Title

Revealed Preferences Models for Reconstructing and Analysing Partnerships in Two-Sided Matching Market

Author

Goyal, Shuchi

Publication Date

2023

UNIVERSITY OF CALIFORNIA

Los Angeles

Revealed Preferences Models for Reconstructing and Analysing Partnerships in Two-Sided Matching Markets

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Statistics

by

Shuchi Goyal

2023

ABSTRACT OF THE DISSERTATION

Revealed Preferences Models for Reconstructing and Analysing Partnerships in Two-Sided Matching Markets

by

Shuchi Goyal
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2023
Professor Mark S. Handcock, Chair

Many social processes studied by demographers can be viewed as two-sided matching markets. For example, heterosexual marriages, job searching, and residency assignments for medical school graduates all require members of two disjoint groups to mutually consent to form a relationship, or “match.” Yet the underlying mechanisms dictating such processes are often opaque. Demographers require statistical models for partnership formation that separate the underlying preferences individuals have for various types of partners from the availability of such partners.

To address this need, in my dissertation I develop a revealed preferences model (RPM) which captures the complex interplay between discrete characteristics, both observed and unobserved, of individuals and the availabilities of potential partners to form a stable set of partnerships in networks of different sizes. The major contribution of this work is the introduction of a model that not only estimates partnership outcomes in a two-sided matching market, but also is flexible enough to handle realistic data drawn from various sampling schemes and population types.

Contextualizing the problem in the heterosexual marriage market setting, in the first two chapters I present background information on the two-sided matching market problem and and introduce the revealed preferences model novel statistical methodology to compute point estimates for preferences parameters. I validate the approaches with multiple simulation studies and demonstrate RPM's key novel contribution, the ability to recover societal preferences for partners independent of the types of partners available. I additionally use these simulation studies to conduct a comprehensive study of model performance under different conditions, such as varying population and sample sizes and different sampling schemes.

To facilitate the use of the model in practical settings, I propose additional novel tools such as bootstrap procedures for bias-correcting parameter estimation and empirical and analytical approaches for computing uncertainty intervals for the preference parameter estimates. I discuss the process of model selection and propose several methods for assessing the goodness-of-fit, including quantitative and visual procedures. To my knowledge this is the first time these procedures have been discussed in literature.

To aid continued development of models for two-sided matching markets, I present a review of the two major frameworks that have been hypothesized as underlying the partnership process. While developments in marriage modeling under these different frameworks have continued in parallel over the last several years, this is the first time that the assumptions and implications of the two frameworks have been compared clearly side-by-side using consistent notation. I bridge the gap in literature between the two settings by demonstrating how RPM can be adapted to model the marriage process in either scenario.

Throughout the dissertation, I continue to validate the proposed procedures through extensive simulation studies, and I show how the model can be applied and interpreted given survey data from the 2008 Survey on Income and Program Participation.

The dissertation of Shuchi Goyal is approved.

Jennie E. Brandt

Chad Hazlett

Megan M. Sweeney

Mark S. Handcock, Committee Chair

University of California, Los Angeles

2023

To Mummy and Papa

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This research was supported by National Science Foundation BIGDATA: Applications 915 program, grant NSF IIS-1546259, and from the Eunice Kennedy Shriver National Institute 916 of Child Health and Human Development, population research infrastructure grants P2C917 HD041041 and P2C-HD041022 and training grant T32-HD007545.

I am also grateful for the support of the UCLA Dissertation Year Fellowship.

I could write a whole other dissertation on all the people I would like to thank for their love and support throughout my time in graduate school. However, the time for writing dissertations has come to an end, and so I will attempt to restrain myself.

First and foremost, I would like to thank Mark Handcock, whose continued guidance in both my personal and professional development has been invaluable, and who seems to have an endless collection of anecdotes about the history of research both within and outside of the field of statistics. Your insights have enriched my own experiences and perspectives as a student. Thank you for providing such a wonderful environment for students to explore their interests in social statistics. I am so thankful for your mentorship, feedback, patience, and continued encouragement during my time in graduate school.

I would like to thank Jennie Brand, Chad Hazlett, and Megan Sweeney for serving on my committee and for their invaluable feedback on my work. I would also like to thank the administrative staff of the UCLA statistics department and, in particular, Chie Ryu and Enrique Reyes, for their commitment to students and for being ready to troubleshoot everything at a moment's notice.

I would like to thank Abhyuday Mandal, who showed great kindness and generosity with his time at a moment when I was struggling to find my direction. Thank you for encouraging me to go to graduate school and, along with Dr. Datta, for giving me my first research experience, which continues to pay dividends today. I am extremely lucky to have had your guidance as a professor and now as a friend.

I would like to thank Michael Rendall and Heide Jackson for welcoming me with such open arms into the world of demography, and for their collaboration and insights on this project over three years of weekly meetings. I have so appreciated working with and learning from you both and hope to continue to do so.

I would like to thank my graduate school cohort-mates and members of the Statistical Relations reading group for their friendship. I am aware that I am extremely lucky to have had such kind and supportive (not to mention extremely talented) classmates during my time at UCLA, and I could not have asked for a better group of people to be surrounded by during my time in graduate school. I look forward to many years of continued friendship.

I would like to thank all the teachers and professors who have played a part in my education between preschool to now.

I would like to thank Shubhi Didi for her love and support, and occasionally her well-reasoned reality checks. I so value the time I have been able to spend with you during the last few years while we have both been residents of California. Thank you for always being ready to talk to me. You are my best friend and my inspiration. Thank you for doing my hair and makeup before big events. I would also like to thank Rahul and Aashna for the laughs, the hugs, the encouragement, and the free medical advice. Your family brings so much joy into my life. I love you.

I would like to thank my parents, Deepa and Dharmender Goyal, for providing me with so much love, for encouraging my curiosity, and for making me the student and person I am. You have provided me with the freedom to live happily, pursue my passions, and be a little reckless, all of which I deeply appreciate.

Lastly, I would like to thank Bhumi and Vikram. Thank you both for coming into my life. You have made it infinitely better. Vikram, thank you for always reminding me to eat, and for reminding me to breathe, and for watching Spongebob Squarepants with me last night. I could not have done this without your love and support. Thank you, Chip.

VITA

2013–2018	B.S. Statistics, B.A. Economics, Magna Cum Laude, University of Georgia
2018-2019	Graduate Teaching Assistant
2019	Graduate Research Assistant
2020–2021	Trainee, California Center for Population Research
2021	Research Data Science Intern, Meta Inc.
2021–2023	Graduate Student Researcher, Civil Rights Project of UCLA
2022	Research Data Science Intern, Meta Inc.
2022–2023	Pathways Intern in Statistical Methodology, U.S. Department of State

PUBLICATIONS

Goyal, S.; Handcock, M.S.; Jackson, H.M.; Rendall, M.S. (2022), *A Practical Revealed Preference Model for Separating Preferences and Availability Effects in Marriage Formation.* Journal of the Royal Statistical Society Series A: Statistics in Society.

Rendall, M.S.; Jackson, H.M.; Goyal, S.; Handcock, M.S.; Weden, M.M.; Zvavitch, P. (2022), *Estimation, Simulation, and Validation of a Two-sex Model of Intergenerational Reproduction of Education*, Submitted.

Losen, D.J.; Goyal, S.; Alam, M.; Salazar, R. (2022), *Unmasking School Discipline Disparities in California*, Center for Civil Rights Remedies at the Civil Rights Project of UCLA.

Dhanjani, S.; Yang, H.H.; Goyal, S.; Zhang, K.; Gee, G.; Cowgill, B. (2021), *Trends in Healthcare Access Disparities Among Asian and Pacific Islander Health Fair Participants in Los Angeles,* 2011-2 019, Public Health Reports.

Goyal, S.; Datta, G. S.; Mandal, A. (2020), *A Hierarchical Bayes Unit-Level Small Area Estimation Model for Normal Mixture Populations*, Sankhya Series B.

Introduction

The objective of this dissertation is to develop a practical model which approximates the dynamics of a two-sided matching market. I consider a scenario in which we observe the partnership decisions by some or all agents in a population, as well as discrete covariates for the agents (and their partners, if any). I hypothesize that given such data, the proposed model "reveals," or allows inference of, preferences of the agents in the market when in the partnership formation process. We therefore refer to this general class of models as *revealed preference models.*

The matching markets considered in this dissertation share two important characteristics. First, they are "two-sided," meaning that network is bipartite, agents in the market must belong to one of two disjoint groups, and partnerships can only be formed between agents of different groups. Second, matchings must be one-to-one, meaning that an agent either chooses exactly one partner or remains single. To facilitate discussion throughout this dissertation, I will frame the development of model theory and simulation studies in the context of the heterosexual monogamous marriage market.

The remainder of this dissertation is organized as follows: Chapters 1 and 2 of this thesis are adapted from Goyal et al. (2023). Chapter 1 introduces the revealed preferences model (RPM). I show how, given a list of pairings and the characteristics of agents in those pairings, RPM can be used to estimate market preferences. In Chapter 3, I introduce methodology for model selection, proposing both quantitative and ad-hoc metrics for comparing goodness-of-fit. I also introduce a modified significance test and tools for visualizing model fit for RPM. In Chapter 4, I discuss the extension of the model to transferable utility frameworks and utilize goodness-of-fit methodology to directly compare the model fit under assumptions of non-transferable and non-transferable utility.

CHAPTER 1

Background

This chapter introduces the proposed revealed preferences model and has been adapted from Goyal et al. (2023).

Many social processes of pair formation can be viewed as two-sided matching problems. These scenarios are prevalent in demography, economics, sociology, political science and education, among other fields. For example, heterosexual marriages, job searching, and residency assignments for medical school gradfuates all require members of two disjoint groups to mutually consent to forming a relationship, or match. Yet the underlying mechanisms which dictate such processes are often opaque.

We consider not only how an actor chose a spouse of the opposing gender, but also the interactions between pairs of actors in a choice situation and the stability of the matching result. Actors from opposing sides have to choose each other voluntarily in order for a "match" to occur. Of particular interest to many researchers is the role individual and societal preferences play in the match-making process.

These preferences are difficult to discern for multiple reasons. First, it is challenging to collect data which records complete information about characteristics of observed pairings and the pool of options from which each individual made a selection. Second, the final observed matchings are as much a result of the availability of different types of individuals as they are of individual preferences.

As a simple example, consider a heterosexual marriage market within a two-sex population of size $N = 100$, as illustrated in Figure 1.1a. The nodes represent individuals in the market. The shape of a node represents the gender of the individual, either male or female, and the color of the node represents the individual's education level (high school diploma or no high school diploma). An edge connects two nodes if the individuals represented by those nodes are married, and edges can only exist between individuals of opposite genders. Nodes with no edges represent a single individual, and nodes cannot have more than one edge, a feature of the monogamous marriage market.

In this example, a researcher may observe that married women with no high school diploma tend to have spouses who also do not have a high school diploma. Theoretically, this phenomenon could be driven by two conflicting factors: 1) Women without a high school diploma *prefer* a partner without a high school diploma to partners with high school diplomas; or 2) women without a high school diploma prefer a partner with a high school diploma but are restricted in their options due to low *availability* of men with high school diplomas in the population.

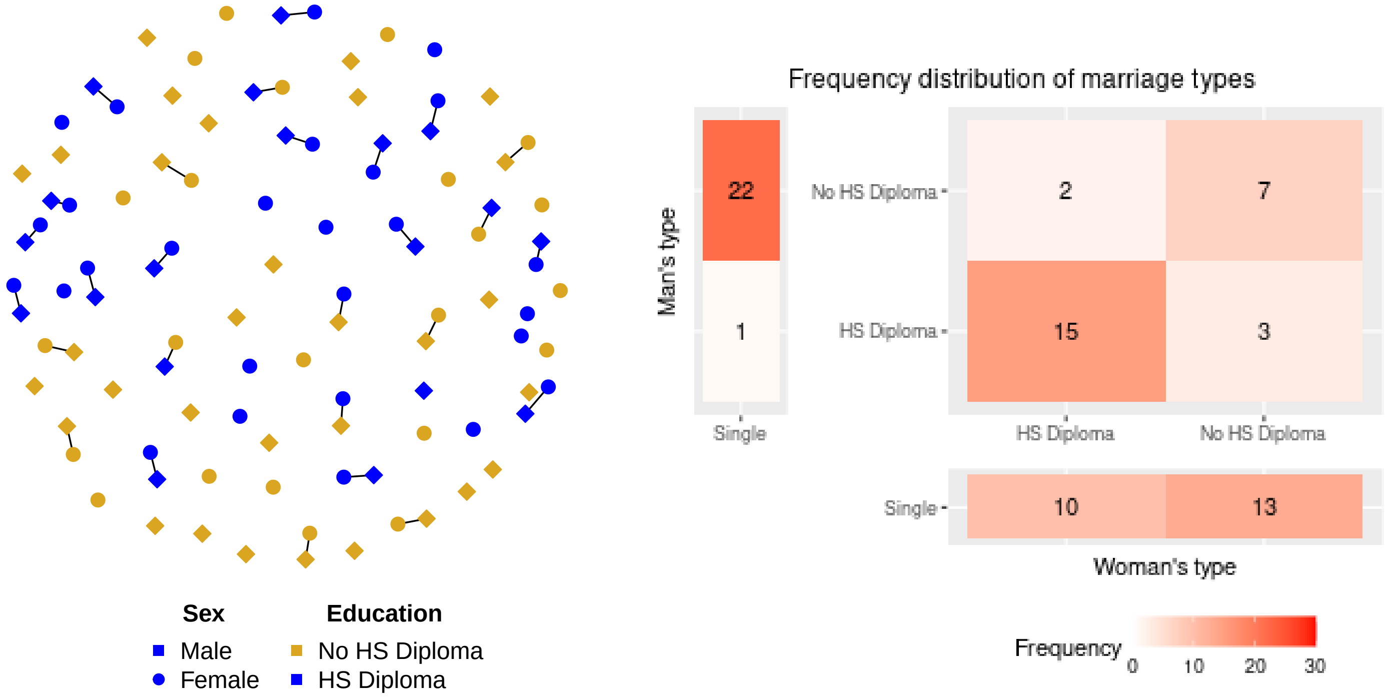
It is important to identify which of these scenarios is correct by distinguishing the effects of preferences and availability in the final realized matching, as the scenarios have differing implications. Suppose, for example, that a new education initiative allows several males in the market who did not previously have a high school diploma to obtain one. Assuming partnership preferences and education levels on the female side stay the same, then in the first scenario, the shift in male education attainment would lead to a decrease in partnership rates among females without a high school diploma. However, under the same conditions in scenario 2, the partnership rate among females without a high school education would increase.

Changing partnership rates can have implications on fertility rates, population growth, and other societal factors. In fact, a major motivation for this work in the dissertation is ongoing research on intergenerational transfers of inequality and poverty, e.g. Rendall et al. (2022). Prior research in this area has largely relied on one-sex models, in which child outcomes are modeled based on the characteristics of only one parent, usually the mother. Without considering the characteristics of the second parent, these models suffer from bias. Fertility rates depend on the characteristics of both parents and, clearly, people do not choose partners completely randomly. Advancements in partnership modeling within two-sided markets, and more specially modeling preferences for partners within the marriage market, are crucial for furthering research in this area.

Menzel (2015) proves a series of new mathematical results related to the asymptotic distribution of matching outcomes in a two-sided market. I develop Menzel's (2015) technical findings for application in demographic studies of two-sided matching processes. We propose a *revealed preferences model* which, given an observed set of stable matchings in a large population, uses a re-parameterised version of Menzel's (2015) equations to recover latent preference parameters in the population. These preference parameters are used to estimate the total utility of a given partnership, given the characteristics of the individuals in that partnership. To measure uncertainty of parameter estimates, we also propose both an analytical and an empirical approach to computing confidence intervals. We conduct simulation studies to show that for realistic populations, the revealed preferences model reconstructs preference parameters that are invariant under different population availabilities. We also show that the proposed confidence intervals achieve appropriate coverage.

The revealed preferences model can be generalized for applications where an individual is permitted to have multiple relationships, as in the case of an employer and its employees (Yeung, 2019). However, for the purposes of this dissertation I focus only on the case in which individuals have at most one partner, also known as one-to-one matching.

The remainder of this chapter provides a review of existing literature exploring the two-sided matching market. In Section 1.1 I provide background information on key concepts related to the revealed preferences problem. In Section 1.2 I review existing literature on attempts to solve the two-sided matching problem and the challenges of identifying individual preferences in such settings.



(a) Observed marriages for N=100

(b) $\bar{c}$ for the network on the left.

Figure 1.1: A synthetically generated example of partnership data from the heterosexual marriage network.

1.1 Concepts

This problem of separating the effects of preferences and availability in partnership formation has long been recognized in demography and has motivated an impressive body of literature without having been satisfactorily resolved (Choo and Siow, 2006; Dagsvik, 2000; Logan et al., 2008).

There are several relevant concepts that we borrow from other fields in the development of the revealed preferences model for a two-sided matching market. We discuss each of these in subsequent sections of this chapter.

1.1.1 Stable Matching

First, we consider the idea of *stable matchings.* In most social settings, relationships are constantly shifting over time. For example, marriages form and dissolve, employees join and leave firms, and students enroll in and drop out of schools. These complex movements are difficult to capture in any data set due to their continuous nature. To circumvent this problem in the context of marriages, we focus on newly-formed partnerships in a given sample at a discrete point and assume that this organization of one-to-one matches is *stable.*

The notion of *stable matchings* has been previously explored in depth by economists and statisticians. A matching is stable at a given timepoint when no two individuals who are not currently partnered with each other exist such that both individuals would prefer each other over their current partner. Furthermore, no person in a partnership would prefer to be single over their current partner. Gale and Shapley (See Roth and Sotomayor, 1990) showed that in large populations, there are various stable matchings that can be realized. By assuming matching stability, we are able to assume that the observed data accurately reflect individual and societal preferences at that time point.

1.1.2 Revealed Preferences Theory

The idea that we can infer preferences from a stable matching stems from the theory of *revealed preferences*, which has been studied extensively in economics and was notably developed in (Samuelson, 1938, 1948) in the context of consumer-goods markets. Samuelson (1938) hypothesizes that a consumer's preferences can be inferred based on their purchasing habits. An agent selects a good from a discrete set of options. Thus, the agent's choice reveals his preference for that option over all others. The inference of these preferences is achieved through the use of *discrete choice models*, which we discuss in 1.1.3.

The idea of revealed preferences becomes more complex when we shift from a consumer-goods market to a matching market. This is primarily because the final choices, or "pairings", achieved reflect the (possibly competing) preferences of *multiple* agents. A man may "select", or propose marriage to, the woman in the marriage market who maximizes his utility, but the woman also assesses her own utility gain from the partnership and can choose to reject the proposal. In this case, the marriage is not realized, and the man must revise his decision with a new, restricted option set which excludes the woman who rejected his proposal. In other words, the passive "choice" from the consumer-goods market has been replaced with an active agent who "chooses back."

A researcher interested in preferences for specific traits in spouses typically only observes the final list of achieved marriages, not the proposal-rejection process which precedes it. The researcher must distinguish whether a marriage occurred because it generated high utility for the individuals involved or because, despite generating low utility, it was simply the best option available to the individuals. Addressing this question requires complex analytical tools.

1.1.3 Discrete Choice Models

Closely related to the development of revealed preferences theory is the study of discrete choice models.

In general, discrete choice models statistically relate the choice decision to the decision maker's attributes and the attributes of the alternatives available. Game theorists and statisticians initially proposed discrete choice models to understand agent preferences in one-sided settings. In these scenarios, each individual has a set of discrete possible choices. Essentially, there is a "chooser" and a "chosen". The agent in the role of chooser is the sole decision maker of their outcome, although his decision may be affected by the decisions of other choosers around them. The one-sided discrete choice model estimates the utility the chooser would derive from every possible choice in his option set and assumes that agents make the utility-maximizing choice. The parameters of interest are the chooser's preferences.

However, the traditional one-sided discrete choice model is unsuitable for use in two-sided scenarios. First, as mentioned earlier, the option set of each agent is rarely observed completely. Second, the observed matchings in two-sided processes are no longer reflective of the preferences of a single individual, as both actors involved in the partnership must consent to the partnership. That is, rather than dividing the population into groups of "choosers" and "chosens," both individuals in the partnership are choosers of each other. Each member of the partnership aims to maximize his or her own utility, and preferences may not necessarily be reciprocated. For example, highly educated women may have a preference for highly educated men, but highly educated men may not have a preference for highly educated women.

One approach to studying two-sided matching scenarios is through the use of *two-sided discrete choice models*, so-called because individuals in the population have a set of discrete options with which they can match. The goal of two-sided matching models is to obtain the frequencies for the different types of partnerships that can occur, where the partnership type is defined by the combination of observable characteristics of the individuals in the partnership (Dagsvik et al., 2001).

1.2 Literature Review

Among others, Schoen (1981), Pollak (1986) and Pollard (1997) approached the two-sided matching problem with the goal of estimating the frequency distribution of match types. However, the methodologies they propose are limited in that they say little about the behavior of agents in the two-sided market. Thus, there is no apparent mechanism for detecting the underlying preferences which motivate the matchings.

In contrast, Logan et al. (2008), Dagsvik (2000) and Menzel (2015) all theorize two-sided versions of the discrete choice model which consider the role of both preference parameters and availability of partners in matching markets and propose methodology which can implicitly be used to estimate said preferences. Logan et al. (2008) propose a model for bipartite populations where each side has a distinct utility function for partnerships with agents on the opposing side. In the case of heterosexual marriages, all women have an identically-defined deterministic component to their utility which depends on the woman's own observed characteristics z and the characteristics of her partner x; similarly, all men have an identically-defined deterministic component to their utility which depends on the man's own observed characteristics z and the observed characteristics x of his partner. Here, $x \in \mathcal{X}$ and $z \in \mathcal{Z}$, where the sample spaces $\mathcal{X}$ and $\mathcal{Z}$ represent the set of observable types of women and men, respectively, and may be continuous or discrete. Unobserved characteristics are accounted for in the utility by including an individual fixed effect term for each actor. Logan et al. (2008) assume that an individual's unobserved option set within the local marriage market can be approximated by the observed sample distribution of characteristics.

Logan et al. (2008) show that their proposed method for small populations could theoretically be used to compute maximum-likelihood estimates (MLEs) of preference parameters. Rather than basing their inference on the true likelihood of the observed match being realized, they propose inferences based on the likelihood that the observed match is stable. For computation of these estimates, they propose Bayesian inference based on Markov chain Monte Carlo (MCMC).

The approach suggested by Logan et al. (2008) is limited in that the Bayesian inference works best for small populations. For example, the authors apply their method to make inferences about gender-based marital preferences using data from the National Survey of Families and Households (NSFH). With a sample containing 314 men and 360 women, they are able to compute parameter estimates for the two-sided model.

However, the method cannot be used with large sample data sets such as the Survey of Income and Program Participation (SIPP), where the number of people of each gender exceeds 16,000 or the American Community Survey (ACS), where the number of people of each gender exceeds 100,000. In such cases, the calculations required to update parameter estimates in each step of the MCMC process are extremely complex and often intractable. Additionally, when large populations with multiple stable matching solutions are studied, the posterior distribution of the parameters may have multiple maxima, thereby also rendering the parameters unidentifiable. Logan et al. (2008) also note limitations in parameter identifiability when certain parallel terms are included in the utility functions.

Dagsvik (2000) focuses on the identification and estimation of preference parameters in a closely related two-sided matching market model. He proposes constructing aggregate supply and demand functions based on preferences on both sides of the matching market. When the asymptotic supply and demand functions are equal, they derive equilibrium equations for the number of partnerships achieved between individuals of specific types. These equations imply that availability of partners and personal preferences are asymptotically separable in their relationship to the distribution of matching outcomes in a large population. This is a significant finding because, intuitively, the ability of people to achieve their preferred partnership outcome is constrained by the existence of partners. Dagsvik (2000) then shows that these equations can be manipulated to obtain point estimates of preference parameters. However, methodology for analytically computing standard errors for these estimates is not presented. In addition the results only apply to discrete agent types.

Nevertheless, the insights by Dagsvik (2000) lay important groundwork for the work done by Menzel (2015). Specifically, Menzel (2015) proves that the relationships suggested by Dagsvik (2000) hold true for large populations. Menzel (2015) derives equations which establish a relationship between the preference parameters and availabilities of men and women of each type in the population and the limiting distribution of types of matches across the possible outcomes. These calculations prove that in a large population, the interdependency between availability and preferences can be accurately modeled, and therefore that preferences can be recovered independently of the population availability context. Menzel (2015) then proposes that the relationship he develops can be used to construct a likelihood function for observing a particular matching. His results also apply to continuous agent characteristics.

We develop the results of Menzel (2015) to derive re-parametrized equations which allow asymptotically stable estimates of the proportions of single and partnered persons of each type in the population. We propose a subclass of two-sided discrete choice models which we refer to as *revealed preference models.* In this subclass of models we, like Logan et al. (2008), Dagsvik (2000) and Menzel (2015), focus on bipartite networks. Actors in the network are divided into two distinct groups. Edges, which represent partnerships, form only between members of opposing groups. Whereas Logan et al. (2008) assume that the full opportunity set of each actor is observed, we allow agents of different observed types to have different opportunity sets (Yeung, 2019). The goal of our study is to extend Menzel's (2015) findings to estimate a set of latent structural parameters that describes the decision-making behavior of a given population which led to the observed matching outcome. The difficulty of this problem is that the set of alternatives for each actor is not generally observed and determined endogenously in the market. Our proposed model utilizes key findings from Menzel (2015) about the limiting distribution of matches in a large population and applies them to estimate preference parameters based on an observed distribution of matches. We extend Menzel (2015) by developing a modification of his estimator that corrects for bias in small populations across a range of sample sizes and sample fractions.

Our study extends from the non-transferable utility assumption following Dagsvik (2000), Logan et al. (2008) and Menzel (2015). Variants of this model have been used to represent decision-making in a matching market that assumes transferable utility (TU) within partnerships, with recent studies, including Dupuy and Galichon (2014), Chiappori et al. (2017), and Galichon and Salanié (2021) building on the TU framework developed by Choo and Siow (2006). We note here only the basic commonalities and differences between the TU model of Choo and Siow (2006) and the NTU model of Menzel (2015).[1] The TU model is grounded in the economic theory of Becker's (1973, 1974) model of marriage. It requires the key assumption that the members of a couple engage in within-couple exchanges of utility-providing goods and services. Choo and Siow (2006) interpret these exchanges as determining "...each spouse's share of responsibilities within a marriage." The major statistical modeling implication is that in a TU model, the choosing individual only considers the prospective match's observable characteristics (Chiappori, 2020). In contrast, within the NTU framework there is no similar exchange of utility-providing goods and services, and the individual is influenced by the prospective match's observable (to the researcher) characteristics and the characteristics that are to the researcher unobservable. In the NTU case, increased availability leads to increased propensity to find a match.

[1] A more detailed discussion comparing the NTU and TU frameworks is found in Chapter 4.

CHAPTER 2

Revealed Preferences Model

To facilitate our discussion of the revealed preferences model, we will discuss the problem within the context of heterosexual marriages within a two-sex population unless otherwise noted. In this set-up, we consider a population with two distinct groups, and individuals are either male or female. At any given point in time, individuals have at most one partner of the opposite sex, and they also have the outside option to remain unpartnered (single). Both the male and the female must agree to the partnership for that partnership, or "marriage," to be observed.

Individuals evaluate their marital options through use of a utility function, which consists of a deterministic and random component. Actors of the same gender are assumed to have deterministic components to their utility functions that depend on their own observed characteristics x and those of their potential partners, z. The random component of the utility function accounts for the fact that agents' characteristics are only partially observed. Agents choose the partner from available options who will maximize their own total utility. The latent parameters in the deterministic component of the utility function which govern this pair formation are commonly known as preference parameters in the sense that they represent how actors would choose among different alternatives if given a choice (Logan, 1996a; Logan et al., 2008).

We also consider the heterosexual monogamous marriage market in our empirical example of application, which relies on data from the 2008 Survey of Income and Program Participation. (U.S. Bureau of the Census, 2020)

2.1 Notation

We consider a population with N_w women and N_m men, so that the total population size is $N = N_w + N_m$. N_h represents the number of households in the population, where a household is defined as an entity consisting of either a single (unpartnered) man or woman or a partnered couple, so that $N_h \leq N$, and $N_h = N$ only when all individuals choose to remain single. A household is either "single" if it contains of a single, unpartnered person or "partnered" if it contains two individuals in an exclusive partnership.

A household holds either exactly one single person of any gender or one married couple, and a household is characterized by the type(s) of the individual(s) in it.[1] Each single household is further differentiated by the gender and type of the individual living in it. Each partnered household is further differentiated by the combination of the type of female and the type of male who live in the household.

The differentiation between N and N_h is important when considering the sampling scheme, as N_h is equal to the total number of sampled *units*, while N is the number of individuals considered *across* the N_h sampled units. We expand on this point in Section 2.4.

Using the same notation introduced in Section 1.2, we observe a $p-$vector of observed covariates $x \in \mathcal{X}^p$ on the women and a $q-$vector of observed covariates $z \in \mathcal{Z}^q$ on the men. For ease, we will generally omit the p and q superscripts when referring to the covariate spaces $\mathcal{X}$ and $\mathcal{Z}$ hereafter. $\mathcal{X}$ and $\mathcal{Z}$ may overlap but are not required to.

Let x_i and z_j denote the observed attributes of woman $i = 1, \ldots, N_w$ and man $j = 1, \ldots, N_m$, respectively. The equations in this section are written generally so that the elements of x and z may be continuous, discrete, or a combination of the two. For ease of presentation, however, in the simulation studies in Section 2.9 where we apply the revealed preferences model, we assume that x and z are discrete.

Actors may perceive potential partners differently based on their own characteristics.

[1] This definition of household is different from the one often utilized in demography work, where households can consist of any combination of unpartnered and partnered individuals, as well as their offspring.

Thus, the perceived utility gained by partnering with a particular opposite-sex individual may differ from one decision-maker to the next. However, all actors are assumed to choose the partner within their respective choice sets that maximizes utility. Given the utility-maximizing behavior of the decision-makers, we define the utility gained by woman i with observed attributes x_i from partnering with man j with observed attributes z_j as

$$U_{ij} = \underbrace{U(x_i, z_j|\underset{\sim}{\theta}^W)}_{\substack{\text{deterministic}\\ \text{component}}} \quad + \quad \underbrace{\eta_{ij}}_{\substack{\text{unobserved random}\\ \text{component}}} \tag{2.1}$$

where $\underset{\sim}{\theta}^W$ is the set of parameters denoting the woman's preferences. The deterministic part of the utility functions depend on variables representing the respective types of women and men. Similarly, we define the utility gained by man j with observed attributes z_j from a partnership with woman i with observed attributes x_i as

$$V_{ij} = \underbrace{V(z_j, x_i|\underset{\sim}{\theta}^M)}_{\substack{\text{deterministic}\\ \text{component}}} \quad + \quad \underbrace{\zeta_{ij}}_{\substack{\text{unobserved random}\\ \text{component}}} \tag{2.2}$$

where $\underset{\sim}{\theta}^M$ is the set of parameters representing men's preferences. From this point forth in the paper, we will use tilde below a Greek letter to refer to a vector.

Following Menzel (2015), we assume that unobserved random components of the utility functions as defined in Equations (2.1) and (2.2) are independently and identically distributed draws from a distribution in the domain of attraction of the extreme-value type-I (Gumbel) distribution. This domain includes Exponential, Gamma, Gaussian, Lognormal, and Weibull distributions. Here we will focus on the Gumbel itself, but note our model and methods are generalizable.

2.2 Model specifications

Having introduced the general setup of a two-sided discrete choice model, we now go into detail about model forms for the deterministic and random utility components. We focus on the special case where the deterministic components of the utilities in (2.1) and (2.2) are additive linear functions; however, other choices of utility functions can also be used (See Dagsvik (1994) for inference of latent preferences under other choices of utility functions).

For additive linear utility functions, let

$$\begin{aligned} U(x_i, z_j|\underset{\sim}{\theta}^W) &= \theta_{w0} + \sum_{k=1}^{K_w} \theta_{wk} X^k(x_i, z_j) \\ V(z_j, x_i|\underset{\sim}{\theta}^M) &= \theta_{m0} + \sum_{k=1}^{K_m} \theta_{mk} Z^k(x_i, z_j) \end{aligned} \tag{2.3}$$

where x_i and z_j are vectors measuring observed characteristics of woman i and man j, respectively. The woman's deterministic utility consists of an intercept term θ_{w0} and K_w functions $X^k(x_i, z_j)$ which represent utility that woman i derives from the partnership based on her perception of her own characteristics and the characteristics of man j. For example, $X^k(x_i, z_j)$ might be an indicator function that represents whether certain observed attributes are identical for the pair (i.e., the partnership is homogamous). The corresponding K_m functions for the man's side are denoted as $Z^k(x_i, z_j)$. Here $\underset{\sim}{\theta}^W = [\theta_{w0}, \theta_{w1}, \ldots \theta_{wK_w}]^T$ and $\underset{\sim}{\theta}^M = [\theta_{m0}, \theta_{m1}, \ldots \theta_{mK_m}]^T$ are the preference parameters.

The random component of the utility model accounts for unobserved information about individuals in the data which may impact partnership choices. The random terms, are assumed to be identically distributed draws from an extreme-value type-I (Gumbel) distribution.

We additionally define the random utility for the choice of remaining single as Menzel (2015) did, so that

$$\begin{aligned} U_{i0} &= 0 + \max_{k=1,\ldots,N_m^\delta} \{\eta_{i0,k}\} \\ V_{j0} &= 0 + \max_{k=1,\ldots,N_w^\delta} \{\zeta_{j0,k}\} \end{aligned} \tag{2.4}$$

for females and males, respectively.

The single household utility specification in Equation (2.4) implies that the deterministic component of the utility for an individual choosing to be unpartnered is 0. The non-deterministic component of the single utility function of females is defined as the maximum of N_m^δ independent draws of $\eta_{i,k}$, the Gumbel-domain-of-attraction distributed random term of the male partnered utility function presented in Equation (2.1). Similarly, the non-deterministic component of the single utility function for males is the maximum of N_w^δ independent draws of $\zeta_{j,k}$ from Equation (2.2). Am interpretation for this formulation is that in a market of N_m men, woman i also considers N_m^δ outside latent non-market alternatives (and vice-versa for men).

We focus on the case where $\eta_{i,k}$ and $\zeta_{i,k}$ are i.i.d. Gumbel. Since the maximum of N_m^δ i.i.d. Gumbel random variables is also Gumbel-distributed with the location parameter increased by $\delta \log N_m$, the hyperparameter δ effectively sets the expected utility for an individual choosing to be unpartnered. We choose δ based on prior expectations of how the proportion individuals in the population who are single will change for different market sizes. For this model, we set $\delta = 1/2$. This specification ensures that the share of singles in the market is stable for different market sizes (Menzel, 2015, Assumption 2.2). Intuitively, increasing the value of δ will make the choice of remaining single more attractive in large populations, while decreasing the value of δ makes the single option less attractive.

2.2.1 Large population approximation

Let $w(x)$ be the number of women in the population with characteristics x and $m(z)$ be the number of men in the population with characteristics z. For notational convenience, let $\bar{w}(x) = w(x)/N$ and $\bar{m}(x) = m(x)/N$.

Consider a population with utilities drawn from the model (2.1), (2.2), (2.3) and (2.4). Then the stable matching induces a probability distribution over the observed characteristics. Consider sampling a random household from the data. Let $f(x, *)$ and $f(*, z)$ be the probability that a sampled household consists of an unmatched woman of type x, and an unmatched man of type z, respectively. Let $f(x, z)$ be the probability a sampled household consists of a woman of type x and a man of type z who are married to each other. Finally let $\bar{f} = \{f(x, z), f(x, *), f(*, z)\}, x \in \mathcal{X}, z \in \mathcal{Z}$. Together, $\bar{f}$ defines a distribution satisfying the overall normalization constraint:

$$\int f(x,z)dxdz + \int f(x,*)dx + \int f(*,z)dz = 1 \tag{2.5}$$

More specifically,

$$\begin{aligned} \bar{w}(x) &= f(x,*) + f(x,\diamond) \\ \bar{m}(z) &= f(*,z) + f(\diamond,z) \end{aligned} \tag{2.6}$$

where $f(x,\diamond)$ is the probability the person is a matched woman of type x:

$$\begin{aligned} f(x,\diamond) &= \int f(x,z)dz \\ f(\diamond,z) &= \int f(x,z)dx \end{aligned}$$

A major result of Menzel (2015) is that, under mild regularity conditions, if the population size is large and the matching is stable, the frequencies approximately satisfy the relations:

$$f(x,z) = e^{W(x,z|\underset{\sim}{\beta})} f(x,*) f(*,z) \qquad \forall x, z. \tag{2.7}$$

$$W(x,z|\underset{\sim}{\beta}) = U(x,z|\underset{\sim}{\theta}^W(\underset{\sim}{\beta})) + V(z,x|\underset{\sim}{\theta}^M(\underset{\sim}{\beta})), \;\; \forall x \in \mathcal{X}, z \in \mathcal{Z}$$

is the sum of the deterministic components of the utilities and $\underset{\sim}{\theta}^W(\underset{\sim}{\beta})$ and $\underset{\sim}{\theta}^M(\underset{\sim}{\beta})$ are functions such that $\underset{\sim}{\beta}$ parameterizes $W(x,z|\cdot)$. The solution must satisfy the population equilibrium conditions on the parameter values, $\underset{\sim}{\beta}$:

$$\begin{aligned} \frac{f(x,\diamond)}{f(x,*)} &= \int e^{W(x,s|\underset{\sim}{\beta})} f(*,s)ds \quad \forall\ x \\ \frac{f(\diamond,z)}{f(*,z)} &= \int e^{W(s,z|\underset{\sim}{\beta})} f(s,*)ds \quad \forall\ z \end{aligned} \tag{2.8}$$

The typical number of stable matchings possible increases exponentially with the population size. However, all of these stable matchings have the same limiting probability distribution $(\bar{f})$ over the observed characteristics.

Together, (2.6) and (2.7) make it possible to obtain estimates $\hat{\underset{\sim}{\beta}}$ of the preference parameters.

2.3 Parametrisation and Identifiability

We say that a parametrisation of the model, $\beta \in B$, is large population identifiable if for each $\beta_1, \beta_2 \in B$ with $\beta_1 \neq \beta_2$ there exists a state of the covariates x and z such that

$$P(\bar{c}|\beta_1) \neq P(\bar{c}|\beta_2)$$

Based on equations (2.7) and (2.8), and the expression

$$W(x, z|\underset{\sim}{\beta}) = U(x, z|\underset{\sim}{\theta}^W(\underset{\sim}{\beta})) + V(z, x|\underset{\sim}{\theta}^M(\underset{\sim}{\beta})), \quad \forall x \in \mathcal{X}, z \in \mathcal{Z}$$

only the sum of the partnered individuals' utilities is identifiable. $U(x, z|\underset{\sim}{\theta}^W)$ and $V(z, x|\underset{\sim}{\theta}^M)$ may not be separably identifiable when they are additive linear functions as in Equation (2.3) and include parallel terms. In general, let $\underset{\sim}{\theta}^W(\underset{\sim}{\beta})$ and $\underset{\sim}{\theta}^M(\underset{\sim}{\beta})$ be functions such that

$$W(x, z|\underset{\sim}{\beta}) = U(x, z|\underset{\sim}{\theta}^W(\underset{\sim}{\beta})) + V(z, x|\underset{\sim}{\theta}^M(\underset{\sim}{\beta})), \quad \forall x \in \mathcal{X}, z \in \mathcal{Z}$$

In this case, $W(x, z)$ can be parameterized in terms of $\underset{\sim}{\beta}$. We will consider parametrisations where $\underset{\sim}{\beta}$ is identifiable. To emphasize the relationship between $\underset{\sim}{\beta}, \underset{\sim}{\theta}^W$, and $\underset{\sim}{\theta}^M$, we refer to the gender-specific preference parameters as $\underset{\sim}{\theta}^W(\underset{\sim}{\beta})$ and $\underset{\sim}{\theta}^M(\underset{\sim}{\beta})$ for the rest of this paper.

2.3.1 Reparametrisation of the model

We can reparametrise these expressions to improve interpretability and ease computation. Define parameters $g(x, *)$ and $g(*, z)$ via the equations:

$$\begin{aligned} f(x, *) &= \frac{\bar{w}(x)e^{g(x,*)}}{(1 + e^{g(x,*)})} \\ f(*, z) &= \frac{\bar{m}(z)e^{g(*,z)}}{(1 + e^{g(*,z)})} \end{aligned} \tag{2.9}$$

so that $g(x, *)$ and $g(*, z)$ both have range the real line.

We interpret $g(x, *)$ as the log-odds that a women with characteristics x is single. Similarly, we interpret $g(*, z)$ as the log-odds that a men with characteristics z is single. Hence this reparametrisation is essentially from probabilities to logits. We will use $g(x,*)$ and $g(*,z)$ in place of $f(x,*)$ and $f(*,z)$ to ease computation and interpretability. Note that

$$\begin{aligned} f(x,\diamond) &= \frac{\bar{w}(x)}{(1+e^{g(x,*)})} \\ f(\diamond,z) &= \frac{\bar{m}(z)}{(1+e^{g(*,z)})} \end{aligned}$$

so that (2.6) is automatically satisfied and (2.7) becomes

$$f(x,z) = \text{pref}(x,z)\bar{w}(x)\bar{m}(z) \qquad \forall x,z \tag{7'}$$

where

$$\text{pref}(x,z) = \frac{e^{W(x,z)+g(x,*)+g(*,z)}}{[1+e^{g(*,z)}][1+e^{g(x,*)}]} \qquad \forall x,z$$

Equation (7') explicitly separates the availability component of the model $(\bar{w}(x)\bar{m}(z))$ from the preferences-related component $(\text{pref}(x,z))$. In this parametrisation (2.8) becomes

$$\begin{aligned} e^{-g(x,*)} &= \int \frac{e^{W(x,s)+g(*,s)}\bar{m}(s)}{1+e^{g(*,s)}} ds \quad \forall\ x \\ e^{-g(*,z)} &= \int \frac{e^{W(s,z)+g(s,*)}\bar{w}(s)}{1+e^{g(s,*)}} ds \quad \forall\ z \end{aligned} \tag{8'}$$

2.4 Data

The analysis depends on the sampling design that produces the data. Let $c(x,*)$ and $c(*,z)$ be the counts of households with unmatched women of type x, and unmatched men of type z, respectively. Let $c(x,z)$ be the sample counts of households which consist of a partnered woman and man, where the woman has observed characteristics x and the man has observed characteristics z. Finally, let $\bar{c} = \{c(x,z), c(x,*), c(*,z)\}, x \in \mathcal{X}, z \in \mathcal{Z}$. Together, $\bar{c}$ defines the empirical version of the distribution $\bar{f}$. Note that

$$\sum_{x\in\mathcal{X}}\sum_{z\in\mathcal{Z}} c(x,z) + \sum_{x\in\mathcal{X}} c(x,*) + \sum_{z\in\mathcal{Z}} c(*,z) = N_h.$$

As an example, we refer again to the simulated population with size $N = 100$ shown in Figure 1.1a. Figure 1.1b visualizes $\bar{c}$ for this population. The columns in the inner grid correspond to different types of women, and the rows represent type of their partner. The number of women of each type who remain single is represented by on the bottom margin to correspond with the columns. Similarly, the rows of the inner grid correspond to types of men, and now the columns represent their partner's type. The frequency of single men of each type is provided on the left margin in alignment with the rows.

Our method can be applied to a broad range of complex survey sampling designs, with the requirement that they produce estimates of $\bar{f}$. The interpretation of c differs slightly based on the sampling design, although there is no difference in the practical implications. In the case of census data, $\bar{c}$ represents the frequency distribution over the total population of households N_h. Similarly, in a stock-stock sampling scheme, the sampling units are households, so $\bar{c}$ is the frequency distribution over all sampled households. By contrast, in a stock-flow sampling scheme, the sampling units are individuals and spouses of partnered sampled individuals are considered "after the fact." In this case, $\bar{c}$ is the frequency distribution over all sampled individuals. However, the sum over the distribution is still effectively the number of households considered based on the sample. By defining $\bar{c}$ as the frequency distribution over sampled *units* rather than sampled *individuals*, we ensure $c(x, z) \in \mathbb{Z}^+$ for all $x \in \mathcal{X}, z \in \mathcal{Z}$.

In this section we focus on the situation where the data are a probability sample of the individuals in a population where the weights are w_i^w for the i^{th} woman and w_j^m for the j^{th} man. It is presumed that the weights are normalized via post-stratification to sum to population quantities over the covariates in the model. It is also presumed that the characteristics of the partner, if any, of sampled individuals are available. We take a super population framework, where the population is sampled from a super population process. Specifically, the N members of the population are independent and identical draws from a super population stochastic process. The sample of women is denoted $\{x_i, z_i, w_i^w\}_{i=1}^{n_w}$, where z_i are the characteristics of the women's partner, if any. If the sampled women is single, formally set z_i to $*$. Similarly, the sample of men is $\{z_j, x_j, w_j^m\}_{j=1}^{n_m}$.

Estimates of $w(x)$ and $m(z)$ may be available from auxiliary surveys. Otherwise, we can use the data alone and standard design-based estimates of $w(x)$ and $m(z)$, written as $\tilde{w}(x)$ and $\tilde{m}(z)$, respectively. Note that these represent *availabilities* and do not depend on the preference parameters. The parameters are then $\underset{\sim}{\psi} = (\underset{\sim}{\beta}, \{g(x,*)\}_{x\in\mathcal{X}}, \{g(*,z)\}_{z\in\mathcal{Z}})$.

2.5 Large-population Likelihood Approach

Had we observed the entire population, the likelihood for $\underset{\sim}{\psi}$ would involve the complex dependencies between the individual choices and matchings in the population. Each of the matchings is interdependent. Our approach is to use as a surrogate for the likelihood for $\underset{\sim}{\psi}$, one based on the likelihood of the observed frequencies of pairings by covariates, $\bar{c}$, and the model (2.7) and (2.8). Specifically, we approximate the exact likelihood for $\underset{\sim}{\psi}$ by:

$$\begin{aligned} &\text{lp-log-lik}(\underset{\sim}{\beta}, g(x,*), g(*,z) | \{x_i, z_i, w_i^w\}_{i=1}^{n_w}, \{z_j, x_i, w_j^m\}_{j=1}^{n_m}) \qquad (2.10)\\ =&\sum_{x\in\mathcal{X}}\sum_{z\in\mathcal{Z}} c(x,z)\log f(x,z) + \sum_{x\in\mathcal{X}} c(x,*)\log f(x,*) + \sum_{z\in\mathcal{Z}} c(*,z)\log f(*,z)\end{aligned}$$

The log-likelihood (2.10) can be written in terms of $g(x,*)$ and $g(*,z)$ using (7').The values $\tilde{w}(x)$ and $\tilde{m}(z)$ replace $w(x)$ and $m(z)$ in these expressions.

To obtain estimates, (2.10) can be maximized subject to the constraints expressed in (8') to produce the maximum large-population likelihood estimator (MLPLE), $\hat{\underset{\sim}{\psi}}$. This was achieved via a sequential quadratic programming (SQP) algorithm for non-linearly constrained gradient-based optimization (Kraft, 1994; Johnson, 2020). The algorithm optimizes successive second-order (quadratic/least-squares) approximations of the objective function (via BFGS updates), with first-order (affine) approximations of the constraints. We note that there are many possible survey sampling schemes in use, and the sampling could be at the individual level or at the household level. These alternative survey designs are straightforward to incorporate into the above equations and we do not explicate it here.

2.6 Correcting the estimator for bias and confidence coverage

It is likely that the estimator of Section 2.5 will be biased because it is based on a large-population approximation to the generating process followed by a number of statistical approximations. As noted in Section 4.1, we take a super population framework, with N specifying the size of the draw from the super population to the population and $n \leq N$ the size of the subsequent draw of the sample from the population. There is added uncertainty associated with both steps (specifically, the large population approximation at the first step and the sampling error at the second step).

The large population approximation does not take into account information in the matching that is not captured by the counts of matches and singles by type. In addition, the super population sampling distribution of these counts is not truly multinomial; While the utilities in equations (2.1) and (2.2) are independent, the matches are interdependent and hence so are the counts. However, the counts are asymptotically (with N) sufficient for parameter estimation (Menzel, 2015), and the bias should be small for large population sizes.

To address potential bias, we propose using bootstrap procedures to estimate the sampling distribution of the estimator and correct for bias and confidence coverage. We propose two versions of this bootstrap: a parametric version that is preferred where computationally feasible and a classical version to be used for large population sizes.

2.6.1 Parametric bootstrap

If the population size is small (e.g., less than 20,000), we can generate the (stochastic) relational utilities for all population members using equations (2.1), (2.2), and (2.4) at the MLPLE parameter values. We can then use the Gale-Shapley algorithm to achieve a stable matching for that population. This matching is from the population-generating process of the data. We follow it with a sampling of size n using the sampling design of the data including survey weights (e.g., stock-stock, stock-flow, census). We repeat this process b times, so that we have b bootstrapped samples. We fit the revealed preferences model to each of the b samples and obtain the bootstrapped parameter estimates for a single parameter $\utilde{\psi}$, which we denote as $\utilde{\psi}^* = [\utilde{\psi}^*_{(1)}, \utilde{\psi}^*_{(2)}, \ldots, \utilde{\psi}^*_{(b)}]$. Doing so requires us to re-solve a constrained maximization problem for each bootstrap sample. This can be computationally expensive, but is simply parallelizable (as we have done in the software associated with this paper (Handcock et al., 2022)).

The empirically estimated bias of $\hat{\utilde{\psi}}$, denoted as $\widehat{\text{bias}}_{\hat{\utilde{\psi}}}$, is equal to the mean of the bootstrapped parameter estimates $\utilde{\psi}^*$ minus $\hat{\utilde{\psi}}$. We then propose as our bias-corrected point estimator $\hat{\utilde{\psi}}_{BC} = 2\hat{\utilde{\psi}} - \frac{1}{b}\sum_{i=1}^{b} \utilde{\psi}^*_{(i)}$.

As we are drawing directly from the super-population generating and sampling processes, we believe this will provide a firm basis for bias-reduction and coverage correction for the census case.

2.6.2 Large-population bootstrap

The computational burden of the Gale-Shapley algorithm is heavy for large populations (e.g., $N > 20,000$). In this case, we consider a classical bootstrap for survey data, simple random resampling b data sets from the original data with replacement so that we have b sets of bootstrapped samples (Shao and Tu, 1995). As before, we fit the revealed preferences model to each of the b samples and obtain the bootstrapped parameter estimates for a single parameter $\utilde{\psi}$, which we denote as $\utilde{\psi}^* = [\utilde{\psi}^*_{(1)}, \utilde{\psi}^*_{(2)}, \ldots, \utilde{\psi}^*_{(b)}]$ and propose a bias-corrected point estimator appropriate for survey data $\hat{\utilde{\psi}}_{BC} = \hat{\utilde{\psi}} - \frac{N-n}{n-1}(\frac{1}{b}\sum_{i=1}^{b} \utilde{\psi}^*_{(i)} - \hat{\utilde{\psi}})$ (McCarthy and Snowden, 1985).

In this scenario, $N \gg n$ so that sampling uncertainty dominates errors from the large-population approximation. We then appeal to survey sampling bootstrap asymptotics as justification (Shao and Tu, 1995, Theorem 6.5).

This procedure and its parametric complement appear to work well, as is borne out in the simulation studies of Sections 2.8 and 6.

2.7 Measuring uncertainty of the estimates

Once we obtain the parameter estimates $\hat{\underset{\sim}{\psi}}$, a natural next step is to measure their uncertainty.

The covariance matrix of the estimates can be approximated by a standard Central Limit Theorem argument. The approximate log-likelihood function, augmented by the constraints, is

$$\begin{aligned} &\text{log-lik}_A(\underset{\sim}{\psi}|\{x_i, z_i, w_{wi}\}_{i=1}^{n_w}, \{z_j, x_i, w_{mj}\}_{j=1}^{n_m}) && (2.11) \\ &= \text{lp-log-lik}(\underset{\sim}{\psi}|\{x_i, z_i, w_{wi}\}_{i=1}^{n_w}, \{z_j, x_i, w_{mj}\}_{j=1}^{n_m}) + \sum_{k=1}^{|\mathcal{X}|+|\mathcal{Z}|} \lambda_k h_k(\underset{\sim}{\psi}) && (2.12) \end{aligned}$$

where $\{h_k(\underset{\sim}{\psi})\}_{k=1}^{|\mathcal{X}|+|\mathcal{Z}|}$ are the constraints (7'). Its Hessian is

$$\mathbb{E}\left(\frac{\partial^2 \text{log-lik}_A}{\partial \underset{\sim}{\psi} \partial \underset{\sim}{\psi}'}\right) = \begin{pmatrix} H & J \\ J^T & 0 \end{pmatrix} \tag{2.13}$$

where H is the Hessian of (2.10) with ij^{th} element $\mathbb{E}\left(\frac{\partial^2 \text{lp-log-lik}}{\partial \underset{\sim}{\psi} \partial \underset{\sim}{\psi}'}\right)$ and J is the Jacobian matrix of the constraints with kj^{th} element $\frac{\partial h_k(\underset{\sim}{\psi})}{\partial \underset{\sim}{\psi}}$. The estimate of the (asymptotic) covariance matrix of the MLPLE of $\underset{\sim}{\psi}$ is the (1,1) block of the Moore-Penrose inverse of this matrix (Hartmann and Hartwig, 1996).

The accuracy of the estimate of the covariance matrix depends on the application-specific accuracy of the various approximations. Thus, the analytically estimated standard errors may not accurately reflect the standard errors of parameter estimates that are observed over repeated samples from the same population. However, they are easy and fast to compute. It is natural to consider robust (sandwich formula) variance estimators for this situation. However these performed poorly as they did not adequately take into account the constraints.

As an alternative, we propose estimating standard errors empirically using the bootstrap procedures of Section 2.6. Most directly, the empirically estimated standard error of $\hat{\underset{\sim}{\psi}}$, denoted as $\widehat{\text{se}}_{\hat{\underset{\sim}{\psi}}}$, is equal to standard error of the bootstrapped parameter estimates $\underset{\sim}{\psi}^*$.

We also consider various methods employing bootstrap procedures to compute confidence intervals for each parameter. The *percentile bootstrap*, is the most straightforward of these methods. We denote $\underset{\sim}{\psi}^*_{(\alpha)}$ as the α percentile of the bootstrap parameter estimates $\underset{\sim}{\psi}^*$. The $(1-\alpha)\%$ percentile bootstrap confidence interval for parameter $\underset{\sim}{\psi}$:

$$(\underset{\sim}{\psi}^*_{(\alpha/2)}, \underset{\sim}{\psi}^*_{(1-\alpha/2)}).$$

The second method we employ is the basic bootstrap confidence interval. For the parameter $\underset{\sim}{\psi}$ with estimate $\hat{\underset{\sim}{\psi}}$, we use the basic bootstrap procedure to obtain a $(1-\alpha)$ confidence interval:

$$(2\hat{\underset{\sim}{\psi}} - \underset{\sim}{\psi}^*_{(1-\alpha/2)}, 2\hat{\underset{\sim}{\psi}} + \underset{\sim}{\psi}^*_{(\alpha/2)}).$$

We also consider a modified version of the studentized t bootstrap confidence interval. Here we obtain a $(1-\alpha)\%$ confidence interval as:

$$(\hat{\underset{\sim}{\psi}} - t^*_{(1-\alpha/2)}\widehat{\mathrm{se}}_{\hat{\underset{\sim}{\psi}}}, \hat{\underset{\sim}{\psi}} + t^*_{(\alpha/2)}\widehat{\mathrm{se}}_{\hat{\underset{\sim}{\psi}}}).$$

We test the performances of the analytical confidence intervals as well as those of all three proposed bootstrap confidence interval methods in Section 2.9.5 as part of our simulation studies.

2.8 Simulation Studies of Model and Inferential Accuracy

In this section we illustrate the statistical properties of the revealed preferences model by conducting three simulation studies which we refer to hereafter as studies I, II, and III. In simulation study I we show that the revealed preferences model accurately estimates underlying preference parameters which partially motivate matching outcomes in a population under different availability scenarios. In simulation study II, we investigate the relationship between the population size N and bias of preference parameter estimates produced by the revealed preferences model when census data is available. In simulation study III, we investigate the relationship between the relative sample proportion n_h/N and bias of preference parameter estimates when data is available for a sample of a population. In all three studies, we show the bias-corrected maximum large-population likelihood estimates (MLPLEs) for the preference parameters, adjusted using the methodology proposed in Section 2.6. In addition, in studies II and III we also show the MLPLEs prior to bias correction and compare them to the bias-corrected MLPLEs, demonstrating that the bias-corrected MLPLEs consistently improve estimate accuracy with little cost to precision.

Together, the simulation studies shown in this paper make a significant contribution to existing literature as they clearly demonstrate the novel ability of our proposed revealed preferences methodology to separate effects of preference and availability on matching outcomes. Previously, Menzel (2015) presented a simulation study with maximum-likelihood estimation of preference parameters. However, his results were extremely limited in that he considers populations that are restricted to size $N \leq 2,000$ and are generated under a single availability scenario. In contrast, we will show that the revealed preferences model recovers preferences for given sample or census data for a wide range of population (sample) sizes and under different availability scenarios. We also demonstrate the use of bias-correction procedures to improve the accuracy of our estimates. For researchers in other fields who will apply our model, we also consider several different specifications for the systematic component of the utility function to demonstrate the flexibility of our proposed approach.

The remainder of this section is structured as follows: we first describe a general procedure for the three simulation studies. We then describe the two availability scenarios considered for generating individuals of different genders and education in each simulated population in Section 2.8.1. In Section 2.8.2 we discuss the choice of $\underset{\sim}{\beta}^0$ and the different utility model specifications considered for the function $W(x_i, z_j|\underset{\sim}{\beta})$. Once we have defined the availability scenarios and utility model specification, we then provide further detail about the different specifications of each study in Section 2.8.3.

The basic procedure for the different simulation studies is the same. We begin by assuming a heterosexual marriage market in which males and females base partnership decisions on their own education level and the education of prospective spouses, as well as some other unobserved characteristics. We assume that the marginal distributions of gender and education within the population are known and represented as availability scenario $\mathcal{A} = \{\bar{w}(x), \bar{m}(z)\}$. We also assume that the form of the partnership utility function $W(x_i, z_j|\underset{\sim}{\beta})$ and the preference parameters $\underset{\sim}{\beta}$ for individuals in the market are both known.

We suppose a population of size N which reflects the gender and education distributions of availability scenario $\mathcal{A}$ and the partnership preferences $\underset{\sim}{\beta}$. In simulation studies I.i and II, we assume the data consists of information on the full simulated population, while for simulation studies I.ii and III we suppose that the data is a sample of n_h households from the simulated population. We then obtain the distribution of partnerships $\bar{c}$, either empirically or via large population approximation described in Equation 2.7. We fit the revealed preferences model to the data to produce estimates $\hat{\underset{\sim}{\beta}}$ of the original preference parameters.

2.8.1 Choice of availability scenarios

We consider two marginal distributions for gender and education as our availability scenarios, referred to hereafter as $\mathcal{A}_1$ and $\mathcal{A}_2$. Both availability scenarios were chosen based on data from the 2008 Panel of the Survey of Income and Program Participation (SIPP), which has been made publicly available by the United States Census Bureau (U.S. Bureau of the Census, 2020). The 2008 SIPP is a nationally representative panel study that followed individuals in sampled households from 2008 through 2012. Individuals responded to a set of core questionnaires administered every 4 months and in 2009, individuals over the age of 15 answered a series of supplemental survey questions on their marital history, and, if currently married, the date their most recent marriage began.

We limit the analytic sample to individuals 18-59 years old who either: 1) at wave 2 had married for the first time in the past year; or 2) were not currently married and never had been married, and were living in households that responded to Waves 1 and 2 of the 2008 SIPP Panel as well as the marital history topical module administered at the Wave 2 interview. We focus on first marriages that initiated no more than a year prior to the survey data to ensure we capture preferences at the time the marriage was initiated and to avoid bias due to marital dissolution, remarriage, or educational upgrading (Schwartz and Mare, 2005; Kalmijn, 1994). With these limitations, our analytic sample consists of 21,567 individuals, 1,040 of whom had married in the last year, and 20,527 who remained single in the last year. The 1,040 newly-married individuals were by survey design married to another sample member, and therefore were in 520 couples in our sample. Within a given year, entering into a marriage is therefore relatively rare, with only 5% of individuals in our analytic sample having entered a new marriage. Thus preferences for marriage, meaning for getting married in a given year, are negative when we run the revealed preferences model in Section 2.9. This 2008 SIPP sample design corresponds to Menzel's (p.913) sample of households that are assumed to be drawn from a population resulting from the stable matching. In our case, we have 21,077 households, of which 520 consist of a married couple.

The maximum education level attained by each individual is a categorical variable coded as 1 for less than a high school education, 2 for a high school degree, 3 for some college, and 4 for a bachelors degree or beyond. The education level of female i is stored as x_i and the education level of male j is stored as z_j.

The first availability scenario $\mathcal{A}_1$ is factual (a population like the 2008 SIPP). In other words, it utilizes the gender and education distributions of the overall population based on the 2008 SIPP sample, and the partnership preferences of individuals are equal to preferences estimated in the 2008 SIPP sample. In this availability scenario, about 49.1% of individuals are women and 51.9% are men.

Availability scenario $\mathcal{A}_2$ has the same marginal distribution of education and availability as the non-Hispanic Black population in the 2008 SIPP data. However, the preferences of individuals in availability scenario are kept the same as those of individuals in scenario $\mathcal{A}_1$. Under availability scenario $\mathcal{A}_2$ about 58.0% of individuals are females and 42.0% are males, which reflects a significant gender skew not seen in scenario $\mathcal{A}_1$. In both $\mathcal{A}_1$ and $\mathcal{A}_2$, women are less likely to have less than a high school degree (education category 1) and are more likely to have completed any college (education category 3 or higher).

In simulation studies I.i and I.ii we simulate populations from both scenarios $\mathcal{A}_1$ and $\mathcal{A}_2$.

Table 2.1: Availability Scenarios

Availability scenario	Source of availability distribution	Type
$\mathcal{A}_1$	2008 SIPP full sample	Total U.S. population in 2008
$\mathcal{A}_2$	2008 SIPP non-Hispanic Black sample	A realistic sub-population availability

Given the utility model specification, we assume that in both scenarios all individuals are characterized the same true preference parameters $\underset{\sim}{\beta}^0$. By fitting the revealed preferences model on data from populations based on both availability scenarios, we show that preference parameter estimates are unbiased even as the availability of potential partners changes. Thereafter, in simulation studies II and III we only simulate populations based on availability scenario $\mathcal{A}_1$.

2.8.2 Utility model specification

We now discuss three different partnership utility specifications under which we test the performance of the revealed preferences model. We first consider a very simple model specification in which a female experiences a shift in utility, relative to her utility had she remained unpartnered, only when she partners with a man whose education level is the same as her own. The tendency for partnered individuals to share similar characteristics is reflected by *homogamous* pairings, and preference for such partnerships is referred to as *homophily.* We designate this specific model as the *uniform homophily model* because the shift in the deterministic component of the utility is uniform for all types (education levels) of individuals. The set of parameters for this model is denoted as $\underset{\sim}{\beta}^{\mathrm{UH}}$. The sum of woman i and man j's utilities if they partnered with each other is

$$W_{ij}(x_i, z_j|\underset{\sim}{\beta}^{\mathrm{UH}}) = \beta_0 + \beta_{\mathrm{UH}}\mathbb{I}\{x_i = z_j\}, \tag{2.14}$$

The uniform homophily model can be extended if we assume that the utility a woman derives from a partnership is based not only on whether she and her partner have equal

Table 2.2: Gender and Education Distributions under the two availability scenarios

	Males		Females	
Education Level	% Population	% of Males	% Population	% of Females
	Availability scenario $\mathcal{A}_1$			
1 (< high school)	7.4	14.5	5.3	10.9
2 (high school)	14.5	28.5	11.2	22.8
3 (some college)	19.5	38.4	21.0	42.9
4 (≥ bachelors)	9.5	18.6	11.5	23.4
Total	**50.9**	**100.0**	**49.1**	**100.0**
	Availability scenario $\mathcal{A}_2$			
1 (< high school)	7.2	17.1	7.1	12.3
2 (high school)	13.8	33.0	15.3	26.4
3 (some college)	15.9	37.8	25.4	43.7
4 (≥ bachelors)	5.1	12.1	10.2	17.6
Total	**42.0**	**100.0**	**58.0**	**100.0**

education levels, but also on the education level itself. Once again, there is a corresponding utility function for males. We refer to this as a *differential homophily model*, where the change in utility depends not only on partners share a particular trait, but also on the value of trait considered. Given the set of education types $\mathcal{T} = \mathcal{X} \cup \mathcal{Z}$, the model is denoted as $\underset{\sim}{\beta}^{\text{DH}}$.

$$W_{ij}(x_i, z_j | \underset{\sim}{\beta}^{\text{DH}}) = \beta_0 + \sum_{t=1}^{\mathcal{T}} \beta_{t,t} \mathbb{I}\{x_i = z_j = t\}. \tag{2.15}$$

The third model we consider is a modified version of the *saturated mix model*, which includes every possible first-order term. In the saturated mix model, women and men both derive a different utility from each possible combination of education levels in the marriage. The full set of parameters is denoted by the vector $\underset{\sim}{\beta}^{\text{SM}}$.

We are able to remove the intercept term β_0 from the utility model because it is a constant value added to the matching utility of every pair. Thus, the sum of the utilities of two individuals in a marriage is given by

$$W(x_i, z_j | \underset{\sim}{\beta}^{\text{SM}}) = \sum_{x,z} \beta_{x,z} \mathbb{I}\{x_i = x, z_j = z\}. \tag{2.16}$$

The term $\beta_{x,z}$ is the coefficient to an indicator which equals 1 if the partnership consists of a woman of type x and a man of type z, and 0 otherwise. The saturated mix model consists of $X \times Z$ first-order parameters where, as previously defined, $X = |\mathcal{X}|$ is the number of possible discrete types for women and $Z = |\mathcal{Z}|$ is the number of possible discrete types for men.

Out of the 21,077 households in the SIPP analytic sample, there is 1 couple which consists of a woman with education level 1 and a man with education level 4, and 1 couple which contains a woman with education level 4 and a man with education level 1. The low counts make estimation of the $\theta_{1,4}$ and $\theta_{4,1}$ parameters difficult, as the joint utility of such couple is perceived as effectively negatively infinite. To facilitate estimation in these cases, we consider pairings between a woman with education level 1 and a man of education level 4 to have equal utility to a pairing between a woman with education level 2 and a man of education level 4. This "reduces" the $\beta_{1,4}$ and $\beta_{2,4}$ parameters to a $\beta_{1 \text{ or } 2,4}$ parameter. Likewise, we can equate pairings between a woman with education 4 and man with education 1 to pairings between a woman with education 4 and a man with education 2, so that $\beta_{4,1}$ and $\beta_{4,2}$ are replaced by $\beta_{4,1 \text{ or } 2}$. Thus, rather than using the fully saturated model with 16 parameters to estimate, we consider a *reduced mix model* with only 14 parameters, represented in vector form as $\underset{\sim}{\beta}^{\text{RM}}$. The situation here is very similar to the "collapsing cells" situation in contingency table modeling (Agresti, 2002, Section 10.1).

We note that mix models are of particular interest to demographers who have access to large samples from populations. When the size of the available data is small as is the case for simulation studies II and III, however, model saturation can result in biased and highly variable parameter estimates and the less parametrised uniform homophily or differential homophily model may be preferable.

The testing procedure for each model specification is the same, and we outline the basic procedure which is used in simulation study I. We first choose a set of preference parameters $\underset{\sim}{\beta}^0$ given the specific model that we assume is the underlying truth. This is done by using RPM to fit that model on the analytic 2008 SIPP data and calculating parameter estimates $\tilde{\underset{\sim}{\beta}}$. We assume that these estimates are equivalent to the true preference parameters of individuals under all availability scenarios, so that $\underset{\sim}{\beta}^0 = \tilde{\underset{\sim}{\beta}}$. In each simulated population, the known preferences $\underset{\sim}{\beta}^0$ are applied to calculate total household utility for every potential partnership and form a stable matching. We fit the revealed preferences model on the observed stable matching outcome from the simulated population, constraining the MLPLEs to lower and upper bounds of -10 and 10, respectively, and utilize the methodology proposed in Section 2.6 to obtain bias-corrected MLPLEs. We compare these estimates to the true underlying true preferences $\underset{\sim}{\beta}^0$. We make minor modifications to this process for simulation studies II and III which are described below.

2.8.3 Details for simulation studies I, II, and III

Having established the availability scenarios and utility models we will consider in this paper, we now provide further detail on each of the simulation studies.

To demonstrate that the revealed preferences model produces unbiased estimates of $\underset{\sim}{\beta}$ given either an observed distribution of partnerships $\bar{c}$ or a large population approximation of $\bar{c}$, we conduct simulation study I in two parts. In study I.i, we simulate populations of size $N = 6,000$. The generating distribution for the populations may be either availability scenario $\mathcal{A}_1$ or $\mathcal{A}_2$, and a population consists of individuals whose partnership utilities are either all determined by the differential homophily utility model (Equation 2.15) or the reduced mix utility model (Equation 2.16). Thus, we consider four possible combinations of availabilities and utility model specifications, and we simulate 1,000 populations of each combination. For every simulated population, based on the utility function and $\underset{\sim}{\beta}^0$ we obtain a stable matching using the Gale-Shapley algorithm. (Gale and Shapley, 1962) We then compute the empirical distribution of partnerships $\bar{c}$ observed in this stable matching. Treating the simulated data as a census, we fit the revealed preferences model to obtain preference parameter estimates.

Ideally, to obtain the distribution of partnerships within a population, we would always use the Gale-Shapley algorithm to first achieve a stable matching for that population. However, a large amount of memory and computational power is required to create stable partnerships for large population sizes (e.g., greater than 20,000), since the household utility matrices $\{W_{ij}\}_{N_w \times N_m}$ and $\{M_{ij}\}_{N_m \times N_w}$ must be calculated for all potential pairings. In such cases, rather than implementing the Gale-Shapley algorithm to achieve a stable matching, we can approximate the empirical distribution of household types in the outcome and estimate preference parameters based on the large population approximation (Equation (2.7)). In general, we suggest using the large population approximation rather than replicating the actual matching process when working with simulated populations with more than 6,000 individuals.

In study I.ii, we show that a large population approximation of $\bar{c}$ is suitable for unbiased estimation of preference parameter estimates. We begin once again assuming that a population can be characterized by the same four combinations of availabilities and utility model specifications considered in study I.i. In this case, however, we suppose that $N = 300$ million within a single population. Rather than simulating the population directly, we approximate the distribution of partnerships that would occur in a stable matching within such a population. We then sample about 20 thousand households from this approximated distribution, fit the revealed preferences model to the sample data, and obtain preference parameter estimates. For each combination of availability and utility model, we take 1,000 samples.

We note here that populations generated using availability scenario $\mathcal{A}_1$ can be considered "factual" in that they resemble the 2008 SIPP sample. In other words, both the underlying marginal distributions of gender and education $\mathcal{A}_1$ and the preferences $\underset{\sim}{\beta}$ used to generate matchings in the simulated population are based on the 2008 SIPP. In contrast, populations generated using availability scenario $\mathcal{A}_2$ are "counter-factual" as the population composition changes while preferences of the 2008 SIPP are maintained.

In simulation study II, we simulate 1,000 populations each of size $N = 60, 600$ and $6,000$ with the assumption that the education and gender for individuals in all populations are generated based on availability $\mathcal{A}_1$ and all individuals have a uniform homophily utility model (Equation 2.14) for partnership. We choose the uniform homophily model for this part of the study to avoid negatively infinite estimates at $N = 60$. We also make a small modification here to the model testing procedure described previously; we do not set the true underlying preferences $\underset{\sim}{\beta}^0$ equal to $\tilde{\underset{\sim}{\beta}}^{\text{UH}}$, the preference estimates obtained by fitting the uniform homophily model on the SIPP data. Instead, we increase the intercept term in $\tilde{\underset{\sim}{\beta}}^{\text{UH}}$ by a magnitude of 4 to increase the number of partnerships and facilitate stable estimation of preference parameters. For each simulated population, we use the Gale-Shapley algorithm to obtain a stable matching and fit the revealed preferences model to the observed $\bar{c}$ for the entire population. We then compare the bias of the median parameter MLPLEs and bias-corrected MLPLEs. at each N as N increases. We also evaluate the effectiveness of using a bootstrap approach for bias correction of $\underset{\sim}{\hat{\beta}}$ at different N.

For simulation study III, we simulate populations of size $N = 6,000$ with the assumption that the education and gender for individuals in all populations are generated based on availability $\mathcal{A}_1$ and all individuals have a differential homophily utility model (Equation 2.15) for partnership. For each stable population, after using the Gale-Shapley algorithm to reach a stable matching, we sample $n_h = 600, 1,200$ or $3,000$ households. Similar to simulation study II, rather than set $\underset{\sim}{\beta}^0 = \underset{\sim}{\tilde{\beta}}^{\text{DH}}$, we increase the intercept term in $\underset{\sim}{\tilde{\beta}}^{\text{DH}}$ by 4 units to increase partnership rates. We fit the revealed preferences model to the sample data and compare the performance of the mean MLPLEs and bias-corrected MLPLEs $\underset{\sim}{\hat{\beta}}$ as n_h increases.

2.9 Results

2.9.1 Simulation study I.i: Population Data

For simulation study I.i, we simulate populations of size $N = 6,000$ from "factual" availability $\mathcal{A}_1$ and "counterfactual" availability $\mathcal{A}_2$ and utilized the Gale-Shapley algorithm to perform stable matching on the individuals in each simulated population. The utility derived from each potential partnership was calculated based on $\underset{\sim}{\beta}^0$ for a specified deterministic utility function and an extreme-value Type-I distributed random error term. The utility a woman achieves by staying single is equal to maximum value of $\sqrt{N_w}$ random draws from an extreme-value Type-I distribution.

The plots in Figure 2.1 show the distribution of the 1,000 bias-corrected MLPLEs for each combination of availability scenario $\mathcal{A} \in \{\mathcal{A}_1, \mathcal{A}_2\}$ and two utility model specifications (differential homophily and reduced mix). The red lines in the plots represent the true β_0 preference values which induced the Gale-Shapley matchings. Negatively infinite estimates are recognized via a point mass at value -6 with an area proportional to the number of such estimates.

The medians and standard deviations, of parameter estimates for the match and reduced mix models are presented in Tables A.1 and A.2. For this and all following simulation studies,

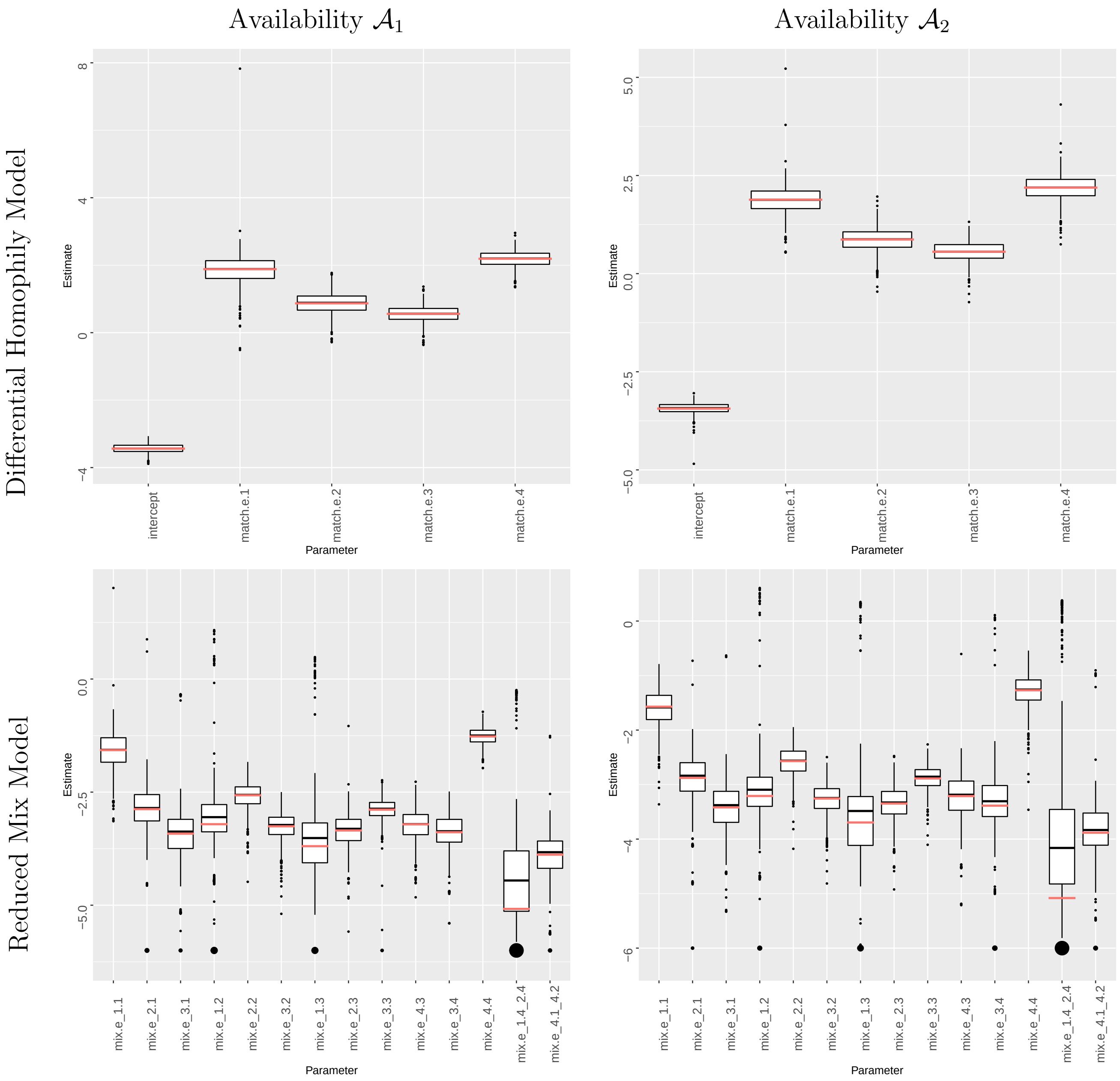


Figure 2.1: Distribution of bias-corrected MLPLEs in simulation study I.i: Population data with $N = 6,000$ (1,000 simulations)

we compute standard deviation as a standardized version of the interquartile range. Tables with numerical results are in Appendix A.1.

Although availability of individuals differs between $\mathcal{A}_1$ and $\mathcal{A}_2$, under both model specifications the revealed preferences model produces estimates of the true preference parameters which are about equal in accuracy and precision. Based on the plots for study I.i in Figure 2.1, the mean estimates of all reduced mix model parameters except $\beta_{1 \text{ or } 2,4}$ appear to align with the true values fairly well in all availability scenarios. Furthermore, the estimates for all parameters, with the exception of $\beta_{1 \text{ or } 2,4}$, resemble a normal distribution.

We note that when using the reduced mix model, for both availability scenarios the distribution of $\hat{\beta}_{1 \text{ or } 2,4}$ displays a right skew. When the population has very few or no pairings of a certain type, the model estimates the total utility of such a pairing as very negative, if not infinitely so. In our implementation of this model, we impose an upper bound of 10 and a lower bound of -10 on all parameters. The high frequency of extremely negative values (≤ -6) in the parameter estimates of $\beta_{1 \text{ or } 2,4}$ indicate that in that specific population, there were very few or no households which contained a matching between a woman with education level 1 or 2 and a man with education level 4.

We ran simulation study I.i with both the differential homophily and reduced mix model specifications on a third availability scenario (results not shown), in which men outnumber women 3:1 and educational attainment was highly asymmetric across genders. We found that in this artificially extreme case, the occurrence of highly negative estimates of $\beta_{1 \text{ or } 2,4}$ increased. Furthermore, the estimates of $\beta_{1,3}$ and $\beta_{2,3}$ also showed a strong right skew. In general, the standard deviation of the parameter estimates tends to increase as the population becomes more skewed.

2.9.2 Simulation study I.ii: Sampling from a large population

In this simulation study, we simulate samples from large populations using availabilities $\mathcal{A}_1$ and $\mathcal{A}_2$, each with a nominal size of $N = 300$ million and a household sample size of $n_h = 21,077$ (equivalent to the size of the analytic SIPP sample). We find that the resulting estimates are very robust to the population size as long as it is modestly large (e.g., $N > 6,000$). We choose to study large populations as they are typical in demography. Brien (1997), for example, compares model performance for three levels of population aggregation of the marriage market: in descending order, state, metropolitan area, and county group. He finds that the highest, state level of aggregation best explains marriage differentials between population subgroups.

We employ a large population approximation of stable matching outcomes in the simulated population that would be observed if individuals had true preferences $\underset{\sim}{\beta}^{0}$, either based on a differential homophily or a reduced mix utility model. The plots in Figure 2.2 show the distribution of the 1,000 parameter estimates $\hat{\underset{\sim}{\beta}}$ for each combination of simulating availability scenario and revealed preferences model specification. The red lines in the plots represent the true values $\underset{\sim}{\beta}^{0}$ which we are attempting to recover.

The first row of Figure 2.2 shows the distributions of the parameter estimates under the differential homophily model given large simulated population. The medians and standard errors of the differential homophily model parameters are presented in Table A.3.

In both availability scenarios, we observe that the mean estimate for each parameter in the differential homophily model is very close to the true value. We also note that when simulating from availability scenarios $\mathcal{A}_1$ and $\mathcal{A}_2$, the standard errors of the parameter estimates stay about the same. However, we also ran this simulation study under the artificially extreme availability scenario described in the results for study I.i (results not shown) and found that in that case the standard error nearly tripled for all parameters.

The second row of Figure 2.2 shows the distributions of the parameter estimates under the reduced mix model when the simulated population size is large. Due to space constraints, we relegate Table A.4, which shows the medians and standard errors of the parameter estimates, to Appendix A.1. The revealed preferences model recovers the true preference parameters $\underset{\sim}{\beta}^{\mathrm{RM},0}$ for all availability scenarios. Furthermore, the standard deviations of all parameter estimates stay similar across the availability scenarios.

2.9.3 Simulation study II: Small population sizes

Simulation study II is carried out for two primary purposes. The first purpose is to illustrate how the revealed preferences model can be used with population data that includes small to very-small population sizes. The second is to show the relationship between population size N and estimate bias and the relationship between population size N and the effectiveness of our proposed bias correction methodology.

We simulate 1,000 populations each of sizes $N = 60, 600$, and $6,000$ from availability scenario $\mathcal{A}_1$. We then use the Gale-Shapley algorithm to obtain a stable matching in the population, with true preference parameters $\underset{\sim}{\beta}^{\mathrm{UH},0}$ based on the uniform homophily model and the inflated intercept. The distributions of the maximum large-population likelihood estimates (MLPLEs) and the bias-corrected MLPLEs for each N are shown in Figure 2.3. The median estimates and standard deviations of the MLPLEs and bias-corrected MLPLEs given in Table A.5, respectively.

The panels in the first column of Figure 2.3 shows model estimates for each parameter when $N = 60$. Each panel corresponds to a single parameter and shows two distributions; the left box plot shows the distribution of the MLPLEs and the right box plot shows the distribution of the bias-corrected MLPLEs. The second and third columns of Figure 2.3 show the same information for $N = 600$ and $N = 6,000$.

At each population size, the MLPLEs for both the intercept term and the uniform homophily preference term underestimates the true value $\underset{\sim}{\beta}^{\mathrm{UH},0}$, though the bias of the latter term is of a much smaller magnitude than of the former. For both parameters, bias decreases as the population N increases. The standard deviation of the MLPLE estimate decreases substantially as N increases; we see in Table A.5 that when N increases by a factor of 10, the standard deviation decreases by a factor of approximately 1/3 for the intercept parameter and 1/4 for the homophily parameter.

When bias correction methodology is applied, the bias in the estimates of both the intercept and the homophily parameter decreases for all population levels. The improvement of the estimates due to bias correction is especially clear for the intercept term. We also notice that for both parameters the difference in the mean MLPLE and mean bias-corrected MLPLE is greatest at $N = 60$. The bias-corrected MLPLEs have a slightly higher standard deviation than the non-bias-corrected MLPLEs, though the magnitude of this different decreases with population size N.

2.9.4 Simulation study III: Increasing Relative Sample Size

In simulation study III, we investigate the relationship between the sample size n_h and the bias of MLPLEs when fitting the revealed preferences model, as well as the impact of bias correction methodology on the estimates as sample size increases.

Figure 2.4 shows the distribution of parameter MLPLEs and bias-corrected MLPLEs at each value of n_h, while the medians and standard deviations of the estimates are given in Tables A.6. At all three values of n_h, the mean MLPLE estimate underestimates the true value. We see much less bias, though still a small amount (< 0.05 units), in the estimates for the matching preferences at each education level. The variance of the MLPLEs decreases as the sample size increases.

After bias correction methods are used, the difference between the truth $\underset{\sim}{\beta}^{\mathrm{DH},0}$ and the bias-corrected MLPLE becomes very small. As with simulation study II, a consequence of bias correction is a slight increase in variance for $n_h = 600$ and 1200. At $n_h = 3,000$, however, the impact of bias correction on the variance of estimates is ambiguous. While the variance increases with bias correction for the intercept parameters and the parameters indicating homogamy on the education levels 2 (high school education) and 3 (some college), the variance of the parameters indicating homogamy at education levels 1 (less than high school) and 4 (college degree or higher) actually decreases.

We repeated this exercise at $N = 1,000$ and $n_h = 100, 200$, and 500 (results not shown) and obtained results that were consistent with the earlier findings. Specifically, the MLPLEs showed some bias at all n_h, with bias in the intercept MLPLEs being much higher than in other parameters. The bias-corrected MLPLEs were closer estimates of the true $\underset{\sim}{\beta}^{\mathrm{DH},0}$. As n_h increased, the variance of both the MLPLEs and the bias-corrected MLPLEs decreased. In general we find that as long as the sample size is large enough to ensure non-zero entries in $\bar{c}$ are rare, the bias-corrected MLPLEs have high accuracy and improve in precision as n_h increases.

2.9.5 Confidence intervals and coverage probabilities

To supplement the findings in simulation study I.ii, we calculate 95% confidence intervals for bias-corrected MLPLEs based on samples from simulated populations of size $N = 300$ million, and we compare the empirical coverage rates of the true parameter values to the 95% threshold.

To calculate empirical coverage rates, we simulate $S = 200$ samples from large populations from availability $\mathcal{A}_1$. For each sample, we fit the reduced mix model and produce analytical 95% confidence intervals based on the approximated Hessian matrix, as detailed in Section 2.7. We additionally implement the basic, percentile, and modified studentized t bootstrap methods also discussed in Section 2.7 to construct empirical 95% confidence intervals. An illustration of the coverage results from a single set of 200 simulations are presented for selected parameters in Figures A.1 and A.2 in Appendix A.2.

The process of simulating 200 populations and constructing confidence intervals for each simulation was repeated 40 times, so that we observed an empirical coverage rate across 200 simulations 40 times. We show the mean coverage rates of the reduced mix model parameters using the various confidence intervals in the right-hand panel of Figure 2.5. The dotted black line at 0.95 denotes the 95% threshold we aim to achieve. The analytical confidence intervals appears to be the most volatile; across the 14 parameters estimated in the reduced mix model, the mean coverage rate of the analytical confidence intervals ranged from 19.3 to 99.2%. The three bootstrap confidence intervals have a more consistent performance; within each interval type, the range of the mean coverage rates across the parameters is about 2 percentage points. The basic and percentile bootstraps both display undercoverage, with mean coverage rates around 90% across parameters. The studentized t interval achieves mean coverage rates closest to the 95% target.

We pay special attention to the coverage rates for the $\beta_{1 \text{ or } 2,4}$ parameter. This parameter corresponds to a preference for couples with a female of education level 1 or 2 and a male of education level 4. As noted earlier, the number of couples of this type in the SIPP data and in the simulated samples was very small. The mean coverage rates of the percentile, basic, and analytical confidence intervals are all lowest for this parameter, likely because of the low count of such couples in the data. We note, however, that the performance of the studentized t interval does not appear to be affected by the low couple count. In fact, the mean coverage rate of the studentized t interval for $\beta_{1 \text{ or } 2,4}$ is 95.3%.

The coverage rates shown Figure 2.5 were produced based on populations simulated from the "factual" availability scenario $\mathcal{A}_1$. We repeated the procedure to evaluate confidence interval coverages using populations simulations from the "counterfactual" availability scenario $\mathcal{A}_2$ (results not shown. We found no evidence that the change in population availabilities impacted the coverage rates of the bootstrap confidence intervals.

We also repeated this process to evaluate the performance of confidence intervals for differential homophily model parameters. In this case, we found that the analytical confidence intervals were two to three times wider than the student t intervals and captured the true value 100% of the time for all parameters, indicating overcoverage. We again observed that the studentized t confidence intervals consistently achieved the highest coverage rate of the bootstrap procedures. The basic and percentile bootstrap 95% confidence intervals show slight undercoverage, falling between 89.6% and 91.3% coverage. A plot of mean coverage rates by analytical and bootstrap confidence intervals for the differential homophily model is provided in the left panel of Figure 2.5 under Appendix A.2. We show coverage results from a single set of 200 simulations for selected parameters in the differential homophily model in Figures A.3 and A.4 in Appendix A.2

2.10 Discussion

The ability to extract preferences separably from availabilities is a key feature of the revealed preferences model and methodology which we propose in this paper. In simulation study I.i we simulate a small population ($N = 6,000$) and run the Gale-Shapley algorithm to obtain a stable matching. Given statistics of the types of matchings, we are able to compute parameter estimates which are very close to the true values. We note that Logan (1996b) was able to show a similar result for his initial special case of the model.

In simulation study I.ii, we simulate a large population and obtain an approximate distribution of household types in a stable matching. We sample couples and individuals from this matching and then maximize (2.10) over the sample data to obtain parameter estimates, showing that the method accurately recovers true preference parameter values even under various different availabilities of prospective partners. In both simulation studies I.i and I.ii, the distribution of the parameter estimates appears Gaussian in most cases. The standard errors decrease when the population size is larger, as in simulation study I.ii.

When there are very few or none of a certain type of couple in the data, the total utility of such a pairing is estimated be negative infinity. As an example, we refer to the estimates of $\beta_{1 \text{ or } 2,4}$ in simulation study I.ii, shown in the first column of Figure 2.2. If we observed no couples in which a woman has education level 1 or 2 and the man has education level 4, then the parameter estimate for the utility model term indicating such a match is negative infinity. This artifact is a form of separation also seen for generalized linear models (Heinze and Schemper, 2002). The high concentration of parameter estimates for $\beta_{1 \text{ or } 2,4}$ under -6 correctly captures this and reflects the lower utility corresponding to such pairings.

For both availability scenarios $\mathcal{A}_1$ ("factual") and $\mathcal{A}_2$ ("counterfactual") under the differential homophily model, the standard errors of the estimates in simulation study I.ii are smaller than the corresponding values in Simulation study I.i (small population scenario). As in simulation study I.i, the distributions of the parameter estimates appear to follow a Gaussian distribution.

In simulation studies II and III, we investigated the performance of the revealed preference model under different population and sample sizes. In simulation study II we assumed access to population data.

We found that for different population sizes N, the bias-corrected MLPLEs provided accurate estimates of preference parameters with the variance of estimates decreasing inversely with N. This is a significant finding as the previous formulation of the model proposed by Menzel (2015) required n/N to be small. We show that even when $n/N = 1$, the bias-corrected MLPLEs obtained using bootstrap methods recover true preference for pairings. The bias-corrected MLPLEs are similarly effective in reduced estimate bias in simulation study III, in which we obtain samples of different sizes from populations of $N = 6,000$ individuals. We note again that bias in the MLPLEs is mitigated through the bootstrap bias correction. Together, the findings of simulation studies II and III provide strong support for the use of bias-corrected MLPLEs to estimate preferences in revealed preferences models and show that accurate estimates can be achieved for a wide range of n/N.

We also evaluate different methods of accounting for uncertainty in our estimates. Based on results in Section 2.9.5, we believe that the approximation of the Hessian matrix leads to volatile analytical confidence intervals which deviate from the threshold coverage rate of 95%. These confidence intervals are often too wide or narrow to be useful. We also find that among the three bootstrap based methods for producing confidence intervals, the mean coverage probabilities of the studentized t interval were the closest to 95%, while the percentile and basic method-based confidence intervals demonstrate slight undercoverage.

The revealed preferences model can be used to make inferences which are particularly useful in demographic studies. For example, the preference parameter estimates when we fit the reduced mix specification of the revealed preferences model to the 2008 SIPP data are given in column 3 of Table A.2. The estimated utility of pairings in which both individuals have the same education level is substantially higher than it is for pairings where individuals have different education levels. Homophilous behavior is expected by researchers who study matching problems. It is also consistent with the findings of Logan et al. (2008), who presented results that implied a preference for homophily in race and religion in heterosexual marriages.

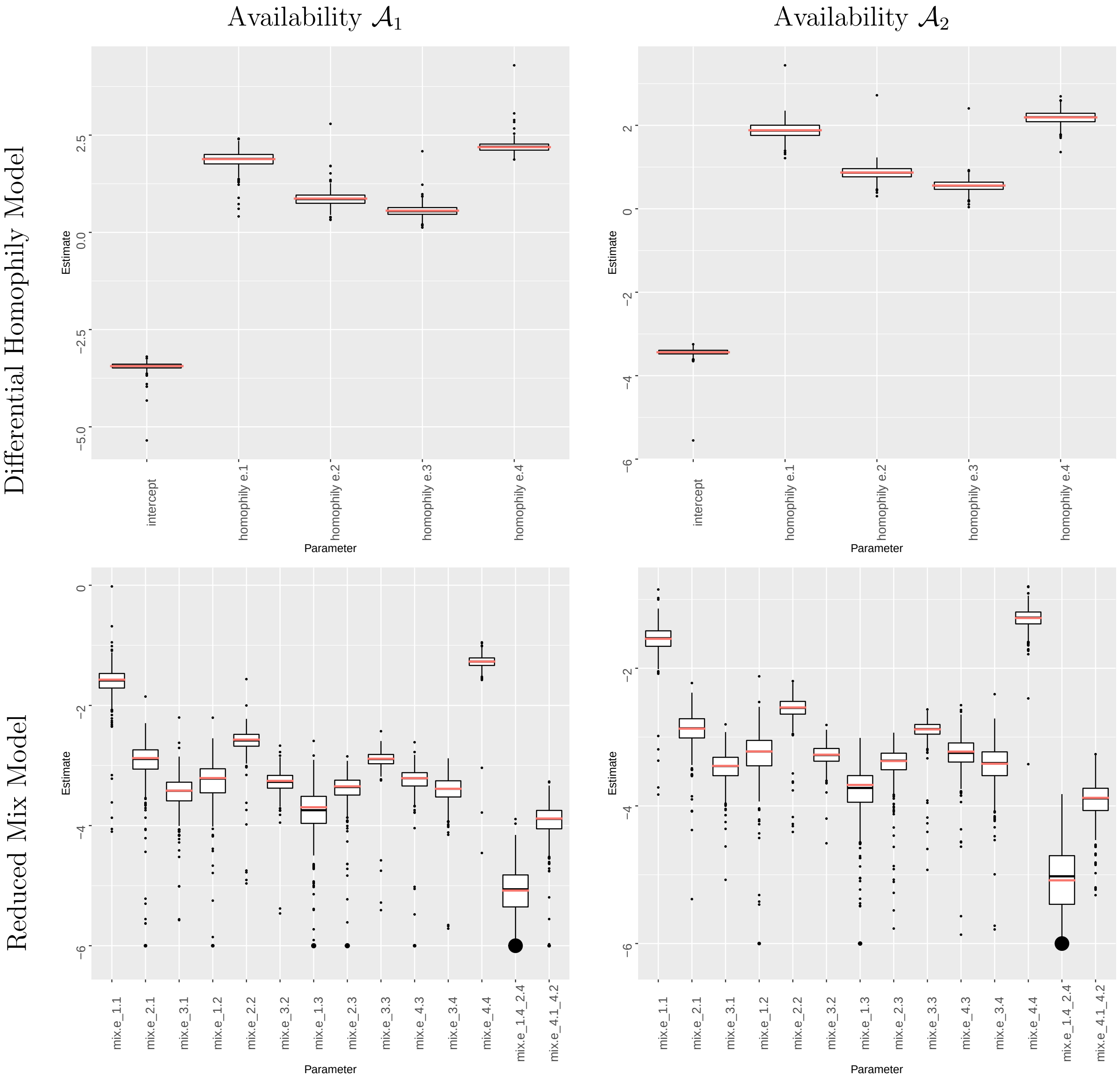


Figure 2.2: Distribution of bias-corrected MLPLEs in simulation study I.ii: Sample data with $n_h = 21,077$ from a population of $N = 300$ million (1,000 simulations)

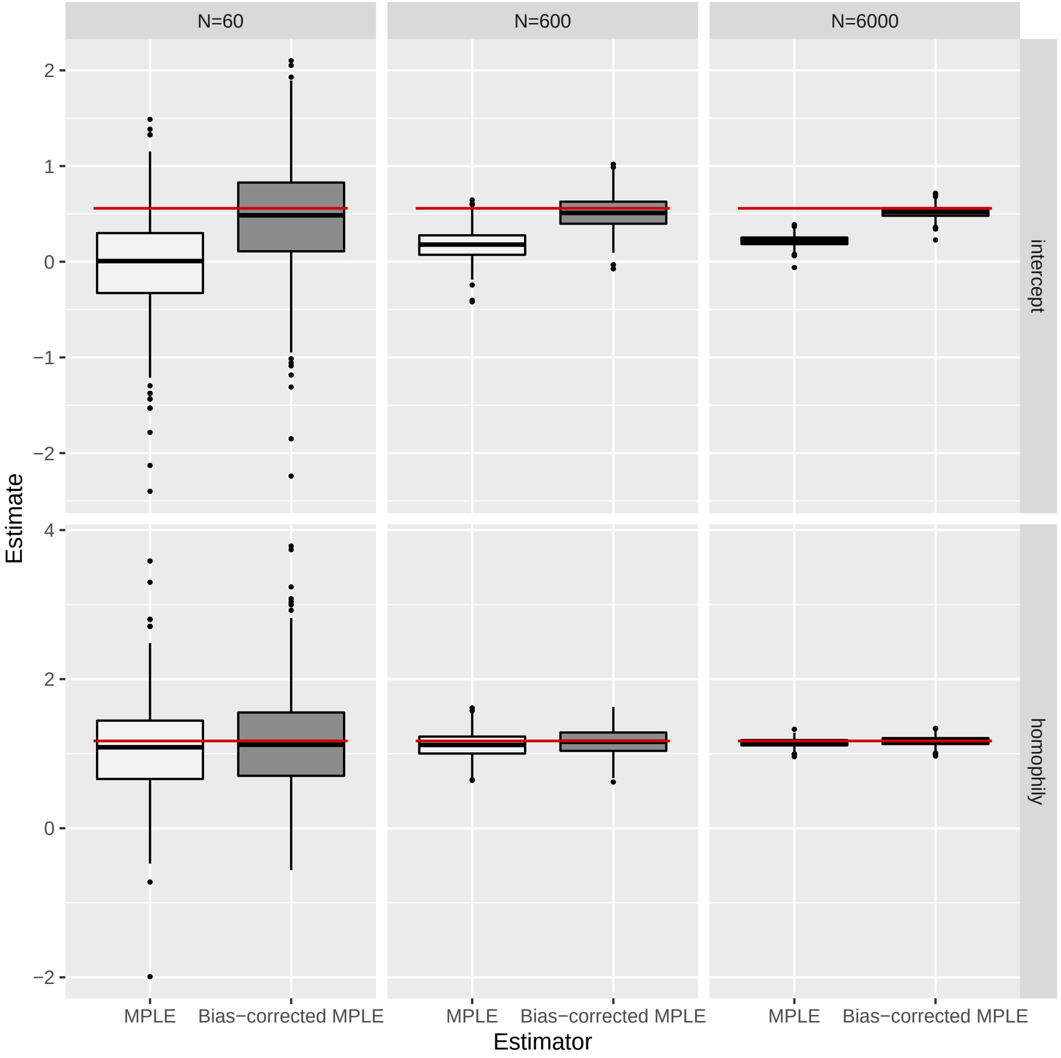


Figure 2.3: Simulation study II: Distribution of uniform homophily MLPLEs and bias-corrected MLPLEs for different population sizes N; 1,000 simulations

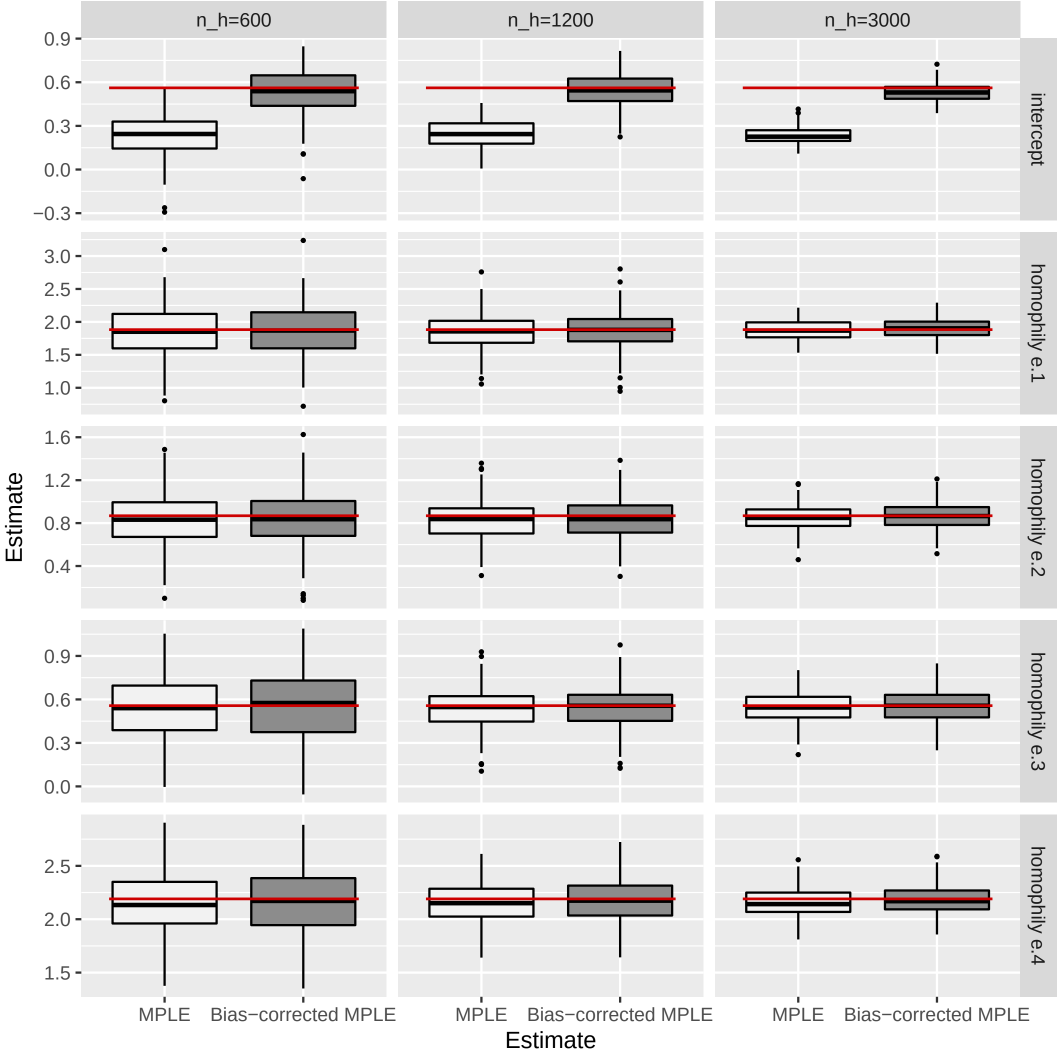


Figure 2.4: Simulation study III: Distribution of differential homophily MLPLEs and bias-corrected MLPLEs for different n_h, where N=6,000; 200 simulations

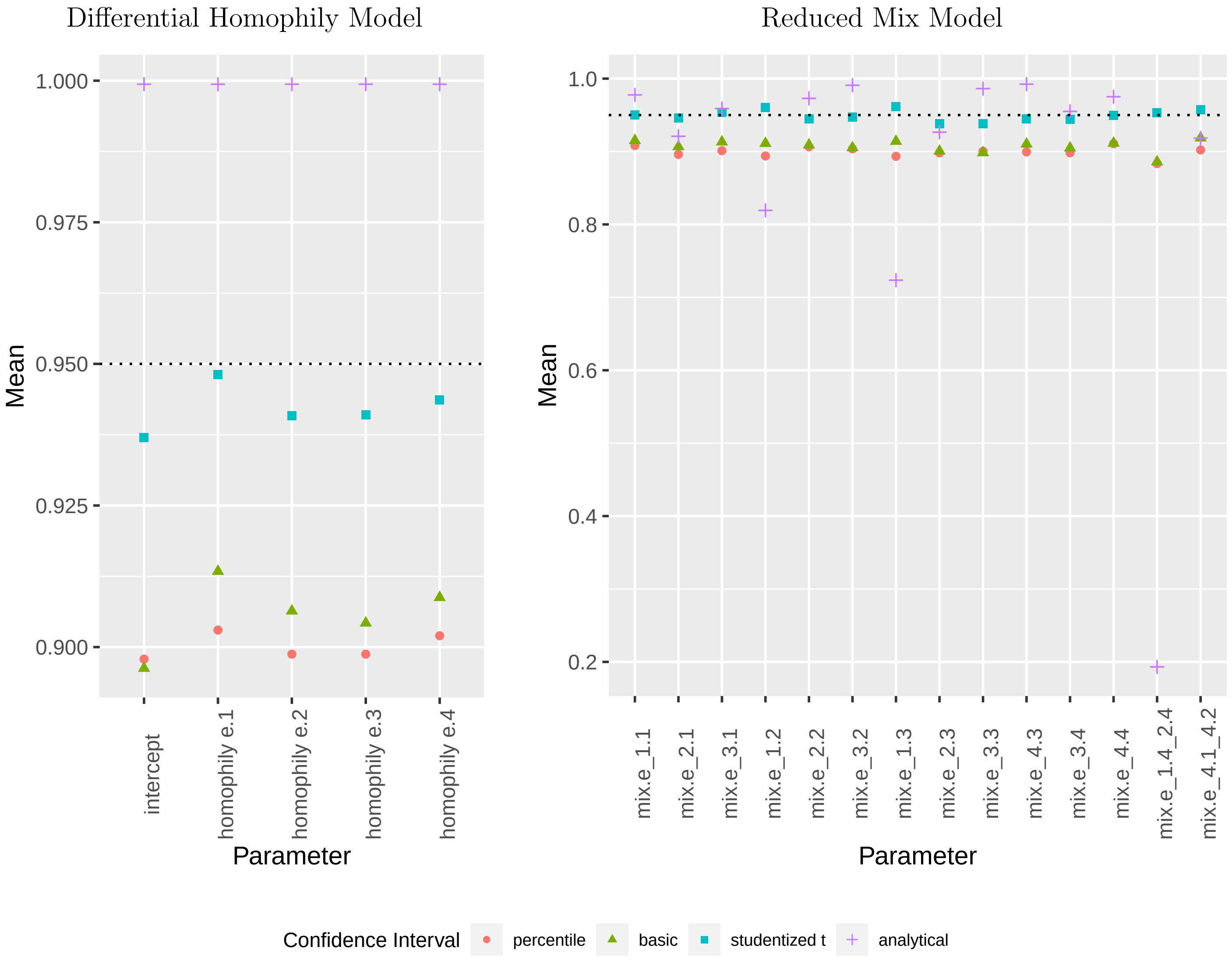


Figure 2.5: Mean empirical coverage probability by bootstrap confidence intervals for model parameters (40 sets of 200 simulations from Availability scenario $\mathcal{A}_1$)

CHAPTER 3

Goodness-of-fit and model selection for revealed preferences models

An advantage of the revealed preferences model proposed in Chapter 2 is that it allows for a broad spectrum of possible model specifications for the partnership surplus utility function W (referred to as the marital surplus model, joint surplus model, or simply surplus model, hereafter) in the two-sided market which motivates the preferences-driven component of the matching. Examples of such model specifications are discussed in Section 2.8.2. The choice of model specification may be driven by prior knowledge, logic, or intuition. By following procedures to obtain the MLPLEs of the preference parameters, we are able to identify the "best" model within that model class based on the observed data.

Often, however, there are multiple plausible specifications that a researcher may consider, and the best one is not immediately clear. A natural follow-up question for researchers is how to evaluate candidate models *across* classes and choose the "best" one, while acknowledging that the "best" model almost surely is not the correct data-generating process, which is too complex to be specified. For example, given the MLPLEs for a uniform homophily surplus model (Equation 2.14) and a differential homophily surplus model (Equation 2.15) for the 2008 SIPP data, how does one choose the best option?

To assess this, we may consider the plausibility of the different models given the data. This plausibility can be measured in a number of ways, such as the likelihood that the hypothesized utility model would produce a matching like the one observed in our data, the difference between some feature of the observed matching and the expectation of that feature under the hypothesized model, or the amount of variation in the observed data that can be explained by the model. Collectively, these attributes all describe the model's"goodness-of-fit" to the data and provide a tool for evaluating a model's performance both individually and relative to competing models, so that researchers can ultimately choose the most appropriate model for their study.

To the best of my knowledge, there are currently no proposed methods for model selection in the existing literature studying two-sided revealed preference models. To address this need, in this chapter I initiate the development of procedures for testing goodness-of-fit for RPMs and propose several procedures for model selection while considering both accuracy in capturing true matching behavior and preference for parsimony.

In Section 3.1 I review the traditional conceptualization of goodness-of-fit and discuss the limitations of the traditional approach in practical settings. In Section 3.2 I develop the concept of model classes in the context of the revealed preferences model as it pertains to model selection. To facilitate the following discussion of goodness-of-fit procedures, in Section 3.3 I suggest various specific statistics that can be used as raw metrics of deviance between the observed data and the expected outcome under a given model. Based on these measures, I propose basic tools for model selection across several candidate models by using information criterion scores and information gain in Section 3.4. In Section 3.5 I propose a re-conceptualization of significance testing in the context of the two-sided matching problem which can be used for additional goodness-of-fit testing when census data is available for small populations. I describe two simulation study procedures in Section 3.6 to assess the performance of the proposed procedures under different conditions and validate approaches. The results of these simulation studies and related discussion are in Sections 3.7 and 3.8. In Section 3.9 I propose additional visual tools for assessing goodness-of-fit which, when applied alongside numerical measures, allow for a better understanding of lack of fit. I close the chapter with a discussion in Section 3.10 of general findings, the state of goodness-of-fitness testing for RPM following the proposed methods in this chapter, and steps for further development.

3.1 Motivation

Goodness-of-fit testing can be broadly defined as a statistical technique used to determine whether a certain data-generating process could have produced an observed set of data. In the traditional goodness-of-fit test, the null hypothesis explicitly states that a particular model W_{H0} is the true data-generating process. The researcher computes a test statistic that measures the deviance d between the observed data and data that would be expected under the null hypothesis. If the test statistic is small, it suggests a good fit between the observed and expected data, while a large test statistic indicates a poor fit. The researcher may also conduct significance testing by comparing d from the observed data to the distribution of deviances D_0 which would occur across samples if model W_{H0} were the true data-generating process and compute a p-value representing the probability of observing the data given the null hypothesis, e.g. $p = \mathbb{P}(d \leq D_0)$. Based on this p-value, the researcher may reject or fail to reject the null hypothesis that model W_{H0} is the true underlying data process. A test with confidence level $(1 - \alpha)\%$ will reject the null hypothesis when $p < \alpha$. If the null hypothesis is correct, the distribution of p values over repeated samples is uniform.

The interpretation of the null hypothesis and subsequent analysis in goodness-of-fit testing varies based on the research question at hand and the conditions under which the study is being conducted. For example, in the heterosexual marriage market within a two-sex population, the researcher may know the distribution of marriage types in a large population and wish to test the hypothesis that an observed sample could have been randomly selected from this population. If all observable agent characteristics are discrete, then this situation is very similar to comparing two categorical distributions. We may use a chi-squared goodness-of-fit test to compare the observed distribution of household types to the expected distribution given the population.[1]

In an alternate and perhaps more interesting case, the researcher may observe a sample

[1] We note here that the assumption of independent observations required for the validity of chi-squared goodness-of-fit tests is violated, as there are subtle interactions and effects of competition between agents in the matching market. This is discussed further in Section 3.3.

of married couples and singletons and wish to test the hypothesis that a certain surplus function motivated the matches in the observed data. In this case, the researcher can derive the expected distribution of partnership types that satisfies Equations 2.7 and 2.8 under the null preferences model and the known availability of different agent types. Once again, the chi-squared goodness-of-fit test can be used to compare the observed partnership distribution to the expected distribution under the model to obtain the likelihood that the observed sample could have resulted from the hypothesized data-generating process.

These examples of applications of goodness-of-fit testing to two-sided matching market problems rely on the assumption that the researcher is able to either observe the distribution of partnership types in a full population or hypothesize the exact utility function that motivated the partnership formation process.

However, once we move beyond these unlikely scenarios, there are major limitations to practical applications of traditional goodness-of-fit testing to two-sided matching markets. In most realistic scenarios, the true utility model is unknown and/or too complex to be specified. Instead, the model specified under the null hypothesis is only an estimate of this truth, often chosen exogenously through inference procedures such as MLPLE. Performing significance testing in the traditional way to test the hypothesis that the null model is correct is not very useful because the null model is already known to be incorrect and thus will almost always be rejected. The goodness-of-fit test described above gives no indication of the potential usefulness of the hypothesized model. Furthermore, even with the rejection of the null model, the direction of the lack of fit remains ambiguous; that is, it can be difficult to ascertain whether the model was over- or under-specified relative to the true data-generating process.

3.2 Choosing $\mathcal{W}_{H0}$ for the null hypothesis

To facilitate the discussion of goodness-of-fit in this chapter, I first discuss the concept of model class[2] in the context of revealed preferences in greater detail. Let $\mathcal{W}$ refer to the set of all possible (linear) surplus models. Then, a "model class" $\mathcal{W}_c \subseteq \mathcal{W}$ is a subset of models defined by some constraint. Examples of model specification are the uniform homophily (UH), differential homophily (DH) and saturated mix (SM) models described in Section 2.8.2. We recall that the MLPLE is computed over such a model class.

Each of these specifications comes with a different set of restrictions. For example, the saturated mix model class $\mathcal{W}_{\text{SM}}$ (Equation 2.16) is the least restrictive linear specification and allows each possible partnership combination of (x, z) for all $x \in \mathcal{X}, z \in \mathcal{Z}$ to have its own deterministic utility.

The differential homophily model class $\mathcal{W}_{\text{DH}}$ (Equation 2.15 introduces the assumption that individuals get no additional deterministic utility by marrying a partner of a different type, relative to staying single. It is only possible to achieve a shift in deterministic utility if the individuals in the marriage have the same type, and the magnitude and direction of the shift depend on the value of the type itself. Formally, $\underset{\sim}{\beta}$ is constrained such that the deterministic shift in partnership utility $\beta_{x,z} = 0$ for all partnerships between women of type x and men of type z where $x \neq z$.

The uniform homophily class $\mathcal{W}_{\text{UH}}$ (Equation 2.14) further constrains the set of models considered by assuming that the additional deterministic utility a partnership between a type x woman and type z man generates when $x = z$ is the same regardless of the value of z or z. Thus, in addition to the constraint which defines the differential homophily class, a second constraint is imposed stating that if $x = z$ and $x' = z'$, then $\beta_{x,z} = \beta_{x',z'}$ for all $(x, x') \in \mathcal{X}$ and $(z, z') \in \mathcal{Z}$.

Each model class is a continuous $\mathcal{N}$-dimensional space, where $\mathcal{N}$ is the number of free model parameters. Then $\mathcal{W}_{\text{UH}}$ is a subspace of $\mathcal{W}_{\text{DH}}$, which is itself a subspace of $\mathcal{W}_{\text{SM}}$.

[2] In this chapter, the terms "model class" and "model specification" will be used interchangeably.

Theoretically, then, if marriages in an observed closed population were truly motivated by a UH surplus model, then the MLPLEs of the preference parameters should be equivalent regardless of whether we fit a UH or SM model to the data. But then, why not always choose to fit the SM model? Computation of the MLPLE becomes more expensive as the number of parameters in the marital surplus model to be estimated increases and requires more data to avoid sparsity issues. Overfit models may also generalize poorly to new data from the same population and lead to erroneous conclusions if data is simulated from the overfit model for future study. A common goal in goodness-of-fit testing is to choose a model that balances good fit to the data with parsimony.

3.3 Deviance metrics for revealed preferences models

An important decision for any goodness-of-fit analysis and for the model selection procedure discussed in the previous two sections is the choice of metric used to measure deviance between observed and expected data. The difference between the observed and expected data in a two-sided matching market can be assessed in several different ways. We recall that in the case of a market where agents have discrete types, we can summarize the observed data using the sufficient statistic $\bar{c}$, representing the frequency of marriage and singleton types across sampled households. (See, for example, Figure 1.1b.) We aim to compare the observed frequency distribution to $\hat{\bar{c}}$, the frequency distribution that which would be expected under the hypothesized model. Alternately, we can compare the observed and expected probability mass distributions, denoted $\bar{f}$ and $\hat{\bar{f}}$, respectively.

Let $X = |\mathcal{X}|$ denote the number of possible observable types for women and $Z = |\mathcal{Z}|$ denote the number of possible observable types for men, We note that the frequency distribution $\bar{c}$ (and the corresponding probability distribution $\bar{f}$) of unique household combinations in the matching market resembles an observation of a vector of counts from a multinomial distribution that has

$$T = X \times Z + X + Z \tag{3.1}$$

discrete outcomes, where $X \times Z$ is the number of possible combinations for types in a married household, and X and Z are the number of possible types of singleton female households and singleton male households, respectively. An important difference between data from a multinomial distribution and matching market data is that in the former, observations are completely independent, so that the outcome for one "trial" (e.g. the marital or singleton type of a sampled household) is unrelated to the outcome of all other trials. Within the two-sided matching market setting where we apply the revealed preferences model, however, there is interaction and competition between individuals so that the count outcomes for each household type share some interdependence. If, as Menzel (2015) asserts, this interaction effect becomes negligible in asymptotically large populations, the consequences of violating the independence assumption are not significant. Thus, we propose that many deviance metrics used for equivalence testing of multinomial distributions can also be applied as test statistics for revealed preferences models, though the distributions of these statistics in two-sided matching markets may differ from the distribution that would occur if observations were truly completely independent.

All of the deviance metrics which I will discuss below broadly measure the distance between observed and expected data but vary in computation. There are some metrics that might be useful in special cases, such as when census data for a population with a fixed composition is available. However, this case requires more sophisticated computational tool and is therefore relegated to a brief discussion in Appendix B.2

Because we mostly deal with empirical distributions of the deviance metrics rather than the asymptotic ones in our proposed procedures, the analytical properties of the statistic are less important for our decision-making and the choice of deviance metric may be more a matter of personal preference. So long as the deviance values used to compare two models are computed using the same metric, they should all lead to similar conclusions. For model selection, as discussed in Section 3.4, we may prefer a metric that additionally considers model complexity. I explore choices for the deviance metric below.

3.3.1 Power-divergence statistics

Suppose we consider a multinomial distribution with n trials and T mutually exclusive possible outcomes. For category $t \in \{1, \ldots, T\}$, $c(t)$ represents the observed frequency in that category and $\hat{c}(t)$ represents the expected frequency so that $\sum_t c(t) = \sum_t \hat{c}(t) = n$, where n is the sample size. The power-divergence class of statistics takes the form

$$2nI^{\lambda} = \frac{2}{\lambda(1+\lambda)} \sum_{t}^{T} c(t) \left[\left(\frac{c(t)}{\hat{c}(t)} \right)^{\lambda} - 1 \right], \qquad \lambda \in \mathbb{R}$$

If the null hypothesis is true, then as $n \to \infty$, the distribution of a statistic in the power divergence class approaches χ^2_{N-p-1}.

The optimal choice of λ is explored by Cressie and Read (1984), who recommend any $\lambda \in [0, 1\frac{1}{2}]$ when there is no information about the alternative hypothesis. Two well-known statistics in the power divergence family which fall in Cressie et al.'s recommended range are the X^2 statistic ($\lambda = 1$) and the likelihood ratio statistic G^2 ($\lambda = 0$). We define these statistics mathematically below:

$$X^2 = \sum_{t}^{T} \frac{(c(t) - \hat{c}(t))^2}{\hat{c}(t)} \tag{3.2}$$

$$G^2 = 2 \sum_{t}^{T} c(t) \ln \frac{c(t)}{\hat{c}(t)} \tag{3.3}$$

The X^2 and G^2 statistics correspond to the χ^2 and G-tests, respectively, that are commonly referenced in goodness-of-fit literature. For large sample sizes, X^2 and G^2 lead to similar conclusions in goodness-of-fit testing. Because the distribution of G^2 under the null hypothesis is a closer approximation to the χ^2 distribution, particularly for small samples, it has been increasingly recommended over the X^2 statistic. However, because we will compare the observed test statistic to the empirical approximation of the distribution under the null rather than a theoretical distribution, the differences between the two test statistics in practice are insubstantial.

A limitation of statistics in the power-divergence family is that they both rely on a large sample size but become useless if the sample size is *too* large. For large samples, tests based on power divergence statistics will reject almost any null hypothesis, even when the deviation between the observed and expected values is small. (Simonoff, 2003, Section 4.4) Furthermore, likelihood-based statistics such as G^2 are sensitive to outliers. (Simonoff, 2003, Section 4.7)

3.3.2 f-divergence statistics

In the previous section we considered statistics which measure the divergence between the observed and expected frequency distributions. Now we consider measures of divergence between observed and expected probability distributions.

The class of f-divergence functions was introduced by Renyi (1961) and usually refers to divergences between continuous distributions, although the statistics in this family can be generalized to discrete distributions. An advantage of statistics in the f-divergence class to power-divergence statistics is that they are more robust to outliers.

Arguably the best-known divergence in the f-divergence family is the Kullback-Leibler (KL) divergence, which measures entropy as the divergence of one probability distribution P from another Q. In our notation, we use $\bar{f}$ in place of P for the observed distribution, and $\hat{\bar{f}}$ in place to Q to represent the model, or expected, distribution. KL divergence was introduced as a measure of relative entropy and can be interpreted as the expected difference in the logarithmic probability of an event between distributions $\bar{f}$ and $\hat{\bar{f}}$, given that the event has true (observed) probability distribution $\bar{f}$. For discrete distributions defined over a sample space of K possible outcomes, the KL divergence is defined as

$$D_{\mathrm{KL}}(\bar{f}||\hat{\bar{f}}) = \sum_{t}^{T} f(t) \log \left(\frac{f(t)}{\hat{f}(t)} \right). \tag{3.4}$$

The terms $f(t)$ and $\hat{f}(t)$ denote the observed and expected probability mass in the tth category, respectively.

KL divergence is not symmetric, e.g. $D_{\mathrm{KL}}(\bar{f}||\hat{\bar{f}}) \neq D_{\mathrm{KL}}(\hat{\bar{f}}||\bar{f})$ and has a range $[0, \infty)$. Additionally, the KL divergence may be inflated when $\hat{f}(t)$ is small for some t. An alternate f-statistic is the squared-Hellinger distance, which is both symmetric and bounded by $[0, 1]$.

The squared-Hellinger distance between two discrete distributions $\bar{f}$ and $\hat{\bar{f}}$ is defined as

$$D_{\mathrm{HS}} = \frac{1}{2}\sum_{t}^{T}(\sqrt{f(t)} - \sqrt{\hat{f}(t)})^2. \tag{3.5}$$

The squared-Hellinger distance is equal to the squared Euclidean distance between $\sqrt{\bar{f}}$ and $\sqrt{\hat{\bar{f}}}$, with $\frac{1}{2}$ acting as the scaling factor that restricts the Hellinger distance to the desired range.

3.3.3 Additional goodness-of-fit metrics

Of common goodness-of-fit criteria for model selection, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) scores are two well-known options that consider model complexity, with smaller scores generally indicating better models. In Section 2.2.1, we presented an approximation of the log-likelihood for a given model in Equation 2.10. For brevity, we notate the pseudo-log-likelihood of the observed data under estimated model $\hat{W}$ as simply $\ell(\hat{W})$. When model-fitting using the proposed revealed preferences procedure from Chapter 2, we choose some model class $\mathcal{W}_c$ and compute the MLPLE as the model $\hat{W} \in \mathcal{W}_c$ which maximizes $\ell(\hat{W})$. The AIC score computation incorporates both $\ell(\hat{W})$ and the model complexity by considering the number of model parameters estimated k.

$$\mathrm{AIC}(\hat{W}) = 2k - 2\ell(\hat{W}). \tag{3.6}$$

Because lower AIC scores are preferable, the k term acts as a penalty for models which have achieved high log-likelihood values by overfitting on multiple nuisance parameters.

BIC is a similar computation which additionally considers the number of sampled units n (either sampled households for stock-stock sampling or sampled individuals for stock-flow sampling) in the data:

$$\mathrm{BIC}(\hat{W}) = \log(n)k - 2\ell(\hat{W}). \tag{3.7}$$

In effect, the penalty imposed by BIC for model complexity is greater than that imposed by AIC when the sample size is greater than 7.

Burnham and Anderson (2004) present a comparison of AIC and BIC scores and their performance in different use cases. AIC is usually suggested for model selection when the goal is prediction, i.e. projecting marital outcomes in a population with changing population characteristics or composition. For more general inference and interpretation purposes, the BIC may be more appropriate. BIC has important theoretical properties when the true data-generating process is represented within the set of candidate models. Because BIC assigns greater penalties to complex models when sample sizes are large, as is often true given two-sided matching data, it should theoretically detect over-parametrized models better than the AIC.

A challenge with both AIC and BIC scores is that the conclusion about the "best" model can be subjective. While we generally prefer models with lower scores, when the difference in the scores of two competing models is small, the choice between models may depend on outside knowledge and other considerations within the context of the research question. In this case, what constitutes a "small" difference may not be defined by an absolute threshold. Burnham and Anderson (2004) suggest soft guidelines for comparing models based on the differences in their AIC scores, proposing that when the difference between the AIC scores of two models is less than or equal to 2, the models may be deemed similar.

3.4 Comparing models by raw deviance score and information gain

Using the deviance metrics just described, we can rank candidate models in terms of usefulness by comparing observed data is to data expected under the model. Given models $\{W_1, W_2, \ldots\}$, we can simply compute deviances $\{d_{W_1}, d_{W_2}, \ldots\}$. A simple analysis might conclude that the model that minimizes d_W is the best one. However, for statistics in the f-divergence and power-divergence classes, when one candidate model class is nested in another, this method favors the more complex class. For example, suppose that $\hat{W}_c$ and $\hat{W}_{c+}$ are the MLPLEs from model classes $\mathcal{W}_c$ and $\mathcal{W}_{c+}$, respectively, for a given sample, and that

$\mathcal{W}_c \subseteq \mathcal{W}_{c+}$. Then it is almost always true that $d_{W_c} \leq d_{W_{c+}}$, and the researcher may almost always choose W_{c+} over W_c.

A slightly more sophisticated analysis will consider the trade-off between decreasing model deviance and increased model complexity. This requires a deviance metric that accounts for model complexity, such as the Akaike information criterion (AIC) or Bayesian information criterion (BIC) score.

Alternatively, we may perform model selection based on a more informal analysis of complexity-deviance trade-off by assessing the decrease in deviation facilitated by a given marital surplus model relative to the *maximum* possible deviance. To compute the maximum possible deviance, we consider a "null" marital surplus model where agents are completely indifferent to all potential partnerships based on the deterministic utility. In other words, the surplus model consists only of a deterministic intercept term and an idiosyncratic component, where the intercept term represents the overall market preference for partnership relative to staying single. [3] After fitting the intercept-only model, the maximum possible deviance $d_{\max}$ is equal to the deviance between the observed data and the data expected under this model.

Let $\mathrm{IG}(W)$ denote the information gain achieved about the data-generating process given model W, relative to a completely naive model. We can decompose $d_{\max}$ into the information gain achieved by Then, we can compute the information gain $\mathrm{IG}(W)$ that model W provides about the data-generating process as the reduction from to maximum possible deviance $d_{\max}$ achieved under model W. We divide $\mathrm{IG}(W)$ by $1/d_{\max}$ to scale the value between 0 and 1. This scaling also allows us to think of $\mathrm{IG}(W)$ as the proportion of the maximum deviance reduced by model W:

$$\mathrm{IG}(W) = \frac{d_{\max} - d_W}{d_{\max}}, \qquad 0 \leq \mathrm{IG}(W) \leq d_{\max}. \tag{3.8}$$

[3] We could also consider a model where agents are completely indifferent to all potential partnerships *and* the choice of remaining single. In this case, the deterministic quantities of the surplus function W_{ij} and the singlehood utility functions U_{i0} and V_{j0} are all fixed at 0, and the model requires no "fitting" because there are no parameters to estimate. In most realistic settings, the choice of "null" model between these two options generally will not impact the conclusion of the analysis, but the intercept-only model is more aligned with traditional conceptualizations of a null model, e.g. when considering F-tests.

The true model W_0 would yield a deviance of 0 between the observed and expected data, so that $\text{IG}(W_0) = 1$. A large $\text{IG}(W)$ corresponds to a smaller deviance between the observed data and the data expected under model W and an overall closer model fit. In general, the smaller the remaining deviance between the observed and expected data given model W, the better we believe the model to be. However, in some cases, a model W_c with a simpler specification may already cover a lot of the information gain. The added benefit of a more complex model W_{c+} may appear to have a high information gain when compared to the maximum possible deviance, but may actually have limited added value when compared to the simpler model. We can think of $\text{rIG}(c+, c) = \text{IG}(W_{c+}) - \text{IG}(W_c)$ as the relative information gain achieved by model W_{c+} compared to W_c. When rIG is small, the researcher may decide that the relative information gain of the more complex model is minimal and that the simpler model is sufficient for their purposes.

There are no firm rules or benchmarks regarding the threshold at which relative information gain becomes small enough that it can be considered "negligible"; this usually depends on the context of the research question. Although the information gain procedure is a more informal way of model selection with consideration of parsimony than the use of AIC and BIC scores, it has better interpretability and can therefore be a useful measure to consider.

3.5 Proposed re-conceptualizations of goodness-of-fit significance testing for two-sided markets

In Section 3.1 I described traditional signifance testing for goodness-of-fit and its limitations when applied to two-sided matching market models. In this section I propose an alternate way to think about goodness-of-fit that considers how well a particular *model class* reflects the true partnership-generating process. The proposed method can be used to refine the set of candidate models to the best choice(s). I first provide a general overview of the test, and then go into detail about the computation of the null distribution of the test statistic, to which the observed test statistic is compared and which is generally difficult to compute analytically. Prior to going into these details, however, it is important to note that this proposed test still suffers from many weaknesses and is only intended as a starting point for developing improved significance testing procedures for RPMs.

In the proposed goodness-of-fit significance test, the null hypothesis asserts that the true partnership surplus model W_0 is a member of some class $\mathcal{W}_{H0}$. The test statistic is the deviance d' between the observed data and the data expected under $\hat{W}_{H0} \in \mathcal{W}_{H0}$, where $\hat{W}_{H0}$ is the optimized model estimate (e.g. MLPLE) from class $\mathcal{W}_{H0}$. Options for computing d' are discussed in Section 3.3. By design, $\hat{W}_{H0}$ is equivalent to the model under the null hypothesis in traditional goodness-of-fit testing. Thus the test statistic d' in this test is equivalent to the test statistic d as defined in Section 3.1.

Next we obtain the distribution of the test statistic D'_0 which would occur if the null hypothesis is true. Each draw from this distribution represents a single instance of the deviation between data observed in a sample drawn from $\hat{W}_0$ and the data expected under model $\hat{W}_s$, where $\hat{W}_s$ represents the MLPLE over $\mathcal{W}_{H0}$ given sample s. Computation of the null distribution of D'_0 is often not feasible and can be estimated empirically, as discussed in Section 3.5.1.

We compare the observed test statistic d' to the distribution under the null hypothesis for D'_0. To proceed with significance testing, we compare the p-value $\mathbb{P}(d' \leq D')$ to some Type I error threshold α and accept or reject the null hypothesis. As previously, when the null hypothesis is true, the p-value theoretically follows a uniform distribution.

3.5.1 Computation of the null distribution of test statistic

The distribution of deviances D' under the null hypothesis is usually difficult to compute analytically and can instead be estimated empirically. (Simonoff, 2003, Section 4.4.3). To do this, a bootstrapping method may be employed. First simulate B draws of n-sized samples from $\hat{W}_0$, where n is the size of the observed sample. By simulating from $\hat{W}_{H0}$, we ensure that all B simulated samples are generated from the hypothesized model class $\mathcal{W}_{H0}$. Define $f_{(b)}$ as the statistic observed in the bth iteration of simulated data - in the case of the matching market, the frequency distribution of household types $\bar{c}$. Now, compute $\hat{W}_{(b)}$ as the MLPLE model from $\mathcal{W}_{H0}$ given the bth simulated sample, and let $\hat{f}_{(b)}$ be the expected distribution under $\hat{W}_{(b)}$.

By repeating this procedure B times, we obtain B observed statistics from the simulated matchings $\{c_{(1)}, c_{(2)}, \ldots, c_{(B)}\}$ and B expected statistics $\{\hat{c}_{(1)}, \hat{c}_{(2)}, \ldots, \hat{c}_{(B)}\}$ given the MLPLEs from the null family over the simulated data. We can now compute the divergence $d'_{(b)}$ between $c_{(b)}$ and $\hat{f}_{(b)}$ for $b \in [1 : B]$. As B grows large, the empirical distribution of $\{d'_1, d'_2, \ldots, d'_B\}$ converges to D'.

3.5.2 Advantages and disadvantages of the proposed goodness-of-fit test

The proposed goodness-of-fit test asks the question, if the model class $\mathcal{W}_{H0}$ includes the true data-generating process for a given sample, how often would we observed a deviation this extreme between the observed data and the data observed under the MLPLE model from that same class? This question is broader than the traditional question of whether the MLPLE model $\hat{W}_{H0}$ itself is correct and more likely to yield useful results. Significance testing is useful for quantifying uncertainty when answering this question and communicating confidence through traditional statistical language, including p-values and confidence levels.

An advantage of significance testing is that theoretically, it should have high sensitivity to over-parametrized models; that is, it should be able to detect cases where the model family can be further restricted. Consider, for example, a market where the true surplus utility model is in the uniform homophily class $\mathcal{W}_{\text{UH}}$:

$$W = W^{\text{UH}}_{ij}(x_i, z_j | \underset{\sim}{\beta}_{\text{UH}}) = \beta_0 + \beta_{\text{UH}} \mathbb{I}\{x_i = z_j\} + \eta_{ij} + \zeta_{ij}. \tag{3.9}$$

The identical surplus utility model can also be written as a member of the differential homophily class $\mathcal{W}_{\text{DH}}$:

$$W^{\text{UH}}_{ij}(x_i, z_j | \underset{\sim}{\beta}_{\text{UH}}) = W^{\text{DH}}_{ij}(x_i, z_j | \underset{\sim}{\beta}_{\text{DH}}) = \beta_0 + \sum_{\mathcal{T}} \beta_{\text{UH}} \mathbb{I}\{x_i = z_j = t\} + \eta_{ij} + \zeta_{ij} \tag{3.10}$$

where $\mathcal{T} = \mathcal{X} \cup \mathcal{Z}$.

Just as $\mathcal{W}_{\text{UH}} \subset \mathcal{W}_{\text{DH}}$, the deviances D'_{UH} that would occur over repeated samples from models in $\mathcal{W}_{\text{DH}}$ are a subset of the deviances W_{DH} that would occur over repeated samples from models in $\mathcal{W}_{\text{DH}}$. Since both classes are technically correct in that they include the true partnership surplus, we might expect that the significance test will usually fail to reject either class under the null hypothesis. However, the distribution of D'_{DH} is narrower than the distribution of the D'_{UH}.

If we compute the same MLPLE for a given data set by fitting models from UH and DIH classes, we would observe the same test statistic d'. However, the distributions of the test statistic under the null hypothesis to which we compare d_{DH} are different. Since D'_{DH} has a narrower distribution, $P(d' \leq D'_{\text{DH}}) \leq P(d' \geq D'_{\text{UH}})$. Therefore, we are more likely to reject the null hypothesis of the DH class than we are the UH class.

I re-iterate here that model specifications like UH, DH and SM are rarely true generating processes themselves, but estimates of a far more complex true surplus function. The proposed significance testing procedure works when the candidate model classes could plausibly include the true match-motivating function, but this scenario is extremely unlikely; in realistic two-sided matching markets that have surplus functions too complex to be reasonably specified. Thus, the procedure may still lead to the rejection of most model classes.

Additionally, in simulation studies, the significance testing using the proposed method appeared to work well when population sizes were small ($N < 1,000$). However, performance deteriorated when samples were drawn from larger populations. Therefore, significance testing using this procedure is only recommended with the population size is small.

3.6 Simulation Studies

The purpose of the simulation studies in this chapter is to evaluate the performance of the different deviance metrics and methods for assessing goodness-of-fit that are proposed in this chapter. I conducted several experiments under different conditions to see how different methods performed in response to such changes. The findings from these experiments have been organized into two separate simulation studies, which we refer to as Simulation Study I and Simulation Study II. Simulation Study I focuses on model selection through comparison of raw deviance statistics, information gain, and information criterion scores. Simulation Study II focuses on significance testing. [4]

All experiments were conducted using synthetic data. In this chapter, I will first describe how I constructed the synthetic data used for the simulation studies and then discuss the procedure for each study in Section 3.6.2. Results and related discussion of the simulation studies are presented in Section 3.7 and Section 3.8.

3.6.1 Synthetic Population Data

The simulation studies use data from synthetically-generated populations. Every synthetic population consists of two distinct sides of the market, female and male. Within each gender, agents come in one of 4 types $\mathcal{T} = \{1, 2, 3, 4\}$. The gender distribution, as well as the marginal distributions of agent type by gender, are parameters of the superpopulation from which simulated populations are realized. These are defined in Table 3.1. To simulate a synthetic population, I use the superpopulation availabilities to sample a random assortment of N individuals. These individuals are in the realized synthetic population.

Within each population, a stable matching is achieved using the Gale-Shapley algorithm, based on either the UH marital surplus utility model W_{UH} described in Table 3.2 or the DH marital surplus utility model W_{DH} described in 3.3. According to model W_{UH}, a partnership between any two individuals results in a marital surplus of -1.5 units, relative to the baseline deterministic utility for remaining single. If the individuals in the partnership have the same type, the deterministic surplus utility generated by the marriage shifts an additional 2.8 units in the positive direction. Formally,

$$W_{\text{UH}}(x_i, z_j | \underset{\sim}{\beta}) = -1.5 + 2.8 \times \mathbb{I}\{x_i = z_j\}. \tag{3.11}$$

[4] As a brief digression, I note that some results from additional simulation studies are not presented in the main body of this chapter as they are still in development and require more work. While I will not detail these experiments formally in this chapter, I include some brief notes on preliminary findings in Appendix B.1.

Table 3.1: Gender and Type Availability Distributions in Superpopulations for synthetic data

	Males		Females	
Type	% Population	% of Males	% Population	% of Females
1	9.00	20.0	8.25	15.00
2	15.75	35.0	16.50	30.0
3	13.50	30.0	19.25	35.0
4	6.75	15.0	11.00	20.0
Total	45.0	100.0	55.0	100.0

Equation 3.11, along with randomly-generated taste-shifters, determines the stable matching in each population in the first set. In this way, I constructed 200 stable matchings motivated by W_{UH} each for populations with size $N \in \{1000, 5000\}$.

I repeated this procedure for synthetic populations in which a stable matching was achieved based on W_{DH}, where

$$\begin{aligned} W_{\text{UH}}(x_i, z_j|\underset{\sim}{\beta}) = &- 1.5 + 3.0\mathbb{I}\{x_i = z_j = 1\} + 2.4\mathbb{I}\{x_i = z_j = 2\} + \\ &2.5\mathbb{I}\{x_i = z_j = 3\} + 3.3\mathbb{I}\{x_i = z_j = 4.4\}. \end{aligned} \tag{3.12}$$

I again generate 200 such synthetic stable matchings for each population size $N \in \{1000, 5000\}$.

In the simulation studies, I denote a single synthetic population using the notation $S^{(b)}_{N,w}$, where $N \in \{1000, 5000\}$ refers to the number of individuals in the population, $w \in \{\text{UH}, \text{DH}\}$ refers to the partnership surplus utility model which motivated the stable matching in the population, and $b \in 1, \ldots, B$ is the index of the simulation. A synthetic population with $w =$ UH has a stable matching motivated by W_{UH} in Equation 3.11, and a synthetic population with $w =$ DH has a matching motivated by W_{DH} in Equation 3.12.

Table 3.2: Parameters of UH partnership surplus utility model W_{UH} for marriage within synthetic populations

Parameter $\underset{\sim}{\beta}_{\text{UH}}$	Value
Intercept	-1.5
β_{UH}	2.8

Table 3.3: Parameters of DH partnership surplus utility model W_{DH} for marriage within synthetic populations

Parameter $\underset{\sim}{\beta}_{\text{DH}}$	Value
Intercept	-1.5
$\beta_{1,1}$	3.0
$\beta_{2,2}$	2.4
$\beta_{3,3}$	2.5
$\beta_{4,4}$	3.3

3.6.2 Procedure

Each simulation study has two parts, the first of which studies populations where matchings were generated based on a UH model, and the second of which studies populations where matchings were generated based on the DH model. For Simulation Study I, these parts are denoted as studies I.i and I.ii, respectively. Similarly, for Simulation Study II, the two parts are denoted as studies II.i and II.ii.

We will begin by described the procedure for Simulation Study I.i. In this study I, I analyze each subset of N-sized populations separately. For the bth synthetic population of size N, I fit models from the uniform homophily and differential homophily classes to obtain MPLEs $\hat{W}_{\text{UH}}(S_{N,\text{UH}})^{(b)}$ and $\hat{W}_{\text{DH}}(S_{N,\text{UH}})^{(b)}$. At each iteration, we compute goodness-of-fit measures of the MLPLE models by using various raw deviance metrics, information gain, and AIC and BIC scores and compare the values for the two hypothesized models. For raw deviance metrics and computation of information gain, we consider four options: the chi-squared divergence, G-squared divergence, squared Hellinger distance, and KL divergence. Since the uniform homophily class is nested in the differential homophily class and the true matching model belongs to both, we expect the raw deviance statistics under $\hat{W}_{\text{DH}}(S_{N,\text{UH}})^{(b)}$ to be lower than the corresponding statistic under $\hat{W}_{\text{UH}}(S_{N,\text{UH}})^{(b)}$. We also expect the DH model estimate to give higher information gain than the UH model estimate, but only slightly, since both models are correct. Lastly, we expect the AIC and BIC scores for the DH model estimates to be better (lower) than the corresponding scores for the UH model. I plot the distribution of the values from the different analyses over the 200 iterations and assess the performance of the different approaches.

For Simulation Study I.ii, I repeat this entire procedure from I.i using synthetic populations that have matchings motivated by W_{DH}. I use corresponding notation for this part of the study. The only change in this study is that I fit the saturated mix model in addition to the differential homophily and uniform homophily models, so that the null hypotheses consider an underparametrized class ($\mathcal{W}_{\text{UH}}$), an overparametrized class ($\mathcal{W}_{\text{SM}}$), and the correct class ($\mathcal{W}_{\text{DH}}$).

In Simulation Studies II.i and II.ii, I attempt to validate the proposed significance test approach. For this study, I only present results for synthetic matching data from populations of size $N = 1,000$ individuals. I use 200 synthetic stable matchings motivated by W_{UH} in study II.i and 200 stable matchings motivated by W_{DH} in study II.ii. In both II.i and II.ii, for each iteration of data in each set, I fit a model from the UH class and a model from the DH class. I compute the deviance using all four metrics for each hypothesized model and then follow the procedure described in Section 3.5 to obtain a p-value for the observed deviance. I plot the distributions of these p-values and assess the performance of the proposed method by computing Type I and Type II error rates under different conditions.

3.7 Results: Simulation Study I

In this section, I present the results from Simulation Study I. While the results are fairly straightforward, I add some analysis and commentary as I introduce different sets of results to track insights gained from the studies.

3.7.1 Simulation Study I.i

For Simulation Study I.i, I have 200 synthetic populations each of sizes $N = 1,000$ and $N = 5,000$ individuals. The composition of all synthetic populations are based on the hyperparameters given in Table 3.1, and a stable matching is simulated in each population based on W_{UH} (Equation 3.11). I then assume census data and fit UH and DH models onto the synthetic populations and compared the fits.

For the first step in assessing goodness-of-fit, I compute the raw deviances between each observed distribution of households and the distribution expected under the MLPLE, using four different metrics for deviance. The distributions of these deviances are shown in Figure 3.1. Each row of the figure shows a different deviance metric, and each column shows the results for a different population size. On each plot, the blue line shows the density of deviances $D'_{N,\text{UH}}(\text{UH})$ when fitting a model from the UH class to $S^{(b)}_{N,\text{UH}}$ for all b, and the red line shows the density of deviances $D'_{N,\text{UH}}(\text{DH})$ when fitting a model from the DH class.

For all deviance metrics and both population sizes, the deviances observed after fitting the UH models are distributed toward higher values than the deviances from the DH models. This makes sense intuitively, as $\mathcal{W}_{\text{UH}} \subset \mathcal{W}_{\text{DH}}$ and the additional parameters in the DH class will almost always result in a closer fitting model.

The magnitudes of the chi-squared and G-squared divergences increase with population size, while the squared Hellinger distances and KL divergences decrease slightly. This is because the former two metrics are calculated based on frequency distributions while the latter two are calculated based on probability distributions. Because the increased population size allows for a closer-fitting model while the scale of the probability distribution remains unchanged, the squared Hellinger distance and KL divergence actually decrease inversely with population size. Although the values and scales of the four deviance metrics are different, their distributions look quite similar within each (N, w). This is particularly true at $N = 5,000$.

In our second step, I assess the relative value of fitting more complex models by looking at information gain scores. While the DH model will produce a closer fit to the data, we know that in reality the UH model is sufficient for the synthetic data in this portion of the study and therefore expect the information gain achieved by the DH model relative to the UH model to be small. I plot the distributions of information gain achieved by each model in Figure 3.2. The values shown in the figure were computed based on raw chi-squared divergences; I found similar results when I computed information gain based on the other three metrics. For brevity, I show only the chi-squared-based information gain scores in this section and relegate the rest to Appendix B.3.

The first row of Figure 3.2 shows information gain results for synthetic populations of size $N = 1,000$, and the second row shows the same for $N = 5,000$. In the plots in the first column, we show the distribution of the information gain achieved by fitting each model to the synthetic population data, relative to the null (intercept-only) model. The left box shows the distribution of information gain achieved by fitting UH models $\mathrm{IG}_N^{(b)}(\hat{W}_{\mathrm{UH}})$, with N referring to the size of the population data and b referring to the simulation iteration. Similarly, the right box shows the distribution of information gain achieved by fitting DIH models $\mathrm{IG}_N^{(b)}(\hat{W}_{\mathrm{DH}})$, with N referring to the size of the population data.

The information gains achieved by fitting the UH model for $N = 1,000$ populations are very high. The average information gain achieved by fitting the UH model $\overline{\mathrm{IG}}_{1000}(\hat{W}_{\mathrm{UH}})$ is approximately 0.96, meaning that the UH model reduces 96% of the maximum possible deviance relative to the naive model. At $N = 5,000$, the average information gain $\overline{\mathrm{IG}}_{5000}(\hat{W}_{\mathrm{UH}})$ increases to 0.988. At both population sizes, while the plots show that the DH models do have slightly higher information gain scores, the distributions of $\overline{\mathrm{IG}}_N(\hat{W}_{\mathrm{UH}})$ and $\overline{\mathrm{IG}}_N(\hat{W}_{\mathrm{DH}})$ show considerable overlap.

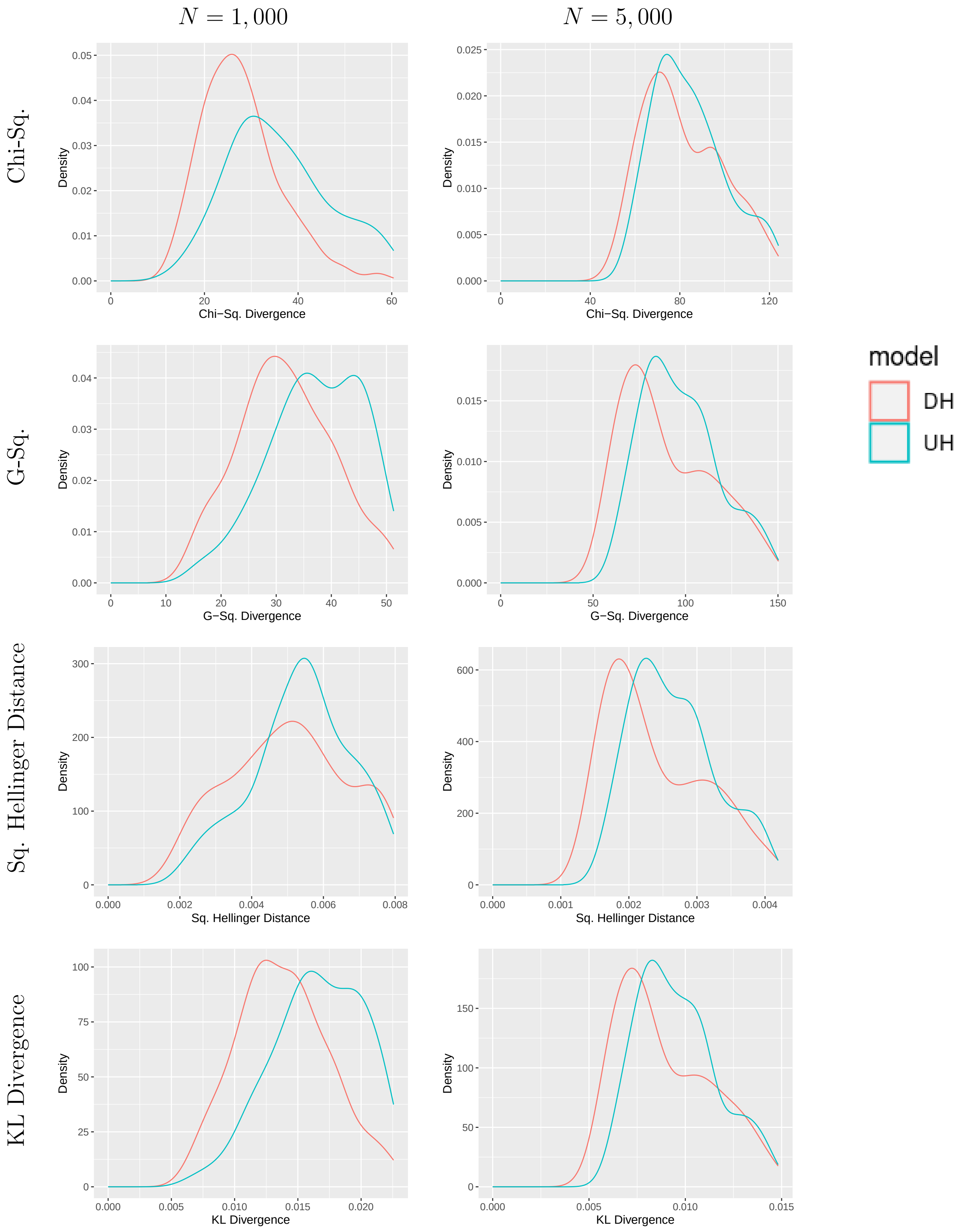


Figure 3.1: Simulation Study I.i: Distribution of raw deviances; true partnership utility model is W_{UH} (200 simulations)

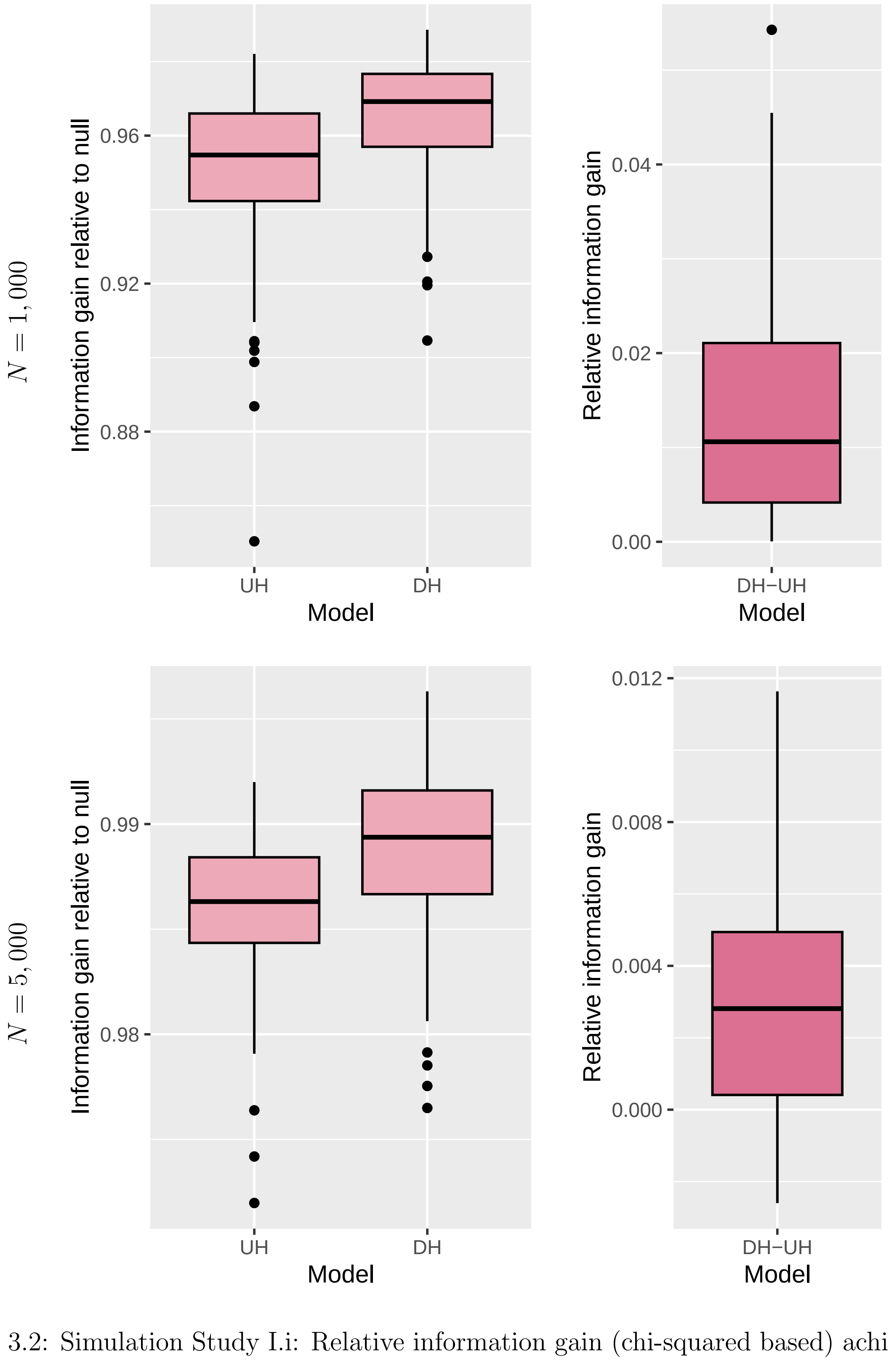


Figure 3.2: Simulation Study I.i: Relative information gain (chi-squared based) achieved by different models; true partnership utility model is W_{UH} (200 simulations)

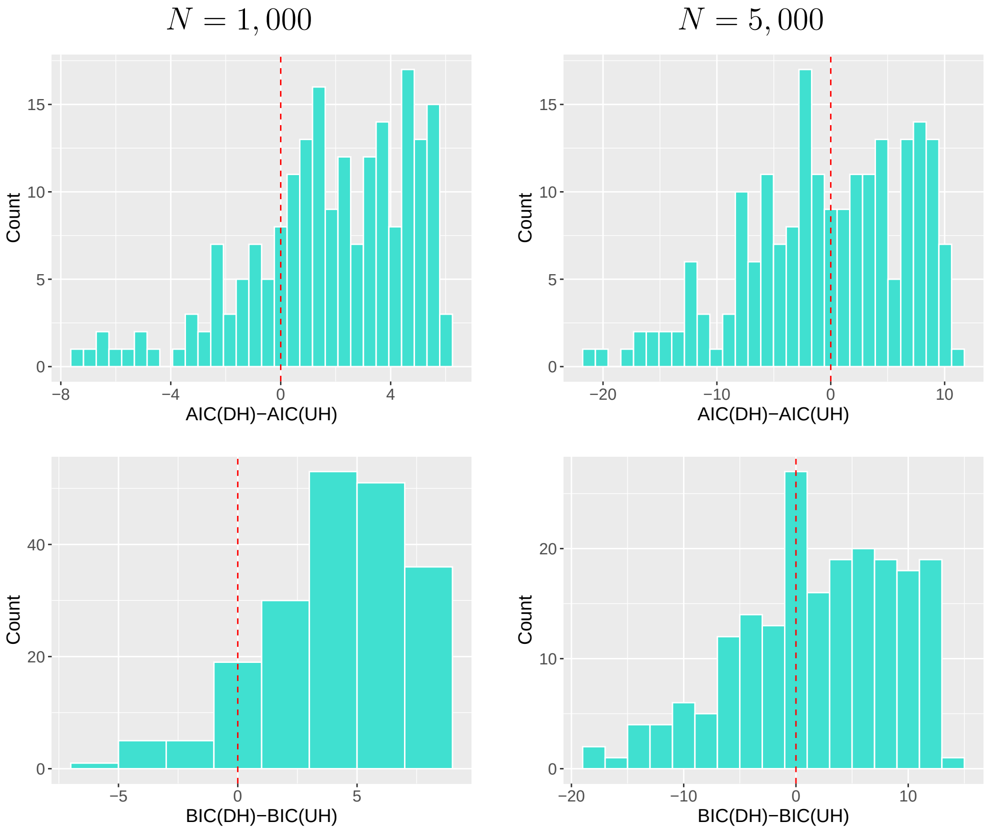


Figure 3.3: Simulation Study I.i: Differences in AIC (top) and BIC (bottom) scores of different models; true partnership utility model is W_{UH} (200 simulations)

In the second column of Figure 3.2 I show the distribution of the relative information gains achieved by fitting the DH model compared to the UH model $\text{rIG}_N^{(b)}(\hat{W}_{\text{DH}}, \hat{W}_{\text{UH}})$ when both models are computed using the same synthetic population data $S_{N,\text{UH}}^{(b)}$. For $N = 5,000$, there was one instance across the 200 iterations in which $\text{rIG}_N^{(b)}(\hat{W}_{\text{DH}}, \hat{W}_{\text{UH}}) < 0$, indicating that the chi-squared divergence between the estimated and observed data in that iteration was actually lower when fitting the UH model relative to the DH model. The actual relative information gain for this iteration is -0.002. It is possible that bias-correction, which has some randomness due to the empirical bootstrapping procedure it employs, perturbed the model count matrices in a way that led the DH model to have a (very slightly) poorer fit than the UH model.

In general, the relative information gain is very small for the DH model relative to the UH model, with $\overline{\text{rIG}}_{1000}(\text{DH}, \text{UH}) = 0.011$ and $\overline{\text{rIG}}_{5000}(\text{DH}, \text{UH}) = 0.0024$. An increase in the population size by a factor of 5 corresponds to a decrease in relative information gain by a factor of a little over 1/5 ($\frac{\overline{\text{rIG}}_{5000}(\text{DH,UH})}{\overline{\text{rIG}}_{1000}(\text{DH,UH})} = 0.218$), although confirming this relationship would require further study.

Notably, while the IG distributions in Figure 3.2 have some low outliers, there are no such outliers in the rIG distributions for either population (though there is a high outlier in the distribution of $\text{rIG}_{1000}(\text{DH}, \text{UH})$). This implies that even when the information gain for the models for a particular iteration is near the tail of the distribution, the relative information gain remains generally consistent.

The third step of the results is a more formal model comparison. Since we know that in truth the UH model is sufficient for the populations in this portion of Simulation Study I and the criterion scores penalize added model complexity, we expect better (lower) AIC and BIC scores from the UH model relative to the DH model; equivalently, the difference in scores (DH $-$ UH) should be positive. In Figure 3.3 I show the histograms of observed AIC and BIC differences of 200 simulations for each population size. The top row shows the difference in AIC scores while the bottom row shows the difference in BIC scores. The left column shows the distribution of differences for synthetic populations of size $N = 1,000$ while the right column shows the same for $N = 5,000$. The dotted red line on each plot marks where the difference equals 0. We expect most of the distribution to fall to the right of this line.

In all four plots in Figure 3.3, a considerable portion of the distributions are in fact on the left side of the red line. The error rate, or rate at which the direction difference in criterion scores is opposite to what we would expect, varies widely across the four plots, with the score difference (DH-UH) less than 0 between 11% and 58% of the time depending on the criterion used and the population size. For both population sizes, the BIC scores gave the expected result more often than the AIC scores, which makes sense because the BIC assigns higher penalties for additional terms. For both AIC and BIC, the mean of the distribution stays consistent as population size changes, while the variance increases with population size.

3.7.2 Simulation Study I.ii

I now present a similar set of results for synthetic populations in which stable matchings were achieved based on W_{DH} (Equation 3.12). I again generate 200 populations each of size $N = 1,000$ and $N = 5,000$ and assume census data. This time I fit UH, DH, and SM models onto the synthetic populations and compared the fits of all three.

I follow the first step of the goodness-of-fit analysis by showing the distributions of the raw deviances between the observed and expected distributions of households and the distribution expected under the model class MLPLE in Figure 3.4. Once again, each row of the figure corresponds to a different deviance metric, and each column shows the results for a different population size. On each plot, the blue line shows the density of deviances $D'_{N,\text{DH}}(\text{UH})$ when fitting a model from the UH class to $S^{(b)}_{N,\text{UH}}$ for all b, the red line shows the density of deviances $D'_{N,\text{DH}}(\text{DH})$ when fitting a model from the DH class, and the green line shows the density of deviances $D'_{N,\text{DH}}(\text{SM})$ when fitting a model from the SM class.

The plots show that for all population sizes and deviance metrics, the variance of $D'_{N,\text{DH}}(\text{SM})$ is far lower than the corresponding variances of $D'_{N,\text{DH}}(\text{UH})$ and $D'_{N,\text{DH}}(\text{DH})$ Within each population size-deviance metric cell, as model complexity (the number of parameters to be estimated) increases, the raw deviances decrease. For the chi-squared and G-squared statistics, the means of the distributions of each model increase with population size. For the squared Hellinger distance and KL divergence the means decrease very slightly with the increased population size. These shifts mirror the ones seen in Figure 3.1 for Simulation Study I.i.

The second step of the results are the distributions of IG and rIG scores, shown in Figure 3.5. The right column shows the information gain scores for the three hypothesized models. In general, the models appear to provide more information gain when $N = 5,000$ as compared to $N = 1,000$. The UH model, which is from the simplest of the three candidate classes considered, achieves average IG scores of 0.942 and 0.965 for $N = 1,000$ and $N = 5,000$, respectively. While these scores are quitehigh, they are lower than the corresponding scores in Simulation Study I.i. This makes sense since in study I.i the UH model actually was the true model, whereas in this study the UH model is underparametrized for the data. The information gain relative to the null model increases with model complexity. The SM model seems to achieve a near-perfect fit and has an extremely small variance, although it does also have a relatively large number of outliers. With the increase in population size, the variance of the information gain distributions for all three models decreases.

The left column of Figure 3.5 shows the relative information gain for increasingly complex models. The first boxplot shows the distribution of the relative gain from the UH to the DH model $\mathrm{rIG}_N(\mathrm{DH}, \mathrm{UH})$, and the second shows the relative gain from DH to SM $\mathrm{rIG}_N(\mathrm{SM}, \mathrm{DH})$. The mean relative gain achieved by the DH model compared to the UH model is 0.310 for both $\mathrm{rIG}_{1000}(\mathrm{DH}, \mathrm{UH})$ and $\mathrm{rIG}_{5000}(\mathrm{DH}, \mathrm{UH})$. This value is larger than the information gain seen for the same pair of models in Simulation Study I.i but is still somewhat smaller than expected. The small relative information gain from UH to DH might be explained by the high IG already achieved by the UH model. The UH model's high IG scores despite being underparametrized in this part of the study may be due to the fact that although the true preferences $\beta_1, \beta_2, \beta_3$ and β_4 are all different from each other, the overall difference between the parameters may be small enough that the true model W_{DH} is perhaps just outside the boundary of the UH class. Thus, the UH model is still able to fit the data fairly well.

At $N = 1,000$, the tail of the distribution of $\mathrm{rIG}_{1000}(SM, DH)$ is actually just below 0, implying that the SM model is actually leading to information loss instead of gain relative to the DH model. Similar to the explanation in Section 3.7.1, the considerable number of rIG values at $N = 1,000$ which fall below 0 may be an artifact of bias correction, as well as the small population size possibly leading to sparse distributions and greater instability in estimation for the saturated model.

When the population size increases to $N = 5,000$, the variance in the distributions of both $\mathrm{rIG}_N(\mathrm{DH}, \mathrm{UH})$ and $\mathrm{rIG}_N(\mathrm{SM}, \mathrm{DH})$ decreases. While the average relative information gain from the UH to DH model $\overline{\mathrm{rIG}}_N(\mathrm{DH}, \mathrm{UH})$ stays stable with the population increase from $N = 1,000$ to $N = 5,000$, $\overline{\mathrm{rIG}}_N(\mathrm{SM}, \mathrm{DH})$ decreases considerably, from $\overline{\mathrm{rIG}}_{1000}(\mathrm{SM}, \mathrm{DH}) = 0.0284$ to $\overline{\mathrm{rIG}}_{5000}(\mathrm{SM}, \mathrm{DH}) = 0.0059$. As the population size increased by a factor of 5, the average

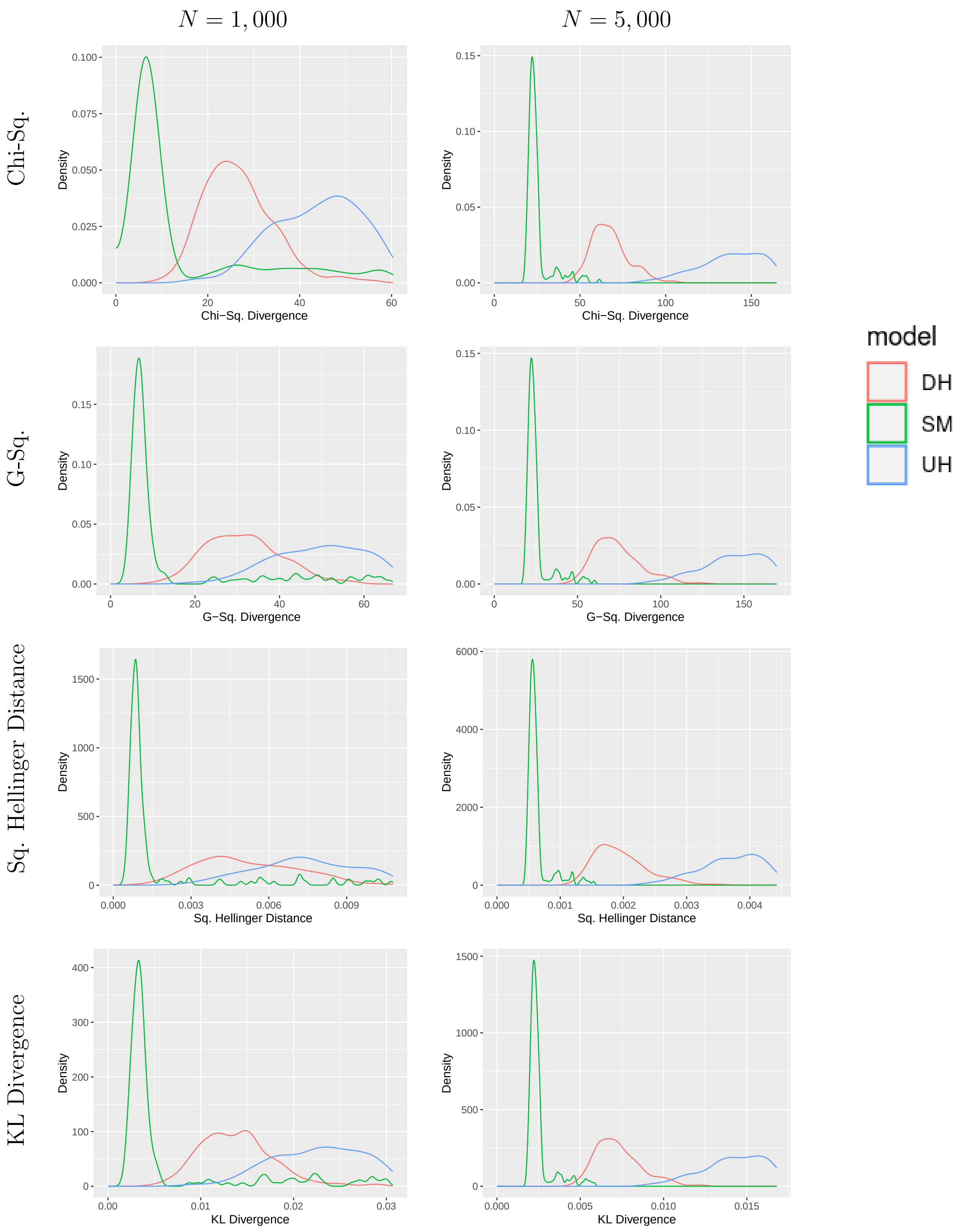


Figure 3.4: Simulation Study I.ii: Distribution of raw deviances; true partnership utility model is W_{DH} (200 simulations)

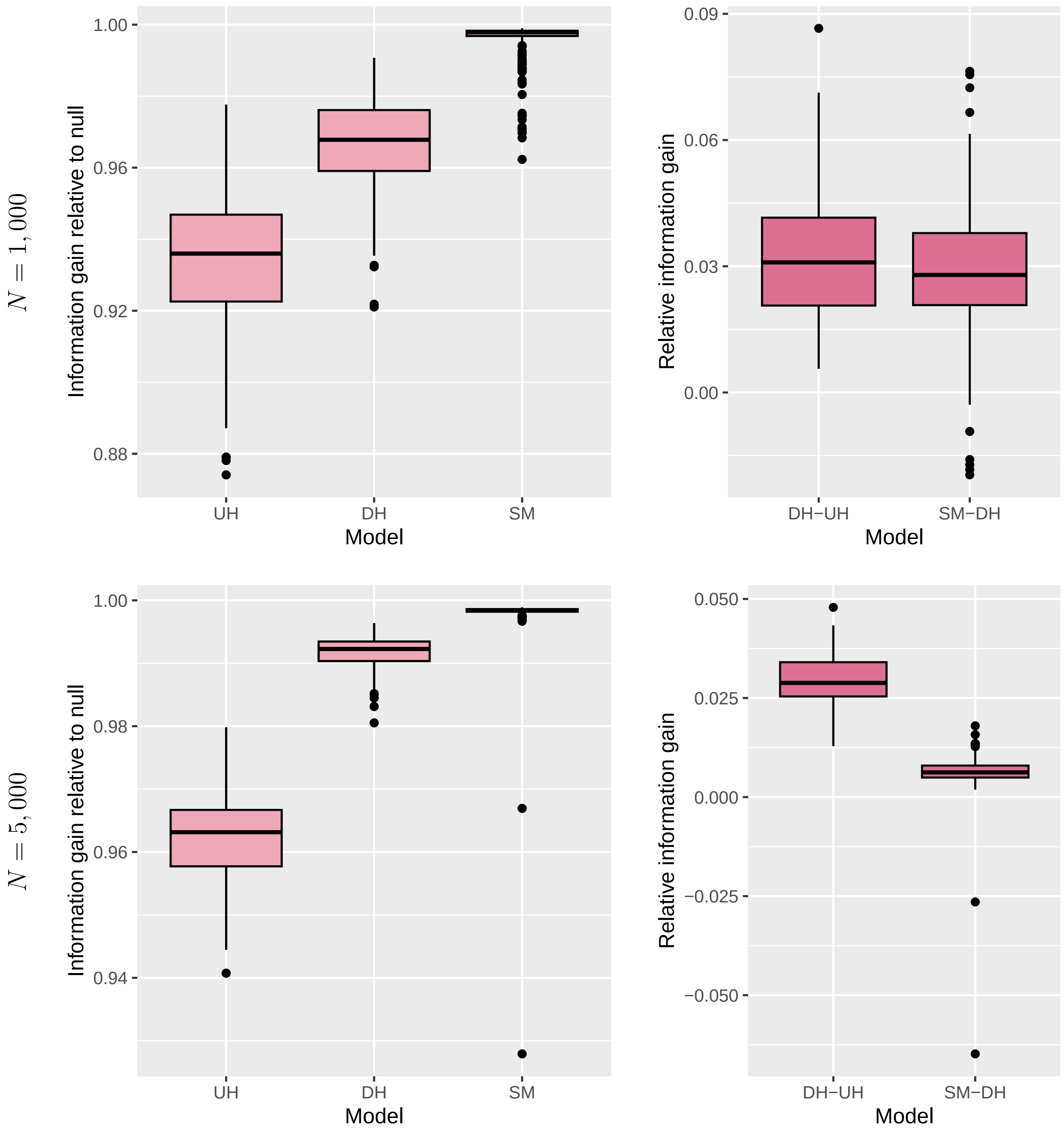


Figure 3.5: Simulation Study I.ii: Relative information gain (chi-squared based) achieved by different models; true partnership utility model is W_{DH} (200 simulations)

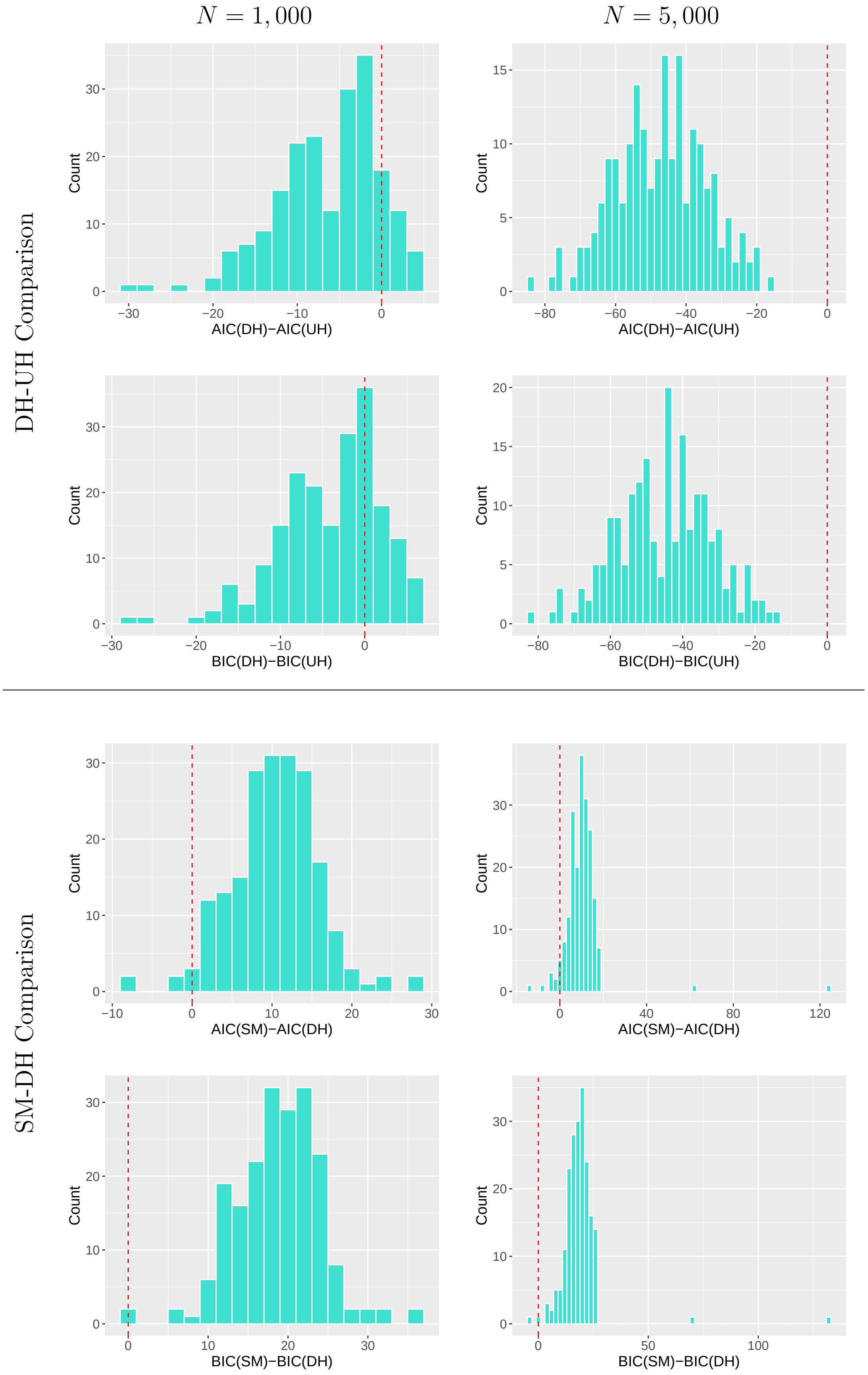


Figure 3.6: Simulation Study I.ii: Differences in AIC and BIC scores of different models; true partnership utility model is W_{DH} (200 simulations)

rIG for the oversaturated model decreased by a factor of slightly over 1/5 ($\frac{\overline{\text{rIG}}_{5000}(\text{SM,DH})}{\overline{\text{rIG}}_{1000}(\text{DH,UH})} = 0.227$). This is similar to the ratio observed in the corresponding results for Simulation Study I.i in Section 3.7.1. However, we reiterate that further study would be required to establish a formal relationship.

In the final part of this portion of Simulation Study I.ii, I present the differences in AIC and BIC scores for the MLPLEs of the different model classes. The distributions of these differences are shown in Figure 3.6. The top panel of four plots shows the distributions of differences in criterion scores when comparing the UH and DH models, with AIC in the top row and BIC in the bottom row. The bottom panel of four panels are organized the same way to compare the DH and SM models. I calculate AIC differences for these comparisons as AIC(DH) − AIC(UH) and AIC(SM) − AIC(DH), respectively, so that the difference is always taken from the model with more parameters. We do the same for BIC. Again, the plots in the right column show criterion scores when the models are fit to populations of size $N = 1,000$ and those on the left show the same for $N = 5,000$.

In the comparison of criterion scores of the UH and DH models, we expect the model from the DH class to have a better (lower) score than the UH model and therefore should observe negative differences. The criterion scores appear to achieve this, with over 90% of the observed differences for both AIC and BIC less than zero when the synthetic populations are of size $N = 1,000$, and 100% of both scores less than zero at $N = 5,000$. For each population size, the two scores appear to perform similarly.

When comparing the criterion scores for the DH and SM models, we expect to see a positive difference since the scores for the over-parametrized SM models should be worse (higher) than the scores for the DH models at the same iteration. For both AIC and BIC and at both $N = 1,000$ and $N = 5,000$, the difference in criterion scores for the SM and DH models is positive in at least 95% of simulations.

The high success rate of the criterion scores in correctly rejecting the SM model when comparing it to the DH model is encouraging. A natural question might be why the criterion scores have a considerably lower success rate when comparing the DH and UH models in this study at $N = 1,000$. This is in part because as the number of agent "types" in the population goes up, the difference between the SM and DH models becomes much greater than the difference between the DH and UH models. The UH model always has exactly two preference parameters requiring estimation. For a population with agents of K types, the DH model will have $K-1$ more parameters than the UH model. The SM model in turn will have $K^2 - K - 1$ more parameters than the DH model. The greater distance between the SM and DH models makes the difference in their performance easier to detect.

3.8 Results: Simulation Study II

The purpose of Simulation Study II is to assess the validity of the proposed significance testing procedure. I compute p-values for the raw deviance statistics computed given the observed synthetic data and hypothesized model distribution. All synthetic populations in this study consist of exactly $N = 1,000$ individuals and are individually sampled from the hyperpopulation with availabilities as described in Table 3.1. I present the results for this study in two parts: Simulation Study II.i shows results when the true partnership surplus model is W_{UH}, and Simulation Study II.ii shows results when the true partnership surplus model is W_{DH}.

3.8.1 Simulation Study II.i

I first show the results for study II.i. The distributions of the p-values are shown as a set of eight plots in Figure 3.7. The plots in the left column show p-values when $\mathcal{W}_{\text{UH}}$ is the hypothesized class, and those in the right column show the p-values when $\mathcal{W}_{\text{DH}}$. The four plots in each column correspond to the four power divergence and f-divergence deviance metrics discussed in this chapter which I used to compute p-values. Each row corresponds to a different metric.

The vertical dashed red line on each plot indicates the threshold where $p = 0.05$. The density to the left of this threshold is equal to the proportion of times we would reject the null hypothesis at $\alpha = 0.05$. If the null hypothesis is correct and the significance testing procedure is working as expected, 5% of the p-values will be less than or equal this value. The red label at the top-right corner of every panel gives the proportion of times over the 200 simulations where p was computed as less than 0.05.

Since the UH class is the correct family of the marital surplus utility function in this part of the study, we expect the p-values under this null hypothesis to follow a uniform distribution. However, the resemblance of all four p-value distributions to the uniform distribution is, at best, weak. Furthermore, when the null hypothesized class is the over-parametrized DH class, the distribution of the p-values actually looks slightly *more* uniform.

Based on the distributions of the p-values in Figure 3.7, we can conclude that the proposed significance testing procedure performed quite poorly overall in this portion of the study. Still, it is worth commenting on some other features of the results. Under the UH null hypothesis, the p-values from the chi-squared, G-squared, and KL statistics all appear to follow a similarly shaped right-skewed distribution. However, the p-value distribution based on the squared Hellinger distance looks markedly different, following a more bell-curved shape. This is also true under the DH class null hypothesis – while the p values based on the other statistics under this hypothesis appear to be clustered mostly below 0.5, the p-values computed based on the Hellinger distance under $H_0 : \mathcal{W}_{\mathrm{DH}}$ again follow a bell curve distribution.

The different shape of the Hellinger-based p-value distributions under both null hypotheses may be attributed to the fact that the Hellinger distance itself is constrained between 0 and 1, whereas p-values are generally computed for test statistics that are unbounded on at least one side. Thus, it may be best to rule out the squared Hellinger distance as a choice for test statistic for significance testing RPMs. We do not discuss it further for this portion of the simulation study.

For completeness, we can look at $P(p \leq 0.05)$ in these plots. At $\alpha = 0.05$ we expect Type I error rate, or the probability of rejecting a correct hypothesis, to be 5%. In reality, however, we see that for the chi-squared, G-squared, and KL-based p-values under the UH null hypothesis, the Type I error rate is much larger, ranging from 0.115 to 0.315.

Conversely, the probability of rejecting the DH null hypothesis at $\alpha = 0.05$ ranges between 0.055 and 0.065 depending on the metric used for the test statistic. This finding is interesting because of the persisting question of how the proposed significance test performs when the class under the null hypothesis is overparametrized. On the one hand, the true data-generating model W_{UH} is in fact the DH class, so the null hypothesis is technically correct. On the other hand, as discussed in Section 3.5.2, the distribution of deviances between estimates from most models in the DH class and the data actually observed based on those models should be quite narrow and therefore results in a smaller p-value.

3.8.2 Simulation study II.ii

In this study, the stable matchings in the simulated populations were motivated by W_{DH}. Similar to study II.i, for each model fit and deviance metric, I show the distributions of the 200 simulated p-values, producing the 8 panels shown in Figure 3.8. The $p-$values based on the UH null hypothesis are in the left column, and those based on the DH null hypothesis fit are in the right column.

Though not quite uniform, the p-values under the DH null are closer to the uniform distribution than those under the UH null. Within each null hypothesis, the distribution of p-values looks similar for the chi-squared, G-squared, and KL divergence metrics but again looks slightly different for the squared Hellinger distance. Under the UH null hypothesis, the distribution of the p-values based on the squared Hellinger distance declines less steeply than the distributions from all the other metrics under the same null hypothesis. Under the DH null hypothesis, the p-values based on the squared Hellinger distance follow a distribution resembling a normal distribution but with a plateau at the peak, whereas the other distributions under that null have much milder peaks and are more right-skewed. Due to the continued inconsistent results produced in significance testing using the squared Hellinger distance relative to the other metrics, I reiterate that it should generally not be used for significance testing.

With the other metrics, I assess error rates. Under the DH null hypothesis, if we assign $\alpha = 0.05$ the empirical Type I error rate is between 0.055 and 0.065, depending on the metric used, over 200 simulations. This aligns fairly closely with what we would expect and is encouraging. However, this comes with the caveat that the observed Type I error rate here only aligns with the expected value at $\alpha = 0.05$; since the distribution of the p-values is not truly uniform, if we increased our α threshold to 0.1, we would likely observed a Type I error rate higher than the expected rate of 10%.

The error rates under the UH null hypothesis on the other hand are quite poor, especially considering that the true stable matching-motivating model is not a member of the UH class. The test fails to reject the incorrect null hypothesis between 68% (based on the chi-squared statistic) and 90% (based on the G-squared statistic) of the time.

3.9 Visual tools for more in-depth goodness-of-fit analysis

In assessing goodness-of-fit for revealed preferences models, a researcher may be interested in how much a model reduces the overall lack-of-fit of the expected data to the observed and how well lack-of-fit is reduced in specific areas of the distribution.

I explored several different options for deviance metrics in Section 3.3. While the interpretations and underlying theoretical properties of these statistics differ, we note the power-divergence and f-divergence statistics for measuring the deviance between distributions P and Q have computations that can be written in the form

$$D(P, Q) = \sum_t d(P(t), Q(t)) \tag{3.13}$$

In other words, there is a natural way to decompose the deviance, or lack-of-fit, into the contributions from each of $t \in \{1, \ldots, T\}$ household categories, where $T = X \times Z + X + Z$. As T increases, the job of comparing these contributions against each other based on the raw numerical data and identifying any patterns therein becomes cumbersome. For this reason, I propose an intuitive visual tool to help researchers evaluate and compare the goodness-of-fit

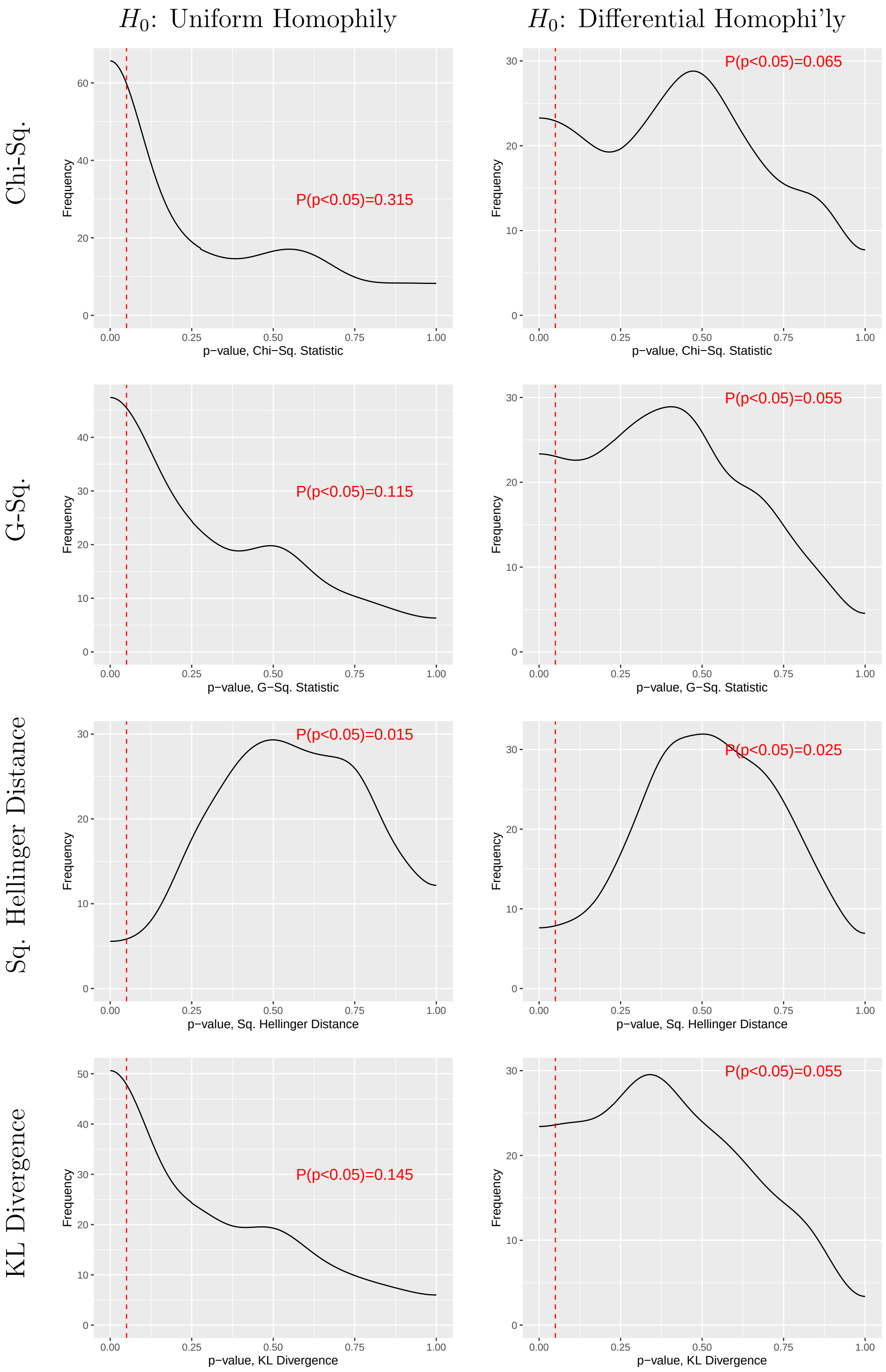


Figure 3.7: Simulation Study II.i: Distribution of p-values; true partnership utility model is W_{UH} (200 simulations)

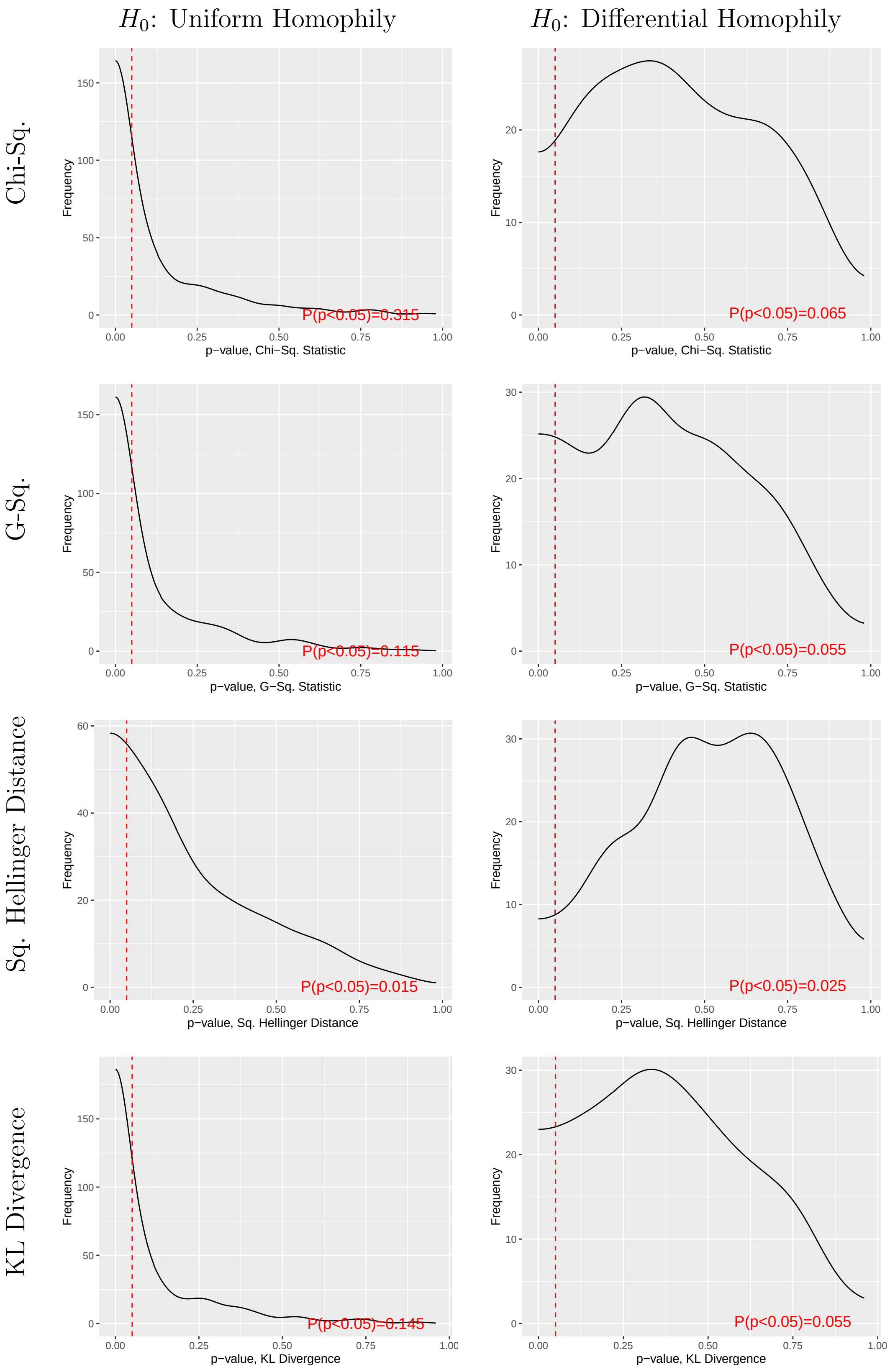


Figure 3.8: Simulation Study II.ii: Distribution of p-values; true partnership utility model is W_{DH} (200 simulations)

		Male's (Female's Partner's) Type					
		1	$\dots$	z	$\dots$	Z	Single
Female's (Male's Partner's) Type	1	$c(1,1)$					$c(1,*)$
	$\vdots$		$\ddots$				$\vdots$
	x			$c(x,z)$			$c(x,*)$
	$\vdots$				$\ddots$		$\vdots$
	X					$c(X,Z)$	$c(X,*)$
	Single	$c(*,1)$	$\dots$	$c(*,z)$	$\dots$	$c(*,Z)$	0

Figure 3.9: Frequency distribution $\bar{c}$ for matching outcomes represented as a $(X+1) \times (Z+1)$ contingency table

for different models at a more granular level.

Rather than treating the frequency distribution $\bar{c}$ (or $\bar{f}$) as a T-length vector, it is helpful to visualize it as a $(X+1) \times (Z+1)$ contingency table, as shown in Figure 3.9. Across the first X rows and Z columns, the (x, z)th element holds $c(x, z)$, representing the frequency of partnerships between a type x woman and type z man in the data. The xth element of the $(Z+1)$th column is $c(x, *)$, representing the frequency of type x women who are single. Similarly, the zth element of the $(X+1)$th row is $c(*, z)$, the frequency of type z men who are single. The $(X+1, Z+1)$th element is a structural 0 because a partnership between a single man and a single woman is, by definition, impossible. Thus, the table has $\mathcal{T}$ cells. The observed matching can be represented in a table this way for any sampling design.

Since the frequency distribution over these outcomes can be decomposed as

$$\bar{c} = \{c(x,z), c(x,*), c(*,z); x \in \mathcal{X}, z \in \mathcal{Z}\}$$

(and similarly for $\bar{f}$), the deviance in Equation 3.13 can be rewritten as

$$D(\bar{c}, \hat{\bar{c}}) = \sum_{x}^{\mathcal{X}} \sum_{z}^{\mathcal{Z}} d(c(x,z), \hat{c}(x,z)) + \sum_{x}^{\mathcal{X}} d(c(x,*), \hat{c}(x,*)) + \sum_{z}^{\mathcal{Z}} d(c(*,z), \hat{c}(*,z)). \quad (3.14)$$

When decomposed in this way, we can more intuitively visualize $d_{xz} = d(c(x,z), \hat{c}(x,z)), x \in \mathcal{X}, z \in \mathcal{Z}$ as a $X \times Z$ grid, where the (x, z)th item represents the divergence between the observed and expected proportions of the units consisting of a partnership between a type x woman and type z man. We then represent the divergences contributions from singleton households $d_{x*} = d(c(x, *), \hat{c}(x, *)), x \in \mathcal{X}$ and $d_{*z} = d(c(*, z), \hat{c}(*, z)), z \in \mathcal{Z}$ as X-vector and Z-vectors on the margins, representing single women and men respectively. Together, the partnership grid and singleton vectors are treated like a heat map, so that the outcomes with the darkest coloring correspond to the areas of the distribution with the poorest fit.

This tool can be applied using any of the deviance metrics discussed in Section 3.3, although some choices have more intuitive interpretations than others. The additive components of the squared Hellinger distance, for example, are each bound between 0 and 1 when the scaling factor $\frac{1}{2}$ is excluded from the calculation. The squared Hellinger contribution from the kth category is 0 if the expected and observed proportions of outcomes in that category are equal. Conversely, the squared Hellinger distance for the kth category is 1 when the observed proportion is 1 and the expected proportion is 0, or vice versa, in that category. The decomposed squared Hellinger distance therefore allows comparison not only of the contributions to the deviance from different household types relative to each other, but also the contribution from a single category to the minimum and maximum possible values.

The KL divergence is also a good measure for decomposing deviance contributions. While the total KL divergence between two contributions must be non-negative, the contributions from each cell have an unrestricted domain across $\mathbb{R}$. A negative KL divergence contribution at the tth category implies a negative log-likelihood ratio at that category, indicating that the observed density at that outcome $P(t)$ is less than the expected density $Q(t)$. Similarly, a positive value implies that $P(t) > Q(t)$. The KL-divergence equals 0 when $P(t) = Q(t)$. Thus the KL-divergence allows investigation not only of where the lack-of-fit has the greatest magnitude, but also into the direction of the lack-of-fit at each location.

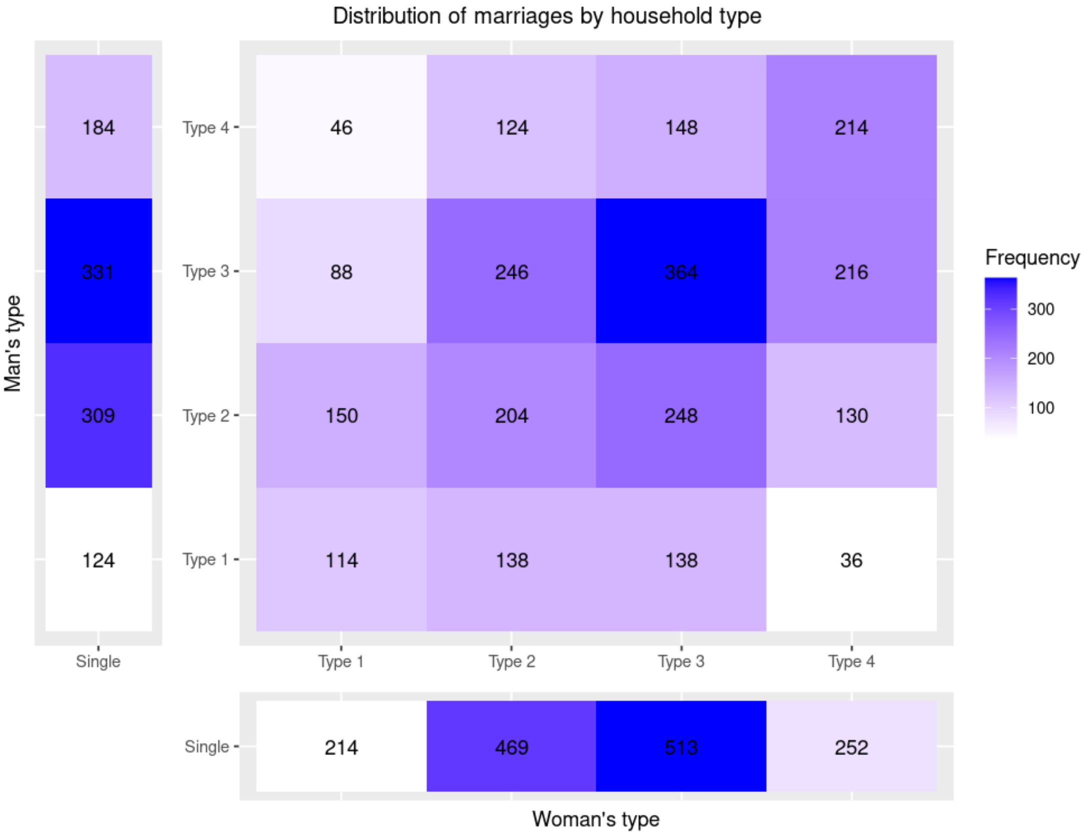


Figure 3.10: Synthetic Population with matching from SM model

3.9.1 Example

As an example, suppose we observe a (synthetic) closed two-sex population of size $N = 5,000$, where each women and men have sets of discrete types $\mathcal{X}$ and $\mathcal{Z}$, such that $\mathcal{X} = \mathcal{Z}$. The observed frequencies of partnership by type for a synthetic population following this set-up is visualized in Figure 3.10. Based on the frequency distribution, partnerships tend to be more common between individuals of the same type, indicating a preference for homophily.

I first fit a DH model to the data and compute deviance between the observed and model frequency distributions in terms of squared Hellinger distance. The categorical decomposition of the squared Hellinger distance under the DH model is shown in Figure 3.11, with the sum total equal to 20.6×10^{-4}. From the figure, among partnered households the contribution to the total deviance is relatively low from household categories where both individuals are the same type. The lack of fit contribution also tends to be greater in categories with low observed or expected probabilities.

To address the relatively large lack-of-fit from partnered household categories where the man and woman have different types, we may then try fitting a model from the SM class. The deviance decomposition of the squared Hellinger distance is shown in Figure 3.12. With the SM model, the total deviance is reduced by a factor of 10. Furthermore, the lack of fit across all categories is now very small. Notably, although $d(1,1)$ and $d(2,2)$ are slightly greater under the SM model than under the DH model, the increased deviance contribution in these categories is more than offset by the reduced error in the others. In fact, the synthetic population in Figure 3.10 over which we are assessing goodness-of-fit was actually matched based on an SM partnership utility function.

I repeated this analysis of lack-of-fit using KL divergence. The KL divergence decompositions for the DH model fit and the SM model fit are shown in Figures 3.13 and 3.14, respectively. In Figure 3.13, the KL divergence contribution $d(4,3)$, representing partnerships between a Type 4 woman and a Type 3 man, has a high magnitude in the positive direction, while the contributions from all other partnered households including a Type 4 woman are all negative. This implies that the DH model underestimates the number of partnerships between Type 4 women and Type 3 men and overestimates all other partnerships involving Type 4 women. When the SM model is fit, not only does the overall deviance reduce by a factor of nearly 8, but the deviance contributions from all partnered outcomes involving a Type 4 woman also decrease in magnitude.

We can supplement these decomposition visualizations with a table showing information gain statistics. Table‘3.4 shows the information gain attained by fitting the UH, DH, and SM models to the data in Figure 3.10

As shown in this example, the visual decomposition allows intuition for how a hypothesized model can be improved, based on the areas of the distribution with the greatest lack-of-fit. However, even if two different models have similar overall deviances, researchers may be more interested in accurately estimating matching behavior of specific types of individuals or more closely modeling the number of individuals who remain single. In this

Table 3.4: Information gain (relative to null) from fitting different models, given population data shown in Figure 3.10 with true matching motivated by SM model

	Sq. Hellinger		KL Divergence	
Model	Deviance	IG	Deviance	IG
Null	0.0064	-	0.026	-
UH	0.0051	0.206	0.020	0.231
DH	0.0023	0.641	0.0091	0.650
SM	0.00026	0.959	0.0010	0.962

case, the visual tool helps differentiate two similarly performing models by identifying the outcomes with the greatest error contributions and allows researchers to select models based on the priorities of the research question at hand.

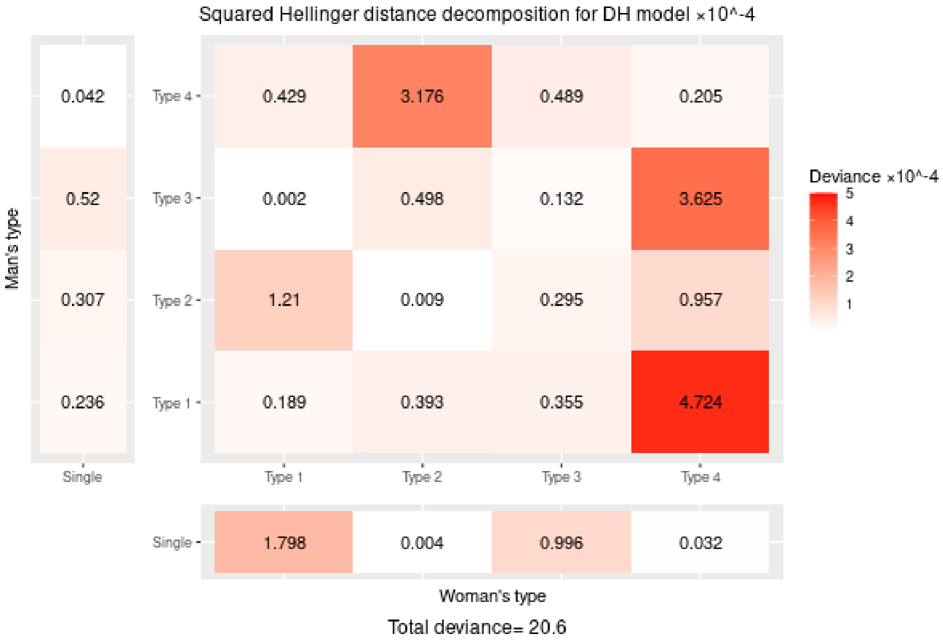


Figure 3.11: Sq. Hellinger distance decomposition when fitting DH model to data from SM model

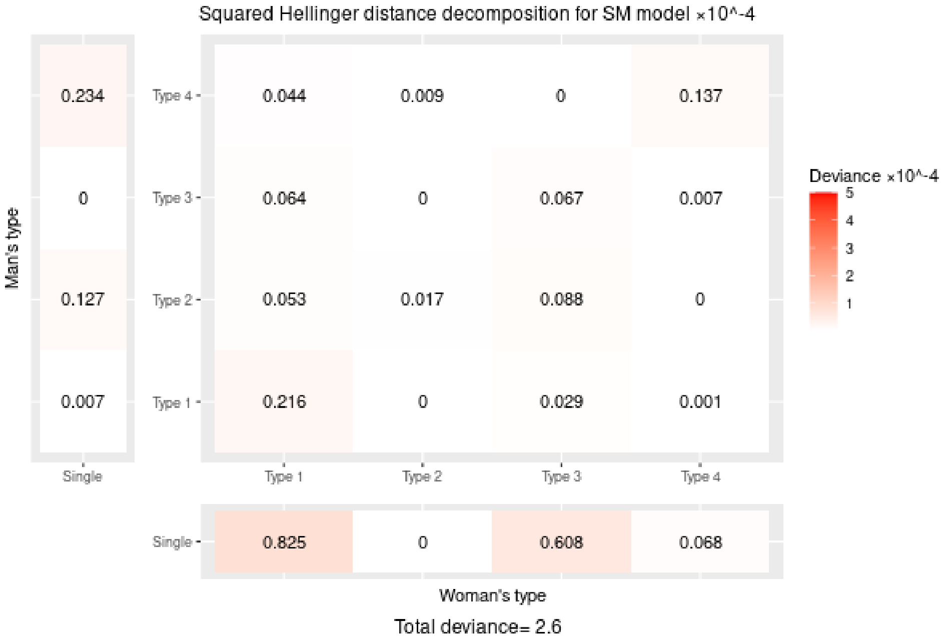


Figure 3.12: Sq. Hellinger distance decomposition when fitting SM model to data from SM model

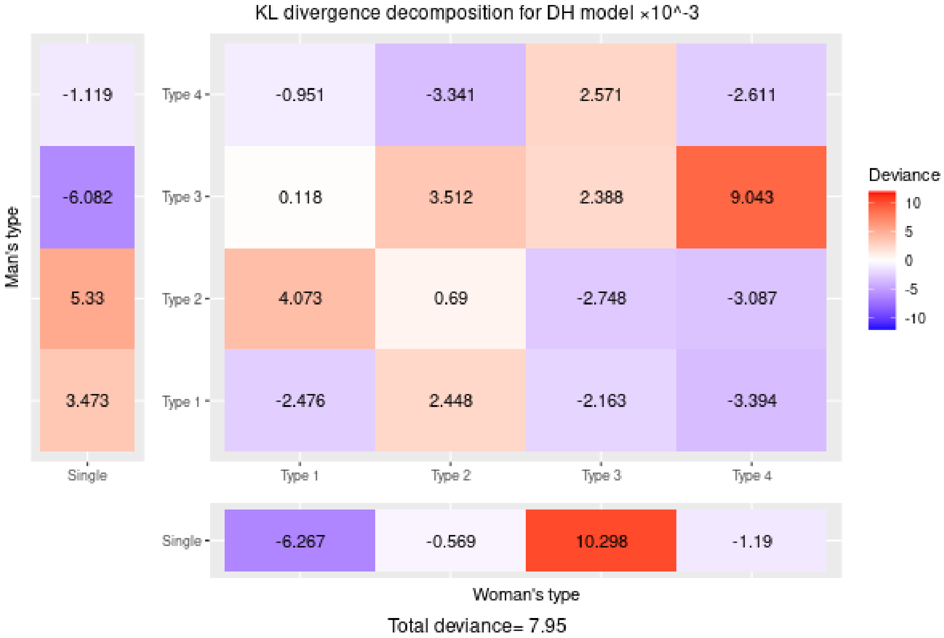


Figure 3.13: KL divergence decomposition when fitting DH model to data from SM model

3.10 Discussion

In this chapter I proposed several tools for model selection and assessing goodness-of-fit when applying RPMs. The results from Simulation Studies I.i and I.ii suggest that criterion scores can be used to select models while balancing error and parsimony and that the BIC is a particularly good metric for this. The studies additionally show how analysis of information gain can also facilitate model selection in a way that is intuitive and interpretable.

I also tested four different metrics for measuring model deviance from observed data and compared their performances. We find that all four metrics have comparable performance when studying information gain but that their performances were more varied when used for significance testing as in Simulation Study II.

While this chapter substantially develops goodness-of-fit procedures for RPM and establishes some baseline findings, there is much room for continued progress and further research. As mentioned in Section 3.5, the current method proposed for significance testing is a meaningful step in communicating confidence about a hypothesized model, but it collapses when population sizes exceed $1,000$ or when sample data is provided. This is clearly a major

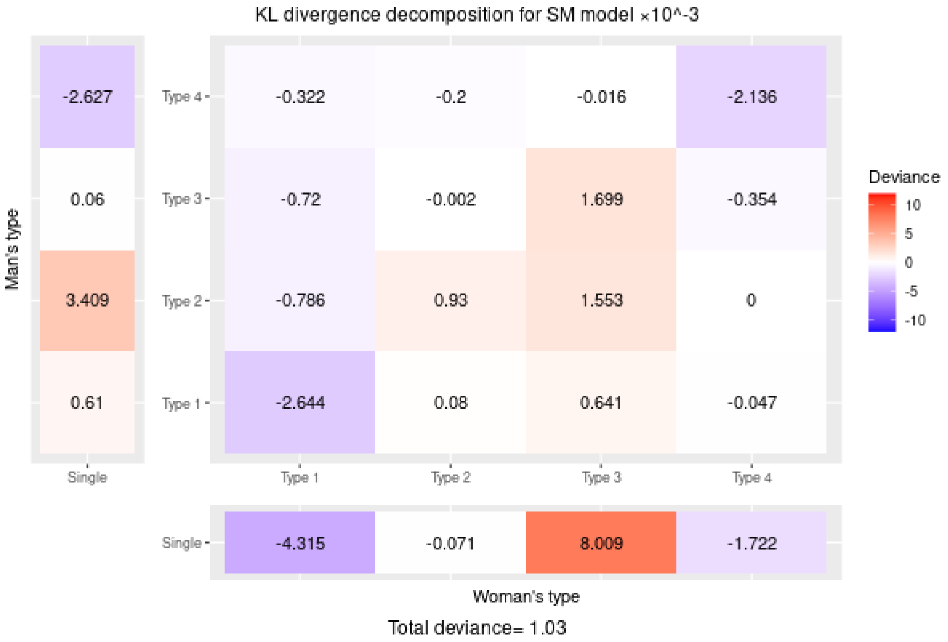


Figure 3.14: KL divergence decomposition when fitting SN model to data from SM model

weakness. Further development of significance testing procedures would be of substantial value to researchers using RPM.

Closely related to this topic is the current lack of procedure for determining the asymptotic distribution of different deviance metrics under a null hypothesis. The asymptotic independence of preferences and availability, as proposed by Menzel (2015), suggests that for large populations the outcomes for individual households should be independent. In actuality, however, asymptotically large populations are not available to researchers, and supplementary simulation studies I conducted (not presented in this dissertation) strongly suggested that for populations as large as $N = 20,000$, a given realized stable matching did not resemble a multinomial distribution with independent units. The chi-squared and G-squared divergence statistics computation between the data-generating model and the samples drawn from that process did not appear to follow a chi-squared distribution, as would be expected if the household outcomes were truly independent, nor were the observed distributions a straightforward transformation of the chi-squared distribution. While empirical estimates of the distributions of the deviance metrics work well in practice, additional research on the analytical distributions of any of these metrics in the RPM setting would strengthen the approach.

CHAPTER 4

Utility transfer in the two-sided matching market

4.1 Introduction

In the previous chapters of this dissertation, all proposed models, methods, and discussions have been set in a framework assuming non-transferable utility (NTU). An alternative to this setting is the transferable utility (TU) framework. These frameworks refer to a different sets of assumptions about how spouses share utility after marriage and, more broadly, about the target utility that agents are trying to maximize during the matching process.

Many economists and sociologists have made considerable contributions over time to modeling preferences under TU assumptions in the two-sided marriage market (e.g. Choo and Siow, 2006; Galichon and Salanié, 2021; Chiappori et al., 2017). Conversely, Dagsvik (1994), Logan et al. (2008), and Menzel (2015) contextualize their research of matching models in the NTU setting. Until now, much of the development of models in these two frameworks has taken place in parallel rather than in conjunction. Consequently, to our knowledge there is currently no recent literature that formally and comprehensively compares the assumptions, both implicit or imposed, of the NTU and TU frameworks as they are applied in the modeling of two-sided matching markets.

In this chapter of my dissertation, I aim to bridge this gap in the literature by providing a consolidated review of the TU and NTU frameworks in one place, using the same notation to describe the assumptions and models to ease direct comparison between the two. The contributions of this chapter are as follows: 1) a formal comparison of the NTU and TU frameworks in two-sided matching market literature, and the assumptions under each; 2) development of a recent TU model and results (Galichon and Salanié, 2021) from economics to a statistical framework; and 3) a comparison of how NTU and TU models compare when fit on real data.

The remainder of the chapter is organized as follows: in Section 4.2, I formally introduce the TU framework and review the NTU framework, explaining how RPM can be extended as an inference procedure for preferences in both cases. I note the commonalities in the initial structure of both models and then show how they deviate in current modeling. In Section 4.3 I show how RPM can be used to estimate educational preferences in spouses using data from the 2008 SIPP Topical Module. (U.S. Bureau of the Census, 2020) I follow this with a simulation study in Section 4.4 to demonstrate that RPM can be used to select between the NTU and TU frameworks for modeling when the researcher wishes to consider both as candidate models. I conclude in Section 4.5 with general comments on the different frameworks and suggestions for further research.

4.2 Settings for the NTU and TU markets

I briefly mentioned the NTU and TU frameworks in Section 1.2 and outlined some basic differences between the two. I now provide a more in-depth explanation of these frameworks and how they apply within the two-sided matching market, first discussing each framework individually and then formally comparing the two.

In Chapters 1 and 2, we discussed the basic setting of two-sided matching markets and how we conceptualize agents and preferences and developed basic notation. To ease continuity for the reader and facilitate the flow of the presentation of material in this chapter, I will review some of these concepts throughout this section. I begin by stating the basic characteristics in the marriage market that apply to both the NTU and TU frameworks.

4.2.1 Common background

Consider a two-sex population of size N with N_w women and N_m men, so that $N_w + N_m = N$. Women are indexed by $i \in \{1, \ldots, N_w\}$ and men are indexed y $j \in \{1, \ldots, N_m\}$.

The observable characteristics of woman i can be represented by some vector $x_i \in \mathcal{X}$ and the observable characteristics of man j can be represented by some vector $z_j \in \mathcal{Z}$. $\mathcal{X}$ and $\mathcal{Z}$ are countable finite sets with lengths X and Z, respectively, so that there are $X = |\mathcal{X}|$ observable types for women and $Z = \mathcal{Z}|$ observable types for men. Let $\bar{w}(x)$ represent the proportion of the *entire population* consisting of type $x \in \mathcal{X}$ women and $\bar{m}(z)$ represent the same proportion for type $z \in \mathcal{Z}$ men.

In the two-sided setting, marriages can only occur between individuals on opposite sides of the market. Individuals can also choose to remain single. We define households as entities that consist of either: 1) exactly one couple (a married woman and man); 2) exactly one single woman; or, 3) exactly one single man. A household is characterized by the observable type, or combination of types, of the individual(s) within it, so that there are $X \times Z + X + Z$ distinct types of households observable. The total number of households in the population is N_h. We note that $N_h \leq N$, with equality only if every woman and man in the population remains single.

A sufficient statistic to summarize the stable matching in this population

$$\bar{c} = \{c(x,z), c(x,*), c(*,z), x \in \mathcal{X}, z \in \mathcal{Z}\}.$$

- $c(x,z)$ is the count of households consisting of a married type x woman and type z man
- $c(x,*)$ is the count of households consisting of a single type x woman
- $c(*,z)$ is the count of households consisting of a single type z man

Effectively, $\bar{c}$ is the realized frequency distribution of household types with the sum of its elements equal to N_h.

Presented with data on the stable matching in the form of $\bar{c}$, we are able to discern the composition, or availabilities, of the individuals in the population:

$$\frac{c(x,*) + \sum_{z \in \mathcal{Z}} c(x,z)}{N} = \bar{w}(x) \quad \forall x \in X \tag{4.1}$$

$$\frac{c(*,z) + \sum_{x \in \mathcal{X}} c(x,z)}{N} = \bar{m}(z) \quad \forall z \in Z \tag{4.2}$$

For inference of preference parameters, we require

$$c(x,*), c(*,z), c(x,z) \geq 0, \quad \forall x \in \mathcal{X}, z \in \mathcal{Z}. \tag{4.3}$$

Let

$$\bar{f} = \{f(x,z), f(x,*), f(*,z), x \in \mathcal{X}, z \in \mathcal{Z}\},$$

with elements referring to the proportion of all N_h households in each category but otherwise similarly defined to $\bar{c}$. $\bar{f}$ represents the probability mass distribution of households in the population given some hypothesized data-generating process (model). Then, $\bar{c}$ is an empirical realization of $\bar{f}$.

Additionally, let $g(x,*)$ represent the log-odds that a type x woman chooses to remain single in the stable matching, so that

$$f(x,*) = \frac{\bar{w}(x) e^{g(x,*)}}{(1 + e^{g(x,*)})} \tag{4.4}$$

$$f(*,z) = \frac{\bar{m}(x) e^{g(*,z)}}{(1 + e^{g(*,z)})}. \tag{4.5}$$

In both the TU and NTU frameworks, it has been shown that as the population size grows large, $\bar{f}$ approaches a limiting distribution that can be defined by separable terms representing preferences parameters and availabilities. These relationships have been derived by Choo and Siow (2006) for the TU setting and by Dagsvik (2000) and Menzel (2015) in the NTU setting.

To proceed with inference, we can approximate the log-likelihood of the realized household counts $\bar{c}$ with a large population pseudo-likelihood function:

$$\begin{aligned} &\text{lp-log-lik}(\underset{\sim}{\beta}, g(x,*), g(*,z) | \{x_i, z_i, w_i^w\}_{i=1}^{n_w}, \{z_j, x_i, w_j^m\}_{j=1}^{n_m}) \qquad (4.6) \\ =& \sum_{x \in \mathcal{X}} \sum_{z \in \mathcal{Z}} c(x,z) \log f(x,z) + \sum_{x \in \mathcal{X}} c(x,*) \log f(x,*) + \sum_{z \in \mathcal{Z}} c(*,z) \log f(*,z), \end{aligned}$$

where $\bar{f}$ represents the model-based probability mass distribution of households in large populations.

We re-write $f(x, z)$ as the product of an availability component and a preferences component as dictated by the asymptotic relationship in our chosen marriage market framework. We additionally substitute $f(x, *)$ and $f(*, z)$ with Equations 4.4 and 4.5, respectively. This gives a pseudo-log-likelihood function which is written in terms of preference parameters $\underset{\sim}{\beta}$, $\{g(x, *), g(*, z),\ \forall(x, z)\}$, and $\{\bar{w}(x), \bar{m}(z)\}$. We derive the maximum large population pseudo-likelihood estimates (MLPLEs) for $\underset{\sim}{\beta}$, $g(x, *)$, and $g(*, z)$ subject to additional asymptotic constraints dependent on the framework . Having established a common set-up and notation, we now proceed to describe the specific assumptions and properties in the TU and NTU frameworks. The next two subsections can be read in any order.

4.2.2 Transferable utility framework

In a work seminal to the modern development of marriage modeling using tools from economics, Becker (1973) suggests that a marriage generate a "total output," or surplus utility, which is shared between the individuals in that marriage. This notion forms the basis of the transferable utility setting, in which researchers assume that in a marriage, one agent can decrease their utility by a specific amount to increase his or her spouse's utility by the exact same amount. In some literature, this is understood as the idea that the members of a couple engage in within-couple exchanges of utility-providing goods and services. Alternatively, Choo and Siow (2006) interpret these exchanges as determining "...each spouse's share of responsibilities within a marriage."

Choo and Siow (2006) developed a marriage model in the TU setting to allow point estimation of societal-level preferences for each possible observable household type. This work served as a basis for later developments in the same setting by other economists, including Dupuy and Galichon (2014), Chiappori et al. (2017), and Galichon and Salanié (2021). As the setting specifications assumed in Choo and Siow (2006) are the ones that much subsequent literature has followed when developing the TU marriage market model, those are the ones we will describe here.

In the TU market, the joint surplus utility of any given woman i of type x_i and any given man j of type z_j is given by

$$W_{ij}^{TU}(x_i, z_j) = W(x_i, z_j|\underset{\sim}{\beta}^{\mathrm{TU}}) + \eta_{iz} + \zeta_{xj} \tag{4.7}$$

This function has a deterministic component $W(x_i, z_j|\underset{\sim}{\beta}^{\mathrm{TU}})$ deterministic component and stochastic components η_{iz} and ζ_{xj}. The random component η_{iz} represents some utility shift that occurs in a partnership between woman i and any man of type z due to the unobserved preferences either of woman i for all type z men. This component is drawn from an extreme-value Type-I (Gumbel) distribution P_{x_i}. Then ζ_{xj} represents the corresponding quantity for man j and any woman of type x and is drawn from an extreme-value Type-I distribution Q_{z_j}. The random error terms η for all type x women are drawn from the same distribution P_x, and the random error terms ζ for all type z men are drawn from the same distribution Q_z.

A key assumption of this set-up is that woman i may have unobserved (to the researcher) idiosyncratic preferences, represented by η_{iz} for the observed characteristics z_j of man j. However, she cannot have unobserved preferences for his *unobserved* characteristics. Thus, in isolation, woman i is indifferent between all men with type z. Furthermore, woman i's idiosyncratic preference for the type z man is drawn from a probability distribution that depends on her own type x. The parallel assumption holds true for men. This assumption is known as *separability*.

In addition to the joint partnership utility function, woman i of type x_i perceives her utility for remaining single as

$$U_{i0}^{\mathrm{TU}} = 0 + \eta_{i0}, \qquad \eta_{i0} \sim P_{x_i} \tag{4.8}$$

and man j perceives his utility for remaining single as

$$V_{0j}^{\mathrm{TU}} = 0 + \zeta_{0j}, \qquad \zeta_{0i} \sim Q_{z_j} \tag{4.9}$$

The deterministic components of the singlehood utilities for both genders are scaled to 0, and the idiosyncratic terms are drawn from type-I extreme-value distributions determined by the agent's type (and the same distribution from which that agent's idiosyncratic term in the partnership utility function was drawn.)

An exogenously-determined surplus splitting function $\tau(x_i, z_j)$ (known to all individuals in the market prior to partnership) determines the share of the fixed component of the joint surplus utility $W_{ij}^{TU}(x_i, z_j|\underset{\sim}{\beta}^{\text{TU}})$ that woman i of type x_i receives after marrying man j of type z_j. The share of the deterministic utility received by woman i is determined only by her type and her spouse's type.

The final realized post-marriage utility for woman i is

$$U_{ij}^{\text{TU}} = \tau(x_i, z_j) + \eta_(i, z_j). \tag{4.10}$$

Similarly, the realized post-marriage utility for man j is

$$V_{ij}^{\text{TU}} = W(x_i, z_j|\underset{\sim}{\beta}^{\text{TU}}) - \tau(x_i, z_j) + \zeta_(x_i, j). \tag{4.11}$$

For every woman i (resp. man j), there is some $z \in \mathcal{Z}$ (resp. $x \in \mathcal{X}$) that maximizes 4.10 (resp. 4.11). This translates to a demand from a type x_i woman (resp. type z_j man) for a type z man (resp. type x woman). An additional important but somewhat opaque assumption made by Choo and Siow (2006) is that there are a large (uncountable infinite) number of woman and men of each type $x \in \mathcal{X}$ and $z \in \mathcal{Z}$ in this population. As a result, the demand for type z men among type x women is met by the supply of z men who demand a marriage with a type x women.

This appears to be a fairly strong assumption. In practical applications, a vital but as yet unanswered question is how large N must be for this assumption to hold true and for the marriage market to clear. Interestingly, subsequent literature that builds on Choo and Siow's (2006) framework either only mentions this assumption briefly or not at all, and there is little elaboration on whether it is possible for compositional changes in the population to create violations of this assumption. (Chiappori et al., 2017; Galichon and Salanié, 2021)

Assuming that the TU marriage market *does* clear, however, Choo and Siow (2006) show that the following relationship holds asymptotically

$$f(x,z) = \underbrace{2e^{W(x,z|\underset{\sim}{\beta})}\sqrt{\frac{e^{g(x,*)+g(*,z)}}{(1+e^{g(*,x)})(1+e^{g(*,z)})}}}_{\substack{\text{preference}\\ \text{component}}} \cdot \underbrace{\sqrt{\bar{w}(x)\bar{m}(z)}}_{\substack{\text{availability}\\ \text{component}}} \quad \forall x, z. \tag{4.12}$$

Thus, in a TU setting $f(x,z)$ can be decomposed into separate preferences-related and availability-related components, suggesting that the interaction between these two components becomes negligible for large populations. We exploit the relationship in Equation 4.19 to infer the joint surplus preference parameters $\underset{\sim}{\beta}^{\text{NTU}}$ alongside the relative likelihoods of remaining single $g(x,*), g(*,z)$ $\forall x \in \mathcal{X}, z \in \mathcal{Z}$.

4.2.3 Non-transferability framework

The core assumption of the NTU framework is that there is no mechanism for utility transfer between spouses once a marriage occurs. In other words, neither spouse can decrease their own utility to increase their partner's utility by an equal amount.

In the application of this framework, we consider a utility function for partnership that any given woman i uses to assess any given man j as potential partners based on the observable characteristics of both the woman i and man j:

$$U_{ij}^{NTU} = U_{ij}(x_i, z_j|\underset{\sim}{\theta}^W) + \eta_{ij} \tag{4.13}$$

The partnership utility function has a deterministic component $U_{ij}(x_i, z_j|\underset{\sim}{\theta}^W)$ and a stochastic component η_{ij}. Previous literature on NTU markets, including Menzel (2015) and the earlier chapters of this dissertation, focus on the case where η_{ij} follows standard extreme-value Type-I (Gumbel) distribution.

Any given man j similarly assesses a potential partnership with any given woman i using the function

$$V_{ij}^{\text{NTU}} = V_{ij}(z_j, x_i|\underset{\sim}{\theta}^M) + \zeta_{ij}, \tag{4.14}$$

which shares the characteristics of the U_{ij} described above.

Woman i also assesses her utility for remaining single based on a singlehood utility function

$$U_{i0}^{\text{NTU}} = 0 + \eta_{i0}. \tag{4.15}$$

The deterministic component of this utility function is scaled to 0. Following earlier NTU work, e.g. Menzel (2015), as well as our proposed model, the random error term η_{i0} in the singlehood utility function is drawn from a Gumbel distribution with location parameter $\ln \sqrt{N_w}$ (and scale parameter 1).

Man j has a similarly defined function for his own singlehood utility

$$V_{0j}^{\text{NTU}} = 0 + \zeta_{0j}, \tag{4.16}$$

with $\zeta_{0j} \sim \text{Gumbel} \ln \sqrt{N_m}, 1)$.

The sum of U_{ij} and V_{ij} gives the total partnership surplus W_{ij} that would exist in a marriage between woman i and man j, e.g.

$$\begin{aligned} W_{ij}^{\text{NTU}} &= U_{ij} + V_{ij} \\ &= U_{ij}(x_i, z_j|\underset{\sim}{\theta}^W) + V_{ij}(z_j, x_i|\underset{\sim}{\theta}^M) + \eta_{ij} + \zeta_{ij} \end{aligned} \tag{4.17}$$

The deterministic component of this surplus function is written as

$$W_{ij}^{\text{NTU}}(x_i, z_j|\underset{\sim}{\beta}^{\text{NTU}}) = U_{ij}(x_i, z_j|\underset{\sim}{\theta}^W(\underset{\sim}{\beta}^{\text{NTU}}) + V_{ij}(z_j, x_i|\underset{\sim}{\theta}^M(\underset{\sim}{\beta}^{\text{NTU}})), \tag{4.18}$$

where $\underset{\sim}{\beta}^{\text{NTU}}$, which parametrizes W_{ij}, is a (in our case linear) combination of $\underset{\sim}{\theta}^W$ and $\underset{\sim}{\theta}^M$.

In Chapter 2, we presented the results that showed for asymptotically large populations, the following relationship approximately holds

$$f(x,z) = \underbrace{2\frac{e^{W_{ij}(x_i,z_j|\underset{\sim}{\beta}^{\text{NTU}})+g(x,*)+g(*,z)}}{(1+e^{g(x,*)})(1+e^{g(*,z)})}}_{\substack{\text{preference}\\ \text{component}}} \cdot \underbrace{\bar{w}(x)\bar{m}(z)}_{\substack{\text{availability}\\ \text{component}}}. \qquad \forall x \in \mathcal{X}, z \in \mathcal{Z} \tag{4.19}$$

In this relationship $f(x,z)$ is decomposed into separate preferences-related and availability-related components, suggesting that the interaction between these two components becomes negligible for large populations. We exploit the relationship in Equation 4.19 to infer the joint surplus preference parameters $\underset{\sim}{\beta}^{\text{NTU}}$ alongside the relative likelihoods of remaining single $g(x,*), g(*,z)\ \forall x \in \mathcal{X}, z \in \mathcal{Z}$.

4.2.4 Comparison of the frameworks

The defining assumption that distinguishes the TU and NTU frameworks is that in the former spouses are able to transfer utility to one another while in the latter they are not. The major implication of this assumption is that agents behave differently during the matching process. In the TU setting, agents choose partnerships to maximize their surplus partnership utility function W_{ij}^{TU}, whereas in the NTU setting women and men aim to maximize their individual utility functions U_{ij}^{NTU} and V_{ij}^{NTU} respectively.

Because of the difference in behavior during the matching process, what constitutes a stable matching in one framework may not be stable in the other. In the big picture, the stable matchings in the two frameworks converge to different limiting distributions.

Alongside this consideration of whether or not utility can be transferred within a partnership, there are several other subtle but important differences in the TU and NTU settings as they have been developed by Choo and Siow (2006) and Menzel (2015), respectively. We list some of them here:

1. In the NTU framework, the idiosyncratic component in the female's partnership utility function η_{ij} is unique to every man j and independent of both her type and his, and all η_{ij} are drawn from a standard Gumbel distribution. In the TU framework, woman i derives the same idiosyncratic utility shift for all type z men and does not distinguish between them. In addition, the distribution underlying the random component is dependent on woman i's type. This difference applies in parallel to the male partnership utility function.

   The major implication of this difference is that in a TU model, the choosing individual only considers the prospective match's observable characteristics (Chiappori, 2020).

In contrast, within the NTU framework, individuals are influenced by the prospective match's observable (to the researcher) characteristics and the characteristics that are to the researcher unobservable.

2. Also in the NTU framework, the idiosyncratic component η_{i0} of the partnership function is drawn from a standard Gumbel distribution, while the idiosyncratic component η_{ij} of the singlehood function is drawn from Gumbel distribution with location parameter $\ln \sqrt{N_w}$. The shift in the distribution of η_{i0} is in place so that the probability of remaining single remains stable with changes in the population size. In the TU framework, η_{ij} and η_{i0} are drawn from the same distribution. This difference applies in parallel to the male partnership utility function.

3. A significant but, in my view, underemphasized assumption of the TU framework is the large population assumption which seems to render the question of population composition and availability of different types of spouses moot. To my knowledge, there are not currently any guidelines for how the TU model can be adapted to work with small populations or in large populations where individuals are still constrained by an option set with a limited number of potential partners of each type. For the NTU model, we have developed bias correction procedures to address this scenario.

4. Early versions of the TU model did not allow for spillover effects - that is, variations in the number of type x' women or type z' men do not impact $c(x, z)$ for $x \neq x', z \neq z'$. (Schoen, 1981) Choo and Siow (2006) indicate that their revision of the TU model addresses this deficiency, but it is not entirely clear how, or how it compares to this property goes unmentioned in subsequent papers by other researchers who build on their work. (Dupuy and Galichon, 2014; Galichon and Salanié, 2021)

   Under the NTU model, however, the entry of type x' women can impact is such that an increased availability of potential partners leads to increased probability of an agent finding a match.

Despite these differences in the TU and NTU settings, RPM can still be used for preference parameter inference assuming either framework. We only need to account for the difference in the asymptotic relationship between $f(x, y)$, preferences, and availabilities, which we then plug into the large population pseudo-likelihood function (Equation 4.6).

4.3 Applications

Whether the TU or NTU framework provides a more accurate model for the two-sided marriage market is still debated. In this section, I show how this question might be answered by applying goodness-of-fit procedures to assess the fits of a TU and NTU model on data from the 2008 Survey on Income and Program Participation.

To review, the sample was drawn using a stock-stock design and was isolated for this study to only include households consisting of single individuals who had never been married or couples who had gotten married for the first time in the prior year at the time of survey. The sample consisted of 21,077 households, of which 520 contained married couples and the rest contained singletons. Researchers recorded the education level of each agent as either: less than high school (LHS), high school (HS), some college (SCO), or college graduate (CO).

I fit SM models from TU and NTU frameworks to it. The parameter preference estimates were constrained to fall between -15 and 15. The parameter estimates for $\beta_{x,z}$, representing the deterministic shift in surplus utility in a partnered household with type x woman and type z, are visualized in Figure 4.1. The left plot shows the parameter estimates under the TU model and the right plot shows the estimates under the NTU model.

We note first that the TU model estimates universally lower preference parameters than the NTU model for the MLPLE in the SM class, with an average difference of about 5 units. Both models indicate that models between individuals with very different education levels produce the most negative shifts in utility. The lightest colored squares lie along the positive diagonal, indicating a preference for partners of the same education level and reflecting a broader preference for educational homogamy in the population.

Despite the difference in parameter estimates, the two models have equal AIC scores

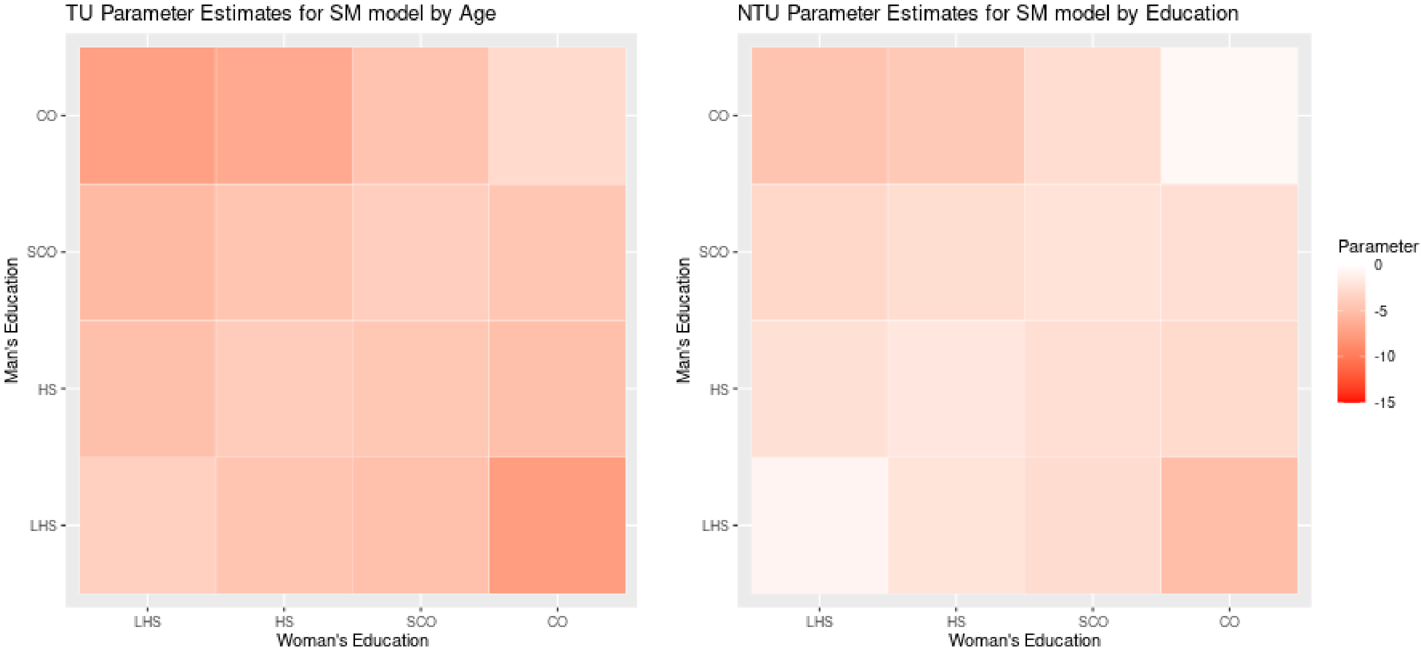


Figure 4.1: SM model parameter estimates when applying TU (L) and NTU (R) frameworks to SIPP data

of 95470.11 and equal BIC scores 95482.47 given the data. As it turns out, the estimated probability mass distribution of households is almost exactly the same between these two models, so that even the difference in the chi-squared scores (TU-NTU) is only -0.07 - extremely small considering the sample size of over 21,000.

Through additional studies (not presented here), I found that for large populations, although the estimated TU and NTU models are clearly different from each other, they achieve almost identical similar fits to the data.

4.4 Simulation Study

To better understand the TU and NTU models and to assess how well goodness-of-fit measures can actually detect the differences between the two models, I conducted a simulation study in two parts. In Part I, I simulated $B = 200$ populations of size $N = 5,000$, where men and women come in four observable types each and are drawn from the same marginal distribution considered in Table 3.1 in Chapter 3. I then used a modified version of the Gale-Shapley algorithm to achieve a stable matching in that population based on a DH marital surplus model with TU assumptions. The DH model is specified identically to the one in 3.12 of Chapter 3.

I use RPM to fit the TU and NTU models to the data, estimating $\hat{W}^{\text{TU},b_1}$ and $\hat{W}^{\text{NTU},b_1}$ for each iteration $b_1 \in \{1, \dots, B\}$. I then compute the AIC scores of the models and subtract AIC($\hat{W}^{\text{NTU},b_1}$) from AIC($\hat{W}^{\text{TU},b_1}$). Because the TU model is actually the correct data-generating process, we expect it to have a greater log-likelihood and thus a smaller AIC score.

The distribution of the differences in Figure 4.2. The red dotted line indicates where the difference in the two AIC scores is equal. The plot shows that the difference in the AIC scores of the TU and NTU models is in fact almost always negative, with a mean of approximately -17.8. It also appears to follow a bell-shaped distribution. The error rate, or probability that the difference in the AIC scores is greater than 0 and thus indicating the

AIC score favors the NTU framework, is quite low 3.36%.

I repeat this procedure for Part II of the simulation study, except this time the stable matching is achieved under NTU assumptions. I again fit RPM under both frameworks and estimate $\hat{W}^{\text{TU},b_2}$ and $\hat{W}^{\text{NTU},b_2}$ for each iteration $b_2 \in \{1, \ldots, B\}$. This time, I compute the difference of AIC scores as $\text{AIC}(\hat{W}^{\text{NTU},b_2}) - \text{AIC}(\hat{W}^{\text{TU},b_2})$. Since we are once again subtracting the AIC of the incorrect model from that of the correct model, we again expect to see differences less than 0.

The distribution of the AIC differences in Part II is shown in Figure 4.2. The majority of the distribution does appear to fall below 0 as expected. The distribution still seems to resemble a bell-curve with a mean at -7.4, though perhaps it is slightly skewed right. This slight skew is reflected in the error rate of 0.168, which is quite a bit higher than the error rate in Part I.

Notably, the magnitude of the differences appears to be much higher in Part I when fitting the NTU model to the TU data, than in Part II when fitting the TU model to NTU data. As Burnham and Anderson (2004) suggests that a difference of even 2 in AIC scores is often enough to imply a considerable difference in model fit, the differences observed in both parts of the simulation study are still large enough usually to suggest that the correct model is definitively better than the incorrect one. However, the lower mean and variance observed in the distribution of differences in Part II, as shown in Figure 4.3, combined with the higher error rate indicates that perhaps TU models fit slightly better to NTU data than the other way around.

4.5 Discussion

This chapter formally compares the two most common frameworks that have developed over the last two decades for modeling two-sided marriage markets. By developing the results from Choo and Siow (2006) and Menzel (2015) into similar notation, we hope to ease comparison of the two frameworks for future researchers and encourage researchers to consider both

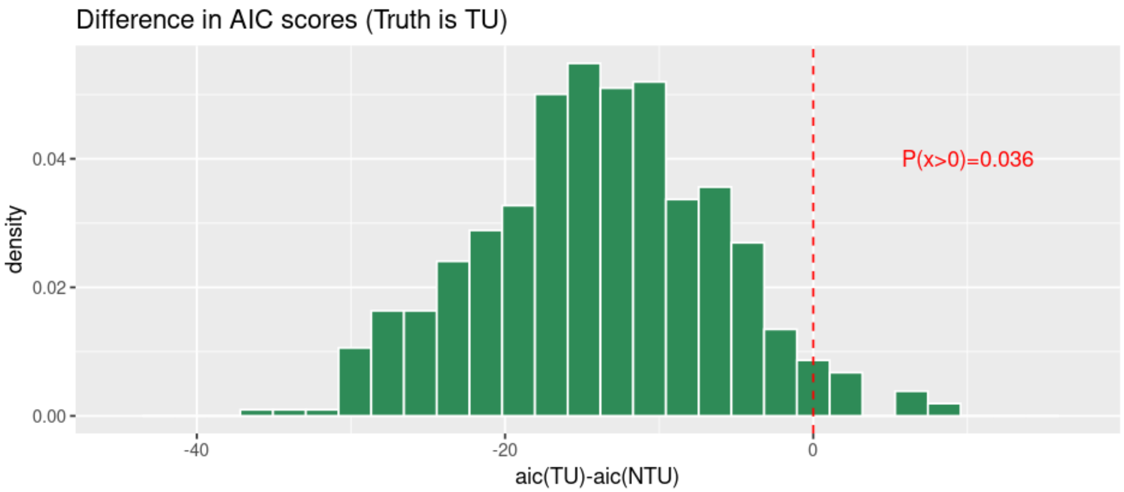


Figure 4.2: Difference in AIC scores of TU and NTU models over simulated data from TU framework (200 simulations)

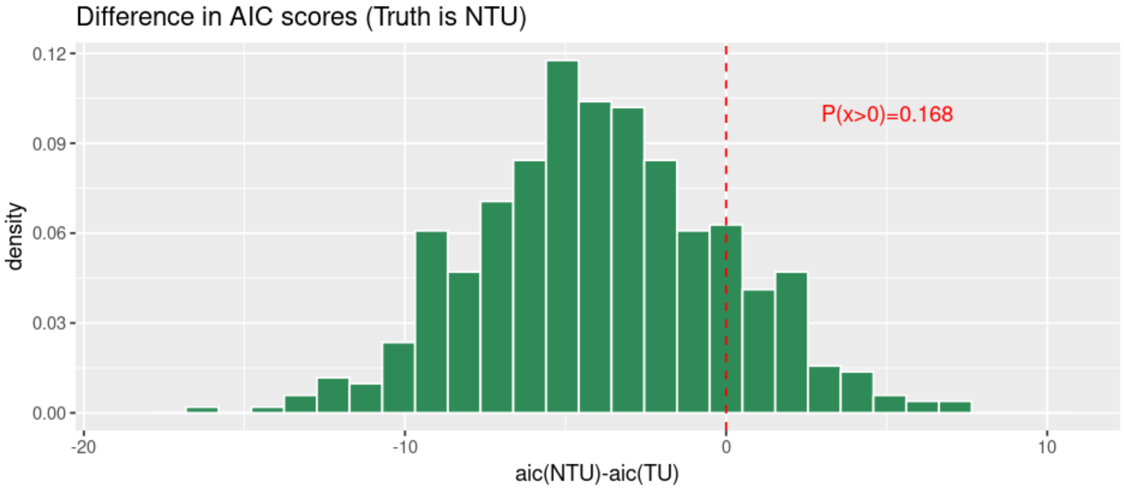


Figure 4.3: Difference in AIC scores of TU and NTU models over simulated data from NTU framework (200 simulations)

models in analyzing market preferences, particularly when there is no strong prior evidence suggesting (non-)transferability of utility.

As noted throughout the chapter, there are some assumptions invoked in the TU market that remain opaque and thus difficult to translate. Further development in this area is necessary. There are some additional weaknesses of the TU framework that were not commented on earlier, such as the unlikely assumptions that transfers are required, unconstrained, and frictionless (i.e. no loss of utility in the transfer process). More likely, the transfer function or mechanism is dependent on observed and unobserved characteristics of the agents in the partnership.

We note that besides TU and NTU, there are additional frameworks that make more relaxed assumptions about the transferability of utility, e.g. imperfectly transferable utility (ITU). For a review of these frameworks, which relax some of the assumptions of the TU framework, I recommend Chiappori (2020).

CHAPTER 5

Parallel and Future Work

This dissertation makes a significant contribution by proposing a revealed preferences model to infer preferences in marriage markets and other two-sided markets. However, many of the methods presented here have significant room for further development. An effective method of significance testing that can be used in practical applications, for example, would be a major step forward and facilitate the usefulness of this model to demographers, economists, and other interested parties. Another potential approach for goodness-of-fit that is not explored in this dissertation is using cross-validation - e.g. splitting data into training and testing sets.

This dissertation also makes a significant step in bridging the widening gap between TU and NTU literature by formally comparing them side-by-side. As noted in Chapter 4, beyond the obvious difference in assumption about utility transferability which defines the TU and NTU frameworks, there are many additional, more subtle differences in the assumptions *imposed* by researchers in each setting to facilitate inference. Future work may continue to further develop one framework to more closely resemble the other so that direct comparison of the models is easier.

An important issue not addressed by this paper is the identification of the effective population that constitutes the market. An important concept that I did not explore in this dissertation is that of awareness, that is, the set of people a person effectively chooses among. There are numerous ways to conceptualize this - perhaps, for example, the social distance between two people. This idea can be pursued in at least two separate ways.

In one case we model a person's aware of a potential partner as a function of observed characteristics of the individuals (e.g., geographic distance, age difference) (Menzel, 2015). Incorporating these characteristics requires a significant expansion of the model (for example, geographic distance is a continuous variable, requiring the integral version of the model) and decisions about the model should incorporate awareness as a deterministic (0-1) or probabilistic, as well as of course choices about how to parameterized the awareness function itself.

Alternatively, we could conceptualize lack of awareness of a potential partner as interchangeable with negative utility for that partnership. The idea here is that there is some reason people are unaware of each other, perhaps due to geographic or social distance. We could add a term to the surplus utility model which then shifts utility based on this "distance" between individuals. This line of research is pursued in Zhang (2022).

APPENDIX A

Simulation Studies from Chapter 2

A.1 Supplementary Tables

Table A.1: Medians and standard deviations (SDs) of differential homophily model bias corrected MLPLEs in simulation study I.i (1,000 simulations, $N = 6,000$)

		Availability			
Parameter	Truth	$\mathcal{A}_1$		$\mathcal{A}_2$	
	$\underset{\sim}{\beta}^{\mathrm{DH},0}$	Median	SD	Median	SD
intercept	-3.439	-3.435	0.136	-3.425	0.133
homophily e.1	1.883	1.887	0.391	1.879	0.332
homophily e.2	0.868	0.886	0.310	0.875	0.290
homophily e.3	0.557	0.561	0.238	0.558	0.256
homophily e.4	2.191	2.198	0.243	2.194	0.308

The homophily t parameter is the coefficient of an indicator which equals 1 if both partners have education level t.

Table A.2: Medians and standard deviations (SDs) of reduced mix model bias corrected MLPLEs in simulation study I.i (1,000 simulations, $N = 6,000$)

Education			Availability			
Parameter		Truth	$\mathcal{A}_1$		$\mathcal{A}_2$	
Female	Male	$\underset{\sim}{\beta}^{\text{RM},0}$	Median	SD	Median	SD
1	1	-1.572	-1.565	0.401	-1.585	0.330
2	1	-2.877	-2.854	0.433	-2.837	0.389
3	1	-3.419	-3.374	0.477	-3.377	0.422
1	2	-3.209	-3.070	0.496	-3.097	0.406
2	2	-2.570	-2.561	0.277	-2.561	0.270
3	2	-3.256	-3.226	0.283	-3.247	0.269
1	3	-3.695	-3.619	0.627	-3.524	0.672
2	3	-3.348	-3.311	0.348	-3.328	0.306
3	3	-2.888	-2.867	0.216	-2.855	0.217
4	3	-3.211	-3.207	0.330	-3.186	0.397
3	4	-3.387	-3.365	0.372	-3.311	0.425
4	4	-1.270	-1.249	0.190	-1.257	0.274
1 or 2	4	-5.082	-5.139	4.128	-4.838	4.349
4	1 or 2	-3.883	-3.829	0.450	-3.839	0.441

Education level codes: 1 =<high school, 2 =high school, 3 =some college, 4 =≥bachelors

Table A.3: Medians and standard deviations (SDs) of differential homophily model bias corrected MLPLEs in simulation study I.ii (1,000 simulations, $n_h = 21,077$)

Parameter	Truth	Availability			
		$\mathcal{A}_1$		$\mathcal{A}_2$	
	$\underset{\sim}{\beta}^{\text{DH},0}$	Median	SD	Median	SD
intercept	-3.439	-3.437	0.072	-3.435	0.064
homophily e.1	1.883	1.889	0.180	1.875	0.181
homophily e.2	0.868	0.854	0.156	0.864	0.145
homophily e.3	0.557	0.544	0.127	0.553	0.127
homophily e.4	2.191	2.200	0.115	2.195	0.149

The homophily e.t parameter is the coefficient of an indicator which equals 1 if both partners have education level t.

Table A.4: Medians and standard deviations (SDs) of reduced mix model bias corrected MLPLEs in simulation study I.ii (1,000 simulations, $N = 21,077$)

Education			Availability			
Parameter		Truth	$\mathcal{A}_1$		$\mathcal{A}_2$	
Female	Male	$\underset{\sim}{\beta}^{\mathrm{RM},0}$	Median	SD	Median	SD
1	1	-1.572	-1.585	0.180	-1.563	0.167
2	1	-2.877	-2.892	0.238	-2.873	0.207
3	1	-3.419	-3.421	0.230	-3.422	0.198
1	2	-3.209	-3.219	0.297	-3.210	0.273
2	2	-2.570	-2.584	0.146	-2.576	0.137
3	2	-3.256	-3.273	0.157	-3.261	0.137
1	3	-3.695	-3.745	0.335	-3.740	0.287
2	3	-3.348	-3.360	0.185	-3.343	0.178
3	3	-2.888	-2.893	0.115	-2.884	0.104
4	3	-3.211	-3.212	0.164	-3.230	0.206
3	4	-3.387	-3.388	0.200	-3.378	0.256
4	4	-1.270	-1.271	0.092	-1.264	0.128
1 or 2	4	-5.082	-5.069	0.410	-5.047	0.529
4	1 or 2	-3.883	-3.889	0.230	-3.891	0.240

Education level codes: 1 =<high school, 2 =high school, 3 =some college, 4 =≥bachelors

Table A.5: Simulation study II: MLPLEs and bias corrected MLPLEs for different N with Availability $\mathcal{A}_1$ and uniform homophily preferences (1,000 simulations)

Parameter	Truth	Bias Correction	$N = 60$		$N = 600$		$N = 6,000$	
	$\utilde{\beta}^{\text{UH},0}$		Median	SD	Median	SD	Median	SD
intercept	0.558	No	0.007	0.465	0.179	0.151	0.218	0.052
		Yes	0.485	0.533	0.509	0.171	0.520	0.059
homophily	1.170	No	1.086	0.580	1.117	0.168	1.147	0.053
		Yes	1.120	0.630	1.159	0.182	1.166	0.058

A.2 Confidence intervals from 200 simulations

Figures A.1 and A.2 show the analytical confidence intervals and the empirical bootstrap confidence intervals produced over 200 simulations for the $\beta_{4,4}$ and $\beta_{1 \text{ or } 2,4}$ parameters in the reduced mix model. These figures coincide with the simulation results related to uncertainty estimates described in Section 2.9.5. The horizontal axis gives the simulation index, and the vertical axis shows the range of the interval. The solid point at the center of each interval indicates the parameter estimate in the bootstrapped sample at that index. The horizontal red line in each plot represents the true parameter value, and intervals in blue are those which failed to include the true value. We provide the empirical coverage rate of the parameter for each method of confidence interval in the top-right corner of the plots.

The first three panels of Figure A.1 show the 200 confidence intervals for $\beta_{4,4}$ produced by each of the three bootstrapping methods which were described in Section 2.7. The three methods for constructing the bootstrapped confidence intervals produce very similar results, with the basic bootstrap method achieving 95% coverage and the percentile and modified studentized t methods achieving 96% coverage. Furthermore, the confidence intervals appear to have similar lengths across the three methods.The bottom-right panel shows the analytical confidence intervals produced for $\beta_{4,4}$ based on the same simulated populations. We note that

Table A.6: Simulation study III, MLPLEs for different n_h with Availability $\mathcal{A}_1$ and differential homophily preferences at $N = 6,000$ (200 simulations)

Parameter	Truth	Bias Correction	$n_h = 600$		$n_h = 1,200$		$n_h = 3,000$	
	$\underset{\sim}{\beta}^{\text{DH},0}$		Median	SD	Median	SD	Median	SD
intercept	0.561	No	0.223	0.142	0.227	0.092	0.218	0.057
		Yes	0.531	0.156	0.537	0.107	0.519	0.064
homophily e.1	1.883	No	1.859	0.363	1.848	0.262	1.844	0.153
		Yes	1.891	0.388	1.884	0.294	1.886	0.170
homophily e.2	0.868	No	0.872	0.262	0.861	0.186	0.870	0.121
		Yes	0.866	0.295	0.846	0.199	0.872	0.129
homophily e.3	0.557	No	0.567	0.216	0.569	0.141	0.581	0.087
		Yes	0.551	0.242	0.541	0.159	0.564	0.097
homophily e.4	2.191	No	2.122	0.269	2.106	0.178	2.110	0.119
		Yes	2.206	0.281	2.173	0.197	2.193	0.126

Education level codes: 1 =<high school, 2 =high school, 3 =some college, 4 =≥bachelors

The homophily e.t parameter is the coefficient of an indicator which equals 1 if both partners have education level t.

the analytical 95% confidence intervals only achieve 83% coverage in this set of simulations, indicating undercoverage.

The performances of the three bootstraps methods are more varied more when evaluating the $\beta_{1 \text{ or } 2,4}$ parameter. The modified studentized t and the percentile bootstrap confidence intervals achieve a coverage rate of 88% and 86.5%, respectively, while the basic bootstrap intervals achieve much lower coverage of 78.5%. Furthermore, the percentile and studentized t methods produce intervals which are generally wider than those produced by the basic bootstrap method. The analytical confidence intervals in the bottom-right panel of the figure are so narrow that few of them capture the true value, resulting in a poor coverage rate of 10.5%.

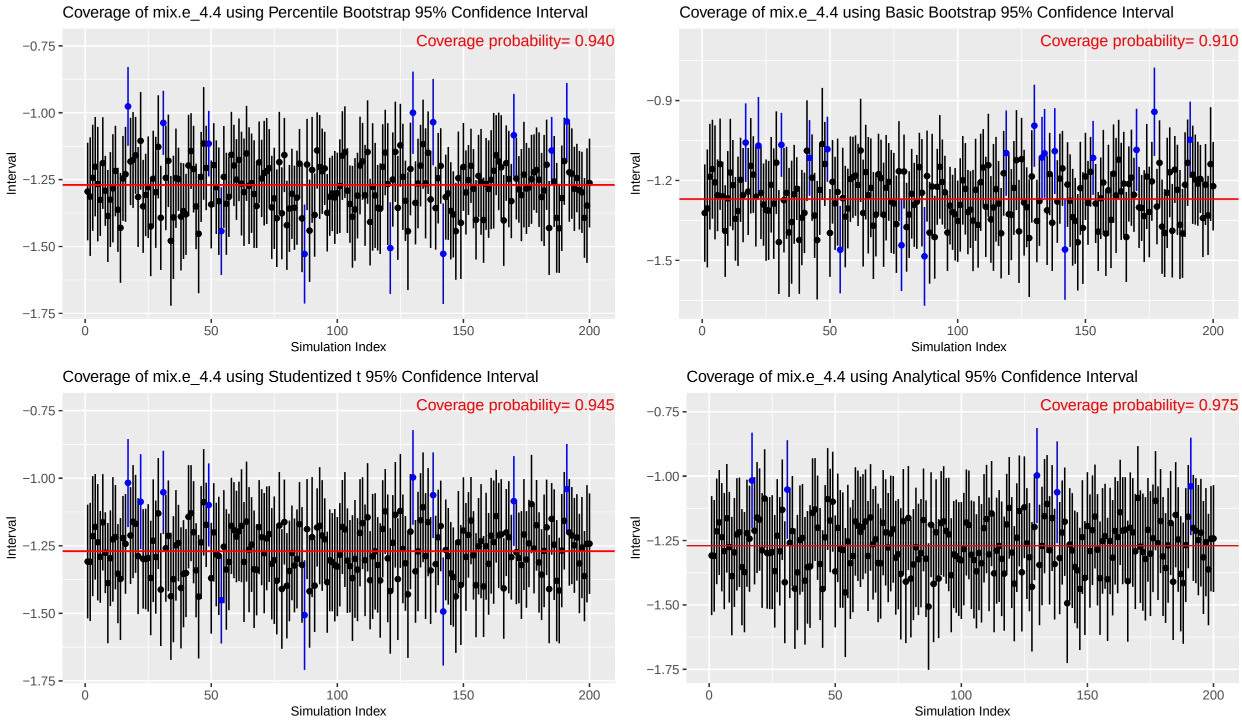


Figure A.1: Coverage of $\beta_{4,4}$ in reduced mix model over 200 simulations

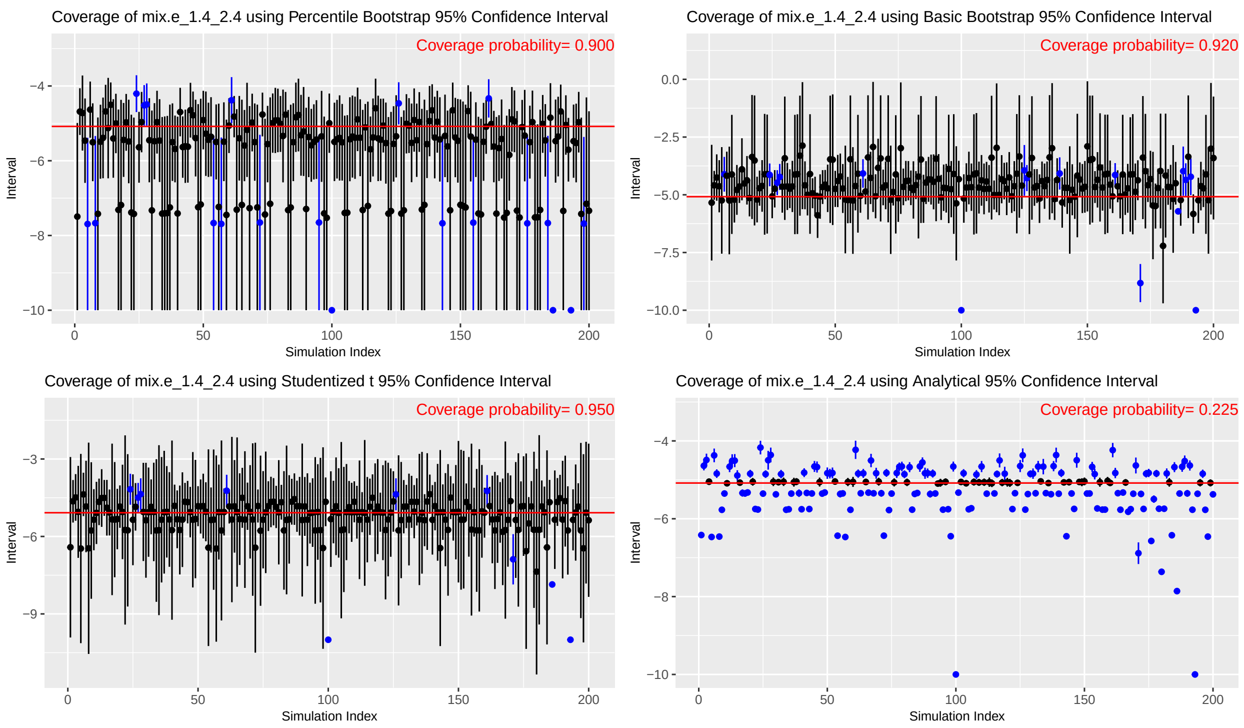


Figure A.2: Coverage of $\beta_{1 \text{ or } 2,4}$ in reduced mix model over 200 simulations

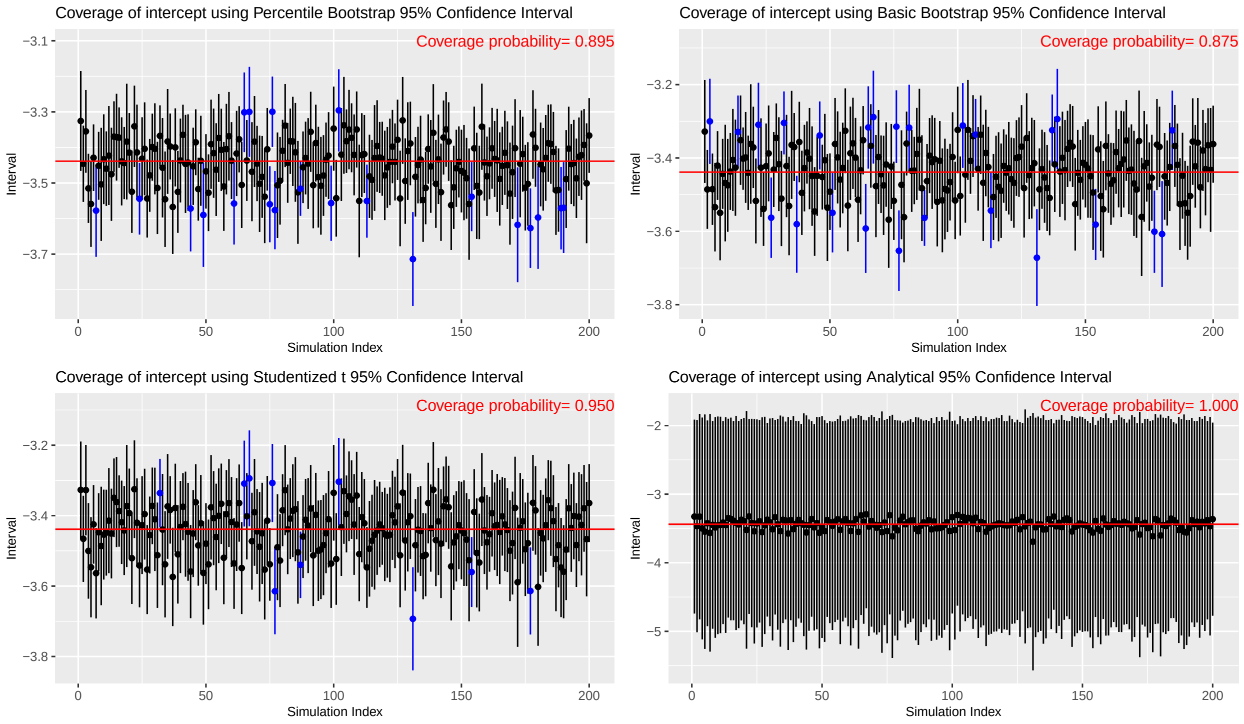


Figure A.3: Coverage of intercept β_0 in differential homophily model over 200 simulations

Figure A.4: Coverage of $\beta_{\text{homophily e.1}}$in differential homophily model over 200 simulations

APPENDIX B

Supplementary material on goodness-of-fit

B.1 Additional simulation study findings

I repeated the simulation study procedures described in Section 3.6 for synthetic population data with $N = 10,000$ and $N = 20,000$ individuals and for synthetic samples with $n = 1,000$ households. For the latter study, I first generated synthetic populations of size $10,000$ using the availabilities from Table 3.1, generated a stable matching based on either W_{UH} or W_{DH}, and then randomly drew sampled $1,000$ households based on a stock-stock sampling design.

At small sizes of B (up to 50), the results for these studies appeared consistent with those discussed in this section. Significance testing continued to perform poorly at large population sizes and for sampled data. The other proposed methods for goodness-of-fit testing, such as comparison of information gain and AIC and BIC scores, performed well for both large population data and for the sample. In the studies of synthetic populations with $N = 10,000$ and $N = 20,000$ individuals, the error rate in the AIC and BIC scores decreased and the information gain achieved by all models increased. However, these studies could not be repeated at a larger number of iterations for inclusion in this dissertation due to resource constraints. For this reason, they have been excluded from formal discussion in the results section. Further efforts in understanding the performance of goodness-of-fit methods given sample or large census data on a stable matching would further understanding of advantages and disadvantages of the methods proposed in this chapter.

B.2 Goodness-of-fit measurement for census data

In rare cases, researchers may obtain data on partnership outcomes within a fully observed closed population. In such a scenario, every woman $i \in \{1, \dots, N_w\}$ is fully aware of every man $j \in \{1, \dots, N_m\}$ in the closed population for consideration as a marital partner, and vice versa. Agents of either gender do not consider any partners outside the market, and the researcher has information on all partnerships formed in the stable matching achieved within this market. Because there is no element of random sampling, the only randomness in the observed matching comes from the stochastic component of the surplus function.

Recall that in Section 3.9 we showed that, for agents with discrete characteristics, matching data $\bar{c}$ can be represented as a contingency table, as in Figure 3.9.

When census data is available, a special case arises where the marginal totals are fixed for rows $1, \dots, X$ and for rows $1, \dots, Z$. Letting $w(x)$ and $m(z)$ equal the number of type x women and type z men in the population, respectively,

$$c(x, *) + \sum_{z \in \mathcal{Z}} c(x, z) = w(x) \quad \forall x \in X \tag{B.1}$$

$$c(*, z) + \sum_{x \in \mathcal{X}} c(x, z) = m(z) \quad \forall z \in Z \tag{B.2}$$

For each of the first X rows, the sum of the xth row is equal to the total number of type x women in the data. In each of the first Z columns, the sum of the zth column equals the total number of type z men in the data.

Whereas with sample data we considered the probability of observing the given sample assuming some hypothesized partnership surplus utility function; with population data we consider instead the probability that the observed count matrix would occur in that population over the distribution of unobserved preferences (idiosyncratic taste-shifters, or random errors). To test the goodness-of-fit of the model in this case, we can adapt the goodness-of-fit statistic suggested for $X \times Z$ contingency tables with fixed margins, which follows a hypergeometric distribution. (Agresti, 2002, Section 3.5.7) However, because the derivation of the distribution under the null hypothesis becomes unfeasible for large population sizes, and because census data is so rarely available, I do not explore this scenario in-depth in this dissertation.

B.3 Relative information gain in simulation studies using other deviance metrics

B.3.1 Simulation Study I.i

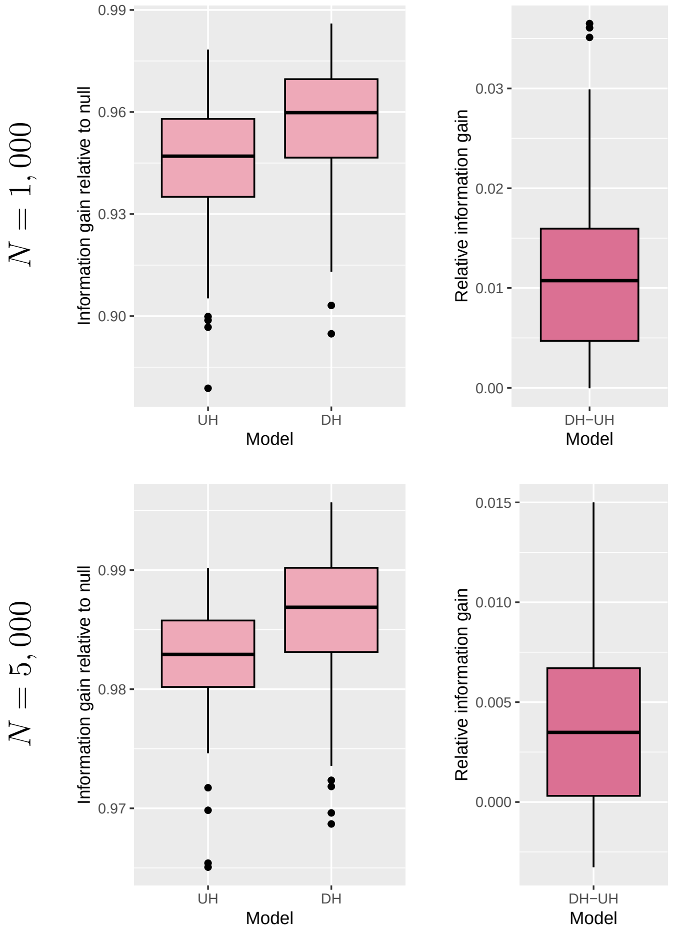


Figure B.1: Simulation Study I.i: Relative information gain (G-squared based) achieved by different models; true partnership surplus utility model is W_{UH} (200 simulations)

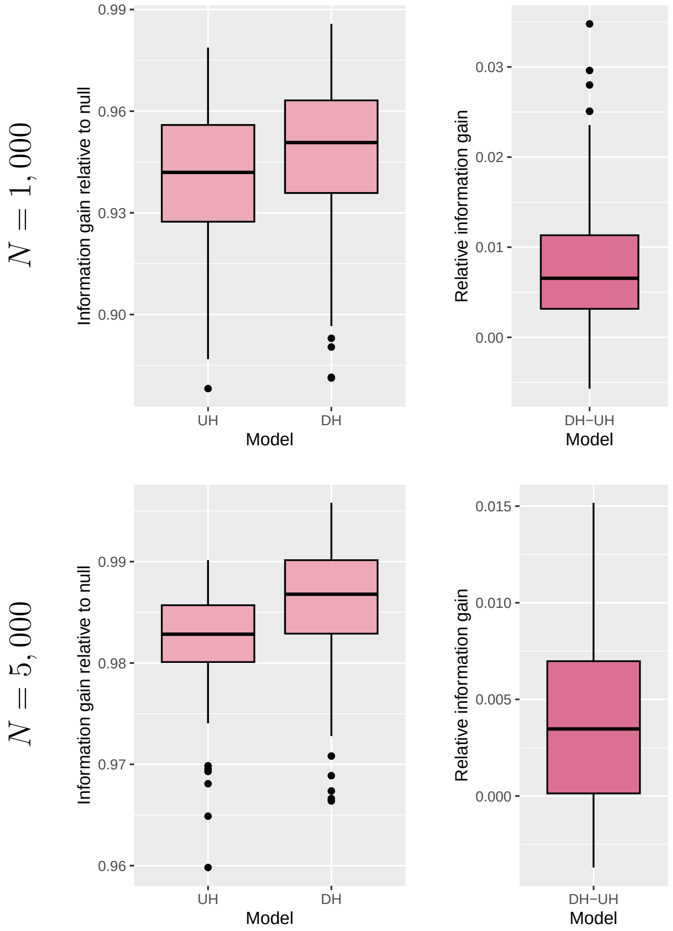


Figure B.2: Simulation Study I.i: Relative information gain (squared Hellinger distance based) achieved by different models; true partnership surplus utility model is W_{UH} (200 simulations)

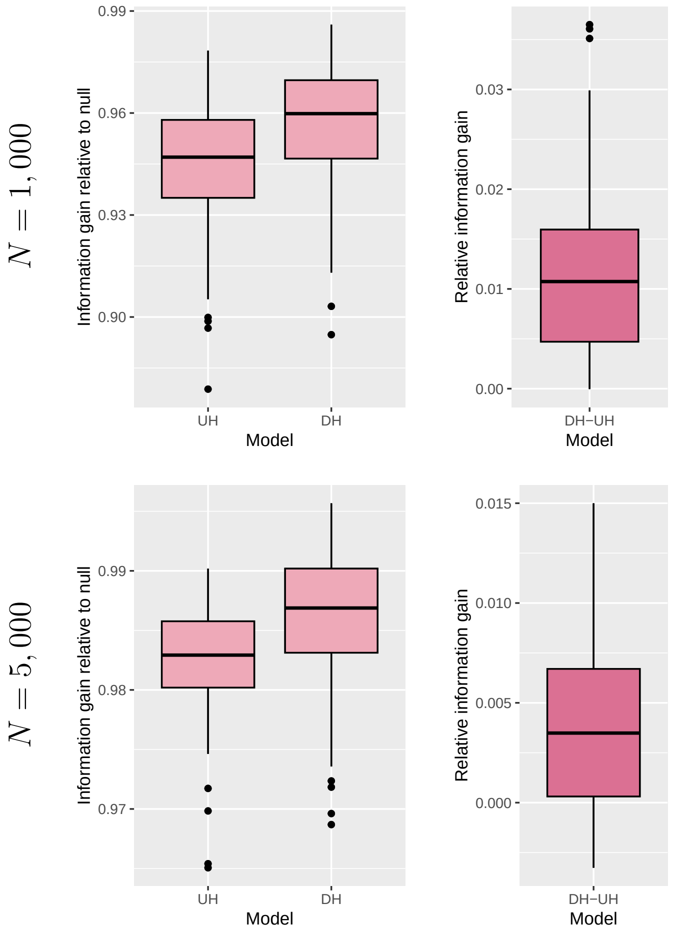


Figure B.3: Simulation Study I.i: Relative information gain (KL divergence based) achieved by different models; true partnership utility model is W_{UH} (200 simulations)

B.3.2 Simulation Study I.ii

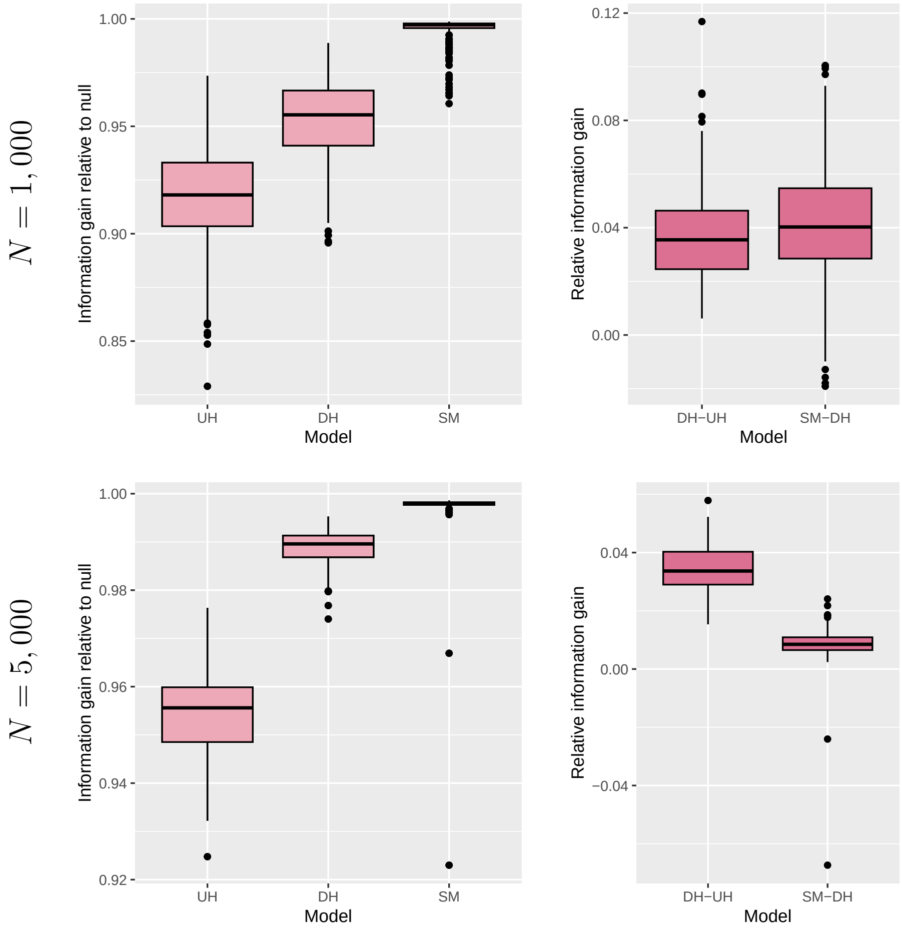


Figure B.4: Simulation Study I.ii: Relative information gain (G-squared based) achieved by different models; true partnership utility model is W_{DH} (200 simulations)

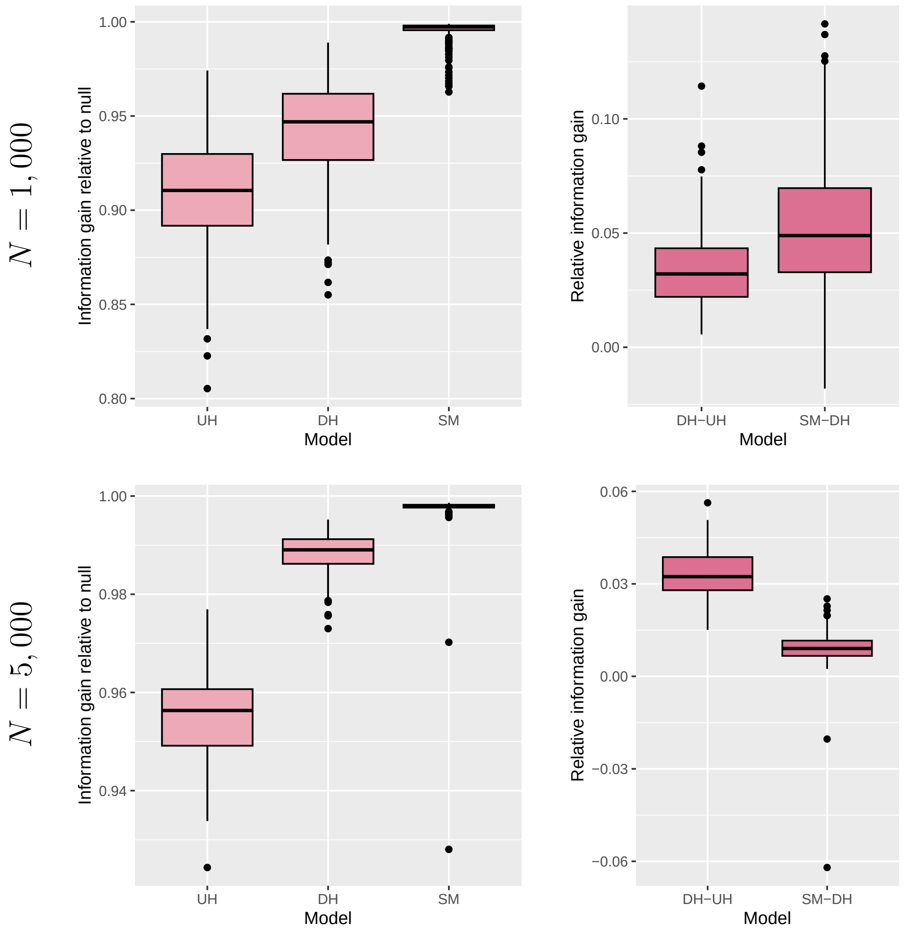


Figure B.5: Simulation Study I.ii: Relative information gain (squared Hellinger distance based) achieved by different models; true partnership utility model is W_{DH} (200 simulations)

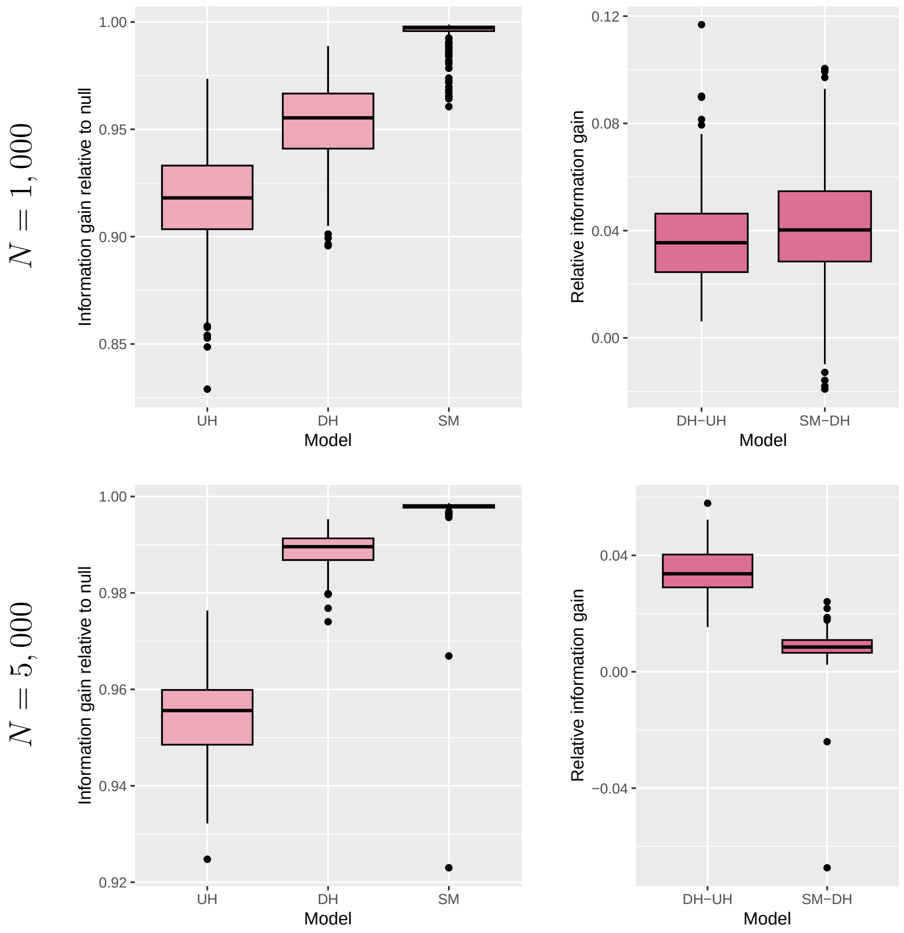


Figure B.6: Simulation Study I.ii: Relative information gain (KL divergence based) achieved by different models; true partnership utility model is W_{DH} (200 simulations)

BIBLIOGRAPHY

Agresti, A. (2002). *Categorical Data Analysis* (2 ed.). Wiley.

Becker, G. S. (1973). A theory of marriage: Part i. *Journal of Political economy 81*(4), 813–846.

Becker, G. S. (1974). A theory of marriage: Part ii. *Journal of Political Economy 82*(2, Part 2), S11–S26.

Brien, M. J. (1997). Racial differences in marriage and the role of marriage markets. *Journal of Human Resources 32*(4), 741–778.

Burnham, K. P. and D. R. Anderson (2004). Multimodel inference: Understanding aic and bic in model selection. *Sociological Methods & Research 33*(2), 261–304.

Chiappori, P.-A. (2020). The theory and empirics of the marriage market. *Annual Review of Economics 12*(1), 547–578.

Chiappori, P.-A., B. Salanié, and Y. Weiss (2017). Partner choice, investment in children, and the marital college premium. *American Economic Review 107*(8), 2109–67.

Choo, E. and A. Siow (2006). Who marries whom and why. *Journal of Political Economy 114*(1), 175–201.

Cressie, N. and T. R. C. Read (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological) 46*(3), 440–464.

Dagsvik, J. K. (1994). Discrete and continuous choice, max-stable processes, and independence from irrelevant attributes. *Econometrica: Journal of the Econometric Society*, 1179–1205.

Dagsvik, J. K. (2000). Aggregation in matching markets. *International Economic Review 41*(1), 27–57.

Dagsvik, J. K., H. Brunborg, and A. S. Flaatten (2001). A behavioral two-sex marriage model. *Mathematical Population Studies 9*(2), 97–121.

Dupuy, A. and A. Galichon (2014). Personality traits and the marriage market. *Journal of Political Economy 122*(6), 1271–1319.

Gale, D. and L. S. Shapley (1962). College admissions and the stability of marriage. *The American Mathematical Monthly 69*(1), 9–15.

Galichon, A. and B. Salanié (2021, 12). Cupid's Invisible Hand: Social Surplus and Identification in Matching Models. *The Review of Economic Studies 89*(5), 2600–2629.

Goyal, S., M. S. Handcock, H. M. Jackson, M. S. Rendall, and F. C. Yeung (2023, 03). A practical revealed preference model for separating preferences and availability effects in marriage formation. *Journal of the Royal Statistical Society Series A: Statistics in Society*, qnad031.

Handcock, M. S., R. M. Admiraal, F. C. Yeung, H. M. Jackson, M. S. Rendall, and S. Goyal (2022). **rpm***: Modeling of Revealed Preferences Matchings*. Los Angeles, CA: University of California, Los Angeles. R package version 0.70.

Hartmann, W. M. and R. E. Hartwig (1996). Computing the Moore-Penrose inverse for the covariance matrix in constrained nonlinear estimation. *SIAM Journal on Optimization 6*(3), 727–747.

Heinze, G. and M. Schemper (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine 21*, 2409–2419.

Johnson, S. G. (2020). *The NLopt nonlinear-optimization package.*

Kalmijn, M. (1994). Assortative mating by cultural and economic occupational status. *American Journal of Sociology 100*(2), 422–452.

Kraft, D. (1994, September). Algorithm 733: Tomp–fortran modules for optimal control calculations. *ACM Trans. Math. Softw. 20*(3), 262–281.

Logan, J. A. (1996a). Opportunity and choice in socially structured labor markets. *American Journal of Sociology 102*(1), 114–160.

Logan, J. A. (1996b). Opportunity and choice in socially structured labor markets. *American Journal of Sociology 102*(1), 114–160.

Logan, J. A., P. D. Hoff, and M. A. Newton (2008). Two-sided estimation of mate preferences for similarities in age, education, and religion. *Journal of the American Statistical Association 103*(482), 559–569.

McCarthy, P. J. and C. B. Snowden (1985). The bootstrap and finite population sampling. *Vital and health statistics. Series 2, Data evaluation and methods research; no 95 January 1985.*

Menzel, K. (2015). Large matching markets as two-sided demand systems. *Econometrica 83*(3), 897–941.

Pollak, R. A. (1986). A reformulation of the two-sex problem. *Demography 23*(2), 247–259.

Pollard, J. H. (1997). Modelling the interaction between the sexes. *Mathematical and Computer Modelling 26*(6), 11–24.

Rendall, M. S., M. M. Weden, and J. Brown (2022). Family and household sources of poverty for black, hispanic, and white newborns. *Journal of Marriage and Family 84*(1), 330–346.

Renyi, A. (1961). On measures of entropy and information. *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability* (3), 547–561.

Roth, A. E. and M. A. O. Sotomayor (1990). *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. Econometric Society Monographs. Cambridge University Press.

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica 5*(17), 61–71.

Samuelson, P. A. (1948). Consumption theory in terms of revealed preference. *Economica 15*(60), 243–253.

Schoen, R. (1981). The harmonic mean as the basis of a realistic two-sex marriage model. *Demography 18*(2), 201–216.

Schwartz, C. R. and R. D. Mare (2005). Trends in educational assortative marriage from 1940 to 2003. *Demography 42*(4), 621–646.

Shao, J. and D. Tu (1995). *The Jackknife and Bootstrap.* Springer Series in Statistics. Springer New York.

Simonoff, J. S. (2003). *Analyzing Categorical Data* (1 ed.). Springer.

U.S. Bureau of the Census (2020). 2008 Survey of Income and Program Participation (SIPP).

Yeung, F. C. (2019). *Statistical Revealed Preference Models for Bipartite Networks.* Ph. D. thesis, University of California at Los Angeles.

Zhang, X. (2022). An awareness model for a two-sided matching market. Master's thesis, University of California at Los Angeles.