Title
Non-Parametric Tests for Treatment Effect Heterogeneity in Randomized Experiments and Observational Studies

Author
Dai, Maozhu

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

Non-Parametric Tests for Treatment Effect Heterogeneity in Randomized Experiments and Observational Studies

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Maozhu Dai

Dissertation Committee:
Chancellor's Professor Hal S. Stern , Chair
Assistant Professor Weining Shen
Professor Daniel L. Gillen

2021

DEDICATION

To my parents, Bin Dai & Xiaobin Chen.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I thank my advisor, Professor Hal S. Stern. Reaching out to Hal four years ago was one of the most correct decisions I have ever made in my life, which laid the foundation of my not only meaningful but also joyful life at UCI. Apart from a great deal of statistical expertise I have learned from Hal, he sets a role model for me as being a real scientist. He is upright, diligent, rigorous, and a real leader. Meanwhile, he is also warm, patient and compassionate. During the past four years, I have received tremendous help from Hal. Without his continued support, this dissertation would not be possible.

I also express my sincere gratitude to my other dissertation committee members Weining Shen and Daniel Gillen, as well as my advancement committee members which further included Yaming Yu and Sherrie Kaplan, for their very insightful suggestions. In addition, I thank Weining Shen, the co-author of matrierals in Chapter 3 and Chapter 4, for his unrelenting dedication.

I thank all other professors in the Department of Statistics for being very generous with their time and wisdom. Because of their kindness, the department has been a paradise for us students. I also thank all my friends for their companion. Without them, my life would never have been so colorful. I thank Rosemary Busta, Lisa Stieler and Kazumo Washizuka for providing me with lots of convenience and making my life easy.

At last, I thank my parents Bin Dai and Xiaobin Chen for their unconditional love and support, which makes me braver in the face of any obstacles.

VITA

Maozhu Dai

EDUCATION

Doctor of Philosophy in Statistics **2016-2021**
University of California, Irvine *Irvine, CA, USA*

Bachelor of Science in Statistics **2012-2016**
University of Science and Technology of China *Hefei, Anhui, China*

RESEARCH EXPERIENCE

Graduate Student Researcher **2017–2021**
University of California, Irvine *Irvine, CA, USA*

TEACHING EXPERIENCE

Teaching Assistant **9/2019-12/2019, 9/2018-12/2018**
University of California, Irvine *Irvine, CA, USA*

INDUSTRY EXPERIENCE

Artificial Intelligence - Machine Learning Engineer Summer Intern **6/2020-9/2020**
LinkedIn *Irvine, CA, USA*

Data Scientist Summer Intern **6/2019–9/2019**
Microsoft *Redmond, WA, USA*

PUBLICATIONS and WORKING PAPERS

Dai, Maozhu; Stern, Hal S. (2020) "A U-Statistic-Based Test of Treatment Effect Heterogeneity" Under review, *Journal of Nonparametric Statistics*.

Kaplan, Sherrie H.; Fortier, Michelle A.; Shaughnessy, Marilou; Maurer, Eva; Vivero-Montemayor, Marla; Masague, Sergio Gago; Hayes, Dylan; Stern, Hal S.; **Dai, Maozhu**; Kain Zeev N. (2020) "Development and initial validation of self-report measures of general health, preoperative anxiety, and postoperative pain in young children using computer-administered animation" Published, *Pediatric Anesthesia*.

Dai, Maozhu; Shen, Weining; Stern, Hal S. (2021) "Nonparametric Tests for Treatment Effect Heterogeneity in Observational Studies" Under review, *Journal of Multivariate Analysis*.

Dai, Maozhu; Shen, Weining; Stern, Hal S. (2021) "Sensitivity Analysis for the Adjusted Mann-Whitney Test with Observational Studies" Under review, *Observational Studies*.

CONFERENCE PRESENTATIONS

Dai, Maozhu and Stern, Hal S. (8/2018). "Exploring Treatment Eect Heterogeneity Using Propensity Scores." Joint Statistical Meetings (JSM), Vancouver, Canada.

SOFTWARE

Programming Languages **Fluent** in R, Python, SQL; **Experienced** in Spark, C, LaTeX

ABSTRACT OF THE DISSERTATION

Non-Parametric Tests for Treatment Effect Heterogeneity in Randomized Experiments and Observational Studies

by

Maozhu Dai

Doctor of Philosophy in Statistics

University of California, Irvine, 2021

Chancellor's Professor Hal S. Stern , Chair

Comprehensively assessing the effect of a treatment usually includes two objectives, estimating the average treatment effect across the whole target population and evaluating variability of the treatment effect across different subpopulations in an effort to provide more precise treatment recommendations. A common way to identify treatment effect heterogeneity is to split the sample into several strata based on one or more baseline covariates which may be relevant to the effect of treatment, and then compare the localized or stratum-specific treatment effects across those strata. Parametric approaches have been proposed to compare average treatment effects across several strata. One approach is testing interactions between treatment indicator and group indicators in linear regressions (Allison, 1977), and another approach is the likelihood ratio test proposed by Gail and Simon (1985). Both of them require parametric assumptions of outcome distributions. When the parametric assumptions fail, the test may be invalid or the power may be negatively impacted. Thus there is a need for non-parametric tests that can better adapt to various outcome distributions.

Randomized experiments are considered to be the gold standard for assessing treatment effects, as all baseline covariates are expected to be well balanced in treatment groups after randomization. However, randomized experiments are not always feasible due to various ob-

stacles, e.g., ethical concerns and high expense. Therefore researchers turn to observational studies. Not only can they avoid the obstacles faced by randomized experiments, but because there are often fewer exclusionary characteristics, the results of observational studies may generalize better to the target population. A main challenge of observational studies is controlling confounding variables. There is a considerable literature on causal inference in observational studies that has been developed targeting this challenge. Many of the proposed procedures balance the observed variables and then rely on the unconfoundedness assumption, i.e., all confounding variables are observed. This assumption is not only strong but also unverifiable. The violation of this assumption can invalidate causal conclusions. A variety of approaches to assessing the sensitivity of causal conclusions to violations of the unconfoundedness assumption have been proposed. By assessing the extent of the assumption violation required to change the conclusion and evaluating the possibility of such a violation based on domain knowledge, it is possible to provide more reliable conclusions.

In this dissertation, we describe three contributions we have made to the goal of comprehensively evaluating the effectiveness of treatments. The first two contributions are non-parametric U-statistic-based tests examining the variability of treatment effects across different subpopulations. The first procedure can be appropriately applied in cases, like randomized experiments, where all baseline covariates are well balanced within each stratum; the second procedure adjusts unbalanced confounding variables using propensity scores. Compared to their parametric counterparts, likelihood ratio tests, our non-parametric tests are more powerful when the distributions of study outcomes depart substantially from the distributions assumed by likelihood ratio tests. The third contribution is a sensitivity analysis that addresses the concern of possible violation of the unconfoundedness assumption for the adjusted Mann-Whitney test, a non-parametric test that evaluates the existence of treatment effects in observational studies.

Chapter 1

Introduction

Researchers from many scientific disciplines are interested in estimating the effects of treatments. For example, biologists study whether a mutation in a gene causes a particular human disease; economists investigate how a welfare policy would change household incomes; and criminologists are interested in how a proposed punishment regime will affect criminals' recidivism rates. This is however a challenging task, as each subject can only be exposed to at most one treatment, and subjects receiving different treatments may have very different characteristics.

The causal inference literature usually focuses on estimating the average treatment effect across a large population (e.g., Rubin, 1974; Rosenbaum and Rubin, 1983). With a large number of subjects in the treatment and control groups, in order to isolate the effect of treatment on the outcome of interest, we only need to balance the distributions of all baseline covariates between the two treatment groups instead of looking for an exact one-to-one match. Randomized experiments are an effective approach for achieving such balance. In randomized experiments, the treatment assignments are completely independent of all other factors, which makes the distributions of all covariates in both the treatment and control

groups equal to their marginal distributions. For this reason randomized experiments are considered to be the gold standard for average treatment effect estimation, and have been commonly used in various cases, including effectiveness evaluation of drugs and vaccines in human clinical trials, e.g, the vaccines for COVID-19 (Voysey et al., 2021).

1.1 Assessing treatment effect heterogeneity in randomized experiments

Apart from average treatment effects, the variability of treatment effects across the population is also of great importance. Though it is possible that a study will identify a treatment that provides similar effect across the entire population of interest, it is often the case that people with different characteristics can respond quite differently to the same treatment. A treatment with positive effect on some people can have a negative effect on others. Focusing on the average treatment effect alone may neglect such important information. For example, Pate and Hamilton (1992) found that in domestic violence cases, arrest of the suspect did not have a significant average effect on subsequent spouse assault. However, when they looked into subgroups of suspects, they found that arrest decreased the recidivism rate among employed suspects whereas it increased the recidivism rate among unemployed suspects. Therefore, quantifying localized or subpopulation effects is crucial. A popular method to learn about localized treatment effects is subgroup analysis (Rothwell, 2005; Cook et al., 2004). We can split the original sample into several strata based on covariates that are expected to be relevant to the effect of the treatment, and then conduct analyses to obtain average treatment effects for each stratum. Though commonly used, subgroup analysis involves various problems, e.g., multiple testing and loss of statistical power (Cook et al., 2004). Thus it is not always recommended. However when there is enough treatment effect heterogeneity across strata, subgroup analysis can be beneficial. Therefore a test identifying

the existence of treatment effect heterogeneity can be valuable.

Lots of tests assessing treatment effect heterogeneity have been published. They focus on different hypotheses to examine different aspects of heterogeneity. For instance, Delgado and Escanciano (2013) and Hsu (2017) focus on testing conditional stochastic dominance; Gail and Simon (1985) and Chang et al. (2015) are interested in checking whether treatment effects in different subpopulations are of the same sign; Crump et al. (2008) and Chang et al. (2015) focus on testing whether treatment effects are consistently equal to zero; Ding et al. (2016) are interested in the sharp null hypothesis that all individuals have the same treatment effects. In this dissertation, we focus on assessing whether average treatment effects are the same across multiple strata. With respect to this goal, parametric approaches have been proposed a long time ago. For instance, we can test interactions between treatment indicators and subgroup indicators in linear regressions (Allison, 1977; Byar, 1985). Also Gail and Simon (1985) proposed the likelihood ratio test (LRT). These approaches relies on parametric assumptions of the outcome distributions, and failure of which may make the tests invalid or impact the power. In some cases, a non-parametric test that does not rely on parametric assumptions can be helpful.

U-statistics (Korolyuk and Borovskich, 2013) have been commonly used in various non-parametric tests, e.g., the Mann-Whitney test (Mann and Whitney, 1947) and the signed rank test (Van der Vaart, 2000). An overview of U-statistics can be found in Section 2.2. We propose a non-parametric test for treatment effect heterogeneity based on U-statistics. It does not rely on parametric assumptions for the distributions of study outcomes and it also bypasses treatment effect estimation for each stratum. Compared to the LRT, the proposed U-statistic-based test could be more powerful when outcome distributions deviate substantially from normal distributions and it is more robust to outliers.

1.2 Assessing treatment effect heterogeneity in observational studies

A major limitation of randomized experiments is that they are not always feasible. One reason is that the expense may be too high if the treatment is expensive and the number of participants required is large. More importantly, equipoise may not hold in some settings and conducting randomized experiments in those situations is unethical. For instance, it would be unethical to ask random subjects to smoke in order to investigate the effect of smoking on lung cancer. Also we would never propose modifying subjects' genes to study the causal effect of a gene mutation on a disease of interest. This leads to a scientific need for causal inferences to be derived from observational studies.

In observational studies, we observe and collect data from subjects in different treatment groups without any manipulation by the experimenter. This procedure is often cheaper and it avoids the ethical issues that randomized experiments can encounter. An additional advantage is that with appropriate sampling, the inference results can be easily generalized to the population of interest, whereas the results from randomized experiments can only be generalized to individuals who are like the participants, which may be different from the target population. Despite these advantages, estimating treatment effects in observational studies involves many challenges, e.g., potential selection bias and existence of confounders (Lu, 2009).

A large number of techniques have been proposed in the last several decades targeting on balancing confounders in observational studies, ranging from exact matching, to regressions, to various propensity-score-based approaches (e.g., Rosenbaum and Rubin, 1983; Dehejia and Wahba, 1999; Rosenbaum, 2002b; Imai and Ratkovic, 2014; Imbens and Rubin, 2015; Vegetabile et al., 2020). Propensity-score-based approaches are more and more commonly used when there are a large number of baseline covariates. The propensity score is the

probability of receiving a particular treatment conditional on observed baseline covariates, which can be estimated by any binary classification models that predict class probabilities. Based on a theorem from Rosenbaum and Rubin (1983), we only need to balance the one-dimensional propensity scores to adjust all observed baseline covaraites. Various approaches can be used to adjust for propensity scores. For example, a multiple linear regression that only adjusts for propensity scores can be used. Matching and subclassification on propensity scores are also popular (Imbens and Rubin, 2015). Apart from them, some researchers also use inverse probability weighting (IPW) (Horvitz and Thompson, 1952) to balance baseline covariates. After weighting subjects by inverse of their group membership probabilities, the distributions of baseline covariates in treatment and control groups would both be equal to the marginal distributions of these covariates. Motivated by Satten et al. (2018), which uses IPW to balance covariates for two-sample U-statistics, we apply IPW to extend our proposed U-statistic-based test for treatment effect heterogeneity to be applicable in observational studies. The adjusted non-parametric test inherits the advantages of the unadjusted version, i.e., the adjusted non-parametric test could be more powerful than its parametric counterpart when the parametric assumptions of the latter fail.

1.3 Sensitivity analysis for the unconfoundedness assumption

The most challenging limitation of propensity-score-based approaches is that they can only adjust for variables that have been observed; causal inferences rely on the unconfoundedness assumption which assumes we have observed all confounders. Even in a carefully designed study, the assumption can be violated and it is untestable with empirical data. There is emerging literature on approaches for addressing this limitation (e.g., Cornfield et al., 1959; Rosenbaum, 2002c; VanderWeele and Ding, 2017; Zhao et al., 2019). These authors propose

various sensitivity analysis approaches to assess the degree of violation of the unconfoundedness assumption that would lead to a change in the conclusion. Different sensitivity frameworks/approaches are implemented in different settings. For example, Rosenbaum (2002a) targets on studies with matched cases, and the sensitivity framework sets threshold for ratios of the probabilities of receiving treatment for the matched subjects. Hosman et al. (2010) and Cinelli and Hazlett (2020) focus on linear regression models and their sensitivity frameworks set thresholds for the associations between unobserved covariates with outcomes and treatment assignments. Zhao et al. (2019) focuses on IPW-based estimators and adopts the marginal sensitivity framework that sets thresholds for the odds ratios between the desired propensity scores that are also conditional on unobserved confounders and the estimated propensity scores.

We develop a sensitivity analysis approach for the adjusted Mann-Whitney test proposed by Satten et al. (2018). The Mann-Whitney test is a popular non-parametric test used in randomized two-treatment experiments to assess treatment effects. Satten et al. (2018) use IPW based on propensity scores to extend this test to be applicable in observational studies. We develop an approach for conducting a sensitivity analysis to assess the robustness of the Satten et al. (2018) test to the violation of the unconfoundedness assumption. We use the marginal sensitivity framework introduced by Tan (2006) and Zhao et al. (2019), which sets a threshold for the absolute value of the log odds ratio between the desired propensity score conditional on unobserved confounders and the estimated propensity score. Based on the bootstrap idea proposed by Zhao et al. (2019), which focuses on sensitivity analysis for mean estimation with missing data, we develop an approach that derives an interval that achieves the nominal coverage probability for the expectation of the adjusted Mann-Whitney test statistic as long as the true propensity scores are within the pre-specified sensitivity ranges. We also extend this approach to more general adjusted multi-sample U-statistics, which includes the test statistic comparing treatment effects between two strata of Chapter 3 and mean estimation with missing data as special examples.

1.4 Outline of this dissertation

This dissertation describes three contributions to the assessment of treatment effect heterogeneity. The first two contributions focus on testing treatment effect heterogeneity while avoiding strong assumptions about the distribution of study outcomes or test statistics. Chapter 2 introduces a nonparametric U-statistic-based test for treatment effect heterogeneity across pre-defined strata in randomized experiments. Chapter 3 extends this test to be applicable in observational studies. The third contribution, described in Chapter 4, addresses an approach to sensitivity analysis for the adjusted Mann-Whitney test and other more general adjusted U-statistics. In Chapter 5, we summarize our conclusions and discuss potential future directions for research in this area.

Chapter 2

A U-Statistic-Based Test of Treatment Effect Heterogeneity

2.1 Introduction

Treatment effect heterogeneity is of great importance, as the average treatment effect across the whole population may neglect important variability of the treatment effect across subpopulations. In health care, the concept of personalized medicine is attracting a great deal of attention (Ginsburg and Willard, 2009; Jain, 2009; Chan and Ginsburg, 2011), as it promises a way to provide treatment recommendations with greater precision based on a patient's baseline characteristics. In social sciences, people are using similar approaches to assess localized effects in order to comprehensively evaluate a policy or a campaign strategy (Bitler et al., 2006; Feller and Holmes, 2009). In one criminology study, arrest can result in effects of opposite sign on recidivism rates for different kinds of criminals (Pate and Hamilton, 1992). In all areas, subgroup analysis (Dixon and Simon, 1991) can be conducted to assess treatment effects at the subpopulation level. Both randomized studies and observational studies

can provide insight into defining relevant subpopulations. In randomized clinical trials, patients can be assigned to different subgroups based on one or several baseline factors. In observational studies, one approach to identify subpopulation is through subclassification on propensity scores (Xie et al., 2012). However, subgroup analysis is not always recommended as it also involves some problems, e.g., multiple testing and loss of statistical power (Cook et al., 2004). Subgroup analysis can be especially valuable when there is enough treatment effect heterogeneity. Therefore, reliable inference about whether there is heterogeneity in treatment effects across strata is usually needed.

There is a great deal of literature on exploring heterogeneity of treatment effects. The published results focus on different aspects of heterogeneity in that they examine different null hypotheses. Some focus on testing conditional stochastic dominance (Delgado and Escanciano, 2013; Hsu, 2017). Other focus on testing whether the treatment effects in subpopulations are of the same sign (Gail and Simon, 1985; Chang et al., 2015). Crump et al. (2008) and Chang et al. (2015) are interested in tesing whether the treatment effects are consistently equal to zero. Ding et al. (2016) focus on the null hypothesis that all individuals have the same treatment effects. In this paper, we focus on testing whether the average stratum-specific treatment effects are constant across different strata.

A common approach to identifying whether the stratum-specific average treatment effects are equal across different strata is through parametric statistical tests. For example, we can test the interaction term between treatment assignment and effect modifiers in multiple linear regressions (Allison, 1977; Byar, 1985). We can also use the likelihood ratio test (LRT) proposed by Gail and Simon (1985). These approaches are widely used but rely on the parametric assumptions of outcome distributions being correct. When the assumptions are not correct, the inference is invalid or the power of the test is impacted.

Non-parametric approaches exist as well. Crump et al. (2008) created a non-parametric approach based on a particular series estimator for treatment effect introduced by Imbens

et al. (2006). Sant'Anna (2020) generalized Crump et al. (2008) by allowing censored data and endogenous treatment selections. In this paper, we propose a U-statistic-based approach to test whether the stratum treatment effects are homogeneous without having to estimate the stratum treatment effects. Our approach relies on an unconfoundedness assumption in each stratum.

U-statistics have been widely used to create distribution-free tests. Examples include the signed rank test and Mann-Whitney test (Van der Vaart, 2000). Compared to their parametric counterparts assuming normal distributions, U-statistic-based tests usually have higher power when the distributions are far from normal. When the normality assumption is satisfied, the parametric test has slightly higher power, but when the U-statistic-based test is more powerful, the advantage can be significant (Hodges et al., 1956; Lehmann and D'Abrera, 1975; Zimmerman, 1998). The non-parametric heterogeneity test we propose here is also based on U-statistics. We use U-statistics to compute a test statistic comparing treatment effects across pairs of strata. The overall test statistic is a combination of the pairwise test statistics. Its performance will be compared to the LRT proposed by Gail and Simon (1985).

The remainder of the paper is structured as follows. In Section 2.2, we provide a review of U-statistics. In section 2.3, we introduce our proposed U-statistic-based non-parametric test for treatment effect heterogeneity in detail. In section 2.4, some simulation studies demonstrate the validity of the test and its comparison with the LRT under several different circumstances. In section 2.5, we apply the proposed method to a randomized study of program effectiveness in labor economics. Additional discussion of this approach can be found in Section 2.6.

2.2 Background: Review of U-Statistics

U-statistics are a class of statistics widely used to construct non-parametric unbiased estimators of estimable parameters with minimum variance. The asymptotic normality property of U-statistics (under some mild conditions) makes it very popular as a non-parametric testing tool.

We start with a review of one-sample U-statistics (Van der Vaart, 2000). Let $X_1, \cdots, X_n$ be a random sample from $F(x)$, and assume there is a symmetric function $\phi(x_1, \cdots, x_m)$ $(m \leq n)$ such that $E[\phi(X_1, \cdots, X_m)] = \theta$, where θ is the parameter of interest. Then the U-statistic for the parameter θ created by kernel ϕ is

$$U(X_1, \cdots, X_n) = \frac{1}{\binom{n}{m}} \sum_{\beta \in B} \phi(X_{\beta_1}, \cdots, X_{\beta_m}), \tag{2.1}$$

where B contains all $\binom{n}{m}$ ordered subsets $\beta = (\beta_1, \cdots, \beta_m)$ of m integers chosen without replacement from the set $\{1, \cdots, n\}$ with $1 \leq \beta_1 < \cdots < \beta_m \leq n$.

The signed rank statistic is an example of a one-sample U-statistic where $\theta = E(I(X_1 + X_2 > 0))$. This can be used to test whether the location (median) of a symmetric distribution is equal to 0 via testing whether $\theta = \frac{1}{2}$. With the symmetric kernel $\phi(x_1, x_2) = I(x_1 + x_2 > 0)$, the corresponding U-statistic estimator of θ is

$$U = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} I(X_i + X_j > 0). \tag{2.2}$$

A key property of the one-sample U-statistic is that it has an asymptotic normal distribution. If $E\phi^2(X_1, \cdots, X_m) < \infty$, then

$$\sqrt{n}(U - \theta) \xrightarrow{D} N(0, \sigma^2), \tag{2.3}$$

where σ^2 is the asymptotic variance of $\sqrt{n}U$.

The multi-sample U-statistic is a natural extension of the one-sample U-statistic. Let $\{X_{1\alpha}, \alpha = 1, \cdots, n_1\}$, $\cdots$, $\{X_{c\delta}, \delta = 1, \cdots, n_c\}$ be c independent random samples from distribution functions $F_1(x), \cdots, F_c(x)$ respectively, and $\phi(x_{11}, \cdots, x_{1m_1}; \cdots; x_{c1}, \cdots, x_{cm_c})$ be a symmetric function within each set of variables $\{x_{j1}, \cdots, x_{jm_j}\}$ $(j = 1, \cdots, c)$ with $E(\phi)$ equal to the parameter of interest θ, where $m_1 \leq n_1, \cdots, m_c \leq n_c$. Then the corresponding c-sample U-statistic is

$$U = [\prod_{j=1}^{c} \binom{n_j}{m_j}]^{-1} \sum_{\alpha_1} \cdots \sum_{\alpha_c} \phi(X_{1\alpha_{1,1}}, \cdots, X_{1\alpha_{1,m_1}}; \cdots; X_{c\alpha_{c,1}}, \cdots, X_{c\alpha_{c,m_c}}) \tag{2.4}$$

where the summation is over all possible sets of subscripts $\alpha_j = (\alpha_{j,1}, \cdots, \alpha_{j,m_j})$ such that $1 \leq \alpha_{j,1} < \cdots < \alpha_{j,m_j} \leq n_j$ for each of the c samples (i.e, $j = 1, \cdots, c$).

The Mann-Whitney statistic is an example of two-sample U-statistic with $m_1 = m_2 = 1$. The parameter of interest is $\theta = E(I(X_{11} < X_{21}))$. The U-statistic with respect to the parameter of interest is

$$U = \frac{1}{\binom{n_1}{1}\binom{n_2}{1}} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I(X_{1i} < X_{2j}). \tag{2.5}$$

The Mann-Whitney statistic is the test statistic of Mann-Whitney test (Mann and Whitney, 1947), which is a consistent test for the null hypothesis that $\theta = \frac{1}{2}$ versus the alternative hypothesis that $\theta \neq \frac{1}{2}$.

There is also an asymptotic normality property for multi-sample U-statistics, even for a vector of several multi-sample U-statistics defined upon the same sets of mutually independent samples with different kernel functions. Lehmann et al. (1963) showed that if there are r multi-sample U-statistics $U^{(1)}, \cdots, U^{(r)}$, each defined as in (4), with corresponding kernel functions $\phi^{(1)}, \cdots, \phi^{(r)}$ such that $E[\phi^{(k)}] = \theta^{(k)}$ and $E([\phi^{(k)}]^2) < \infty$ for $k \in \{1, \cdots, r\}$, and

if there also exists positive constants λ_j $(0 < \lambda_j < 1)$ such that $\frac{n_j}{N} \to \lambda_j$ as $N = \sum_{j=1}^{c} n_j \to \infty$ for $j \in \{1, \cdots, c\}$, then

$$\sqrt{N} \begin{pmatrix} U^{(1)} - \theta^{(1)} \\ U^{(2)} - \theta^{(2)} \\ \vdots \\ U^{(r)} - \theta^{(r)} \end{pmatrix} \xrightarrow{D} N(0, \Sigma), \tag{2.6}$$

where Σ is the asymptotic covariance matrix of $\sqrt{N}(U^{(1)}, U^{(2)}, \cdots, U^{(r)})$.

In order to apply (6) to hypothesis testing with regard to the parameter $\theta = (\theta^{(1)}, \cdots, \theta^{(r)})$, we need to identify the form of Σ. This can be addressed using Hájek projection principle (Hájek, 1968) to derive the asymptotic normality property of the U-statistics.

For one c-sample U-statistic of degree $(m_1, \cdots, m_c)$, if $E(\phi^2) < \infty$, the Hájek projection of $U - \theta$ onto the space $\mathcal{V} = \{V | V = \sum_{i=1}^{n_1} f_1(X_{1i}) + \cdots + \sum_{i=1}^{n_c} f_c(X_{ci})$ where f_j $(j \in \{1, \cdots, c\})$ are some real-valued functions$\}$ is

$$\hat{U} = \frac{m_1}{n_1} \sum_{i=1}^{n_1} h_1(X_{1i}) + \cdots + \frac{m_c}{n_c} \sum_{i=1}^{n_c} h_c(X_{ci}), \tag{2.7}$$

where the h functions are defined as

$$h_j(x) = E[\phi(X_{11}, \cdots, X_{1m_1}; \cdots; X_{c1}, \cdots, X_{cm_c} | X_{j1} = x] - \theta, \quad j \in \{1, .., c\}. \tag{2.8}$$

Then it can be proved (Korolyuk and Borovskich, 2013) that

$$\sqrt{N}(U - \theta - \hat{U}) \xrightarrow{P} 0 \text{ as } N \to \infty. \tag{2.9}$$

This shows that $U - \theta$ and $\hat{U}$ have the same asymptotic distribution. By the Central Limit

Theorem,

$$\sqrt{N}\hat{U} \xrightarrow{d} N(0, \frac{m_1^2}{\lambda_1}Var(h_1(X_1)) + \cdots + \frac{m_c^2}{\lambda_c}Var(h_c(X_c))) \text{ as } N \to \infty, \tag{2.10}$$

provided the variance terms are finite. Thus we have

$$\sqrt{N}(U - \theta) \xrightarrow{D} N(0, \frac{m_1^2}{\lambda_1}Var(h_1(X_1)) + \cdots + \frac{m_c^2}{\lambda_c}Var(h_c(X_c))) \text{ as } N \to \infty. \tag{2.11}$$

With the list of U-statistics $(U^{(1)}, \cdots, U^{(r)})$, there is a list of Hájek projection $(\hat{U}^{(1)}, \cdots, \hat{U}^{(r)})$ corresponding to each of them with

$$\hat{U}^{(k)} = \frac{m_1^{(k)}}{n_1}\sum_{i=1}^{n_1} h_1^{(k)}(X_{1i}) + \cdots + \frac{m_c^{(k)}}{n_c}\sum_{i=1}^{n_c} h_c^{(k)}(X_{ci}) \text{ for } k \in \{1, \cdots, r\}, \tag{2.12}$$

where $h_j^{(k)}(x) = E[\phi^{(k)}(X_{11}, \cdots, X_{1n_1}; \cdots; X_{c1}, \cdots, X_{cn_c}|X_{j1} = x] - \theta^{(k)}$ for $j \in \{1, \cdots, c\}$. By the multidimensional Central Limit Theorem, we know

$$\sqrt{N}\begin{pmatrix} \hat{U}^{(1)} \\ \hat{U}^{(2)} \\ \vdots \\ \hat{U}^{(r)} \end{pmatrix} \xrightarrow{D} N(0, \frac{1}{\lambda_1}\Sigma_1 + \cdots + \frac{1}{\lambda_c}\Sigma_c) \tag{2.13}$$

where $\Sigma_j = Cov[m_j^{(1)}h_j^{(1)}(X_j), \cdots, m_j^{(r)}h_j^{(r)}(X_j)]$ for $j \in \{1, \cdots, c\}$. Since

$$\sqrt{N}\begin{pmatrix} U^{(1)} - \theta^{(1)} - \hat{U}^{(1)} \\ U^{(2)} - \theta^{(2)} - \hat{U}^{(2)} \\ \vdots \\ U^{(r)} - \theta^{(r)} - \hat{U}^{(r)} \end{pmatrix} \xrightarrow{P} 0 \text{ as } N \longrightarrow \infty, \tag{2.14}$$

we have

$$\sqrt{N}\begin{pmatrix} U^{(1)} - \theta^{(1)} \\ U^{(2)} - \theta^{(2)} \\ \vdots \\ U^{(r)} - \theta^{(r)} \end{pmatrix} \xrightarrow{D} N(0, \frac{1}{\lambda_1}\Sigma_1 + \cdots + \frac{1}{\lambda_c}\Sigma_c). \tag{2.15}$$

2.3 Testing for Treatment Effect Heterogeneity

Suppose we are focused on a study population comprised of S strata. For each stratum $s, s \in \{1, \cdots, S\}$, let Y_s^t denote the outcomes of subjects in the treatment group where $Y_s^t = \{Y_{si}^t, i = 1, \cdots, n_s^t\}$, and Y_s^c denotes the outcomes in the control group where $Y_s^c = \{Y_{si}^c, i = 1, \cdots, n_s^c\}$. Define $N_s = n_s^t + n_s^c$ as the total sample size in strata s, and $N = \sum_{s=1}^{S} N_s$ as the overall sample size across all strata. We develop a non-parametric U-statistic-based test (U test) for the null hypothesis of no treatment effect heterogeneity against the alternative hypothesis that not all treatment effects are equal. In the derivation and in our studies, we focus on an assumed additive treatment effect. Alternative methods of the treatment effect can be considered, they would require alternative choices for the U-statistic kernel functions.

The technique to be discussed here relies on two assumptions: (1) $Y_1^t, \cdots, Y_S^t, Y_1^c, \cdots, Y_S^c$ are mutually independent; (2)There exist constants $\lambda_s^\omega \in (0, 1)$ such that $\frac{n_s^\omega}{N} \to \lambda_s^\omega$ for all $s \in \{1, \cdots, S\}$ and $\omega \in \{t, c\}$.

2.3.1 Comparing Treatment Effects Between the First Two Strata

We start by constructing a U-statistic comparing the treatment effects of the first two strata. The hypotheses we focus on are

$$\begin{aligned} H_0 : P(Y_1^t - Y_1^c < Y_2^t - Y_2^c) + \frac{1}{2}P(Y_1^t - Y_1^c = Y_2^t - Y_q^c 2) = \frac{1}{2} \\ \Longleftrightarrow H_a : P(Y_1^t - Y_1^c < Y_2^t - Y_2^c) + \frac{1}{2}P(Y_1^t - Y_1^c = Y_2^t - Y_q^c 2) \neq \frac{1}{2}. \end{aligned} \tag{2.16}$$

The term $\frac{1}{2}P(Y_1^t - Y_1^c = Y_2^t - Y_2^c)$ is used to account for possible ties for discrete distributions. Under the null hypothesis H_0, $E[I(Y_1^t - Y_1^c < Y_2^t - Y_1^c) + \frac{1}{2}I(Y_1^t - Y_1^c = Y_2^t - Y_2^c)]$ will be equal to $\frac{1}{2}$. Then the 4-sample U-statistic based on kernel function $\phi^{(1,2)}(y_1^t; y_1^c; y_2^t; y_2^c) = I(y_1^t - y_1^c < y_2^t - y_2^c) + \frac{1}{2}I(y_1^t - y_1^c = y_2^t - y_2^c)$ is

$$U^{(1,2)} = \frac{1}{n_1^t n_1^c n_2^t n_2^c} \sum_{i=1}^{n_1^t} \sum_{j=1}^{n_1^c} \sum_{k=1}^{n_2^t} \sum_{l=1}^{n_2^c} I(Y_{1i}^t - Y_{1j}^c < Y_{2k}^t - Y_{2l}^c) + \frac{1}{2} I(Y_{1i}^t - Y_{1j}^c = Y_{2k}^t - Y_{2l}^c). \tag{2.17}$$

Denoting $\theta^{(1,2)} = E(U^{(1,2)})$ and using the background results about multi-sample U-statistics with $r = 1$, with the assumption that $\frac{n_s^\omega}{N} \to \lambda_s^\omega (0 < \lambda_s^\omega < 1)$ as $N \to \infty$ and the fact that $E[(\phi^{(1,2)})^2] \leq 1$, we have

$$\sqrt{N}(U^{(1,2)} - \theta^{(1,2)}) \xrightarrow{D} N(0, \sigma_{1,2}^2), \tag{2.18}$$

where

$$\sigma_{1,2}^2 = \frac{1}{\lambda_1^t} Var(h_1^{t,(1,2)}(Y_1^t)) + \frac{1}{\lambda_1^c} Var(h_1^{c,(1,2)}(Y_1^c)) + \frac{1}{\lambda_2^t} Var(h_2^{t,(1,2)}(Y_2^t)) + \frac{1}{\lambda_2^c} Var(h_2^{c,(1,2)}(Y_2^c)) \in (0, \infty),$$

$$h_s^{\omega,(1,2)}(x) = E[\phi^{(1,2)}(Y_1^t; Y_1^c; Y_2^t; Y_2^c) | Y_s^\omega = x] - \theta^{(1,2)},$$

and assuming $Var\left(h_s^\omega(Y_s^\omega)\right) > 0$ for $s \in \{1,2\}, \omega \in \{t,c\}$.

The test based on this is consistent for hypotheses H_0 vs H_1 in (2.16).

To apply this method, we first estimate $h_s^{\omega,(1,2)}(x)$ ($s \in \{1,2\}$ and $\omega \in \{t,c\}$), an expectation, by the method of moments. For instance, $h_1^{t,(1,2)}(x)$ is estimated by the sample mean $\hat{h}_1^{t,(1,2)}(x) = \frac{1}{n_1^c n_2^t n_2^c} \sum_{j=1}^{n_1^c} \sum_{k=1}^{n_2^t} \sum_{l=1}^{n_2^c} I(x - Y_{1j}^c < Y_{2k}^t - Y_{2l}^c)$. Note that this calculation is repeated with each data value Y_{1i}^t ($i = 1, \cdots, n_1^t$) taking the place of x. Likewise for other h terms. Then we estimate $Var(h_s^{\omega,(1,2)}(Y_s^\omega))$ by the sample variance of $\hat{h}_s^{\omega,(1,2)}(Y_s^\omega)$ as $\frac{1}{n_s^\omega - 1} \sum_{i=1}^{n_s^\omega} [\hat{h}_s^{\omega,(1,2)}(Y_{si}^\omega) - \frac{1}{n_s^\omega} \sum_{j=1}^{n_s^\omega} \hat{h}_s^{\omega,(1,2)}(Y_{sj}^\omega)]^2$, for $s \in \{1,2\}$ and $\omega \in \{t,c\}$, and take the weighted sum of them to approximate $\sigma_{1,2}^2$.

2.3.2 Testing Treatment Effect Heterogeneity Across Multiple Strata

With S strata, the hypotheses we focus on are

$$H_0 : P(Y_p^t - Y_p^c < Y_q^t - Y_q^c) + \frac{1}{2} P(Y_p^t - Y_p^c = Y_q^t - Y_q^c) = \frac{1}{2} \text{ for any } 1 \le p < q \le S$$

$$\Longleftrightarrow H_a : \text{the equation does not hold for at least one pair of } (p,q). \tag{2.19}$$

For any pair of strata, we can construct a test statistic like (2.17). Denote $U^{(p,q)}(p < q)$ as the U-statistic comparing strata p and q with kernel $\phi^{(p,q)}(y_p^t; y_p^c; y_q^t; y_q^c) = I(y_p^t - y_p^c < y_q^t - y_q^c) + \frac{1}{2} I(y_p^t - y_p^c = y_q^t - y_q^c)$ and expectation $\theta^{(p,q)}$. By applying the Hájek projection

principle to a vector of multi-sample U-statistics as in Section 2.2, we have

$$\sqrt{N}\begin{pmatrix} U^{(1,2)} - \theta^{(1,2)} \\ U^{(1,3)} - \theta^{(1,3)} \\ \vdots \\ U^{(S-1,S)} - \theta^{(S-1,S)} \end{pmatrix} \xrightarrow{D} N(0,\Sigma) \tag{2.20}$$

where $\Sigma = \frac{1}{\lambda_1^t}\Sigma_1^t + \frac{1}{\lambda_1^c}\Sigma_1^c + \cdots + \frac{1}{\lambda_S^t}\Sigma_S^t + \frac{1}{\lambda_S^c}\Sigma_S^t$ and $\Sigma_s^\omega = Cov(\tilde{h}_s^{\omega,(1,2)}(Y_s^\omega), \cdots, \tilde{h}_s^{\omega,(1,S)}(Y_s^\omega), \tilde{h}_s^{\omega,(2,3)}(Y_s^\omega), \cdots, \tilde{h}_s^{\omega,(S-1,S)}(Y_s^\omega))$ for all $s \in \{1, \cdots, S\}$ and $\omega \in \{t, c\}$. Here

$$\tilde{h}_s^{\omega,(p,q)}(x) = \begin{cases} h_s^{\omega,(p,q)}(x) = E[\phi^{(p,q)}(Y_p^t; Y_p^c; Y_q^t; Y_q^c)|Y_s^\omega = x] - \theta^{(p,q)} & \text{if } s = p \text{ or } s = q, \\ 0 & o.w. \end{cases}$$

for $\{(p,q)|1 \le p < q \le S\}$. Under the null hypothesis H_0 in (2.19), all θ's are equal to $\frac{1}{2}$.

Estimation of Σ is carried out using a similar approach as described for estimation of $\sigma_{1,2}^2$ in Section 2.3.1. We first construct an empirical estimate for each h function (as in the paragraph below (2.18)) and then use the sample covariance matrix of $[\tilde{h}_s^{\omega,(1,2)}(Y_s^\omega), \cdots, \tilde{h}_s^{\omega,(1,S)}(Y_s^\omega), \tilde{h}_s^{\omega,(2,3)}(Y_s^\omega), \cdots, \tilde{h}_s^{\omega,(S-1,S)}(Y_s^\omega)]$ with $h's$ replaced by their corresponding estimates to get $\hat{\Sigma}_s^\omega$. Then $\hat{\Sigma}$, the estimate of Σ, is the sum of the $\hat{\Sigma}_s^\omega$ over $s \in \{1, \cdots, S\}$ and $\omega \in \{t, c\}$ with each term weighted by $\frac{1}{\lambda_s^\omega}$.

The vector of pairwise test statistics $U = (U^{(1,2)}, U^{(1,3)}, \cdots, U^{(S-1,S)})^T$ can be combined into a single overall test statistic using any function of U. Here we focus on $U_h = N \cdot \sum_{1 \le p < q \le S} (U^{(p,q)} - \frac{1}{2})^2$. The asymptotic distribution of U_h is not available in analytic form, but a simulation approach can be used to assess U_h. A large number of independent samples of $\sqrt{N}(U - \frac{1}{2})$ are generated from the null $N(0, \hat{\Sigma})$ distribution, and U_h is computed for each

sample to generate the empirical null distribution of U_h. For α level test, we reject H_0 when U_h is greater than or equal to the $100(1-\alpha)$ percentile of the empirical null distribution. Note that other test statistics are also possible, e.g., $\sqrt{N} \cdot \max_{1 \le p < q \le S} |U^{(p,q)} - \frac{1}{2}|$, and simulation is always an option for deriving the reference distribution. In our simulation study, we use the statistic U_h because it proved reliable. This test based on U_h is consistent for the hypotheses in (2.19).

Another test statistic that might seem natural is $T = N(U - \frac{1}{2}\mathbf{1})\hat{\Sigma}^{-}(U - \frac{1}{2}\mathbf{1})^T$ whose reference distribution is χ^2_k, where k is the rank of $\hat{\Sigma}$. $\hat{\Sigma}^{-} = \sum_{i=1}^{k} \frac{1}{\alpha_i} q_i q_i^t$ is a generalized inverse of $\hat{\Sigma}$, where $\{\alpha_i, i = 1, \cdots, k\}$ are the non-zero eigenvalues of $\hat{\Sigma}$ and $\{q_i, i = 1, \cdots, k\}$ are the corresponding orthogonal eigenvectors. However, as U is a vector of all of the pairwise U-statistics, the determinant of its covariance matrix can be very close to 0. Then $\hat{\Sigma}$ can have an eigenvalue α very close to 0, for which a tiny rounding error would have a large impact on $\hat{\Sigma}^{-}$ and thus on T. So even though the reference distribution of the test statistic T has known distribution, we prefer using U_h.

2.3.3 Three Particular Cases to Apply

The approach described above is non-parametric, it does not make any specific assumption about the shapes of the distributions of $Y_1^t, \cdots, Y_S^t, Y_1^c, \cdots, Y_S^c$ or the relationship among them other than independence. Our U-statistic is testing whether the probability $P(Y_p^t - Y_p^c < Y_q^t - Y_q^c) + \frac{1}{2}P(Y_p^t - Y_p^c = Y_q^t - Y_q^c)$ $(p \neq q)$ is equal to one half for all p, q. This does not however provide much insight into the feature of outcomes that is being tested. When there is no further assumptions of the outcome distributions at all, confusion may arise. We describe three cases here, with a set of semi-parametric assumptions for each of them, under which the interpretation of the test is clear. Many real-life problems may fit into those cases.

Case A: We assume all outcomes in different strata and different treatment groups follow a

common distribution F up to a location shift, which is comprised of the additive treatment effects τ_s within each stratum and assumed additive stratum effects Δ_s for $s \in \{1, \cdots, S\}$. In this case, $P(Y_p^t - Y_p^c < Y_q^t - Y_q^c) + \frac{1}{2}P(Y_p^t - Y_p^c = Y_q^t - Y_q^c)$ is equal to a half if and only if the within-stratum location shifts (τ_p and τ_q) are the same. It can then be easily shown that our test is consistent with identifying whether τ_s is the same across all strata. Since the shift can be considered as the difference in means or difference in any percentiles between two different treatment groups, the test is consistent with respect to the alternative hypothesis that the difference of means (or percentiles) are unequal for at least one pair of strata. This case is equivalent to testing the interaction in a two-factor factorial design when one factor has two levels; in our case, one factor is stratum and the other is treatment which takes two levels. The ANOVA F-test can be used to address this scenario, and non-parametric tests have been proposed as an alternative (Patel and Hoel, 1973; De Neve and Thas, 2017). The De Neve and Thas (2017) approach is similar to our approach in that they use the same basic U-statistic that we do. They use a different summary statistic to aggregate the pairwise comparison. A limitation is that their approach applies only to this case (*Case A*) and not to *Case B* or *Case C* that are described next.

Case B: We assume that all outcomes in the treatment groups follow a common distribution F^t up to a stratum-specific location shift Δ_s^t and all outcomes in the control groups follow a possibly different common distribution F^c up to a stratum-specific location shift Δ_s^c for all $s \in \{1, \cdots, S\}$. Then the within-stratum treatment effect can be depicted as the difference of the location shift $\Delta_s^t - \Delta_s^c$ plus the difference of F^t and F^c. No matter what metric is used to describe the difference of F^t and F^c, it is consistent across all strata. So in this case, $P(Y_p^t - Y_p^c < Y_q^t - Y_q^c) + \frac{1}{2}P(Y_p^t - Y_p^c = Y_q^t - Y_q^c)$ is $\frac{1}{2}$ if and only if $\Delta_p^t - \Delta_p^c = \Delta_q^t - \Delta_q^c$. So our test is identifying whether the difference of location shifts $\Delta_s^t - \Delta_s^c$ are consistent across strata, which can also be considered as the consistency of the difference of means or percentiles between different treatment groups. We can easily prove the test is consistent with the alternative hypothesis that the difference of means (or percentiles) between the

treatment and control groups are not all identical across strata.

Case C: We assume that the outcomes in the same strata $s \in \{1, \cdots, S\}$ follow the same distribution F_s up to a stratum-specific location shift τ_s, the additive treatment effect. Similarly, in this case, our method is a consistent test with respect to the alternative hypothesis that there are at least two strata such that the difference of means (or percentiles) between the treatment and control groups are different.

The semi-parametric assumption in *Case A* is considered to be reasonable in practice when the factor we use to stratify subjects, as well as the treatment, only shifts the location of the outcome distributions but cannot change the shape. This is a common assumption. The assumption in *Case B* is considered to be reasonable when the treatment changes the shape of the outcome distributions, whereas the factor used to stratify subjects would only shift the location. Suppose we are studying whether the effect of a welfare reform policy on household income differs across several different geographic regions. Usually a welfare reform plan has different magnitude of impact on families who are on different economic levels. Thus we may expect the distribution of income after the reform to differ from the control condition. If the shape of income distribution is similar across geographic regions or just differs by scale among different regions, then this scenario could fit in our *Case B*. The assumption in *Case C* is considered to be reasonable when the treatment effects are constant within each stratum, but the distribution of outcomes differs across strata. Back to the welfare reform example. If we stratify subjects according to their economic levels, then income distributions would be expected to vary among different strata, but it is possible that treatment effects could be constant.

2.4 Simulation Study

We demonstrate the U-statistic test of treatment effect heterogeneity via a simulation study. There are some computational challenges that are addressed first, then the simulation study is described and results are provided. Simulations compare the U test to the LRT in a range of scenarios, some of which match the assumptions in the parametric test and others do not. Additionally, we show the relationship among power, sample size and effect size in two scenarios.

2.4.1 Computational Issues

When the sample sizes are large, the computation of the U-statistic is computationally expensive. Let's take $U^{(1,2)}$ as an example. We need to compute the average over all the combinations of $I(Y_{1i}^t - Y_{1j}^c < Y_{2k}^t - Y_{2l}^c) + \frac{1}{2}I(Y_{1i}^t - Y_{1j}^c = Y_{2k}^t - Y_{2l}^c)$, denoted by $\phi^{(1,2)}(i,j,k,l)$, which includes $n_1^t \times n_1^c \times n_2^t \times n_2^c$ terms. As this computation can be done in parallel for different (i,j,k,l), it should not be a big problem when applying the method for a single data set. In simulations, we need to generate thousands of data sets and compute U-statistics for each of them. So in the simulation study, instead of computing the average through exhaustive enumeration, we generate approximate U-statistics by randomly selecting some of the combinations with replacement and use this average to approximate the U-statistic. We randomly selected $M = 10^3 N$ samples with replacement from each treatment subgroup as $\{y_{1i}^t, y_{1i}^c, y_{2i}^t, y_{2i}^c, i = 1, \cdots, M\}$ to approximate $U^{(1,2)}$, here N is the total sample size of the two strata. The sampling size M was determined by considering a range of scenarios and estimating the variance of the test statistic $U_h = N \cdot \sum_{1 \le i < j \le S} (U^{(i,j)} - \frac{1}{2})^2$ constructed by approximate U-statistics within each scenario via simulation. The variance increases as N increases. In the simulation, as N ranged from 60 to 3000, the maximum estimated variances for this choice of M ranged from 0.0058 to 0.104, which was judged to

provide sufficient precision. One requirement of this sampling is that all subjects have to be selected at least once, because we also used the sampled indicators to estimate $h_s^{\omega,(1,2)}(Y_{si}^{\omega})$ $(s \in \{1,2\}, \omega \in \{t,c\}, i \in \{1,\cdots,n_s^{\omega}\})$. For instance, the estimate of $h_1^{t,(1,2)}(Y_{11}^t)$ is computed as the average of all selected $\phi^{(1,2)}(i,j,k,l)$ with Y_{1i}^t equal to Y_{11}^t. Though this requirement is not a challenge due to the large sampling size M, in the rare events that it occurs we would need to redo the sampling procedure. As for the empirical reference distribution of the test statistic U_h, we generated 10^5 random samples $r_i = (r_i^{(1,2)},\cdots,r_i^{(S-1,S)})$ $(i = 1,\cdots,10^5)$ from the multivariate normal distribution $N(0,\hat{\Sigma})$ for each simulation, and got the distribution of $||r_i||^2$ as the empirical reference distribution under H_0. The empirical p-value is the percentage of generated $||r_i||^2$ greater than U_h. Fixing the type I error as $\alpha = 0.05$, we reject H_0 when the p-value is smaller than α. We determined the sample size 10^5 by considering a range of scenarios and estimating the variance associated with a simulation-based 95th percentile. The sample size of 10^5 in the various cases makes the variance less than 0.01.

2.4.2 Review of the Likelihood Ratio Test for Treatment Effect Heterogeneity

The likelihood ratio test for treatment effect heterogeneity was developed by Gail and Simon (1985). Let τ_s denote the treatment effect in subgroup s $(s \in \{1,\cdots,S\})$. The test assesses the null hypotheses $H_0 : \tau_1 = \cdots = \tau_S$ versus the alternative that at least two of the subgroup treatment effects are unequal. Under the assumption that $\hat{\tau}_s (s \in \{1,\cdots,S\})$ follows a normal distribution with

$$\hat{\tau}_s \overset{indep}{\sim} N(\tau_s, \sigma_s^2), s \in \{1,\cdots,S\}, \tag{2.21}$$

we have heterogeneity test statistic

$$H = \sum_{s=1}^{S} (\hat{\tau}_s - \bar{\hat{\tau}})^2 / s_s^2 \overset{H_0}{\sim} \chi^2_{S-1} \tag{2.22}$$

where $\bar{\hat{\tau}} = (\sum_{s=1}^{S} \hat{\tau}_s / s_s^2) / (\sum_{s=1}^{S} 1/s_s^2)$, and s_s^2 is a consistent estimator of σ_s^2.

With fixed type I error α, we reject the null hypotheses H_0 when the test statistic H is greater than or equal to the $100(1-\alpha)th$ percentile of χ^2_{S-1}.

For additive treatment effects, the treatment effect estimates $\hat{\tau}_s$ can be the difference between the sample means of the two treatment groups within strata s. When the subgroup sample sizes are large, according to the Central Limit Theorem, $\hat{\tau}_s$ will approximate to a normal distribution. However, when the distributions of the outcomes differ from normality and the validity of the test relies on large sample sizes, the power of the test will be impacted as with other parametric tests (Lehmann, 2004).

2.4.3 Simulation Study Design

Assuming we have three strata, we generated n_s^ω random samples from treatment subgroup ω within strata s from a distribution F_s^ω ($s \in 1, 2, 3; \omega \in \{t, c\}$). The choices of n_s^ω and F_s^ω are described below. The hypothesis of no heterogeneity was tested via our non-parametric U test and the LRT reviewed in the previous section. For each simulation scenario (choices of n_s^ω and F_s^ω), we generate $L = 2000$ data sets and carry out the tests on each. This yields rejection rates and the empirical distribution of p-values.

We developed 17 different scenarios for the choice of the distributions F_s^ω. These scenarios are listed in Table 2.1. They are organized according to the three application cases outlined in Section 2.3.3. For each scenario, the true treatment effects were varied to provide the null

and alternative instances. The upper half of Table 2.1 lists the null cases and the lower half lists the alternative cases.

For *Case A*, outcomes in the three strata and two treatment groups follow a common distribution F up to location shifts. The first scenario ($A1$) is that $F = N(0, 1)$. For the next two scenarios ($A2$ and $A3$), F are still symmetric distributions, but with tails lighter ($A2$) or heavier ($A3$) than normal distribution. Next we consider three skewed distributions χ^2_1, $Exp(1)$ and χ^2_4 (labeled as $A4 - A6$) with their skewnesses decreasing in that order. The support of these distributions are all positive. We are generally more interested in comparing the scales of the treatment and control groups instead of location shifts in these scenarios. So we suppose there are constants c^ω_s such that $\frac{Y^\omega_s}{c^\omega_s} \sim F$ $(s \in 1, 2, 3; \omega \in \{t, c\})$. Here we use the logarithm of those outcomes in our test statistic U_h instead of using the original outcomes directly. Now the problem of testing the consistency of ratio of scales is changed into a problem of testing the consistency of the location shift $log(c^t_s) - log(c^c_s)$ $(s = 1, \cdots, S)$. The final *Case A* example is a bimodal distribution $0.5N(-5, 1) + 0.5N(5, 1)$ (labeled as $A7$).

In *Case B*, all outcomes in the treatment groups follow a common distribution F^t up to a location shift and all outcomes in the control group follow a different common distribution F^c up to a location shift. To create examples here, we choose two of the distributions used in *Case A* whose supports are the whole real line ($A1 - A3$, $A7$) and randomly assign them to the treatment and control group. We try all $\binom{4}{2}$ combinations, they are labeled as $B1 - B6$.

In *Case C* where all outcomes in the same stratum follow a common distribution F_s $(s = 1, \cdots, S)$ up to a location shift, we select three distribution from *Case A* with support on the whole real line and randomly assign them to the three strata. So we have $\binom{4}{3}$ combinations and they are labeled as scenarios $C1 - C4$.

For each scenario described above, we vary the true treatment effects to get the null and alternative cases, and also consider a range of different sample sizes. In each scenario, there

Scenario	F_1^t	F_1^c	F_2^t	F_1^c	F_3^t	F_3^c
Null Cases						
$A1$	$N(0,1)$	$N(0,1)-1$	$N(0,1)+1$	$N(0,1)$	$N(0,1)+2$	$N(0,1)+1$
$A2$	$U(-2,2)$	$U(-2,2)-1$	$U(-2,2)+1$	$U(-2,2)$	$U(-2,2)+2$	$U(-2,2)+1$
$A3$	t_4	t_4-1	t_4+1	t_4	t_4+2	t_4+1
$A4$	$e^1 \cdot \chi_1^2$	χ_1^2	$e^2 \cdot \chi_1^2$	$e^1 \cdot \chi_1^2$	$e^3 \cdot \chi_1^2$	$e^2 \cdot \chi_1^2$
$A5$	$e^1 \cdot Exp(1)$	$Exp(1)$	$e^2 \cdot Exp(1)$	$e^1 \cdot Exp(1)$	$e^3 \cdot Exp(1)$	$e^2 \cdot Exp(1)$
$A6$	$e^1 \cdot \chi_4^2$	χ_4^2	$e^2 \cdot \chi_4^2$	$e^1 \cdot \chi_4^2$	$e^3 \cdot 6\chi_4^2$	$e^2 \cdot \chi_4^2$
$A7$	$0.5N(-5,1)+0.5N(5,1)$	$0.5N(-5,1)+0.5N(5,1)-1$	$0.5N(-5,1)+0.5N(5,1)+1$	$0.5N(-5,1)+0.5N(5,1)$	$0.5N(-5,1)+0.5N(5,1)+2$	$0.5N(-5,1)+0.5N(5,1)+1$
$B1$	$N(0,1)$	$U(-2,2)$	$N(0,1)+1$	$U(-2,2)+1$	$N(0,1)+2$	$U(-2,2)+2$
$B2$	$N(0,1)$	t_4	$N(0,1)+1$	t_4+1	$N(0,1)+2$	t_4+2
$B3$	$N(0,1)$	$0.5N(-5,1)+0.5N(5,1)$	$N(0,1)+1$	$0.5N(-5,1)+0.5N(5,1)+1$	$N(0,1)+2$	$0.5N(-5,1)+0.5N(5,1)+2$
$B4$	$U(-2,2)$	t_4	$U(-2,2)+1$	t_4+1	$U(-2,2)+2$	t_4+2
$B5$	$U(-2,2)$	$0.5N(-5,1)+0.5N(5,1)$	$U(-2,2)+1$	$0.5N(-5,1)+0.5N(5,1)+1$	$U(-2,2)+2$	$0.5N(-5,1)+0.5N(5,1)+2$
$B6$	t_4	$0.5N(-5,1)+0.5N(5,1)$	t_4+1	$0.5N(-5,1)+0.5N(5,1)+1$	t_4+2	$0.5N(-5,1)+0.5N(5,1)+2$
$C1$	$N(0,1)$	$N(0,1)-1$	$U(-2,2)$	$U(-2,2)-1$	t_4	t_4-1
$C2$	$N(0,1)$	$N(0,1)-1$	$U(-2,2)$	$U(-2,2)-1$	$0.5N(-5,1)+0.5N(5,1)$	$0.5N(-5,1)+0.5N(5,1)-1$
$C3$	$N(0,1)$	$N(0,1)-1$	t_4	t_4-1	$0.5N(-5,1)+0.5N(5,1)$	$0.5N(-5,1)+0.5N(5,1)-1$
$C4$	$U(-2,2)$	$U(-2,2)-1$	t_4	t_4-1	$0.5N(-5,1)+0.5N(5,1)$	$0.5N(-5,1)+0.5N(5,1)-1$
Alternative Cases						
$A1$	N(0,1)	N(0,1)-1	N(0,1)+1	N(0,1)+1-1.25	N(0,1)+2	N(0,1)+2-1.5
$A2$	$U(-2,2)$	$U(-2,2)-1$	$U(-2,2)+1$	$U(-2,2)+1-1.1$	$U(-2,2)+2$	$U(-2,2)+2-1.2$
$A3$	t_4	t_4-1	t_4+1	$t_4+1-1.25$	t_4+2	$t_4+2-1.5$
$A4$	$e^1 \cdot \chi_1^2$	χ_1^2	$e^{2.5} \cdot \chi_1^2$	$e^1 \cdot \chi_1^2$	$e^4 \cdot 9\chi_1^2$	$e^2 \cdot \chi_1^2$
$A5$	$e^1 \cdot Exp(1)$	$Exp(1)$	$e^{2.25} \cdot Exp(1)$	$e^1 \cdot Exp(1)$	$e^{3.5} \cdot Exp(1)$	$e^2 \cdot Exp(1)$
$A6$	$e^1 \cdot \chi_4^2$	χ_4^2	$e^{2.25} \cdot \chi_4^2$	$e^1 \cdot \chi_4^2$	$e^{3.5} \cdot \chi_4^2$	$e^2 \cdot \chi_4^2$
$A7$	$0.5N(-5,1)+0.5N(5,1)$	$0.5N(-5,1)+0.5N(5,1)-1$	$0.5N(-5,1)+0.5N(5,1)+1$	$0.5N(-5,1)+0.5N(5,1)+1-2$	$0.5N(-5,1)+0.5N(5,1)+2$	$0.5N(-5,1)+0.5N(5,1)+2-3$
$B1$	$N(0,1)$	$U(-2,2)$	$N(0,1)+1$	$U(-2,2)+1-0.25$	$N(0,1)+2$	$U(-2,2)+2-0.5$
$B2$	$N(0,1)$	t_4	$N(0,1)+1$	$t_4+1-0.25$	$N(0,1)+2$	$t_4+2-0.5$
$B3$	$N(0,1)$	$0.5N(-5,1)+0.5N(5,1)$	$N(0,1)+1$	$0.5N(-5,1)+0.5N(5,1)+1-1$	$N(0,1)+2$	$0.5N(-5,1)+0.5N(5,1)+2-2$
$B4$	$U(-2,2)$	t_4	$U(-2,2)+1$	$t_4+1-0.25$	$U(-2,2)+2$	$t_4+2-0.5$
$B5$	$U(-2,2)$	$0.5N(-5,1)+0.5N(5,1)$	$U(-2,2)+1$	$0.5N(-5,1)+0.5N(5,1)+1-1$	$U(-2,2)+2$	$0.5N(-5,1)+0.5N(5,1)+2-2$
$B6$	t_4	$0.5N(-5,1)+0.5N(5,1)$	t_4+1	$0.5N(-5,1)+0.5N(5,1)+1-1$	t_4+2	$0.5N(-5,1)+0.5N(5,1)+2-2$
$C1$	$N(0,1)$	$N(0,1)-1$	$U(-2,2)$	$U(-2,2)-1.25$	t_4	$t_4-1.5$
$C2$	$N(0,1)$	$N(0,1)-1$	$U(-2,2)$	$U(-2,2)-1.5$	$0.5N(-5,1)+0.5N(5,1)$	$0.5N(-5,1)+0.5N(5,1)-2$
$C3$	$N(0,1)$	$N(0,1)-1$	t_4	$t_4-1.5$	$0.5N(-5,1)+0.5N(5,1)$	$0.5N(-5,1)+0.5N(5,1)-2$
$C4$	$U(-2,2)$	$U(-2,2)-1$	t_4	$t_4-1.5$	$0.5N(-5,1)+0.5N(5,1)$	$0.5N(-5,1)+0.5N(5,1)-2$

Table 2.1: Simulation scenarios

are stratum-specific location shifts ($\Delta_1 = 0, \Delta_2 = 1, \Delta_3 = 2$ for scenarios in *Case A* and *Case B*). For all null cases, the stratum treatment effects are the same across the three strata. For alternative cases, the treatment effects (τ_1, τ_2, τ_3) form an arithmetic series with $\tau_2 = \tau_1 + \Gamma$ and $\tau_3 = \tau_1 + 2\Gamma$, where $\Gamma > 0$. Simulations were carried out for a range of values of Γ. If Γ is too large, then both tests always reject the null hypothesis for almost any sample size. Results are presented for a representative choice of Γ where this does not occur. All simulation scenarios were investigated with six assumptions regarding sample sizes. First, all subgroup sample sizes are the same ($n_s^\omega = n$) with n equal to 10, 50, 100 and 500. Second, the sample sizes of treatment and control within each stratum are the same, but sample size varies across strata. We tried $(n_1^\omega, n_2^\omega, n_3^\omega)$ ($\omega \in \{t, c\}$) equal to $(50, 100, 150)$ and $(150, 100, 50)$. The former corresponds to the case that the strata with higher effect sizes have larger sample sizes. The latter is the opposite case.

2.4.4 Simulation Study Results

The rejection rates of both the U test and the LRT assuming $\alpha = 0.05$ for all null cases of the different scenarios are provided in Table 2.2. The first column of Table 2.2 shows the labels of all scenarios. Each of the remaining columns corresponds to one sample size setting for all scenarios. As expected all rejection rates are close to the 0.05 level. Since empirical type I errors were computed by generating data for $L = 2000$ times, the standard error for each is approximately 0.005. The table shows that when $n_s^\omega = 10$ ($s \in \{1, 2, 3\}; \omega \in \{t, w\}$), the type I errors of both tests are a bit too high in all scenarios except for $A4$ and $C4$. For scenario $C3$, when (n_1, n_2, n_3) is equal to $(50, 50, 50)$ or $(150, 100, 50)$, type I errors of both tests are a bit too high. In all other settings, the type I errors are well controlled for both tests.

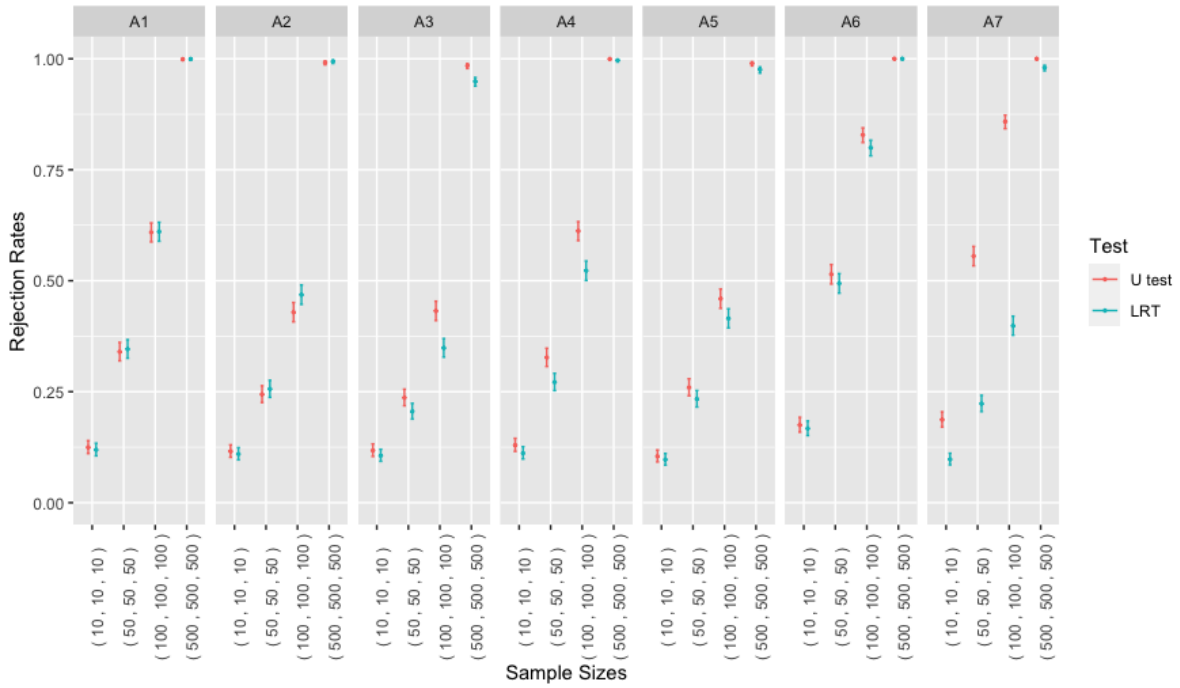
We then compare the power of the two tests by comparing their rejection rates for the alter-

$(n_1^\omega = n_1, n_2^\omega = n_2, n_2^\omega = n_3)$	(10,10,10)		(50,50,50)		(100,100,100)		(500,500,500)		(50,100,150)		(150,100,50)	
Test	U test	LRT	U test	LRT	U test	LRT	U test	LRT	U test	LRT	U test	LRT
Scenario												
$A1$	0.075	0.068	0.048	0.045	0.048	0.044	0.053	0.052	0.054	0.048	0.054	0.051
$A2$	0.071	0.07	0.048	0.046	0.055	0.054	0.051	0.051	0.054	0.06	0.049	0.049
$A3$	0.078	0.065	0.056	0.059	0.056	0.051	0.051	0.056	0.064	0.055	0.052	0.046
$A4$	0.058	0.052	0.044	0.045	0.051	0.047	0.049	0.05	0.052	0.051	0.053	0.056
$A5$	0.067	0.064	0.06	0.056	0.056	0.056	0.052	0.053	0.052	0.051	0.054	0.054
$A6$	0.07	0.064	0.045	0.044	0.057	0.054	0.046	0.044	0.052	0.054	0.048	0.046
$A7$	0.074	0.073	0.049	0.052	0.047	0.051	0.06	0.056	0.055	0.052	0.048	0.048
$B1$	0.082	0.076	0.05	0.052	0.051	0.049	0.052	0.055	0.057	0.053	0.061	0.062
$B2$	0.069	0.066	0.049	0.048	0.046	0.044	0.044	0.049	0.058	0.05	0.053	0.046
$B3$	0.07	0.089	0.042	0.052	0.044	0.044	0.04	0.046	0.057	0.06	0.053	0.048
$B4$	0.082	0.074	0.054	0.051	0.057	0.056	0.059	0.058	0.052	0.048	0.05	0.054
$B5$	0.079	0.09	0.056	0.052	0.049	0.052	0.06	0.053	0.052	0.052	0.054	0.047
$B6$	0.066	0.079	0.056	0.056	0.044	0.048	0.043	0.048	0.052	0.054	0.053	0.059
$C1$	0.076	0.07	0.06	0.06	0.046	0.044	0.052	0.055	0.052	0.046	0.054	0.052
$C2$	0.072	0.073	0.062	0.062	0.051	0.05	0.054	0.054	0.05	0.048	0.052	0.05
$C3$	0.066	0.073	0.066	0.067	0.05	0.053	0.051	0.045	0.052	0.049	0.072	0.066
$C4$	0.064	0.06	0.044	0.05	0.052	0.052	0.054	0.054	0.059	0.059	0.052	0.055

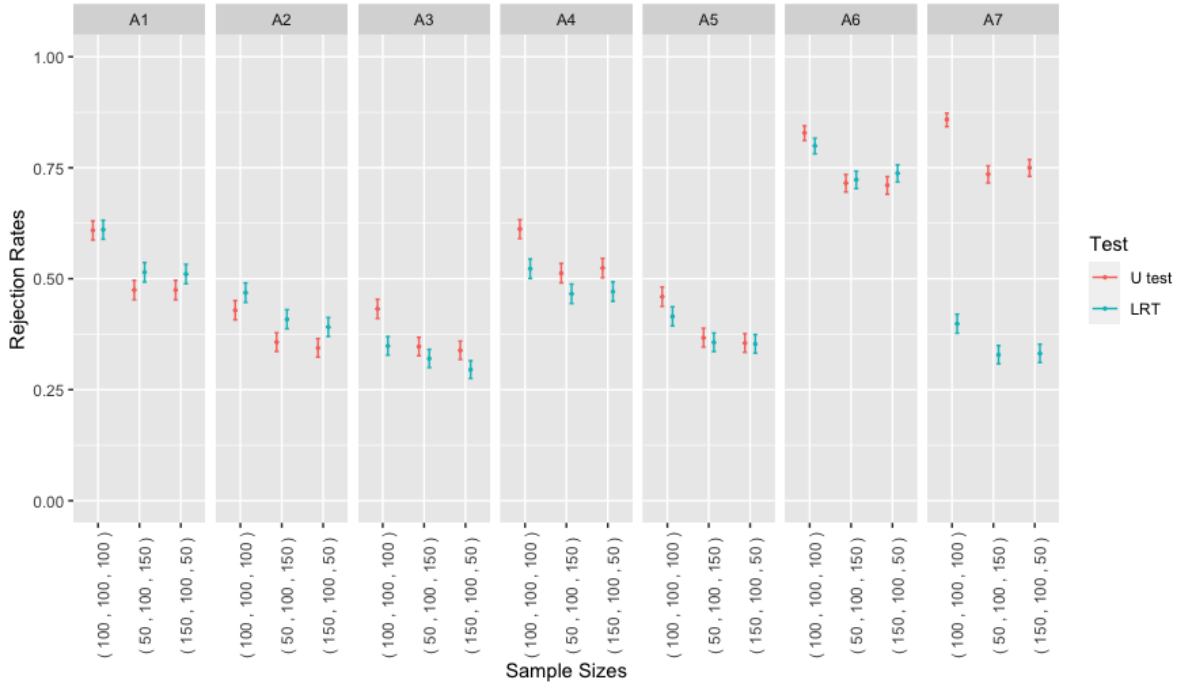
Table 2.2: Rejection rates of null cases under various settings

native cases in all sample size settings. Figure 2.1, 2.2 and 2.3 show the results for scenarios in *Case A*, *Case B* and *Case C* separately. Each figure is comprised of two subfigures (a) and (b). Subfigure (a) shows rejection rates for all cases where sample sizes n_s^ω ($s \in \{1, 2, 3\}$) are equal. Subfigure (b) focuses on the three cases whose stratum-specific sample sizes can vary. Within each subfigure, there is a set of panes, each of which corresponds to a scenario. Within each pane, the vertical axis indicates the rejection rate and the horizon axis indicates the sample size setting. For each type of test, we plot a point showing the empirical rejection rate and a line showing the corresponding 95% confidence interval. The red ones are for our proposed U test, and the blue ones are for the LRT.

For *Case A*, Figure 2.1(a) shows that when stratum-specific sample sizes are equal, the powers of both tests increase as n increases. When the common distribution F is normal ($A1$) or F is symmetric with tails lighter than normal ($A2$), the power of the LRT is a bit higher than the U test. When F is symmetric with tails heavier than normal ($A3$), or F is skewed ($A4$ - $A6$) or bimodal ($A7$), the U test is more powerful than the LRT. Also as F departs more from the Gaussian distribution, the advantage of the U test over the LRT is more substantial. Figure 2.1(b) shows the cases when subgroup sample sizes average 100 but

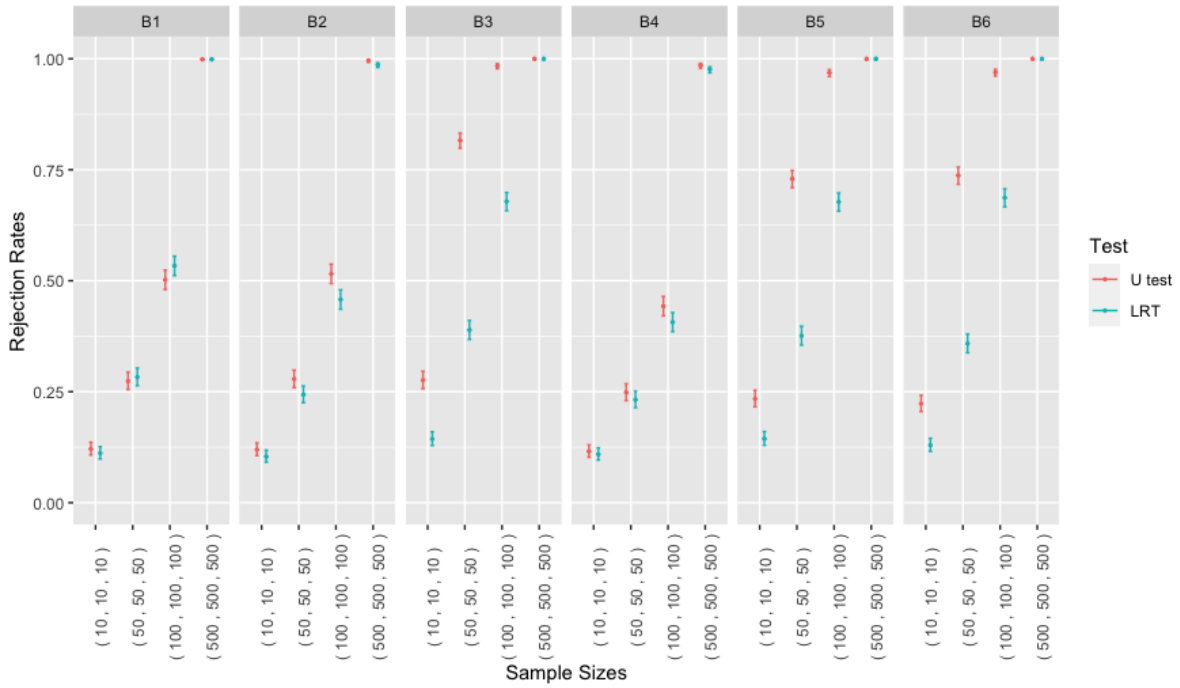


(a) Cases where sample sizes are consistent across strata

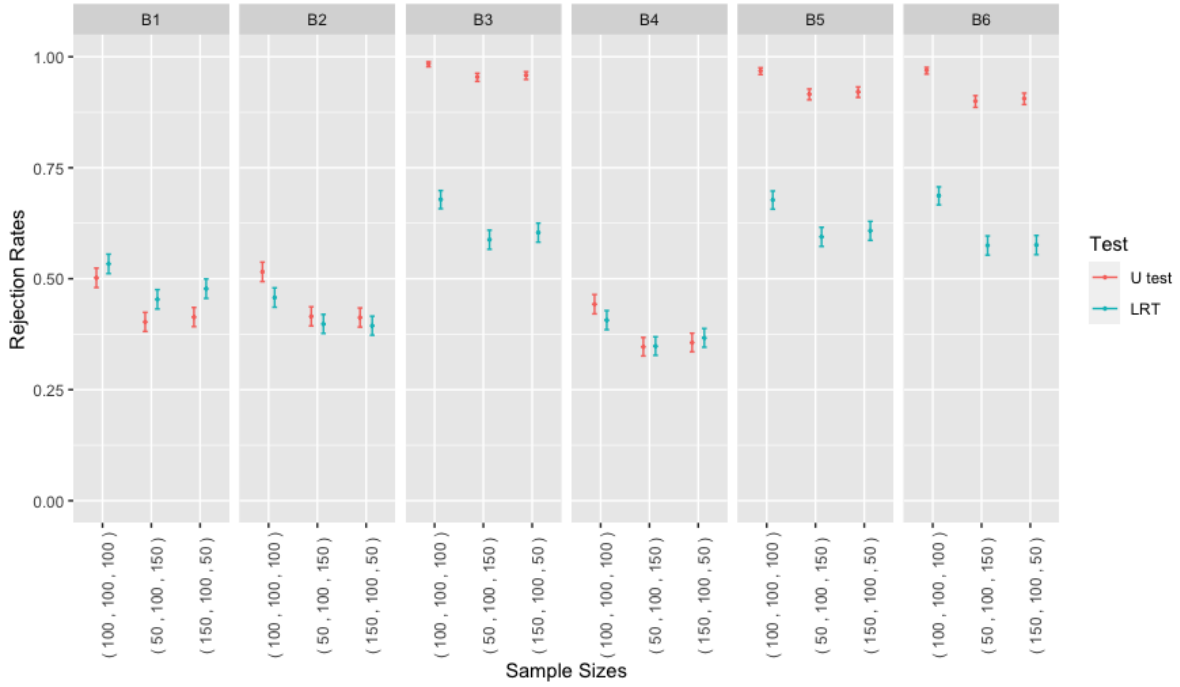


(b) Cases where sample sizes can change across strata

Figure 2.1: Rejection rates and their 95% confidence intervals of alternative cases in *Case A*

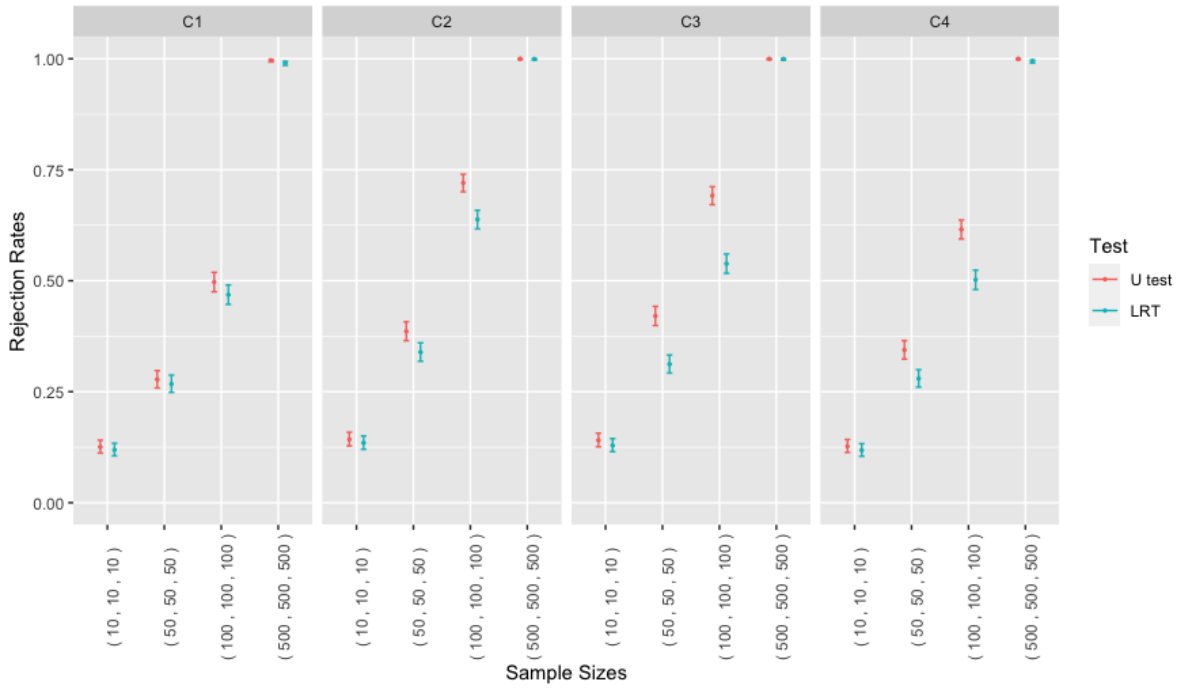


(a) Cases where sample sizes are consistent across strata

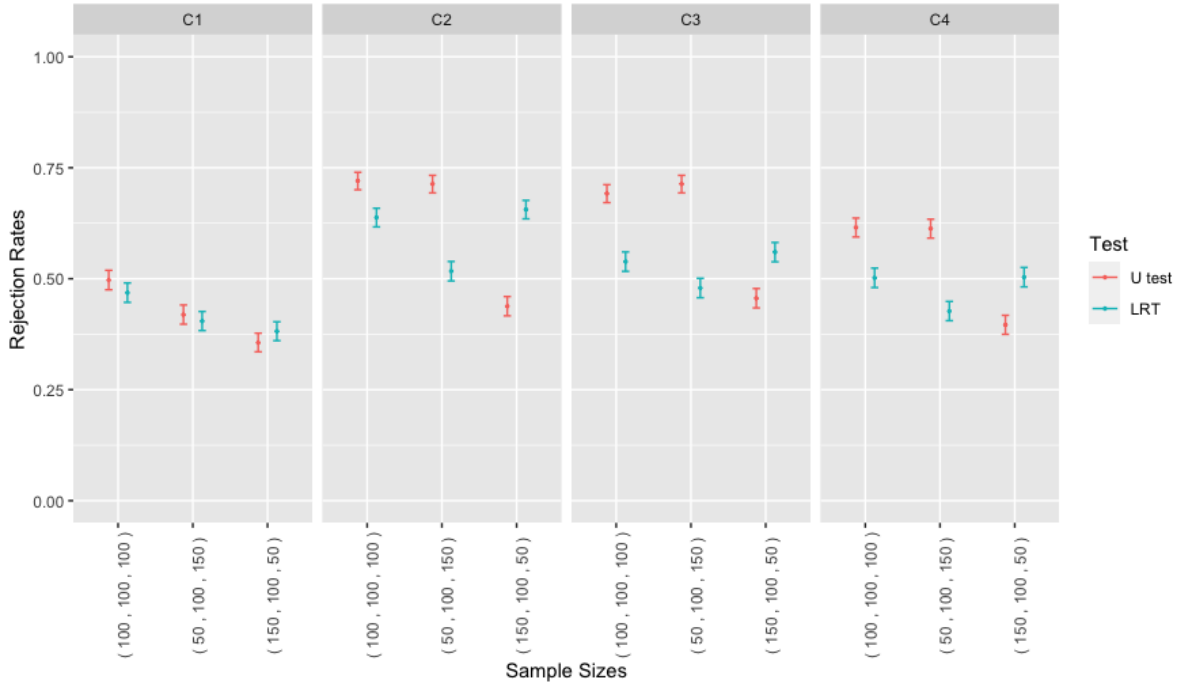


(b) Cases where sample sizes can change across strata

Figure 2.2: Rejection rates and their 95% confidence intervals of alternative cases in *Case B*



(a) Cases where sample sizes are consistent across strata



(b) Cases where sample sizes can change across strata

Figure 2.3: Rejection rates and their 95% confidence intervals of alternative cases in *Case C*

vary across the strata. Compared to the cases with equal sample sizes, the power of both the U test and the LRT drop, and the U test power drops a bit more than the LRT.

Figure 2.2 shows the rejection rates of the U test and the LRT for *Case B*. The results of cases with equal sample sizes across strata are in Figure 2.2(a). As the sample size increases, the powers of both tests increase in all scenarios. When the distributions in both treatment groups are close to normal ($B1$), the LRT is more powerful, otherwise the U test is more powerful. When one of the distributions is very far from normal ($B3$, $B5$ and $B6$), the advantage of the U test over the LRT is large. Next we compare the cases with average subgroup sample sizes all equal to 100 but where sample sizes can vary across strata (Figure 2.2(b)). As with *Case A*, the results with different stratum-specific sample sizes indicate less power than the case with equal stratum-specific sample sizes for both the U test and the LRT.

The empirical power of the two tests for *Case C* are displayed in Figure 2.3. When the stratum-specific sample sizes are equal (Figure 2.3(a)), the power of both tests increase as sample size increases. In all scenarios, the U test outperforms the LRT, and when there is a bimodal distribution ($C2$, $C3$, $C4$), the advantage of the U test is substantial. Figure 2.3(b) shows the effect of varying sample sizes across strata. When the distributions of the data in the three strata have similar variances and none of them are too far from normal ($C1$), the comparison result is similar to that in *Case A* and *Case B*. The setting with consistent stratum-specific sample sizes has largest power for both the U test and the LRT, and the difference of powers between balanced sample size setting and unbalanced sample size setting is larger for the U test than the LRT. When one of the strata follows a distribution that is mixture normal ($0.5N(-5,1) + 0.5N(5,1)$) which has a lot larger variance than the other two distributions and also departs more from the normal, the performances of the two tests are very different. For the U test, if the stratum with large-variance distribution has the smallest sample size, it is least powerful. For the LRT, when the stratum with distribution

very far from normal has the largest sample size, it is least powerful.

2.4.5 Investigating the Power of the Tests

The results in Section 2.4.4 focus on only a single non-null example for each scenario. This section investigate the power as a function of sample size for different treatment effects. We use the scenarios $A3$ and $A4$, and carried out simulations as described in the previous section.

For scenario $A3$, we generated three strata with $n_s^\omega = n$ ($s \in \{1,2,3\}, \omega \in \{t,c\}$), and generated random samples from t_4 distribution with location shifts comprised of strata effects $\Delta_1 = 0, \Delta_2 = 1, \Delta_3 = 2$ and additive treatment effects $\tau_1 = 1$, $\tau_2 = 1 + \Gamma$ and $\tau_3 = 1 + 2\Gamma$. Here the sequence of the treatment effects $\{\tau_1, \tau_2, \tau_3\}$ is arithmetic and we treat Γ as the effect size. For each fixed effect size, we explore the relationship between the sample size n and the rejection rates for both tests, and the results are shown in Figure 2.4 with Γ ranging from 0 to 0.5, and n ranging from 10 to 1000. In alternative cases when $\Gamma > 0$, with each fixed Γ, as n increases, the rejection rates of both tests increase, and the power of the U test is always higher than the LRT for each n.

For scenario $A4$, again we generated three strata with $n_s^\omega = n$ ($s \in \{1,2,3\}, \omega \in \{t,c\}$). With $\frac{Y_s^\omega}{c_s^\omega} \sim \chi_1^2$, we took the logarithm of Y_s^ω as the outcome. So the treatment effect is defined as $\tau_s = log(c_s^t) - log(c_s^c)$. We focus on the case where the sequence of the treatment effects $\{\tau_1, \tau_2, \tau_3\}$ are arithmetic with $\tau_2 = \tau_1 + \Gamma$ and $\tau_3 = \tau_1 + 2\Gamma$, and Γ is the effect size. By fixing $log(c_1^c) = 0, log(c_2^c) = 1, log(c_3^c) = 2$ and $log(c_1^t) = 1$, we have $\tau_1 = 1$. Then we can get different values of Γ by changing the values of c_2^t and c_3^t. Then for each Γ, we can explore the relationship between the rejection rates of our U test and the LRT as the sample size n varies. Figure 2.5 shows the rejection rates with n ranging from 10 to 1000 when Γ ranging from 0 to 1. As we would expect, with fixed n, larger effect size leads to larger rejection rates. The rejection rates of the U test is always larger than the LRT for each fixed n and

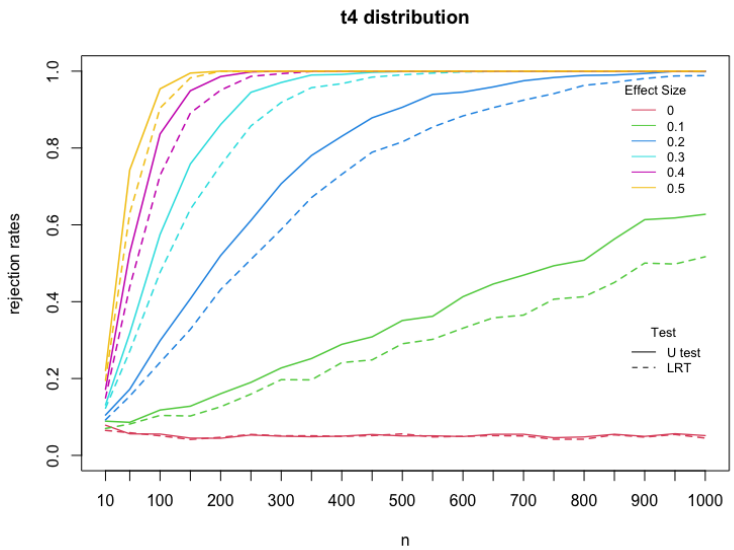


Figure 2.4: Relationship between rejection rates and sample sizes for t_4 distribution

Γ.

2.5 Case Study

In this section, we apply our proposed U test to a randomized data set from a program evaluation study in labor economics, an evaluation of the National Supported Work (NSW) Demonstration. The NSW is a labor training program conducted in the mid-1970s aiming at providing work experience to people with economic difficulties. Please refer to LaLonde (1986) and Dehejia and Wahba (1999) for details about the program. We use a subset of the LaLonde (1986) data that was created and used by Dehejia and Wahba (1999). The data are available at `https://users.nber.org/~rdehejia/data/.nswdata2.html`. These data described results from male participants with earnings information available for 1974. Earnings in 1978 were treated as outcome, and several pretreatment variables were recorded. There are 185 subjects in treatment group and 260 subjects in control group.

In this randomized study, pretreatment variables should have the same distribution between

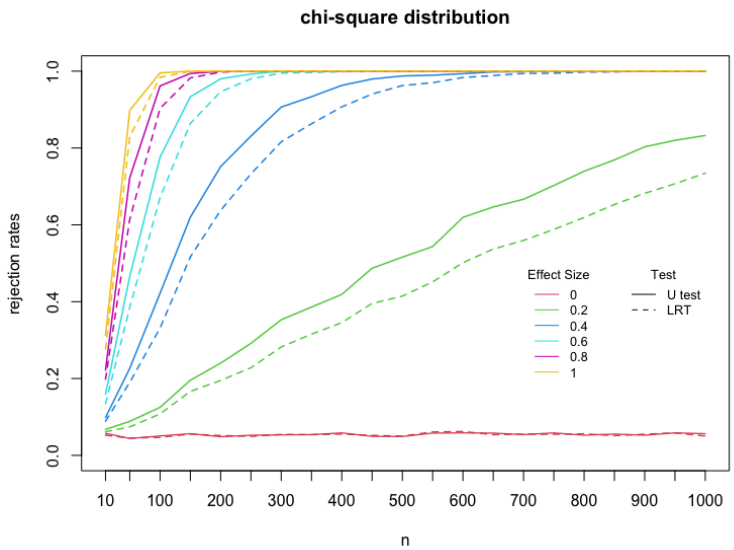


Figure 2.5: Relationship between rejection rates and sample sizes for χ_1^2 distribution

the treatment and control groups. So we can directly compare the distributions of the outcome, 1978 earnings, for the treatment and control groups to get the treatment effect. The outcome distributions of the treatment and control groups are shown in Figure 2.6. Both of them are heavily right-skewed and have an excess of 0 values. Because the distributions are far from a normal distribution, a non-parametric test is more appropriate than parametric test assuming normality. The p-value of Mann-Whitney test is 0.01, and the test statistic is 0.43. Here the expectation of the test statistic is the probability that a random outcome in the treatment group is smaller than a random outcome in the control group. The result indicates that there is a positive treatment effect.

Next we construct strata based on two important pretreatment variables, age and 1974 earnings, separately, and then apply our proposed U test to identify whether there is treatment effect heterogeneity across the strata. We first split all subjects by quartiles of age. The subgroup sample sizes are in Table 2.3, and the pairwise U-statistics are in Table 2.4. Here the expectation of $U^{(p,q)}$ is the probability that the difference between treatment and control outcomes in stratum p are smaller than the difference in stratum q. As the U values are

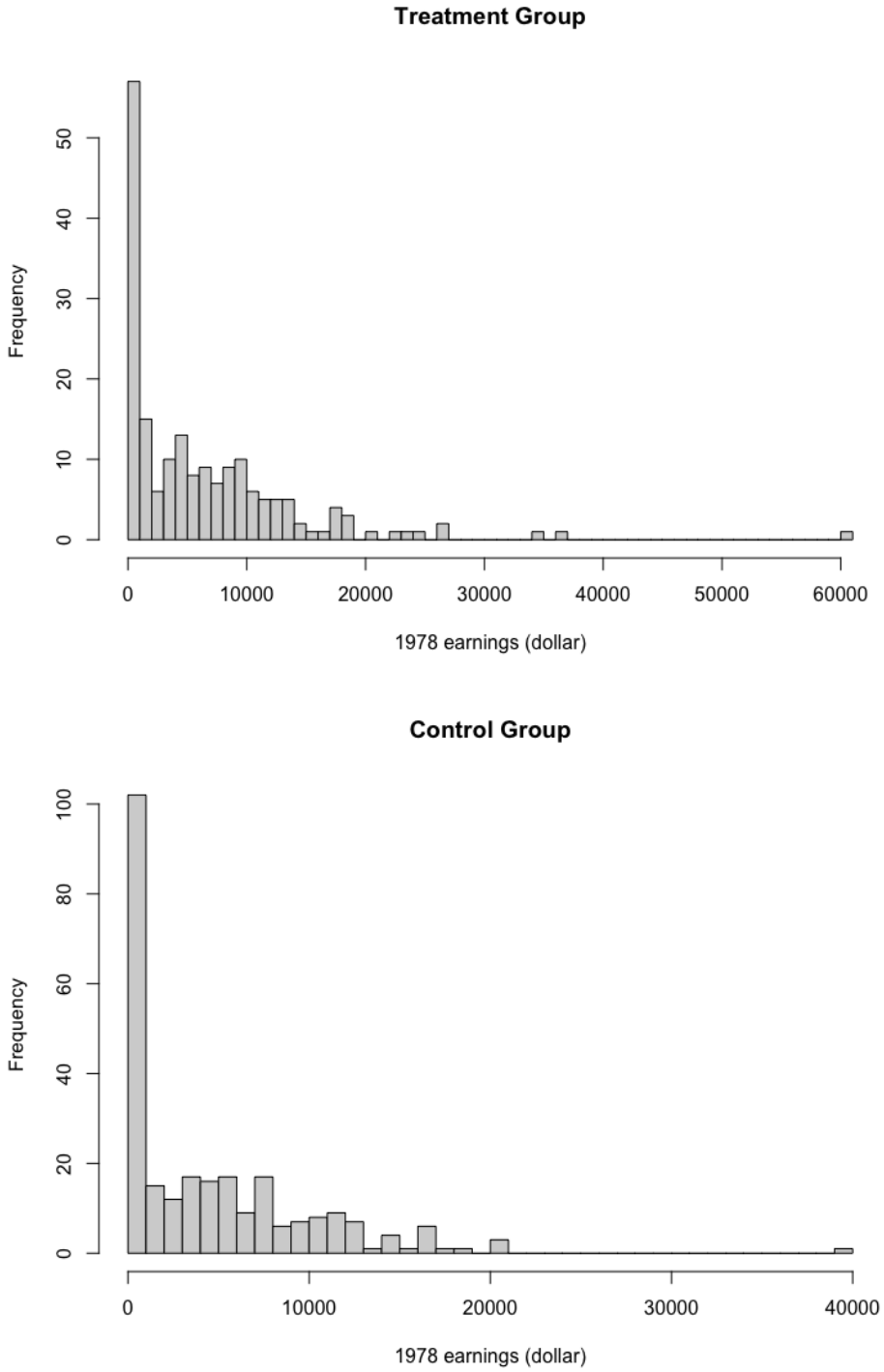


Figure 2.6: Distribution of 1978 earnings in the treatment and control groups

Stratum	1	2	3	4
Age	[17, 20]	(20,24]	(24,28]	(28, 55]
Treatment	47	41	49	48
Control	83	56	60	61

Table 2.3: Sample sizes in different treatment and age groups

$U^{(1,2)}$	$U^{(1,3)}$	$U^{(1,4)}$	$U^{(2,3)}$	$U^{(2,4)}$	$U^{(3,4)}$
0.52	0.55	0.57	0.53	0.55	0.51

Table 2.4: Pairwise U-statistics comparing treatment effects between age groups

greater then 0.5, the treatment effects in younger strata are generally smaller than those in older strata. However, the p-value of our proposed heterogeneity test is 0.58, so the observed heterogeneity is not statistically significant.

Then we explore whether the treatment effect differs between participants with and without positive incomes in 1974. The first stratum is for participants without income in 1974 and the second stratum is for those with positive income. The subgroup sample sizes are shown in Table 2.5. The U-statistic comparing their treatment effects $U^{(1,2)}$ is 0.409, and the p-value of our heterogeneity test is 0.032, which indicates this program has greater impact for participants who did not have any income in 1974 than those who had some income.

2.6 Discussion

Identifying the existence of treatment effect heterogeneity is a key element of attempts to provide more precise treatment recommendations for individuals. We have described a U-

Stratum	1	2
1974 Income	1974 Income = 0	1974 Income > 0
Treatment	131	54
Control	195	65

Table 2.5: Sample sizes of different treatment and income groups

statistic-based approach to formally test the hypothesis of homogeneous treatment effects without assuming a particular parametric form of the outcome distributions, and compared its performance with the LRT when both of these tests consistent with testing whether the differences of outcome averages between treatment and control groups are the same across different strata. The LRT requires the distribution of treatment effect estimates to be normal, which can be satisfied if the outcomes are normal or the sample sizes are large (by the Central Limit Theorem). Our results show that, as expected, when the outcome distributions are close to normal, the power of the LRT is a little better than the U test. However when at least one of the outcome distributions departs substantially from the normal distribution, the power of our non-parametric test can be significantly larger than the LRT. As the departure increases, the advantage of the U test increases. And obviously, the U test is more robust to outliers and the LRT. These observations are similar to the comparison between Mann-Whitney test and t-test (Lehmann, 2004).

A major problem of our non-parametric approach is that it is a consistent test assessing whether the probability $P(Y_p^t - Y_p^c < Y_q^t - Y_q^c) + \frac{1}{2}P(Y_p^t - Y_p^c = Y_q^t - Y_q^c) \quad (p \neq q)$ is equal to one half for all p, q, which however does not provide much insight in practice. Besides, the test is non-transitive, a significant result cannot help us identify the strata with largest or smallest treatment effects. With a result showing there is treatment effect heterogeneity, pairwise test statistics can be used to investigate pairwise comparisons of treatment effects. Other approaches, like summary statistics, linear regression can also be used to explore treatment effects within each stratum and their relationships across different strata. In the cases where the semi-parametric assumptions in Section 2.3 are considered to be satisfied, the U test is most easily interpreted. Under those assumptions, the test is consistent with respect to the null hypothesis that the difference of outcome means or percentiles between treatment and control groups are the same across all strata versus the alternative hypothesis that not all of them are identical.

One limitation of this method is that it requires the distributions of all confounding variables are the same between the treatment group and the control group within stratum, which is true in randomized experiments. However, in observational studies, we will need to adjust for confounding variables. Even if the strata are created based on estimated propensity scores in an effort to balance the baseline covariates (Xie et al., 2012), some further adjustments for remained imbalance may be needed.

Chapter 3

Nonparametric Tests for Treatment Effect Heterogeneity in Observational Studies

3.1 Introduction

Understanding treatment effect heterogeneity has attracted a great deal of attention in various research areas, including social sciences (Bitler et al., 2006; Feller and Holmes, 2009), health care (Kent et al., 2016; Ginsburg and Willard, 2009) and criminology (Na et al., 2015; Pate and Hamilton, 1992). It is well recognized that "one size does not fit all" in disease studies since subjects with different characteristics could respond quite differently to the same treatment. To better account for patient heterogeneity while evaluating the treatment effect and providing accurate personalized treatment recommendation, subgroup analysis (Cook et al., 2004) has been commonly used to identify subpopulations among subjects and examine the localized treatment effects within subpopulations. In some studies, subjects

may be divided into several strata based on baseline characteristics that are expected to be associated with treatment effects, and recommendations are made based on inference conducted within each stratum. However, subgroup analysis involves some problems including multiple testing and loss of statistical power (Cook et al., 2004). Thus it is not always recommended. In the cases where there is enough evidence showing the existence of treatment effect heterogeneity across those strata, subgroup analysis can be especially valuable. Thus before conducting stratum-specific analysis, a test examining existence of treatment effect heterogeneity is often needed.

There has been a large amount of literature on developing hypothesis testing approaches for examining treatment effect heterogeneity (e.g., Chang et al., 2015; Ding et al., 2016; Hsu, 2017) under different definitions of heterogeneity and different modeling assumptions. In this paper we focus on testing whether the average treatment effects across multiple pre-specified subpopulations are identical to each other. Parametric approaches towards this goal were proposed a long time ago. Regression methods have been considered (e.g., Allison, 1977; Byar, 1985), where the heterogeneity of treatment effects is tested by examining interaction terms between treatment assignment and potential effect modifiers. The likelihood ratio test (LRT) was also developed by Gail and Simon (1985) under normality assumptions for the stratum-specific treatment effect estimates. More recently, several nonparametric approaches have been proposed in the literature. Crump et al. (2008) proposed a test based on sieve estimation for treatment effects. This method was later generalized by Sant'Anna (2020) to test for heterogeneity in duration outcomes under endogenous treatment assignment. More recently, Dai and Stern (2020) proposed a U-statistic-based test (U test) which does not require estimating stratum-specific treatment effects. Compared to the LRT and other parametric tests, the nonparametric tests in general require weaker modeling assumptions on the outcome distributions. However, they still either require specifying a model for estimating the treatment effects (Crump et al., 2008; Sant'Anna, 2020), or only consider situations where baseline covariates are well balanced within each stratum (Dai and Stern,

2020). Motivated by these observations, we propose a nonparametric test that bypasses the need for estimating treatment effects while still being applicable to observational studies where there exist confounding variables that need to be addressed.

In this paper, we focus on testing the equality of the average treatment effects across multiple strata while adjusting for potential confounding variables in observational studies. We propose a new testing procedure based on an adjusted four-sample U-statistic that can be viewed as a weighted version of the original U-statistic developed by Dai and Stern (2020). Assuming the strata are mutually independent, the main idea is to first construct an adjusted U-statistic for comparing the treatment effects between two strata, and then formulate an overall test statistic as a function of those pairwise adjusted U-statistics. For each stratum, the weights in the adjusted U-statistic are carefully chosen by covariate matching and propensity score estimation (Li et al., 2018) such that the baseline covariate distributions for both the treatment and control groups are the same as the marginal distribution for the target population. To derive the asymptotic distribution for the proposed test, we find the main challenge is that our adjusted U-statistic no longer belongs to the generalized U-statistic family, therefore classical projection theory is not directly applicable. To solve this problem, we use the idea in Satten et al. (2018), which studies adjusted two-sample U-statistics, to obtain an asymptotic normality result. Based on the derived asymptotic theory, we then conduct several numerical studies to compare the performance of our proposed test with that of the LRT (Gail and Simon, 1985) and the unadjusted U test (Dai and Stern, 2020). Numerical results confirm the excellent operating characteristics for the proposed method even under propensity score model misspecification, and also clearly demonstrate the advantage of our method over the LRT and the unadjusted U test when the data is generated from a non-Gaussian distribution or the baseline covariates are not well balanced.

The remainder of the paper is structured as follows. In Section 3.2, we provide a review of the U test that assesses treatment effect heterogeneity across strata with balanced baseline

covariates. In Section 3.3, we introduce our adjusted U test for treatment effect heterogeneity that allows for the existence of confounding variables. In Section 3.4, we conduct simulation studies to demonstrate the asymptotic validity and efficiency of the adjusted U test, and also explore the impact of model misspecification. In Section 3.5, we further demonstrate the use of our method by two case studies, including an employment program evaluation study in labor economics, and another study on the evaluation of China's one-child policy on children's mental health. We conclude with some remarks in Section 3.6.

3.2 Review of Unadjusted U-Statistic-Based Test for Treatment Effect Heterogeneity

Dai and Stern (2020) (hereafter DS) proposed a U-statistic-based test (U test) to assess the consistency of average treatment effects in several independent strata, assuming there are no confounding variables. Compared to its parametric counterpart, the Likelihood Ratio Test (LRT) introduced by Gail and Simon (1985), their proposed U test can have a significant improvement in power especially when the outcomes are deviating far away from a normal distribution. Since the method we propose in this paper is based on their U test, we start with a review of their method.

Assume there are S strata. Within each stratum s ($s \in \{1, ..., S\}$), let τ_s be the additive treatment effect, $Y_s^t = \{Y_{si}^t, i = 1, ..., n_s^t\}$ be the outcomes of subjects in the treatment group, and $Y_s^c = \{Y_{si}^c, i = 1, ..., n_s^c\}$ be the outcomes of subjects in the control group. The total sample size across all strata is denoted as $N = \sum_{s=1}^{S}(n_s^t + n_s^c)$. Two assumptions are made in DS: (1) the outcomes $(Y_1^t, \cdots, Y_S^t, Y_1^c, \cdots, Y_S^c)$ are mutually independent; and (2) there exist positive constants $0 < \lambda_s^\omega < 1$ for every $s \in \{1, ..., S\}$ and $\omega \in \{t, c\}$ such that $\frac{n_s^\omega}{N} \to \lambda_s^\omega$ as $N \to \infty$.

To test for treatment effect heterogeneity across all strata, DS considers the null hypothesis H_0: $(Y_p^t - Y_p^c < Y_q^t - Y_q^c) + \frac{1}{2}P(Y_p^t - Y_p^c = Y_q^t - Y_q^c) = \frac{1}{2}$ for any $p \neq q$ $(p, q \in \{1, \cdots, S\})$ versus the alternative hypothesis H_a: there is at least one pair of (p, q) such that the equation does not hold. Their U test is consistent for this pair of hypotheses. Under the three sets of semi-parametric assumptions discussed in Section 2.3.3, e.g., $Y_s^t - Y_s^c$ $(s = 1, \cdots, S)$ follow a common distribution up to a location shift, the U test is also a consistent test for the null hypothesis $\tau_1 = ... = \tau_S$ versus the alternative hypothesis that at least two of them are unequal, where $\tau_s = \mathrm{E}(Y_s^t) - \mathrm{E}(Y_s^c)$. The test statistic is constructed by combining all pairwise U-statistics that compare treatment effects in two strata. To compare the treatment effects in the first two strata, a four-sample U-statistic is constructed as

$$U^{(1,2)} = \frac{1}{n_1^t n_1^c n_2^t n_2^c} \sum_{i=1}^{n_1^t} \sum_{j=1}^{n_1^c} \sum_{k=1}^{n_2^t} \sum_{l=1}^{n_2^c} \phi^{(1,2)}(i, j, k, l), \tag{3.1}$$

where the kernel function $\phi^{(1,2)}(i, j, k, l) = I(Y_{1i}^t - Y_{1j}^c < Y_{2k}^t - Y_{2l}^c) + \frac{1}{2}I(Y_{1i}^t - Y_{1j}^c = Y_{2k}^t - Y_{2l}^c)$. The latter term is used to account for possible ties for discrete distributions. Although DS focuses on additive treatment effect, other forms of treatment effects, such as the ratio of outcomes between different treatment groups, can also be incorporated. DS shows that

$$\sqrt{N}(U^{(1,2)} - \theta^{(1,2)}) \xrightarrow{D} \mathcal{N}(0, \sigma_{1,2}^2), \quad \text{when } N \to \infty, \tag{3.2}$$

where $\sigma_{1,2}^2 = \frac{1}{\lambda_1^t}\mathrm{Var}(h_1^{t,(1,2)}(Y_1^t)) + \frac{1}{\lambda_1^c}\mathrm{Var}(h_1^{c,(1,2)}(Y_1^c)) + \frac{1}{\lambda_2^t}\mathrm{Var}(h_2^{t,(1,2)}(Y_2^t)) + \frac{1}{\lambda_2^c}\mathrm{Var}(h_2^{c,(1,2)}(Y_2^c))$ is the asymptotic variance of $\sqrt{N}U^{(1,2)}$, and $h_s^{\omega,(1,2)}(x) = \mathrm{E}[\phi^{(1,2)}(1, 1, 1, 1)|Y_{s1}^{\omega} = x] - \theta^{(1,2)}$ for $s \in \{1, 2\}$ and $\omega \in \{t, c\}$. Under the null hypothesis that the difference of potential outcomes are identically distributed across strata, the expectation of $\phi^{(1,2)}(i, j, k, l)$ is $\frac{1}{2}$, thus $\theta^{(1,2)} \triangleq E(U^{(1,2)})$ is also $\frac{1}{2}$.

With S strata, all pairwise U-statistics $U^{(p,q)}$ $(1 \leq p < q \leq S)$ can be constructed in the exactly same way. Specifically, for every pair of (p, q), we can define $U^{(p,q)}$, $\theta^{(p,q)}$ and $h_s^{\omega,(p,q)}$

$(\omega \in \{t, w\}, s \in \{p, q\})$ similarly with $U^{(1,2)}$, $\theta^{(1,2)}$ and $h_s^{\omega,(1,2)}$ by replacing $(1, 2)$ with (p, q). Under the assumption that $\frac{n_s^\omega}{N} \to \lambda_s^\omega$ $(0 < \lambda_s^\omega < 1)$ as $N \to \infty$ for $s \in \{1, \cdots, S\}$ and $\omega \in \{t, c\}$, DS shows that

$$\sqrt{N}(U^{(1,2)}-\theta^{(1,2)}, U^{(1,3)}-\theta^{(1,3)}, \cdots, U^{(S-1,S)}-\theta^{(S-1,S)})^T \xrightarrow{D} \mathcal{N}(0, \Sigma), \quad \text{when } N \to \infty, \quad (3.3)$$

where $\Sigma = \frac{1}{\lambda_1^t}\Sigma_1^t + \frac{1}{\lambda_1^c}\Sigma_1^c + \cdots + \frac{1}{\lambda_S^t}\Sigma_S^t + \frac{1}{\lambda_S^c}\Sigma_S^c$ and Σ_s^ω is the covariance matrix of $\left(\tilde{h}_s^{\omega,(1,2)}(Y_s^\omega), \tilde{h}_s^{\omega,(1,3)}(Y_s^\omega), \cdots, \tilde{h}_s^{\omega,(S-1,S)}(Y_s^\omega)\right)$ for all $s \in \{1, \cdots, S\}$ and $\omega \in \{t, c\}$. Here $\tilde{h}_s^{\omega,(p,q)}(x) = h_s^{\omega,(p,q)}(x) I(s = p \text{ or } s = q)$.

To apply this method, Σ is estimated by a weighted average of Σ_s^ω $(s \in \{1, ..., S\}, \omega \in \{t, c\})$, and Σ_s^ω can be estimated by the corresponding sample covariance matrix. As $\tilde{h}$ terms are unknown, they need to be estimated as well. Though $h_s^{\omega,(p,q)}(x) = \mathrm{E}[\phi^{(p,q)}(i, j, k, l)|Y_s^\omega = x]-\theta^{(p,q)}$, the constant term $\theta^{(p,q)}$ can be ignored when calculating the covariance matrices. So they take the method-of-moment estimator for the expectation term $\mathrm{E}[\phi^{(p,q)}(i, j, k, l)|Y_s^\omega = x]$ as the estimator of $h_s^{\omega,(p,q)}(x)$. For instance, the estimator of $h_1^{t,(1,2)}(x)$ is $\hat{h}_1^{t,(1,2)}(x) = \frac{1}{n_1^c n_2^t n_2^c} \sum_{j=1}^{n_1^c} \sum_{k=1}^{n_2^t} \sum_{l=1}^{n_2^c} I(x-Y_{1j}^c < Y_{2k}^t-Y_{2l}^c)$. Similar calculation is repeated for all other h functions, and then used for computing the sample covariance $\hat{\Sigma}_s^\omega$ $(s \in \{1, \cdots, S\}, \omega \in \{t, c\})$, which leads to the final estimator of Σ as $\hat{\Sigma} = \frac{1}{\lambda_1^t}\hat{\Sigma}_1^t + \frac{1}{\lambda_1^c}\hat{\Sigma}_1^c + \cdots + \frac{1}{\lambda_S^t}\hat{\Sigma}_S^t + \frac{1}{\lambda_S^c}\hat{\Sigma}_S^c$.

To test the null hypothesis $H_0 : \theta = \frac{1}{2}\mathbf{1}_{S(S-1)/2}$, where $\theta = (\theta^{(1,2)}, \theta^{(1,3)}, ..., \theta^{(S-1,S)})^T$, DS focuses on a one-dimensional overall test statistic $U_h = N \cdot \sum_{1 \le p < q \le S} (U^{(p,q)} - \frac{1}{2})^2$. Though the asymptotic reference distribution of U_h does not have an analytic form, it can be approximated by simulation, that is, after generating a large number of independent samples $\{r_1, \cdots, r_L\}$ from $\mathcal{N}(0, \hat{\Sigma})$, the empirical distribution of $\{||r_1||^2, \cdots, ||r_L||^2\}$ approximates the asymptotic reference distribution of U_h.

3.3 Adjusted U Test of Treatment Effect Heterogeneity

The test described in Section 3.2 can only be used in situations where all baseline covariates are well balanced between different treatment groups in each stratum, e.g., stratified randomized experiments. In observational studies, directly applying that method may lead to misleading conclusions due to the existence of potential confounding variables. Even in the situations where the strata are constructed based on propensity scores, which is the probability of getting treatment (Rosenbaum and Rubin, 1983), in hope of balancing baseline covariates (Xie et al., 2012), there may remain imbalance that needs to be adjusted. So in this paper, we propose an approach that extends the U test reviewed in Section 3.2 to be applicable to situations with unbalanced baseline covariates.

3.3.1 Notation and setup

We introduce some additional notations here. For each stratum s, where $s \in \{1, ..., S\}$, we use $X_s^t = \{X_{si}^t, i = 1, ..., n_s^t\}$ to denote the collection of baseline covariates for subjects in the treatment group where the first element of each vector X_{si}^t is 1, corresponding to an intercept term. Similarly $X_s^c = \{X_{si}^c, i = 1, ..., n_s^c\}$ is used to denote the covariates for subjects in the control group. Let $X_s = X_s^t \cup X_s^c$ be the collection of covariates for all subjects in stratum s, where we assume the first n_s^t subjects are from the treatment group, and the rest are from the control. We use $T_s = \{T_{si}, i = 1, ..., n_s\}$ to denote the indicators of treatment, i.e., the first n_s^t elements are 1's and the rest are 0's. The within-stratum propensity score, $P(T_s = 1|X_s)$, is denoted by $e(X_s) = \{e(X_{si}), i = 1, ..., n_s\}$. Similarly, $e(X_s^t) = \{e(X_{si}^t), i = 1, ..., n_s^t\}$ denotes the first n_s^t elements in $e(X_s)$ and $e(X_s^c) = \{e(X_{si}^c), i = 1, ..., n_s^c\}$ denotes the rest. We assume $0 < e(X_s) < 1$ for all $s \in \{1, \cdots, S\}$.

3.3.2 Balancing baseline covariates within one stratum

To balance confounding variables, one way is to weight the subjects such that within each stratum all baseline covariates from the two treatment groups have the same distributions. As we assume the strata are mutually independent, here we only focus on how to balance the covariates in one stratum, and the same process can be applied to the others. For simplicity, here we omit the stratum indicator s in the subscript. In one stratum, for baseline covariate X, let its marginal density function (or probability mass function if X is discrete) be $f(x)$, and its conditional density functions (or probability mass functions) in the treatment and control groups be $f^t(x)$ and $f^c(x)$, respectively. Our goal is to find weight functions, $w^t(x)$ and $w^c(x)$, in the treatment and control group such that $f^t(x)w^t(x) = f^c(x)w^c(x)$. As discussed in Li et al. (2018), different choices of weight functions will lead to different target populations of interest. They propose to use a general function $h(x)$ to define the population of interest with $h(x)f(x)$ as its marginal distribution. For example, when $h(x) = 1$, the target population has a marginal distribution of $f(x)$, which corresponds to the distribution of X in the combined population of treatment and control groups. When $h(x)$ is $e(x)$ or $1 - e(x)$, the target population refers to the subjects in the treatment or control groups. And when $h(x) = e(x)(1 - e(x))$, the target population is the so-called overlap population (Li et al., 2018).

For a given $h(x)$, the weight functions $w^t(x)$ and $w^c(x)$ should satisfy

$$w^t(x)f^t(x) \propto w^c(x)f^c(x) \propto f(x)h(x). \tag{3.4}$$

Since $f^t(x) \propto f(x)e(x)$ and $f^c(x) \propto f(x)(1 - e(x))$, (3.4) implies

$$w^t(x) \propto \frac{h(x)}{e(x)}, \quad \text{and} \quad w^c(x) \propto \frac{h(x)}{1 - e(x)}. \tag{3.5}$$

When $h(x) = 1$, the induced weight functions yield the classical inverse probability weighting (Horvitz and Thompson, 1952).

The aforementioned weighting method can be incorporated in U-statistics as well. For example, Satten et al. (2018) adopted it to adjust two-sample U-statistics with the goal of testing for the existence of treatment effect in observational studies. For our study, we also use this method to adjust the pairwise U-statistics introduced in Section 3.2 in order to test for treatment effect heterogeneity in observational studies. We take $U^{(1,2)}$ in equation (3.1) as an example, which is the average of several kernel functions. Each kernel function $\phi^{(1,2)}(i, j, k, l)$ is constructed by the outcomes of four independent subjects, and each subject needs to be weighted. Since the outcomes are mutually independent, $\phi^{(1,2)}(i, j, k, l)$ should be weighted by the product of the weights for the four subjects, i.e., $w^t(X_{1i}) \cdot w^c(X_{1j}) \cdot w^t(X_{2k}) \cdot w^c(X_{2l})$.

The choice of the weight functions depends on $h(x)$, which in principle can be chosen as any positive function. However, we further require $h(x)$ to be a constant or a function of $e(x)$, and we require it to be differentiable with respect to $e(x)$. These requirements will later greatly help with the efficient estimation of the asymptotic reference distribution for the adjusted U-statistics without requiring approximation/sampling methods such as bootstrap. In general, the choice of $h(x)$ is flexible. For example, in our simulation study in Section 3.4 and the application study on only children's mental health in Section 3.5.2 , we focus on $h(x) = 1$. In the employment program evaluation study in Section 3.5.1, we choose $h(x) = e(x)$.

In practice, the propensity scores are unknown, and it is common to use a logistic regression model between treatment indicators and associated covariates X_s for their estimation. Formally, within stratum s ($s \in \{1, \cdots, S\}$), we consider the following model with parameter β_s,

$$\log\left(\frac{e(X_{si})}{1 - e(X_{si})}\right) = \beta_s^T X_{si}, \quad i = 1, ..., n_s. \tag{3.6}$$

Note the model specification here is flexible and can be extended to include quadratic (or nonlinear) functions of X_s and interaction terms as needed. The model does not impose any assumptions on the response variable, and in practice it is convenient to conduct model diagnostics for (3.6) based on Austin (2008). The estimate of β_s, denoted by $\hat{\beta}_s$, can be obtained by solving the estimating equation of logistic regression, denoted as $\sum_{j=1}^{n_s} S_{sj}(\hat{\beta}_s) = 0$.

As the propensity scores are functions of β_s, for simplicity, we denote the weights for subjects in the treatment and control groups by $w_{si}^t(\beta_s)$ $(i = 1, ..., n_s^t)$ and $w_{si}^c(\beta_s)$ $(i = 1, ..., n_s^c)$, respectively for $s \in \{1, \cdots, S\}$. In practice, these weights can be estimated by their plug-in estimates.

3.3.3 Testing treatment effect heterogeneity between two strata

We start with constructing a test statistic that compares the treatment effects between the first two strata. After weighting, the U-statistic in (3.1) becomes

$$\begin{aligned} U_a^{(1,2)} &= \frac{\sum_{i=1}^{n_1^t}\sum_{j=1}^{n_1^c}\sum_{k=1}^{n_2^t}\sum_{l=1}^{n_2^c} w_{1i}^t(\hat{\beta}_1)w_{1j}^c(\hat{\beta}_1)w_{2k}^t(\hat{\beta}_2)w_{2l}^c(\hat{\beta}_2)\phi^{(1,2)}(i,j,k,l)}{\sum_{i=1}^{n_1^t} w_{1i}^t(\hat{\beta}_1)\cdot\sum_{j=1}^{n_1^c} w_{1j}^c(\hat{\beta}_1)\cdot\sum_{k=1}^{n_2^t} w_{2k}^t(\hat{\beta}_2)\cdot\sum_{l=1}^{n_2^c} w_{2l}^c(\hat{\beta}_2)} \\ &= \frac{1}{n_1^t n_1^c n_2^t n_2^c}\sum_{i=1}^{n_1^t}\sum_{j=1}^{n_1^c}\sum_{k=1}^{n_2^t}\sum_{l=1}^{n_2^c}\frac{w_{1i}^t(\hat{\beta}_2)w_{1j}^c(\hat{\beta}_1)w_{2k}^t(\hat{\beta}_2)w_{2l}^c(\hat{\beta}_1)\phi^{(1,2)}(i,j,k,l)}{\bar{w}_1^t(\hat{\beta}_1)\bar{w}_1^c(\hat{\beta}_1)\bar{w}_2^t(\hat{\beta}_2)\bar{w}_2^c(\hat{\beta}_2)}, \end{aligned} \tag{3.7}$$

where $\bar{w}_s^\omega(\hat{\beta}_s) = \frac{1}{n_s^\omega}\sum_{i=1}^{n_s^\omega} w_{si}^\omega(\hat{\beta}_s)$ for $s \in \{1, \cdots, S\}$ and $\omega \in \{t, c\}$.

Though $U_a^{(1,2)}$ looks like a generalized U-statistic (Korolyuk and Borovskich, 2013), unfortunately it is not, because $\hat{\beta}_1$ and $\hat{\beta}_2$ are functions of all outcomes in the corresponding strata. Therefore the classical projection theorem cannot be directly applied to $U_a^{(1,2)}$. The key observation is that, if we replace $\hat{\beta}_1$ and $\hat{\beta}_2$ by their estimands, β_1 and β_2, then we obtain a generalized U-statistic. Moreover, if $\hat{\beta}_1$ and $\hat{\beta}_2$ are consistent estimates, we would expect the

asymptotic properties (e.g., normality) of the generalized U-statistics will still hold for our adjusted U-statistic. This is indeed the case by the following theorem. The proof is based on the idea in Satten et al. (2018) where they derived the asymptotic normality for adjusted two-sample U-statistics.

Theorem 3.1. *Suppose that $\hat{\beta}_1$ and $\hat{\beta}_2$ are consistent estimators for β_1 and β_2 and assume that (1) the outcomes $(Y_1^t, Y_2^t, Y_1^c, Y_2^c)$ are mutually independent; (2) there exist positive constants $0 < \lambda_s^\omega < 1$ for every $s \in \{1, 2\}$ and $\omega \in \{t, c\}$ such that $\frac{n_s^\omega}{n_1+n_2} \to \lambda_s^\omega$ as $n_1 + n_2 \to \infty$ and (3) $0 < e(X_s) < 1$ for all $s \in \{1, 2\}$ where $e(X_s)$ is defined in Section 3.3.1 . Then as $n_1 + n_2 \to \infty$, we have*

$$\sqrt{(n_1 + n_2)}(U_a^{(1,2)} - \theta_a^{(1,2)}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_{1,2}^2), \tag{3.8}$$

where $\theta_a^{(1,2)} = \lim_{n_1+n_2\to\infty} E[U_a^{(1,2)}]$ and
$\sigma_{1,2}^2 = \lim_{n_1+n_2\to\infty} \left\{ n_1^t \, Var[\eta_1^{t,(1,2)}(Y_1^t)] + n_1^c \, Var[\eta_1^{c,(1,2)}(Y_1^c)] + n_2^t \, Var[\eta_2^{t,(1,2)}(Y_2^t)] + n_2^c \, Var[\eta_2^{c,(1,2)}(Y_2^c)] \right\}$
is the asymptotic variance of $\sqrt{(n_1 + n_2)}U_a^{(1,2)}$, and the η functions are defined in the proof which can be found in Appendix A. Also we assume $Var[\eta_s^{\omega,(1,2)}(Y_s^\omega)] > 0$ for $s \in \{1, 2\}$ and $\omega \in \{t, c\}$.

Theorem 3.1 establishes the asymptotic distribution for our proposed adjusted U-statistic. Assumptions (1)–(3) are mild and commonly used in the literature. For example, Assumption (2) requires that within stratum, the proportion of treatment/group is not negligible, which is satisfied in most applications. Assumption (3) requires the propensity score to be bounded away from 0 and 1, which is called probabilistic assignment and is commonly used in the causal inference literature (Imbens and Rubin, 2015).

To estimate the asymptotic variance $\sigma_{1,2}^2$, we first estimate the η values, denoted by $\{\hat{\eta}_s^{\omega,(1,2)}(Y_{si}^\omega), i = 1, \cdots, n_s^\omega\}$ for $s \in \{1, 2\}$, by replacing all β's by their consistent estimates and replacing functions by their method-of-moment estimators in the same way as discussed in

Section 3.2. Then we use the sample variance of each set $\{\hat{\eta}_s^{\omega,(1,2)}(Y_{si}^{\omega}), i = 1, \cdots, n_s^{\omega}\}$ to estimate $\mathrm{Var}[\hat{\eta}_s^{\omega,(1,2)}(Y_s^{\omega})]$, i.e., $\widehat{\mathrm{Var}}[\hat{\eta}_s^{\omega,(1,2)}(Y_s^{\omega})] = \frac{1}{n_s^{\omega}-1}\sum_{i=1}^{n_s^{\omega}}(\hat{\eta}_s^{\omega,(1,2)}(Y_{si}^{\omega}) - \bar{\hat{\eta}}_s^{\omega,(1,2)}(Y_s^{\omega}))^2$, where $\bar{\hat{\eta}}_s^{\omega,(1,2)}(Y_s^{\omega})$ is the average of $\{\hat{\eta}_s^{\omega,(1,2)}(Y_{si}^{\omega}), i = 1, \cdots, n_s^{\omega}\}$. Then $\sigma_{1,2}^2$ can be consistently estimated by $\widehat{\sigma}_{1,2}^2 = n_1^t\widehat{\mathrm{Var}}[\hat{\eta}_1^{t,(1,2)}(Y_1^t)] + n_1^c\widehat{\mathrm{Var}}[\hat{\eta}_1^{c,(1,2)}(Y_1^c)] + n_2^t\widehat{\mathrm{Var}}[\hat{\eta}_2^{t,(1,2)}(Y_2^t)] + n_2^c\widehat{\mathrm{Var}}[\hat{\eta}_2^{c,(1,2)}(Y_2^c)]$.

3.3.4 Testing treatment effect heterogeneity in multiple strata

Next we consider testing for treatment effect heterogeneity in multiple strata, i.e., $1, 2, \cdots, S$, with $S > 2$, by extending the adjusted U-statistic in the previous section. For every pair of strata p and q satisfying $1 \le p < q \le S$, we can define an adjusted U-statistic $U_a^{(p,q)}$ in the same way as $U_a^{(1,2)}$. Then it is natural to consider a vector of all pairwise adjusted U-statistics $U_a = (U_a^{(1,2)}, U_a^{(1,3)}, ..., U_a^{(S-1,S)})^T$. In the next theorem, we derive its joint asymptotic distribution.

Theorem 3.2. *Suppose that Assumptions (1)–(3) in Theorem 3.1 are satisfied for every stratum. Then as the total sample size $N \to \infty$,*

$$\sqrt{N}(U_a - \theta_a) \xrightarrow{D} \mathcal{N}(0, \Sigma_a), \tag{3.9}$$

where $\theta_a = \lim_{N\to\infty} E(U_a)$ and $\Sigma_a = \frac{1}{\lambda_1^t}\Sigma_1^t + \frac{1}{\lambda_1^c}\Sigma_1^c + ... + \frac{1}{\lambda_S^t}\Sigma_S^t + \frac{1}{\lambda_S^c}\Sigma_S^c$ is the asymptotic covariance matrix of $\sqrt{N}U_a$, Σ_s^{ω} is the covariance matrix of $(\tilde{\eta}_s^{\omega,(1,2)}, ..., \tilde{\eta}_s^{\omega,(S-1,S)})$ for $s \in \{1, \cdots, S\}$ and $\omega \in \{t, c\}$, where $\tilde{\eta}_s^{\omega,(p,q)} = \eta_s^{\omega,(p,q)} I(s = p \text{ or } s = q)$, assuming $Var(\eta_s^{\omega,(p,q)}) > 0$.

The proof of this theorem can be found in Appendix A. The asymptotic covariance matrix Σ_a in Theorem 3.2 can be conveniently estimated in a similar way as for the univariate case in Theorem 3.1. That is, we first estimate the η terms and $\tilde{\eta}$ functions, and then use the

sample covariance matrix of estimated $\tilde{\eta}_s^{\omega,(p,q)}$ to estimate Σ_s^ω for $s \in \{1, ..., S\}$ and $\omega \in \{t, c\}$.

Given the estimated covariance $\hat{\Sigma}_a$, we can construct a global test statistic by considering a transformation on U_a. For instance, under $H_0 : Y_s^t - Y_s^c$ are identically distributed for $s \in \{1, \cdots, S\}$, we have $\theta_a = \frac{1}{2}\mathbf{1}$; therefore a one-dimensional test statistic can be constructed as $T_a = N(U_a - \frac{1}{2}\mathbf{1})^T(U_a - \frac{1}{2}\mathbf{1})$. Though the analytic form of its reference distribution is not available, we can still approximate it via simulations. This can be done by drawing a large number of samples $\{r_1, \cdots, r_L\}$ from $\mathcal{N}(0, \hat{\Sigma}_a)$, and then use $\{||r_1||^2, \cdots, ||r_L||^2\}$ as the empirical reference distribution. Other functions of U_a, e.g., $\sqrt{N} \max_{1 \le p < q \le S} |U^{(p,q)} - \frac{1}{2}|$, can also be used as the global test statistic, whose reference distribution can be approximated by simulations. In the numerical studies, we focus on using T_a, and propose to reject the null hypothesis when T_a is greater than or equal to the $100(1-\alpha)th$ percentile of $\{||r_1||^2, \cdots, ||r_L||^2\}$, where α is the significance level.

3.3.5 Trimming Sample

In the causal inference literature, it is common to exclude subjects with estimated propensity scores too close to 0 or 1 (Dehejia and Wahba, 1999; Crump et al., 2009; Imbens and Rubin, 2015). This trimming procedure has been shown to effectively improve the covariate balance between different treatment groups for several reasons. One is that those subjects whose true propensity scores that are equal to 0 or 1 should not be used since there are no counterparts in the alternative group. Another reason is that for those subjects whose estimated propensity scores are very close to 0 or 1, their counterparts will be associated with extremely large weights, which will then lead to a large variance for the estimated treatment effects.

There are two popular trimming rules. One is to set a hard threshold for propensity scores to be included in treatment effect estimates, e.g., $[\gamma, 1-\gamma]$ $(0 < \gamma < \frac{1}{2})$ (Crump et al., 2009), i.e., subjects with propensity scores outside this range should be removed. The

other is that we only use the subjects whose propensity scores are within the overlap region (Dehejia and Wahba, 1999). Specifically, we remove all subjects in the control group whose propensity scores are smaller than the minimum propensity score in the treatment group, and remove all subjects in the treatment group whose propensity scores are larger than the maximum propensity score in the control group. In practice, those two rules can be applied simultaneously.

It is worth mentioning that although the trimming procedure in general improves the treatment effect estimation accuracy, the reference population has changed. Hence there is a trade-off. Under this trade-off, people usually still prefer trimming because a reliable estimate for a subpopulation is generally considered more valuable than an estimate for the original population based on extrapolation or with large variance. In the numerical studies, we present both results with and without trimming to demonstrate the effect of trimming. More specifically, when implementing trimming, we first remove subjects outside of the propensity score overlap region, and then re-run the same propensity score model for the remaining subjects to obtain the weights for our adjusted U tests. We have conducted several numerical experiments and found that the type I error is better controlled with the new propensity scores. Therefore we choose to implement this trimming procedure for all numerical studies in this paper.

3.4 Simulation

We conduct simulation studies to evaluate the empirical performance of the proposed adjusted U-statistic test and compare it with the likelihood ratio test (LRT) and the U test developed in Dai and Stern (2020). Here, we focus on the case where the target population is the combination of the treatment and control groups, i.e., $h(x) = 1$. We consider the adjusted U tests with and without the trimming procedure, and denote them as AUT-T and

AUT, respectively.

3.4.1 Implementation Details

We first discuss the computational implementation of both our proposed U tests and the LRT. The U test statistic in (3.9) is a function of $S(S-1)/2$ pairwise adjusted U-statistics, and the computation of each adjusted U-statistic can be expensive in simulation studies. Therefore instead of calculating the complete adjusted U-statistics, we randomly sample some of the ϕ functions in each of the adjusted U-statistics. Take $U_a^{(1,2)}$ in (3.7) as an example, for each stratum $s \in \{1,2\}$ and treatment group $\omega \in \{t,c\}$, we randomly choose $M = 1000N$ (N is the total sample size over all strata) subjects with replacement, denoted by $\{(y_{1i}^t, y_{1i}^c, y_{2i}^t, y_{2i}^c), i = 1, \cdots, M\}$. Then we calculate the kernel function ϕ_i based on $(y_{1i}^t, y_{1i}^c, y_{2i}^t, y_{2i}^c)$ and use the weighted average of $\{\phi_i, i = 1, \cdots, M\}$ to approximate $U_a^{(1,2)}$. As we also use the weighted kernel functions to estimate $\tilde{h}_s^\omega(Y_{si}^\omega)$ for $i \in \{1, \cdots, n_s^\omega\}$, $s \in \{1,2\}$ and $\omega \in \{t,c\}$, which are required to obtain $\hat{\Sigma}_a$, we need to make sure that each subject is sampled at least once. This is usually satisfied given a large sampling size M, and we redo the sampling process on the rare occasion that this requirement is not met. The sampling size $M = 1000N$ was selected by running a series of different simulation scenarios with 3 strata and N ranging from 60 to 3000; this choice of M ensured the variance of the approximated test statistic $\frac{T_a}{N} = \sum_{1 \le p < q \le S} (U^{(p,q)} - \frac{1}{2})^2$ to be smaller than 0.003. In order to approximate the reference distribution of $\frac{T_a}{N}$, 10^5 samples $\{r_i, i = 1, \cdots, 10^5\}$ are generated independently from the estimated reference distribution $\mathcal{N}(0, \frac{1}{N}\hat{\Sigma}_a)$. Then $\{||r_i||^2, i = 1, \cdots, 10^5\}$ are used to obtain the empirical reference distribution $\frac{T_a}{N}$. The sample size of 10^5 is chosen to ensure that the variance of the 95th percentile of $\{||r_i||^2, i = 1, \cdots, 10^5\}$ is below 0.0001.

Next we give a brief review of the competitive approach for testing the treatment effect homogeneity, i.e., the LRT proposed by Gail and Simon (1985). With S strata, they test the

null hypothesis that the average treatment effects τ_s ($s \in \{1, \cdots, S\}$) are the same across all of the strata versus the alternative that at least two of them are unequal. Assuming the treatment effect estimates $\hat{\tau}_s$ ($s \in \{1, \cdots, S\}$) follow normal distributions as $\hat{\tau}_s \overset{indep}{\sim} \mathcal{N}(\tau_s, \sigma_s^2)$, then a test statistic is constructed as

$$H = \sum_{s=1}^{S} (\hat{\tau}_s - \bar{\hat{\tau}})^2 / s_s^2 \overset{H_0}{\sim} \chi^2_{S-1}, \tag{3.10}$$

where $\bar{\hat{\tau}} = \frac{\sum_{s=1}^{S} \hat{\tau}_s / s_s^2}{\sum_{s=1}^{S} 1/s_s^2}$, and s_s^2 is a consistent estimator of σ_s^2 for $s \in \{1, \cdots, S\}$. For an α level test, we reject the null hypothesis when H is greater than or equal to the $100(1-\alpha)^{\text{th}}$ percentile of χ^2_{S-1}.

In randomized experiments where we can directly compare the outcomes of different treatment groups to estimate the treatment effect, $\hat{\tau}_s$ can be the difference of the outcome averages. In observational studies, a method for estimating $\hat{\tau}_s$ that adjusts for confounding variables should be used. Any methods that can provide normally distributed $\hat{\tau}_s$ and consistent estimator for σ_s^2 in stratum s for $s \in \{1, \cdots, S\}$ can be used. For instance, when the outcome follows a continuous distribution, a linear regression model between the outcome and the treatment indicator and other confounding variables can be fitted within each stratum. Under the assumption that the outcomes are mutually independent, the normality assumption for $\hat{\tau}_s$ will be satisfied when the stratum sample size n_s goes to infinity. In this simulation, we fit a linear regression in each stratum s ($s \in \{1, \cdots, S\}$) to obtain $\hat{\tau}_s$ and $\hat{\sigma}_s^2$. We focus on the case that $Y_s^t - Y_s^c$ ($s \in \{1, \cdots, S\}$) follow the same distribution up to a location shift. Thus the hypotheses of the adjusted U tests are equivalent to those of the LRT; hence those two tests are directly comparable.

3.4.2 Simulation Design

We consider three strata ($S = 3$), where each stratum has the same sample size, i.e., $n_1 = n_2 = n_3 = n$. For each stratum s, we generate the data from an outcome model $Y_s = 1 + \beta_{s,t}T_s + Z_s + \epsilon_s$ for $s \in \{1, 2, 3\}$, where the treatment indicator $T_s \sim \text{Bern}(p_s)$, and the residual terms ϵ_s follow a common distribution F_ϵ across all strata. The probability of being assigned to the treatment group p_s is also a function of the confounding variable Z_s, for which we assume $\text{logit}(p_s) = \gamma_s Z_s$. In the following simulations, we fix $Z_1 \sim \mathcal{N}(0, 1)$, $Z_2 \sim \mathcal{N}(0, 1)$, $Z_3 \sim \text{Unif}(-0.5, 0.5)$, and choose $\gamma_1 = 1$, $\gamma_2 = -1$, $\gamma_3 = 1$, such that the confounding variables either follow different distributions or satisfy different relationships with the treatment assignment among the three strata. Also, the treatment effects are set as $\beta_{1,t} = 1$, $\beta_{2,t} = 1 + \Delta$, $\beta_{3,t} = 1 + 2\Delta$, where the constant Δ is treated as the effect size. Note that when $\Delta = 0$, there still exists a treatment effect within each stratum although there is no treatment effect heterogeneity, i.e., the null hypothesis is true. For all of the simulation scenarios, we fix the significance level at 0.05, and repeat the data generating mechanism for $L = 2000$ times to obtain the empirical rejection rates.

In addition to the simulation design described above, we also consider several other designs with unequal sample size and different error distributions across the three strata. The simulation designs and results are very similar to those in Dai and Stern (2020), so we choose not to present them in this chapter.

3.4.3 Simulation results

We first check the type I error of our proposed adjusted U-test with and without trimming (AUT-T and AUT) when $\Delta = 0$, $n = 200$, $F_\epsilon = \mathcal{N}(0, 1)$, and compare them to the unadjusted U test reviewed in Section 3.2. Based on 2000 Monte-Carlo replications, the type I error rates for the AUT-T and AUT are 0.051 and 0.058, both are very close to the nominal

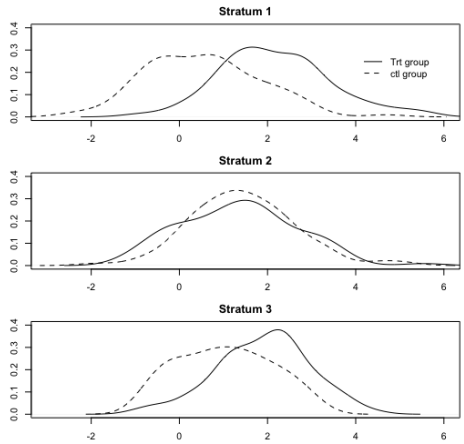


Figure 3.1: Density plots for the unadjusted outcomes in the treatment and control groups.

Stratum 1		Stratum 2		Stratum 3	
Treatment	Control	Treatment	Control	Treatment	Control
7.11	7.09	7.30	7.14	1.64	1.62

Table 3.1: Validity: average number of removed subjects in each subgroup for trimmed U test.

level of 0.05, whereas the unadjusted U test has a rejection rate of 1.000. The invalidity of the unadjusted U test is not surprising, because the unweighted outcome distributions are quite different between treatment and control groups in each stratum, as shown in Figure 3.1. This finding clearly demonstrates the need for confounder adjustment when testing for treatment effect heterogeneity. In Figure 3.2, we plot the empirical p-values with the expected uniformly distributed p-values for both the AUT-T and AUT methods. We find that the empirical distribution for the p-values is very close to the uniform distribution under the null hypothesis, which confirms both the validity of the asymptotic null distribution derived in Theorem 3.2 and the accuracy of random sampling when calculating the test statistics. Compared to AUT, the results for AUT-T is less perfect due to the fact that the population has changed after trimming the propensity score. To demonstrate the effect of trimming, we present the average number of removed subjects for each strata in Table 3.1, and find that the effect of trimming is minor since less than 8% of the subjects are removed from each stratum.

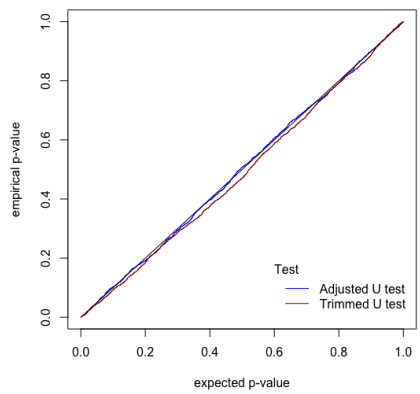


Figure 3.2: Empirical and expected p-values for proposed U tests under the null hypothesis.

Next we investigate the power for the proposed adjusted U tests under different values for the sample size n, effect size Δ and error distributions F_ϵ. We also use the results from the regression-based LRT as a benchmark for power comparison.

We choose four distributions for F_ϵ: $\mathcal{N}(0,1)$, Unif$(-2,2)$, t_4 and $0.5\mathcal{N}(-5,1)+0.5\mathcal{N}(5,1)$. For each of them, we consider four effect sizes (including 0), and then present the empirical rejection rates for the adjusted U tests and the LRT in Figure 3.4. We first note that under all four scenarios, the type I error rates are very close to the nominal level 0.05. There is a minor discrepancy for the trimmed U test, especially when the sample size is small. This is expected because trimming changes the reference population, although the number of trimmed subjects (see Figure 3.3) is quite small (between 2% and 15%). Therefore it is fair to compare the power of those three tests given that their type I errors are at the same level.

When $\Delta > 0$, we first notice that the power increases quickly to one as either the sample size n or the effect size Δ increases. By comparing the power between the two adjusted U tests (AUT-T and AUT), we find that overall ATU-T has a larger power although the advantage is not significant. This is expected because we only remove a minor percentage of subjects by trimming. We then compare the power of the AUT and LRT, and find that

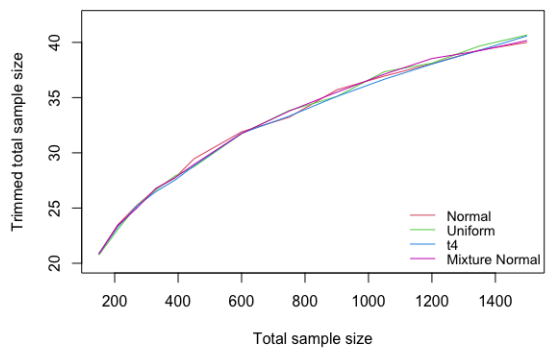


Figure 3.3: Power analysis: average number of trimmed subjects for four error distributions based on 2000 Monte-Carlo replications.

LRT is more powerful than AUT if the error distribution F_ϵ is normal or having lighter tails than normal distribution (e.g., uniform distribution). On the other hand, our proposed AUT is more powerful than LRT when F_ϵ has heavy tails (e.g., t_4) or deviates far away from a normal distribution (e.g., a bimodal distribution as $0.5\mathcal{N}(-5,1)+0.5\mathcal{N}(5,1)$). Those findings confirm that the LRT is still the most powerful test under the normality assumption. However, our proposed method will gain efficiency in testing against the null hypothesis as the true error distribution starts to move away from a normal distribution, with a more significant improvement in power over LRT when the error distribution is bimodal.

3.4.4 Sensitivity Analysis

Because our proposed adjusted U test is based on a propensity score model, in this section, we conduct a sensitivity analysis to evaluate the performance of our method under misspecification of the propensity score model. It is worth mentioning that despite recent advances in propensity score model diagnosis (Imbens and Rubin, 2015; Vegetabile et al., 2020) that rely on measuring the degree of covariance balance from the weighted samples in the treatment and control groups, measuring covariate balance still remains challenging especially when the number of covariates is large. Therefore it remains important to explore the sensitivity of the

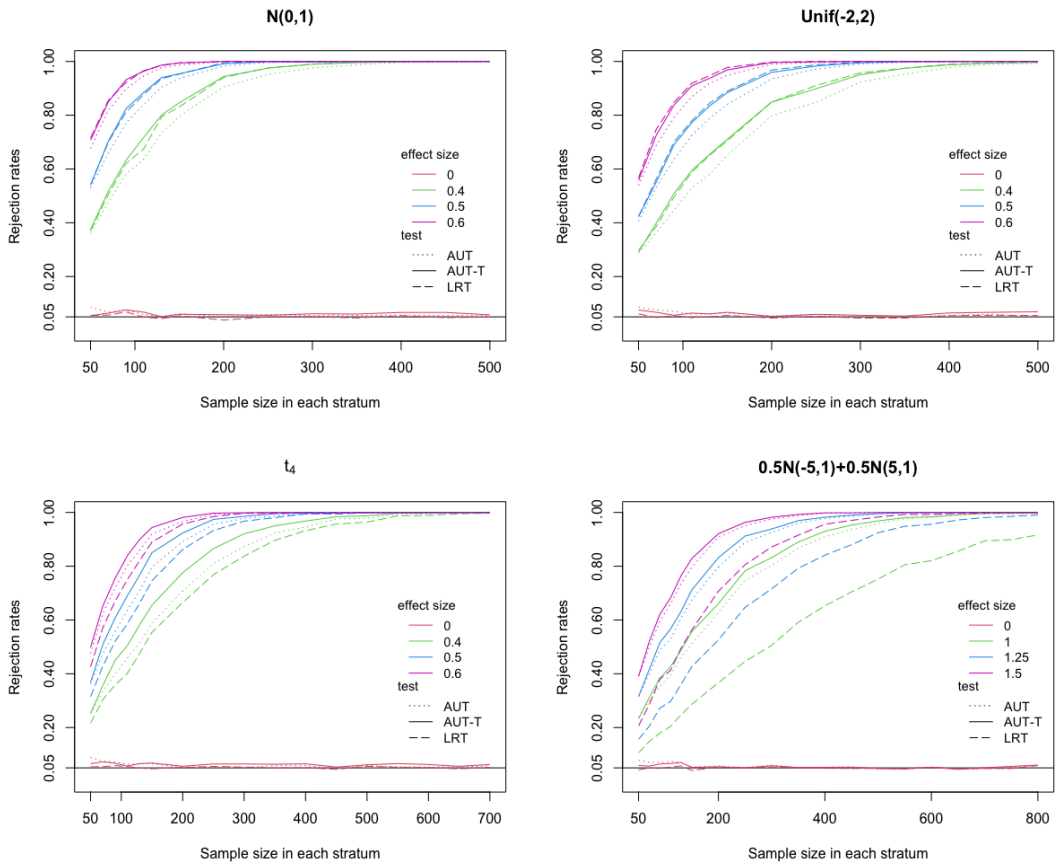


Figure 3.4: Power analysis: empirical rejection rates for three tests under various error distributions, sample sizes, and effect sizes, based on 2000 Monte-Carlo replications.

proposed test to misspecification. We consider several different null cases where there is no treatment effect heterogeneity, and explore the sensitivity of the adjusted U tests with and without trimming by checking the distributions of empirical p-values when the propensity score models are misspecified.

For data generation, we consider three strata ($S = 3$), each with a sample size of 200, and a confounding variable Z_s ($s = 1, 2, 3$) in each stratum satisfying $Z_1 \sim \mathcal{N}(0, 0.5^2)$, $Z_2 \sim \mathcal{N}(0, 0.5^2)$, $Z_3 \sim \text{Unif}(-0.5, 0.5)$. We add a quadratic term of Z_s to both the outcome model and propensity score model $Y_s = T_s + Z_s + \beta_{s,2} Z_s^2 + \epsilon_s$ and $\text{logit}(p_s) = \gamma_{s,0} + Z_s + \gamma_{s,2} Z_s^2$ with $T_s \sim \text{Bern}(p_s)$ and $\epsilon_s \sim F_\epsilon$ for $s \in \{1, 2, 3\}$. Note that there is no treatment effect heterogeneity in this scenario, i.e., the null hypothesis is true. Furthermore, we set $\beta_{1,2} = \gamma_{1,2} = 2$, $\beta_{2,2} = \gamma_{2,2} = -2$, $\beta_{3,2} = \gamma_{3,2} = 2$, $\gamma_{1,0} = -0.5$, $\gamma_{2,0} = 0.5$, $\gamma_{3,0} = -1/6$ to make the coefficients for Z and Z^2 the same in both outcome and propensity score models in every stratum. Values of $\gamma_{s,0}$ ($s = 1, 2, 3$) are chosen to avoid propensity scores being too close to 0 or 1. Here we explore the extent to which the empirical distributions of p-values for the adjusted U tests deviate from the expected uniform distribution when the propensity model is fitted without the quadratic term. As with earlier simulations, we consider four choices for the error distribution F_ϵ as $\mathcal{N}(0, 1)$, $\text{Unif}(-2, 2)$, t_4 and $0.5\mathcal{N}(-2, 1) + 0.5\mathcal{N}(2, 1)$.

Figure 3.5 shows the relationship between the empirical p-values versus the expected uniform p-values for the adjusted U tests with and without trimming under each of the four error distributions. AUT-T is always more robust to model misspecification than AUT. This finding suggests subject trimming based on propensity scores may improve model robustness. We also show the average number of trimmed subjects for each stratum in Table 3.2, and find that proportion to be reasonably small ($< 5\%$).

A possible explanation for the advantage of trimming in this simulation scenario is that after removing subjects with extreme propensity scores, a linear function can better approximate the relationship between the log odds of the propensity scores and the confounders for the

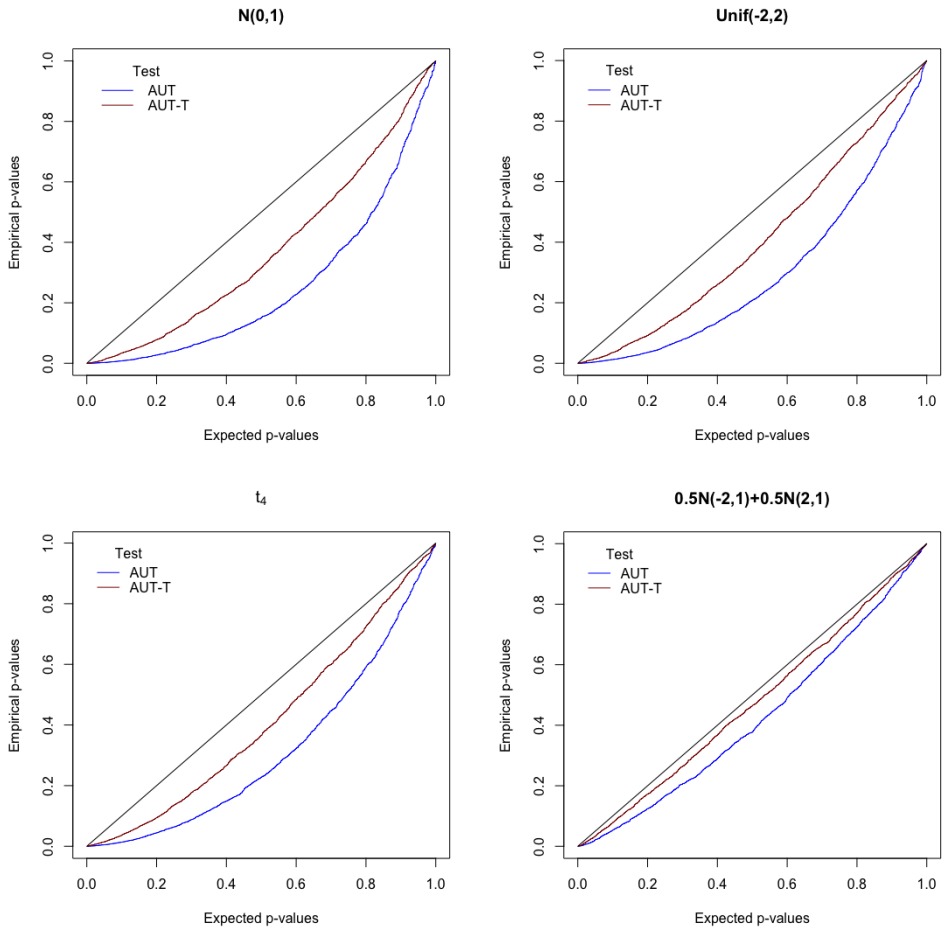


Figure 3.5: Empirical p-values of misspecified adjusted U test, trimmed U test and LRT vs expected p-values.

	Stratum 1		Stratum 2		Stratum 3	
	Treatment	Control	Treatment	Control	Treatment	Control
$\mathcal{N}(0,1)$	9.00	0.21	0.23	9.12	2.21	1.18
$U(-2,2)$	8.95	0.23	0.20	8.95	2.16	1.17
t_4	8.85	0.22	0.23	9.33	2.18	1.20
$0.5\mathcal{N}(-5,1)+0.5\mathcal{N}(5,1)$	9.03	0.19	0.21	8.99	2.20	1.19

Table 3.2: Sensitivity analysis: average number of trimmed subjects for each stratum by trimmed U test based on 2000 Monte-Carlo replications.

Stratum 1		Stratum 2		Stratum 3	
Untrimmed	Trimmed	Untrimmed	Trimmed	Untrimmed	Trimmed
0.88	0.79	0.88	0.79	0.97	0.88

Table 3.3: Sensitivity analysis: average R^2 of linear regression $\text{logit}(p_s) \sim Z_s$ for untrimmed and trimmed samples within each stratum based on 2000 Monte-Carlo replications.

remaining subjects. To examine this conjecture, we consider a different scenario where the R^2 of the linear regression $\text{logit}(p_s) \sim Z_s$ drops after trimming subjects. For data generation, again we consider three strata and the sample size for each strata is 200. The confounder Z_s within each stratum satisfies $Z_1 \sim \mathcal{N}(0,1)$, $Z_2 \sim \mathcal{N}(0,1)$ and $Z_3 \sim \text{Unif}(-3,3)$. The outcome and propensity score models are $Y_s = T_s + \alpha_s W + \epsilon_s$ and $\text{logit}(p_s) = \alpha_s W$, where $W = (-1.875 + Z_s)I(Z_s \leq -1.5) + (1.875 + Z_s)I(Z_s \geq 1.5) + Z_s^3 I(-1.5 < Z_s < 1.5)$, $T_s \sim \text{Bern}(p_s)$ and $\epsilon_s \sim F_\epsilon$ for $s \in \{1,2,3\}$. We set $\alpha_1 = \alpha_3 = 1$ and $\alpha_2 = -1$ to make the confounders either follow different distributions or have different relationships with outcomes and treatment assignments across the three strata. For the misspecified propensity score model, we fit logistic regressions regressing T_s only on Z_s for $s \in \{1,2,3\}$. We consider four choices for F_ϵ, $\mathcal{N}(0,1)$, $\text{Unif}(-2,2)$, t_4 and $0.5\mathcal{N}(-2,1) + 0.5\mathcal{N}(2,1)$.

Table 3.3 shows the average R^2 of linear regression $\text{logit}(p_s) \sim Z_s$ for the original and trimmed samples within each stratum over 2000 Monte-Carlo replications. The values are the same for the four scenarios. It clearly shows that trimming decreases R^2 in all strata. Figure 3.6 shows the relationships between the empirical p-values of the AUT and AUT-T with misspecifed propensity score models and the expected uniform p-values when the tests are valid. It indicates AUT is pretty sensitive, but trimming again minimizes the effect in all scenarios.

For the two simulation studies the test with trimming was less sensitive to misspecification. There is no guarantee that this will always be true, however our attempt to find a counter-example was not successful.

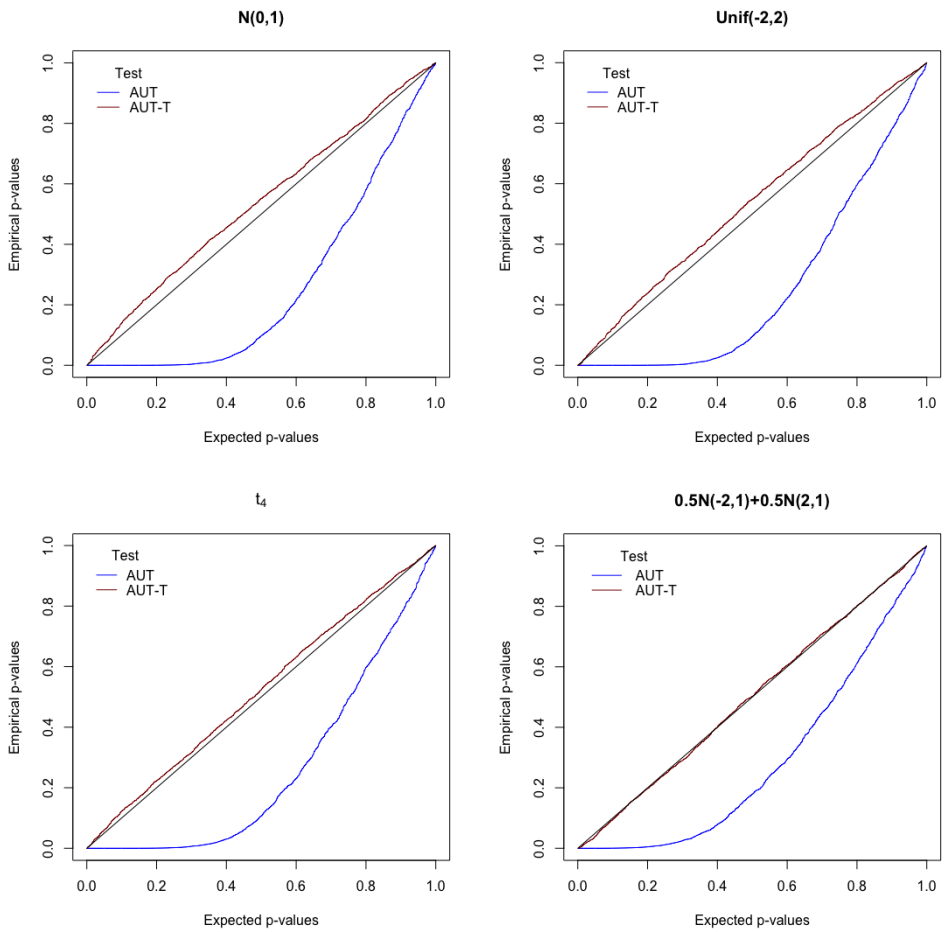


Figure 3.6: Empirical p-values of misspecified adjusted U test, trimmed U test and LRT vs expected p-values.

3.5 Case Study

3.5.1 Comparing effects of an employment program on people with different ages

We apply the proposed method to an employment program evaluation study in labor economics, which evaluates the effect of the National Support Work (NSW) Demonstration on trainee earnings. The NSW was conducted in the mid-1970s with the goal of helping disadvantaged workers gain working experience. More details about this program can be found in LaLonde (1986) and Dehejia and Wahba (1999). In this program, applicants were randomly assigned to the treatment and control groups; and the treatment effect can be easily assessed by directly comparing the outcomes between those two groups. In order to evaluate whether observational studies can replicate results from randomized experiments, LaLonde (1986) compared the treated subjects in the experiment to two nonexperimental comparison groups, namely, the Panel Study of Income Dynamics (PSID-1) and Current Population Survey-Social Security Administration File (CPS-1), as well as several subsets of them. The collected pretreatment covariates include age, education, marital status, indicator of "no degree", race indicators, earnings in 1974 (RE74) and 1975 (RE75). The outcome of interest is earnings in 1978.

We focus on the data set constructed by Dehejia and Wahba (1999), which is a subset of the original data set in LaLonde (1986) that includes data collected from male participants who have earnings information in 1974. The data is available at `https://users.nber.org/~rdehejia/data/.nswdata2.html`. It has been shown by Dehejia and Wahba (1999) that there is a positive treatment effect. Our goal here is to investigate whether there is treatment effect heterogeneity across different age groups for the treated subjects. Two strata are created based on the median age (25 years old) of the treatment group, that is,

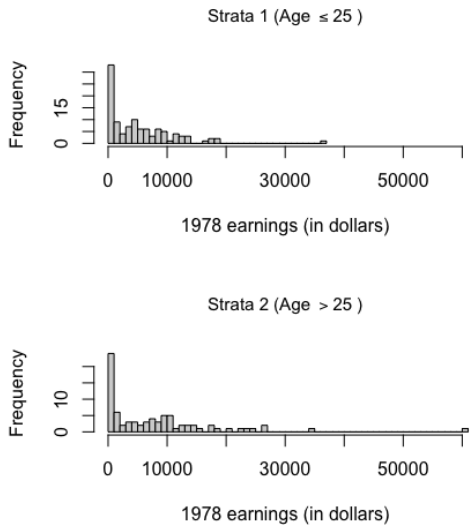


Figure 3.7: Distribution of earnings in 1978 for participants in the treatment group.

stratum 1 for subjects with age ≤ 25 and stratum 2 for age > 25. Figure 3.7 shows the outcome distributions of the treated subjects in the two strata, and it is clear that both distributions are highly right-skewed, which suggests that nonparametric U tests should be preferred to the LRT.

We compare the NSW treatment group to the NSW control group and CPS-1 separately. The first three columns of Table 3.4 shows the summary statistics of baseline covariates in both strata for the three groups. To compare the NSW treatment group with its control, we notice that the baseline covariates between groups are similarly distributed, so the unadjusted U test can be applied to assess the treatment effect heterogeneity between the two strata. We obtain an estimated unadjusted U-statistic of 0.554 with a p-value of 0.181, which suggests that the treatment effect in the younger group (stratum 1) is smaller than that in the elder group (stratum 2), although this difference is not statistically significant (note that a U-statistic value of 0.5 means no heterogeneity between those two strata, and a value larger than 0.5 means stratum 1 has a smaller treatment effect than that of stratum 2).

We then study the comparison between the NSW treatment group and CPS-1 group. Ta-

	NSW Treated	NSW Control	CPS-1	Weighted and Trimmed CPS-1
Stratum 1				
Sample size	106	161	4676	2169
Age	21.09 (2.76)	20.75 (2.75)	20.82 (2.82)	20.97 (2.51)
Education	10.29 (1.77)	9.93 (1.43)	11.91 (2.14)	10.2 (1.54)
Black	0.82 (0.39)	0.8 (0.4)	0.08 (0.28)	0.85 (0.36)
Hispanic	0.08 (0.26)	0.13 (0.33)	0.07 (0.26)	0.06 (0.24)
Married	0.11 (0.32)	0.09 (0.28)	0.36 (0.48)	0.1 (0.3)
Nodegree	0.72 (0.45)	0.89 (0.3)	0.34 (0.47)	0.78 (0.41)
RE74	2129.02 (4809.7)	2195.81 (6240.8)	7044.39 (7156.6)	1845.71 (4032.9)
RE75	1215.97 (2140.9)	1125.32 (3037.3)	7665.79 (7251.4)	1068.04 (2379.4)
Stratum 2				
Sample size	79	99	11316	1668
Age	32.15 (6.24)	32.05 (6.24)	38.35 (8.9)	32.25 (5.97)
Education	10.42 (2.28)	10.35 (1.84)	12.07 (3.12)	10.47 (2.1)
Black	0.87 (0.33)	0.87 (0.33)	0.07 (0.26)	0.89 (0.32)
Hispanic	0.04 (0.2)	0.07 (0.26)	0.07 (0.26)	0.03 (0.17)
Married	0.29 (0.46)	0.26 (0.44)	0.86 (0.35)	0.24 (0.42)
Nodegree	0.7 (0.46)	0.74 (0.44)	0.28 (0.45)	0.67 (0.47)
RE74	2050.7 (4957.2)	1962.64 (4611.6)	16897.94 (8936.6)	1993.3 (4772.0)
RE75	1956.17 (4204.0)	1497.18 (3178.3)	16123.93 (8876.9)	1909.62 (4093.3)

Table 3.4: Sample means (standard deviations) of baseline characteristics for NSW and CPS-1 data in two age strata.

ble 3.4 suggests that the baseline covariate distributions in those two groups seem to differ quite a lot. Therefore we apply the proposed adjusted U test with trimming. In both strata, we use logistic regressions to estimate propensity scores. For stratum 1 we use the following covariates: age, age^2, age^3, education, education^2, I(married), I(no degree), I(black), I(Hispanic), RE74, RE75, I(RE74 $= 0$), I(RE75 $= 0$), RE74 $*$ I(married) and RE74 $*$ I(no degree). In stratum 2, we consider age, age^2, age^3, education, education^2, I(married) $+I$(no degree), I(black), I(Hispanic), RE74, RE75, I(RE74 $= 0$), I(RE75 $= 0$) and education $*$ RE74. Most of those covariates are also included in the study of Dehejia and Wahba (1999). Subjects are weighted according to (3.5) with $h(x) = e(x)$. We present summary statistics for the baseline covariates after trimming and weighting as in the fourth column of Table 3.4. The weighted distributions of baseline covariates in CPS-1 are very similar to the NSW treatment group. Due to the large sample size, we randomly sample $M = 1000N$ ($N = 4022$ is the total sample size) weighted kernel functions to approximate the adjusted U-statistics as illustrated in Section 3.4.1. The estimated adjusted U-statistic comparing the treatment effects in the two strata is 0.541 with a p-value of 0.508, which

leads to the same conclusion as the randomized data comparison (NSW treatment versus its control). Meanwhile, if an unadjusted U test is applied to conduct the same comparison, then the estimated U-statistic would be 0.426 with a p-value of 0.004, which will lead to an opposite conclusion. This finding confirms the benefit of our proposed methodology and also highlights the necessity of appropriately adjusting for covariate balance between groups when testing for a treatment heterogeneity effect.

3.5.2 Assessing heterogeneity of the effect of being an only child on mental health

From 1979 to 2015, China's one-child policy was implemented to slow down the rapid growth of the nation's population. Though the policy has led to economic benefits for China, it has been criticized for introducing a series of social problems, e.g., forced abortions, female infanticide, and a heavy burden of elderly support (Hesketh and Zhu, 1997). Apart from these problems, the psychological wellbeing of the massive number of only children resulting from the policy has been a great concern because it has been widely recognized that siblings have a large impact on children's social behavior and mental health (e.g., Dunn, 1988; McHale et al., 2012). Only children in China are generally perceived to be more self-centered and less trustworthy. However the difference between only and non-only children may vary with geographic area and gender for two reasons. First, parents living in urban and rural areas differ in many aspects including education level, family income and lifestyle. Second, a preference for male children was prevalent at that time, especially in rural areas. For these reasons, the literature assessing the effects of being an only child are typically carried out in different strata that are determined by the type of region (urban/rural) and gender (male/female). For example, Wu (2014) found that only children have worse mental health than children with siblings on average in China, but this negative effect mainly came from rural males, whereas Zeng et al. (2020) found that the negative effects were more significant

in urban areas. It is hence of interest to apply the adjusted U test to study whether there is significant treatment effect heterogeneity among the four subpopulations: urban males, urban females, rural males and rural females.

The data we use is obtained from the Chinese Family Panel Studies (CFPS) (Xie and Hu, 2014), which is a longitudinal survey aiming at documenting changes in various aspects of Chinese society. The baseline survey was conducted in 2010. It covers 25 provinces/municipalities/autonomous regions that represent 95% of the Chinese population. The data set we focus on is a subset of the CFPS baseline sample constructed by Zeng et al. (2020). It consists of children born after 1979 with ages between 20 and 31. The data set is available a `https://rss.onlinelibrary.wiley.com/pb-assets/hub-assets/rss/Datasets/RSSA%20183.4/A1595Zeng-1600084584507.zip`. For families with more than one child, only the oldest child is included in the data set. Baseline covariates include age, ethnicity (Han or not), parents' education level (in years), family income in 2010, parents' marital status (divorced or not), parents' ages when the child was born, region type (urban/rural) and gender. The responses include three self-rated psychological measures: confidence, anxiety and desperation. All measures take integer values from 1 to 5, with a higher value indicating better mental health. We treat the only children as the treatment group and the other children as the control group. We also remove subjects with obviously erroneous information, e.g., a parent's age below 14 at the time of the child's birth or any response measure outside the range of the scale. Three children with family annual incomes higher than two million Chinese Yuan are removed because these are dramatically larger than the remainder. The final data set has 4187 subjects, with 971 in the treatment group (only children). The distributions of baseline coavariates and outcomes are summarized in the left-hand side of Table 3.5. We find that parents with only one child have higher average education level and family income. Among only children, there are large proportions of male or urban subjects compared to children with siblings. With respect to the three responses, their summary statistics are very similar between the two treatment groups. Figure 3.8

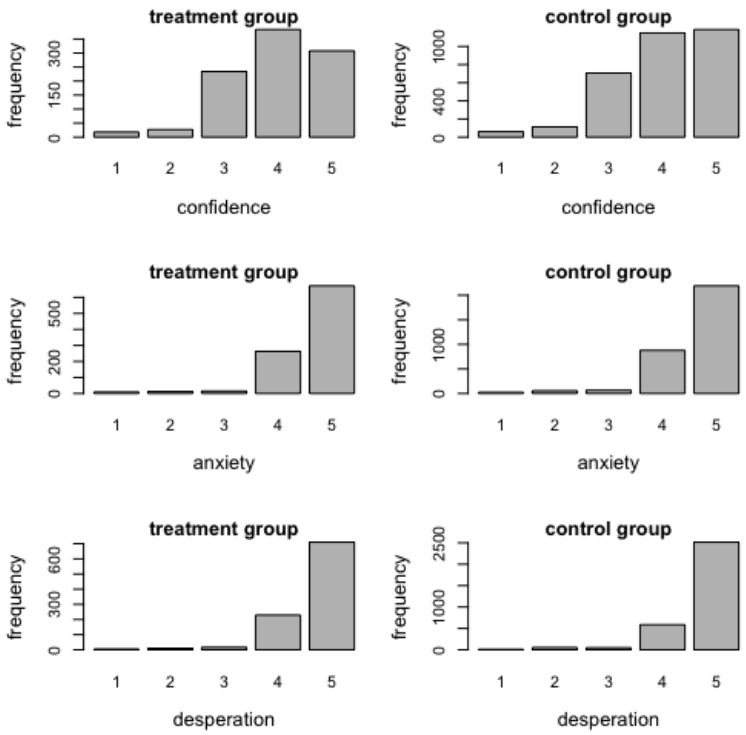


Figure 3.8: Distributions of confidence, anxiety and desperation measures in the treatment and control groups.

shows the distributions of the three responses in each treatment group. Apart from the fact that every outcome is similarly distributed in the treatment and control groups, they are all heavily left-skewed.

We first apply the weighted version of the Mann-Whitney test introduced by Satten et al. (2018) to assess overall average treatment effects with respect to the three outcomes. We standardize all baseline covariates and then fit a logistic regression to estimate propensity scores. After trimming subjects whose estimated propensity scores are outside of the overlap region, we fit the same logistic regression again with the remaining subjects and use the newly estimated propensity scores for weighting. The weights are based on formulas in (3.5). As we focus on estimating the average treatment effects, we use $h(x) = 1$. Summary statistics for the baseline and response variables after trimming and weighting are presented in the right-hand side of Table 3.5. There is a clear improvement in covariate balance, though

	Unweighted		Trimmed and Weighted	
	Only children	Children with siblings	Only children	Children with siblings
All[a]				
Sample size	971	3216	968	3216
Baseline covariates				
Maternal education (yrs)	7.95 (4.28)	4.19 (4.29)	4.44 (4.64)	4.98 (4.49)
Paternal education (yrs)	8.72 (3.98)	6.41 (4.36)	6.21 (4.56)	6.88 (4.36)
Age (yrs)	24.99 (3.38)	25.19 (3.51)	25.38 (3.65)	25.17 (3.50)
Han ethnicity	0.96 (0.20)	0.89 (0.32)	0.88 (0.32)	0.91 (0.30)
Family anuual income (Chinese Yuan)	56957.5 (58152.7)	37403.1 (44133.1)	41324.5 (51470.2)	42793.1 (54362.3)
Parental age at birth (yrs)	26.83 (3.81)	27.66 (5.11)	27.92 (5.68)	27.45 (4.99)
Maternal age at birth (yrs)	25.09 (3.44)	25.7 (4.54)	25.9 (4.67)	25.55 (4.44)
Divorce	0.03 (0.17)	0.01 (0.10)	0.01 (0.10)	0.01 (0.10)
Urban area	0.78 (0.41)	0.39 (0.49)	0.43 (0.50)	0.48 (0.50)
Male	0.59 (0.49)	0.47 (0.50)	0.49 (0.50)	0.50 (0.50)
Outcomes				
Confidence	3.96 (0.92)	4.02 (0.95)	3.95 (0.95)	4.02 (0.94)
Anxiety	4.62 (0.67)	4.60 (0.69)	4.63 (0.69)	4.61 (0.68)
Desperation	4.68 (0.62)	4.72 (0.61)	4.69 (0.62)	4.73 (0.61)

Table 3.5: Unweighted and weighted sample means (standard deviations) of baseline characteristics and responses in treatment and control groups of the overall sample

[a] P(only child) is modeled by a logistic regression with all baseline covariates, (maternal age at birth)^3, (paternal age at birth)^3, and (family income)^3.

the summary statistics of the responses do not change much. The adjusted U-statistics and corresponding 95% confidence intervals are given in the first row of Table 3.7. Here the expectation of the adjusted U-statistic is the probability that outcome in treatment group is smaller than that in control group. Thus a value larger than 0.5 indicates a negative treatment effect, i.e., worse outcomes for only children. The U-statistics show that only children are less confident, less anxious and more desperate than children with siblings, with the 95% confidence intervals showing that none of these findings are statistically significant.

We then split the data into four strata based on gender and region type. The sample sizes and distributions of baseline and response variables in the treatment and control groups are summarized in the left-hand sides of Table 3.6. It shows that the baseline characteristics vary among strata. For instance, urban parents have higher education levels and incomes than rural parents. The proportion of males are higher among only children than children with siblings, especially in rural areas. With respect to the response variables, there is no obvious difference among these subgroups. Adjusted Mann-Whitney tests are implemented in each strata separately based on the same weighting procedure described above. The

	Unweighted		Trimmed and Weighted	
	Only children	Children with siblings	Only children	Children with siblings
Urban males[a]				
Sample size	430	580	423	558
Baseline covariates				
Maternal education (yrs)	8.66 (4.05)	5.15 (4.37)	6.58 (4.70)	6.63 (4.43)
Paternal education (yrs)	9.44 (3.76)	7.29 (4.37)	8.05 (4.14)	8.20 (4.25)
Age (yrs)	25.38 (3.33)	25.66 (3.53)	25.62 (3.46)	25.71 (3.50)
Han ethnicity	0.98 (0.14)	0.93 (0.24)	0.96 (0.20)	0.96 (0.20)
Family anuual income (Chinese Yuan)	59449.4 (60477.7)	45266.9 (45263.7)	51219.8 (56528.3)	54606.3 (64632.9)
Paternal age at birth (yrs)	26.94 (3.48)	27.90 (4.90)	27.24 (4.35)	27.36 (4.52)
Maternal age at birth (yrs)	25.16 (3.23)	26.29 (4.32)	25.61 (4.00)	25.69 (3.76)
Divorce	0.03 (0.17)	0.01 (0.10)	0.02 (0.14)	0.02 (0.14)
Outcomes				
Confidence	4.00 (0.92)	3.96 (0.98)	3.98 (0.94)	3.97 (0.93)
Anxiety	4.61 (0.7)	4.64 (0.62)	4.63 (0.67)	4.65 (0.58)
Desperation	4.67 (0.62)	4.74 (0.58)	4.67 (0.61)	4.75 (0.56)
Urban females[b]				
Sample size	331	690	330	634
Baseline covariates				
Maternal education (yrs)	9.05 (3.74)	5.83 (4.28)	7.03 (4.41)	7.2 (4.21)
Paternal education (yrs)	9.54 (3.39)	7.43 (4.13)	8.35 (3.74)	8.42 (3.91)
Age (yrs)	25.01 (3.37)	25.69 (3.55)	25.43 (3.47)	25.44 (3.46)
Han ethnicity	0.95 (0.22)	0.93 (0.24)	0.93 (0.26)	0.93 (0.24)
Family anuual income (Chinese Yuan)	64914.3 (59182.4)	50261.4 (61045.9)	55548.9 (51572.5)	55414.0 (53315.4)
Paternal age at birth (yrs)	27.07 (3.41)	27.86 (4.99)	27.03 (3.74)	27.21 (3.78)
Maternal age at birth (yrs)	25.47 (3.04)	25.91 (4.19)	25.33 (3.34)	25.54 (3.44)
Divorce	0.03 (0.17)	0.01 (0.10)	0.02 (0.14)	0.02 (0.14)
Outcomes				
Confidence	3.89 (0.87)	3.94 (0.92)	3.87 (0.87)	3.99 (0.91)
Anxiety	4.67 (0.57)	4.62 (0.67)	4.69 (0.55)	4.61 (0.67)
Desperation	4.68 (0.60)	4.73 (0.59)	4.69 (0.57)	4.73 (0.60)
Rural males[c]				
Sample size	146	942	146	927
Baseline covariates				
Maternal education (yrs)	4.92 (4.00)	2.92 (3.96)	3.54 (3.82)	3.24 (4.10)
Paternal education (yrs)	5.94 (3.90)	5.7 (4.30)	5.79 (4.00)	5.73 (4.28)
Age (yrs)	24.13 (3.37)	24.97 (3.44)	24.85 (3.56)	24.83 (3.42)
Han ethnicity	0.92 (0.28)	0.87 (0.35)	0.91 (0.28)	0.88 (0.32)
Family anuual income (Chinese Yuan)	37539.9 (39878.4)	31437.3 (38706.5)	34498.6 (32817.1)	31936.8 (31243.0)
Paternal age at birth (yrs)	25.66 (4.59)	27.66 (5.28)	27.02 (5.20)	27.27 (5.11)
Maternal age at birth (yrs)	24.08 (4.13)	25.7 (4.79)	24.93 (4.43)	25.39 (4.68)
Divorce	0.01 (0.10)	0.01 (0.10)	0.01 (0.10)	0.01 (0.10)
Outcomes				
Confidence	4.07 (0.96)	4.11 (0.93)	4.06 (0.92)	4.12 (0.94)
Anxiety	4.57 (0.75)	4.57 (0.73)	4.59 (0.77)	4.58 (0.73)
Desperation	4.73 (0.59)	4.72 (0.63)	4.71 (0.59)	4.72 (0.63)
Rural Females[d]				
Sample size	64	1004	62	950
Baseline covariates				
Maternal education (yrs)	4.48 (4.11)	3.68 (4.05)	3.55 (3.99)	3.76 (4.07)
Paternal education (yrs)	6.05 (4.41)	5.88 (4.34)	5.83 (4.46)	5.95 (4.29)
Age (yrs)	24.23 (3.27)	24.77 (3.48)	25.18 (3.63)	24.9 (3.47)
Han ethnicity	0.92 (0.26)	0.87 (0.33)	0.91 (0.28)	0.92 (0.26)
Family anuual income (Chinese Yuan)	43360.5 (59802.2)	29620.9 (29073.6)	28334.0 (25023.5)	30437.5 (29195.6)
Paternal age at birth (yrs)	27.47 (5.22)	27.40 (5.14)	27.71 (5.40)	27.39 (5.15)
Maternal age at birth (yrs)	24.97 (4.41)	25.23 (4.60)	25.35 (4.63)	25.14 (4.54)
Divorce	0.03 (0.17)	0.01 (0.10)	0.01 (0.10)	0.01 (0.10)
Outcomes				
Confidence	3.86 (0.95)	4.03 (0.95)	3.77 (0.97)	4.04 (0.95)
Anxiety	4.56 (0.68)	4.6 (0.69)	4.55 (0.65)	4.60 (0.71)
Desperation	4.61 (0.68)	4.71 (0.62)	4.64 (0.64)	4.71 (0.62)

Table 3.6: Unweighted and weighted sample means (standard deviations) of baseline characteristics and responses in treatment and control groups of four strata

[a] P(only child) is modeled by a logistic regression with all baseline covariates, (maternal age at birth)^2, and (family income)^2.

[b] P(only child) is modeled by a logistic regression with all baseline covariates, (maternal age at birth)^2, (paternal age at birth)^2, and (family income)^2.

[c] P(only child) is modeled by a logistic regression with all baseline covariates, (maternal age at birth)^2, (family income)^2, and divorce * family income.

[d] P(only child) is modeled by a logistic regression with all baseline covariates, (maternal age at birth)^2, (family income)^2, and Han * age.

Population	confidence	anxiety	desperation
All	0.523 (0.486, 0.559)	0.489 (0.464, 0.515)	0.519 (0.495, 0.542)
Urban Males	0.497 (0.457, 0.538)	0.500 (0.465, 0.535)	**0.537 (0.505, 0.569)**
Urban Females	**0.542 (0.503, 0.582)**	0.480 (0.446, 0.514)	0.526 (0.494, 0.559)
Rural Males	0.523 (0.471, 0.575)	0.492 (0.447, 0.537)	0.507 (0.466, 0.549)
Rural Females	**0.581 (0.511, 0.651)**	0.536 (0.472, 0.599)	0.540 (0.481, 0.600)

Table 3.7: Adjusted Mann-Whitney test statistics (95% CI) for different populations with respect to different response measures

baseline covariates are clearly better balanced in all strata. The adjusted Mann-Whitney test statistics and corresponding 95% confidence intervals are listed in the second to fifth rows of Table 3.7. Most tests show insignificant results except for testing desperation among urban males, and confidence among urban females and rural females. All these significant results suggest that only children's mental health is worse than children with siblings. Even these should be interpreted with caution given the large number of tests being carried out. The findings here are related but not exactly the same as those reported in Zeng et al. (2020), which found significantly negative treatment effects among both urban female and male strata for almost all responses (except for anxiety of urban females). It is worth noting that the statistical significant findings in both papers are close to the boundary of statistical insignificance, e.g., the confidence intervals of our significant tests and the credible intervals in Zeng et al. (2020) are very close to including the null value, 1/2, in the intervals.

Interpretation of the results here is challenging due to the number of strata and outcomes. A further challenge is that the results in Table 3.7 suggest similar results across strata in each column but with some attaining significance and others not. It is natural to ask whether these are significant differences across strata (see, e.g.,Gelman and Stern (2006)). The question can be addressed by assessing treatment effect heterogeneity among the four strata. We implement our proposed adjusted U test and calculate the test statistic by randomly selecting $M = 1000N$ ($N = 4030$) kernel terms with replacement as described in Section 3.4.1. The obtained p-values are respectively 0.142, 0.411 and 0.738, for the

response variables confidence, anxiety, and desperation, which indicates that there is no significant treatment effect heterogeneity among the four subpopulations for each of the three outcomes. Pairwise tests among the four strata to examine treatment effect heterogeneity regarding the three response variables are also conducted, and the p-values of the 18 tests are almost uniformly distributed, which further demonstrates that there does not appear to be treatment effect heterogeneity across gender and region types.

3.6 Discussion

In this paper, we propose a new nonparametric U test for heterogeneity of treatment effects in observational studies. Our method extends the U test in Dai and Stern (2020) for randomized experiments to observational studies by adjusting for the confounding variables using propensity score modeling. Our approach is adaptive to various choices of target population, as long as the general function $h(x)$ used to define the target population is a constant or a differentiable function of propensity score. Many target populations of interest in practice satisfy this requirement, including subjects in treatment and control groups combined, treated subjects and subjects under control.

Compared to its parametric counterpart, the LRT, the proposed adjusted U test inherits the advantages of nonparametric tests: it requires weaker modeling assumptions about the distribution of the outcome and provide a significant improvement in power for non-normally distributed data. Several simulation scenarios suggest that subject trimming based on propensity scores may improve robustness of the adjusted U test to model misspecification. There is no analytic proof of this latter result; more exploration needs to be done.

Several future working directions remain open. Firstly, we assume that for our method, all confounding variables are observed, which is untestable and may be subject to violation

in practice. It will be of interest to conduct a sensitivity analysis to address this issue. Secondly, we assume there are no missing values of the confounding variables. Multiple imputation (Schafer, 1997) can be used to resolve the issue if the values are missing at random. If they are missing not at random, it will be of interest to extend our work based on ideas from Yang et al. (2019). Thirdly, the calculation of U-statistics is based on a random sampling procedure over all pairwise comparison between strata for our method. Developing a more efficient sampling method for faster U-statistic computation will be an interesting future working direction. Fourthly, it will be of interest to extend our test statistic for high-dimensional covariates based on the results in He et al. (2021).

Chapter 4

Sensitivity Analysis for the Adjusted Mann-Whitney Test with Observational Studies

4.1 Introduction

In randomized experiments, the Mann-Whitney test (Mann and Whitney, 1947) is a popular U-statistic-based nonparametric test to assess the significance of treatment effect. Compared to its parametric counterpart, the two-sample t-test, the Mann-Whitney test is preferred especially when the outcome distributions deviate far from normal distributions (Zimmerman, 1998; Lehmann, 2004). Recently, Satten et al. (2018) proposed an adjusted Mann-Whitney test for testing the existence of treatment effects in observational studies. Their test statistic is based on using inverse probability weighting (IPW) to control for confounding variables. Similar to many other approaches (e.g., propensity score matching) that seek to extract causation from observational studies (e.g., Rosenbaum and Rubin, 1983; Vegetabile et al.,

2020; Imbens and Rubin, 2015; Imai and Ratkovic, 2014; Li and Li, 2019), the adjusted Mann-Whitney test relies heavily on the unconfoundedness assumption (Rubin, 1990), which assumes that all confounding variables are observed.

The possible violation of the unconfoundedness assumption has led to lots of criticism and debates about the aforementioned approaches in the past few decades, since this assumption is not testable by empirical data and is unlikely to be satisfied in all applications. In practice, one may naturally worry that an unobserved confounder could possibly overturn the conclusion of a test or analysis. Some early discussion on this topic dates back to Fisher (1958), where Fisher raises the question that the observed "causation" between lung cancer and tobacco smoking may come from another agent, such as a genetic component, which contributes to both smoking behaviour and lung cancer. In a follow-up publication, Cornfield et al. (1959) presented a sensitivity analysis solution to Fisher's questions by showing that if there exists an agent that can explain away the causal relationship between smoking and lung cancer, that agent has to be nine-fold more prevalent in smokers than non-smokers. In the absence of such important agent, the casual relationship between smoking and lung cancer stands. And of course subsequent research has confirmed that relationship.

Since the pioneering work by Cornfield et al. (1959), there is a vast literature on developing sensitivity analysis approaches for mean estimation problems with missing data and treatment effect estimation in observational studies (e.g., Rosenbaum, 2002c; Imbens, 2003; Rosenbaum, 2002a; VanderWeele and Ding, 2017; Yang and Lok, 2018; Cinelli and Hazlett, 2020; Zhao et al., 2019). Different sensitivity parameters are often chosen for different studies. For instance, Rosenbaum (2002a) focuses on studies with matched cases and chooses the sensitivity parameter as the threshold for ratios of the probabilities of receiving treatment for each pair of matched subjects from different treatment groups. Hosman et al. (2010) and Cinelli and Hazlett (2020) study linear regression models and consider sensitivity parameters as thresholds for the associations between unobserved covariates with the treatment

assignments and outcomes. Zhao et al. (2019) focuses on IPW-based estimators and choose sensitivity parameters as thresholds of odds ratios between estimated propensity scores and the desired probabilities, which are the probabilities of receiving a certain treatment conditional on covariates and outcomes.

The main goal of this paper is to extend the ongoing methodological development of sensitivity analysis for the adjusted Mann-Whitney test. The results can hence be helpful in quantifying and understanding the impact of violation of the unconfoundedness assumption on results of the test. We utilize Zhao et al. (2019)'s sensitivity framework, which was also considered by Tan (2006). Under this framework, there is no need to specify/model the exact relationships between the unobserved confounders with the treatment assignment and the outcome. Instead, we consider a set of sensitivity models by setting thresholds for the odds ratio between the propensity scores obtained with and without unobserved confounders. Then we develop a bootstrap approach that efficiently generates the range of point estimates for the parameter underlying the test over all sensitivity models, and obtain a sensitivity interval that covers the true parameter with a desired nominal coverage probability, as long as the data generating mechanism is included in the set of pre-specified sensitivity models. This framework can naturally apply to cases where there exist more than one unobserved confounder. Compared to the bootstrap approach used in Zhao et al. (2019), the optimization problem involved in our work is quite different and challenging due to a large number of variables and conditions involved. Based on the characteristics of the adjusted Mann-Whitney test statistic, we derive several theorems that can help solve the optimization problem in an efficient way. Furthermore, we generalize our approach to handle a broad class of propensity-score-based adjusted U-statistics, which includes the missing data problem considered in Zhao et al. (2019) as a special example.

The remainder of the paper is structured as follows. Section 4.2 provides a brief review of the adjusted Mann-Whitney test. Section 4.3 describes our method that assesses the

robustness of the adjusted Mann-Whitney test to the violation of the unconfoundedness assumption. Section 4.4 extends this approach to more general propensity-score-based adjusted U-statistics. A simulation study is conducted in Section 4.5. In Section 4.6, we apply our approach to two case studies, a labor program effectiveness evaluation and a mental health evaluation of China's one-child policy. We conclude with a discussion of future work in Section 4.7.

4.2 Review of the adjusted Mann-Whitney Test

We adopt Neyman and Rubin's potential outcome framework (Rubin, 1974) and review the adjusted Mann-Whitney Test in this section. Suppose there are n independent study subjects, and consider $\big(T, Y(1), Y(0), X\big) = \{\big(T_i, Y_i(1), Y_i(0), X_i\big);\ i = 1, \cdots, n\}$, where $T \in \{0, 1\}$ denotes the treatment indicator, $X \in \mathcal{X}$ denotes the observed pre-treatment covariates, and $(Y(1), Y(0)) \in \mathcal{R}^2$ denotes the potential outcomes of units under treatment and control. The observed outcome is $Y = Y(1)T + Y(0)(1 - T)$. We use $(Y^t, X^t) = \{(Y_i^t, X_i^t);\ i = 1, \cdots, n_t\}$ and $(Y^c, X^c) = \{(Y_i^c, X_i^c);\ i = 1, \cdots, n_c\}$ to denote the observed outcomes and pre-treatment covariates of subjects in the treatment and control groups, respectively, where n_t and n_c are the number of subjects in treatment and control groups, and we assume there exist constants $\lambda^\omega \in (0, 1)$ such that $\frac{n_\omega}{n} \to \lambda^\omega$ as $n \to \infty$ for $\omega \in \{t, c\}$.

In order to measure the treatment effect, we compare the marginal distributions of $Y(1)$ and $Y(0)$. This can be done conveniently for randomized experiments, where T is independent of $(Y(1), Y(0))$ and hence the distribution of $(Y|T = a)$ is equal to the marginal distribution of $Y(a)$ for $a \in \{0, 1\}$. When the outcome distributions deviate from normal distributions, the Mann-Whitney test is usually preferred compared to its parametric counterpart, the

two-sample t-test. The test statistic of the Mann-Whitney test is a two-sample U-statistic

$$U = \frac{1}{n_t n_c} \sum_{i=1}^{n_t} \sum_{j=1}^{n_c} \phi(Y_i^t; Y_j^c), \tag{4.1}$$

with kernel function of $\phi(y^t; y^c) = I(y^t < y^c) + \frac{1}{2} I(y^t = y^c)$, and the second term is used to account for possible ties from discrete distributions. Under the null hypothesis that $Y(1)$ and $Y(0)$ follow the same distribution, the test statistic U has an expectation of 0.5 and approximately follows a normal distribution as n goes to infinity.

In observational studies with confounding variables, the independence assumption between T and $(Y(1), Y(0))$ does not hold anymore. As a result, the distributions of the observed outcomes in the treatment and control groups, $(Y|T = 1)$ and $(Y|T = 0)$, are different from the marginal distributions of $Y(1)$ and $Y(0)$. Directly applying the Mann-Whitney test can be misleading. One solution is to utilize the inverse probability weighting mechanism (Horvitz and Thompson, 1952), that is, each subject is weighted by the inverse of their group membership probability, and then the weighted distributions of $(Y, X)|T = 1$ and $(Y, X)|T = 0$ become the same as the marginal distributions of $(Y(1), X)$ and $(Y(0), X)$. Specifically, we weight subjects in the treatment group by $\frac{1}{e_1(X,Y(1))}$ and those in the control group by $\frac{1}{1-e_0(X,Y(0))}$, where $0 < e_a(X, Y(a)) = P(T = 1|Y(a), X) < 1$ for $a \in \{1, 0\}$. To ensure the identifiability of $e_a(X, Y(a))$, the unconfoundedness assumption, that is, $T \perp\!\!\!\perp (Y(1), Y(0))|X$, is usually imposed (Rubin, 1990). Under this assumption, we have $e_a(X, Y(a)) = P(T = 1|Y(a), X) = P(T = 1|X) := e(X)$ for $a \in \{0, 1\}$. The conditional probability $e(X)$ is called the propensity score in the observational studies literature (Rosenbaum and Rubin, 1983) and is identifiable with the available data.

Under the unconfoundedness assumption, Satten et al. (2018) applied the inverse probability weighting method to study a general family of two-sample U-statistics, which includes the Mann-Whitney test statistic as a special example. They proposed to use a logistic regression

model to obtain estimators of $e(X)$, denoted by $\hat{e}(X)$. The weights for treated subjects are $w_i^t = \frac{1}{\hat{e}(X_i^t)}$, and the weights for subjects under control are $w_j^c = \frac{1}{1-\hat{e}(X_j^c)}$. Due to the mutual independence among the subjects, each item $\phi(Y_i^t; Y_j^c)$ in the U-statistic is then weighted by the product of the weights for subjects i and j. The adjusted U-statistic is

$$U_A = \frac{1}{(\sum_{i=1}^{n_t} w_i^t)(\sum_{j=1}^{n_c} w_j^c)} \sum_{i=1}^{n_t} \sum_{j=1}^{n_c} w_i^t w_j^c \phi(Y_i^t; Y_j^c). \tag{4.2}$$

Under the unconfoundedness assumption, Satten et al. (2018) showed that if $Y(0)$ and $Y(1)$ follow the same distribution (i.e., no treatment effect) and $\hat{e}(X)$ is a consistent estimator of $e(X)$, then U_A would converge in distribution to a normal distribution with an expectation of $\frac{1}{2}$ as the sample size n goes to infinity.

4.3 Sensitivity analysis for the adjusted Mann-Whitney test

In practice, a major concern for the adjusted Mann-Whitney test described above is the possible violation of the unconfoundedness assumption, which is not testable by the available data due to the unobserved potential outcomes. The main goal of this paper is to address this concern by conducting a formal sensitivity analysis to assess the reliability of the adjusted Mann-Whitney test under the violation of the unconfoundedness assumption. In particular, we adopt the marginal sensitivity framework introduced by Zhao et al. (2019) and Tan (2006) to measure the robustness of the adjusted Mann-Whitney test.

4.3.1 Asymptotic sensitivity interval for the adjusted Mann-Whitney test

The violation of the unconfoundedness assumption can be depicted as the difference between $e_a(x,y) = P(T=1|Y(a)=y, X=x)$ and $e(x) = P(T=1|X=x)$ for each $a \in \{0,1\}$. As the true value of $e(x)$ is unknown and we can only estimate it, we instead focus on the difference between $e_a(x,y)$ and the estimate of $e(x)$, denoted by $\hat{e}(x)$. This difference captures the degree of violation of the unconfoundedness assumption as well as the estimation bias of $e(x)$. Let $\hat{g}(x) = \log\left(\frac{\hat{e}(x)}{1-\hat{e}(x)}\right)$ and $g_a(x,y) = \log\left(\frac{e_a(x,y)}{1-e_a(x,y)}\right)$ be the log odds of $\hat{e}(x)$ and $e_a(x,y)$, and let $h_a(x,y) = \hat{g}(x) - g_a(x,y)$ be the log odds ratio. By assigning bounds to $|h_a(x,y)|$, we can control the deviation of $\hat{e}(x)$ from $e_a(x,y)$. In the literature, $h_a(x,y)$ is called a sensitivity model and we define $e_a^{(h_a)}(x,y) = \frac{1}{1+\exp\{h_a(x,y)-\hat{g}(x)\}}$ for $a \in \{0,1\}$. Then for each $a \in \{0,1\}$, we consider a set of sensitivity models for $e_a(x,y)$ with sensitivity parameter λ_a ($\lambda_a \geq 0$) defined as

$$\mathcal{H}_a(\lambda_a) = \{h_a(x,y) : |h_a(x,y)| \leq \lambda_a \text{ for all } \ x \in \mathcal{X}, y \in \mathcal{R}\}. \tag{4.3}$$

These notations are set to be consistent with Zhao et al. (2019). When $\lambda_a = 0$, there is only one sensitivity model in $\mathcal{H}_a$, that is, $h_a(x,y) = 0$, and $e_a^{(h_a)}(x,y) = \hat{e}(x)$. As λ_a becomes larger, more models are included in $\mathcal{H}_a(\lambda_a)$, that is, $\mathcal{H}_a(\lambda_{a,1}) \subseteq \mathcal{H}_a(\lambda_{a,2})$ for any $0 \leq \lambda_{a,1} \leq \lambda_{a,2} \leq \infty$.

For each pair of sensitivity models (h_1, h_0), we weight each subject in the treatment group by $w_i^{t,(h_1)} = \frac{1}{e_1^{(h_1)}(x_i^t, y_i^t)} = 1 + \exp\{h_1(x_i^t, y_i^t) - \hat{g}(x_i^t)\}$ for $i = 1, \cdots, n_t$, and each subject in the control group by $w_j^{c,(h_0)} = \frac{1}{1-e_0^{(h_0)}(x_j^c, y_j^c)} = 1 + \exp\{\hat{g}(x_j^c) - h_0(x_j^c, y_j^c)\}$ for $j = 1, \cdots, n_c$. Then

the adjusted Mann-Whitney test statistic with a given pair of sensitivity models (h_1, h_0) is

$$\begin{aligned}
U_A^{(h_1,h_0)} =& \frac{1}{\left(\sum_{i=1}^{n_t} w_i^{t,(h_1)}\right)\left(\sum_{j=1}^{n_c} w_j^{c,(h_0)}\right)} \sum_{i=1}^{n_t}\sum_{j=1}^{n_c} w_i^{t,(h_1)} w_j^{c,(h_0)} \phi(Y_i^t; Y_j^c) \\
=& \frac{1}{\sum_{i=1}^{n_t}[1 + z_i^t \exp\{-\hat{g}(x_i^t)\}] \sum_{j=1}^{n_c}[1 + z_j^c \exp\{\hat{g}(x_j^c)\}]} \cdot \\
& \sum_{i=1}^{n_t}\sum_{j=1}^{n_c}[1 + z_i^t \exp\{-\hat{g}(x_i^t)\}][1 + z_j^c \exp\{\hat{g}(x_j^c)\}]\phi(Y_i^t; Y_j^c), \qquad (4.4)
\end{aligned}$$

where $z_i^t = \exp\{h_1(x_i^t, y_i^t)\}$ for $i \in \{1, \cdots, n_t\}$ and $z_j^c = \exp\{-h_0(x_j^c, y_j^c)\}$ for $j \in \{1, \cdots, n_c\}$. By restricting $h_1(x_i^t, y_i^t) \in \mathcal{H}_1(\lambda_1)$ and $h_0(x_j^c, y_j^c) \in \mathcal{H}_0(\lambda_0)$, we have $z_i^t \in [\Lambda_1^{-1}, \Lambda_1]$ for $i \in \{1, ..., n_t\}$, and $z_j^c \in [\Lambda_0^{-1}, \Lambda_0]$ for $j \in \{1, ..., n_c\}$, where $\Lambda_a = \exp(\lambda_a) \in [1, \infty]$ for $a \in \{0, 1\}$.

For a set of chosen sensitivity parameters λ_1 and λ_0, our sensitivity analysis seeks to find a $(1-\alpha)$-level sensitivity interval $[Lo, Up]$ such that for any data-generating mechanism with $h_1 \in \mathcal{H}_1(\lambda_1)$ and $h_0 \in \mathcal{H}_0(\lambda_0)$, we have $P\left(E[U_A^{(h_1,h_0)}] \in [Lo, Up]\right) \geq 1-\alpha$, i.e., the sensitivity interval is guaranteed to have a desired coverage under the model misspecification allowed for by the choice of λ_1 and λ_0. Moreover, if $[Lo, Up]$ satisfies $\liminf_{n\to\infty} P(E[U_A^{(h_1,h_0)}] \in [Lo, Up]) \geq 1-\alpha$, we call it a $(1-\alpha)$-level asymptotic sensitivity interval. The sensitivity interval and the asymptotic sensitivity interval have been studied and found useful for sensitivity analysis in the literature (e.g., Rosenbaum, 2002c; Zhao et al., 2019).

One obvious way to construct a $(1-\alpha)$-level asymptotic sensitivity interval for the expected value of the adjusted Mann-Whitney test statistic is by taking the union of all $[Lo^{(h_1,h_0)}, Up^{(h_1,h_0)}]$ over $(h_1, h_0) \in (\mathcal{H}_1(\lambda_1), \mathcal{H}_0(\lambda_0))$, where $[Lo^{(h_1,h_0)}, Up^{(h_1,h_0)}]$ is a $(1-\alpha)$-level asymptotic confidence interval for the parameter $\mu^{(h_1,h_0)} = E(U_A^{(h_1,h_0)})$. For each fixed pair of (h_1, h_0), $[Lo^{(h_1,h_0)}, Up^{(h_1,h_0)}]$ can be obtained by bootstrap. As both $\mathcal{H}_1(\lambda_1)$ and $\mathcal{H}_0(\lambda_0)$ are continuous sets and $U_A^{(h_1,h_0)}$ is a continuous function of h_1 and h_0, the $(1-\alpha)$-

level asymptotic sensitivity interval $[\widetilde{Lo}, \widetilde{Up}]$ can be obtained by choosing

$$\widetilde{Lo} = \inf_{(h_1,h_0)\in(\mathcal{H}_1(\lambda_1),\mathcal{H}_0(\lambda_0))} Lo^{(h_1,h_0)}, \qquad \widetilde{Up} = \sup_{(h_1,h_0)\in(\mathcal{H}_1(\lambda_1),\mathcal{H}_0(\lambda_0))} Up^{(h_1,h_0)}. \tag{4.5}$$

However, obtaining $[\widetilde{Lo}, \widetilde{Up}]$ based on (4.5) is computationally infeasible, as we are not able to enumerate all values of (h_1, h_0) in $(\mathcal{H}_1(\lambda_1), \mathcal{H}_0(\lambda_0))$.

Motivated by the method used to construct sensitivity intervals for the mean response with missing data in Zhao et al. (2019), we propose the following approach based on bootstrap sample quantiles. We first generate B bootstrap samples for the treatment and control groups separately as $\{(Y^t_{i,b}, X^t_{i,b})^{n_t}_{i=1};\ b = 1, \cdots, B\}$ and $\{(Y^c_{j,b}, X^c_{j,b})^{n_c}_{j=1};\ b = 1, \cdots, B\}$, and then combine them to obtain bootstrap samples for the overall data set as $\{[(Y^t_{i,b}, X^t_{i,b})^{n_t}_{i=1}, (Y^c_{j,b}, X^c_{j,b})^{n_c}_{j=1}];\ b = 1, \cdots, B\}$. For each bootstrap sample b, we further take the infimum and supremum values of $U_A^{(h_1,h_0)}$ with (h_1, h_0) ranging over $(\mathcal{H}_1(\lambda_1), \mathcal{H}_0(\lambda_0))$, and denote them by $\text{Inf}(U_{A,b})$ and $\text{Sup}(U_{A,b})$. Our approach to finding these values is described below. Given these values, we then define Lo as the $100\alpha/2$-percentile of $\{\text{Inf}(U_{A,b});\ b = 1, \cdots, B\}$, and Up as the $100(1-\alpha/2)$ percentile of $\{\text{Sup}(U_{A,b});\ b = 1, \cdots, B\}$. By the max-min inequality, we have $Lo \leq \widetilde{Lo}$ and $Up \geq \widetilde{Up}$. Therefore $[Lo, Up]$ is indeed a valid $(1-\alpha)$-level asymptotic sensitivity interval for $E[U_A]$, and hence can be used for sensitivity analysis.

4.3.2 Computing the optimums of test statistics over sensitivity models

To obtain the asymptotic sensitivity interval $[Lo, Up]$, a key step is to obtain $\text{Inf}(U_{A,b})$ and $\text{Sup}(U_{A,b})$ for each bootstrap sample b ($b \in \{1, \cdots, B\}$). This optimization problem is challenging because it involves a large number $(n_t + n_c)$ of variables, i.e., $\{z^t_i, i = 1, \cdots, n_t\}$

and $\{z_i^c, i = 1, \cdots, n_c\}$. With pre-selected sensitivity parameters λ_1 and λ_0, every z_i^t takes values in $[\Lambda_1^{-1}, \Lambda_1]$ and every z_j^c takes values in $[\Lambda_0^{-1}, \Lambda_0]$, where $\Lambda_a = \exp(\lambda_a)$ for $a \in \{1, 0\}$. To alleviate the computational burden due to the large number of variables, we identify properties of the adjusted U-statistic that will greatly simplify the optimization. The following theorems are stated for the original sample, but also apply to the bootstrap samples.

Theorem 4.1. *Consider minimizing or maximizing $U_A^{(h_1,h_0)}$ in (4.4) with $h_0 \in \mathcal{H}_0(\lambda_0)$ and $h_1 \in \mathcal{H}_1(\lambda_1)$. There exists a solution $\{(z_i^t)_{i=1}^{n_t}, (z_j^c)_{j=1}^{n_c}\}$ that satisfies $z_i^t \in \{\Lambda_1, \Lambda_1^{-1}\}$ and $z_j^c \in \{\Lambda_0, \Lambda_0^{-1}\}$ for every $i \in \{1, \cdots, n_t\}$ and $j \in \{1, \cdots, n_c\}$.*

Theorem 4.2. *Let $\tilde{\phi}(Y_{i_\omega}^\omega)$ be the collection of all kernel terms ϕ in (4.4) with $Y_{i_\omega}^\omega$ included for $i_\omega \in \{1, \cdots, n_\omega\}$ and $\omega \in \{t, c\}$. To maximize $U_A^{(h_1,h_0)}$ in (4.4) with $h_1 \in \mathcal{H}_1(\lambda_1)$ and $h_0 \in \mathcal{H}_0(\lambda_0)$, there exists a solution $\{(z_i^t)_{i=1}^{n_t}, (z_j^c)_{j=1}^{n_c}\}$ such that for every pair of $k_\omega, l_\omega \in \{1, \cdots, n_\omega\}$, $\omega \in \{t, c\}$, the following results hold: (i) $z_{k_\omega}^\omega = z_{l_\omega}^\omega$ if $\tilde{\phi}(Y_{k_\omega}^\omega) = \tilde{\phi}(Y_{l_\omega}^\omega)$; and (ii) $z_{k_\omega}^\omega \geq z_{l_\omega}^\omega$ if $\tilde{\phi}(Y_{k_\omega}^\omega) \succ \tilde{\phi}(Y_{l_\omega}^\omega)$. Here $\succ$ is defined as $(a_1, \ldots, a_k) \succ (b_1, \ldots, b_k)$ if $a_i \geq b_i$ for every $i = 1, \ldots, k$ and at least one of the inequalities strictly holds.*

Proofs of these two theorems are given in the Appendix B. Theorem 4.1 can be obtained by directly taking the partial derivative of $U_A^{(h_1,h_0)}$ with respect to each of the z terms. Theorem 4.2 can be proved in a similar way to that used in Section A.3. in Zhao et al. (2019).

Theorem 4.1 implies that we only need to search over the endpoints of the feasible region for each of the z_i^t and z_i^c variables. Theorem 4.2 further provides an efficient algorithm for maximizing $U_A^{(h_1,h_0)}$ with $(h_1, h_0) \in (\mathcal{H}_1(\lambda_1), \mathcal{H}_0(\lambda_0))$. First, in the treatment and control groups, we separately sort all unique outcome values as $\tilde{Y}_1^\omega > \cdots > \tilde{Y}_{\alpha_\omega}^\omega$ for $\omega \in \{t, c\}$, where α_t and α_c are the numbers of unique values in the treatment and control groups. As $\tilde{\phi}$ is non-increasing with outcome values in the treatment group, and non-decreasing with outcome values in the control group, for every threshold value $p \in \{\tilde{Y}_1^t, \cdots, \tilde{Y}_{\alpha_t}^t, -\infty\}$ and

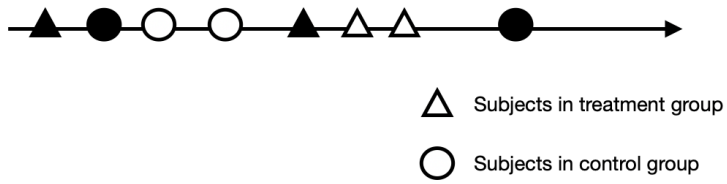


Figure 4.1: Algorithm demonstration

$q \in \{\tilde{Y}_1^c, \cdots, \tilde{Y}_{\alpha_c}^c, -\infty\}$, we set $z_i^t = \Lambda_1^{-1} I(Y_i^t \leq p) + \Lambda_1 I(Y_i^t > p)$ for $i \in \{1, \cdots, n_t\}$ and $z_j^c = \Lambda_0 I(Y_j^c \leq q) + \Lambda_0^{-1} I(Y_j^c > q)$ for $j \in \{1, \cdots, n_c\}$. We consider all $(\alpha_t + 1)(\alpha_c + 1)$ possible combinations of p and q. Then the maximizer of $U_A^{(h_1,h_0)}$ obtained from the $(\alpha_t + 1)(\alpha_c + 1)$ combinations will also maximize $U_A^{(h_1,h_0)}$ for $(h_1, h_0) \in (\mathcal{H}_1(\lambda_1), \mathcal{H}_0(\lambda_0))$. The minimizer of $U_A^{(h_1,h_0)}$ can be obtained in the same way since minimizing $U_A^{(h_1,h_0)}$ is equivalent to maximizing $1 - U_A^{(h_1,h_0)}$, which is the adjusted U-statistic obtained by replacing Y with $-Y$. This algorithm can be further improved by considering fewer thresholds when the degree of interlacement of units from treatment and control groups is small. We could sort all outcomes in the treatment and control groups together. For a list of adjacent subjects from the same group, their values of $\tilde{\phi}$ are the same. So they share the same z values, and we only need to choose one of them as a threshold. For instance, if the outcomes from treatment and control groups are sorted as in Figure 4.1, for the treatment (control) group, we only need to consider the thresholds that are outcome values of subjects represented by filled triangles (circles) and $-\infty$.

The aforementioned approach can help us obtain the range of adjusted U statistic values across all sensitivity models. To obtain the sensitivity interval, we can apply this approach to each bootstrap sample b ($b \in \{1, \cdots, B\}$) and generate $\text{Inf}(U_{A,b})$ and $\text{Sup}(U_{A,b})$. Then the $(1 - \alpha)$-level asymptotic sensitivity interval $[Lo, Up]$ can be obtained by taking the $100\alpha/2$ percentile of $\{\text{Inf}(U_{A,b})_{b=1}^B\}$ and the $100(1 - \alpha/2)$ percentile of $\{\text{Sup}(U_{A,b})_{b=1}^B\}$.

4.3.3 Testing the treatment effect in treatment group

In some situations, there is an interest in studying the treatment effect within the treatment group, that is, the difference between the conditional distributions of $Y(1)|T = 1$ and $Y(0)|T = 1$. We discuss how the adjusted Mann-Whitney test and its sensitivity analysis can be conducted in this case. Note that although the target population has changed, a similar weighting technique can still be applied for observational studies, with the primary difference that we only need to reweight subjects in the control group as the conditional distribution of observed outcomes in the treatment group, $Y|T = 1$, is the same with that of $Y(1)|T = 1$. After weighting subjects in the control group by $\frac{e_0(X,Y(0))}{1-e_0(X,Y(0))}$, the weighted distribution of $(Y, X)|T = 0$ would be equal to $(Y(0), X)|T = 1$.

We still assume the unconfoundedness assumption, i.e., $e(X) = e_0(X, Y(0))$ and use the estimator of $e(X)$, denoted by $\hat{e}(X)$, to replace $e_0(X, Y(0))$ in the weights for subjects under control. The adjusted Mann-Whitney test statistic measuring the treatment effect for treated subjects then has the same form as in (4.2), except with weights $w^t_{ATT,i} = 1$ for $i = 1, \cdots, n_t$ and $w^c_{ATT,j} = \frac{\hat{e}(X_j)}{1-\hat{e}(X_j)}$ for $j = 1, \cdots, n_c$. Those weights were also considered in Satten et al. (2018). Our test statistic in the average treatment effect on treated (ATT) case is

$$U_{ATT} = \frac{1}{n_t \sum_{j=1}^{n_c} w^c_{ATT,j}} \sum_{i=1}^{n_t} \sum_{j=1}^{n_c} w^c_{ATT,j} \phi(Y^t_i; Y^c_j). \tag{4.6}$$

For sensitivity analysis, we adopt the same framework as introduced above. Because we only need to study the deviation of $\hat{e}(X)$ from $e_0(X, Y(0))$ in this case, we only consider sensitivity models for $e_0(X, Y(0))$. After specifying the sensitivity parameter $\lambda_0 \geq 0$, the set of sensitivity models are

$$\mathcal{H}_0(\lambda_0) = \{h_0(x, y) : |h_0(x, y)| \leq \lambda_0 \text{ for all } \ x \in \mathcal{X}, y \in R\}. \tag{4.7}$$

For a fixed sensitivity model $h_0 \in \mathcal{H}_0(\lambda_0)$, we define $e_0^{(h_0)}(x,y) = \frac{1}{1+\exp\{h_0(x,y)-\hat{g}(x)\}}$, where $\hat{g}(x) = \log\left(\frac{\hat{e}(x)}{1-\hat{e}(x)}\right)$. The adjusted U-statistic under the sensitivity model h_0 then becomes

$$\begin{aligned} U_{ATT}^{(h_0)} =& \frac{1}{n_t \sum_{j=1}^{n_c} w_{ATT,j}^{c,(h_0)}} \sum_{i=1}^{n_t} \sum_{j=1}^{n_c} w_{ATT,j}^{c,(h_0)} \phi(Y_i^t; Y_j^c) \\ =& \frac{1}{n_t \sum_{j=1}^{n_c} z_j^c \exp\{\hat{g}(x_j^c)\}} \cdot \sum_{i=1}^{n_t} \sum_{j=1}^{n_c} z_j^c \exp\{\hat{g}(x_j^c)\} \phi(Y_i^t; Y_j^c), \end{aligned} \tag{4.8}$$

where $w_{ATT,j}^{c,(h_0)} = \frac{e_0^{(h_0)}(x_j^c, y_j^c)}{1-e_0^{(h_0)}(x_j^c, y_j^c)}$, $z_j^c = \exp\{-h_0(x_j^c, y_j^c)\} \in [\Lambda_0^{-1}, \Lambda_0]$ for $j \in \{1, \cdots, n_c\}$ and $\Lambda_0 = \exp(\lambda_0)$.

To obtain a sensitivity interval, we use the same bootstrap approach described in Section 4.3.1. Then similarly with Section 4.3.2, we can obtain the optimums of $U_{ATT}^{(h_0)}$ with h_0 ranging over $\mathcal{H}_0(\lambda_0)$ in a computationally efficient way by showing that $U_{ATT}^{(h_0)}$ has similar properties as $U_A^{(h_1,h_0)}$ in (4.4). The next two theorems give the details.

Theorem 4.3. *To maximize or minimize $U_{ATT}^{(h_0)}$ in (4.8) with h_0 ranging over $\mathcal{H}_0(\lambda_0)$, there is a solution $\{z_j^c, j = 1, \cdots, n_c\}$ such that $z_j^c \in \{\Lambda_0, \Lambda_0^{-1}\}$ for every $j = 1, \cdots, n_c$.*

Theorem 4.4. *Let $\tilde{\phi}(Y_j^c)$ be the collection of all kernel terms ϕ in (4.8) with Y_j^c included for every $j \in \{1, \cdots, n_c\}$. To maximize $U_{ATT}^{(h_0)}$ in (4.8) with $h_0 \in \mathcal{H}_0(\lambda_0)$, there exists a solution $\{z_j^c, j = 1, \cdots, n_c\}$ such that for every pair of $k, l \in \{1, \cdots, n_c\}$, the following results hold: (i) $z_k^c = z_l^c$ if $\tilde{\phi}(Y_k^c) = \tilde{\phi}(Y_l^c)$; and (ii) $z_k^c \geq z_l^c$ if $\tilde{\phi}(Y_k^c) \succ \tilde{\phi}(Y_l^c)$.*

Proofs of these two theorems are very similar to those for Theorem 4.1 and Theorem 4.2 and hence are omitted. Based on the two theorems, we can conveniently obtain the maximum value of $U_{ATT}^{(h_0)}$ by first sorting the unique values of outcomes in the control group as $\tilde{Y}_1^c > \cdots > \tilde{Y}_{\alpha_c}^c$, where α_c is the number of unique outcome values in the control group. As $\tilde{\phi}$ is non-decreasing, for each threshold value $q \in \{\tilde{Y}_1^c, \cdots, \tilde{Y}_\alpha^c, -\infty\}$, we set $z_j^c = \Lambda_0 I(Y_j^c \leq q) + \Lambda_0^{-1} I(Y_j^c > q)$ for $j = 1, \cdots, n_c$. The maximizer of $U_{ATT}^{(h_0)}$ over $q \in \{0, \cdots, n_c\}$ yields the maximizer of $U_{ATT}^{(h_0)}$ with $h_0 \in \mathcal{H}_0(\lambda_0)$. Moreover, minimizing $U_{ATT}^{(h_0)}$ is equivalent to

maximizing $1-U_{ATT}^{(h_0)}$, which is equal to $U_{ATT}^{(h_0)}$ with all outcomes Y replaced by $-Y$. Similarly to Section 4.3.2, the optimization process can be further improved by trying fewer threshold values. After sorting the outcomes from both treatment groups, the adjacent subjects from the control group can be assumed to have the same z values since the $\tilde{\phi}$ values for them are the same.

4.4 Extensions to other adjusted multi-sample U-statistics

The proposed sensitivity analysis framework can be extended to a more general scenario involving adjusted S-sample ($S \geq 1$) U-statistics of degree $\underbrace{(1, \cdots, 1)}_{\text{S}}$ (i.e., the kernel function $\phi(y_1; \cdots; y_S)$ only has 1 argument for each sample) where subjects in some samples are weighted. More specifically, consider S independent samples $\{(Y_{s,i}, X_{s,i}, T_{s,i})_{i=1}^{n_s}, s = 1, \cdots, S\}$ and let the first S' ($0 < S' \leq S$) samples be weighted by some functions of propensity score estimators $\hat{e}_s(X_s) = \hat{P}(T_s = 1|X_s)$. Then we can specify sensitivity models for every $e_s(X_s, Y_s) = P(T_s = 1|Y_s, X_s)$ with $s \in \{1, \cdots, S'\}$ as in (4.3), i.e., with pre-specified sensitivity parameters λ_s ($\lambda_s \geq 0$) for $s = 1, \cdots, S'$, the sensitivity models are

$$\mathcal{H}_s(\lambda_s) = \{h_s(x, y) : |h_s(x, y)| \leq \lambda_s \text{ for all } \ x \in \mathcal{X}, y \in \mathcal{R}\}, \tag{4.9}$$

where $h_s(x, y) = \hat{g}_s(x) - g_s(x, y)$, $\hat{g}_s(x) = \log\left(\frac{\hat{e}_s(x)}{1-\hat{e}_s(x)}\right)$ and $g_s(x, y) = \log\left(\frac{e_s(x,y)}{1-e_s(x,y)}\right)$.

Let f_s be a real-valued positive function for $s = 1, \cdots, S'$, with weights $w_{i_s}^s = f_s(\hat{e}(X_s))I(s \leq S') + I(s > S')$ for $i_s = 1, \cdots, n_s$ and $s = 1, \cdots, S$, the adjusted U-statistic is

$$U_S = \frac{1}{\sum_{i_1=1}^{n_1} w_{i_1}^1 \cdots \sum_{i_S=1}^{n_S} w_{i_S}^S} \sum_{i_1=1}^{n_1} \cdots \sum_{i_S=1}^{n_S} w_{i_1}^1 \cdots w_{i_S}^S \phi(Y_{1,i_1}; \cdots; Y_{S,i_S}). \tag{4.10}$$

For a set of fixed sensitivity models $(h_1, \cdots, h_{S'})$, the weights become $w_{i_s}^{s,(h_s)} = f_s(e^{(h_s)}(X_s, Y_s))I(s \leq$

$S') + I(s > S')$, where $e^{(h_s)}(X_s, Y_s) = \frac{1}{1+\exp\{h_s(X_s,Y_s)-\hat{g}_s(X_s)\}}$. Thus the adjusted U-statistic under the sensitivity models becomes

$$U_S^{(h_1,\cdots,h_{S'})} = \frac{1}{\sum_{i_1=1}^{n_1} w_{i_1}^{1,(h_1)} \cdots \sum_{i_S=1}^{n_S} w_{i_S}^{S,(h_S)}} \sum_{i_1=1}^{n_1} \cdots \sum_{i_S=1}^{n_S} w_{i_1}^{1,(h_1)} \cdots w_{i_S}^{S,(h_S)} \phi(Y_{1,i_1}; \cdots ; Y_{S,i_S}). \tag{4.11}$$

Before we present how to obtain the sensitivity interval for the expectation of U_S among sensitivity models $\mathcal{H}_s(\lambda_s)$ for $s \in \{1, \cdots, S'\}$, we first discuss two relevant examples.

The first example is conducting sensitivity analysis for mean response estimation with missing data, considered by Zhao et al. (2019). In this case, $S = S' = 1$. Suppose that there are n i.i.d. observations $(T, X, Y) = \{(T_i, X_i, Y_i), i = 1, \cdots, n\}$, where T is an indicator of the response being observed, X is a vector of observed baseline covariates, and Y is the collection of potential outcomes. Without loss of generalization, we assume the outcomes of the first n' $(0 < n' \leq n)$ subjects are observed. An unbiased estimator of $\mu = E(Y)$ is $\left(\sum_{i=1}^{n'} \frac{1}{e(X_i,Y_i)}\right)^{-1} \sum_{i=1}^{n'} \frac{Y_i}{e(X_i,Y_i)}$, where $e(X_i, Y_i) = P(T_i = 1|X_i, Y_i)$. As $e(X, Y)$ is unidentifiable with available data, in practice, a commonly used estimator for μ is

$$\hat{\mu} = \left(\sum_{i=1}^{n'} \frac{1}{\hat{e}(X_i)}\right)^{-1} \sum_{i=1}^{n'} \frac{Y_i}{\hat{e}(X_i)}. \tag{4.12}$$

This estimator is a one-sample adjusted U-statistic with kernel function $\phi(y) = y$ and weights $w_i = \frac{1}{\hat{e}(X_i)}$ for $i = 1, \cdots, n'$.

Another example is comparing treatment effects of two independent strata in observational studies considered by Dai et al. (2021). In this case, $S = 4$. We use $(T_s^\omega, X_s^\omega, Y_s^\omega) = \{(T_{s,i}^\omega, X_{s,i}^\omega, Y_{s,i}^\omega), i = 1, \cdots, n_s^\omega\}$ to denote i.i.d. subjects in stratum s $(s \in \{1, 2\})$ and treatment group ω $(\omega \in \{t, c\})$. Here X_s^ω denotes the observed baseline covariates, Y_s^ω denotes the outcomes, and T_s^ω denotes the indicator of belonging to treatment group ω

conditional on being in stratum s. Thus $P(T_s^t|X) = 1 - P(T_s^c|X)$. The propensity scores are defined as $e_s^\omega(X_s^\omega) = P(T_s^\omega = 1|X_s^\omega) = \{e_s^\omega(X_{s,i}^\omega); i = 1, \cdots, n_s^\omega\}$. To adjust for the imbalanced baseline covariates within each stratum and compare the treatment effects, Dai et al. (2021) proposes a test statistic under the unconfoundedness assumption as

$$U_S = \frac{\sum_{i=1}^{n_1^t} \sum_{j=1}^{n_1^c} \sum_{k=1}^{n_2^t} \sum_{l=1}^{n_2^c} w_{1i}^t w_{1j}^c w_{2k}^t w_{2l}^c \phi(Y_{1,i}^t; Y_{1,j}^c; Y_{2,k}^t; Y_{2,l}^c)}{\sum_{i=1}^{n_1^t} w_{1i}^t \cdot \sum_{j=1}^{n_1^c} w_{1j}^c \cdot \sum_{k=1}^{n_2^t} w_{2k}^t \cdot \sum_{l=1}^{n_2^c} w_{2l}^c}, \tag{4.13}$$

which is an adjusted four-sample U-statistic with kernel function $\phi(y_1^t; y_1^c; y_2^t; y_2^c) = I(y_1^t - y_1^c < y_2^t - y_2^c) + \frac{1}{2} I(y_1^t - y_1^c = y_2^t - y_2^c)$. If one is interested in comparing the average treatment effects between two strata, the weights can be chosen as $w_{si}^\omega = \frac{1}{\hat{e}_s^\omega(X_{s,i}^\omega)}$ for $i \in \{1, \cdots, n_s^\omega\}$, $s \in \{1, 2\}$ and $\omega \in \{t, c\}$, that is, $S' = 4$. If the average treatment effects of the treated subjects are of interest, the weights are $w_{si}^\omega = I(\omega = t) + \frac{\hat{e}_s^\omega(X_{s,i}^\omega)}{1 - \hat{e}_s^\omega(X_{s,i}^\omega)} I(\omega = c)$ for $i \in \{1, \cdots, n_s^\omega\}$, $s \in \{1, 2\}$ and $\omega \in \{t, c\}$. In this case, $S' = 2$.

Now we apply the same bootstrap approach to obtain sensitivity intervals for $E(U_S)$ over sensitivity models in (4.9). We show in the following theorems that, under some conditions, $U_S^{(h_1, \cdots, h_{S'})}$ in (4.11) have similar properties to $U_A^{(h_0, h_1)}$ in Section 4.3.2, which makes the required optimization procedure computationally feasible.

Theorem 4.5. *For maximization or minimization of $U_S^{(h_1, \cdots, h_{S'})}$ in (4.11) with $h_s \in \mathcal{H}_s(\lambda_s)$ $(s = 1, \cdots, S')$, there exists a solution $\{\big(h_s(X_{s,i_s}, Y_{s,i_s})\big)_{i_s=1}^{n_s};\ s = 1, \cdots, S'\}$ such that every $h_s(X_{s,i_s}, Y_{s,i_s})$ maximizes or minimizes $w_{i_s}^s$ for $i_s = 1, \cdots, n_s$ and $s = 1, \cdots, S'$.*

Theorem 4.6. *For each sample s $(s = 1, \cdots, S')$, assume that the weights $w_{i_s}^{s,(h_s)}$ in (4.11) satisfy $w_{i_s}^{s,(h_s)} = a_s + z_{s,i_s} b_{s,i_s}$ for constants a_s and b_{s,i_s} $(b_{s,i_s} \geq 0)$. Then there exists a solution $\{(z_{s,i_s})_{s=1}^{n_s};\ s = 1, \cdots, S'\}$ that maximizes $U_S^{(h_1, \cdots, h_{S'})}$ in (4.11) with $h_s \in \mathcal{H}_s(\lambda_s)$ $(s = 1, \cdots, S')$. In particular, for any pair $k_s, l_s \in \{1, \cdots, n_s\}$ $(s \in \{1, \cdots, S'\})$, the following results holds: (i) $z_{s,k_s} = z_{s,l_s}$ if $\tilde{\phi}(Y_{s,k_s}) = \tilde{\phi}(Y_{s,l_s})$; and (ii) $z_{s,k_s} \geq z_{s,l_s}$ if $\tilde{\phi}(Y_{s,k_s}) \succ \tilde{\phi}(Y_{s,l_s})$.*

Proofs of these two theorems are given in the Appendix B, and they are based on similar ideas as in Theorems 4.1 and 4.2. Theorem 4.5 and 4.6 are applicable to the two aforementioned examples and the adjusted Mann-Whitney tests discussed in Section 4.3, as their weights satisfy the form required by Theorem 4.6. For example, consider the problem of mean response estimation with missing data. In this case, $w_i^{(h)} = 1+\exp\{h(X_i, Y_i) - \hat{g}(X_i)\}$, where $\hat{g}(X_i) = \log(\frac{\hat{e}(X_i)}{1-\hat{e}(X_i)})$. Thus we have $a = 1$, $z_i = \exp\{h(X_i, Y_i)\}$ and $b_i = \exp\{-\hat{g}(X_i)\}$, for $i = 1, \cdots, n'$. There is no subscript s for this example because there is only one weighted stratum. Note that Theorems 4.5 and 4.6 include Theorems 4.1–4.4 as special cases.

4.5 Simulation

We conduct simulation studies to demonstrate the finite-sample performance of our approach. Here we focus on the average treatment effect. Suppose there are $n = 200$ independent subjects randomly assigned to treatment and control groups. We generate observed baseline covariates $\tilde{X} = \{\tilde{X}_i, i = 1, \cdots, n\}$ i.i.d. from $\mathrm{N}(0, 1)$. Then we generate an unobserved covariate $\tilde{Z} = \{\tilde{Z}_i, i = 1, \cdots, n\}$ such that $\tilde{Z} = X + \epsilon_z$, where X is the collection of standardized values of $\tilde{X}$ (such that X has an exact variance of 1) and $\epsilon_z \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma_z^2)$. Let $R_z^2 = \frac{1}{1+\sigma_z^2}$ be the amount information in $\tilde{Z}$ that can be explained by the observed covariate X. The treatment indicators $T = \{T_i, i = 1, \cdots, n\}$ are independently generated from Bernoulli distributions with probabilities $p = \{p_i, i = 1, \cdots, n\}$ with $\log\left(\frac{p}{1-p}\right) = c(X - Z)$, where Z is the standardized $\tilde{Z}$, and the constant c is chosen to yield $\mathrm{E}[c(X - Z)] = \frac{1}{2}$ so that p will not be too close to 0 or 1. The outcomes $Y = \{Y_i, i = 1, \cdots, n\}$ satisfy $Y = X + Z + \beta T + \epsilon$ where $\epsilon \overset{\text{i.i.d.}}{\sim} N(0, 1)$ and β is the treatment effect. As the unobserved covariate Z is related to both the outcome Y and the treatment assignment T, it is a confounding variable that should be adjusted for in an analysis.

Here we investigate the relationship between R_z^2 and the robustness of the adjusted Mann-

Whitney test when Z is unobserved. We choose $R_z^2 \in \{0.99, 0.5, 0.01\}$. Smaller value of R_z^2 indicates that Z contains more unique unobserved information. For each R_z^2, we choose coefficients β such that the adjusted Mann-Whitney test statistic is around 0.25 when the propensity score model is correctly specified (i.e., Z is included). Under each simulation scenario, adjusted Mann-Whitney tests are conducted based on estimated propensity scores with and without Z separately. The point estimates and 95% confidence intervals for the probability of an outcome in the treatment group being smaller than an outcome in the control group (the estimand of the U-statistic) are summarized in the rows of the upper panel of Table 4.1 where the sensitivity parameter $\Lambda = 1$. In the table, we find that when the propensity score models are correctly specified, the point estimates and 95% confidence intervals indicate statistically significant positive treatment effects in all three scenarios. When the propensity score models are misspecified, the point estimates start to deviate from the true value of 0.25 and becomes closer to 0.5 as R_z^2 decreases, although the conclusions of a significant positive treatment effect are maintained as can be seen by checking the CIs. Next, we conduct sensitivity analyses by setting Z as missing with sensitivity parameters $\Lambda_0 = \Lambda_1 = \Lambda > 1$ and present the results in the same table. For each R_z^2, the smallest Λ's are found such that the 95% sensitivity intervals cover 0.5, which means the sensitivity analysis with this choice of Λ indicates the possibility that observed treatment effect does not remain significant over this full range of sensitivity models. We also find the smallest Λ such that the range of point estimates cover 0.5, which indicates that the observed test statistic may support the opposite sign of treatment effect over the set of sensitivity models. For example, when $\sigma_z = 0.1$, we find the smallest value for Λ is 2.0 such that the 95% CI starts to include .5, and Λ has to be at least 2.8 for the range of point estimates to include .5. As R_z^2 decreases, the value of Λ that is required to overturn the conclusion decreases. This is expected because the model becomes more vulnerable to model misspecification as $\tilde{Z}$ becomes more important, and it hence becomes easier to see different conclusions over the space of sensitivity models.

Propensity score model: $\log\left(\frac{p}{1-p}\right) = c(X - Z)$							
				Correct PS model		Incorrect PS model	
σ_z	R_z^2	β	Λ	point estimate	95% CI	point estimate	95% CI
0.1	0.99	2.05	1	0.250	(.178, .312)	0.258	(.188, .322)
			2.0			(.140, .420)	(.096, .506)
			2.8			(.101, .507)	(.068, .592)
1	0.5	1.79	1	0.250	(.181, .314)	0.298	(.219, .371)
			1.6			(.206, .410)	(.146, .503)
			2.3			(.150, .504)	(.103, .596)
10	0.01	1.61	1	0.250	(.182, .315)	0.316	(.234, .386)
			1.5			(.232, .413)	(.164, .504)
			2.2			(.168, .511)	(.115, .602)
Propensity score model: $\log\left(\frac{p}{1-p}\right) = c(X + Z)$							
				Correct PS model		Incorrect PS model	
σ_z	R_z^2	β	Λ	point estimate	95% CI	point estimate	95% CI
0.1	0.99	2.02	1	0.250	(.177, .317)	.248	(.174, .314)
			2.0			(.132, .412)	(.085, .509)
			2.8			(.093, .501)	(,058, .596)
1	0.5	1.67	1	0.249	(.174, .324)	.237	(.164, .308)
			2.1			(.119, .408)	(.074, .503)
			3.0			(.081, .501)	(.049, .597)
10	0.01	1.56	1	0.249	(.168, .320)	.200	(.131, .260)
			2.5			(.080, .405)	(.049, .501)
			3.6			(.054, .501)	(.031, .597)

Table 4.1: Sensitivity analysis for simulated data. The rightmost columns show sensitivity analysis results when the incorrect sensitivity model is used. Simulation scenarios are defined by R_z^2, β as defined in the Section 4.5.

We also consider another simulation setting, where $\log\left(\frac{p}{1-p}\right) = c(X - Z)$ is replaced by $\log\left(\frac{p}{1-p}\right) = c(X + Z)$, while keeping other parameter settings the same. The results are summarized in the lower panel of Table 4.1. Again when the propensity score model is correctly specified, the results indicate statistically significant positive treatment effects for all values of R_z^2. The point estimates based on incorrectly estimated propensity scores deviate more from the correct estimates (approximately 0.25) as R_z^2 decreases. This time, the point estimates become further away from 0.5 and the thresholds of Λ that are needed to include models that modify the conclusion take larger values. This is because now the unobserved covariate $\tilde{Z}$ has an opposite sign in the logistic model for p compared to the previous case; hence the model is "pushed" towards the other direction under the propensity score misspecification as $\tilde{Z}$ becomes more important.

4.6 Case Study

4.6.1 Assessing the effectiveness of a labor program

We first apply our method to a labor program study that evaluated the effect of the National Support Work (NSW) Demonstration, which was an employment program that provided working experience to disadvantaged workers. More details of this program evaluation study are given in LaLonde (1986) and Dehejia and Wahba (1999). Briefly, the program was conducted in the mid-1970s, and candidates were randomly assigned to treatment and control groups. Post-treatment earnings in 1978 (RE78) is treated as the outcome to assess the effectiveness of the program. The pre-treatment covariates include age, education, indicator of married, indicator of "no degree", race indicators, and earnings in 1974 (RE74) and 1975 (RE75). In order to examine whether techniques used in the analyses of observational studies can replicate the results from randomized experiments, LaLonde

(1986) replaced the randomized control group by two nonexperimental comparison groups, namely the Panel Study of Income Dynamics (PSID-1) cohort and the Current Population Survey-Social Security Administration File (CPS-1) cohort. In addition, two subsets from each group (PSID-2, PSID-3, CPS-2, CPS-3) were also considered. Dehejia and Wahba (1999) used a subset of data used by LaLonde (1986) which only includes male candidates with earnings in 1974 recorded. They used a propensity-score-based approach to compare the experiment treatment group with the nonexperiment comparison groups and found that their propensity-score-based approach was more robust to model misspecification than a linear regression approach to adjusting for possible confounders. In our study, we focus on the data subset constructed by Dehejia and Wahba (1999), which is available at `https://users.nber.org/~rdehejia/data/.nswdata2.html`.

Preliminary data analysis suggests that the outcome RE78 has many zero values and is heavily right-skewed. Hence the Mann-Whitney test is more appropriate than the usual two sample t-test. We first use an unadjusted Mann-Whitney test to compare the randomized study treatment group to the randomized study control group and the six non-experimental comparison groups separately. The point estimates (for the U-statistic) and the corresponding 95% confidence intervals are presented in Table 4.2. Note that the expectation of the U-statistic is the probability of an outcome in the treatment group being smaller than an outcome in the control group. Therefore the result from the randomized experiment indicates that there is a statistically significant positive treatment effect since the confidence interval $(0.388, 0.473)$ is below 0.5. Interestingly, opposite results are found in the comparisons with the nonexperimental groups. All the U-statistics except PSID-3 are uniformly greater than 0.5, with most of the confidence intervals (except CPS-3) above 0.5 as well. Thus the nonexperimental control groups provide misleading results without adjustment.

Next, we apply the adjusted Mann-Whitney test to balance the baseline covariates when non-experimental comparison groups are used. Here we are interested in the treatment effect

Comparison Group	Unadjusted U-statistic	95% Confidence Interval
NSW (randomized)	.430	(.388, .473)
PSID-1	.815	(.791, .840)
PSID-2	.558	(.504, .611)
PSID-3	.405	(.338, .471)
CPS-1	.744	(.716, .771)
CPS-2	.627	(.590, .664)
CPS-3	.527	(.479, .576)

Table 4.2: Labor program data analysis: Point estimates and 95% confidence intervals of unadjusted Mann-Whitney tests comparing treatment group to a list of comparison groups.

	NSW Treated	PSID-1		CPS-1	
	Unweighted	Unweighted	Weighted	Unweighted	Weighted
Sample Size	185	2490	2490	15992	15992
Age(yrs)	25.82 (7.14)	34.85 (10.44)	24.49 (4.86)	33.23 (11.04)	26.31 (7.15)
Education(yrs)	10.35 (2.00)	12.12 (3.08)	10.39 (1.69)	12.03 (2.87)	10.31 (1.84)
I(Black)	0.84 (0.36)	0.25 (0.44)	0.92 (0.28)	0.07 (0.26)	0.87 (0.35)
I(Hispanic)	0.06 (0.24)	0.03 (0.17)	0.03 (0.17)	0.07 (0.26)	0.05 (0.20)
I(Married)	0.19 (0.39)	0.87 (0.35)	0.10 (0.30)	0.71 (0.46)	0.16 (0.37)
I(No Degree)	0.71 (0.46)	0.31 (0.46)	0.75 (0.44)	0.30 (0.46)	0.73 (0.45)
RE74(dollars)	2095.57 (4873.4)	19428.75 (13404.2)	1395.47 (4021.4)	14016.80 (9569.5)	1929.51 (4412.2)
RE75(dollars)	1532.06 (3210.5)	19063.34 (13594.2)	1097.75 (2816.5)	13650.80 (9270.1)	1384.89 (3151.9)

Table 4.3: Labor program data analysis: weighted and unweighted mean(SD) of baseline covariates in NSW treated sample, PSID-1 and CPS-1.

on the treated candidates, so we use the weighting mechanism described in Section 4.3.3. For each non-experimental comparison group, the propensity scores are estimated using a logistic regression model with the same set of baseline covariates as in Dehejia and Wahba (1999). The unweighted and weighted distributions of baseline covariates of PSID-1 and CPS-1, as well as the covariate distributions of the randomized (NSW) treatment group are summarized in Table 4.3. It clearly shows that after weighting, the covariate distributions of PSID-1 and CPS-1 are much more similar to those in the treatment group. We report the adjusted Mann-Whitney test statistics and their associated 95% confidence intervals in Table 4.4. For all non-experimental comparison groups, after adjustment, the test statistics indicate positive treatment effects of the employment program, but only PSID-1 and CPS-1 lead to statistically significant results as shown in confidence intervals.

Although the weighting procedure balances the observed covariates well, it still remains

unclear how unobserved confounding variables may affect the conclusions. Therefore we conduct a sensitivity analysis for the treatment group in comparisons with PSID-1 and CPS-1 separately. In each comparison, we gradually increase the sensitivity parameter $\Lambda_0 = \exp(\lambda_0)$ from 1 with a step size of 0.1 to find the minimum Λ_0 such that the 95% sensitivity interval covers 0.5 and the minimum Λ_0 such that the range of point estimates covers 0.5. The results are summarized in Table 4.5. When $\Lambda_0 = 1$, the sensitivity interval is the confidence interval as the set of sensitivity models only include the estimated propensity score model. When the comparison group is PSID-1, Λ_0 has to be at least 1.2 to let the sensitivity interval cover 0.5. This indicates that if there exist unobserved confounders that can potentially modify the conclusion, then there must exist at least one study subject such that the absolute value of the log odds ratio between the subject's estimated propensity score and the subject's true propensity score (based on the knowledge of those unobserved confounders) achieves $\log(1.2)$. Similarly, we will need Λ_0 to be at least 2.0 to have a point estimates higher than 0.5 be plausible over the range of sensitivity models. As for CPS-1, we will need $\Lambda_0 \geq 1.1$ to make the sensitivity interval cover 0.5, and when $\Lambda_0 \geq 1.4$, there starts to have point estimates being greater than 0.5 among all sensitivity models. These results provide a useful quantification for the sensitivity of statistically significant findings regarding the comparison of the NSW treated group and the nonexperimental control, PSID-1 and CPS-1. A larger value of Λ_0 implies a higher level of robustness to the violation of unconfoundedeness assumption, which suggests that using the PSID-1 group as the comparison group seems to be a more robust choice than using the CPS-1 group as the comparison group.

4.6.2 Evaluating the effect of a one-child policy on children's mental health

To alleviate the rapid growth of the population in China, the Chinese government implemented the one-child policy from 1979 to 2010 to limit the number of children a family could

Comparison Group	Unadjusted U-statistic	95% Confidence Interval
PSID-1	.359	(.281, .470)
PSID-2	.399	(.293, .537)
PSID-3	.387	(.273, .530)
CPS-1	.422	(.366, .489)
CPS-2	.433	(.332, .535)
CPS-3	.478	(.369, .564)

Table 4.4: Labor program data analysis: Point estimates and 95% confidence interval of adjusted Mann-Whitney tests.

Comparison Group	Sensitivity Parameter Λ	Range of Point Estimates	95% Sensitivity Interval
PSID-1	1	.359	(.281, .470)
	1.2	(.322, .396)	(.248, .505)
	2.0	(.235, .511)	(.180, .628)
CPS-1	1	.422	(.366, .489)
	1.1	(.398, .447)	(.340, .509)
	1.4	(.340, .511)	(.287, .569)

Table 4.5: Labor program data analysis: Sensitivity analyses results for estimates with PSID-1 and CPS-1 being the comparison group.

have. Though there were exceptions, very large number of parents were allowed to have only one child. There is a rich literature studying the economic benefits and social impact of the one-child policy, including the resulting gender imbalance and forced abortions (Hesketh and Zhu, 1997). Evaluating the impact of being an only child on children's mental health has received considerable attention, as there has been a growing literature demonstrating the positive effect of siblings on children's well-being (e.g., Dunn, 1988; Gass et al., 2007).

We consider a data set derived from the Chinese Family Panel Studies (Xie and Hu, 2014), which is a national representative longitudinal survey started in 2010, aimed at documenting both economic and non-economic information about the Chinese population over time. Zeng et al. (2020) used a subset of the baseline survey to study the impact of being an only child on children's mental health. Their data set is available at `https://rss.onlinelibrary.wiley.com/pb-assets/hub-assets/rss/Datasets/RSSA%20183.4/A1595Zeng-1600084584507.zip`. It includes children born after 1979, which is the year when the one-child policy was initially implemented. For families with more than one children, the data only keeps the

oldest one. The data covers 25 provinces/municipalities/autonomous regions representing 95% of the Chinese population. The response variables are three self-rated measures for mental health: confidence, anxiety and desperation. Each measure takes integer values from 1 to 5, and a higher value indicates a better mental health condition. Baseline covariates include age, indicator of Han ethnicity, education years of parents, family income in 2010, whether the parents are divorced, parents' ages at the child's birth, indicator of urban area and gender. Only children are treated as treatment group, and non-only children as control.

Several different versions of this data set have been analyzed in previous studies (Zeng et al., 2020; Dai et al., 2021; Wu, 2014) to assess the causal effect of the one-child policy on the three response variables. It is common to divide the data into four strata for analysis based on gender and region types, that is, urban males, urban females, rural males and rural females, to account for the large disparities between families in urban and rural areas and the fact that a preference for male children was prevalent during that era, especially in rural areas. In this paper, we focus on the data set used in Dai et al. (2021); they deleted a small number of cases with anomalous values for outcome variables, parents' age or income. Dai et al. (2021) use the adjusted Mann-Whitney test to assess the treatment effect for the full data set and then separately within each of the four strata. The primary focus in that paper was on testing for heterogeneity of treatment effects across strata, and the primary conclusion favored the null hypothesis of no heterogeneity. Our goal is to conduct a sensitivity analysis for some of their findings for individual strata.

We start with a brief review of the results in Dai et al. (2021). For each of the strata, the authors use logistic regressions to estimate propensity scores and perform trimming and reweighting to balance the baseline covariates in the treatment and control groups. They then perform adjusted Mann-Whitney tests. They found significant comparison results for confidence among urban females (point est: 0.542; 95% CI: (0.503, 0.582)) and rural females (point est:0.581; 95% CI: (0.511, 0.651)), as well as desperation among urban males (point

est: 0.537; 95% CI: (0.505, 0.569)). Here point estimate is the estimated value for the adjusted U-statistic whose expectation is the probability of an only child's mental health score being smaller (i.e., worse) than that of a child with siblings. In other words, these significant point estimates (above .5) and confidence intervals indicate significantly worse mental health levels in only children compared to the levels in children with siblings.

Although the available covariates are seen in Dai et al. (2021) to be reasonably well balanced after weighting by propensity scores, it is still possible that there may be unobserved confounders. For instance, studies have found that mothers of only children tend to be less affiliative (Falbo, 1978), and a lower level of affiliation could possibly affect the personality and mental health of children. It is hence of interest to apply our proposed sensitivity analysis approach to assess the robustness of the statistically significant findings in Dai et al. (2021). In our analysis, we set sensitivity parameters for both the treatment and control groups at the same value, $\Lambda_1 = \Lambda_0 = \Lambda$. For each test, we increase Λ from 1 with a step size of 0.1 to find the minimum Λ such that the 95% sensitivity interval covers 0.5 and the minimum Λ such that the range of point estimates covers 0.5. The results are presented in Table 4.6. Note that when $\Lambda = 1$, the sensitivity interval is the same confidence interval obtained by Dai et al. (2021). For all cases, when Λ is equal to 1.1, the 95% sensitivity intervals cover 0.5, which means that the conclusions can be easily modified even by a mild violation of the unconfoundedness assumption. For the minimum Λ value such that the range of point estimates cover 0.5, we find the values for urban and rural females' confidence measures are respectively 1.3 and 1.4, and the value for urban males' desperation measure is 1.3, which means that the statistically significant result for the rural females' confidence measure is a bit less sensitive to the violation of the unconfoundedness assumption than the others. All these results suggest the need for cautious interpretation of the causality conclusions in Dai et al. (2021).

Response	Sensitivity Parameter Λ	Range of Point Estimates	95% Sensitivity Interval
Urban Females			
confidence	1	0.542	(0.503, 0.582)
confidence	1.1	(0.522, 0.563)	(0.480, 0.600)
confidence	1.2	(0.504, 0.582)	(0.462, 0.619)
confidence	1.3	(0.487, 0.599)	(0.444, 0.637)
Rural Females			
confidence	1	0.581	(0.511, 0.651)
confidence	1.1	(0.559, 0.604)	(0.479, 0.678)
confidence	1.2	(0.538, 0.625)	(0.457, 0.697)
confidence	1.3	(0.518, 0.644)	(0.437, 0.714)
confidence	1.4	(0.499, 0.661)	(0.417, 0.728)
Urban Males			
desperation	1	0.537	(0.505, 0.569)
desperation	1.1	(0.519, 0.554)	(0.487, 0.585)
desperation	1.2	(0.504, 0.571)	(0.472, 0.602)
desperation	1.3	(0.489, 0.586)	(0.458, 0.618)

Table 4.6: One-child policy study: Sensitivity analyses for statistically significant treatment effects.

4.7 Discussion

In this paper, we propose a sensitivity analysis approach that assesses the impact of violating the unconfoundedness assumption for the adjusted Mann-Whitney test developed by Satten et al. (2018). The analysis is based on the marginal sensitivity framework introduced by Zhao et al. (2019) and Tan (2006), which bypasses the need for modeling unobserved confounders and instead focuses on the deviation of the estimated propensity scores from the truth, i.e., the truth assuming the knowledge of unobserved confounders. A discussion of the relationship between this sensitivity framework and others can be found in Section 7 of Zhao et al. (2019). Our approach is also extended to treat general adjusted S-sample ($S \geq 1$) U-statistics of degree $\underbrace{(1, \cdots, 1)}_{\text{S}}$, which applies to the adjusted four-sample U-statistic in Dai and Stern (2020) and includes the missing data problem in Zhao et al. (2019) as a special example.

Several future working directions remain open. First, the proposed bootstrap approach requires solving a complicated optimization problem for each bootstrap sample; when the sample size is large and the outcomes are continuous, the computational cost can be quite

expensive. For an S-sample U-statistic with sample sizes $(n_1, \cdots, n_S)$, the complexity of the U-statistic is $O(\prod_{s=1}^{S} n_s)$. Even with the help of Theorems 4.5 and 4.6, we may need to check as many as $\prod_{s=1}^{S'}(n_s+1)$ solution candidates to obtain optimums of the U-statistic. Therefore the computational complexity for each optimization process is $O(\prod_{s=1}^{S'} n_s^2 \prod_{s=S'+1}^{S} n_s)$, which makes the total time complexity $O(B \prod_{s=1}^{S'} n_s^2 \prod_{s=S'+1}^{S} n_s)$ given B bootstrap samples. Parallel computing and more efficient algorithms will be very helpful when S is large. Second, we restrict the degree of the S-sample U-statistics to be $\underbrace{(1, \cdots, 1)}_{\text{S}}$ in this paper. A more efficient optimization procedure is needed to relax this requirement. Third, it is of interest to extend our work to study test statistics constructed by combining multiple adjusted U-statistics together. For instance, in order to assess treatment effect heterogeneity, Dai et al. (2021) combines several four-sample adjusted U-statistics to compare treatment effects across more than two strata. Our approach cannot be directly applied to this example.

Chapter 5

Conclusion and Future Directions

This dissertation presented three contributions to assessing heterogeneity of treatment effects. Chapters 2 and 3 focus on non-parametric statistical tests for heterogeneity of treatment effects across subpopulations, which is a key element of attempts to assess localized treatment effects and provide more precise treatment recommendations. A non-parametric U-statistic-based test for treatment effect heterogeneity is described in Chapter 2. It is only applicable to cases, like randomized experiments, where covariates are well balanced within each stratum. Compared to its parametric counterpart, the Likelihood Ratio Test (LRT), the U-statistic-based test is more powerful when the parametric assumption of the LRT fails. Chapter 3 extends the U-statistic-based test to observational studies by using propensity scores to adjust for observed confounders. The idea holds on the work of Satten et al. (2018) which uses inverse probability weighting to adjust for confounders for two-sample U-statistics. The adjusted U-statistic test inherits the advantages of the unadjusted test from Chapter 2, and also holds the advantages of propensity-score-based approaches (Dehejia and Wahba, 1999; Imbens and Rubin, 2015), i.e., it is more robust to model misspecification compared to regression-based approaches for adjusting for confounders.

Chapter 4 focuses on addressing a concern associated with propensity score approaches like the adjusted Mann-Whitney test proposed by Satten et al. (2018) and the adjusted U-statistic-based test for two strata in Chapter 3. The adjusted tests are robust to various distributions of the outcomes, but heavily rely on the assumption that we have observed all covariates that are correlated with both the treatment assignment and outcome. Chapter 4 describes an approach to assessing the sensitivity of the adjusted Mann-Whitney test to the violation of this assumption, and also generalizes this approach to general adjusted S-sample ($S \geq 1$) U-statistics of degree $(\underbrace{1, \cdots, 1}_{S})$.

There are several open questions related to our work that invite further study.

- The sensitivity analysis introduced in Chapter 4 is only applicable to adjusted S-sample ($S \geq 1$) U-statistics with degree $(\underbrace{1, \cdots, 1}_{S})$. However, there are many other U-statistics with higher degrees that may be of interest. For example, the signed rank test (Van der Vaart, 2000) is a one-sample U-statistic with degree two, and it is commonly used to test hypotheses the location of a distribution. If observations are missing at random in the sense of Little and Rubin (2019), an adjusted version of the signed-rank U-statistic can be constructed based on inverse probability weighting. However, the approach in Chapter 4 cannot be applied in this case, and the reason is that the efficient algorithm used to solve the optimization problem in Chapter 4 does not apply. Therefore alternative optimization strategies are needed in order to extend this approach to be applicable for higher-degree U-statistics.

- All three contributions encounter computational issues under some situations as they all involve computationally expensive U-statistics. With respect to the heterogeneity tests addressed in Chapters 2 and 3, the test statistics are functions of all pairwise unadjusted and adjusted U-statistics. As the number of strata increases, the number of required U-statistics exhibits quadratic growth. With respect to the sensitivity anal-

ysis for adjusted S-sample U-statistics in Chapter 4, the optimization procedure for each bootstrap sample requires computing the adjusted S-sample U-statistics approximately $\prod_{s=1}^{S} n_s$ times in some cases, where $(n_1, \cdots, n_S)$ are the sample sizes of the S independent samples. Since the time complexity of computing a S-sample U-statistic of degree $(\underbrace{1, \cdots, 1}_{\text{S}})$ is $O(\prod_{s=1}^{S} n_s)$, the time complexity of each optimization procedure in such cases is $O(\prod_{s=1}^{S} n_s^2)$. Though we can randomly select some terms of the U-statistics to approximate their true values to alleviate the computation burden, when the number of samples S is large, the computational time can still be problematic. Parallelization can help address this challenge.

- The approach in Chapter 4 can be directly applied to a single adjusted multi-sample U-statistics, including the adjusted four-sample U-statistic described in Chapter 3 which compares the treatment effects for two strata in observational studies. However when there are more than two strata, the test statistic is a function of multiple adjusted U-statistics, and the sensitivity analysis in Chapter 4 cannot deal with that situation. Therefore, we hope to extend the approach and find sensitivity intervals for test statistics that are functions of multiple adjusted U-statistics.

Bibliography

P. D. Allison. Testing for interaction in multiple regression. *American Journal of Sociology*, 83(1):144–153, 1977.

P. C. Austin. Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. *Pharmacoepidemiology and drug safety*, 17(12):1202–1217, 2008.

M. P. Bitler, J. B. Gelbach, and H. W. Hoynes. What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review*, 96(4):988–1012, 2006.

D. P. Byar. Assessing apparent treatment—covariate interactions in randomized clinical trials. *Statistics in Medicine*, 4(3):255–263, 1985.

I. S. Chan and G. S. Ginsburg. Personalized medicine: Progress and promise. *Annual review of genomics and human genetics*, 12:217–244, 2011.

M. Chang, S. Lee, and Y.-J. Whang. Nonparametric tests of conditional treatment effects with an application to single-sex schooling on academic achievements. *The Econometrics Journal*, 18(3):307–346, 2015.

C. Cinelli and C. Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67, 2020.

D. I. Cook, V. J. Gebski, and A. C. Keech. Subgroup analysis in clinical trials. *Medical Journal of Australia*, 180(6):289, 2004.

J. Cornfield, W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22(1):173–203, 1959.

R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, 90(3):389–405, 2008.

R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

M. Dai and H. S. Stern. A U-statistic-based test of treatment effect heterogeneity. *arXiv preprint arXiv:2012.03432*, 2020.

M. Dai, W. Shen, and H. S. Stern. Nonparametric tests for treatment effect heterogeneity in observational studies. *arXiv preprint arXiv:2103.15023*, 2021.

J. De Neve and O. Thas. A mann–whitney type effect measure of interaction for factorial designs. *Communications in Statistics-Theory and Methods*, 46(22):11243–11260, 2017.

R. H. Dehejia and S. Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448): 1053–1062, 1999.

M. A. Delgado and J. C. Escanciano. Conditional stochastic dominance testing. *Journal of Business & Economic Statistics*, 31(1):16–28, 2013.

P. Ding, A. Feller, and L. Miratrix. Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):655–671, 2016.

D. O. Dixon and R. Simon. Bayesian subset analysis. *Biometrics*, pages 871–881, 1991.

J. Dunn. Sibling influences on childhood development. *Journal of Child Psychology and Psychiatry*, 29(2):119–127, 1988.

T. Falbo. Reasons for having an only child. *Journal of population*, 1(2):181–184, 1978.

A. Feller and C. C. Holmes. Beyond toplines: Heterogeneous treatment effects in randomized experiments. *Unpublished manuscript, Oxford University*, 2009.

R. Fisher. Cigarettes, cancer, and statistics. *The Centennial Review of Arts & Science*, 2: 151–166, 1958.

M. Gail and R. Simon. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, 41(2):361–372, 1985.

K. Gass, J. Jenkins, and J. Dunn. Are sibling relationships protective? a longitudinal study. *Journal of Child Psychology and Psychiatry*, 48(2):167–175, 2007.

A. Gelman and H. Stern. The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, 60(4):328–331, 2006.

G. S. Ginsburg and H. F. Willard. Genomic and personalized medicine: foundations and applications. *Translational research*, 154(6):277–287, 2009.

J. Hájek. Asymptotic normality of simple linear rank statistics under alternatives. *The Annals of Mathematical Statistics*, pages 325–346, 1968.

Y. He, G. Xu, C. Wu, W. Pan, et al. Asymptotically independent u-statistics in high-dimensional testing. *Annals of Statistics*, 49(1):154–181, 2021.

T. Hesketh and W. Zhu. The one child family policy: the good, the bad, and the ugly; health in china, part 3. *British Medical Journal*, 314:1685–1692, 1997.

J. L. Hodges, E. L. Lehmann, et al. The efficiency of some nonparametric competitors of the t-test. *The Annals of Mathematical Statistics*, 27(2):324–335, 1956.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

C. A. Hosman, B. B. Hansen, P. W. Holland, et al. The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *Annals of Applied Statistics*, 4(2):849–870, 2010.

Y.-C. Hsu. Consistent tests for conditional treatment effects. *The Econometrics Journal*, 20 (1):1–22, 2017.

K. Imai and M. Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 243–263, 2014.

G. W. Imbens. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132, 2003.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press, 2015.

G. W. Imbens, W. K. Newey, and G. Ridder. Mean-square-error calculations for average treatment effects. *IRP discussion paper*, 2006.

K. K. Jain. *Textbook of Personalized Medicine.* Springer, 2009.

D. M. Kent, J. Nelson, I. J. Dahabreh, P. M. Rothwell, D. G. Altman, and R. A. Hayward. Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials. *International journal of epidemiology*, 45(6):2075–2088, 2016.

V. S. Korolyuk and Y. V. Borovskich. *Theory of U-statistics*, volume 273. Springer Science & Business Media, 2013.

R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, pages 604–620, 1986.

E. Lehmann et al. Robust estimation in analysis of variance. *The Annals of Mathematical Statistics*, 34(3):957–966, 1963.

E. L. Lehmann. *Elements of Large-Sample Theory.* Springer Science & Business Media, 2004.

E. L. Lehmann and H. J. D'Abrera. *Nonparametrics: Statistical Methods Based on Ranks.* Holden-day, 1975.

F. Li and F. Li. Propensity score weighting for causal inference with multiple treatments. *Annals of Applied Statistics*, 13(4):2389–2415, 2019.

F. Li, K. L. Morgan, and A. M. Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

R. J. Little and D. B. Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

C. Y. Lu. Observational studies: a review of study designs, challenges and strategies to reduce confounding. *International Journal of Clinical Practice*, 63(5):691–697, 2009.

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, pages 50–60, 1947.

S. M. McHale, K. A. Updegraff, and S. D. Whiteman. Sibling relationships and influences in childhood and adolescence. *Journal of Marriage and Family*, 74(5):913–930, 2012.

C. Na, T. A. Loughran, and R. Paternoster. On the importance of treatment effect heterogeneity in experimentally-evaluated criminal justice interventions. *Journal of Quantitative Criminology*, 31(2):289–310, 2015.

A. M. Pate and E. E. Hamilton. Formal and informal deterrents to domestic violence: The dade county spouse assault experiment. *American Sociological Review*, pages 691–697, 1992.

K. M. Patel and D. G. Hoel. A nonparametric test for interaction in factorial experiments. *Journal of the American Statistical Association*, 68(343):615–620, 1973.

P. R. Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327, 2002a.

P. R. Rosenbaum. Overt bias in observational studies. In *Observational studies*, pages 71–104. Springer, 2002b.

P. R. Rosenbaum. Sensitivity to hidden bias. In *Observational studies*, pages 105–170. Springer, 2002c.

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

P. M. Rothwell. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet*, 365(9454):176–186, 2005.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

D. B. Rubin. Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25(3):279 – 292, 1990.

P. H. Sant'Anna. Nonparametric tests for treatment effect heterogeneity with duration outcomes. *Journal of Business & Economic Statistics*, pages 1–17, 2020.

G. A. Satten, M. Kong, and S. Datta. Multisample adjusted u-statistics that account for confounding covariates. *Statistics in Medicine*, 37(23):3357–3372, 2018.

J. L. Schafer. *Analysis of Incomplete Multivariate Data*. CRC press, 1997.

Z. Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

T. J. VanderWeele and P. Ding. Sensitivity analysis in observational research: introducing the e-value. *Annals of Internal Medicine*, 167(4):268–274, 2017.

B. G. Vegetabile, D. L. Gillen, and H. S. Stern. Optimally balanced gaussian process propensity scores for estimating treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(1):355–377, 2020.

M. Voysey, S. A. C. Clemens, S. A. Madhi, L. Y. Weckx, P. M. Folegatti, P. K. Aley, B. Angus, V. L. Baillie, S. L. Barnabas, Q. E. Bhorat, et al. Safety and efficacy of the chadox1 ncov-19 vaccine (azd1222) against sars-cov-2: an interim analysis of four randomised controlled trials in brazil, south africa, and the uk. *The Lancet*, 397(10269):99–111, 2021.

L. Wu. Are only children worse off on subjective well-being?: evidence from china's one-child policy. Master's thesis, Hong Kong University of Science and Technology, 2014.

Y. Xie and J. Hu. An introduction to the china family panel studies (cfps). *Chinese Sociological Review*, 47(1):3–29, 2014.

Y. Xie, J. E. Brand, and B. Jann. Estimating heterogeneous treatment effects with observational data. *Sociological Methodology*, 42(1):314–347, 2012.

S. Yang and J. J. Lok. Sensitivity analysis for unmeasured confounding in coarse structural nested mean models. *Statistica Sinica*, 28(4):1703, 2018.

S. Yang, L. Wang, and P. Ding. Causal inference with confounders missing not at random. *Biometrika*, 106(4):875–888, 2019.

S. Zeng, F. Li, and P. Ding. Is being an only child harmful to psychological health?: evidence from an instrumental variable analysis of china's one-child policy. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(4):1615–1635, 2020.

Q. Zhao, D. S. Small, and B. B. Bhattacharya. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4):735–761, 2019.

D. W. Zimmerman. Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *The Journal of Experimental Education*, 67(1):55–68, 1998.

Appendix A

Supplementary materials for Chapter 3

A.1 Proof of Theorem 3.1

Proof. We prove the asymptotic normality of the adjusted U-statistic $U_a^{(1,2)}$ in (3.7) via approximating by four independent sets of $i.i.d.$ random variables. The asymptotic normality then holds by the Central Limit Theorem. This can be directly generalized to any $U_a^{(p,q)}$ with $1 \leq p < q \leq S$. For simplicity, we omit the superscript $(1,2)$ in the following proof and use $\approx$ to denote the equalities up to $o_p(n^{-\frac{1}{2}})$, where $n = n_1 + n_2$. Some of the notations we use here are similar to what Satten et al. (2018) used in their appendix section. Throughout the proof, we use plim to denote the limit under convergence in probability.

We set $\theta^* = \frac{1}{n_1^t n_1^c n_2^t n_2^c} E[\sum_{i=1}^{n_1^t} \sum_{j=1}^{n_1^c} \sum_{k=1}^{n_2^t} \sum_{l=1}^{n_2^c} w_{1i}^t(\beta_1) w_{1j}^c(\beta_1) w_{2k}^t(\beta_2) w_{2l}^c(\beta_2) \phi(i,j,k,l)]$,
$\theta_s^\omega = \text{plim } \bar{w}_s^\omega(\beta_s) = \text{plim } \frac{1}{n_s^\omega} \sum_{j=1}^{n_s^\omega} w_{sj}^t(\beta_s)$ $(s \in \{1,2\},\ \omega \in \{t,c\})$, and $\theta_a = \frac{\theta^*}{\theta_1^t \theta_1^c \theta_2^t \theta_2^c}$. By

first-order Taylor expansion in four variables at $(\theta^*, \theta_1^t, \theta_1^c, \theta_2^t, \theta_2^c)$, we have

$$\begin{aligned}
U_a - \theta_a =& \frac{1}{n_1^t n_1^c n_2^t n_2^c} \sum_{i=1}^{n_1^t} \sum_{j=1}^{n_1^c} \sum_{k=1}^{n_2^t} \sum_{l=1}^{n_2^c} \frac{w_{1i}^t(\hat{\beta}_1) w_{1j}^c(\hat{\beta}_1) w_{2k}^t(\hat{\beta}_2) w_{2l}^c(\hat{\beta}_2) \phi(i,j,k,l)}{\bar{w}_1^t(\hat{\beta}_1) \bar{w}_1^c(\hat{\beta}_1) \bar{w}_2^t(\hat{\beta}_2) \bar{w}_2^c(\hat{\beta}_2)} - \frac{\theta^*}{\theta_1^t \theta_1^c \theta_2^t \theta_2^c} \\
\approx& c_{11}^t(\bar{w}_1^t(\hat{\beta}_1) - \theta_1^t) + c_{11}^c(\bar{w}_1^c(\hat{\beta}_1) - \theta_1^c) + c_{12}^t(\bar{w}_2^t(\hat{\beta}_2) - \theta_2^t) + c_{12}^c(\bar{w}_2^c(\hat{\beta}_2) - \theta_2^c) \\
&+ c_2[\frac{1}{n_1^t n_1^c n_2^t n_2^c} \sum_{i=1}^{n_1^t} \sum_{j=1}^{n_1^c} \sum_{k=1}^{n_2^t} \sum_{l=1}^{n_2^c} w_{1i}^t(\hat{\beta}_1) w_{1j}^c(\hat{\beta}_1) w_{2k}^t(\hat{\beta}_2) w_{2l}^c(\hat{\beta}_2) \phi(i,j,k,l) - \theta^*]
\end{aligned} \tag{A.1}$$

where $c_{1s}^\omega = -\frac{\theta_a}{\theta_s^\omega}$ for $s \in \{1,2\}$ and $\omega \in \{t,c\}$, $c_2 = \frac{1}{\theta_1^t \theta_1^c \theta_2^t \theta_2^c}$.

Then by first-order Taylor expansion again, we have

$$\begin{aligned}
\bar{w}_s^\omega(\hat{\beta}_s) - \theta_s^\omega &= \frac{1}{n_s^\omega} \sum_{i=1}^{n_s^\omega} w_{si}^\omega(\hat{\beta}_s) - \theta_s^\omega \\
&= \frac{1}{n_s^\omega} \sum_{i=1}^{n_s^\omega} w_{si}^\omega(\hat{\beta}_s) - \frac{1}{n_s^\omega} \sum_{i=1}^{n_s^\omega} w_{si}^\omega(\beta_s) + \frac{1}{n_s^\omega} \sum_{i=1}^{n_s^\omega} w_{si}^\omega(\beta_s) - \theta_s^\omega \\
&\approx c_{3s}^\omega(\hat{\beta}_s - \beta_s) + \frac{1}{n_s^\omega} \sum_{i=1}^{n_s^\omega} w_{si}^\omega(\beta_s) - \theta_s^\omega \quad \text{for } s = 1,2; \omega = t,c
\end{aligned} \tag{A.2}$$

where $c_{3s}^\omega = \text{plim } \frac{1}{n_s^\omega} \sum_{i=1}^{n_s^\omega} \frac{\partial w_{si}^\omega(\beta_s)}{\partial \beta_s}$.

As $\hat{\beta}_1$ and $\hat{\beta}_2$ are obtained by solving estimating equations $\sum_{i=1}^{n_1} S_{1j}(\hat{\beta}_1) = 0$ and $\sum_{i=1}^{n_2} S_{2j}(\hat{\beta}_2) = 0$ respectively, again via first-order Taylor expansion,

$$\hat{\beta}_s - \beta_s \approx -J_s^{-1} \frac{1}{n_s} \sum_{i=1}^{n_s} S_{si}(\beta_s) \quad \text{for } \ s = 1,2 \tag{A.3}$$

where $J_s = \text{plim } \frac{1}{n_s} \sum_{j=1}^{n_s} \frac{\partial S_{sj}(\beta_s)}{\partial \beta_s}$. For the last component of (A.1), by first-order Taylor

expansion in two variable at the point (β_1, β_2), we have

$$\begin{aligned}
&\frac{1}{n_1^t n_1^c n_2^t n_2^c}\sum_{i=1}^{n_1^t}\sum_{j=1}^{n_1^c}\sum_{k=1}^{n_2^t}\sum_{l=1}^{n_2^c} w_{1i}^t(\hat{\beta}_1)w_{1j}^c(\hat{\beta}_1)w_{2k}^t(\hat{\beta}_2)w_{2l}^c(\hat{\beta}_2)\phi(i,j,k,l) - \theta^* \\
\approx & c_{41}(\hat{\beta}_1 - \beta_1) + c_{42}(\hat{\beta}_2 - \beta_2) + \\
&+ \frac{1}{n_1^t n_1^c n_2^t n_2^c}\sum_{i=1}^{n_1^t}\sum_{j=1}^{n_1^c}\sum_{k=1}^{n_2^t}\sum_{l=1}^{n_2^c} w_{1i}^t(\beta_1)w_{1j}^c(\beta_1)w_{2k}^t(\beta_2)w_{2l}^c(\beta_2)\phi(i,j,k,l) - \theta^*, \qquad \text{(A.4)}
\end{aligned}$$

where

$$c_{4s} = \text{plim}\ \frac{1}{n_1^t n_1^c n_2^t n_2^c}\sum_{i=1}^{n_1^t}\sum_{j=1}^{n_1^c}\sum_{k=1}^{n_2^t}\sum_{l=1}^{n_2^c} \frac{\partial w_{1i}^t(\beta_1)w_{1j}^c(\beta_1)w_{2k}^t(\beta_2)w_{2l}^c(\beta_2)}{\partial \beta_s}\phi(i,j,k,l), \quad s = 1,2 \qquad \text{(A.5)}$$

Note in (A.4), $\frac{1}{n_1^t n_1^c n_2^t n_2^c}\sum_{i=1}^{n_1^t}\sum_{j=1}^{n_1^c}\sum_{k=1}^{n_2^t}\sum_{l=1}^{n_2^c} w_{1i}^t(\beta_1)w_{1j}^c(\beta_1)w_{2k}^t(\beta_2)w_{2l}^c(\beta_2)\phi(i,j,k,l)$ is a 4-sample generalized U-statistic with kernel function $\tilde{\phi}(i,j,k,l) = w_{1i}^t(\beta_1)w_{1j}^c(\beta_1)w_{2k}^t(\beta_2)w_{2l}^c(\beta_2)\phi(i,j,k,l)$. So by the classical projection theorem (Hájek, 1968; Van der Vaart, 2000), we have

$$\begin{aligned}
&\frac{1}{n_1^t n_1^c n_2^t n_2^c}\sum_{i=1}^{n_1^t}\sum_{j=1}^{n_1^c}\sum_{k=1}^{n_2^t}\sum_{l=1}^{n_2^c} w_{1i}^t(\beta_1)w_{1j}^c(\beta_1)w_{2k}^t(\beta_2)w_{2l}^c(\beta_2)\phi(i,j,k,l) - \theta^* \\
\approx & \frac{1}{n_1^t}\sum_{i=1}^{n_1^t}\tilde{h}_1^t(Y_{1i}^t) + \frac{1}{n_1^c}\sum_{i=1}^{n_1^c}\tilde{h}_1^c(Y_{1i}^c) + \frac{1}{n_2^t}\sum_{i=1}^{n_2^t}\tilde{h}_2^t(Y_{2i}^t) + \frac{1}{n_2^c}\sum_{i=1}^{n_2^c}\tilde{h}_2^c(Y_{2i}^c) - 4\theta^*,
\end{aligned}$$

where $\tilde{h}_s^\omega(x) = E[\tilde{\phi}(1,1,1,1)|Y_{s1}^\omega = x]$ for $s \in \{1,2\}$ and $\omega \in \{t,c\}$.

Finally, back to Equation (A.1), we have

$$
\begin{aligned}
U_a - \theta_a \approx & \frac{c_{11}^t}{n_1^t}\sum_{i=1}^{n_1^t}[w_{1i}^t(\beta_1) - \theta_1^t] + \frac{c_{11}^c}{n_1^c}\sum_{i=1}^{n_1^c}[w_{1i}^c(\beta_1) - \theta_1^c] \\
& + \frac{c_{12}^t}{n_2^t}\sum_{i=1}^{n_2^t}[w_{2i}^t(\beta_2) - \theta_2^t] + \frac{c_{12}^c}{n_2^c}\sum_{i=1}^{n_2^c}[w_{2i}^c(\beta_2) - \theta_2^c] \\
& - (c_{11}^t c_{31}^t + c_{11}^c c_{31}^c + c_2 c_{41}) J_1^{-1}\frac{1}{n_1}\sum_{i=1}^{n_1} S_{1i}(\beta_1) \\
& - (c_{12}^t c_{32}^t + c_{12}^c c_{32}^c + c_2 c_{42}) J_2^{-1}\frac{1}{n_2}\sum_{i=1}^{n_2} S_{2i}(\beta_2) \\
& + \frac{c_2}{n_1^t}\sum_{i=1}^{n_1^t}\tilde{h}_1^t(Y_{1i}^t) - c_2\theta^* + \frac{c_2}{n_1^c}\sum_{i=1}^{n_1^c}\tilde{h}_1^c(Y_{1i}^c) - c_2\theta^* + \frac{c_2}{n_2^t}\sum_{i=1}^{n_2^t}\tilde{h}_2^t(Y_{2i}^t) - c_2\theta^* \\
& + \frac{c_2}{n_2^c}\sum_{i=1}^{n_2^c}\tilde{h}_2^c(Y_{2i}^c) - c_2\theta^* \\
& \triangleq \sum_{i=1}^{n_1^t}\eta_1^t(Y_{1i}^t) + \sum_{i=1}^{n_1^c}\eta_1^c(Y_{1i}^c) + \sum_{i=1}^{n_2^t}\eta_2^t(Y_{2i}^t) + \sum_{i=1}^{n_2^c}\eta_2^c(Y_{2i}^c), \qquad \text{(A.6)}
\end{aligned}
$$

where

$$
\begin{aligned}
\eta_{1i}^t &= \frac{c_{11}^t}{n_1^t}[w_{1i}^t(\beta_1) - \theta_1^t] + \frac{c_2}{n_1^t}[\tilde{h}_1^t(Y_{1i}^t) - \theta^*] - (c_{11}^t c_{31}^t + c_{11}^c c_{31}^c + c_2 c_{41}) J_1^{-1}\frac{1}{n_1} S_{1i}^t(\beta_1), \\
&\quad \text{for } i = 1, ..., n_1^t, \\
\eta_{1i}^c &= \frac{c_{11}^c}{n_1^c}[w_{1i}^c(\beta_1) - \theta_1^c] + \frac{c_2}{n_1^c}[\tilde{h}_1^c(Y_{1i}^c) - \theta^*] - (c_{11}^t c_{31}^t + c_{11}^c c_{31}^c + c_2 c_{41}) J_1^{-1}\frac{1}{n_1} S_{1i}^c(\beta_1), \\
&\quad \text{for } i = 1, ..., n_1^c, \\
\eta_{2i}^t &= \frac{c_{12}^t}{n_2^t}[w_{2i}^t(\beta_2) - \theta_2^t] + \frac{c_2}{n_2^t}[\tilde{h}_2^t(Y_{2i}^t) - \theta^*] - (c_{12}^t c_{32}^t + c_{12}^c c_{32}^c + c_2 c_{42}) J_2^{-1}\frac{1}{n_2} S_{2i}^t(\beta_2), \\
&\quad \text{for } i = 1, ..., n_2^t, \\
\eta_{2i}^c &= \frac{c_{12}^c}{n_2^c}[w_{2i}^c(\beta_2) - \theta_2^c] + \frac{c_2}{n_2^c}[\tilde{h}_2^c(Y_{2i}^c) - \theta^*] - (c_{12}^t c_{32}^t + c_{12}^c c_{32}^c + c_2 c_{42}) J_2^{-1}\frac{1}{n_2} S_{2i}^c(\beta_2), \\
&\quad \text{for } i = 1, ..., n_2^c.
\end{aligned}
$$

As we always assume in each stratum s ($s \in \{1, 2\}$), the first n_s^t subjects are in the treatment group, and the last n_s^c subjects are in the control group, here $\{S_{si}^t, i = 1, ..., n_s^t\}$ are the first n_s^t elements of $\{S_{si}, i = 1, ..., n_s\}$, and $\{S_{si}^c, i = 1, ..., n_s^c\}$ are the rest elements of it. Since the expectation of the right hand side of (A.6) is 0, the limit expectation of U_a is θ_a. By the Central Limit Theorem, Theorem 3.1 is obtained. □

A.2 Proof of Theorem 3.2

Proof. Following the proof of Theorem 3.1 for $U_a^{(1,2)}$, we define $\hat{U}_a^{(1,2)}$ as

$$\hat{U}_a^{(1,2)} = \sum_{i=1}^{n_1^t} \eta_1^{t,(1,2)}(Y_{1i}^t) + \sum_{i=1}^{n_1^c} \eta_1^{c,(1,2)}(Y_{1i}^c) + \sum_{i=1}^{n_2^t} \eta_2^{t,(1,2)}(Y_{2i}^t) + \sum_{i=1}^{n_2^c} \eta_2^{c,(1,2)}(Y_{2i}^c). \tag{A.7}$$

Thus we have

$$\sqrt{n_1 + n_2}(U_a^{(1,2)} - \theta_a^{(1,2)} - \hat{U}_a^{(1,2)}) \xrightarrow{P} 0, \quad \text{as } (n_1 + n_2) \to \infty. \tag{A.8}$$

For each $U_a^{(p,q)}$ with $1 \leq p < q \leq S$, we have $\hat{U}_a^{(p,q)}$ with the same form of (A.7) satisfying (A.8). Specifically,

$$\hat{U}_a^{(p,q)} = \sum_{j=1}^{n_p^t} \eta_p^{t,(p,q)}(Y_{pj}^t) + \sum_{j=1}^{n_p^c} \eta_p^{c,(p,q)}(Y_{pj}^c) + \sum_{j=1}^{n_q^t} \eta_q^{t,(p,q)}(Y_{qj}^t) + \sum_{j=1}^{n_q^c} \eta_q^{c,(p,q)}(Y_{qj}^c) \tag{A.9}$$

$$\sqrt{n_p + n_q}(U_a^{(p,q)} - \theta_a^{(p,q)} - \hat{U}_a^{(p,q)}) \xrightarrow{P} 0, \quad \text{as } (n_p + n_q) \to \infty. \tag{A.10}$$

Under the assumption that $\frac{n_s^\omega}{N} \to \lambda_s^\omega$ $(0 < \lambda_s^\omega < 1)$ when $N \to \infty$, for $s \in \{1, \cdots, S\}$ and $\omega \in \{t, c\}$, we have

$$\sqrt{N}\begin{pmatrix} U_a^{(1,2)} - \theta_a^{(1,2)} - \hat{U}_a^{(1,2)} \\ U_a^{(1,3)} - \theta_a^{(1,3)} - \hat{U}_a^{(1,3)} \\ \vdots \\ U_a^{(S-1,S)} - \theta_a^{(S-1,S)} - \hat{U}_a^{(S-1,S)} \end{pmatrix} \xrightarrow{p} 0, \text{ as } N \to \infty, \tag{A.11}$$

and by the multivariate Central Limit Theorem,

$$\sqrt{N}\begin{pmatrix} \hat{U}_a^{(1,2)} \\ \hat{U}_a^{(1,3)} \\ \vdots \\ \hat{U}_a^{(S-1,S)} \end{pmatrix} \xrightarrow{D} \mathcal{N}(0, \Sigma_a), \tag{A.12}$$

where $\Sigma_a = \frac{1}{\lambda_1^t}\Sigma_1^t + \frac{1}{\lambda_1^c}\Sigma_1^c + ... + \frac{1}{\lambda_S^t}\Sigma_S^t + \frac{1}{\lambda_S^c}\Sigma_S^c$, and Σ_s^ω is the covariance matrix of $(\tilde{\eta}_s^{\omega,(1,2)}, ..., \tilde{\eta}_s^{\omega,(S-1,S)})$ with

$$\tilde{\eta}_s^{\omega,(p,q)} = \begin{cases} \eta_s^{\omega,(p,q)} & \text{if } s = p \text{ or } s = q, \\ 0 & \text{o.w.} \end{cases}$$

Therefore we have

$$\sqrt{N}(U_a - \theta_a) \xrightarrow{D} \mathcal{N}(0, \Sigma_a) \text{ as } N \to \infty. \tag{A.13}$$

Theorem 3.2 is obtained. □

Appendix B

Supplementary materials for Chapter 4

B.1 Proof of Theorem 4.1

Proof. First we take the partial derivative of $U_a^{(h_1,h_0)}$ with respect to a $z_k (1 \leq k \leq n_t)$ and obtain

$$\begin{aligned}
\frac{\partial U_a^{(h_1,h_0)}}{\partial z_k^t} &= \frac{\exp\{-\hat{g}(x_k^t)\}}{\left(\sum_{i=1}^{n_t} w_i^{t,(h_1)}\right)^2 \sum_{j=1}^{n_c} w_j^{c,(h_0)}} \cdot \sum_{i=1}^{n_t}\sum_{j=1}^{n_c} \left[\phi(Y_k^t;Y_j^c) - \phi(Y_i^t;Y_j^c)\right] \cdot w_i^{t,(h_1)} w_j^{c,(h_0)} \\
&= \frac{\exp\{-\hat{g}(x_k^t)\}}{\left(\sum_{i=1}^{n_t} w_i^{t,(h_1)}\right)^2 \sum_{j=1}^{n_c} w_j^{c,(h_0)}} \sum_{i\in\{1,\ldots,n_t\}/\{k\}} \sum_{j=1}^{n_c} \left[\phi(Y_k^t;Y_j^c) - \phi(Y_i^t;Y_j^c)\right] \cdot w_i^{t,(h_1)} w_j^{c,(h_0)}.
\end{aligned}$$

Since all $w's$ are positive and $\sum_{i\in\{1,\ldots,n_t\}/\{k\}} \sum_{j=1}^{n_c} \left[\phi(Y_k^t;Y_j^c) - \phi(Y_i^t;Y_j^c)\right] \cdot w_i^{t,(h_1)} w_j^{c,(h_0)}$ does not involve z_k^t, the sign of $\frac{\partial U_a^{(h_1,h_0)}}{\partial z_k^t}$ does not depend on z_k^t. Thus given the other $z's$, $U_a^{(h_1,h_0)}$ can achieve minimum and maximum when z_k is equal to $1/\Lambda_1$ or Λ_1 for every $k = 1, \cdots, n_t$.

Similarly, we can take the partial derivative of $U_a^{(h_1,h_0)}$ with respect to z_l^c $(1 \le l \le n_c)$ as

$$\frac{\partial U_a^{(h_1,h_0)}}{\partial z_l^c} = \frac{\exp\{\hat{g}(x_l^c)\}}{\sum_{i=1}^{n_t} w_i^{t,(h_1)} (\sum_{j=1}^{n_c} w_j^{c,(h_0)})^2} \cdot \sum_{i=1}^{n_t} \sum_{j=\{1,\cdots,n_c\}/\{l\}}^{n} \Big[\phi(Y_i^t; Y_l^c) - \phi(Y_i^t; Y_j^c)\Big] \cdot w_i^{t,(h_1)} w_j^{c,(h_0)},$$

and get the conclusion that given the other $z's$, $U_a^{(h_1,h_0)}$ achieves minimum and maximum when z_l is equal to $1/\Lambda_0$ or Λ_0 for every $l = 1, \cdots, n_c$. □

B.2 Proof of Theorem 4.2

Proof. We start by proving part (i). Without loss of generalization, suppose in the treatment group, there exists a pair $k_t, l_t \in \{1, \cdots, n_t\}$ with $\tilde{\phi}(Y_{k_t}^t) = \tilde{\phi}(Y_{l_t}^t)$ and $z_{k_t}^t \neq z_{l_t}^t$. We can find another set $\{(\tilde{z}_i^t)_{i=1}^{n_t}, (\tilde{z}_j^t)_{j=1}^{n_c}\}$ with

$$\begin{aligned}
\tilde{z}_{k_t}^t = \tilde{z}_{l_t}^t &= \frac{z_{k_t}^t \exp\{\hat{g}(x_{k_t}^t)\} + z_{l_t}^t \exp\{\hat{g}(x_{l_t}^t)\}}{\exp\{\hat{g}(x_{k_t}^t)\} + \exp\{\hat{g}(x_{l_t}^t)\}}, \\
\tilde{z}_i^t &= z_i^t \quad \text{for } i \in \{1, \cdots, n_t\}/\{k_t, l_t\}, \\
\tilde{z}_j^c &= z_j^c \quad \text{for } j = 1, \cdots, n_c,
\end{aligned}$$

that also achieves the maximum value.

We then prove part (ii) by contradiction. Without loss of generalization, suppose in the treatment group there exists a pair $k_t, l_t \in \{1, \cdots, n_t\}$ with $\tilde{\phi}(Y_{k_t}^t) \succ \tilde{\phi}(Y_{l_t}^t)$ and $z_{k_t}^t < z_{l_t}^t$, and suppose there is another set $\{(\tilde{z}_i^t)_{i=1}^{n_t}, (\tilde{z}_j^t)_{j=1}^{n_c}\}$ such that

$$\begin{aligned}
\tilde{z}_i^t &= z_i^t + \epsilon \exp\{\hat{g}(x_i^t)\} \cdot I(i = k_t) - \epsilon \exp\{\hat{g}(x_i^t)\} \cdot I(i = l_t) && \text{for } i = 1, \cdots, n_t, \\
\tilde{z}_j^c &= z_j^c && \text{for } j = 1, \cdots, n_c.
\end{aligned}$$

Here ϵ is a positive small constant such that both $\tilde{z}_{k_t}^t$ and $\tilde{z}_{l_t}^t$ are within the range $[1/\Lambda_1, \Lambda_1]$.

The denominator of $U_a^{(h_1,h_0)}$ is the same under $\{(z_i)_{i=1}^{n_t},(z_j)_{j=1}^{n_c}\}$ and $\{(\tilde{z}_i)_{i=1}^{n_t},(\tilde{z}_j)_{j=1}^{n_c}\}$. With $\{(\tilde{z}_i)_{i=1}^{n_t},(\tilde{z}_j)_{j=1}^{n_c}\}$, the numerator is

$$\begin{aligned}
&\sum_{i=1}^{n_t}\sum_{j=1}^{n_c}\phi(Y_i^t;Y_j^c)[1+\tilde{z}_i^t\exp\{-\hat{g}(x_i^t)\}][1+\tilde{z}_j^t\exp\{\hat{g}(x_j^c)\}]\\
=&\sum_{i=1}^{n_t}\sum_{j=1}^{n_c}\phi(Y_i^t;Y_j^c)[1+z_i^t\exp\{-\hat{g}(x_i^t)\}][1+z_j^c\exp\{\hat{g}(x_j^c)\}]\\
&+\epsilon\sum_{j=1}^{n_c}[\phi(Y_{k_1}^t;Y_j^c)-\phi(Y_{k_2}^t;Y_j^c)]w_j^{c,(h_0)}.
\end{aligned}$$

Since $\tilde{\phi}(Y_{k_t}^t)\succ\tilde{\phi}(Y_{l_t}^t)$ and $\epsilon>0$, $w_j^{c,(h_0)}>0$ for every $j\in\{1,\cdots,n_c\}$, $\epsilon\sum_{j=1}^{n_c}[\phi(Y_{k_1}^t;Y_j^c)-\phi(Y_{k_2}^t;Y_j^c)]w_j^{c,(h_0)}>0$. So the set $\{(\tilde{z}_i^t)_{i=1}^{n_t},(\tilde{z}_j^t)_{j=1}^{n_c}\}$ leads to a larger $U_a^{(h_1,h_0)}$, which contradicts with the statement that $\{(z_i^t)_{i=1}^{n_t},(z_j^c)_{j=1}^{n_c}\}$ maximize $U_a^{(h_1,h_0)}$. □

B.3 Proof of Theorem 4.5

Proof. This theorem can be proved directly by taking the partial derivative of $U_S^{(h_1,\ldots,h_{S'})}$ with respect to each weight function $w_k^{s,(h_s)}$ ($k\in\{1,\cdots,n_s\}, s\in\{1,\cdots,S'\}$) respectively. We find that each of the partial derivatives with respect to the weight $w_k^{s,(h_s)}$ is free of $w_k^{s,(h_s)}$ itself, therefore the theorem holds.

Here we take $w_1^{1,(h_1)}$ as an example for illustration. The partial derivative of $U_S^{(h_1,\ldots,h_{S'})}$ with respect to $w_1^{1,(h_1)}$ is

$$\begin{aligned}
\frac{\partial U_S^{(h_1,\ldots,h_{S'})}}{\partial w_1^{1,(h_1)}}=&\frac{1}{\left(\sum_{i_1=1}^{n_1}w_{i_1}^{1,(h_1)}\right)^2\sum_{i_2=1}^{n_2}w_{i_2}^{2,(h_2)}\cdots\sum_{i_S=1}^{n_S}w_{i_S}^{S,(h_S)}}\cdot\\
&\sum_{i_1=2}^{n_1}\sum_{i_2=1}^{n_2}\cdots\sum_{i_S=1}^{n_S}w_{i_1}^{1,(h_1)}\cdots w_{i_S}^{S,(h_S)}\phi(Y_{1,1};Y_{2,i_2};\cdots;Y_{S,i_S}),
\end{aligned}$$

which does not involve $w_1^{1,(h_1)}$.

□

B.4 Proof of Theorem 4.6

Proof. We first prove part (i). Suppose there exists a pair $k_r, l_r \in \{1, \cdots, n_r\}$ ($r \in \{1, \cdots, S'\}$) with $\tilde{\phi}(Y_{r,k_r}) = \tilde{\phi}(Y_{r,l_r})$ and $z_{r,k_r} \neq z_{r,l_r}$. Then there is another set $\{(\tilde{z}_{s,i_s})_{i_s=1}^{n_s};\ s = 1, \cdots, S'\}$ that also achieves the maximum with $\tilde{z}_{r,k_r} = \tilde{z}_{r,l_r}$ and other z values unchanged. In particular,

$$\tilde{z}_{s,i_s} = z_{s,i_s} + I(s = r, i_s \in \{k_r, l_r\}) \left(\frac{z_{r,k_r} b_{r,k_r} + z_{r,l_r} b_{r,l_r}}{b_{r,k_r} + b_{r,l_r}} - z_{s,i_s} \right),$$

for $i_s = 1, \cdots, n_s$ and $s = 1, \cdots, S'$.

We next prove part (ii) by contradiction. Without loss of generalization, suppose there exists a pair $k_1, l_1 \in \{1, \cdots, n_r\}$ ($1 \in \{1, \cdots, S'\}$) with $\tilde{\phi}(Y_{1,k_1}) \succ \tilde{\phi}(Y_{1,l_1})$ and $z_{1,k_1} < z_{1,l_1}$. We construct another set $\{(\tilde{z}_{s,i_s})_{i_s=1}^{n_s};\ s = 1, \cdots, S'\}$ such that $\tilde{z}_{s,i_s} = z_{s,i_s} + \epsilon \frac{1}{b_{s,i_s}} I(s = 1, i_s = k_1) - \epsilon \frac{1}{b_{s,i_s}} I(s = 1, i = l_1)$, where $\epsilon > 0$ and is small enough such that $\tilde{z}_{1,k_1} \leq z_{1,l_1}$ and $\tilde{z}_{1,l_1} \geq z_{1,k_1}$. We use $U_S(z)$ and $U_S(\tilde{z})$ to denote the values of $U_S^{(h_1,\ldots,h_{S'})}$ with solution $\{(z_{s,i_s})_{i_s=1}^{n_s};\ s = 1, \cdots, S'\}$ and $\{(\tilde{z}_{s,i_s})_{i_s=1}^{n_s};\ s = 1, \cdots, S'\}$. Then

$$U_S(\tilde{z}) = U_S(z) + \epsilon \sum_{i_2=1}^{n_2} \cdots \sum_{i_S=1}^{n_S} w_{i_2}^{2,(h_2)} \cdots w_{i_S}^{S,(h_S)} \left(\phi(Y_{1,k_1}; \cdots; Y_{S,i_S}) - \phi(Y_{1,l_1}; \cdots; Y_{S,i_S}) \right).$$

As $\tilde{\phi}(Y_{1,k_1}) \succ \tilde{\phi}(Y_{1,l_1})$, $\epsilon > 0$ and all w's are positive, $U_S(\tilde{z}) > U_S(z)$, which contradicts with the statement that $\{(z_{s,i})_{i=1}^{n_s};\ s = 1, \cdots, S'\}$ maximizes $U_S^{(h_1,\cdots,h_{S'})}$. This completes the proof.

□