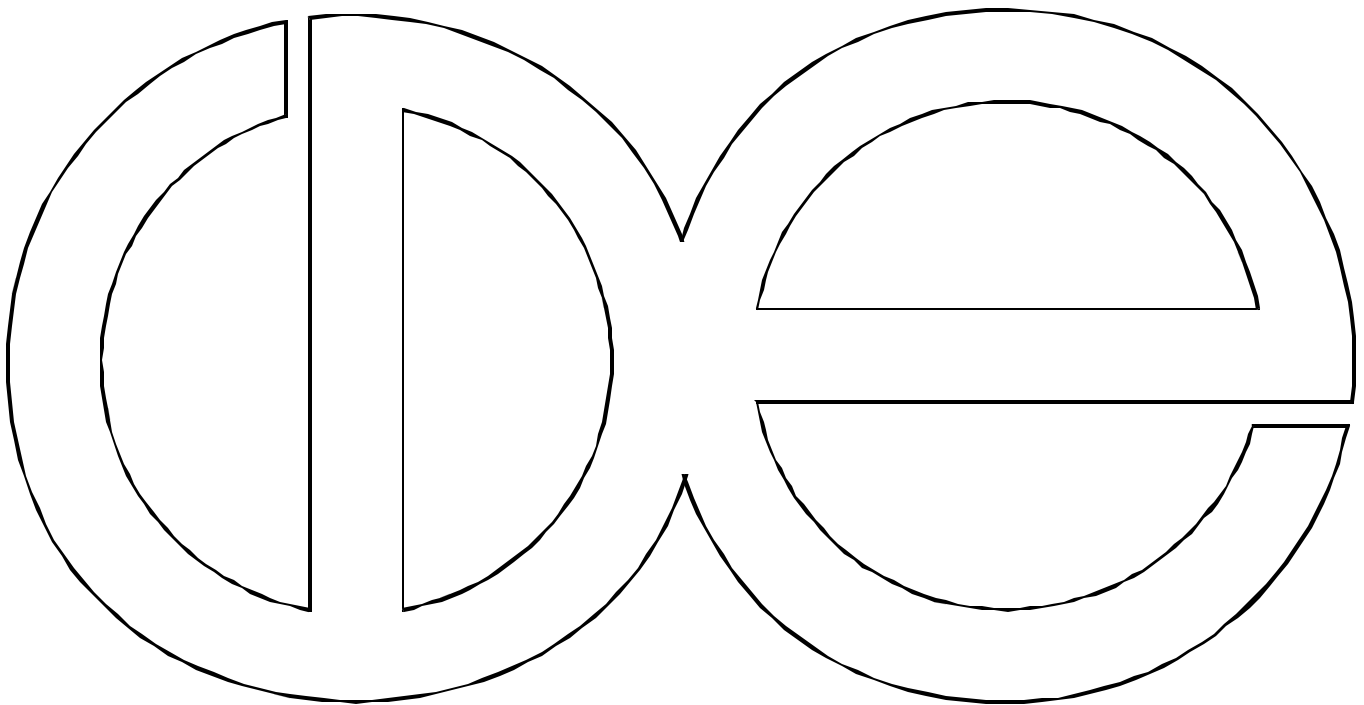


**Center for Demography and Ecology
University of Wisconsin-Madison**

Estimation for the Marriage Model

**John Allen Logan
Peter D. Hoff
Michael A. Newton**

CDE Working Paper No. 99-31



Estimation for the Marriage Model

John Allen Logan, Peter D. Hoff, and Michael A. Newton

August 5, 1999

Abstract

We consider a parametric version of the two-sided matching model of marriage described in Roth and Sotomayor (1990). There are two matching populations, men and women. Members of each population have utilities for pairing with members of the other population, which results in a stable set of matches among the two populations. We present a method of estimating the degree to which characteristics of individuals affect the utilities upon which the matches are based.

Social outcomes often result from individual choices made under constraints that are not observed by researchers. Marriage, for example, is typically recorded in social science data sets by describing the marital status of each individual and listing selected characteristics of the marital partner, if one is present. The choice context in which the marriage took place, however, is not recorded because of the difficulty of determining the sets of potential partners that were available to individuals prior to making their choices. This lack of contextual information makes it difficult to infer the preferences of individuals which led to the choices. Put simply, it is unclear to what degree observed choices were constrained by the unavailability of more desirable partners.

Employment data are similar. The characteristics of currently- or previously-held jobs are recorded for individuals, but not, typically, the characteristics of other jobs that could have been taken. From the employer's side, the characteristics of employees may be recorded, but usually not the characteristics of all other workers who might have been hired. Not knowing what alternatives were available makes it difficult to determine which characteristics of available alternatives were important in the choices that were made.

Statistical inference is well-developed for situations where the choice context is known or can be approximated a priori. The familiar logit and probit models have been derived from underlying utility maximization models for this case. Such "discrete choice" models have been extensively elaborated and widely applied in suitable empirical domains, notably transportation and consumer-choice studies (e.g., Ben-Akiva and Lerman 1985, Anderson, De Palma and Thisse 1992, Pudney 1989). These areas differ from marriage and employment in that it is easy to learn which modes of transportation or which brands of products were available to the choosing individuals, and therefore possible to infer preferences from their observed choices.

Logan (1996a) extended the discrete choice model for the employment situation by simultaneously estimating logit models for the preferences of workers and employers, using data only on the characteristics of their current employment partners. In this model, the missing information about the choices available for actors on either side of the job market is in effect replaced by using estimates of the preferences of the actors on the other side. Logan's two-sided logit model treated workers and employers asymmetrically, in that characteristics of individual workers were used directly in the estimation, while employers' offered job characteristics entered into the model only as occupational category means. The motivation for using occupational category means was an exponential increase in computation time with the number of distinct employment alternatives available to the workers. Ten or fewer alternatives is a reasonable upper limit for this model in most computing environments.

The use of occupational category means in Logan's two-sided logit model carries some important limitations. The number of employer characteristics that can be included must be no more than the number of categories, minus one. The use of only a few categories implies that the choices available to the workers can only be differentiated crudely, so that, for example, all managerial jobs might be considered equivalent, with the same (mean) salary and other attributes. This limitation extends to the characterization of the worker's own job: available information in the data concerning the specific attributes of the current job must be discarded, even though this information is obviously often important in understanding why the current job is being held. Some problems do not lend themselves as naturally to categorization as does employment. In particular, in marriage the use of categories of men in combination with individual-level data on women, or the reverse, would seem unnatural and would also preclude evaluating certain hypotheses about symmetry between the preferences of the sexes. Finally, the loss of information involved in using category means may be associated with the appearance of multiple local maxima of roughly equal magnitude in the likelihood function (see Logan 1998).

The present paper considers the estimation of a completely symmetrical, non-categorized variant of Logan's two-sided model, which is intended to remove the limitations discussed in the preceding paragraph. The goal is to analyze marital data without categorization, meaning that a few thousand potential partners for both men and women must be considered in a typical data set. Such a problem is far beyond the capability of the quasi-Newton and EM algorithms considered in Logan (1998). Though the techniques to be considered here have applicability to either employment or marriage, we will consider them in the context of marriage for specificity. The discussion section will consider applications to employment.

1 The Marriage Model

The two-sided model of Logan (1996a, 1998) is composed of two basic utility functions and a decision rule. As adapted for the marriage problem the utility for man i 's evaluation of woman j is this function of the

characteristics of the woman x_j and the shared male preferences α :

$$U_{i,j} = \alpha' x_j + \epsilon_{i,j} \tag{1}$$

The term $\epsilon_{i,j}$ is a disturbance representing unobserved characteristics influencing the utility. When the same parameters are applied to evaluate the option of remaining or becoming single, index $j = 0$ designates the self-matching characteristics available in that state, which may vary from man to man:

$$U_{i,0} = \alpha' x_0^i + \epsilon_{i,0}$$

Similarly, the utilities woman i sees in man j and in a self-match are:

$$V_{i,j} = \beta' y_j + \gamma_{i,j} \tag{2}$$

$$V_{i,0} = \beta' y_0^i + \gamma_{i,0}$$

Vector y_j contains measured characteristics (resources) of man j , with $j = 0$ again indicating singleness, and the vector β contains preference coefficients over those characteristics, shared among all women. The decision rule followed is utility maximization within the set of opportunities provided by other actors: Each man takes the highest-utility woman among those available to him, or else remains single if his utility for singleness exceeds that for his highest-rated available woman. Women do the same, with the roles reversed.

It is assumed that the disturbances in (1) and (2) are i.i.d. continuous random variables. This assumption is tantamount to assuming that each man and each woman has a complete preference ordering over each member of the opposite sex, an assumption which is also explicitly adopted. For given realizations of the disturbance terms, the system of men and women just described is equivalent to a two-sided matching model known simply as the marriage model in the economic game theory literature. Among the many theorems that have been proved for this system (see Roth and Sotomayor 1990), it is known that at least one stable matching must exist, no matter what the pattern of preferences. A stable matching is an assignment of men and women to matches, either with members of the opposite sex or with themselves (in which case they are single), such that no man can find a woman he prefers to his current match who also prefers him to her match, and such that the same is true for women, with the roles reversed. Since no man and no woman can make a move that will better his or her situation, the matching is stable. More than one stable matching may exist for a system.

Stable matchings are of central interest in the application of two-sided matching models. Since a matching that is not stable could be improved on by the voluntary actions of system members, it is reasonable to expect that real systems will approximate stable matchings. In addition it is known (Roth and Vande Vate 1990) that a random flow of information about potential partners will lead to a stable matching with probability 1 in the long run. Such arguments as these make the analysis of stable matchings relevant to understanding real systems, even though it is usually the case that stability cannot be directly confirmed.

We assume that the data to be analyzed are drawn from a stable matching. The stability in question should be thought of as inherently transitory, subject to change as persons enter or leave the system, or as

their characteristics or preferences change over time through aging processes or for exogenous reasons. Our assumption is that system members are generally knowledgeable about the characteristics of other actors, and that their observed matches are entered and maintained voluntarily. Though existing matches of long duration may have given rise to particularistic mutual evaluations or may for legal or other reasons entail large potential costs if the relationship were to be broken, such considerations do not argue against the voluntary nature of the continuing relationships nor suggest that an unstable matching is somehow enforced against the wills of the system members, when all relevant factors are considered. Stability in the sense adopted here means mutually voluntarily maintained, rather than unchanging as circumstances change.

The marriage model is composed of individuals but has a systemic quality. Each man observes the set of women available to him or her, and then chooses the woman that has the highest utility for him. Since particular men do not generally have all women available to them, their opportunities are constrained by the preferences of women. In addition, since the availability of any given woman to a particular man depends on the available pool of men that would be her willing partners, most men's opportunities are also constrained by the preferences of other men. These same points apply for women, with the roles reversed. Both men's and women's preferences weight the actual resources of possible partners, meaning that opportunity is in part a function of the distribution of these resources, and not just of preferences as such. The marriage model is therefore a system model of the interwoven effects of the preferences and resources of all men and women.

2 Estimation Methods

Our goal is to estimate α and β , the marital preferences of men and women, using data D on matchings and the characteristics of men and women. The maximum likelihood estimation methods of Logan (1998) would be applicable to the present model with minor alterations, except the estimation time associated with those methods makes them impractical here. Instead we consider two Markov chain Monte Carlo (MCMC) algorithms. An excellent overview of MCMC methods is provided by the introductory chapter of Gilks, Richardson and Spiegelhalter (1996).

2.1 MCMC Estimation

Maximum likelihood estimates of the parameters $\theta = (\alpha, \beta)$ are those values which maximize the likelihood function, $L(\theta) = \Pr(D|\theta)$. However, in contrast to many applications of likelihood, in our model the contributions to $L(\theta)$ from different matches are not independent; constraints induced by the assumption of stability connect the matches in a complicated way. Furthermore, none of the utilities is observed, making it necessary to integrate $\Pr(D, U, V|\theta)$ over all possible values of the utilities in order to obtain $\Pr(D|\theta)$. As an alternative to maximum likelihood estimation, we take a Bayesian perspective. A formal Bayesian analysis requires a prior density function $\pi(\theta)$ to represent uncertainty about θ before data have been analyzed. One

often attempts an objective assessment of the parameters by using the improper uniform prior $\pi(\theta) = 1$. With this prior, the posterior density $\pi(\theta|D)$ is proportional to the likelihood function, and it forms the basis of inference about components of θ . Note that maximum likelihood estimation amounts to finding the mode of this joint posterior distribution, which is impractical with methods currently available. On the other hand, Bayesian inference concerning any single component of θ is based on the marginal posterior distribution of that component, the distribution obtained by integrating the full posterior distribution over all other parameters.

Markov chain Monte Carlo provides a feasible method for approximating such integrals: A computer is used to generate a random sample from a distribution approximating the joint posterior distribution, and the values taken by any one component of θ can be used to approximate the corresponding marginal posterior distribution of that component. This technique can be used to bypass the problem of integrating over the unknown utilities: While direct sampling from $\pi(\theta|D)$ is extremely difficult, sampling from values of θ, U, V from a distribution approximating $\pi(\theta, U, V|D)$ is feasible. The resulting sampled θ values can be examined marginally to estimate $\pi(\theta|D)$. This use of data augmentation to simplify sampling has been discussed in Tanner and Wong (1987) and Albert and Chib (1993).

Let $S = (U, V, \theta)$ represent the collection of unknowns from our model, which includes both the unobserved utilities as well as the unknown parameters. Ideally we would approximate $\pi(S|D)$ from the marginal distribution of a set of independently sampled values S^1, \dots, S^B from $\pi(S|D)$. However, for our problem there is no direct mechanism which produces such samples. Instead, we use the Metropolis-Hastings algorithm to generate a Markov chain Monte Carlo solution (Tierney 1994). The sampled values S^1, \dots, S^B are not independent, but form a Markov chain whose stationary distribution is equal to the desired distribution $\pi(S|D)$. The Metropolis-Hastings algorithm is defined by a collection of *proposal* distributions which are used to sample proposal states S^* given the current state $S = S^b$. For a given proposal distribution having transition density $q(S^*|S)$, we sample S^* from q and compute the Metropolis-Hastings ratio

$$r(S^*, S) = \min \left\{ 1, \frac{\pi(S^*|D)q(S|S^*)}{\pi(S|D)q(S^*|S)} \right\}. \quad (3)$$

The next state of the chain S^{b+1} is S^* with probability $r(S^*, S)$; otherwise the next state is the same as the current state S . A *scan* of the algorithm is complete after modifications have been attempted with each proposal distribution in the collection, following some specified order.

2.2 Two MH Algorithms

We now describe in detail the particular Markov chains used to generate a sequence of values for the unknowns S . Our Metropolis-Hastings algorithms will consist of a collection of separate updates for each unobserved utility, as well as separate updates for the two parameters α and β . The proposal distributions for each utility will generate a proposal value S^* which matches S for all component values except the utility in

question. Similarly, the proposal distribution for α or β will generate values of S^* in which only the value of α or β is altered.

2.2.1 Utilities

We first consider a proposal distribution for the matching utilities of males. For a particular male i and female j , we sample a proposed utility $U_{i,j}^*$ from its *full conditional* distribution, which is the distribution of the utility conditional upon the data and the current state of all other utilities and parameters. The proposed utility induces a proposal state S^* which is identical to the current state S except for the value $U_{i,j}^*$. Sampling from a full conditional distribution is known as Gibbs sampling, and the acceptance probability $r(S^*, S)$ of such a sample is always unity.

For logistic and normal disturbance terms, the full conditional distribution of $U_{i,j}^*$ given all other unknowns is simply a logistic or normal distribution constrained by the other utilities and the observed matching. For example, if man i is matched to woman w_i , then he must have a higher utility for matching with w_i than the utility for being single or matching with any other woman available to man i . Similarly, if man i is single then $U_{i,0}$ must be higher than the utility of being matched to any woman available to man i . More precisely, the constrained sampling is as follows: Letting w_i denote the wife of man i ($w_i = 0$ if i is single) and h_k denote the husband of woman j ,

- if $j \neq w_i$
 1. if $j = 0$ or $V_{j,i} > V_{j,h_j}$ (i.e. j is available to i), sample $U_{i,j}^*$ from the logistic or normal distribution with mean $\alpha'x_j$, conditional upon $U_{i,j}^* < U_{i,w_i}$.
 2. if $V_{j,i} \leq V_{j,h_j}$ (j is not available to i), sample $U_{i,j}^*$ from the logistic or normal distribution with mean $\alpha'x_j$ (unconstrained).
- if $j = w_i$, sample $U_{i,j}^*$ from the logistic or normal distribution with mean $\alpha'x_j$ conditional upon $U_{i,j}^* \geq U_{i,0}$ and $U_{i,j}^* > U_{i,k}$ for any $k : V_{k,i} > V_{k,h_k}$.

The utilities $V_{i,j}$ of women for men are sampled the same way, except with V, h interchanged with U, w .

2.2.2 Preferences

Normal Disturbances If we assume disturbances are standard normal and the preferences α and β are a priori independent and normally distributed, $\alpha \sim \mathcal{N}(\mu_1, \Sigma_1)$, $\beta \sim \mathcal{N}(\mu_2, \Sigma_2)$, straightforward calculation shows the full conditionals are again normal:

1. $\alpha |_{\beta, U, V, D} \sim \mathcal{N}(\hat{\mu}_1, \hat{\Sigma}_1)$, where
 - $\hat{\Sigma}_1^{-1} = n_m X X' + X_0 X_0' + \Sigma_1^{-1}$
 - $\hat{\mu}_1 = \hat{\Sigma}_1 (X U' \mathbf{1} + Y_0 U_0 + \Sigma_1 \mu_1')$

2. $\beta|_{\alpha,U,V,D} \sim \mathcal{N}(\hat{\mu}_2, \hat{\Sigma}_2)$, where

- $\hat{\Sigma}_2^{-1} = n_f Y Y' + Y_0 Y_0' + \Sigma_2^{-1}$
- $\hat{\mu}_2 = \hat{\Sigma}_2 (Y V' \mathbf{1} + Y_0 V_0 + \Sigma_2 \mu_2')$

where

- n_m and n_f are the male and female population sizes,
- X and Y are the matrices of female and male characteristics,
- X_0 and Y_0 are the matrices of self matching characteristics,
- U and V are the matrices of utilities of men for women and women for men,
- U_0 and V_0 are the self-matching utilities.

The improper uniform prior $\pi(\alpha, \beta) = 1$ also gives normal full conditionals, corresponding to zeros for values of $\mu_1, \Sigma_1, \mu_2, \Sigma_2$ above. Such normal distributions are easy to sample from, and so we use them to generate our proposal values α^* and β^* . Since these are Gibbs samples, the acceptance probabilities of the proposals is unity.

Logistic Disturbances Unlike the normal case with a uniform prior, the marriage model with logistic disturbances has no nice closed form expression for the full conditionals of α or β . Therefore, instead of generating parameter updates from a full conditional distribution, we use a random walk proposal, that is, we sample a proposal value uniformly from a box around the current value:

$$\begin{aligned} q_1(\alpha^*|\alpha) &= A^{-1} \quad \text{for } \alpha^* \in \alpha \pm \delta_1 \\ &= 0 \quad \text{otherwise} \\ q_2(\beta^*|\beta) &= B^{-1} \quad \text{for } \beta^* \in \beta \pm \delta_2 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

where δ_1 and δ_2 are vectors of the same dimension as α and β respectively, and A and B are the volumes of the boxes spanned by $\alpha \pm \delta_\alpha$ and $\beta \pm \delta_\beta$. Because these proposal distributions are not the full conditionals, the acceptance probability is not necessarily unity, and needs to be calculated. Since updating the utilities given the data and parameters is equivalent to updating the disturbances ϵ and γ , the acceptance probability for a parameter update can be based on the conditional distribution given the disturbances instead of the utilities. For example, when updating the α parameter the Metropolis-Hastings ratio (3) is

$$\begin{aligned} \frac{\pi(\alpha^*, \beta, \epsilon, \gamma|D) q(S|S^*)}{\pi(\alpha, \beta, \epsilon, \gamma|D) q(S^*|S)} &= \frac{\pi(D|\alpha^*, \beta, \epsilon, \gamma) \pi(\alpha^*) \pi(\beta) \pi(\epsilon) \pi(\gamma) q_1(\alpha|\alpha^*)}{\pi(D|\alpha, \beta, \epsilon, \gamma) \pi(\alpha) \pi(\beta) \pi(\epsilon) \pi(\gamma) q_1(\alpha^*|\alpha)} \\ &= \frac{\pi(D|\alpha^*, \beta, \epsilon, \gamma)}{\pi(D|\alpha, \beta, \epsilon, \gamma)} \end{aligned} \tag{4}$$

$$= \pi(D|\alpha^*, \beta, \epsilon, \gamma). \tag{5}$$

Equality (4) holds if we use the uniform prior on α (since $\pi(\alpha^*) = \pi(\alpha)$), and because the proposal distribution is symmetric ($q_1(\alpha^*|\alpha) = q_1(\alpha|\alpha^*)$). Equality (5) holds because $\pi(D|\alpha, \beta, \epsilon, \gamma)$ is simply the zero-one indicator variable of whether or not D is a stable matching under the values of $\alpha, \beta, \epsilon, \gamma$, which must be one since α was accepted at a previous step of the chain. Therefore, the probability of accepting α^* is simply $\pi(D|\alpha^*, \beta, \epsilon, \gamma)$, which is unity if D is a stable match under $\alpha^*, \beta, \epsilon, \gamma$, and zero otherwise. Similarly, the acceptance probability when updating β is $\pi(D|\alpha, \beta^*, \epsilon, \gamma)$, which also takes values zero or one.

3 Simulation Results

A simulation study was performed to test the ability of the above MCMC algorithms to estimate the preference parameters. Ten computer-generated datasets were created and the algorithms were run on each one. For each dataset, 220 mean-zero variance-one normally distributed deviates were generated to form a 2×110 matrix X of female characteristics, i.e. two characteristics for each of $n_f = 110$ females. Similarly, a 2×90 matrix Y was generated representing the characteristics of $n_m = 90$ males. The self matching characteristics were generated as normal variates with a mean of minus one and variance of one for all individuals of both sexes. The utilities were generated using the parameters $\alpha' = (1, 3)$ for male preferences and $\beta' = (2, 2)$ for female preferences. In summation, we have

- independently for each $j = 1, \dots, 110$, $x_j \sim \mathcal{N}((0, 1)', I)$.
- independently for each $i = 1, \dots, 90$, $y_i \sim \mathcal{N}((0, 1)', I)$.
- independently for each $j = 1, \dots, 110$, $y_0^j \sim \mathcal{N}((-1, 1)', I)$.
- independently for each $i = 1, \dots, 90$, $x_0^i \sim \mathcal{N}((-1, 1)', I)$.
- $U_{i,j} = \alpha' x_j + \epsilon_{i,j}$
- $U_{i,0} = \alpha' x_0^i + \epsilon_{i,0}$
- $V_{i,j} = \beta' y_j + \gamma_{i,j}$
- $V_{i,0} = \beta' y_0^i + \gamma_{i,0}$,

where the disturbances ϵ, γ are sampled independently from the standard normal distribution, and I represents the identity matrix.

From these utilities, two (possibly identical) stable matchings were generated for each dataset according to the Gale-Shapley algorithm (Gale and Shapley 1962, Roth and Sotomayor 1990). The Gale-Shapley algorithm entails each member of one sex proposing to their most preferred member of the opposite sex. Each member of the opposite sex then makes a temporary engagement with their most preferred proposer, and each rejected proposer proposes to the next member on their ranked list of opposite sex members. The

algorithm iterates until no further rejections or proposals are forthcoming. At this point all proposers are either engaged or have been rejected by all members of the opposite sex whom they prefer to self-matching. This algorithm was run on the computer-generated datasets twice, once with men proposing and once with women proposing.

In systems containing more than one possible stable matching the men-proposing and women-proposing matchings are polar opposites in the sense that the former is at least as good for every man as any other stable matching and the latter is the same for every woman (Roth and Sotomayor 1990). The two matchings are called man-optimal and woman-optimal, respectively. The model outlined in section 1 did not specify which stable matching is being observed in the data. Instead, we have assumed only that *some* stable matching is observed. For this reason, we expect the estimation algorithms to be insensitive to the particular stable matching that has been achieved. We compute estimates from the man- and woman-optimal matchings to help demonstrate this property. In systems of real men and women there are likely to be many possible stable matchings, so insensitivity of the estimation algorithm to the particular identity of the achieved matching is crucial.

Figures 1 and 2 show the estimation results for the α and β parameters respectively. The figures contain separate graphs for each of the four parameters. Within each graph there are 10 clusters of results corresponding to the 10 sets of simulated data. In each cluster there are typically four error bars, two labelled L for the logistic results and two labelled N for the normal. The first error bar in each pair of L or N results is for the men-proposing matching obtained on a particular set of simulated data, and the second is for the women-proposing matching on the same set of simulated data. In data sets 0, 1, 4, and 7 the men- and women-proposing matchings are identical. In these cases only a single logistic and a single normal estimation are presented, since there is only one matching.

Since the logistic and normal algorithms are based on different disturbance distributions their parameters are not directly comparable. This kind of difficulty is well-known in the discrete choice literature, where logit and probit models are often compared. Two different arguments about the logit and probit disturbances lead to adjustment factors of $1/1.6$ and $1/1.8$ for the conversion of logit model estimates to probit model equivalents (Maddala 1983: 23). We adopt an average of the two adjustments, $1/1.7$, for the logit/probit conversion. In addition, we adjust for a difference between our logistic disturbance model and the extreme value disturbances underlying the standard logit model (see Ben-Akiva and Lerman 1985). Since the logistic disturbances have twice the variance of extreme value disturbances, we divide the logit/probit conversion factor by the square root of 2, obtaining a total adjustment of $1/2.4$. The logistic algorithm results in Figures 1 and 2 have been multiplied by this factor. The normal-disturbance algorithm results were not adjusted.

Each of the sets of results shown in figures 1 and 2 summarizes scans 50,001 through 200,000 of a Markov chain, the first 50,000 scans having been discarded to allow the chain to approach its stationary distribution. The midpoint of each error bar is the mean of the marginal posterior distribution for the parameter, while the bars themselves extend plus and minus two standard deviations of the distribution. These bars are

Bayesian confidence intervals. Each graph includes a horizontal line at the location of the parameter value used in the simulations.

The agreement of both the logistic and normal model estimates with the parameter values is reasonably good. The error bars generally overlap the parameter values. In total, 115 of the 128 intervals in the graphs include the parameter values. In general, the logistic and normal model results are similar from simulation to simulation. Standard deviations are similar between the two models on average.

We believe the results in Figures 1 and 2 show fairly convincing evidence that the algorithms work effectively. The similarity of results between the two independently derived and programmed methods is reassuring. The poorer runs may be the result of sampling error due to the small number of simulated matches or of Monte Carlo error due to the relatively small number of scans. The results in the figures are runs on the first sets of simulated data meeting the programs' requirements. We plan to generate data with more information for additional sets of estimations by increasing the sample size. In addition to adding information to the data, we will do further experiments in tuning the logistic algorithm by varying the volume of the boxes from which new parameter values are drawn. (The normal algorithm has no tuning parameters as such since it is a Gibbs sampler.)

4 Random Blocking Algorithms

The algorithms of the previous section have execution times that are roughly quadratic in the number of cases. Though a great improvement on the exponential rule associated with the methods of Logan (1998), execution times measured in days would still be the rule for marriage data with thousands of cases using these algorithms. To achieve higher speed we have devised a modification of the algorithms based on an idea of random blocking. Though this method has not been investigated thoroughly, we will present the basic idea along with some suggestive results using 2500 cases from the Panel Study on Income Dynamics to give an indication of the direction of our efforts. This modified version of the algorithms has execution times that are linear in the number of cases, meaning, for example, that 2500 cases can be analyzed in about an hour and a half on a 300 MHz computer.

The idea of random blocking follows from observing that any subset of matches from a stable matching is necessarily also stable. (*Proof*: In a stable matching no man can find a woman he prefers to his match such that she prefers him to her match. Reducing the women potentially available to the man by forming subsets of matches cannot produce new women actually available to him since each woman is placed in a subset together with her current match, whom she still prefers. The same argument applies to the women, with roles reversed.) Our procedure simply assigns the matches randomly to small blocks of constant size, discarding any remaining matches that would not make up an additional fixed-sized block. The stability property necessarily applying within each block, we execute the algorithms of the preceding sections within blocks, using a single set of working parameter values across blocks.

Figures 3 and 4 show alpha and beta estimates obtained using the logistic algorithm on the same ten data sets considered previously. Only the men-proposing matches were used. Each group of three error bars shows the unblocked logistic estimates (seen previously), estimates obtained using a block size of 22 matches, and estimates obtained using a block size of 5. The blocked estimates have greater variability than the unblocked, which is to be expected since blocking discards information in the interest of speed. The blocked estimates generally track the unblocked estimates, though there are notable exceptions in simulations 5 and 7. We note that the blocksize-22 estimates are not uniformly better than the blocksize-5 estimates. As in the unblocked results, we intend to increase the information in the data when performing additional experiments.

To demonstrate the execution speed that can be obtained by blocking, we analyzed 2500 cases from the 1993 final release family data file of the Panel Study of Income Dynamics. The sample was restricted to households in which the head was between 25 and 44 years old. A very simple specification was used for the purpose of the demonstration. The age and education of the man and of the woman were entered as the x and y variables, and the self-matching values of (spouse's) age and education were set to zero. Using a block size of 5, we obtained 200,000 scans in roughly 1.5 hours. Further runs of 500,000 and 1,000,000 scans produced similar results, with linear increases in execution time.

Figure 5 shows posterior kernel density estimates for scans 50,000 through 500,000 of the second run. Vertical lines enclose 95 percent highest posterior density (HPD) regions, the conventional interval estimates for Bayesian analyses. We note that the estimates seem substantively reasonable. The preference of women for men's education (β_2) is estimated at about 0.19, while the preference of men for women's education (α_2) is about 0.11. Values of α_2 exceed β_2 in only 0.001 of the sampled scans, suggesting that women's educational preferences for their mates are distinctly higher than men's. The preference of women for higher ages in their mates (β_1) exceeds the reverse preference of men for higher ages in women (α_1), with no sampled value contradicting this pattern. Both of the age preferences are smaller than the education preferences, a comparison that is meaningful since both variables are measured in years.

5 Discussion

The marriage model specified in equations (1) and (2) is both an extension of Logan's (1998) two-sided logit model and a direct parameterization of the widely-known marriage model described in Roth and Sotomayor (1990). The model is attractive in its simplicity, representing the preference orderings of men and women as linear functions of one another's characteristics or resources. A system model of opportunity, it derives constraints on individual choices solely by reference to properties of other individuals. Though the same model is analyzed non-statistically in economics, it is important to emphasize that it does not conflict with sociological ideas of norms, customs and structures, but rather can be used to represent and investigate such things (see Logan 1996). It is emphatically a model suited to the sociologist's point of view.

The marriage model is also a practical solution for the employment problem addressed previously by the

two-sided logit model. With suitable adaptations, the same algorithms can be used to take advantage of information on individual job characteristics, instead of category means. The categories used in the two-sided logit model can still appear in the new type of model, but would only be used to differentiate employers' preference structures by occupation or industry instead of being used to reduce job characteristic data to means. This will greatly expand the capabilities seen in the earlier model.

The MCMC methods described here show the promise of making the marriage model applicable to data sets of moderate to large size. We close by reemphasizing that this is still a work in progress, and that our results are tentative.

References

- Albert, James H., and Siddhartha Chib. 1993. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association* 88(422):669-679.
- Anderson, S., A. De Palma, and J.-F. Thisse. 1992. *Discrete Choice Theory of Product Differentiation*. Cambridge, MA: MIT Press.
- Ben-Akiva, Moshe, and Steven R. Lerman. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, MA: MIT Press.
- Gale, David, and Lloyd S. Shapley. 1962. "College Admissions and the Stability of Marriage." *American Mathematical Monthly* 69:9-15.
- Gilks, W.R., S. Richardson, and N.G. Spiegelhalter (eds.). 1996. *Markov Chain Monte Carlo in Practice*. New York, London: Chapman and Hall.
- Logan, John Allen. 1996. "Opportunity and Choice in Socially Structured Labor Markets." *American Journal of Sociology* 102(1; July):114-160.
- Logan, John Allen. 1998. "Estimating Two-Sided Logit Models." *Sociological Methodology* 28:139-173.
- Maddala, G.S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.
- Pudney, Stephen. 1989. *Modelling Individual Choice*. Oxford: Basil Blackwell.
- Roth, Alvin E., and Marilda A. Oliveira Sotomayor. 1990. *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. Cambridge; New York: Cambridge University Press.
- Roth, Alvin E., and John H. Vande Vate. 1990. "Random Paths to Stability in Two-Sided Matching." *Econometrica* 58:1475-80.

Tanner, Martin A., and W.H. Wong. 1987. "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association* 82:528-540.

Tierney, L. 1994. "Markov Chains for Exploring Posterior Distributions" (with discussion). *The Annals of Statistics* 22:1701-1762.

Figure 1. Logistic and Normal Estimates of Alpha Parameters on Simulated Data. (Logistic estimates are rescaled by 1/2.4. Error bars are \pm two standard deviations.)

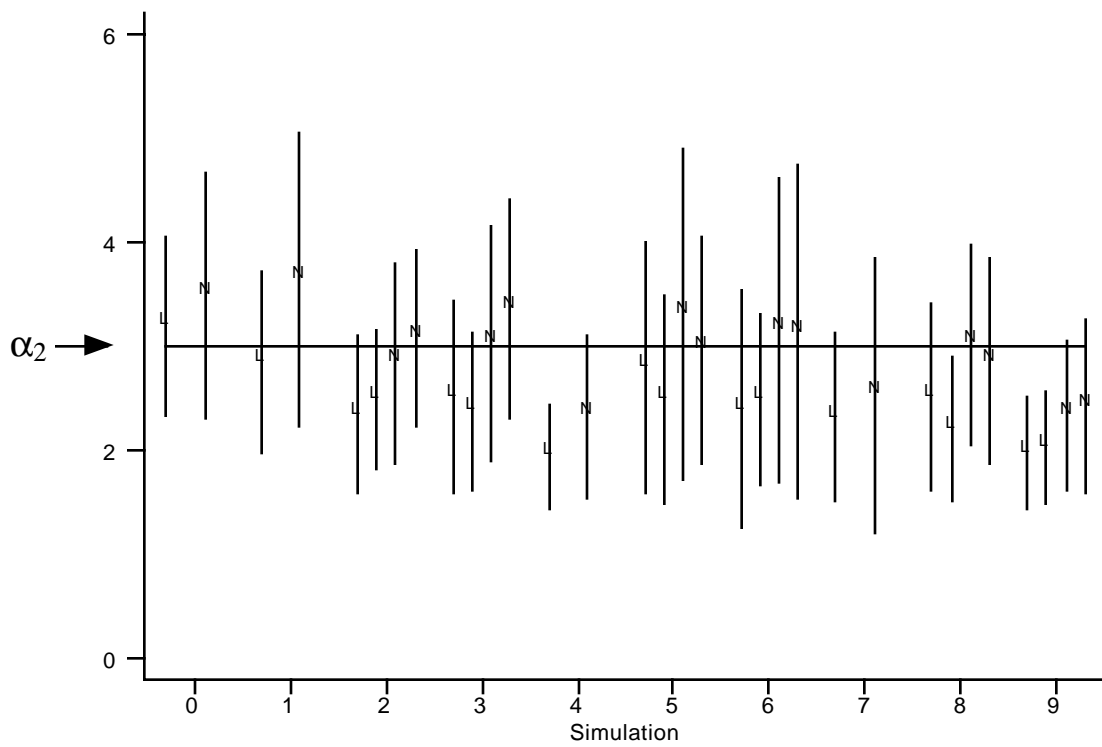
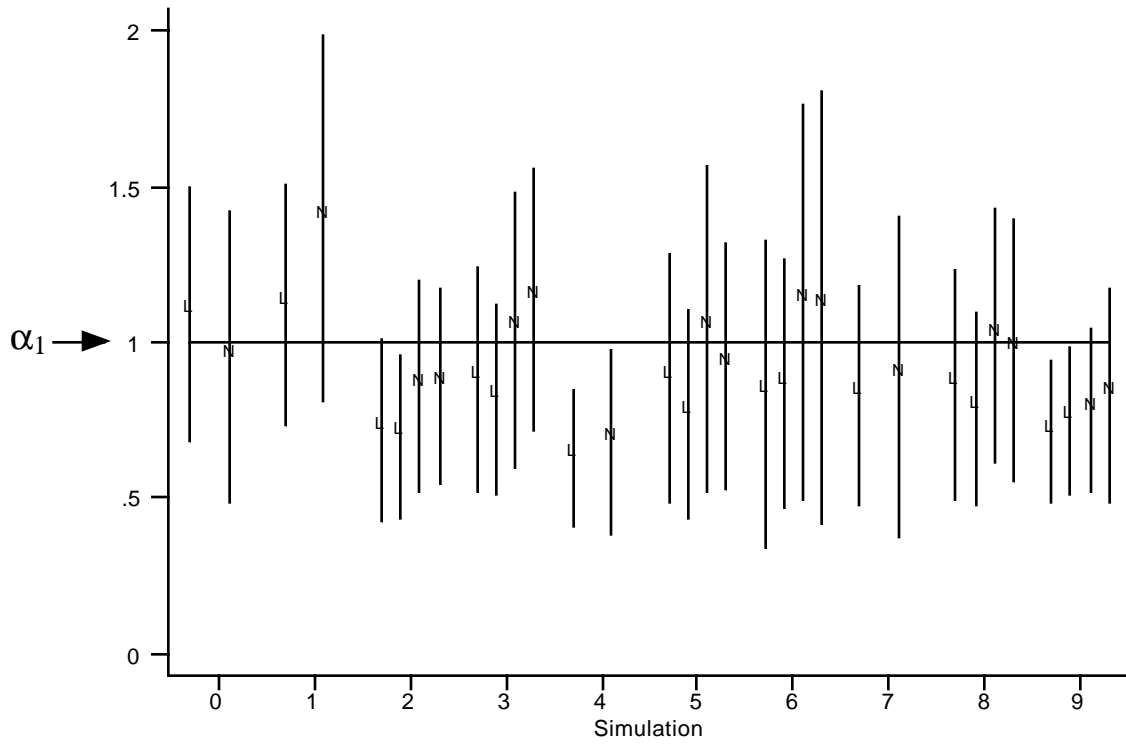


Figure 2. Logistic and Normal Estimates of Beta Parameters on Simulated Data. (Logistic estimates are rescaled by 1/2.4. Error bars are \pm two standard deviations.)

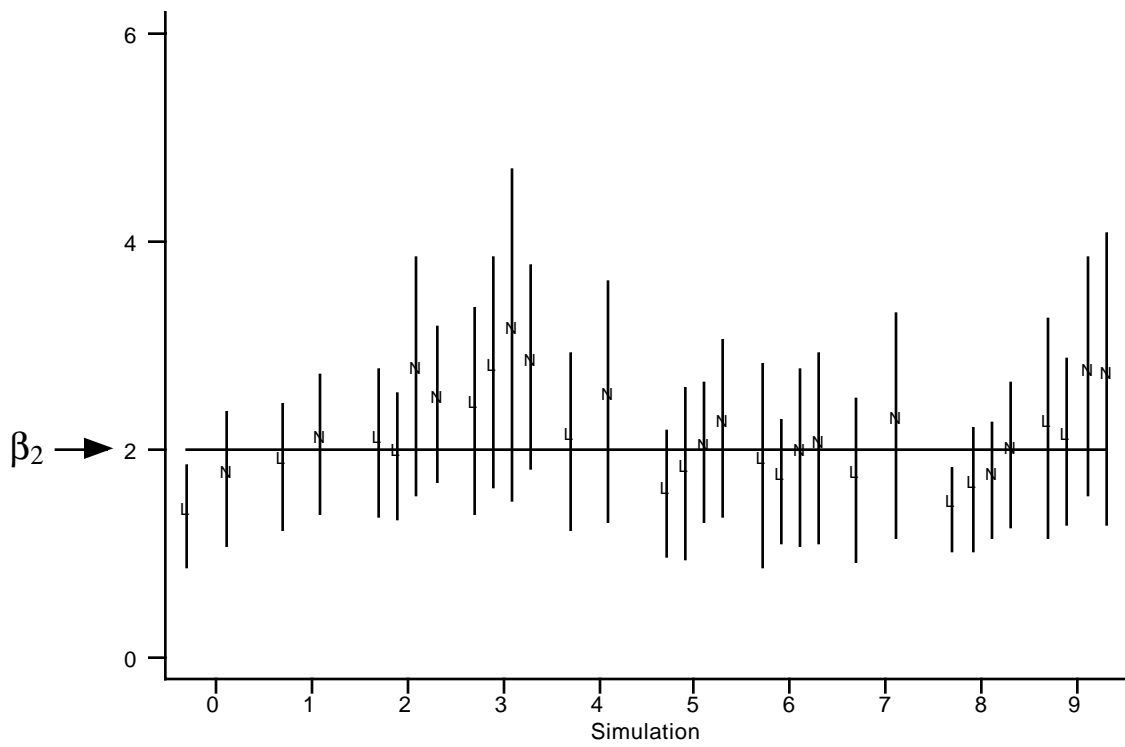
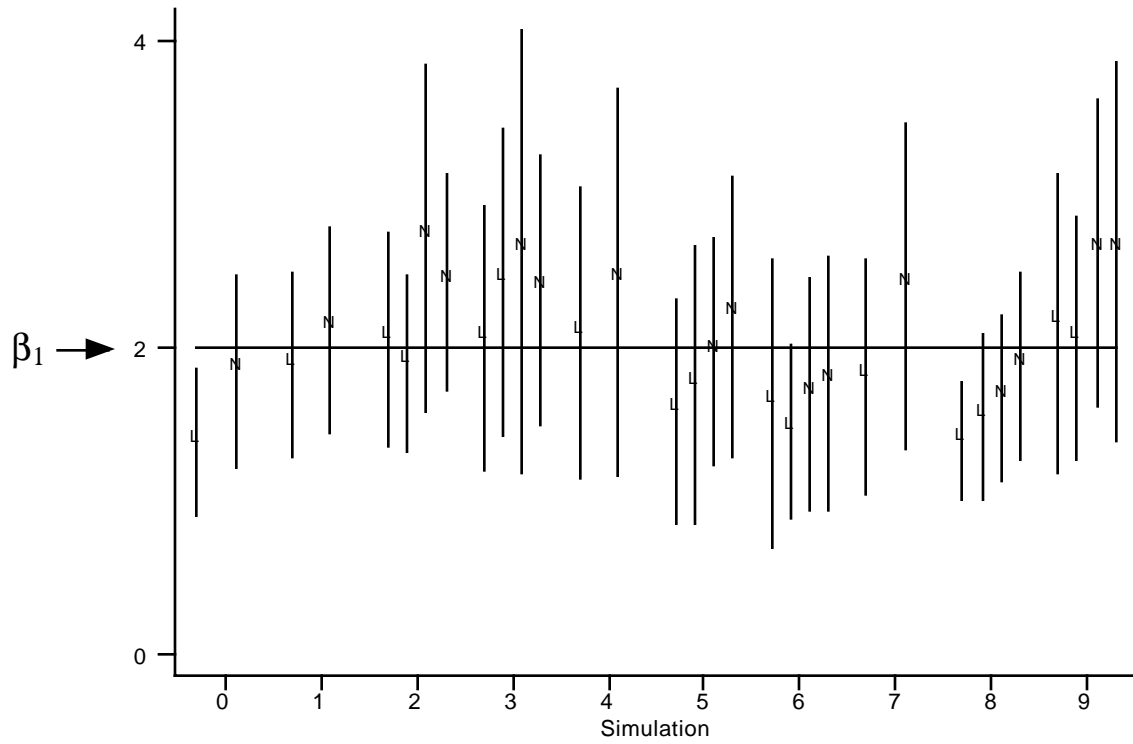


Figure 3. Unblocked and Randomly Blocked Logistic Model Alpha Estimates on Simulated Men-Proposing Data. (Unblocked, and block sizes 22 and 5, in order.)

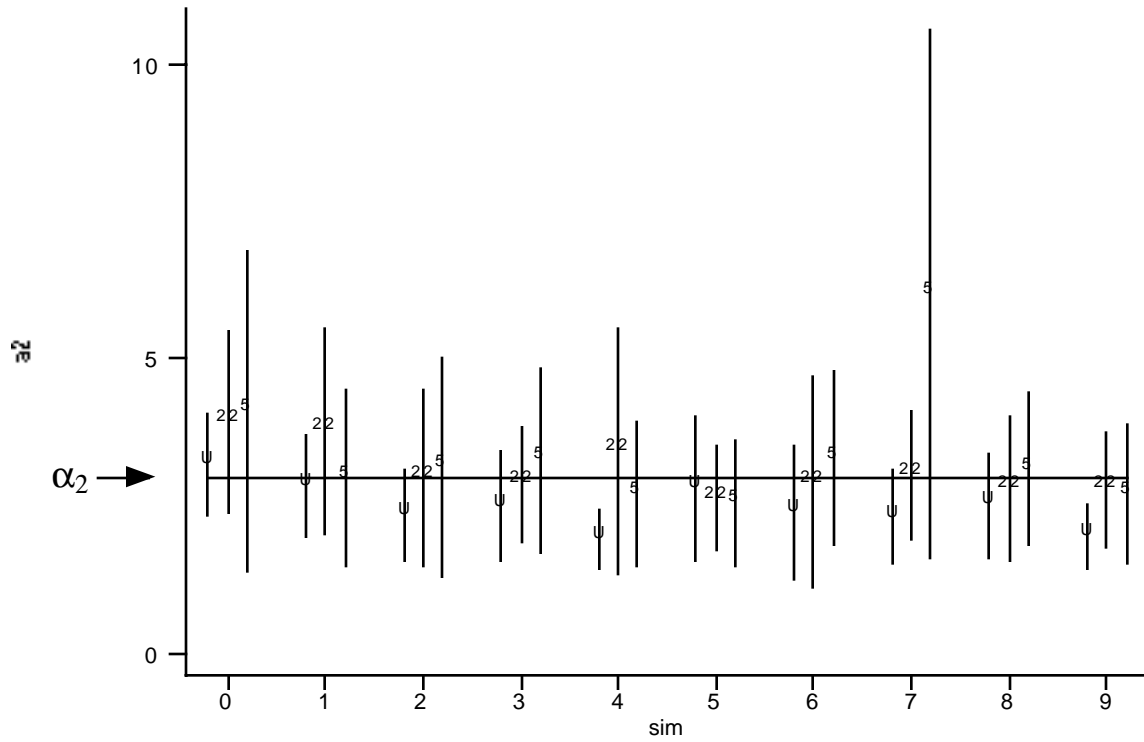
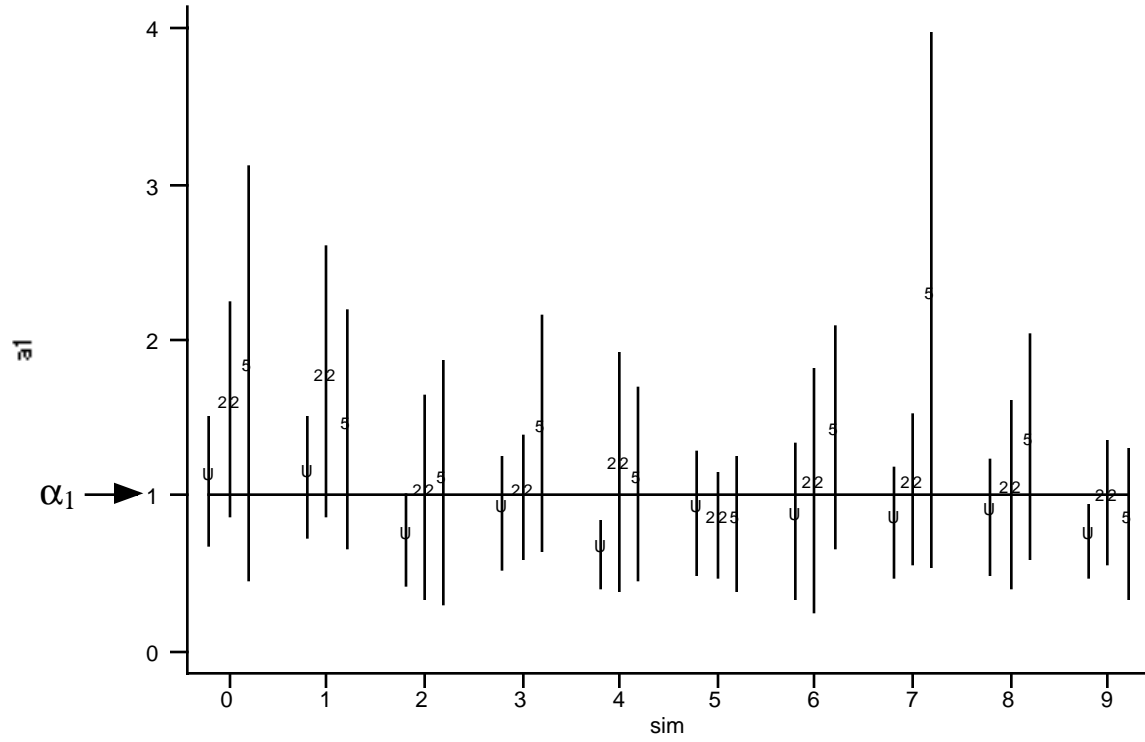


Figure 4. Unblocked and Randomly Blocked Logistic Model Beta Estimates on Simulated Men-Proposing Data. (Unblocked, and block sizes 22 and 5, in order.)

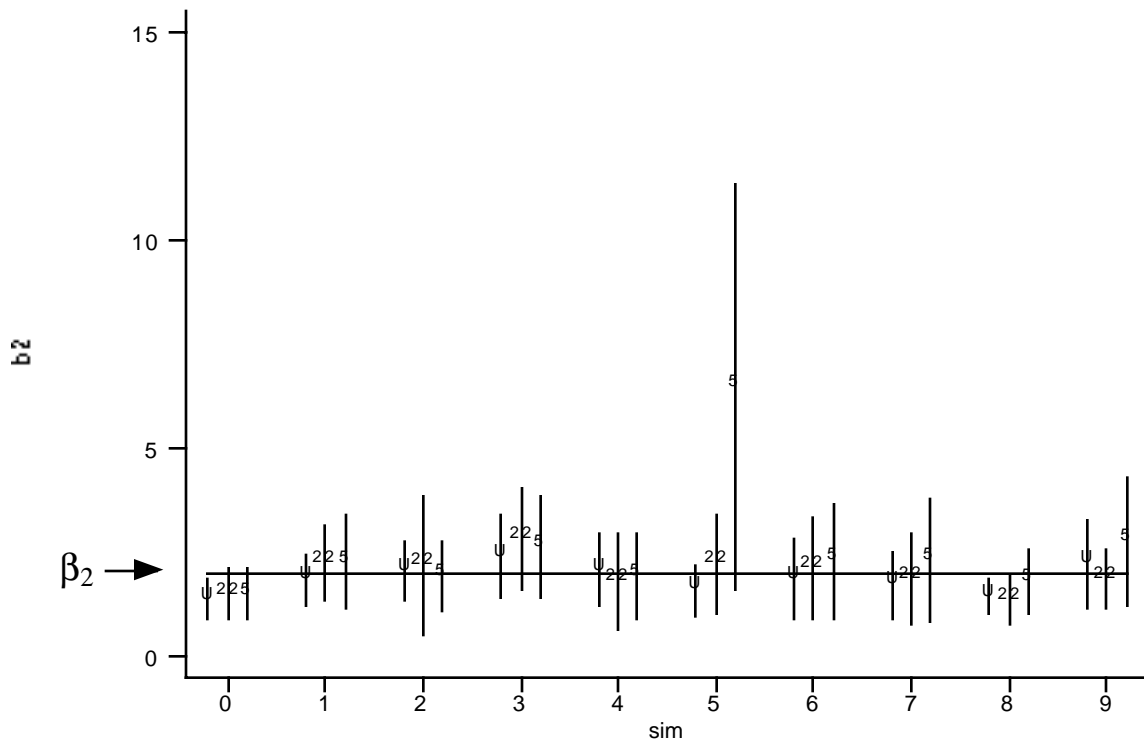
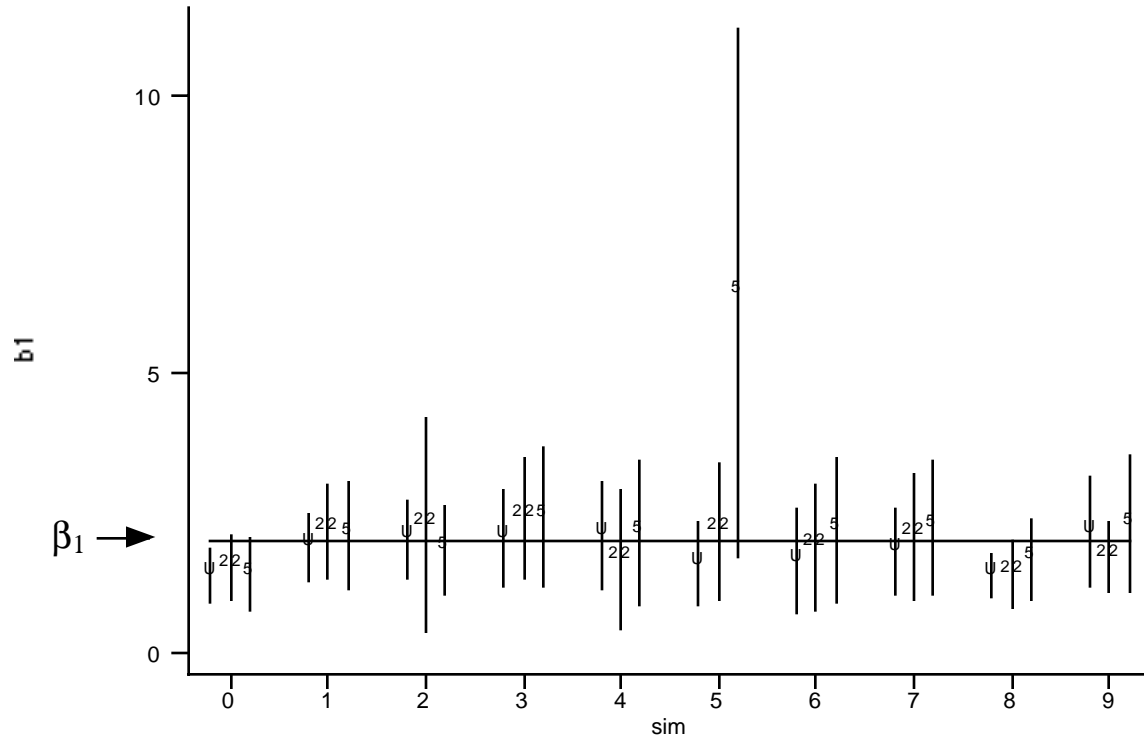
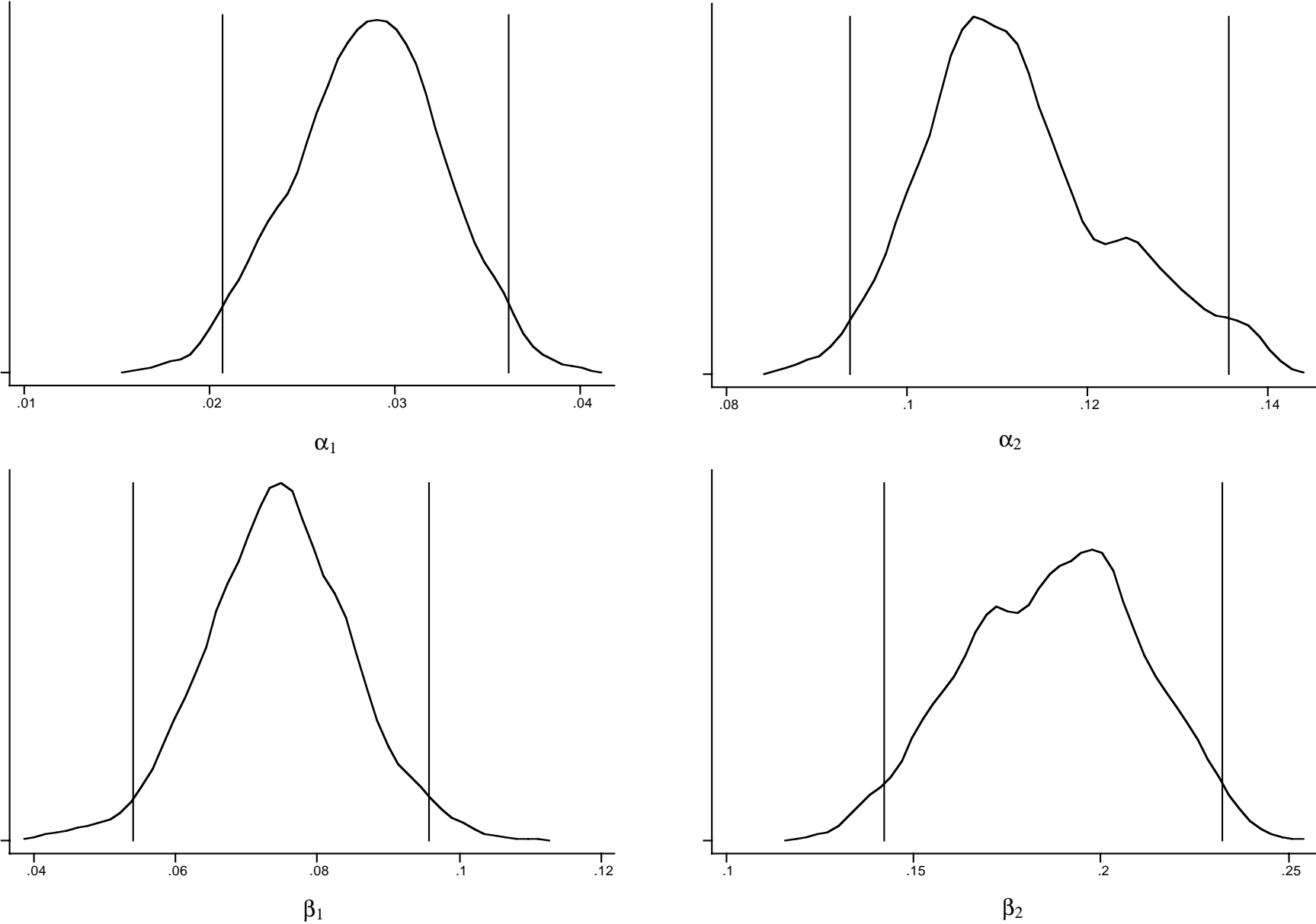


Figure 5. Posterior Kernel Density Estimates, Scans 50,000-500,000 of 500,000, for 2500 PSID Matches. Logistic Algorithm, Blocksize = 5.
(Vertical lines enclose 95% highest posterior density regions. Logistic estimates rescaled by 1/2.4.)



Center for Demography and Ecology
University of Wisconsin
1180 Observatory Drive Rm. 4412
Madison, WI 53706-1393
U.S.A.
608/262-2182
FAX 608/262-8400
comments to: logan@ssc.wisc.edu
requests to: cdepubs@ssc.wisc.edu