Matching individuals in the Current Population Survey

Stuart Craig[1] with Jacob S. Hacker[1], Gregory A. Huber[1], Austin Nichols[2], Philipp Rehm[3], and Mark J. Schlesinger[1]

[1]Yale University [2]Urban Institute [3]Ohio State University

Stata Conference, 2012

Outline

Outline

The Economic Security Index

- $ESI = \sum w_i L_i / \sum w_i$
- where $L_{it} = \left(\frac{y_{it} - M_{it} - D_{it}}{e_{it}} < \left(\frac{3}{4}\right) \frac{y_{it-1} - M_{it-1} - D_{it-1}}{e_{it-1}} \right) (W_{it} < W^*_{it})(1 - R_{it})$
- A comprehensive measure of economic risk based on the realized losses of household resources.
- Accounts for:
  - Income (adjusted for family size)
  - Out of pocket medical expenses
  - Liquid financial resources (wealth and debt)

The Economic Security Index

- $ESI = \sum w_i L_i / \sum w_i$
- where $L_{it} = \left(\frac{y_{it} - M_{it} - D_{it}}{e_{it}} < \left(\frac{3}{4}\right) \frac{y_{it-1} - M_{it-1} - D_{it-1}}{e_{it-1}} \right) (W_{it} < W^*_{it})(1 - R_{it})$
- A comprehensive measure of economic risk based on the realized losses of household resources.
- Accounts for:
  - Income (adjusted for family size)
  - Out of pocket medical expenses
  - Liquid financial resources (wealth and debt)

Data limitations and use of multiple surveys

- No survey captures all of these
- Closest thing we had at the beginning was the SIPP which provided:
  - Short mini-panels
  - Income
  - Medical expenditure data*
  - Wealth/debt data
- Medical expenditure data in the SIPP was not continuous so we used a model based imputation
- For more information on construction of the index, see (Hacker et al., 2011)

Data limitations and use of multiple surveys

- No survey captures all of these
- Closest thing we had at the beginning was the SIPP which provided:
  - Short mini-panels
  - Income
  - Medical expenditure data*
  - Wealth/debt data
- Medical expenditure data in the SIPP was not continuous so we used a model based imputation
- For more information on construction of the index, see (Hacker et al., 2011)

Data limitations and use of multiple surveys

- No survey captures all of these
- Closest thing we had at the beginning was the SIPP which provided:
  - Short mini-panels
  - Income
  - Medical expenditure data*
  - Wealth/debt data
- Medical expenditure data in the SIPP was not continuous so we used a model based imputation
- For more information on construction of the index, see (Hacker et al., 2011)

Transition to the CPS

- Big attrition in the SIPP
- Break between 2004 and 2008 panels coincided with the Great Recession
- SIPP waves and years did not line up

Pros:

- Attrition is at least relatively consistent in the CPS
- Reference period in the March Supplement is the preceding calendar year
- Available for (nearly) every year and extending earlier than the 1980's
- CPS designed to produce geographic estimates

Cons:

- No medical spending or wealth data
- *Only two year panels*

Transition to the CPS

- Big attrition in the SIPP
- Break between 2004 and 2008 panels coincided with the Great Recession
- SIPP waves and years did not line up

Pros:

- Attrition is at least relatively consistent in the CPS
- Reference period in the March Supplement is the preceding calendar year
- Available for (nearly) every year and extending earlier than the 1980's
- CPS designed to produce geographic estimates

Cons:

- No medical spending or wealth data
- *Only two year panels*

Outline

Census Bureau guidance

Years: 1968-1971
Variables: Random Cluster Code (F6-10) and Serial Number (F11-14)

Years: 1971-1972
Changes in CPS clustering procedures and the accompanying change of household identification numbers prevent matching 1971 and 1972 March CPS files.

Years: 1972-1973
The 1972 file uses 1960 random cluster codes while the 1973 file uses 1970 random cluster codes, thus precluding the matching of records.

Years: 1973-1975
Variables: Random Cluster Code (F7-11). Segment Number (F12-16), and Serial Number (F217-218)

Years: 1975-1976
Variables: 1975: Random Cluster Code (F7-11) Segment Number (F12-16), and Serial Number (F217-218)
1976: Random Cluster Code (H35-39), Segment Number (H40-43), and Serial Number (H44-45)

Years: 1976-1977
Matching is not possible because variables required for matching are in a different format each year.

Years: 1977-1985
Variable: Household Identification Number (H18-29)

Years: 1985-1986
Matching is not possible because the 1986 file is based entirely on the 1980 census design sample.

Years: 1986-1993
Variable: Household Identification Number (H18-29)

Years: 1994-1995
(See CPS, March 1995 User Note 1.)

Years: 1995-1996
Matching is not possible because the March 1996 file is based entirely on the 1990 Census design sample.

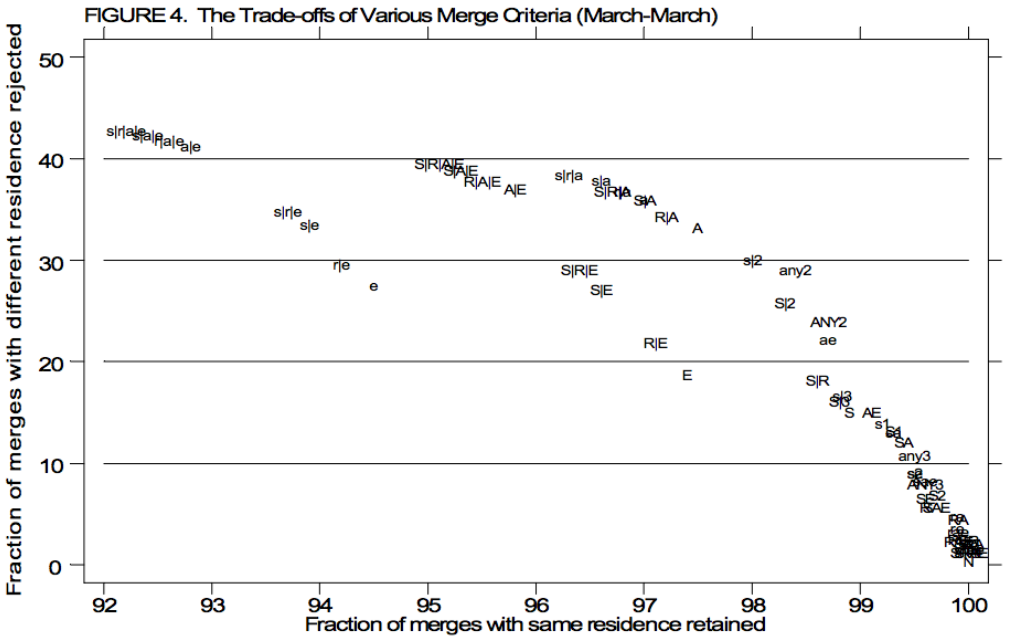
Years: 1996-2010
Variable: Household Identification Number (H344-358)

Need for a matching algorithm

- Household identifiers are helpful, but the survey is one of geographic residences (no effort to follow respondents)
- Especially in early years, there was little effort to keep flag changes in occupants
- There is a migration flag, but that too is error prone
- Introduction of non-rotation group individuals in the March Supplement starting in early 2000's

Madrian and Lefgren

- Often cited - widely used as the "model"
- Generates potential matches based on identifiers - verifies using characteristics
- Code hosted at NBER
- Only provides guidance on a limited set of years (1980-1998)
- Hightlights tradeoff between matches and mismatches



Others

- Welch (1993)***—emphasized importance of selecting match criteria based on parameters to be measured
  - You would not want to use relationship to household head as a validating variable if changes in family structure are the object of interest
- Feng (2001) and (2008) - Probabilistic matching and observation that household IDs did not uniquely identify households

Outline

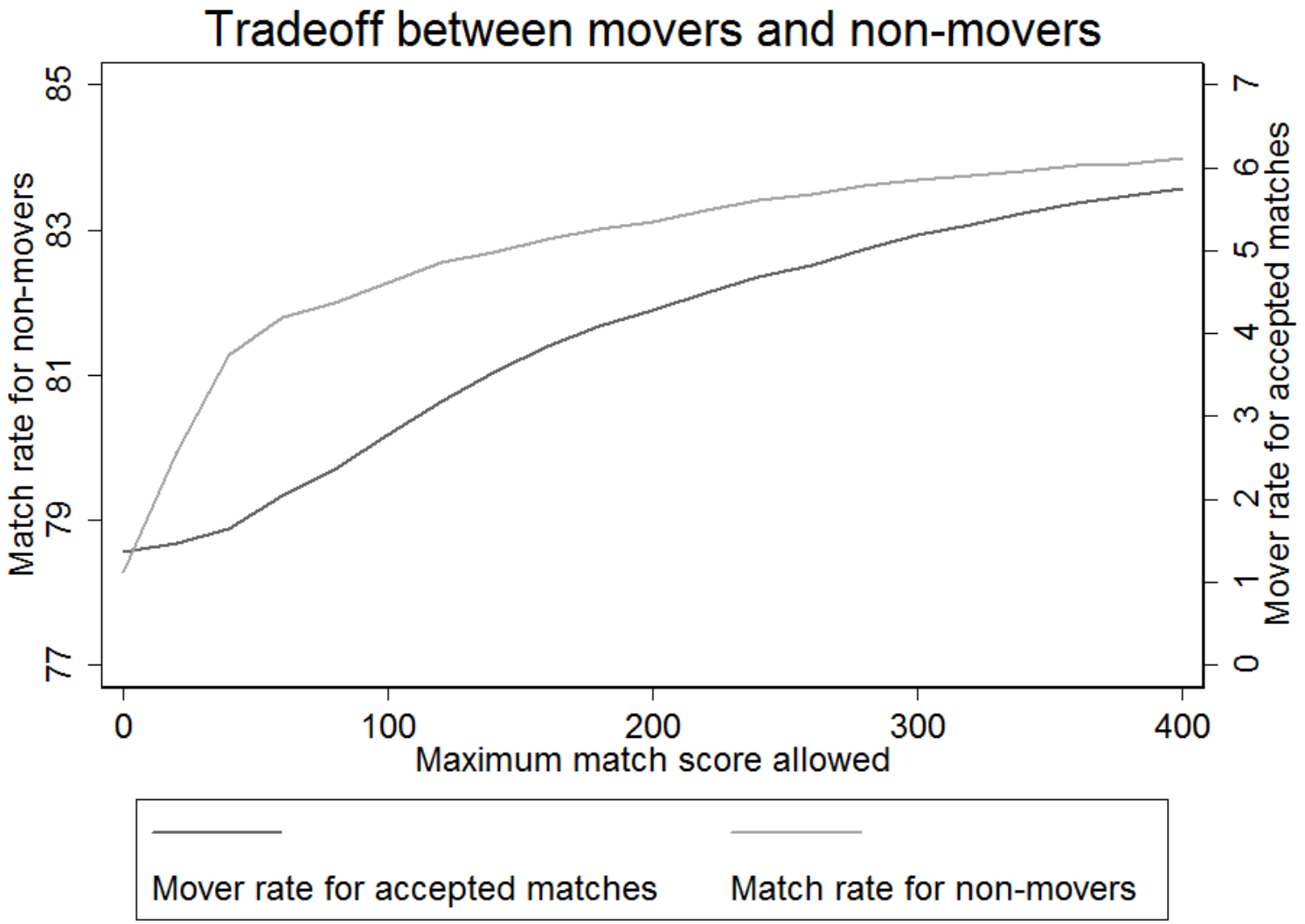
Goals

- Maximize potential matches
- Minimize any bias created by the matching process
- As continuous a series as possible (minimize missing years)
- Handle the differing demands of changes to the CPS

The algorithm

1. Create all pairwise combinations within household IDs
2. Generate a match score based on weighted set of characteristics (increasing with difference)
3. Exclude those pairs with unacceptable match scores
4. Match individuals who minimize each other's distances (both directions)
5. Stipulate a minimum which at least one person in the household must meet (0 or 10)
6. Small number of ties (usually <10) are dropped as duplicates in one year or new residents

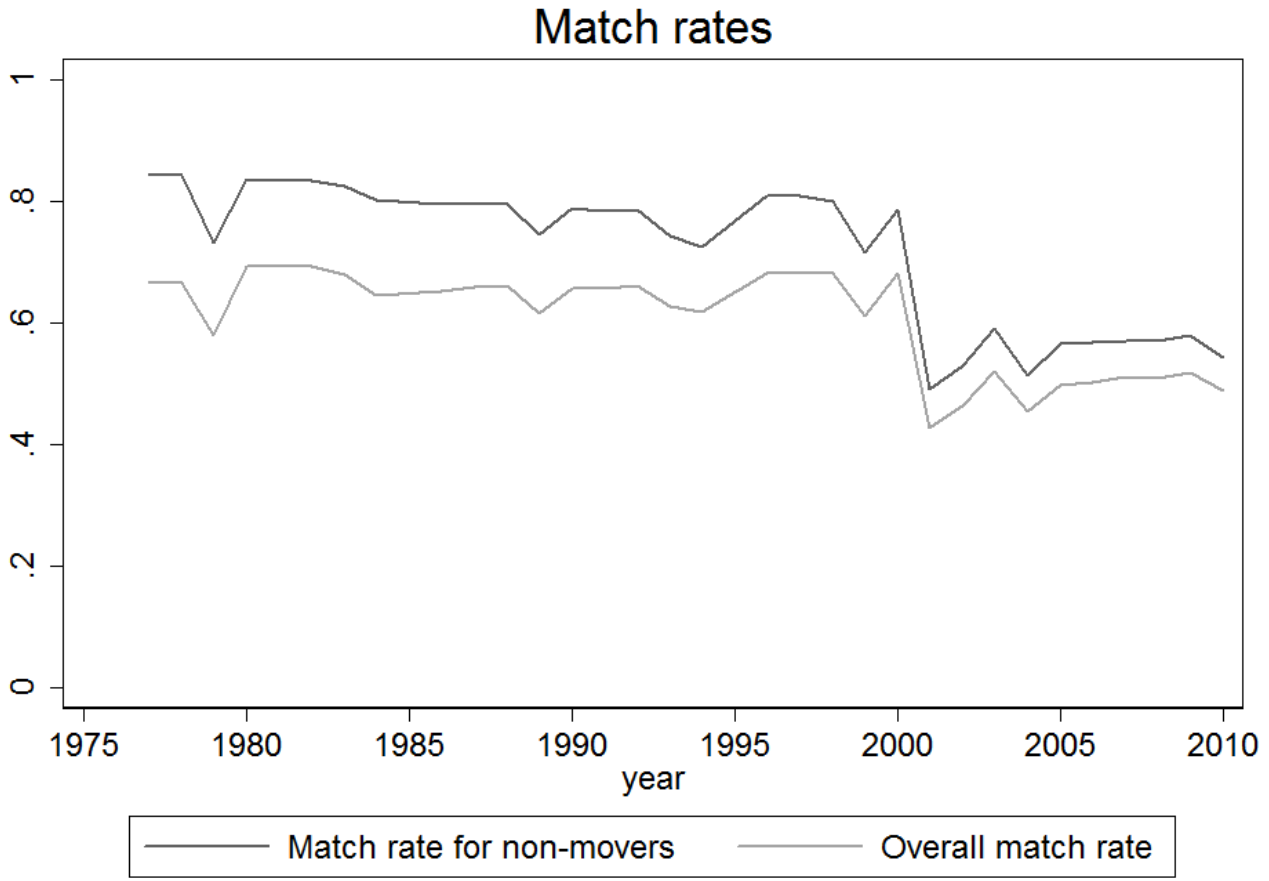
Setting the maximum match score



Advantages to this approach

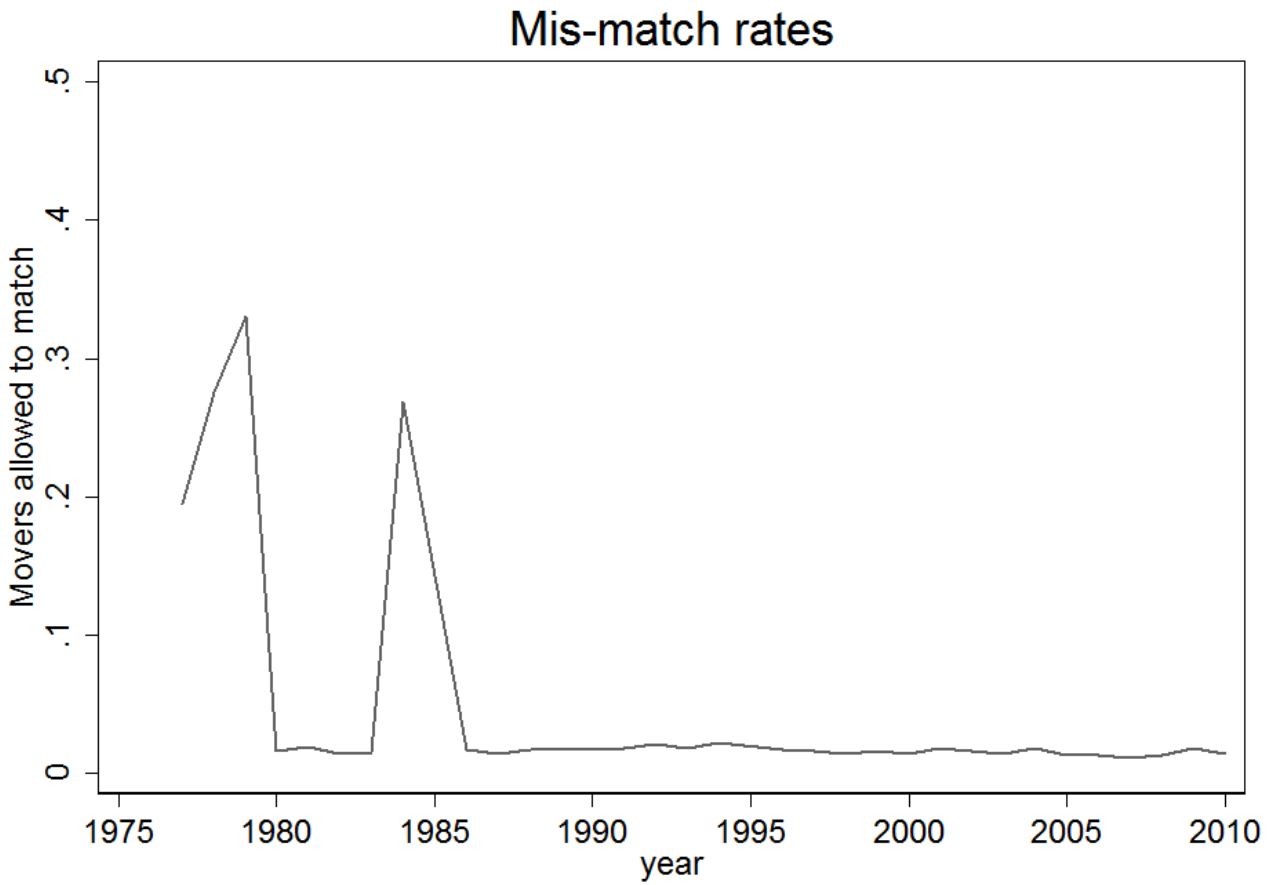
- Because we do not require that line numbers match exactly, we can match individuals even in years for which line numbers are absent.
- Use of distance matching provides an elegant solution to the problem of non-uniue household IDs
- Any bias introduced by this method is at least applied to the entire series
- This method produces consistent match and mis-match rates.

Analyzing the performance



Note: Decline in match rates in early 2000's are a result of SCHIP and "rotation group 9" oversample (See Feng, 2008).

Analyzing the performance pt. 2



Note: Mover flags in late 70's refer to migration since 1975 and 1985 flag refers to migration since 1980.

Outline

Longitudinal weights

- As per Nichols (2007), we reweight the matched group to the full year-2 sample using propensity scores
- Not usually discussed in the volatility literature - Hertz 2007 reweights, but only to adjust for dropping imputations
- Two stage process
  - Generate probabilities of match based on
  - adjust resulting weights to match the proportions of full sample by age, race, and sex

Further work and Wrap up

- I hope to create a flexible set of programs to allow users to adopt this approach of creating matches in ways that are sensitive to their needs
- Match Outgoing Rotation Groups of the monthly CPS for more timely and frequent estimates

Contact: stuart.craig@yale.edu
Questions?