A Note: The Effect of Assortative Mating on Income Inequality

Karl Harmenberg*

November 24, 2014

Minimally revised: June 13, 2017

1 Introduction

The effect of assortative mating on income inequality has recently been the subject of some debate. This note was originally written to contest the claim by Greenwood et al. (2014) that marital sorting was a key contributor to the increase in household income inequality in the US from 1960 to 2005. The claim was reported by many news outlets[1] and made a big impact in the research community.[2] In this note, I tried to replicate the results of Greenwood et al. (2014) using the exact same data. Instead of finding large effects of marital sorting on household income inequality, I found small effects in line with the results of other concurrent studies, Eika et al. (2014) and Hryshko et al. (2014).

A year after the originally published result, a corrigendum was published by the authors (Greenwood et al. (2015)) and a scientific consensus (about the particular claim of Greenwood et al. (2014)) has been reached. Since this note has been circulated and cited in the scientific community, I am keeping it on my webpage.

Greenwood et al. (2014), this note, and the related literature, all ask the question: What would happen to household income inequality if spouses matched randomly instead of the current matching? Greenwood et al. (2014) argue that randomizing household formation would generate a large drop in Gini inequality (from 0.43 to 0.34) in the US (ACS data, 2005). Eika et al. (2014) report that randomizing household formation generates a drop from 0.403 to 0.384 for the US (CPS, 2007) and a drop from 0.244 to 0.235 for Norway (Statistics Norway, 2007). Hryshko et al. (2014) find that randomizing household formation generates a drop from 0.295 to 0.290 (US data, 2004/05 SIPP-SSA) and from 0.273 to 0.264 (US data, 2004/05 PSID). To summarize, Greenwood et al. (2014) find a large effect (a drop in Gini inequality by 0.09) of randomizing

*Institute for International Economic Studies, Stockholm University. The author would like to thank the IIES Macro Group, Philippe Aghion and Markus Jäntti for helpful suggestions and comments.

[1] For example, *Sex, Brains and Inequality* (The Economist, February 8th, 2014), *How When Harry Met Sally Explains Inequality* (The Atlantic, February 3rd, 2014), and *The Marriages of Power Couples Reinforce Income Inequality* (New York Times, December 24th, 2015)

[2] Greenwood et al. (2014) have 119 citations as of June 13, 2017 on Google Scholar.

household formation while Eika et al. (2014) and Hryshko et al. (2014) find a small effect (drops of magnitude around 0.01).

In this note, I employ two different methodologies. The first method, addition randomization, is similar to the approach of Hryshko et al. (2014), and gives similar results to theirs (randomizing matches generate a drop in Gini smaller than 0.01). The second method, imputation randomization, is similar to the approaches of Eika et al. (2014) and Greenwood et al. (2014). With this approach, I find a non-negligible effect of assortative mating on household inequality. However, the effect is much smaller than reported by Greenwood et al. (2014) (a drop in Gini of $0.01 - 0.02$ rather than 0.09), in line with the results presented by Eika et al. (2014).

1.1 Two distinct approaches to randomization

There are two different approaches to randomizing, and they differ in what they keep constant when randomizing. Either individual incomes are kept constant or household incomes are kept constant. I call the first approach *addition* and the second approach *imputation*. Both approaches randomize household formation but differ in how they assign household income to a randomly formed household (which I call a pseudo-household). The addition approach computes pseudo-household income by summing individual incomes. The imputation approach imputes pseudo-household income by assuming that it follows the same distribution as the distribution of income for (actual) households with the same characteristics.

Let x_i, x_j be individual characteristics (age, education, ...) of men and women respectively. Let y_i, y_j be individual incomes of men and women respectively. Household income of a household (i, j), y_{ij} is naturally defined as $y_{ij} := y_i + y_j$. Let $dF(z)$ denote the (true) distribution of a given variable z.

The addition approach, keeping attributes x_i, x_j in couples constant, amounts to computing the distribution

$$(dF(y_i|x_i) * dF(y_j|x_j))dF(x_i, x_j).$$

where $*$ denotes the convolution operator. That is, it is assumed that agents with characteristics i (j) are randomly assigned a partner, so that $y_i|x_i$ and $y_j|x_j$ are independent. Under unconditional randomization the attributes kept constant, x_i, x_j, are empty, so this reduces to

$$dF(y_i) * dF(y_j).$$

The imputation approach, using observable characteristics x_i, x_j, instead computes

$$dF(y_{ij}|x_i, x_j)dF(x_i)dF(x_j).$$

Table 1: *Marginal distributions: True distribution, under addition randomization and imputation randomization.*

Actual distribution	$dF(y_{ij}\|x_i, x_j)dF(x_i, x_j)$
Addition randomization	$(dF(y_i\|x_i) * dF(y_j\|x_j))dF(x_i, x_j)$
Imputation randomization	$dF(y_{ij}\|x_i, x_j)dF(x_i)dF(x_j)$

Under this scheme, household income is imputed from the observable characteristics of the household.

These two approaches to randomization are conceptually distinct, and it makes little sense in general to compare results across the approaches.

The benefit of addition randomization is its directness: It takes the existing population and randomizes it. There is no loss in information. The drawback is that the method takes labor supply and income as exogenous to household formation.

The benefit of imputation randomization is that it does take into account that labor supply and income are endogenous to household formation. The drawback with the methodology in practice is that the attributes available give less than perfect imputation. There are reasons to be worried that e.g., young men with no high school degree who are married to old women with more than a college degree are systematically different from young men with no high school degree married to young women with no high school degree.

1.2 Methodological comparison to the literature

Greenwood et al. (2014) and Eika et al. (2014) both use (in my terminology) imputation methods. Greenwood et al. (2014) approximate the income distribution by deciles, observe the income distribution for each combination of educational attainment (for both spouses), and computes the counterfactual random distribution of educational attainment for couples. Using the counterfactual educational attainment as weights, they sum the income distributions (approximated by deciles) over educational attainment and get a new counterfactual household income distribution. Eika et al. (2014) use the semiparametric decomposition approach proposed by DiNardo et al. (1996), which in this context amounts to an imputation method.

Hryshko et al. (2014) use (in my terminology) an addition method. They construct randomized couples and compute the pseudo-couples' income as the sum of the two individual incomes.

2 Empirical analysis

2.1 Data

The data and the restrictions on the sample are identical to the data and restrictions of Greenwood et al. (2014). For the analysis, I use data publicly available at the Integrated Public Use Microdata Series (IPUMS)

webpage. For 1960, 1970, 1980, 1990 and 2000, I use the one percent sample of the US Census. For 2005, I use the American Community Survey (ACS).

The population of households is restricted to households with singles or married couples. Only households with the adults and their own children (younger than 19) are considered. Families with e.g. grandparents, aunts, uncles or friends living in the household are not considered. Widows, widowers and individuals with their spouse missing are not included, but separated individuals are. Income variables are restricted to be non-negative.

2.2 Methodology

The methodology will be micro oriented. We will consider two randomization approaches, addition randomization and imputation randomization.

The first approach, addition randomization, takes individual income as given and randomizes matches (pseudo-couples) on the micro level. The income of a randomized pseudo-couple is the sum of the two individual incomes. This approach has the advantage that it is transparent, but the drawback that it treats labor supply as strictly exogenous. I do four variations of this approach.

1. The first randomizes fully as described above.

2. The second approach keeps the age distribution within the couple and single population constant, by randomizing each age cohort separately. For example, a real couple with a 28-year-old man and a 44-year-old woman is replaced by a randomized couple consisting of a 28-year-old man and a 44-year-old woman, and a 51-year-old single woman is replaced by a 51-year-old woman (who may be single or married).

3. The third approach keeps the marital status of each individual fixed, so that the pseudo-couples are formed by married individuals.

4. The fourth approach keeps the marital status and family size of each individual fixed, thereby allowing for a reasonable interpretation of household income adjusted for family size.

In all four variations, the Gini coefficient of labor income, total income, adjusted (OECD equivalence scale) labor income and total income are computed. Adjusted income is calculated using the family size of the pseudo-household. For example, a couple with seven children is replaced by two individuals and it is assumed that these individuals have to support the seven children. When randomizing conditional on marital status and family size, this become a non-issue since the family sizes of the original family, the man's real family and the woman's real family are all equal.

The second approach, imputation randomization, also randomizes matches on the micro level but recognizes that individual income is endogenous. Instead of computing the pseudo-household's income as the sum

of the individual incomes, the pseudo-household's income is imputed by the characteristics of the pseudo-household. Concretely, the pseudo-household is characterized by the educational levels of the man and woman, as well as the age groups (25-29, 30-39, 40-49, 50-54) of the man and woman. The income of the pseudo-household is imputed by randomly drawing a real household with the same characteristics, and using its income.[3] For incomes adjusted by household size, the adjusted income is drawn directly and the family size is never explicitly used in the algorithm. The Gini coefficients of labor income, total income, adjusted (OECD equivalence scale) labor income and total income are computed.

2.3 Results

The results of the two experiments for 2005 and 1960 can be seen in Table 2 and Table 3. An immediate observation is that under any specification, the effect of randomization on inequality is small, in contrast with the sizeable effects reported by Greenwood et al. (2014). For 2005, no change in Gini is larger than 0.015.

Keeping individual incomes fixed, randomizing reduces inequality in the 2005 sample except if conditioned on marital status. Randomizing among married couples forms more dual-earner and no-earner households than in the real data, it is therefore not that strange that randomizing increases inequality when conditioned on marital status.

For 1960 the general trend is reversed, addition randomization actually increases inequality. Given the low level of female labor force participation and the negative correlation between male income and female labor force participation, this is not surprising.

Under imputed household incomes, randomization does lessen inequality. The fall in Gini, for different income measures, is about 0.01-0.02 for 2005, while substantially smaller (less than 0.005) for 1960.

Table 2: *Effects of randomization, keeping individual incomes fixed. The Gini coefficient for the data, under full randomization, conditional on age, conditional on marital status and conditional on marital status and family size.*

Gini 2005	Data	Full	Age	Mar.	Mar. and fam.
Labor income	0.4900	0.4823	0.4838	0.4902	0.4855
Total income	0.4605	0.4568	0.4592	0.4653	0.4596
Labor income adjusted	0.4736	(0.4792)	(0.4783)	(0.4745)	0.4686
Total income adjusted	0.4401	(0.4506)	(0.4510)	(0.4451)	0.4383
Gini 1960	Data	Full	Age	Mar.	Mar. and fam.
Labor income	0.3976	0.4124	0.4109	0.4037	0.3972
Total income	0.3352	0.3599	0.3591	0.3497	0.3423
Labor income adjusted	0.4195	(0.4372)	(0.4361)	(0.4210)	0.4178
Total income adjusted	0.3570	(0.3852)	(0.3849)	(0.3656)	0.3631

[3]For some pseudo-households, there is no actual couple with those characteristics in the sample. These pseudo-households, amounting to less than 1 percent of the pseudo-sample, are dropped.

Table 3: *Effects of randomization, imputing household incomes. The Gini coefficient for the data and under full randomization.*

2005	Data	Imputed
Labor income	0.4900	0.4738
Total income	0.4605	0.4410
Labor income adjusted	0.4736	0.4642
Total income adjusted	0.4401	0.4276
1960	**Data**	**Imputed**
Labor income	0.3976	0.3938
Total income	0.3352	0.3306
Labor income adjusted	0.4195	0.4167
Total income adjusted	0.3570	0.3543

3 Discussion

Assortative mating contributes to Gini measured inequality. However, the magnitude of the contribution appears to be much smaller than reported by Greenwood et al. In this paper, we find under various specifications no effect larger than a change in Gini of 0.015, in comparison with Greenwood et al's change in Gini of 0.09.

Under imputation randomization, the effect of assortative mating is for US 2005 data $0.01 - 0.02$ (depending on the measure of income). This result is in line with the findings by Eika et al. (2014) who find a drop in Gini of 0.02 for US 2007 data with similar methodology.

Under addition randomization, the effect of assortative mating is smaller, since endogenous labor supply dampens the effect of assortative mating. Keeping marital status constant, randomizing actually increases measured inequality. Under all other specifications, income inequality is lessened by randomization, but the effect is less than 0.01 under all specifications. This is in line with the results of Hryshko et al. (2014), who find an effect of individual randomization of 0.05 and 0.09 for different data sources for 2004/05 US data.

In conclusion: There is an effect of assortative mating on income inequality, but this study indicates that the effect is much smaller than argued by Greenwood et al. (2014). This is in line with the results of Eika et al. (2014) and Hryshko et al. (2014).

References

DiNardo, J., Fortin, N. M., and Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica*, 64(5):1001–1044.

Eika, L., Mogstad, M., and Zafar, B. (2014). Educational Assortative Mating and Household Income Inequality.

Greenwood, J., Guner, N., Kocharkov, G., and Santos, C. (2014). Marry Your Like: Assortative Mating and Income Inequality. *American Economic Review: Papers and Proceedings*, 104(5):3–5.

Greenwood, J., Guner, N., Kocharkov, G., and Santos, C. (2015). Corrigendum to Marry your like: Assortative mating and income inequality.

Hryshko, D., Chinhui, J., and Mccue, K. (2014). Trends in Earnings Inequality and Earnings Instability among U. S. Couples: How Important is Assortative Matching?