Changes in Assortative Matching: Theory and Evidence for the US

by Pierre-André Chiappori, Mónica Costa-Dias, Costas Meghir

James J. Heckman

THE UNIVERSITY OF CHICAGO

Econ 350, Winter 2023

1. Introduction

- Measuring the extent to which such assortative matching differs between two economies is challenging when the marginal distributions of the characteristic along which sorting takes place (e.g., education) also changes for either or both sexes.

- Drawing from the statistics literature we define simple conditions that any index has to satisfy to provide a measure of change in sorting that is not distorted by changes in the marginal distributions of the characteristic.

- While most indices used in the literature satisfy these conditions, one of them—the likelihood ratio of Eika et al. (2019)—does not.

- We attribute their empirical result of declining homogamy at the top of the distribution to the violation of our conditions.

2. Measuring Assortative Matching and Changes in Assortativeness

Defining Assortative Matching

We start with a simple example. Consider an economy where an equal mass of men and women match by their level of education, which can only take two values - say, High School versus College. We will abstract from singles - everybody gets matched. The matching patterns in this population are summarized by the 2×2 matching Table (a, b, c, d) shown below. In the Table, $a + b$ and $a + c$ are the number of female and male college graduates respectively, while a is the number of couples where both spouses are college graduates. In what follows, we assume that $b \geq c$, i.e. that women are more educated than men; the alternative case obtains by switching b and c.

Table (a, b, c, d)

$w \backslash h$	C	HS
C	a	b
HS	c	d

We say that Table (a, b, c, d) exhibits Positive Assortative Matching (PAM) if *the number of couples with equal education* (the 'diagonal' of the Table) *is larger than what would obtain under random matching.* Under random matching the number of couples where both spouses are college graduates will simply be $\frac{(a+b)(a+c)}{a+b+c+d}$. Then we have PAM if and only if $a\,(a+b+c+d) \geq (a+b)\,(a+c)$ or equivalently if $ad \geq bc$; this also implies that more High School graduates marry each other than would be implied by random matching: $d\,(a+b+c+d) \geq (b+d)\,(c+d)$.

Defining “Increases in Assortative Matching”: Required Properties

Scale Invariance Any index I should not depend on the total size of the population, i.e.:

$$I(\lambda a, \lambda b, \lambda c, \lambda d) = I(a, b, c, d) \text{ for all } \lambda > 0.$$

Symmetry Any index I should treat the two categories (here, C and HS) identically; in other words, the index characterizes the level of assortativeness of the whole table. In practice, thus: $I(a, b, c, d) = I(d, c, b, a)$.

Monotonicity Consider the case where the two tables T and T' represent the same total population ($a + b + c + d = a' + b' + c' + d'$) with *identical marginal distributions* - i.e., the same proportion of educated men and women so that $a + b = a' + b'$ and $a + c = a' + c'$. In this case, Monotonicity requires that more assortativeness is equivalent to more people in each diagonal cell. So T is more assortative than T' if a larger fraction of individuals marry their own type in Table T than in Table T'. Formally:

$$a \geq a' \quad \text{or equivalently} \quad d \geq d'.$$

Perfect PAM We introduce two forms of this condition, weak and strong. The strong form relates to the polar case that obtains when all educated men marry educated women. This implies $c = 0$ and the sorting Table is $T = (a, b, 0, d)$. T is *perfectly (positive) assortative* in the sense that all educated individuals in the less educated population (men in our example) marry a spouse in the same education group as themselves; the only 'cross-marriages' result from the fact that educated women are more numerous than educated men.

The strong version of Perfect PAM states that T displays maximum degree of assortativeness, so no other Table can display strictly more assortativeness than T. So if a ranking of Tables by assortativeness is defined by an index, this index should reach its maximum value for T.

A weaker version of the condition applies to populations with an equal proportion of educated men and women. In that case, Perfect PAM obtains when $c = 0$ *and* $b = 0$, so all educated individuals marry their own. In statistical terms, the association between spouses' educations is then *absolute* in the sense of Kendall and Stuart (1961). Weak Perfect PAM therefore states that no Table can display strictly more assortativeness than Table $(a, 0, 0, d)$. Clearly, Perfect PAM implies the Weak version, while the converse is not true.

Our main claim is that *any acceptable measure of assortativeness must satisfy (at least) Scale Invariance, Symmetry, Monotonicity and Weak Perfect PAM*. In particular, it cannot be the case that a Table in which all individuals marry their own is found to be strictly less assortative than a Table that includes cross-marriages. As we shall see, imposing these conditions has bite, since some existing indices violate some of them.

Existing indices

Odds Ratio This is probably the most widely used index:

$$I_O(a, b, c, d) = \ln\left(\frac{ad}{bc}\right)$$

The odds ratio is popular in the demographic literature, as it can be directly derived from the log-linear approach (see for instance Mare (2001), Mare and Schwartz (2005), Bouchet-Valat (2014)); in economics, it was used by Siow (2015) ('local odds ratio') Chiappori et al. (2017), Ciscato and Weber (2020), Chiappori, Costa-Dias, Crossman and Meghir (2020) among many others.

Correlation A natural index is the correlation between wife's and husband's educations, each considered as a Bernoulli random variable taking the value C with probability $\frac{a+b}{a+b+c+d}$ (resp. $\frac{a+c}{a+b+c+d}$) and HS with probability $\frac{c+d}{a+b+c+d}$ (resp. $\frac{b+d}{a+b+c+d}$). This has been used in various contributions (for instance Greenwood et al., 2003, 2014), either explicitly or through a linear regression framework. Here:

$$I_{Corr}(a,b,c,d) = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

One can readily check that, in our case, the correlation index also coincides with Spearman's rank correlation, which exploits the natural ranking of education levels (C > HS). Equivalently, one can consider the χ^2 index, which is $\chi^2(a,b,c,d) = [I_{Corr}(a,b,c,d)]^2$

Minimum Distance In the minimum distance approach of Fernández and Rogerson (2001) and Abbott et al. (2019), one constructs the convex combination of two extreme cases (random and perfectly assortative) that minimizes the distance with the Table under consideration, and defines the weight of the perfectly assortative component as the index. Here, it is equal to:

$$I_{MD}(a,b,c,d) = \frac{ad - bc}{A}$$

where $A = (c+d)(a+c)$ if $b \geq c$ and $A = (b+d)(a+b)$ if $c \geq b$. This coincides in our context, with the 'perfect-random normalization' of Liu and Lu (2006) and Shen (2019).

Likelihood Ratio Eika et al. (2019) measure marital sorting between men of education level I and women of education level J *"as the observed probability that a husband with education level I is married to a wife with education level J, relative to the probability under random matching with respect to education"*. In practice, this leads to a specific index, namely the likelihood ratio:

$$I_L(a,b,c,d) = \frac{a(a+b+c+d)}{(a+b)(a+c)}$$

Properties of the indices of assortativeness

Result *Among indices just described:*

- *The odds ratio and the minimum distance index both satisfy Scale Invariance, Symmetry, Monotonicity and Perfect PAM (therefore Weak Perfect PAM).*
- *Correlation and Spearman rank correlation both satisfy Scale Invariance, Symmetry, Monotonicity and Weak Perfect PAM but not Perfect PAM.*
- *The Likelihood ratio satisfies Scale Invariance and Monotonicity; it does not satisfy Symmetry nor Weak Perfect PAM.*

We argue that the likelihood ratio is not an adequate measure because it violates two of the basic properties that an index of assortativeness should satisfy. It violates symmetry by singling out one category (here College); applied to the alternative category (High School), it would result in a different value for the index. It also violates Weak Perfect PAM as demonstrated by the following simple example. We compare Tables A and B corresponding, for instance, to two different cohorts in the same economy (population sizes normalized to 1):

Table $A = (.03, .07, .07, .83)$

$w \backslash h$	C	HS
C	.03	.07
HS	.07	.83

Table $B = (.5, 0, 0, .5)$

$w \backslash h$	C	HS
C	.5	0
HS	0	.5

The distribution of education is independent of gender in both A and B, but the number of educated people has increased from 10% to 50% between A an B. Cohort A exhibits PAM in the usual sense (more people on the diagonal than would obtain under random matching); yet, 70% of educated people marry an uneducated spouse. Cohort B displays perfect PAM, with all college educated individuals marrying together. Consistently, all indices discussed above satisfying Weak Perfect PAM conclude that B displays more assortativeness than A. However, the likelihood ratio yields $I_L(A) = 3$ and $I_L(B) = 2$, suggesting that assortativeness has *decreased* from A to B - a conclusion that is intuitively quite difficult to accept.

Marginal Independence

To explain this paradox, it is interesting to refer to an older statistical literature that discusses the properties of measures of association in the case of paired attributes (i.e., in our case, husband's and wife's education). The Marginal Independence requirement posed by Edwards (1963) states that the association should not be 'influenced by the relative sizes of the marginal totals' (p. 110). That is, the measure should not change if one starts from a Table $T(a, b, c, d)$ and doubles the number of couples where the man is educated (while keeping unchanged the ratio of educated versus uneducated wives). Formally, for any non negative (a, b, c, d) and any positive λ, it should hold that:

$$I(\lambda a, \lambda b, c, d) = I(\lambda a, b, \lambda c, d) = I(a, \lambda b, c, \lambda d) = I(a, b, \lambda c, \lambda d) = I(a, b, c, d)$$

Among the indices just reviewed, only one - the odds ratio - satisfies Edwards's marginal invariance. It is interesting to consider *how* the other indices violate this requirement. Consider Table $T_\lambda = (\lambda a, \lambda b, c, d)$ with $ad > bc$ and $\lambda \geq 1$. Suppose λ increases. Then:

- The minimum distance index *increases* since $\partial I_{MD}/\partial\lambda > 0$;
- The likelihood ratio *decreases* since $\partial I_{LR}/\partial\lambda < 0$;
- The correlation and Spearman correlation may increase or decrease depending on parameters.

Structural interpretations

Finally, it is important to note that among the various indices, the odds ratio has the additional advantage of having a known structural interpretation. Specifically, assume that the observed matching behavior constitutes the stable equilibrium of a frictionless matching model under transferable utility. Assume, furthermore, that the surplus generated by a match between woman i belonging to category I and man j belonging to category J takes the separable form:

$$s(i,j) = Z^{IJ} + \alpha_i^J + \beta_j^I$$

where Z is a deterministic component depending only on individual educations and the α, β are random shocks reflecting unobserved heterogeneity among individuals.

3. Empirical Example: Educational Homogamy among Educated People in the US

Table 1: Distribution of education among married men and women – birth cohorts 1950-59 and 1970-75

	High School and below	Some College	4+ years College degree
	Birth cohort 1950-59		
Men	43.1	25.6	31.3
Women	46.8	27.2	26.0
	Birth cohort 1970-75		
Men	37.0	25.3	37.7
Women	30.7	27.0	42.3

Data source: March extract of the US Current Population Survey, subsample of married individuals observed when aged 35-44 and born in 1950-59 or 1970-75.

That increase in education is illustrated in Table 1. The figures in the Table are estimated on the same data from the March extract of the US Current Population Survey that is used in Eika et al. (2019). Here we consider two birth cohorts, 1950-59 and 1970-75. The educational choices of these two cohorts were taken under very different scenarios about the college premium, with the later cohort facing a much higher market return to college education than the earlier one. Consistently with that change, the Table shows that the number of college educated men and women increased between the two cohorts and that the change was especially marked among women, who became more educated than men. Moreover, the increasing concentration of married individuals among college graduates comes entirely at the expense of fewer individuals in the bottom education group.

Theory predicts that preferences for homogamy should increase as a result of the increase in the college premium (Chiappori et al., 2017). Estimates in Table 2 show how the various indices describe changes in PAM at the top of the education distribution across the two cohorts. They refer to three different 2×2 tables, comparing College graduates with the education group just below (column 1) and with everyone who did not graduate from College (column 2), as well as those who attended College with those who did not (column 3).

Table 2: Marital assortativeness at the top of the distribution of education – comparing birth cohorts 1950-59 and 1970-75

	College vs Some College	College vs less than College	At least Some College vs High School and below
	Panel A: Odds ratio		
diff. across cohorts	0.214	0.061	0.293
adjusted p-value	0.000	0.044	0.000
	Panel B: χ^2		
diff. across cohorts	0.029	0.040	0.032
adjusted p-value	0.000	0.000	0.000
	Panel C: Minimum Distance		
diff. across cohorts	0.001	0.026	0.071
adjusted p-value	0.888	0.000	0.000
	Panel D: Likelihood ratio		
diff. across cohorts	-0.046	-0.421	-0.125
adjusted p-value	0.000	0.000	0.000

Notes: Columns identify each of the 2×2 sorting matrices. In each panel, row 1 shows estimates of the difference in the respective index between the latest and earliest cohorts; row 2 shows p-values for 2-sided significance testing adjusted for multiple hypothesis using the stepdown method for the three outcomes on the row. (Romano and Wolf (2005), Romano et al. (2008), Romano and Wolf (2016)). *Data source*: March extract of the US Current Population Survey, subsample of married individuals observed when aged 35-44 and born in 1950-59 or 1970-75.

We see that, in all cases, the odds ratio, the χ^2 and the minimum distance index, which all satisfy Symmetry and Weak Perfect PAM, conclude that assortativeness significantly increased between the two cohorts. On the contrary, the likelihood ratio shows a significant reduction.

4. Concluding Remarks

- It is relatively simple to estimate whether there is positive assortative matching in a stochastic marriage market along the dimensions of a characteristic such as education.
- However, measuring the extent to which such assortative matching differs between two economies or between two points in time for the same economy is challenging when the marginal distribution of the characteristics also change.
- Drawing from the statistics literature we define simple conditions that any index should satisfy to provide a measure of change in sorting that is not distorted by changes in the marginal distributions of the characteristic.
- We show that most frequently used indices satisfy these conditions but that the likelihood criterion of Eika et al. (2019) does not.
- This difference in properties may underlie the contrasting conclusions about the change in homogamy over the recent decades.