}_0} &= -2\mathbb{E}(\boldsymbol{V}^T) + 2\boldsymbol{\lambda}_0^T + 2\mathbb{E}(\boldsymbol{X}^T)\boldsymbol{\Lambda}_1^T, \\ \frac{\partial L}{\partial \boldsymbol{\Lambda}_1} &= -2\mathbb{E}(\boldsymbol{X}\boldsymbol{V}^T) + 2\mathbb{E}(\boldsymbol{X})\boldsymbol{\lambda}_0^T + 2\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^T)\boldsymbol{\Lambda}_1^T. \end{aligned}$$

Hence, the score equations of $(\boldsymbol{\lambda}_0, \boldsymbol{\Lambda}_1)$ are

$$\begin{aligned} S(\boldsymbol{\lambda}_0) &= \boldsymbol{\lambda}_0 - \mathbb{E}(\boldsymbol{V}^T) + \mathbb{E}(\boldsymbol{X}^T)\boldsymbol{\Lambda}_1^T = \mathbf{0}, \\ S(\boldsymbol{\Lambda}_1) &= \left[\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^T) - \mathbb{E}(\boldsymbol{X})\mathbb{E}(\boldsymbol{X}^T)\right]\boldsymbol{\Lambda}_1^T - \left[\mathbb{E}(\boldsymbol{X}\boldsymbol{V}^T) - \mathbb{E}(\boldsymbol{X})\mathbb{E}(\boldsymbol{V}^T)\right] \\ &= \mathrm{Cov}(\boldsymbol{X})\boldsymbol{\Lambda}_1^T - \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{V}) = \mathbf{0} \end{aligned}$$

Therefore, $\boldsymbol{\Lambda}_1^T = \mathrm{Cov}(\boldsymbol{X})^{-1}\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{V})$ and $\boldsymbol{\lambda}_0^T = \mathbb{E}(\boldsymbol{V}^T) - \mathbb{E}(\boldsymbol{X}^T)\mathrm{Cov}(\boldsymbol{X})^{-1}\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{V})$.

□

Theorem 3.3.2 (Bias of Stage 2 Treatment Effect Estimators)**.** *Assume that $\boldsymbol{V}_{20}$ is a vector of unmeasured covariates that are independent of A_2 and $Cov(\boldsymbol{X}_2)$ is invertible. The estimators of stage 2 heterogeneous treatment effects are unbiased if and only if at least one of the following conditions is satisfied:*

- $\mathbb{E}(A_2) = 0$*;*
- $\boldsymbol{V}_{20}$ *is correlated with neither* $\boldsymbol{X}_{20}$ *nor* $\boldsymbol{X}_{21}$*.*

Proof. By Theorem 3.3.1, we know that the bias of $\hat{\boldsymbol{\beta}}_2 = \begin{pmatrix} \hat{\boldsymbol{\beta}}_{201}^T & \hat{\beta}_{210} & \hat{\boldsymbol{\beta}}_{211}^T \end{pmatrix}^T$ is

$$\text{Bias}(\hat{\boldsymbol{\beta}}_2) = \text{Cov}(\boldsymbol{X}_2)^{-1}\text{Cov}(\boldsymbol{X}_2, \boldsymbol{V}_{20})\boldsymbol{\gamma}_{20} = \text{Cov}(\boldsymbol{X}_2)^{-1} \begin{pmatrix} \text{Cov}(\boldsymbol{X}_{20}, \boldsymbol{V}_{20}) \\ \boldsymbol{0}^T \\ \mathbb{E}(A_2)\text{Cov}(\boldsymbol{X}_{21}, \boldsymbol{V}_{20}) \end{pmatrix} \boldsymbol{\gamma}_{20}.$$

Suppose $\mathbb{E}(A_2) \neq 0$. Since $\text{Cov}(\boldsymbol{X}_2)$ is invertible, we have

$$\text{Bias}(\hat{\boldsymbol{\beta}}_2) = \boldsymbol{0} \iff \begin{pmatrix} \text{Cov}(\boldsymbol{X}_{20}, \boldsymbol{V}_{20}) \\ \boldsymbol{0}^T \\ \mathbb{E}(A_2)\text{Cov}(\boldsymbol{X}_{21}, \boldsymbol{V}_{20}) \end{pmatrix} = \boldsymbol{0} \iff \text{Cov}(\boldsymbol{X}_{20}, \boldsymbol{V}_{20}) = \boldsymbol{0} \text{ and } \text{Cov}(\boldsymbol{X}_{21}, \boldsymbol{V}_{20}) = \boldsymbol{0}.$$

Suppose $\mathbb{E}(A_2) = 0$. Note that $\text{Cov}(\boldsymbol{X}_{20}, A_2\boldsymbol{X}_{21}) = \mathbb{E}(A_2)\text{Cov}(\boldsymbol{X}_{20}, \boldsymbol{X}_{21}) = \boldsymbol{0}$.
Rewrite $\text{Cov}(\boldsymbol{X}_2)$ as a partitioned matrix:

$$\text{Cov}(\boldsymbol{X}_2) = \text{Cov}\begin{pmatrix} \boldsymbol{X}_{20} \\ A_2 \\ A_2\boldsymbol{X}_{21} \end{pmatrix} = \begin{pmatrix} \text{Cov}(\boldsymbol{X}_{20}) & \text{Cov}(\boldsymbol{X}_{20}, A_2) & \text{Cov}(\boldsymbol{X}_{20}, A_2\boldsymbol{X}_{21}) \\ \text{Cov}(A_2, \boldsymbol{X}_{20}) & \text{Var}(A_2) & \text{Cov}(A_2, A_2\boldsymbol{X}_{21}) \\ \text{Cov}(A_2\boldsymbol{X}_{21}, \boldsymbol{X}_{20}) & \text{Cov}(A_2\boldsymbol{X}_{21}, A_2) & \text{Cov}(A_2\boldsymbol{X}_{21}) \end{pmatrix}$$

$$= \begin{pmatrix} \text{Cov}(\boldsymbol{X}_{20}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0}^T & \text{Var}(A_2) & \text{Var}(A_2)\mathbb{E}(\boldsymbol{X}_{21}^T) \\ \mathbf{0}^T & \text{Var}(A_2)\mathbb{E}(\boldsymbol{X}_{21}) & \text{Cov}(A_2\boldsymbol{X}_{21}) \end{pmatrix}.$$

Let $P = \begin{pmatrix} \text{Var}(A_2) & \text{Var}(A_2)\mathbb{E}(\boldsymbol{X}_{21}^T) \\ \text{Var}(A_2)\mathbb{E}(\boldsymbol{X}_{21}) & \text{Cov}(A_2\boldsymbol{X}_{21}) \end{pmatrix}$.

Then $\text{Cov}(\boldsymbol{X}_2)^{-1} = \begin{pmatrix} \text{Cov}(\boldsymbol{X}_{20})^{-1} & \mathbf{0}^T \\ \mathbf{0} & P^{-1} \end{pmatrix}$. It follows that

$$\mathbb{E}(A_2) = 0 \iff \text{Bias} \begin{pmatrix} \hat{\beta}_{210} \\ \hat{\boldsymbol{\beta}}_{211} \end{pmatrix} = P^{-1} \begin{pmatrix} 0 \\ \mathbb{E}(A_2)\text{Cov}(\boldsymbol{X}_{21}, \boldsymbol{V}_{20}) \end{pmatrix} = \mathbf{0}.$$

Therefore, the estimators of stage 2 treatment effects, $\hat{\beta}_{210}$ and $\hat{\boldsymbol{\beta}}_{211}$, are unbiased if and only if $\mathbb{E}(A_2) = 0$ or $\text{Cov}(\boldsymbol{X}_{20}, \boldsymbol{V}_{20}) = \text{Cov}(\boldsymbol{X}_{21}, \boldsymbol{V}_{20}) = \mathbf{0}$. □

Theorem 3.3.3 (Bias of Stage 2 Main Effect Estimators)**.** *Assume that $\boldsymbol{V}_{20}$ is a vector of unmeasured covariates that are independent of A_2 and $\mathbb{E}(A_2) = 0$. Suppose that $\boldsymbol{V}_{20}$ is correlated with $\boldsymbol{X}_{20}$ and $Cov(\boldsymbol{X}_2)$ is invertible. Then the estimators of stage 2 main effects are biased and the bias is $\mathcal{B}'\boldsymbol{\gamma}_{20}$, where*

$$\mathcal{B}' = \begin{pmatrix} \mathbb{E}(\boldsymbol{V}_{20}^T) - \mathbb{E}(\boldsymbol{X}_2^T)Cov(\boldsymbol{X}_2)^{-1}Cov(\boldsymbol{X}_2, \boldsymbol{V}_{20}) \\ Cov(\boldsymbol{X}_{20})^{-1}Cov(\boldsymbol{X}_{20}, \boldsymbol{V}_{20}) \end{pmatrix}. \tag{3.7}$$

Proof. By Theorem 3.3.1, we know that the bias of $\begin{pmatrix} \hat{\beta}_{200} \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix}$ is

$$\mathcal{B} = \begin{pmatrix} \mathbb{E}(\boldsymbol{V}_{20}^T) - \mathbb{E}(\boldsymbol{X}_2^T)\mathrm{Cov}(\boldsymbol{X}_2)^{-1}\mathrm{Cov}(\boldsymbol{X}_2, \boldsymbol{V}_{20}) \\ \mathrm{Cov}(\boldsymbol{X}_2)^{-1}\mathrm{Cov}(\boldsymbol{X}_2, \boldsymbol{V}_{20})\gamma_{20} \end{pmatrix}.$$

Note that $\mathrm{Cov}(\boldsymbol{X}_{20}, A_2\boldsymbol{X}_{21}) = \mathbb{E}(A_2)\mathrm{Cov}(\boldsymbol{X}_{20}, \boldsymbol{X}_{21}) = \mathbf{0}$. Rewrite $\mathrm{Cov}(\boldsymbol{X}_2)$ as a partitioned matrix:

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{X}_2) = \mathrm{Cov}\begin{pmatrix} \boldsymbol{X}_{20} \\ A_2 \\ A_2\boldsymbol{X}_{21} \end{pmatrix} &= \begin{pmatrix} \mathrm{Cov}(\boldsymbol{X}_{20}) & \mathrm{Cov}(\boldsymbol{X}_{20}, A_2) & \mathrm{Cov}(\boldsymbol{X}_{20}, A_2\boldsymbol{X}_{21}) \\ \mathrm{Cov}(A_2, \boldsymbol{X}_{20}) & \mathrm{Var}(A_2) & \mathrm{Cov}(A_2, A_2\boldsymbol{X}_{21}) \\ \mathrm{Cov}(A_2\boldsymbol{X}_{21}, \boldsymbol{X}_{20}) & \mathrm{Cov}(A_2\boldsymbol{X}_{21}, A_2) & \mathrm{Cov}(A_2\boldsymbol{X}_{21}) \end{pmatrix} \\ &= \begin{pmatrix} \mathrm{Cov}(\boldsymbol{X}_{20}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0}^T & \mathrm{Var}(A_2) & \mathrm{Var}(A_2)\mathbb{E}(\boldsymbol{X}_{21}^T) \\ \mathbf{0}^T & \mathrm{Var}(A_2)\mathbb{E}(\boldsymbol{X}_{21}) & \mathrm{Cov}(A_2\boldsymbol{X}_{21}) \end{pmatrix}. \end{aligned}$$

Let $P = \begin{pmatrix} \mathrm{Var}(A_2) & \mathrm{Var}(A_2)\mathbb{E}(\boldsymbol{X}_{21}^T) \\ \mathrm{Var}(A_2)\mathbb{E}(\boldsymbol{X}_{21}) & \mathrm{Cov}(A_2\boldsymbol{X}_{21}) \end{pmatrix}$.

Then $\mathrm{Cov}(\boldsymbol{X}_2)^{-1} = \begin{pmatrix} \mathrm{Cov}(\boldsymbol{X}_{20})^{-1} & \mathbf{0}^T \\ \mathbf{0} & P^{-1} \end{pmatrix}$ and the bias of $\hat{\boldsymbol{\beta}}_{201}$ is

$$\mathrm{Cov}(\boldsymbol{X}_{20})^{-1}\mathrm{Cov}(\boldsymbol{X}_{20}, \boldsymbol{V}_{20})\boldsymbol{\gamma}_{20}.$$

□

Theorem 3.4.1. *Suppose* $X \sim N(\mu, \sigma^2)$. *Then*

$$\mathbb{E}(|X|) = \left(\frac{2\sigma^2}{\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \mu\left\{1 - 2\Phi\left(-\frac{\mu}{\sigma}\right)\right\}.$$

Proof. Let $f(x)$ and $F(x)$ be the probability density function and the cumulative distribution function of X respectively. Then

$$\begin{aligned}
\mathbb{E}(|X|) &= \int_{-\infty}^{0} -xf(x)dx + \int_{0}^{\infty} xf(x)dx \\
&= 2\int_{0}^{\infty} xf(x)dx - \mu \\
&= 2\int_{0}^{\infty} (x-\mu)f(x)dx + 2\mu\int_{0}^{\infty} f(x)dx - \mu \\
&= 2\int_{0}^{\infty} (x-\mu)\left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx + 2\mu\Big\{1 - F(0)\Big\} - \mu \\
&= 2\left[-\sigma^2\left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\right]_{0}^{\infty} + 2\mu\left\{1 - \Phi\left(-\frac{\mu}{\sigma}\right)\right\} - \mu \\
&= \left(\frac{2\sigma^2}{\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \mu\left\{1 - 2\Phi\left(-\frac{\mu}{\sigma}\right)\right\}.
\end{aligned}$$

□

B.2 Additional Results for Data Analysis

Method	n_1	Stage 1 Prediction		n_2	Stage 2 Prediction	
		$\hat{d}_1^{\text{opt}} = -1$	$\hat{d}_1^{\text{opt}} = 1$		$\hat{d}_2^{\text{opt}} = -1$	$\hat{d}_2^{\text{opt}} = 1$
			`max_drinks`			
sQ	490	219 (44.7%)	271 (55.3%)	140	90 (64.3%)	50 (35.7%)
mQ	490	224 (45.7%)	266 (54.3%)	140	89 (63.6%)	50 (36.4%)
IQ	490	187 (38.2%)	303 (61.8%)	140	90 (64.3%)	50 (35.7%)
mIQ	490	181 (37.0%)	309 (63.0%)	140	90 (64.3%)	50 (35.7%)
			`byaacq`			
sQ	496	298 (60.1%)	198 (39.9%)	142	78 (54.9%)	64 (45.1%)
mQ	496	261 (52.6%)	235 (47.4%)	142	77 (54.2%)	65 (45.8%)
IQ	496	237 (47.8%)	259 (52.2%)	142	79 (55.6%)	63 (44.4%)
mIQ	496	229 (46.2%)	267 (53.8%)	142	77 (54.2%)	65 (45.8%)

Table B.1: Data analysis results using all four methods: sQ, mQ, IQ, and mIQ.

For the primary outcome `max_drinks`, standard Q-learning and the modified counterpart with Murphy's regret function generate similar results, indicating that the heterogeneity of stage 1 intervention effects is weak. However, there is a significant difference in the results between standard and interactive Q-learning, which might indicate that the heterogeneity of stage 2 intervention effects is relatively stronger. For the secondary outcome `byaacq`, the results are distinctive across all four methods and may indicate the intervention at both stages has a heterogeneous effect across students.

Appendix C

Supplementary Materials for mLQ

C.1 The Manifest Distribution

In the latent variable formulation of probit regression, the joint distribution of Y and Y^* is:

$$\begin{aligned} f(y_i, y_i^*|\mu, \sigma) &= f(y_i^*|\mu_i, \sigma)p(y_i|y_i^*) \\ &= (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(y_i^* - \mu_i)^2\right\} \Big[\mathbb{1}\{y_i = 1\}\mathbb{1}\{y_i^* > \xi\} \\ &\qquad + \mathbb{1}\{y_i = 0\}\mathbb{1}\{y_i^* \leq \xi\}\Big] \\ &= (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(y_i^* - \mu)^2\right\} \Big[y_i\mathbb{1}\{y_i^* > \xi\} + (1 - y_i)\mathbb{1}\{y_i^* \leq \xi\}\Big] \end{aligned}$$

where f denotes the probability density function and p denotes the probability mass function.

The manifest distribution, or the marginal distribution of the response variable Y,

conditional on model covariates H_2 and A_2, is

$$\begin{aligned} p(y_i|H_{2i}, A_{2i}) &= \int_{-\infty}^{\infty} p(y_i|y_i^*, H_{2i}, A_{2i}) f(y_i^*|H_{2i}, A_{2i}) dy_i^* \\ &= y_i \int_{\xi}^{\infty} f(y_i^*|H_{2i}, A_{2i}) dy_i^* + (1 - y_i) \int_{-\infty}^{\xi} f(y_i^*|H_{2i}, A_{2i}) dy_i^* \\ &= y_i \left\{ 1 - \Phi \left(\frac{\xi - \boldsymbol{x}_{20,i}^T \boldsymbol{\beta}_{20} - a_{2i} \boldsymbol{x}_{21,i}^T \boldsymbol{\beta}_{21}}{\sigma} \right) \right\} \\ &+ (1 - y_i) \Phi \left(\frac{\xi - \boldsymbol{x}_{20,i}^T \boldsymbol{\beta}_{20} - a_{2i} \boldsymbol{x}_{21,i}^T \boldsymbol{\beta}_{21}}{\sigma} \right), \end{aligned}$$

where Φ is the cumulative distribution function of a standard normal random variable.

C.2 Conditional Moments of the Latent Variable

In order to construct the EM algorithm, we need to derive the first and second moments of the latent variable Y^* conditional on the manifest variable Y. As the value of Y implies the truncation range of Y^*, the distribution of Y^* conditional on Y is a truncated normal distribution. The first moment, i.e. the expectation, of Y^* conditional on Y is

$$\mathbb{E}\left(Y^* \middle| Y\right) = \mu + \frac{(2Y - 1)\sigma^2 \phi(\mu, \sigma^2; \xi)}{Y\left\{1 - \Phi(\mu, \sigma^2; \xi)\right\} + (1 - Y)\Phi(\mu, \sigma^2; \xi)}, \tag{C.1}$$

and the second moment of Y^* conditional on Y is

$$\mathbb{E}\left(Y^{*2} \middle| Y\right) = \mu^2 + \sigma^2 + \frac{(2Y - 1)(\mu + \xi)\sigma^2 \phi(\mu, \sigma^2; \xi)}{Y\left\{1 - \Phi(\mu, \sigma^2; \xi)\right\} + (1 - Y)\Phi(\mu, \sigma^2; \xi)}, \tag{C.2}$$

where $\phi(\mu, \sigma^2)$ and $\Phi(\mu, \sigma^2)$ denote the probability density function and cumulative distribution function of $N(\mu, \sigma^2)$ respectively.

We show the derivation of Equations (C.1) and (C.2) for the case of $Y = 1$ only, and the derivation for the case of $Y = 0$ can be done similarly.

$$
\begin{aligned}
\mathbb{E}\left(Y^*\middle|\, Y=1\right) &= \frac{1}{1-\Phi(\mu,\sigma^2;\xi)}\int_{\xi}^{\infty} y^*(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(y^*-\mu)^2\right\}dy^* \\
&= \frac{1}{1-\Phi(\mu,\sigma^2;\xi)}\left\{-\sigma^2\left[(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(y^*-\mu)^2\right\}\right]_{\xi}^{\infty}\right. \\
&\qquad\qquad \left. +\int_{\xi}^{\infty}\mu f(y^*)dy^*\right\} \\
&= \frac{\sigma^2\phi(\mu,\sigma^2;\xi)+\mu\left\{1-\Phi(\mu,\sigma^2;\xi)\right\}}{1-\Phi(\mu,\sigma^2;\xi)} \\
&= \mu+\sigma^2\frac{\phi(\mu,\sigma^2;\xi)}{1-\Phi(\mu,\sigma^2;\xi)}
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\left(Y^{*2}\middle|\, Y=1\right) &= \frac{1}{1-\Phi(\mu,\sigma^2;\xi)}\int_{\xi}^{\infty} y^{*2}(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(y^*-\mu)^2\right\}dy^* \\
&= \frac{1}{1-\Phi(\mu,\sigma^2;\xi)}\left\{\int_{\xi}^{\infty} y^*(y^*-\mu)(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(y^*-\mu)^2\right\}dy^*\right. \\
&\qquad\qquad \left. +\mu\Big[\mu\big\{1-\Phi(\mu,\sigma^2;\xi)\big\}+\sigma^2\phi(\mu,\sigma^2;\xi)\Big]\right\} \\
&= \frac{1}{1-\Phi(\mu,\sigma^2;\xi)}\left\{\left[-\sigma^2 y^*(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(y^*-\mu)^2\right\}\right]_{\xi}^{\infty}\right. \\
&\qquad\qquad +\int_{\xi}^{\infty}\sigma^2(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(y^*-\mu)^2\right\}dy^* \\
&\qquad\qquad \left. +\mu\Big[\mu\big\{1-\Phi(\mu,\sigma^2;\xi)\big\}+\sigma^2\phi(\mu,\sigma^2;\xi)\Big]\right\} \\
&= \frac{1}{1-\Phi(\mu,\sigma^2;\xi)}\Big[\sigma^2\xi\phi(\mu,\sigma^2;\xi)+\sigma^2\left\{1-\Phi(\mu,\sigma^2;\xi)\right\} \\
&\qquad\qquad +\mu^2\left\{1-\Phi(\mu,\sigma^2;\xi)\right\}+\mu\sigma^2\phi(\mu,\sigma^2;\xi)\Big] \\
&= \mu^2+\sigma^2+(\mu+\xi)\sigma^2\frac{\phi(\mu,\sigma^2;\xi)}{1-\Phi(\mu,\sigma^2;\xi)}
\end{aligned}
$$