DEMOGRAPHIC RESEARCH
A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 44, ARTICLE 22, PAGES 513–536
PUBLISHED 17 MARCH 2021
https://www.demographic-research.org/Volumes/Vol44/22/
DOI: 10.4054/DemRes.2021.44.22

Research Article

A counterfactual choice approach to the study of partner selection

Aaron Gullickson

Contents

Demographic Research: Volume 44, Article 22
Research Article

A counterfactual choice approach to the study of partner selection

Aaron Gullickson[1]

Abstract

BACKGROUND
Research on assortative mating – how partner characteristics affect the likelihood of union formation – commonly uses the log-linear model, but this approach has been criticized for its complexity and limitations.

OBJECTIVE
The objective of this paper is to fully develop and illustrate a counterfactual model of assortative mating and to show how this model can be used to address specific limitations of the log-linear model.

METHODS
The model uses a sample of alternate counterfactual unions to estimate the odds of a true union using a conditional logit model. Recent data from the United States are used to illustrate the model.

RESULTS
Results show important biases can result from assumptions about the marriage market implicit in existing methods. Assuming that spouses are drawn from a national-level marriage market leads to underestimates of racial exogamy and educational heterogamy, while the exclusion of the unmarried population (the unmarried exclusion bias) leads to overestimates of these same parameters. The results also demonstrate that controls for birthplace and language endogamy substantially affect our understanding of racial exogamy in the United States, particularly for Asian and Latino populations.

CONCLUSIONS
The method gives the researcher greater control of the specification of the marriage market and greater flexibility in model specification than the more standard log-linear model.

CONTRIBUTION
This paper offers researchers a newly developed technique for analyzing assortative mating that promises to be more robust and flexible than prior tools. Further, it demonstrate best practices for using this new method.

[1] Department of Sociology, University of Oregon, USA. Email: aarong@uoregon.edu.

1. Introduction

Demographers and other social scientists have long been interested in the process of partnering up in marriage, cohabitation, and other romantic relationships. Much of the interest within this broad literature is on assortative mating – who partners with whom based on characteristics such as race, religion, education, age, and attractiveness. Answering these questions can address issues of growing inequality between families, the strength and durability of social boundaries, and the implicit valuation of partner characteristics (Kalmijn 1998; Schwartz 2013).

Methodological concerns about how to measure assortative mating have long been an issue for scholars working in this field. Partnering up is complicated by the fact that it is a joint decision made simultaneously by both individuals involved. Implicitly, this decision is made by each participant from among a set of other individuals with whom they could have alternatively partnered, in addition to the possibility of not forming a union at all. Scholars have tackled this problem using a variety of approaches including Schoen's harmonic mean model, agent based modeling, and the use of novel data, but the standard workhorse for the field has been the log-linear model. While log-linear models have many advantages, they also suffer from certain limitations.

In this article, I illustrate and further develop a newer methodology for measuring patterns of assortative mating, using a technique that more closely matches the implicit counterfactual involved in partnering up. Why do participants end up with who they do rather than the alternative partners available to them? The approach is simple and intuitive. For each real union, the researcher generates a set of alternate unions for existing partners and then uses these counterfactual unions to determine how partner characteristics affect the likelihood of real partnering. This approach is akin to the well known discrete choice model used in economics to determine what choice consumers make based on the characteristics of the choice being selected (McFadden 1973).

Although more computationally intensive and data demanding, the approach can address limitations inherent to log-linear models and has much greater flexibility in model specification. Specifically, the model allows for a more focused identification of the "market" from which partners are drawn than is feasible in a log-linear model and the specification of independent variables is much simpler due to the model's familiar linear structure.

A smattering of existing research on assortative mating has used some version of this model framework (Dalmia and Lawrence 2001; Jepsen and Jepsen 2002; Nielsen and Svarer 2009; Qian and Lichter 2018). However, the method has received relatively little attention and has never been fully developed as an alternate method. In this article, I develop the model more systematically, address unresolved issues around sampling procedures, and demonstrate its strengths and weaknesses relative to existing methods.

I use this model to examine patterns of racial exogamy and educational heterogamy in marriage in the United States.[2] Using this method, I show that patterns of racial exogamy, in particular, are sensitive to the specification of the relevant marriage market, the identification of alternate partners, and controls for birthplace and language endogamy.

2. The limitations of log-linear models

Log-linear models have become the standard method among scholars studying assortative mating because they resolve important limitations with simpler binary and multinomial logit model approaches. In the binary/multinomial logit framework, the researcher treats one characteristic of the respondent's union (e.g., an interracial union, educational hypogamy/hypergamy) as the outcome that can be predicted by other characteristics. The most consequential problem with this approach is that it does not adequately address differences in group size that affect the likelihood of union formation even when underlying propensities remain the same. For example, because of group size, Asian Americans will have far fewer opportunities to marry members of the same race than will White Americans. Thus, we would expect that even if the propensity to marry across racial lines is the same for both groups, Asian Americans will have a much higher likelihood of marrying someone of a different race. To address this limitation, researchers have utilized multilevel models that include group composition measures at some higher geographic level (De Hauw, Grow, and Van Bavel 2017; Qian, Lichter, and Tumin 2018). However, this approach assumes, without justification, a linear functional form between the composition variable and the log-odds of an outcome. Furthermore, compositional measures are typically not included for covariates in the same model and it is unclear how to even correct for such compositional issues within this model framework. A secondary limitation of the binary/multinomial logit approach is the difficulty of accounting for other patterns of assortative mating as covariates within the model framework. Because the distribution of education often varies by race, we might, for example, be interested in how a tendency toward educational homogamy affects the likelihood of racial exogamy.

Log-linear models resolve both of these issues in their intrinsic model structure. Differences in group size are factored out in the marginal effects of all variables and researchers then examine more informative patterns of association in the multi-way effects. Controlling for other patterns of assortative mating is as simple as including

[2] Assortative mating scholars often use the terms exogamy/endogamy to identify marriage across lines of clearly bounded identities such as race or religion and heterogamy/homogamy to refer to marriage along less bounded status characteristics such as education.

additional multi-way effects into the model. However, log-linear models can be complicated to specify and difficult to interpret, leading some scholars to criticize their "inscrutable complexity" (Rosenfeld 2005: 1287). Outside of mobility research in sociology, log-linear models have few other applications and because they lack a dependent variable in the same sense as most linear models, they are less intuitive to nonspecialists. Because most terms in log-linear models are specified through interaction terms, increasing the number of variables rapidly increases the complexity of the model. Significant debate among scholars has persisted over models with as few as four variables (Rosenfeld 2010; Gullickson and Fu 2010; Kalmijn 2010). Additionally, log-linear models were originally developed to handle only categorical variables, limiting their flexibility. Researchers have developed methods for including quantitative variables into the log-linear model, but the technique is limited to averaging scores across the groups used to construct the basic contingency table (Hout 1984; Hout and Goldstein 1994).

Aside from these issues of complexity and interpretation, log-linear models suffer from two important substantive limitations relevant to the study of assortative mating. First, log-linear models only account for group size at the geographic level that data are aggregated. Substantial dissimilarity in the distribution of relevant characteristics at a lower geographic level that more accurately captures the market of potential partners may bias the results of models estimated at the higher geographic level. Typically, researchers use nationally representative data and do not take account of variation at the level of states, counties, or cities that could affect results. Racial groups in the United States, for example, have different geographic distributions across the country. These different patterns of settlement limit opportunities for interracial contact that would facilitate partner formation and thus deflate estimates of the likelihood of interracial marriage when the country as a whole is considered as the marriage market (Harris and Ono 2005). One could add in geographic level as another variable in the model, but this adds considerable complexity and can lead to issues of sparseness in model fitting.

Second, log-linear models only include individuals who are currently partnered. Thus, these models implicitly compare existing partners to the set of potential partners one could have had that are currently partnered with someone else. Individuals who are currently single are excluded from the model. I refer to this limitation as the *unmarried exclusion bias*. It ignores the potential that unattractive characteristics may lead to a particular partner not being chosen at all and thus may significantly bias our understanding of the assortative mating process. Schoen's harmonic mean method is an alternate method that adjusts for both compositional issues in terms of group size and the unmarried exclusion bias, by using population level data in the denominator of its "magnitude of marriage attraction" parameter (Schoen 1986; Schoen, Wooldredge, and Thomas 1989). However, because the harmonic mean method does not use a linear model framework, estimating assortative mating simultaneously along multiple dimensions

requires the researcher to manually create a compound set of categories from all the possible combinations of given variables. For example, Schoen and Wooldredge (1989) use a two-step procedure in which they estimate a very large array of age, race, and education specific marriage rates using the harmonic mean and then enter these rates into a linear model framework for further analysis. The method thus requires considerable effort on the part of the researcher, especially when considering alternate specifications of variables. Consequently, this method has fallen out of favor relative to log-linear models which, for all of their complexity, offer a more routine approach to including multiple variables.

Another approach to addressing the unmarried exclusion bias is to use the multinomial logit model approach and include unmarried as a category of the outcome variable (De Hauw, Grow, and Van Bavel 2017). However, this approach retains all of the limitations associated with a multinomial logit model described above.

In the remainder of this article, I will outline an alternative approach that maintains all of the advantages of a log-linear model in its intrinsic framework, while at the same time accounting for the problems of geographic scope and the unmarried exclusion bias.

3. A counterfactual approach

The analysis of assortative mating can be seen as a question about a counterfactual case: why was this union chosen instead of other unions that could have been chosen? The model I use here explicitly builds on this intuition by generating a set of counterfactual unions for each actual union. Counterfactual unions are created by sampling partners from a list of available partners within the same defined marriage market. For every actual union i, a choice set of J unions is created that includes the actual union as well as $J-1$ counterfactual unions. A vector of characteristics $\mathbf{x}_{ij}$ can then be defined for every possible union j in the choice set i. These characteristics will typically be defined based on the joint characteristics of the partners in a potential union, such as age differences, educational differences, or racial differences. P_{ij} is the probability that a given union within the choice set is the actual union. The relationship between $\mathbf{x}_{ij}$ and P_{ij} can then be estimated via a conditional logit model:

$$P_{ij} = \frac{e^{\mathbf{x}_{ij}\beta}}{\sum_{k=1}^{J} e^{\mathbf{x}_{ik}\beta}}$$

The estimated β parameter provides log-odds ratios indicating how the log-odds of union formation change as a function of the covariates. In practice, this conditional logit

model can be estimated using a fixed-effects logistic regression model with dummies for each choice set i. These fixed effects ensure that estimates are driven only by the choice made within each choice set.

This relatively simple model has several advantages over the more familiar log-linear model approach. First, the results are more directly interpretable and intuitive as the change in the log-odds of a given match based on some function of partner characteristics. Second, because this approach uses a generalized linear model, researchers can easily incorporate a variety of categorical and quantitative variables as predictors, along with nonlinear terms, quadratics, interaction terms, and the like. Third, because researchers can define the parameters for what defines a marriage market from which alternate partners will be selected, this model implicitly controls for differences in partner availability in different marriage markets. Finally, because the definition of alternate partners can include unmarried individuals, this model incorporates information about the unmarried population.

This model framework bears some similarity to the two-sided probit model proposed by Logan, Hoff, and Newton (2008). Both approaches use a counterfactual set of fictional unions to estimate the probability of a match. The main difference is that Logan, Hoff, and Newton (2008) use a more complex theoretical model arising from game theory to estimate separate parameters for men and women. The result is a model that is far more difficult to estimate using standard techniques. The model proposed here makes less assumptions and does not attempt to differentiate distinct parameters for men and women.

This model, like other common approaches to measuring assortative mating, does not explicitly distinguish between individual preferences in mate selection and constraints on those preferences that affect the opportunity individuals have to meet certain potential partners. For example, structural impediments such as neighborhood and occupational segregation will limit access to potential partners of a different race. Thus, the results from these models should not be construed purely as reflecting the preferences or choices made by individuals, but rather the combination of preferences and constraints operating on the marriage market.

Nonetheless, this model does have the ability to easily include constraints when data allow for it. First, by limiting the constructed choice sets based on some selection criteria, this model has the capability to factor out some basic constraints. For example, by limiting choice sets to individuals in the same metropolitan area, I can eliminate the constraint on potential partners that live far from one another. Second, when additional data on constraints are available, the linear framework of this model makes it easy to incorporate such constraints. For example, Nielsen and Svarer (2009), use this modeling approach on detailed Danish data to include a dummy variable indicating whether potential partners attended the same educational institution. They find that accounting for this form of constraint substantially reduces the strength of educational homogamy.

In the sections that follow, I describe how to set up the data for this model and describe the model and some of its limitations in more detail. I then show how the model can be used to model assortative mating by age, education, and race in the United States. Finally, I show how the specification of the model specifically improves our understanding of assortative mating with an analysis of how sensitive estimates are to the specification of the marriage market and alternate partners.

4. Data

To demonstrate this modeling approach, I use data from the 2012–2017 American Community Survey (ACS) samples. The ACS is an annual 1-in-100 survey of the United States population, conducted by the United States Census Bureau. Since some of the combinations I analyze below are relatively rare, I pool the annual samples over five years to increase the overall sample size. All data were accessed using the Integrated Public Use Microdata Series (IPUMS), which provides some additional constructed variables that I utilize for the analysis (Ruggles et al. 2020).

I use the pooled ACS samples to construct a dataset of married couples and a dataset of all potential marriage partners, both of which are stratified by geographic marriage markets. In order to capture information about the current marriage market, I limit the sample of married couples to those couples who were married in the previous year. The sample of potential partners includes all of the partners in actual marriages from the previous year as well as single individuals who were neither widowed nor divorced in the previous year. Because of the difficulty of identifying potential partners for same-sex unions, the analysis is limited to heterosexual married couples and the pool of potential partners will include some gay and lesbian individuals.

I define marriage markets by the current metropolitan statistical area (MSA) of the respondent. Due to privacy limitation on the public-use Census data, I can only identify 260 distinct MSAs among the respondents. As a sensitivity analysis, I also consider an alternate marriage market of the entire United States, but respondents are still limited in this case to the 260 distinct MSAs identified in the data in order to isolate the effect of marriage market classification on model parameter estimates.

I apply restrictions to eliminate individuals who may have not been married or eligible to be married in their current marriage market for a full prior year. I remove all respondents from both samples who either migrated to the United States within the last year or moved from a different state within the last year. It is not possible with the existing data to eliminate individuals who may have moved in or out of the MSA, but within the same state. After all sample restrictions, I have a total of 63,114 actual marriages, 1,846,030 alternate male partners and 2,255,952 alternate female partners.

For the analysis I simplify educational attainment for all respondents into a set of four ordinal categories of less than high school completion, high school completion, some college but less than a four-year degree, and a four-year college degree or more. For race, I collapse race responses into White, Black, American Indian, Asian/Pacific Islander, and Latino. The model can become difficult to estimate on smaller racial categories so I exclude those who responded as non-Latino other race and those who identified as belonging to multiple races. Hispanicity is asked as a separate question from race in the ACS, but I classify all respondents with a positive response to Hispanicity as Latino by race.

5. The model

The analytic dataset is created by constructing choice sets for each actual marriage in which the marriage is combined with several counterfactual marriages. For each actual marriage, I randomly choose to sample either alternative husbands or wives. I sample a certain number (n) of alternate partners within the same marriage market. The final choice set is the actual couple along with n counterfactual marriage alternatives that did not happen. This procedure is repeated for all actual marriages.

Independent variables are constructed from each partner's characteristics and/or the combination of characteristics. The estimated effects from this model can be interpreted as the change in the log-odds of union formation as a function of partner characteristics.

Because the model follows a standard linear model structure, researchers have great flexibility in specifying this model. As a baseline example, I fit a model that accounts for age differences between the spouses, educational heterogamy, and racial exogamy. To estimate the effect of age differences between partners on the likelihood of union formation, I take the linear age difference and its square. This allows for a nonlinear parabolic effect from which I can calculate the age difference between spouses where the likelihood of union formation is maximized.

I calculate the effect of educational heterogamy on union formation using the common educational crossing model. This creates three dummy variables at the level of high school completion, some college completion, and college completion. Cases where one partner is above this threshold and the other is below are coded as 1 while cases where both partners are on the same side of the threshold are coded as 0. These parameters indicate how the log-odds of union formation change when the educational differences between potential partners cross a given boundary. These terms are cumulative such that the log-odds of a union between one partner with less than high school education and another partner with a college degree would include all three terms. I model these thresholds as gender-symmetric, but I also include separate dummy variables for whether

the potential union would be educationally hypergamous (the woman has less education than the man) or educationally hypogamous (the woman has more education than the man).

I model racial exogamy through a series of dummy variables that capture racial exogamy between every possible exogamous combination (e.g., Black/White, Latino/Asian). These parameters estimate the log-odds of a union involving this form of racial exogamy relative to a union in which the partners are racial endogamous. I use gender-symmetric terms here, so that the model implicitly averages across the exogamy effects for each possible gender combination (e.g., Black husband/White wife, White husband/Black wife), but the model can easily be extended to create gender-specific dummy variables instead.

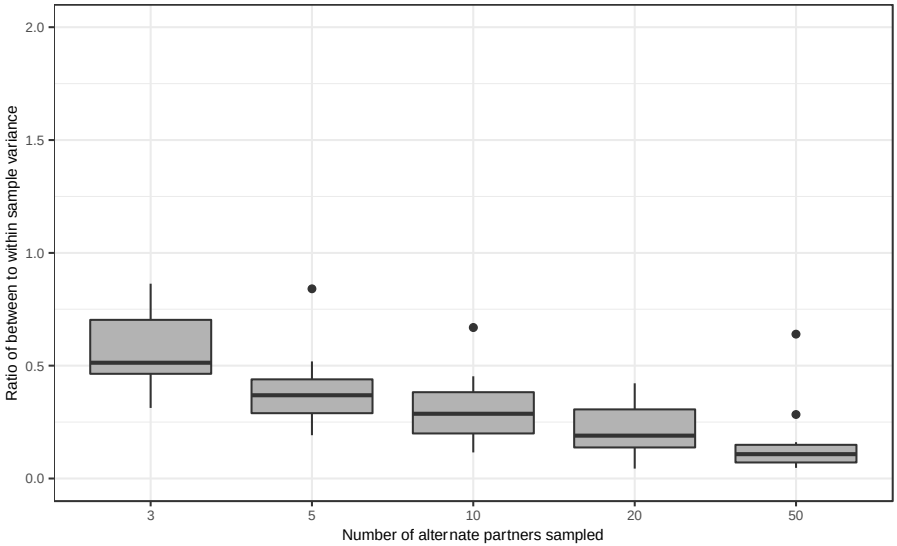
Finally, I consider several other variables that capture immigration-related factors that are likely to affect the racial exogamy of Asian and Latino individuals. First, I account for birthplace endogamy by coding a dummy variable indicating whether the potential spouses were born in the same country. If, for example, Mexican-born immigrants tend to prefer other Mexican-born immigrants as partners, this will indirectly increase Latino exogamy, even though it reflects a form of endogamy distinct from racial endogamy. Similarly, I account for language endogamy, based on the language that each partner reports speaking at home.[3] I also account for a general 'nativity endogamy' by including separate gender-specific exogamy terms for potential unions where one partner is an immigrant and the other partner is native-born.

Two technical issues need to be addressed before estimating this model. First, how large should n be? Jepsen and Jepsen (2002) argue, based on prior work with discrete choice models, that small numbers of three or five alternate choices are often sufficient for precise model estimation. To test this assertion, I estimated the models detailed above using varying numbers of sampled alternate partners (n) of 3, 5, 10, 20, and 50. For each value of n, I conducted twenty separate draws of alternate partners and estimated parameters for each draw. I then estimated the ratio of variance within (the mean standard error squared across all twenty estimates) and between (the variance of the parameter estimate across all twenty estimates) for each parameter. Figure 1 shows boxplots of this ratio across all fifteen parameters for all twenty draws for each sample size. The share of between sample variability diminishes significantly in larger samples. In samples of size three the variability between samples is nearly 50% of the variability within samples for the median parameter. This diminishes substantially in larger samples but is still sizable for samples of size 20. The only drawback to drawing a larger number of alternate partners is the computation time required. Therefore, while it may be useful to develop exploratory results on smaller samples, I would recommend that researchers sample at

[3] Because the language spoken at home is recorded after the union, this form of endogamy may be somewhat overstated in the model. Some partners may have shifted their language of choice after union formation.

least twenty alternate partners for final analysis. In the analysis that follows, I sample fifty alternate partners for each actual union.

Figure 1: Boxplots of the ratio of variance between samples to within samples for each parameter in a model with fifteen parameters, based on the number of alternate partners sampled for each marital union



Note: Models are estimated on twenty independent draws for each value.

Second, how do we account for the additional form of variability as a result of sampling alternate partners? Each time we make different draws, we will get somewhat different outcomes. This variability is quite similar to the extra variability that arises in multiple imputation of missing values. It can be handled in an analogous fashion by drawing m distinct analytic samples and then pooling results by taking the mean of all estimates and a pooled standard error that accounts for the added variation due to randomly sampling alternate partners. The standard error for a given parameter, β, from the model can be calculated by:

$$SE_\beta = \sqrt{W + B + \frac{B}{m}}$$

where W is the average squared standard error from each model and B is the variance in β between models.

In all of the models that follow, I sample 50 alternate partners ($n = 50$) and draw five ($m = 5$) distinct analytic samples. Final results are pooled as described above.

The labor-intensive component of this analytic technique is the creation of the analytic dataset by sampling alternate partners. As part of this project, I have created an *R* package that will sample a given number of alternate partners for each union and construct a full analytic dataset based on input of actual marriages and a full set of alternative partners.[4]

6. Results

Table 1 below shows three different conditional logit models predicting the log-odds ratio of an actual union. Model 1 includes only age differences and racial exogamy terms. I include educational heterogamy terms in Model 2 in order to see how much educational differences might affect the racial exogamy estimates. Model 3 then includes additional immigration-related factors.

The age effects included in every model highlight the method's ability to easily include both quantitative variables and nonlinear effects on those quantitative variables. In this case, I have fit the effects of age difference with a linear and squared term, to create a parabolic function. Figure 2 graphs the odds ratios from this parabola for age differences up to twenty years in either direction based on the results from Model 3. The model predicts that the odds of union formation are maximized when the husband is 2.76 years older than the wife. These odds drop precipitously for unions that stray even by as much as five years from this peak. At ten years of age difference, the odd ratio is very low (around 0.12). For couples more than twenty years apart in age, the odds of union formation are minuscule.

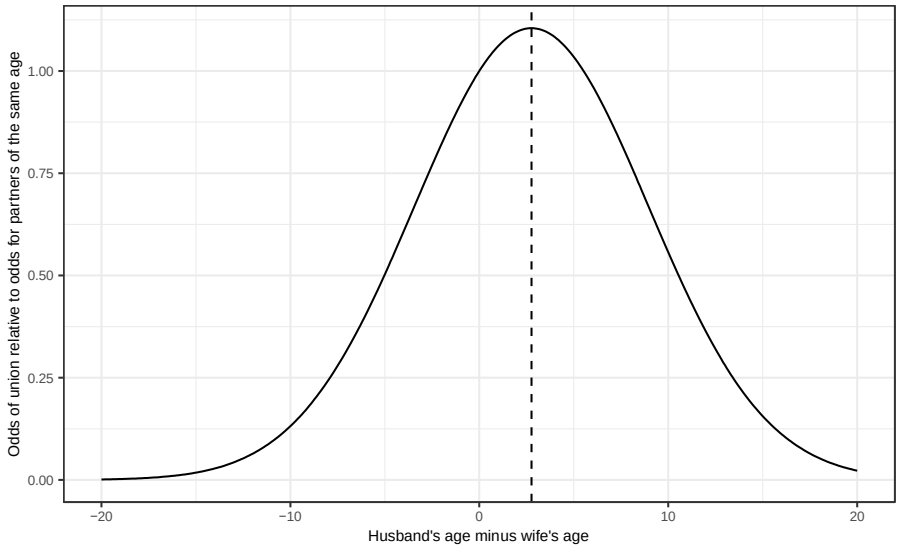
[4] This package is available at https://github.com/AaronGullickson/fakeunion The package also has functions that assist in pooling model estimates across multiply sampled analytic datasets.

Table 1: Estimated log-odds ratios from conditional logit models predicting a marital union by age differences, racial exogamy, educational heterogamy, and immigration-related factors

	Model 1	Model 2	Model 3
Age difference			
Husband's age – wife's age	0.072 (0.001)	0.072 (0.001)	0.072 (0.001)
Square of husband's age - wife's age	–0.014 (0.000)	–0.013 (0.000)	–0.013 (0.000)
Racial exogamy (ref. endogamy)			
Black/White	–3.051 (0.031)	–3.011 (0.032)	–2.988 (0.032)
American Indian/White	–1.550 (0.104)	–1.477 (0.106)	–1.379 (0.106)
Asian/White	–2.029 (0.034)	–2.009 (0.033)	–1.294 (0.036)
Latino/White	–1.694 (0.018)	–1.576 (0.019)	–0.892 (0.022)
Black/American Indian	–2.576 (0.262)	–2.540 (0.266)	–2.424 (0.262)
Black/Asian	–3.921 (0.114)	–3.794 (0.113)	–3.054 (0.113)
Black/Latino	–2.930 (0.049)	–2.873 (0.050)	–2.109 (0.051)
Asian/American Indian	–2.802 (0.376)	–2.683 (0.374)	–1.866 (0.373)
Latino/American Indian	–1.966 (0.157)	–1.887 (0.157)	–1.244 (0.161)
Asian/Latino	–2.962 (0.059)	–2.765 (0.059)	–1.771 (0.062)
Educational heterogamy (ref. homogamy)			
Educational crossing, LHS to HS		–0.751 (0.023)	–0.691 (0.023)
Educational crossing, HS to SC		–0.606 (0.014)	–0.597 (0.015)
Educational crossing, SC to C		–0.734 (0.019)	–0.720 (0.019)
Female educational hypergamy		–0.237 (0.021)	–0.242 (0.022)
Female educational hypogamy		0.097 (0.022)	0.091 (0.021)
Immigrant related factors			
Husband immigrant/wife native			–0.005 (0.039)
Wife immigrant/husband native			0.167 (0.042)
Birthplace endogamy			0.674 (0.038)
Language endogamy			1.659 (0.023)
BIC (relative to null)	–155415	–172299	–185003
Number of married couples	63,114	63,114	63,114
Number of alternate unions per actual union	50	50	50

Note: Standard errors are shown in parenthesis. Results are based on five separate datasets generated by sampling fifty alternate partners for each existing marital union. Results across models pooled and standard errors adjusted for variance between estimates.

Figure 2: Odds ratio of a marital union by the difference in spousal age, relative to a union in which both spouses are the same age. The maximum odds ratio is shown by the dotted line at an age difference of 2.76



Note: Model parameters based on Model 3 of Table 1.

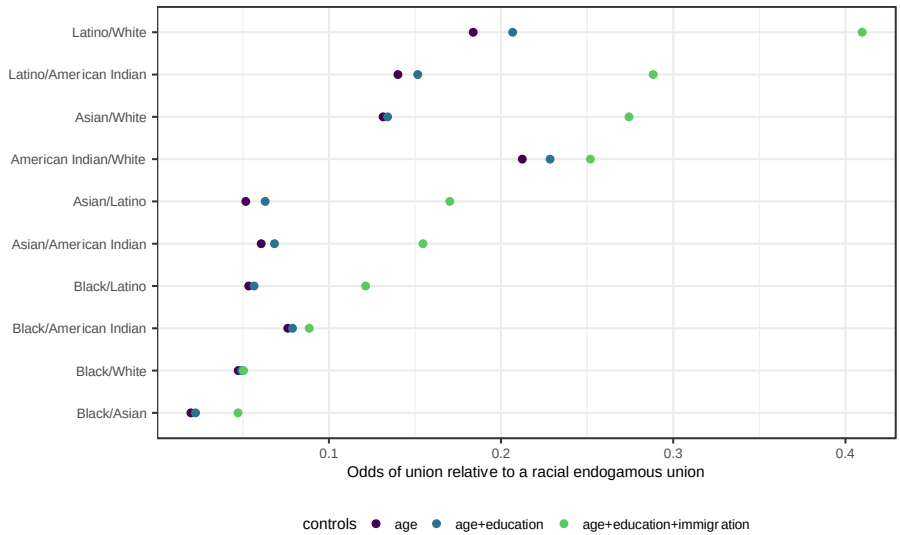
Racially exogamous unions are much less likely than racially endogamous unions, as evidenced by the large negative values for the racial exogamy terms across all of the models. The educational crossing parameters are also consistently negative, although not as large as the racial exogamy terms in absolute effect. Consistent with prior work, the results for the middle crossing term between those with a high school diploma or less and those with some college education or more is somewhat weaker in effect than the two other educational boundaries (Schwartz and Mare 2005). The model indicates about a 9.5% increase in the odds of union formation when the woman is downwardly mobile in comparison to the case when partners are educationally homogamous. Union formation is least likely in cases where women are upwardly mobile. Although the positive results for educational hypogamy are somewhat surprising, the overall findings are consistent with prior work suggesting a global decline in the tendency toward female educational hypergamy in marriage (Esteve, García-Román, and Permanyer 2012).

The results for Model 3 show strong positive effects of birthplace and language endogamy. Interestingly, potential matches in which the husband is native-born and the

wife is an immigrant are more likely (18%) to result in a union than cases where both spouses are either native or foreign born, all else being equal. The same is not true of unions in which the wife is native-born and the husband is an immigrant, which have the same odds as unions that are endogamous with respect to nativity.

A key question that I can address with these models is how the racial exogamy parameters change as we account for other characteristics of potential partners. Figure 3 shows how the odds of racial exogamy change for each interracial combination of spouses across different models. As expected, controlling for educational heterogamy increases the odds of racial exogamy, but the increases are relatively small for most groups. In contrast, controlling for the immigration-related factors substantially increases the odds of racial exogamy for all groups involving an Asian or Latino partner. Part of the low rate of racial exogamy for Asians and Latinos reflects birthplace and language endogamy.

Figure 3: Odds of racial exogamy relative to racial endogamy based on three different models



Note: Results are ordered from largest to smallest based on values from most complex model.

Importantly, the controls in Model 3 substantially shift the rank-ordering of the odds of racial exogamy by specific interracial combinations and reveal important patterns that are otherwise obscured. All four cases involving a Black partner and a non-Black partner have the lowest odds of union formation. Among the remaining cases, those involving a

White partner tend to have the highest odds of union formation, and interracial unions involving Latinos tend to have higher odds than those involving Asians. The placement of American Indians within this hierarchy is harder to classify, but this may reflect the greater uncertainty in model estimates for this group due to small sample size.

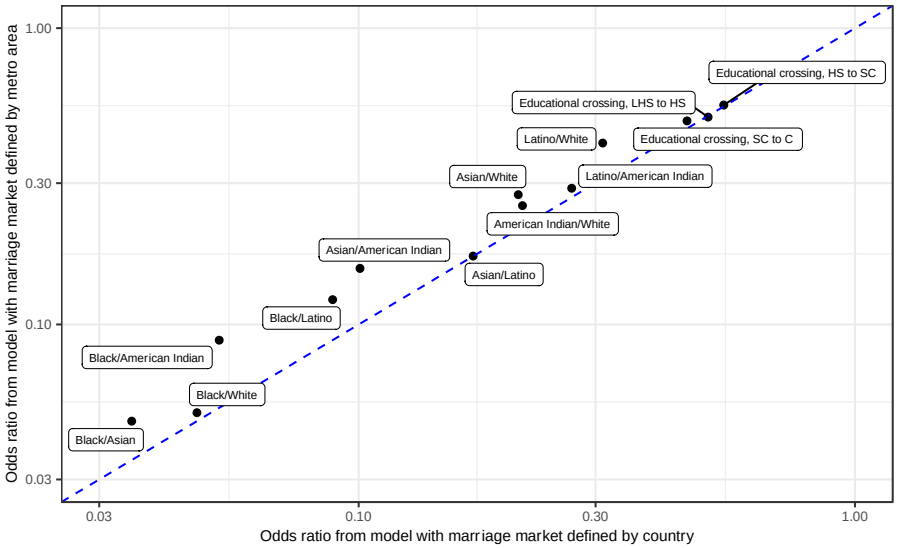
6.1 Marriage market sensitivity

The models above are based on samples where alternate partners are drawn from the same metropolitan area as the extant married couple. No geographic definition of the marriage market will be perfect because partner selection does not always depend on spatial proximity. Nonetheless, metropolitan area is almost certainly a better proxy for the marriage market than large spatial units such as the state or the entire country. However, most prior work implicitly uses the entire country as the marriage market. How much does shifting the definition of the marriage market change our understanding of assortative mating?

In order to address this question, I estimated models identical to Model 3 from Table 1, but using samples that drew alternate partners from the entire country rather than a specific marriage market. Figure 4 shows a comparison of the racial exogamy and educational heterogamy terms for the different sample specifications. The racial exogamy terms are more heavily affected by the marriage market specification than the educational heterogamy terms. The country-based models underestimate racial exogamy relative to the metro area-based models for most pairings, although the Black/White and Asian/Latino terms are similar across both models. Among the educational parameters, the educational crossing parameter from some college to college is also slightly underestimated by the country-based models, but the remaining educational parameters are very similar.

In general, the country-based specification systematically underestimates exogamy/heterogamy across groups when a discrepancy exists. Why would this be the case? Racial and educational groups are not evenly distributed across metropolitan areas in the United States. This unequal distribution somewhat limit's one access to partners of another race or education level and thus deflates exogamy/heterogamy parameters when estimates are generated at the national level.

Figure 4: Scatterplot showing a comparison of the racial exogamy and educational heterogamy parameters from a model using the metro area to sample alternate partners compared to a model using the entire country to sample alternate partners



Note: The dotted line indicates the point where both parameters are identical. Both axes are drawn using a logarithmic scale.

To demonstrate this phenomenon more formally, I calculated the index of dissimilarity for each racial and educational comparison among alternate spouses across each metropolitan area.[5] For race, the index of dissimilarity ranged from a high of 61.7 between Blacks and American Indians to a low of 31.1 for Blacks and Whites. The relative dissimilarity for the educational crossing parameters was much lower. The highest index of dissimilarity among the educational groups was the 12.7 index of dissimilarity between the college-educated and those with less than a four-year college degree.

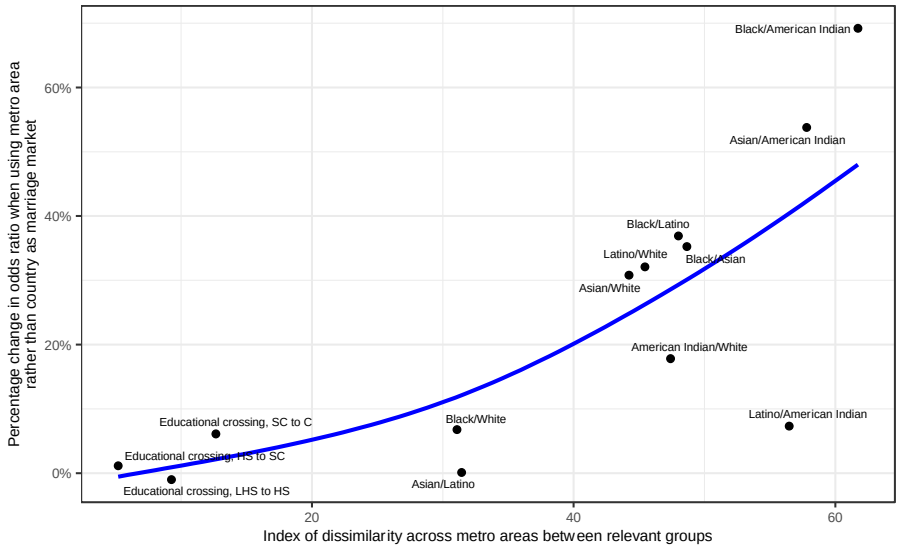
[5] The index of dissimilarity D is calculated as

$$D = 50 * \sum_{\text{all i}} |a_i/A - b_i/B|$$

where a_i and b_i are the total number of members of each group in metropolitan area i and A and B are the total number of members of each group in the United States as a whole. D gives the percent of either group that would have to change metropolitan areas to produce similar distributions by group.

Figure 5 plots these indices of dissimilarity by the relative increase in the odds of exogamy/heterogamy for each term in the model. The model clearly shows a strong positive trend between the two measures. When the index of dissimilarity between potential partners in the marriage market is higher, the bias from a country-based specification of the marriage market is greater. The Latino/American Indian case stands out as an outlier, but this may simply be a result of the large uncertainty on this estimate due to small sample size.

Figure 5: Relative increase in racial exogamy and educational heterogamy when using the metropolitan area as the marriage market rather than the entire country as a function of the index of dissimilarity across metropolitan areas for the relevant groups



Note: The smoothed trend line is calculated using a general additive model smoother.

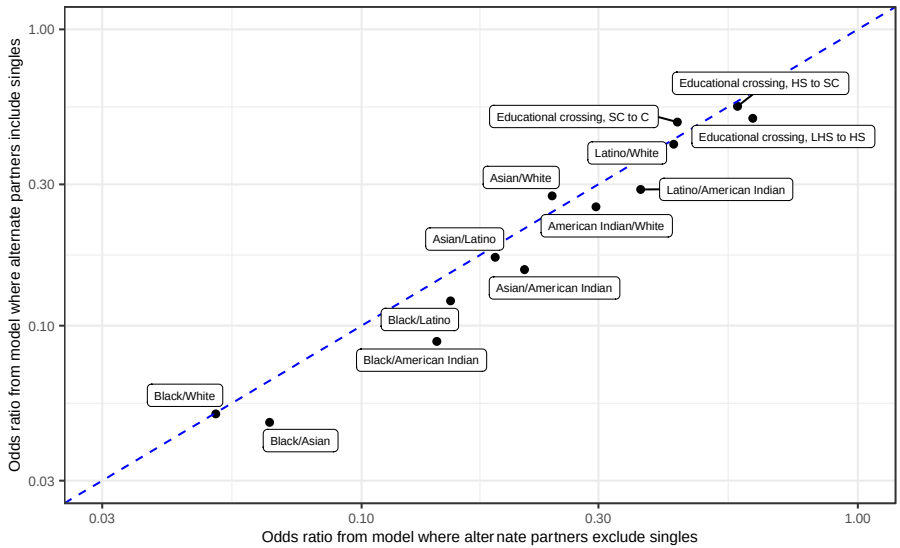
6.2 Unmarried exclusion bias

Another form of bias comes in terms of how alternate partners are defined. Log-linear models only include married couples and thus implicitly exclude unmarried individuals from the definition of alternate partners. This exclusion may create a bias. This bias may be particularly problematic if certain characteristics are more universally valued on the

marriage market (e.g., attractiveness, wealth, education, lighter skin tone) because those with low values on these characteristics may be less likely to partner up at all.

To estimate the potential effect of this bias, I estimate models at the metropolitan area level in which I exclude currently unmarried individuals from the pool of alternate partners. Figure 6 compares the estimates for racial exogamy and educational heterogamy from this restricted model to the estimates from the full model. In cases where a discrepancy exists, the restricted models generally overestimate the odds of racial exogamy and educational heterogamy, although the magnitude and consistency of this bias is generally smaller than that produced by the shift in marriage market definition explored above.

Figure 6: Scatterplot showing a comparison of the racial exogamy and educational heterogamy parameters from a model using both married partners and singles as eligible counterfactual partners vs a model that excludes singles



Note: The dotted line indicates the point where both parameters are identical. Both axes are drawn using a logarithmic scale.

7. Conclusions

In this article, I have outlined and demonstrated an alternative modeling approach to studying how partner characteristics affect union formation. The modeling approach is highly flexible, allowing for complex models that can account for a variety of categorical and quantitative variables. I use this model to estimate patterns of racial exogamy and educational heterogamy in the United States that both affirm prior findings and extend them in important ways.

The models I employ here illustrate how to easily incorporate quantitative variables, nonlinear effects, and how to control for multiple confounding characteristics. In this latter case, the results demonstrate that failure to account for birthplace, nativity, and language endogamy substantially biases estimates of Latino and Asian exogamy downward. Accounting for these factors importantly changes the rank-ordering of different forms of racial exogamy and produces a clearer picture of how the United States racial hierarchy relates to the realm of interracial marriage.

Furthermore, the model used here also accounts for two important sources of bias in existing methods, such as log-linear models. First, most existing studies implicitly use the entire country as the marriage market. As I show, this approach tends to bias estimates of exogamy and heterogamy downward because individuals are not evenly distributed across the country with regard to race, education, or other characteristics of potential interest. The greater the dissimilarity across actual marriage markets in terms of these characteristics, the greater the bias will be.

Second, existing studies typically restrict samples to only married individuals and disregard the unmarried population in the consideration of what predicts union formation. Thus, studies implicitly use only spouses in other existing unions as the counterfactual alternate partners. I describe this as the unmarried exclusion bias. My results show that this bias tends to inflate estimates of racial exogamy and educational heterogamy in models that exclude unmarried individuals. I hypothesize that this bias may arise because when certain characteristics are valued broadly in the marriage market (e.g., education, wealth, attractiveness, lighter skin), those with the least assets in this regard will be more likely to not form a union at all.

This model offers promising and fertile avenues for future research. However, it is not without its limitations. The primary limitation of this model is that is it more demanding of the data than a typical log-linear model. The method relies on having some good geographical identifier for a marriage market. Here I have used metro area, but this leads to an important limitation in restricting the sample only to individuals living in the 260 (typically more populous) metropolitan areas identified in the American Community Survey. Thus, the results might not be representative of populations in less populous metropolitan areas or nonmetropolitan areas. An alternative approach for the United

States that is more inclusive but less precise would use states as the marriage market. Usage of the model in other national contexts will also have to carefully consider how well available geographic identifiers match the theoretical concept of the marriage market.

This model also relies upon having data on alternate partners that includes the unmarried population. This may not be possible in some data sources, particularly when derived from historical data. Furthermore, this requirement highlights two complications with this model that might extend its usefulness: cohabitation and same-sex unions. Theoretically, this model could easily be extended to include cohabitations. A dummy variable could be used to identify cohabiting couples and interacted with relevant variables to allow for differences in the effects of spousal characteristics on cohabitation rather than marriage. However, most large-scale data sources such as the American Community Survey do not contain information on the timing of cohabitation which complicates analysis when marriages are restricted by the timing of marriage. Some other restriction could be used instead, such as the age of partners, but this needs to be done for both marriages and cohabitations.

Defining alternate partners for same-sex unions is also difficult as the sexuality of unmarried individuals is typically not recorded in most large-scale data sources. In this article, I assume that all unmarried individuals are eligible for opposite-sex unions. This assumption is clearly wrong, but any bias this assumption induces will be minimal because the heterosexual population is much larger than the gay and lesbian population. The bias could be large for same-sex unions if the characteristics of the gay/lesbian population are significantly different from the heterosexual population within marriage markets. An alternative approach would be to use a model in which only currently married partners from same-sex couples form the pool of alternate partners. Although this approach will come with some bias, we at least have a sense of how large and in what direction this bias is likely to run.

In addition to the demand it places on data, this model, like other standard methods, cannot generally separate preferences from constraints in partner selection and offers no information about the search process that leads to mate selection. However, in cases where researchers have additional data on potential constraints in the marriage market, the linear framework of this model offers an easy way to control for such constraints.

Due to these limitations, the model described here offers a complement rather than a replacement for log-linear models and other approaches. In situations with certain data limitations, a log-linear model may be a more reasonable, or even the only viable, option. In this case, however, it is important for researchers to be aware of the potential issues of bias arising from marriage market identification and the exclusion of the unmarried population.

References

Dalmia, S. and Lawrence, P.G. (2001). An empirical analysis of assortative mating in India and the U.S. *International Advances in Economic Research* 7(4): 443–458. doi:10.1007/BF02295773.

De Hauw, Y., Grow, A., and Van Bavel, J. (2017). The reversed gender gap in education and assortative mating in Europe. *European Journal of Population* 33(4): 445–474. doi:10.1007/s10680-016-9407-z.

Esteve, A., García-Román, J., and Permanyer, I. (2012). The gender-gap reversal in education and its effect on union formation: The end of hypergamy? *Population and Development Review* 38(3): 535–546. doi:10.1111/j.1728-4457.2012.00515.x.

Gullickson, A. and Fu, V.K. (2010). Comment: An endorsement of exchange theory in mate selection. *American Journal of Sociology* 115(4): 1243–1251. doi:10.1086/649049.

Harris, D.R. and Ono, H. (2005). How many interracial marriages would there be if all groups were of equal size in all places? A new look at national estimates of interracial marriage. *Social Science Research* 34(1): 236–251. doi:10.1016/j.ssresearch.2004.01.002.

Hout, M. (1984). Status, autonomy, and training in occupational mobility. *American Journal of Sociology* 89(6): 1379–1409. doi:10.1086/228020.

Hout, M. and Goldstein, J.R. (1994). How 4.5 million Irish immigrants became 40 million Irish Americans: Demographic and subjective aspects of the ethnic composition of white Americans. *American Sociological Review* 59(1): 64–82. doi:10.2307/2096133.

Jepsen, L.K. and Jepsen, C.A. (2002). An empirical analysis of the matching patterns of same-sex and opposite-sex couples. *Demography* 39(3): 435–453. doi:10.1353/dem.2002.0027.

Kalmijn, M. (1998). Intermarriage and homogamy: Causes, patterns, trends. *Annual Review of Sociology* 24: 395–421. doi:10.1146/annurev.soc.24.1.395.

Kalmijn, M. (2010). Educational inequality, homogamy, and status exchange in black-white intermarriage: A comment on Rosenfeld. *American Journal of Sociology* 115(4): 1252–1263. doi:10.1086/649050.

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(482): 559–569. doi:10.1198/016214507000000996.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (ed.). *Frontiers in econometrics.* New York: Academic Press: 105–142.

Nielsen, H.S. and Svarer, M. (2009). Educational homogamy: How much is opportunities? *Journal of Human Resources* 44(4): 1066–1086. doi:10.3368/jhr.44.4.1066.

Qian, Z. and Lichter, D.T. (2018). Marriage markets and intermarriage: Exchange in first marriages and remarriages. *Demography* 55(April): 849–875. doi:10.1007/s13524-018-0671-x.

Qian, Z., Lichter, D.T., and Tumin, D. (2018). Divergent pathways to assimilation? Local marriage markets and intermarriage among U.S. Hispanics: Marriage markets and intermarriage. *Journal of Marriage and Family* 80(1): 271–288. doi:10.1111/jomf.12423.

Rosenfeld, M.J. (2005). A critique of exchange theory in mate selection. *American Journal of Sociology* 110(5): 1284–1325. doi:10.1086/428441.

Rosenfeld, M.J. (2010). Still weak support for status-caste exchange: A reply to critics. *American Journal of Sociology* 115(4): 1264–1276. doi:10.1086/649051.

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J., and Sobek, M. (2020). *IPUMS USA: Version 10.0 [Dataset]*. Minneapolis: IPUMS. doi:10.18128/D010.V10.0.

Schoen, R. (1986). A methodological analysis of intergroup marriage. *Sociological Methodology* 39: 49–78. doi:10.2307/270919.

Schoen, R. and Wooldredge, J. (1989). Marriage choices in North Carolina and Virginia, 1969–71 and 1979–81. *Journal of Marriage and the Family* 51(2): 465–481. doi:10.2307/352508.

Schoen, R., Wooldredge, J., and Thomas, B. (1989). Ethnic and educational effects on marriage choice. *Social Science Quarterly* 70(3): 617–630.

Schwartz, C.R. (2013). Trends and variation in assortative mating: causes and consequences. *Annual Review of Sociology* 39(1): 451–470. doi:10.1146/annurev-soc-071312-145544.

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