California Center for Population Research
University of California - Los Angeles

A Practical Revealed Preference Model for Separating Preferences and Availability Effects in Marriage Formation

Shuchi Goyal and Mark S. Handcock and Fiona C. Yeung†
Department of Statistics, University of California
Heide M. Jackson
Maryland Population Research Center, University of Maryland
Michael S. Rendall
Department of Sociology and Maryland Population Center, University of Maryland College

PWP-CCPR-2020-010

November 12, 2020

California Center for Population Research
On-Line Working Paper Series

A Practical Revealed Preference Model for Separating Preferences and Availability Effects in Marriage Formation

Shuchi Goyal and Mark S. Handcock and Fiona C. Yeung†
Department of Statistics, University of California, Los Angeles, CA, USA
Heide M. Jackson
Maryland Population Research Center, University of Maryland, College Park, Maryland USA
Michael S. Rendall
Department of Sociology and Maryland Population Center, University of Maryland College Park, Maryland, USA

Summary.
Many problems in demography require models for partnership formation that separate latent preferences for partners from the availability of partners. We consider a model for matchings within a bipartite population where individuals have utility for people based on known and unknown characteristics. People can form a partnership or remain unpartnered. The model represents both the availability of potential partners of different types and preferences of individuals for such people. We develop Menzel's (2015) framework to estimate preference parameters based on sample survey data on partnerships and population composition. We conduct simulation studies based on new marriages observed in the Survey for Income and Program Participation (SIPP) to show that, for realistic population sizes, the model recovers preference parameters that are invariant under different population availabilities. We also develop confidence intervals that have correct coverage. This model can be applied in family demography to understand individual preferences given different availabilities.

1. Introduction to the Two-sided Matching Market

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 graduates all require members of two disjoint groups to mutually consent to form a relationship, or match. Yet the underlying mechanisms which dictate such processes are often opaque.

We consider not only how an actor chooses from a set of actors from the opposite side, 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. For example, in the heterosexual marriage market,

†*Address for correspondence:* Department of Statistics, University of California, Los Angeles, Los Angeles, CA 90095-1554, USA. Email: sgoyal25@ucla.edu

women may prefer men who are highly educated. However, a limit in the supply of men with this characteristic means that some women must either choose a partner with lower education levels or remain single. It is important to distinguish the effects of preferences from those of availability in the final matchings realized. This problem has long been recognized in demography without having been satisfactorily resolved (Choo and Siow, 2006; Pollak, 1986; Schoen, 1981; Pollard, 1997).

Menzel (2015) proves a series of new mathematical results related to the asymptotic distribution of matching outcomes in a two-sided market. In this paper we 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-parametrized 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 large 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 paper we focus only on the simpler case in which individuals have at most one partner, also known as one-to-one matchings.

The paper is organized as follows: in Section 2 we provide background information on the general two-sided matching problem and review existing literature which addresses the challenges of identifying individual preferences in such settings. In Section 3 we detail the proposed revealed preferences model and introduce relevant mathematical notation. We also address how we overcome challenges in the identifiability of certain preference parameters. In Section 5 we discuss parameter inference using a pseudo empirical likelihood approach which depends on the sampling process through which the data was obtained. We also describe methods of computing standard errors for parameter estimates and constructing confidence intervals. In Section 6, we demonstrate application of the revealed preferences model. We provide details on two simulation studies in which we attempt to recover known preferences using our proposed method. We present the results of these simulation studies in Section 7 which demonstrate the model's accurate estimation of parameters. We conclude in Section 8 with a discussion regarding the implications of the results and examples of ways the revealed preferences model might be useful in other fields.

2. Background

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, we record the status of all partnerships in a given sample at a discrete time point and assume that this organization of matches is *stable*.

The concept of *stable matchings* has been previously explored in depth by economists and statisticians. Stability is achieved 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. Roth and Sotomayor (1990) show 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 is an accurate reflection of individual and societal preferences at that time point.

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. 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 his outcome, although his decision may be affected by the decisions of other choosers around him. 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 the 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 reciprocal. 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.

Logan et al. (2008) and Menzel (2015) both propose a two-sided version of the discrete choice model to estimate preference parameters in matching markets. Logan et al. (2008) propose a model where disjoint groups in the population have distinct, though possibly parallel, utility functions. For example, in the case of heterosexual marriages, all men have the same deterministic utility function which depends on the man's observed characteristics x and the characteristics of his partner z, and all women have the same deterministic utility function which depends on the woman's observed characteristics z and the characteristics of her partner x. Here, $x \in \mathcal{X}$ and $z \in \mathcal{Z}$. The sample spaces $\mathcal{X}$ and $\mathcal{Z}$ represent the set of possible "types" of men and women, 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. By supposing a small population, Logan et al. (2008) are able to assume that the full opportunity sets for all actors are known.

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. However, since the computation of the actual MLE is often complex and involves an integral which may be intractable, they suggest approximating MLEs using 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 for 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.

Menzel (2015) studies the two-sided matching problem with a goal of analyzing the distribution of observable outcomes. Here, observable outcomes are the possible matchings which may occur. For example, in the case of the marriage market, we may conceptualize outcomes as different households. Households are broadly characterized as either "single" or "partnered," depending on whether they hold a single person or a married couple. 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. Each 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.

An important result of Menzel (2015) is the derivation of equations which allow asymptotically stable estimates of the proportions of single and partnered households of each type in the population. 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. In a small population, there is an interaction effect between preferences and partner availabilities which influences the observed matching. For example, a man's preference for a highly educated female spouse may result in more females pursuing higher education. We extend the results of Menzel (2015) to derive equations which establish a relationship between the preferences θ and availabilities of men and women of each type in the population and the limiting distribution of households across the possible outcomes. These calculations prove that in a large population, the dependency between availability and preferences is negligible, and therefore that preferences can be recovered independently of the population availability context.

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) 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 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 matchings in a large population and applies them to estimate preference parameters based on an observed distribution of matching.

We note that previous work on decision-making in a matching market have assumed transferable utility among agents (e.g. Choo and Siow, 2006). For this paper, we follow Logan et al. (2008) and Menzel (2015) and assume a non-transferable utility (NTU) framework. In NTU setting, an agent's observed attributes remain unchanged upon match formation and dissolution. This assumption is not only realistic, but also greatly simplifies the discussion that follows.

3. 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 must be 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 single (unpartnered). Both the male and the female must agree to the partnership for that partnership, or "marriage," to be observed.

Individuals evaluate their marital options using a utility function, which contains a deterministic and random component. Actors of the same gender are assumed to have identically specified utility functions. 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 utility. The latent parameters of the utility function which govern this pair formation are commonly known as "preference" parameters in the broad sense that they represent how actors would choose among different alternatives if given a choice (Roth and Sotomayor, 1990).

We consider a population with N_w women and N_m men, so that the total population size is $N = N_w + N_m$. Using the same notation introduced in Section 2, we observe a $p-$vector of covariates $x \in \mathcal{X}$ on the women and a $q-$vector of covariates $z \in \mathcal{Z}$ on the men. 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 later simulation study examples where we apply the revealed preferences model, we assume that x and z are discrete and have length 1.

Actors may perceive potential partners differently based on their own characteristics. Thus, the perceived utility gained by partnering with the same individual of the opposite sex may differ from one decision maker to the next. However, all actors are assumed to choose the partner within their respective choice sets that can provide the maximum gain in 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

$$U_{ij} = \underbrace{U(x_i, z_j|\theta_W)}_{\substack{\text{deterministic}\\ \text{component}}} \quad + \quad \underbrace{\eta_{ij}}_{\substack{\text{unobserved random}\\ \text{component}}} \tag{1}$$

where θ_W is the set of parameters denoting the woman's preferences. They can be individually specific, and we focus on the case where the parameters are common to all women. Similarly, we define the utility gained by man j with observed attributes z_j from partnering with woman i with observed attributes x_i as

$$V_{ji} = \underbrace{V(z_j, x_i|\theta_M)}_{\substack{\text{deterministic}\\ \text{component}}} \quad + \quad \underbrace{\zeta_{ji}}_{\substack{\text{unobserved random}\\ \text{component}}} \tag{2}$$

where θ_M is the set of parameters representing men's preferences.

Following Menzel (2015), we assume that unobserved random components of the utility functions as defined in Equations (1) and (2) are independently and identically distributed draws from a distribution in the domain of attraction of the extreme-value type-I (Gumbel) distribution. This includes Exponential, Gamma, Gaussian, Lognormal, and Weibull. Here we will focus on the Gumbel itself, but note our model and methods are more general.

3.1. 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 (1) and (2) are additive linear functions; however, other choices of utility functions can also be used.‡

For additive linear utility functions, let

$$\begin{aligned} U(x_i, z_j|\theta_W) &= \theta_{w0} + \sum_{k=1}^{K_w} \theta_{wk} X^k(x_i, z_j) \\ V(z_j, x_i|\theta_M) &= \theta_{m0} + \sum_{k=1}^{K_m} \theta_{mk} Z^k(x_i, z_j) \end{aligned} \tag{3}$$

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_{m0} and K_w additive linear functions. Each of these functions $X^k(x_i, z_j)$ represents a portion of woman i's total utility which is derived from 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 (e.g. homophilous). The corresponding K_m functions for the man's side are denoted as $Z^k(x_i, z_j)$. Here $\theta_W = [\theta_{w0}, \theta_{w1}, \ldots \theta_{wK_w}]^T$ and $\theta_M = [\theta_{m0}, \theta_{m1}, \ldots \theta_{mK_m}]^T$ are the preference parameters, which are vectors of the scalar coefficients in the utility functions.

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

$$\begin{aligned} U_{i0} &= 0 + \max_{k=1,\ldots,N_m^\delta} \{\eta_{i0,k}\} \\ V_{j0} &= 0 + \max_{k=1,\ldots,N_w^\delta} \{\zeta_{j0,k}\} \end{aligned} \tag{4}$$

for females and males, respectively.

The single household utility specification in Equation (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 (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).

We choose the hyperparameter δ based on prior expectations of how the proportion individuals in the population who are single will change as the market size increases. For this model, we set $\delta = 1/2$. This specification ensures that the share of singles in the market stays constant as the market grows large (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.

3.2. 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 $\bar{m}(x) = m(x)/N$.

‡See Dagsvik (1994) for latent choice set derivation for other choices of utility functions.

Consider a population with utilities drawn from the the model (1), (2), (3) and (4). Then the stable matching induces a probability distribution over the observed characteristics. Consider sampling a random person from the population and their classification of matched or single. Let $f(x,*)$ and $f(*,z)$ be the densities of unmatched women of type x, and unmatched men of type z, respectively. Let $f(x,z)$ be the joint density of the matches between women of observed characteristics x and men of type z. 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 satisfing the overall normalization constraint:

$$\int f(x,z)dxdz + \int f(x,*)dx + \int f(*,z)dz = 1 \tag{5}$$

More specifically,

$$\begin{aligned} \bar{w}(x) &= f(x,*) + f(x,\diamond) \\ \bar{m}(x) &= f(*,z) + f(\diamond,z) \end{aligned} \tag{6}$$

where $f(x,\diamond)$ is the probability of being partnered:

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

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:

$$f(x,z) = 2e^{W(x,z|\boldsymbol{\beta})} f(x,*)f(*,z) \qquad \forall x,z \tag{7}$$

where

$$W(x,z|\boldsymbol{\beta}) = U(x,z|\theta_W(\boldsymbol{\beta})) + V(z,x|\theta_M(\boldsymbol{\beta})), \quad \forall x \in \mathcal{X}, z \in \mathcal{Z}$$

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

$$\begin{aligned} \frac{f(x,\diamond)}{f(x,*)} &= \int e^{W(x,s|\boldsymbol{\beta})} f(*,s)ds \quad \forall\ x \\ \frac{f(\diamond,z)}{f(*,z)} &= \int e^{W(s,z|\boldsymbol{\beta})} f(s,*)ds \quad \forall\ z \end{aligned} \tag{8}$$

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

Together, (6) and (7) make it possible to obtain estimates $\hat{\boldsymbol{\beta}}$ of the preference parameters.

4. Data

The analysis depends on the sampling design that produces the data. Let $c(x,*)$ and $c(*,z)$ be the design-based estimates of the numbers of unmatched women of type x, and unmatched men of type z in the population, respectively. Let $c(x,z)$ be the design-based estimates of the number of matches between women of observed characteristics x and men of type z in the population. 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 $\bar{f}$. Our method can be applied with a broad range of complex survey sampling designs, with the requirement that they produce estimates of $\bar{f}$. Here we focus on the situation where the data are a probability sample of the individuals in a population where the weights are w_{wi} for the i^{th} woman and w_{mj} 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 $\{z_j, x_i, w_j^m\}_{j=1}^{n_m}$. In our analysis we use the standard Hájek estimator (Hájek, 1971).

If the population size N is large and the sample fraction is not high, we will focus inference on the near sufficient statistics $\bar{c}$ for the distribution $\bar{f}$. In our experience, we offer as benchmarks $N > 7000, n < N/2$ as sufficient to have this approximation be very accurate. We provide evidence for these guidelines in Section 6.

4.1. Parametrization and Identifiability

Following Logan et al. (2008), we say that a parametrization 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 (7) and (8), and the expression

$$W(x, z|\boldsymbol{\beta}) = U(x, z|\theta_W(\boldsymbol{\beta})) + V(z, x|\theta_M(\boldsymbol{\beta})), \quad \forall x \in \mathcal{X}, z \in \mathcal{Z}$$

only the sum of the partnered individuals' utilities is identifiable, and the individual components $U(x, z|\theta_W)$ and $V(z, x|\theta_M)$ are not. For example, suppose that men and women both have a preference for homophily, meaning that an individual gains additional utility from a partner of the same "type" as him- or herself. The deterministic component of the utility for woman i when she partners with man j is given by

$$U(x_i, z_j|\theta_w) = \theta_w X(x_i, z_j),$$

where $X(x_i, z_j) = \mathbb{I}\{x_i = z_j\}$ is an indicator function that equals 1 if woman i and man j have the same observed characteristics and 0 otherwise. Furthermore, the deterministic utility for man j when he partners with woman i is given by

$$V(z_j, x_i|\theta_m) = \theta_m Z(z_j, x_i),$$

where $Z(z_j, x_i) = \mathbb{I}\{x_i = z_j\}$ is also an indicator function that equals 1 if man j and woman i have the same observed characteristics and 0 otherwise. Then, it is always true that $X(x_i, z_j) = Z(z_j, x_i)$. In this case where individuals show preference for homophily, the deterministic value of the total household utility is

$$\begin{aligned} W(x_i, z_j|\boldsymbol{\beta}) &= U(x_i, z_j|\theta_w) + V(z_j, x_i|\theta_m) \\ &= \theta_w X(x_i, z_j) + \theta_m Z(z_j, x_i) \\ &= (\theta_w + \theta_m)\mathbb{I}\{x_i = z_j\} \\ &= \beta\mathbb{I}\{x_i = z_j\}. \end{aligned} \tag{9}$$

We see that while θ_w and θ_m are not separately identifiable, their sum $\beta = \theta_w + \theta_m$ is. More broadly, $U(x, z|\theta_W)$ and $V(z, x|\theta_M)$ may not be separably identifiable when they are additive linear functions as in Equation (3) and include parallel terms. In general, let $\theta_W(\boldsymbol{\beta})$ and $\theta_M(\boldsymbol{\beta})$ be functions such that

$$W(x,z|\boldsymbol{\beta}) = U(x,z|\theta_W(\boldsymbol{\beta})) + V(z,x|\theta_M(\boldsymbol{\beta})), \quad \forall x \in \mathcal{X}, z \in \mathcal{Z}$$

For example, if the utility functions are additive and linear (Equation 3), $\boldsymbol{\beta} = \theta_W(\boldsymbol{\beta}) + \theta_M(\boldsymbol{\beta})$. In this case, $W(x,z)$ can be parameterized in terms of $\boldsymbol{\beta}$. We will consider paramatrizations where $\boldsymbol{\beta}$ is identifiable. To emphasize the relationship between $\boldsymbol{\beta}, \theta_W$, and θ_M, we refer to the gender-specific preference parameters as $\theta_W(\boldsymbol{\beta})$ and $\theta_M(\boldsymbol{\beta})$ for the rest of this paper.

4.2. Reparametrization of the model

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

$$f(x,*) = \frac{\bar{w}(x)e^{g(x,*)}}{(1+e^{g(x,*)})} \tag{10}$$
$$f(*,z) = \frac{\bar{m}(z)e^{g(*,z)}}{(1+e^{g(*,z)})}$$

so that $g(x,*)$ and $g(*,z)$ both have range the real line. We can interpret $g(x,*)$ as the log-odds that a women with characteristics x is single. Similarly, we can interpret $g(*,z)$ as the log-odds that a men with characteristics z is single. We will use $g(x,*)$ and $g(*,z)$ in place of $f(x,*)$ and $f(*,z)$ to ease computation and interpretability. Note that

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

so that (6) is automatically satisfied and (7) becomes

$$f(x,z) = 2\frac{e^{W(x,z)+g(x,*)+g(*,z)}}{[1+e^{g(*,z)}][1+e^{g(x,*)}]}\bar{w}(x)\bar{m}(z) \qquad \forall x,z \tag{7'}$$

so that

$$2\frac{e^{W(x,z)+g(x,*)+g(*,z)}}{[1+e^{g(*,z)}][1+e^{g(x,*)}]} \qquad \forall x,z$$

expresses the preferences related component of the model. In this parametrization (8) becomes

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

5. Inference

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 $\boldsymbol{\psi} = (\boldsymbol{\beta}, \{g(x,*)\}_{x\in\mathcal{X}}, \{g(*,z)\}_{z\in\mathcal{Z}})$.

5.1. Pseudo Likelihood Approach

Had we observed the entire population, the likelihood for $\boldsymbol{\psi}$ 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 $\boldsymbol{\psi}$, one based on the likelihood of the observed frequencies of pairings by covariates, $\bar{c}$, and the model (7) and (8). Specifically, the population likelihood for $\boldsymbol{\psi}$ is:

$$\text{log-lik}_{pop}(\boldsymbol{\psi}|\{x_i, z_i, w_i^w\}_{i=1}^{N_w}, \{z_j, x_i, w_j^m\}_{j=1}^{N_m}) = \sum_{i=1}^{N_w} \log f(x_i, z_j) + \sum_{j=1}^{N_m} \log f(x_i, z_j) \quad (11)$$

However, we do not observe the full population and so we approximate the population likelihood by the design-based estimator:

$$\begin{aligned} &\text{p-log-lik}(\boldsymbol{\psi}|\{x_i, z_i, w_i^w\}_{i=1}^{n_w}, \{z_j, x_i, w_j^m\}_{j=1}^{n_m}) \quad (12) \\ &= \sum_{x\in\mathcal{X}} \sum_{z\in\mathcal{Z}} c(x,z) \log f(x,z) + \sum_{x\in\mathcal{X}} c(x,*) \log f(x,*) + \sum_{z\in\mathcal{Z}} c(*,z) \log f(*,z) \end{aligned}$$

This approach is based on the arguments of Godambe and Thompson (1986). The log-likelihood (12) 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, the pseudo log-likelihood can be maximized subject to the constraints expressed in (8') to produce the pseudo maximum likelihood estimator (PMLE), $\hat{\boldsymbol{\psi}}$. This was achieved via a sequential quadratic programming (SQP) algorithm for non-linearly constrained gradient-based optimization (Kraft, 1994; Johnson, 2020). 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.

5.2. Measuring uncertainty of the estimates

Once we obtain the parameter estimates $\hat{\boldsymbol{\psi}}$, 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 pseudo log-likelihood function, argumented by the constraints, is

$$\begin{aligned} &\text{log-lik}_A(\boldsymbol{\psi}|\{x_i, z_i, w_{wi}\}_{i=1}^{n_w}, \{z_j, x_i, w_{mj}\}_{j=1}^{n_m}) \quad (13) \\ &= \text{p-log-lik}(\boldsymbol{\psi}|\{x_i, z_i, w_{wi}\}_{i=1}^{n_w}, \{z_j, x_i, w_{mj}\}_{j=1}^{n_m}) + \sum_{k=1}^{|\mathcal{X}|+|\mathcal{Z}|+1} \lambda_k h_k(\boldsymbol{\psi}) \quad (14) \end{aligned}$$

and its Hessian is

$$\mathbb{E}\left(\frac{\partial^2 \text{log-lik}_A}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'}\right) = \begin{pmatrix} H & J \\ J^T & 0 \end{pmatrix} \quad (15)$$

where H is the Hessian of the pseudo log-likelihood with ij^{th} element $\mathbb{E}\left(\frac{\partial^2 \text{p-log-lik}}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'}\right)$ and J is the matrix Jacobian of the constraints with kj^{th} element $\frac{\partial h_k(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}$. The estimate of the (asymptotic) covariance matrix of pseudo MLE of $\boldsymbol{\psi}$ 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. As an alternative, we propose estimating standard errors empirically using bootstrap procedures. We first sample the households of k individuals with repetition from the observed sample, where k is equal to the number of directly sampled individuals in the original sample. We repeat this process b times, so that we have b sets of bootstrapped samples. We fit the revealed preferences model to each of the b samples and obtain the bootstrapped parameter estimates for a single parameter $\boldsymbol{\psi}$, which we denote as $\boldsymbol{\psi}^* = [\boldsymbol{\psi}^*_{(1)}, \boldsymbol{\psi}^*_{(2)}, \ldots, \boldsymbol{\psi}^*_{(b)}]$. The empirically estimated standard error of $\hat{\boldsymbol{\psi}}$, denoted as $\widehat{\text{se}}_{\hat{\boldsymbol{\psi}}}$, is equal to standard error of the bootstrapped parameter estimates $\boldsymbol{\psi}^*$.

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 $\boldsymbol{\psi}^*_{(\alpha)}$ as the α percentile of the bootstrap parameter estimates $\boldsymbol{\psi}^*$. The $(1-\alpha)\%$ percentile bootstrap confidence interval for parameter $\boldsymbol{\psi}$:

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

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

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

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

$$(\hat{\boldsymbol{\psi}} - t^*_{(1-\alpha/2)}\widehat{\text{se}}_{\hat{\boldsymbol{\psi}}}, \hat{\boldsymbol{\psi}} - t^*_{(\alpha/2)}\widehat{\text{se}}_{\hat{\boldsymbol{\psi}}}).$$

We test the performances of the analytical confidence intervals as well as those of all three proposed bootstrap confidence interval methods in Section 7.3 as part of our simulation studies.

6. Simulation Studies of Model and Inferential Accuracy

In this section we describe two simulation studies which demonstrate that the revealed preferences model is able to accurately estimate the underlying preferences which partially motivate matching outcomes in a population. The basic procedure for both 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 then simulate a population from an availability scenario with a known marginal distribution of gender and education and create stable partnerships among the simulated individuals based on utilities computed using known preference parameters $\boldsymbol{\beta}$. We fit the revealed preferences model to the observed matching outcomes in the simulated population and show that the model reconstructs the original preference parameters.

To achieve a stable matching in a simulated population, we would ideally use the Gale-Shapley algorithm. However, a large amount of memory and computational power is required to create stable partnerships for large population sizes (greater than 7,000), 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 Simulation study B, we suppose a population whose size is arbitrarily large. In this case, rather than implementing the Gale-Shapley algorithm to achieve a stable matching, we approximate the distribution of household types in the outcome and estimate preference parameters based on the large population approximation (Equation (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 7,000 individuals. To show that the revealed preferences model can still recover true parameters given an observed, rather than approximated, distribution of outcomes, we also run a second simulation study, which we call Simulation study A, under a small population setting such that the population size is $N = 7,000$.

For each simulation study, we consider three distinct availability scenarios with differing marginal availabilities from which populations are simulated. These scenarios are described further in Section 6.1. Additionally, for each availability scenario, we consider two different specifications of the deterministic total partnership utility $W(x_i, z_j|\boldsymbol{\beta})$. These models are detailed in Section 6.2.

Both the known marginal availability distributions of the availability scenarios and the known underlying preference parameters $\boldsymbol{\beta}_0$ for each model specification are determined based on data from the 2008 Survey on Income and Program Participation (SIPP), which has been made publicly available by the United States Census Bureau (U.S. Bureau of the Census, 2020b,a). 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 at wave 2 had married in the past year or were not currently 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 new marriages so as to measure 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). Within a given year, entering into a marriage is relatively rare, only 5% of individuals in our analytic sample entered a new marriage and thus preferences for marriage are negative when we run the revealed preferences model in Section 7.

The maximum education level attained by each individual is a categorical variables 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.

6.1. Description of Availability Scenarios

We assume three separate availability scenarios, referred to hereafter as availability scenario 1, availability scenario 2 and availability scenario 3. The marginal availabilities in each population are provided fully in Table 2. For each setting we describe a population generating process. One of these scenarios is factual (a populations like the 2008 SIPP), and the two others are counter-factual (i.e., changing the population composition while retaining preferences). In the latter two cases, we reconstruct matchings using the preferences of the 2008 SIPP sample while changing the availabilities of the population. *SG:* [Check placement of Table 1.]

Table 1: The three availability scenarios

Availability scenario	Source of availability distribution	Type
1	2008 SIPP full sample	Total U.S. population in 2008
2	2008 SIPP non-Hispanic Black sample	A realistic sub-population availability
3	Artificial	An extremely mismatched population

Table 2: Gender and Education Distributions under the three availability scenarios

	Males		Females	
Education Level	% Population	% of Males	% Population	% of Females
	Availability scenario 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 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**
	Availability scenario 3			
1 (< high school)	45.0	60.0	2.5	10.0
2 (high school)	15.0	20.0	2.5	10.0
3 (some college)	7.5	10.0	5.0	20.0
4 (≥ bachelors)	7.5	10.0	15.0	60.0
Total	**75.0**	**100.0**	**25.0**	**100.0**

Availability scenario 1 utilizes the gender and education distributions of the overall population based on the restricted 2008 SIPP sample. In this availability scenario, about 49.1% of individuals are women and 51.9% are male.

Availability scenario 2 has the same marginal distribution of education and availability as the non-Hispanic Black population in the restricted 2008 SIPP data. In Availability scenario 2, about 58.0% of the individuals are females and 42.0% are males, which reflects a significant gender skew not seen in Availability scenario 1. In both Availability scenarios 1 and 2, women are more likely to have completed any college (education category 3 or higher) and are less likely to have less than a high school degree (education category 1).

Availability scenario 3 is not based on any known sample and is extremely unrealistic. 25% of individuals are female, and 75% of individuals are male. Females tend to have high education levels, with 60% categorized as having education level 4 and 20% categorized as education level 3. Conversely, men are more likely to have lower education levels, with 60% being categorized as having education level 1 and 20% being categorized as education level 2. This asymmetry in gender and education availabilities is highly unusual in observed populations and creates incongruity in the types of partners who are preferred versus those who are available. The study of Availability scenario 3 is to test if the revealed preferences model can successfully recovers preference parameters even in cases where the availability of individuals in the population is highly skewed.

6.2. Utility model specification

For each availability scenario, we test the performance of the revealed preferences model under two different model specifications. The testing procedure for each model specification is similar. We first obtain a set of preference parameters $\boldsymbol{\beta}_0$ which we assume is the underlying truth. This is done by running the specified model on the 2008 SIPP data and calculating parameter estimates $\tilde{\boldsymbol{\beta}}$. We assume that these estimates are equivalent to the true preference parameters of individuals simulated from every availability scenario, so that $\boldsymbol{\beta}_0 = \tilde{\boldsymbol{\beta}}$. In each simulated population, the known preferences $\boldsymbol{\beta}_0$ are applied to calculate total household utility for every potential partnership and form a stable matchings. We fit the revealed preferences model on the observed stable matching outcome from the simulated population and compare the parameter estimates $\hat{\boldsymbol{\beta}}$ to the underlying true preferences $\boldsymbol{\beta}_0$.

We first consider a model specification assuming that the utility a woman derives from a partnership is based on her own education level and whether her partner shares that same education level. There is a corresponding utility function for males. We refer to this as a *type-based match model,* because preference is based on an individual's own type and whether or not their partner's type matches theirs. The set of parameters for this model is denoted as $\boldsymbol{\beta}_{match}$.

Let

$$X^k(x_i, z_j) = Z^k(z_j, x_i) = \mathbb{I}\{x_i = z_j = k\}.$$

The deterministic component of woman i's utility when she is partnered with man j is

$$U(x_i, z_j|\theta_W(\boldsymbol{\beta}_{match})) = \theta_{w0} + \sum_{k=1}^{4} \theta_{wk} X^k(x_i, z_j). \tag{16}$$

Similarly, the deterministic component of the utility of man j when partnered with woman i is

$$V(z_j, x_i|\theta_M(\boldsymbol{\beta}_{match})) = \theta_{m0} + \sum_{k=1}^{4} \theta_{mk} Z^k(z_j, x_i). \tag{17}$$

Then, the total utility of woman i and man j if they partnered with each other is given by the sum of Equations 16 and 17:

$$\begin{aligned} W_{ij}(x_i, z_j|\boldsymbol{\beta}_{match}) &= \theta_{w0} + \theta_{m0} + \sum_{k=1}^{4}(\theta_{wk} + \theta_{mk})\mathbb{I}\{x_i = z_j = k\} \\ &= \beta_0 + \sum_{k=1}^{4} \beta_k \mathbb{I}\{x_i = z_j = k\}, \end{aligned} \tag{18}$$

where $\beta_t = \theta_{wt} + \theta_{mt}$.

The second 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 $\boldsymbol{\beta}_{mix}$. In this case, woman i's utility from partnering with man j is

$$U(x_i, z_j|\theta_W(\boldsymbol{\beta}_{mix})) = \theta_{w0} + \sum_{p=1}^{4}\sum_{q=1}^{4} \theta_{w(p,q)} X^{(p,q)}(x_i, z_j), \tag{19}$$

where

$$X^{(p,q)}(x_i, z_j) = \mathbb{I}\{x_i = p, z_j = q\}.$$

Similarly, man j's utility for partnering with woman i is

$$V(z_j, x_i|\theta_M(\boldsymbol{\beta}_{mix}) = \theta_{m0} + \sum_{q=1}^{4}\sum_{p=1}^{4} \theta_{m(p,q)} Z^{(q,p)}(x_i, z_j), \tag{20}$$

where

$$Z^{(q,p)}(z_j, x_i) = \mathbb{I}\{z_j = q, x_i = p\}.$$

We are able to remove the intercept terms θ_{m0} and θ_{w0} in Equations 19 and 20 because they are constant values added to the matching utility of every individual. Thus, the total utility of the individuals in a marriage is

$$W(x_i, z_j|\boldsymbol{\beta}_{mix}) = \sum_{p,q} \beta_{p,q}\mathbb{I}\{x_i = p, z_j = q\}. \tag{21}$$

The term $\beta_{p,q}$ is the coefficient to an indicator which equals 1 if the household pairing consists of a woman of type p and a man of type q, and 0 otherwise. The full mix model consists of $P \times Q$ first-order parameters, where there are P possible types for women and Q possible types for men.

Out of the 21,077 households in the SIPP analytic sample, there is 1 household which contains a woman with education level 1 and a man with education level 4, and 1 household 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 households is perceived as effectively negatively infinite. To faciliate 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. Similarly, 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.

7. Results

7.1. Simulation study A: A Small Population

In this simulation study, we simulate 1,000 populations of size $N = 7,000$ from each availability scenario. We use the Gale-Shapley algorithm to perform stable matching on the individuals in each simulated population. The utility derived from each potential partnership is calculated based on β_0 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 1 show the distribution of the 1,000 parameter estimates for each combination of availability scenario and revealed preferences model specification. The red lines in the plots represent the true values β_0 which induced the Gale-Shapley matchings.

The box plots in Figure 1 were constructed to include negatively infinite estimates via a point mass at value -6 with area proportional to the number of negative infinite estimates. This was done to ensure they were recognized in the results.

The means and standard errors of parameter estimates for the match and reduced mix models are presented in Tables 3 and 4, respectively, under Appendix A. We note that although availability of individuals differs between Availability scenario 1 and Availability scenario 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 small population plots in Figure 1, the median 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, in Availability scenarios 1 and 2, the estimates for all parameters, with the exception of $\beta_{1 \text{ or } 2,4}$, resemble a normal distribution.

We note that for all the availability scenarios, the distribution of $\hat{\beta}_{1 \text{ or } 2,4}$ displays a right skew. When the population has very few or no households of a certain type, the model estimates the total utility of such a household as very negative, if not infinitely so.

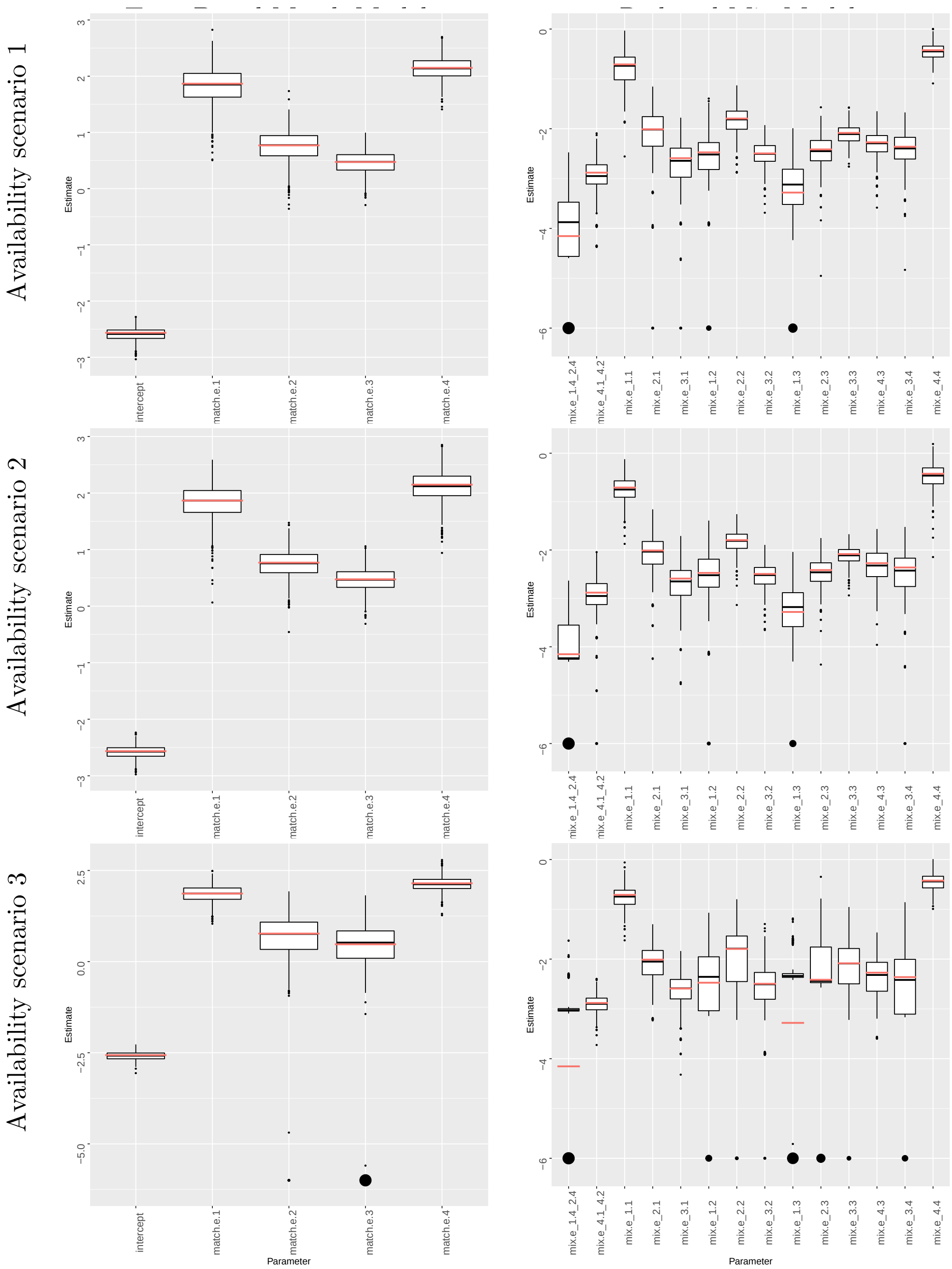

Fig. 1: Distribution of parameter estimates in Simulation study A (small populations); 1,000 simulations, $N = 7,000$

In our implementation of this model, we impose an upper bound of 6 and a lower bound of -6 on all parameters. The high frequency of extremely negative values (< 4) 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 note that the occurrence of highly negative estimates of $\beta_{1 \text{ or } 2,4}$ increases as the gender and education distributions become more skewed. Furthermore, in Availability scenario 3, where men far outnumber women, the estimates of $\beta_{1,3}$ and $\beta_{2,3}$ also develop a right skew. Table 4 in Appendix A shows that the standard errors of these parameter

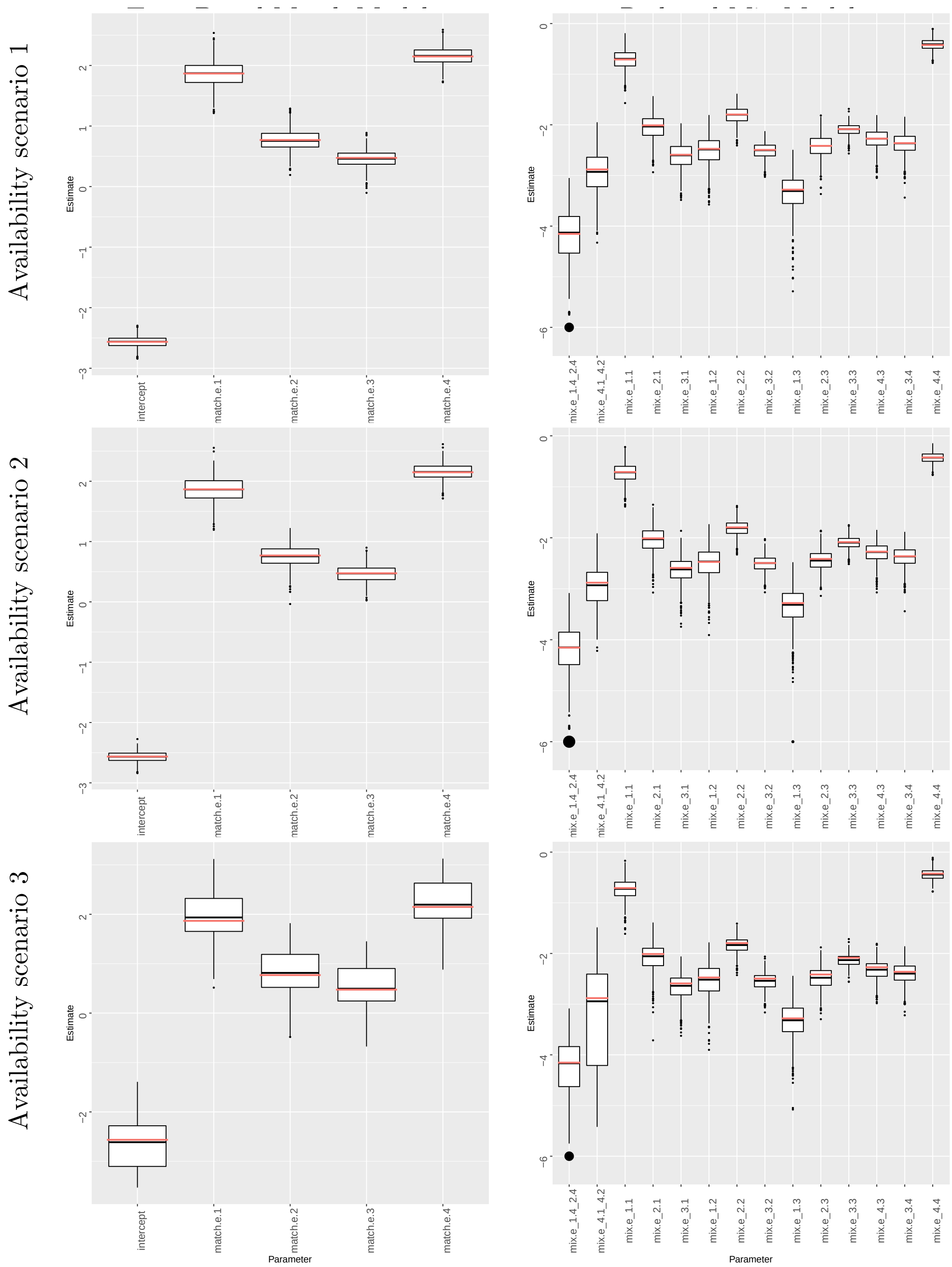

Fig. 2: Distribution of parameter estimates in Simulation study B (large populations); 1,000 simulations, $N = 300$ million

estimates tends to increases as the population becomes more skewed.

The means and standard errors of the match model parameter estimates are provided in Table 3 of Appendix A. We note that although availability of individuals differs between the three availability scenarios, the revealed preferences model produces estimates of the true preference parameters which are comparable in accuracy and precision.

7.2. Simulation study B: A Large Population

In this simulation study, we simulate samples from 1,000 large populations using the specified availabilities, each with a nominal size of $N = 300$ million. The results are very robust to the population size as long as it is modestly large (e.g., $N > 7000$). We choose to study large populations as they are typical in demography.

We employ a large population approximation of stable matching outcomes in the simulated population that would be observed if individuals had true preferences $\boldsymbol{\beta}_0$. The plots in Figure 2 show the distribution of the 1,000 parameter estimates $\hat{\boldsymbol{\beta}}$ for each combination of simulating availability scenario and revealed preferences model specification. The red lines in the plots represent the true values $\boldsymbol{\beta}_0$ which we are attempting to recover.

The first column of Figure 2 shows the distributions of the parameter estimates under the type-based match model given large simulated population. The means and standard errors of the match model parameters are presented in Table 5. To compute these numerical summaries, we again exclude the negative infinite parameter estimates.

In all three availability scenarios, we observe that the mean estimate for each parameter is very close to the true value. We also note that when simulating from Availability scenarios 1 and 2, the standard errors of the parameter estimates stay about the same. However, the standard error nearly triples when the simulated populations are drawn from Availability scenario 3.

The second column of Figure 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 6, which shows the means and standard errors of the parameter estimates, to Appendix A. The revealed preferences model recovers the true preference parameters $\boldsymbol{\beta}_{mix,0}$ for all availability scenarios. Furthermore, the standard errors of all parameter estimates except $\hat{\beta}_{4,1 \text{ or } 2}$ stay similar across the availability scenarios. The standard error of $\hat{\beta}_{4,1 \text{ or } 2}$ is 0.388 and 0.385 for Availability scenarios 1 and 2, respectively, but more than doubles to 0.890 in the Availability scenario 3 setting.

7.3. Confidence intervals and coverage probabilities

To supplement the findings in Simulation study B, we calculate 95% confidence intervals for parameter estimates based on simulations with population size $N = 300$ million and compare the empirical coverage rates of the true parameter values to the 95% threshold.

To calculate empirical coverage rates, we simulate $S = 200$ large populations from scenario 1. For each simulated population, we fit the reduced mix model and produce analytical 95% confidence intervals based on the approximated Hessian matrix, as detailed in Section 5.2. We additionally implement the basic, percentile, and modified studentized t bootstrap methods also discussed in Section 5.2 to construct empirical 95% confidence intervals. An illustration of the coverage results from a single set of 200 simulations are presented in Appendix B.

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. The analytical confidence intervals appeared 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 10 to 90%.

We show the mean coverage rates of the reduced mix model parameters by the bootstrap confidence intervals in Figure 3. The dotted black line at 0.95 denotes the 95% threshold we aim to achieve. For all parameters other than $\beta_{1 \text{ or } 2,4}$ and $\beta_{4,1 \text{ or } 2}$, the mean coverage rates from all three confidence interval types are generally close to 95%. We note, for example, that the mean coverage rate of the confidence intervals for these parameters ranges between 91.7% and 96.2%.

All three bootstrap methods have relatively poor coverage probabilities of $\beta_{1 \text{ or } 2,4}$ and $\beta_{4,1 \text{ or } 2}$. While the studentized t method has a mean coverage probability of 90.2% for $\beta_{1 \text{ or } 2,4}$, the remaining mean coverage probabilities for these two parameters all fall below 90%.

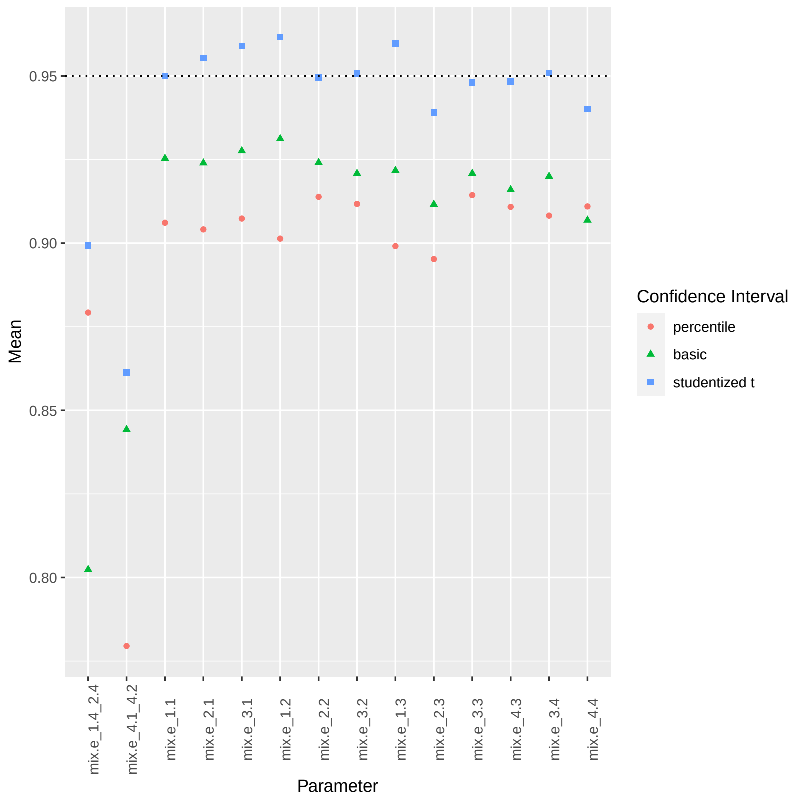

Fig. 3: Mean empirical coverage probability by bootstrap confidence intervals for reduced mix model parameters (40 sets of 200 simulations from Availability scenario 1)

The studentized t interval consistently produces the highest mean coverage rates among the three methods and is also the closest to the 95% threshold. The percentile method generally has the weakest performance of the bootstrap methods.

The mean coverage rates shown in Figure 3 were produced based on populations simulated from Availability scenario 1. We repeated the procedure to evaluate confidence interval coverages using populations simulations from Availability scenario 2. 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 match 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, 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 underperformed slightly, generally falling between 90% and 94% coverage. A plot of mean coverage rates by analytical and bootstrap confidence intervals for the match model is provided in Figure 7 under Appendix B.

8. Discussion

The ability to extract preferences separably from availabilities is a key feature of the revealed preferences model which we propose in this paper. In Simulation study A we simulate a small population ($N = 7,000$) and run the Gale-Shapley algorithm to obtain a stable matching. Given an observed distribution of outcomes rather than just an approximation, we are still able to compute parameter estimates which are very close to the true values.

In Simulation study B, we simulate an large population and determine an approximate stable matching from which we sample matching outcomes. We maximize a pseudo likelihood to obtain parameter estimates and show that the method accurately recovers true preference parameter values even under various different availabilities of prospective partners. In both simulation studies, 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 B.

We note that when there are very few or none of a certain type of matching outcome, the total utility of such a household is assessed to be negative infinity. As an example, we refer to the estimates of $\beta_{1 \text{ or } 2,4}$ in Simulation study B, shown in the first column of Figure 2. If we observed no pairings where a woman has education level 1 or 2 and the man has education level 4, then the estimate negatively infinity. This is a form of separation as also seen for generalized linear models (Heinze and Schemper, 2002). The the high concentration of parameter estimates for $\beta_{1 \text{ or } 2,4}$ around -6 correctly captures this and reflects the lower utility corresponding to such household pairings.

For Availability scenarios 1 and 2 under the type-based match model, the standard errors in Simulation study B (large population scenario) are smaller than the corresponding values in Simulation study A (small population scenario). However, the standard errors under Availability scenario 3 in Simulation study B are about three times larger than the standard errors for Availability scenarios 1 or 2. We suspect that the asymmetrical gender and education availabilities in Availability scenario 3 results in some model degeneracy when the large population approximation of the outcome distribution is used. As in Simulation study A, the distributions of the parameter estimates appear to follow a Gaussian distribution.

We evaluated different methods of accounting for uncertainty in our estimates. Based on results in Section 7.3, we believe that the approximation of the Hessian matrix leads to volatile analytical confidence intervals which deviate from the threshold coverage rate of 95%. We also show that in almost all cases, the modified version of the studentized t procedure for construction confidence intervals performed as well as or better than the percentile and basic methods. Additionally, while the percentile and basic method-based confidence intervals demonstrated slight undercoverage, the average coverage probabilities of the studentized t confidence interval for almost all parameters were centered around 95%. All three bootstrap methods produced confidence intervals which displayed significant undercoverage for the $\beta_{1 \text{ or } 2,4}$ and $\beta_{4,1 \text{ or } 2}$ parameters. This is not surprising, as these categories of households had low counts in populations simulated from Availability scenario 1.

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 restricted 2008 SIPP data are given in column 3 (β_0) of Table 4. The estimated utility of households in which both individuals have the same education level is substantially higher than it is for households where individuals have different education levels. This preference of homophily is expected by researchers who study matching problems. It is also consistent with the findings of Logan et al. (2008), who presented results which implied a preference for homophily in race and religion in heterosexual marriages.

In this paper, we applied the revealed preferences model to SIPP data. However, the model is novel in that the parameterization is well suited for even larger samples and census type data.

An open-source R package implementing the methods developed in this paper, `rpm`, (Handcock et al., 2020), was used to do the simulation studies and analyze the case-studies. We intend to make code available for these procedures in the R package `rpm` on CRAN (R Core Team, 2020).

Acknowledgements

We are grateful for support from the National Science Foundation BIGDATA: Applications program, grant NSF IIS-1546259, and from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, population research infrastructure grants P2C-HD041041 and P2C-HD041022 and training grant T32-HD007545.

A. Supplementary Tables

Table 3: Means and standard errors (SEs) of match model parameter estimates $\hat{\boldsymbol{\beta}}$ in Simulation study A (1,000 simulations, $N = 7,000$)

Parameter	Truth	Availability Scenario 1		Availability Scenario 2		Availability Scenario 3	
	$\beta_{match,0}$	Mean	SE	Mean	SE	Mean	SE
intercept	-2.564	-2.589	0.110	-2.582	0.116	-2.587	0.118
match 1	1.867	1.826	0.324	1.840	0.309	1.861	0.231
match 2	0.769	0.751	0.273	0.743	0.249	0.530	1.123
match 3	0.474	0.469	0.208	0.464	0.211	1.180	1.392
match 4	2.148	2.135	0.198	2.115	0.268	2.124	0.199

Table 4: Means and standard errors (SEs) of reduced mix model parameter estimates $\hat{\boldsymbol{\beta}}$ in Simulation study A (1,000 simulations, $N = 7,000$)

Education Parameter Female	Male	Truth $\beta_{mix,0}$	Availability Scenario 1 Mean	SE	Availability Scenario 2 Mean	SE	Availability Scenario 3 Mean	SE
1 or 2	4	-4.154	-4.008	0.522	-3.898	0.451	-2.884	0.311
4	1 or 2	-2.881	-2.954	0.380	-2.961	0.400	-2.905	0.188
1	1	-0.709	-0.790	0.319	-0.764	0.262	-0.755	0.209
2	1	-2.011	-2.086	0.437	-2.087	0.390	-2.065	0.362
3	1	-2.591	-2.690	0.411	-2.667	0.379	-2.642	0.333
1	2	-2.474	-2.619	0.546	-2.593	0.508	-2.429	0.538
2	2	-1.796	-1.848	0.268	-1.824	0.229	-1.919	0.559
3	2	-2.495	-2.523	0.256	-2.529	0.253	-2.634	0.554
1	3	-3.281	-3.325	0.574	-3.348	0.564	-2.219	0.348
2	3	-2.415	-2.474	0.308	-2.472	0.295	-2.094	0.456
3	3	-2.084	-2.115	0.184	-2.117	0.182	-2.152	0.548
4	3	-2.272	-2.327	0.287	-2.329	0.340	-2.353	0.382
3	4	-2.362	-2.416	0.333	-2.454	0.430	-2.390	0.558
4	4	-0.424	-0.451	0.158	-0.473	0.253	-0.455	0.166

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

B. Confidence intervals from 200 simulations

Figures 5 and 4 show the analytical confidence intervals and the empirical boostrap confidence intervals produced over 200 simulations. These figures coincide with the simulation

Table 5: Mean and standard errors (SEs) of match model parameter estimates $\hat{\boldsymbol{\beta}}$ in Simulation study B (1,000 simulations, $N = 300$ million)

Parameter	Truth	Availability Scenario 1		Availability Scenario 2		Availability Scenario 3	
	$\beta_{match,0}$	Mean	SE	Mean	SE	Mean	SE
intercept	-2.564	-2.566	0.091	-2.570	0.090	-2.670	0.422
match 1	1.867	1.859	0.210	1.863	0.212	1.957	0.448
match 2	0.769	0.758	0.175	0.754	0.181	0.831	0.420
match 3	0.474	0.457	0.142	0.466	0.146	0.534	0.401
match 4	2.148	2.161	0.143	2.158	0.137	2.236	0.427

Table 6: Means and standard errors (SEs) of reduced mix model parameter estimates $\hat{\boldsymbol{\beta}}$ in Simulation study B (1,000 simulations, $N = 300$ million)

Education Parameter		Truth	Availability Scenario 1		Availability Scenario 2		Availability Scenario 3	
Female	Male	$\beta_{mix,0}$	Mean	SE	Mean	SE	Mean	SE
1 or 2	4	-4.154	-4.271	0.616	-4.196	0.500	-4.260	0.572
4	1 or 2	-2.881	-2.939	0.388	-2.955	0.385	-3.227	0.890
1	1	-0.709	-0.711	0.190	-0.726	0.196	-0.735	0.203
2	1	-2.011	-2.048	0.238	-2.047	0.250	-2.079	0.262
3	1	-2.591	-2.615	0.252	-2.636	0.253	-2.660	0.254
1	2	-2.474	-2.511	0.295	-2.503	0.305	-2.533	0.316
2	2	-1.796	-1.807	0.160	-1.815	0.153	-1.838	0.156
3	2	-2.495	-2.511	0.158	-2.505	0.155	-2.549	0.165
1	3	-3.281	-3.344	0.371	-3.345	0.357	-3.338	0.356
2	3	-2.415	-2.423	0.220	-2.444	0.207	-2.486	0.214
3	3	-2.084	-2.093	0.113	-2.099	0.121	-2.135	0.115
4	3	-2.272	-2.284	0.194	-2.289	0.189	-2.330	0.190
3	4	-2.362	-2.375	0.207	-2.379	0.209	-2.397	0.207
4	4	-0.424	-0.414	0.109	-0.430	0.107	-0.445	0.107

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

results related to uncertainty estimates described in Section 7.3. 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 4 show the 200 confidence intervals for $\beta_{4,4}$ produced by each of the three bootstrapping methods which were described in Section 5.2. 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 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%.

We note that several of the confidence intervals shown in Figure 5 include -6, which was the lower bound we imposed on preference parameters in our study. These intervals effectively have no lower bound, since any preference parameter value of -6 or below is interchangeable with negative infinity.

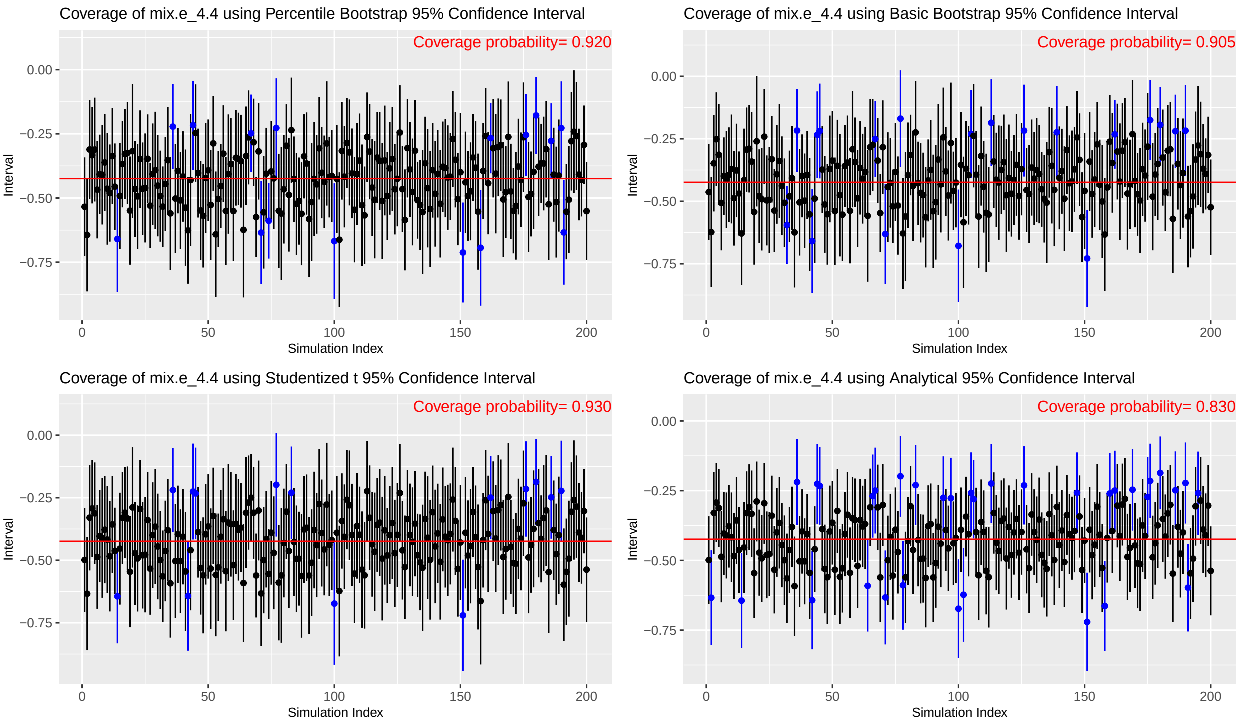

Fig. 4: Coverage of $\beta_{4,4}$ over 200 simulations

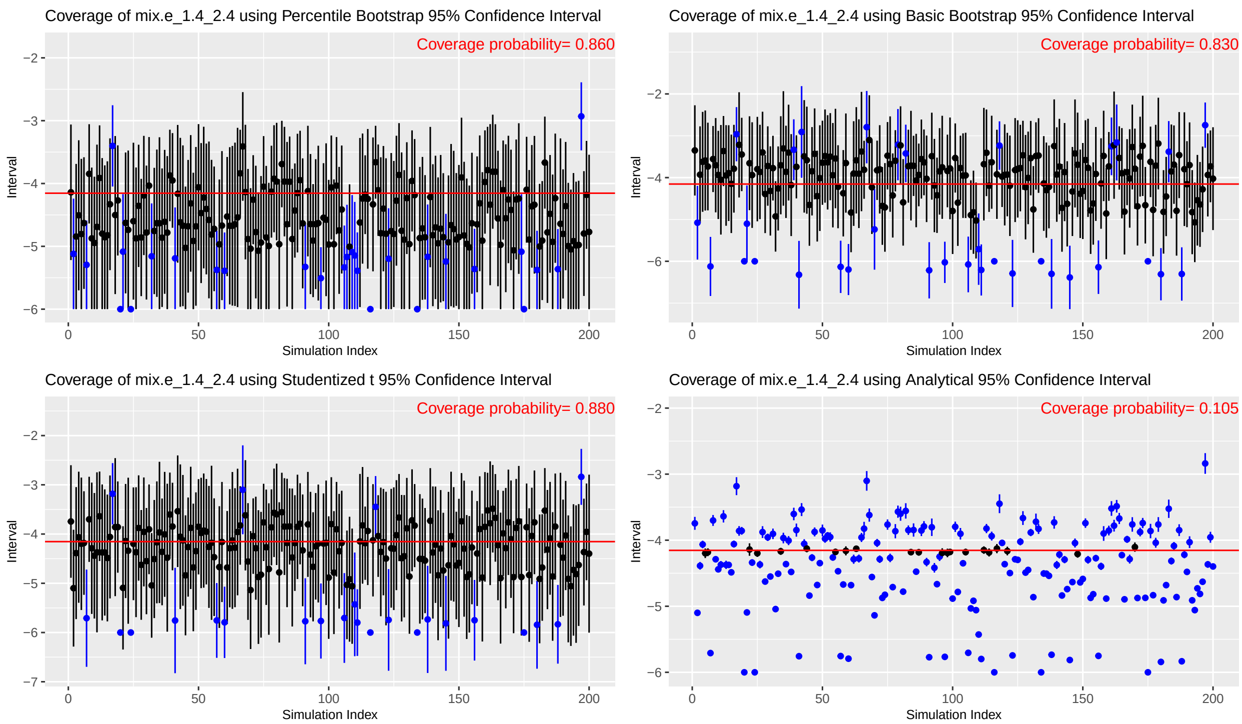

Fig. 5: Coverage of $\beta_{1 \text{ or } 2,4}$ over 200 simulations

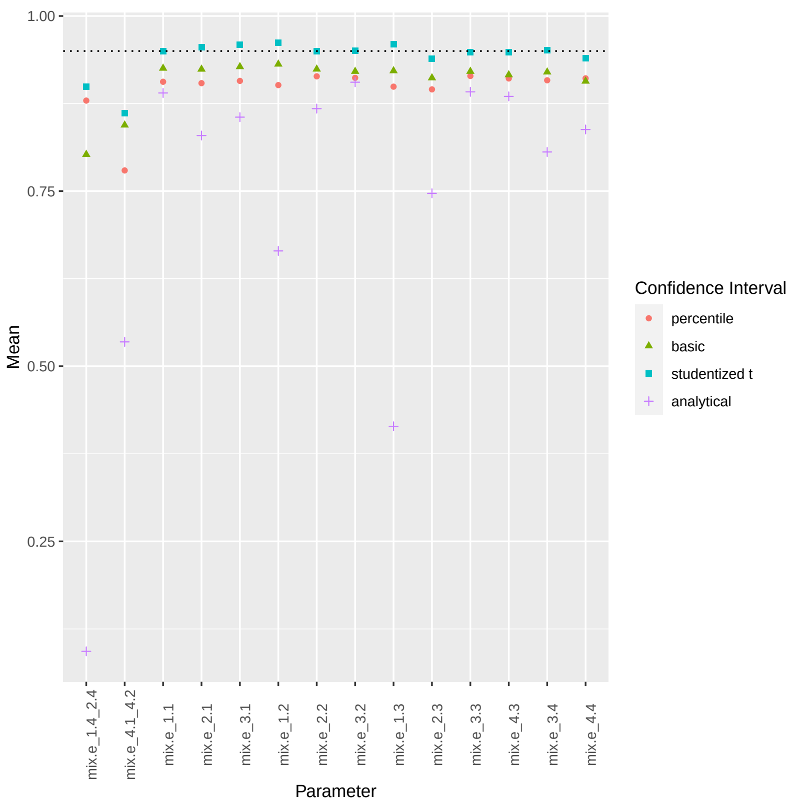

Fig. 6: Mean empirical coverage probability by confidence intervals for reduced mix model parameters (40 sets of 200 simulations from Availability scenario 1)

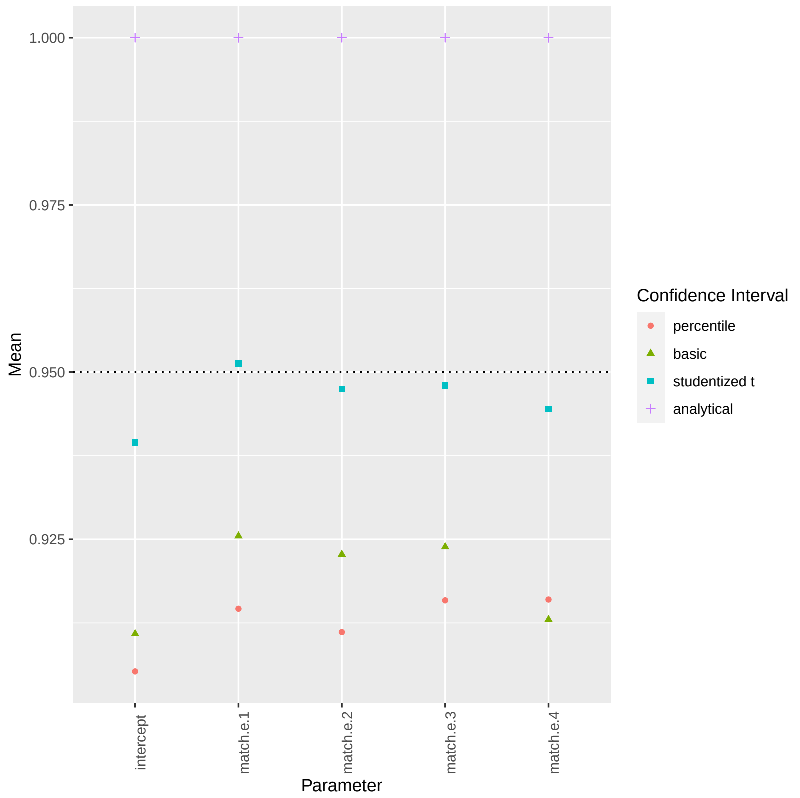

Fig. 7: Mean empirical coverage probability by confidence intervals for type-based match model parameters (40 sets of 200 simulations from Availability scenario 1)

References

Choo, E. and Siow, A. (2006) Who marries whom and why. *Journal of Political Economy*, **114**, 175–201. URL: `http://www.jstor.org/stable/10.1086/498585`.

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

Godambe, V. P. and Thompson, M. E. (1986) Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Review / Revue Internationale de Statistique*, **54**, 127–138. URL: `http://www.jstor.org/stable/1403139`.

Hájek, J. (1971) Comment on "an essay on the logical foundations of survey sampling, part one" by d. basu. In *Foundations of Statistical Inference* (eds. P. Godambe and D. A. Sprott), 236. Holt, Rinehart and Winston.

Handcock, M. S., Admiraal, R., Yeung, F. C. and Goyal, S. (2020) **rpm**: *Statistical estimation of revealed preference models from data collected on bipartite matchings*. Los Angeles, CA. R package version 0.40.

Hartmann, W. M. and Hartwig, R. E. (1996) Computing the Moore-Penrose inverse for the covariance matrix in constrained nonlinear estimation. *SIAM Journal on Optimization*, **6**, 727–747. URL: `https://doi.org/10.1137/S1052623494260794`.

Heinze, G. and Schemper, M. (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*. URL: `http://github.com/stevengj/nlopt`.

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

Kraft, D. (1994) Algorithm 733: Tomp–fortran modules for optimal control calculations. *ACM Trans. Math. Softw.*, **20**, 262–281. URL: `https://doi.org/10.1145/192115.192124`.

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

Menzel, K. (2015) Large matching markets as two-sided demand systems. *Econometrica*, **83**, 897–941. URL: `http://www.jstor.org/stable/43616957`.

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

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

R Core Team (2020) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: `https://www.R-project.org/`.

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

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

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

U.S. Bureau of the Census (2020a) 2008 survey of income and program participation (sipp). URL: `https://www.census.gov/programs-surveys/sipp/data/datasets.2008.html`.

— (2020b) Survey of income and program participation (sipp). URL: `https://www.census.gov/programs-surveys/sipp.html`.

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