June 6-9, 2012. Lausanne Conference on Sequence Analysis

Attrition & counterfactuals:

New applications for sequence analysis?

Matissa Hollister . Sociology Department . Dartmouth College

Attrition & counterfactuals

- Counterfactuals
  - Would like to know what values would be if person had not been exposed to a "treatment" (job training, unemployment, neighborhood effects, etc)
  - Typical tools:
    - Synthetic controls by matching on observed variables before the treatment
      - Propensity score matching
      - Mahalanobis distance
- Attrition from panel data
  - Would like to know what values would be if the surveyors had been able to maintain contact
    - "treatment" is attrition
  - Typical tools:
    - Weighting
    - Multiple imputation
- All methods assume that selection to the "treatment" is random after controlling for observables
  - May be particularly problematic for attrition

How can SA help?

- Underlying belief: Sequence as a whole captures more than its individual parts
  - Including "unobserved" factors behind attrition/selection?
- Career types
  - Different career paths have different employment practices
- Using SA as a similarity measure (no clustering)
- OM distances based upon sequences before attrition→ identify similar individuals remaining in the sample → synthetic counterfactual
- Challenges in implementation, but not necessarily more than other methods

An example/real world simulation

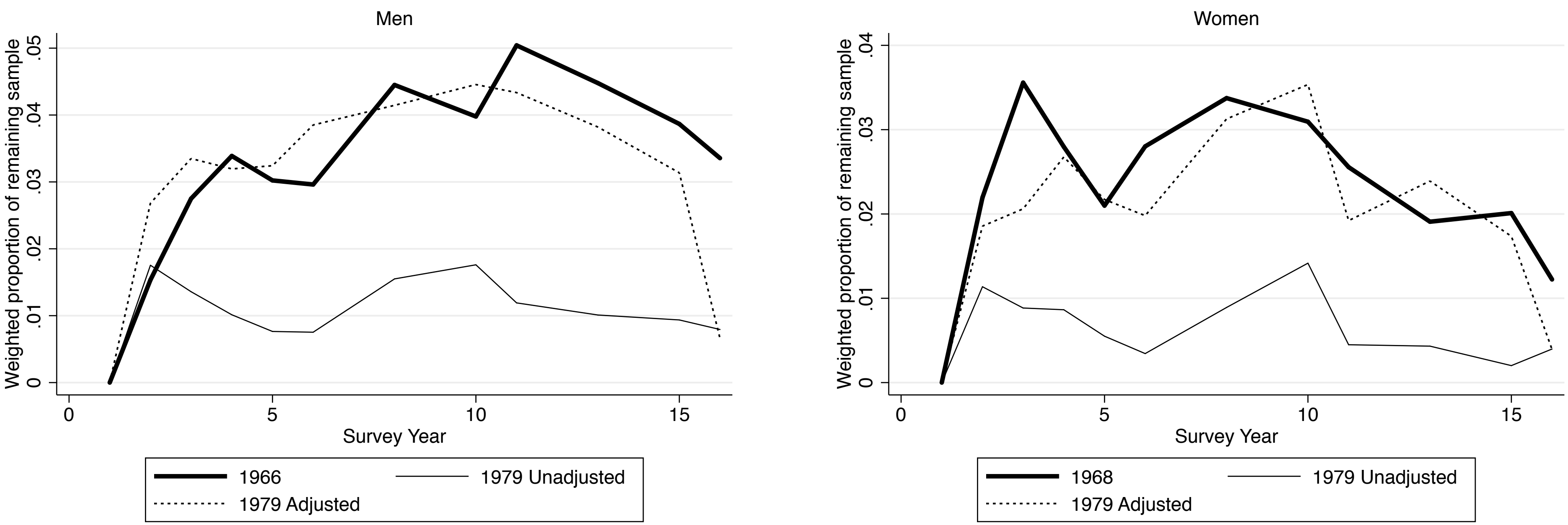
- National Longitudinal Surveys: U.S. survey of young men and women as they transition into the workplace. Starts at age 14-22
  - NLSY79: National Longitudinal Survey of Youth 1979
  - Original cohorts: 1966 Young Men, 1968 Young Women
- Original cohorts had much higher attrition rates
  - Original cohorts: 32% (men) & 31% (women) lost by 16th year
  - NLSY79: 14% (men) & 12% (women)

Differences in survey procedures

- Causes of higher attrition rate in original cohorts
  - Fewer resources to find difficult cases

- Simulate attrition in NLSY79
  - Number of attempts required to contact
    - >20 attempts→ unlikely would have been surveyed under original cohort conditions

Unable to contact



Differences in survey procedures

- Causes of higher attrition rate in original cohorts
  - Fewer resources to find difficult cases
  - Dropped:
    - Anyone who refused a survey
    - Anyone who missed two surveys in a row
- Simulate attrition in NLSY79
  - Number of attempts required to contact
    - >20 attempts→ unlikely would have been surveyed under original cohort conditions
  - Apply rules on refusals & two-in-a-row
  - Results in similar rates of attrition over time

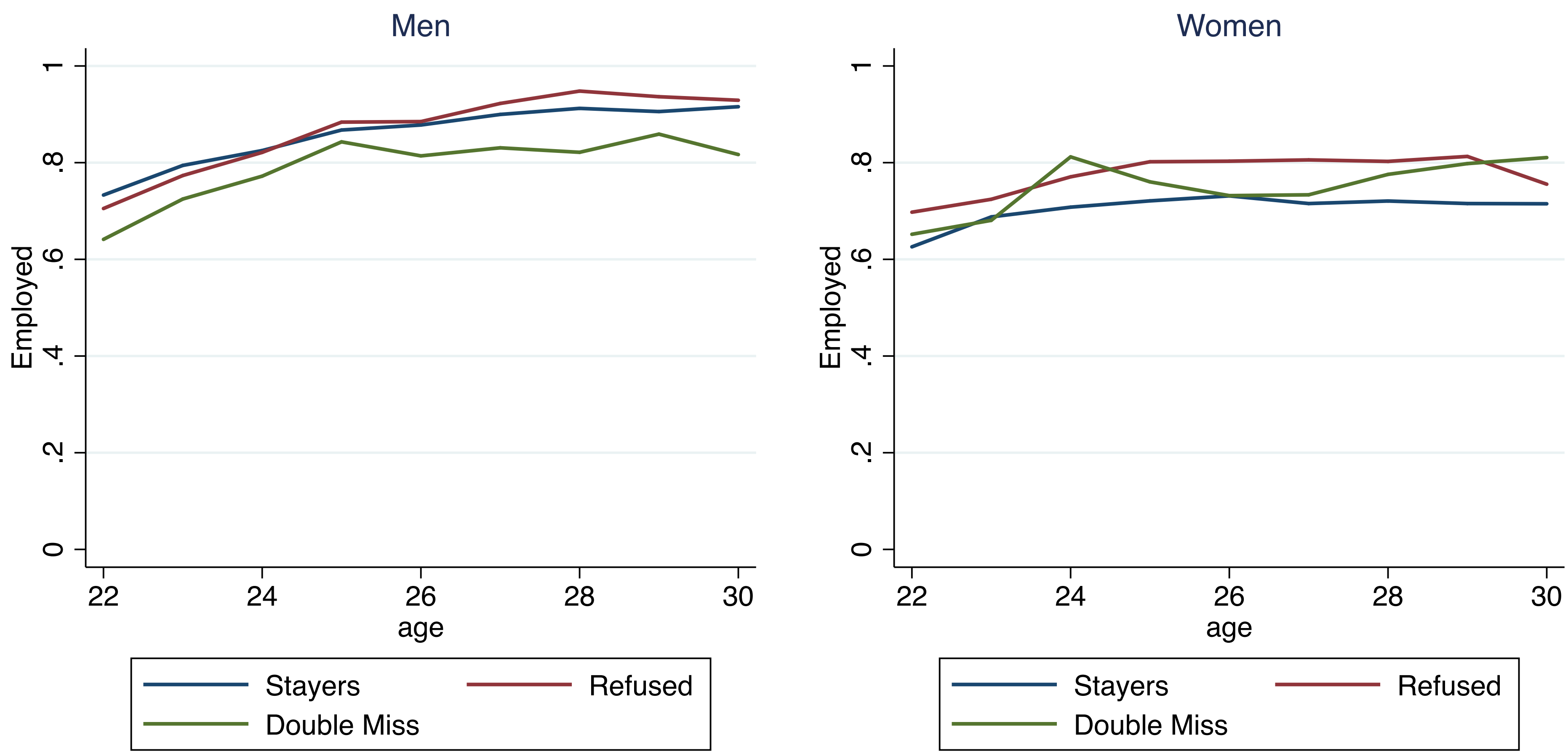
Simulation data

- NLSY79 work histories ages 22-30
- Sample:
  - Remained in the **NLSY79** until age 30
  - at least one observation age 22+ under original cohort rules
- Treatment vs control
  - Treatment: attrition before age 30 **under original cohort rules**
  - Pool of potential controls: remained through age 30

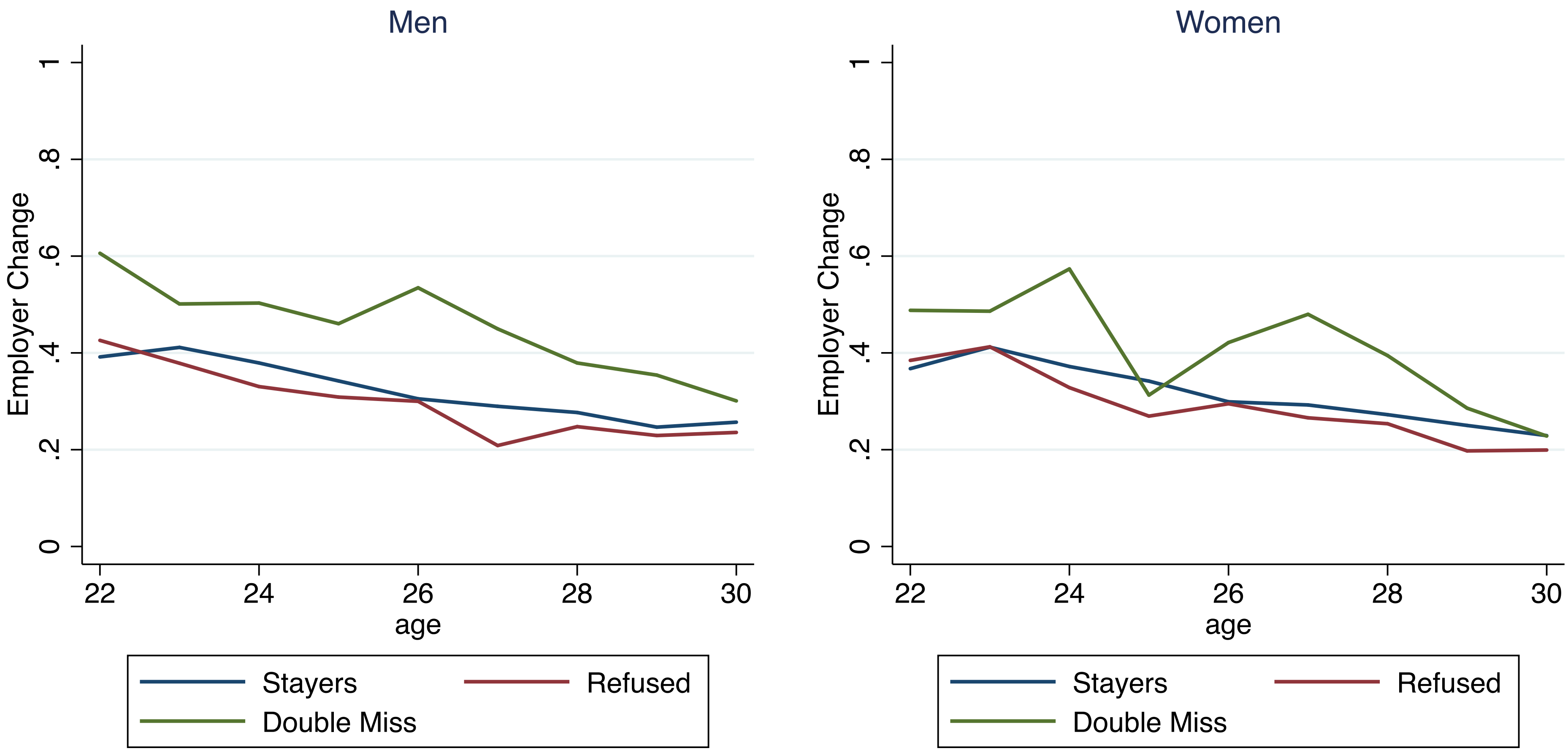
NLSY79 Simulated Attrition: Number of cases

	Men	Women	Total	%
Stayer	3,454	4,292	7,746	0.87
Refused	356	369	725	0.08
Double Miss	283	130	413	0.05
Total	4,093	4,791	8,884	

Employment by attrition status



Employer changing by attrition status



1) Reweighting

- Often provided by survey
  - Really just crude matching on observables
  - Usually based upon a limited set of demographic variables
- Original Cohort reweighting scheme
  - Divide respondents into cells based upon:
    - Black: yes or no
    - Years of residence in initial survey: <9, 10+, N/A
    - Father's occupation: white collar, service, blue collar, farm, N/A
  - Increase weights of remaining members of each cell

2) Optimal Matching

- Challenges
  - Some individuals have short or no sequences
  - How to represent the sequences
    - Can't have missing values
  - Multi-dimensionality
  - Defining substitution costs
- Not unique to SA

Optimal matching setup

- Alphabet: 33 work/occupation/employment states
  - 6 nonworking states: unemployed, school, military, jail, out of the labor force, missing
  - 27 working states: occupation x employment status
    - 9 occupation groups: professional/technical, manager, sales, clerical, craft, operative, laborer, service, farm
    - 3 employment statuses: newly employed, same employer, new employer
- Substitution costs set by transition rates
- Localized OM: x=0.1, y=0.8

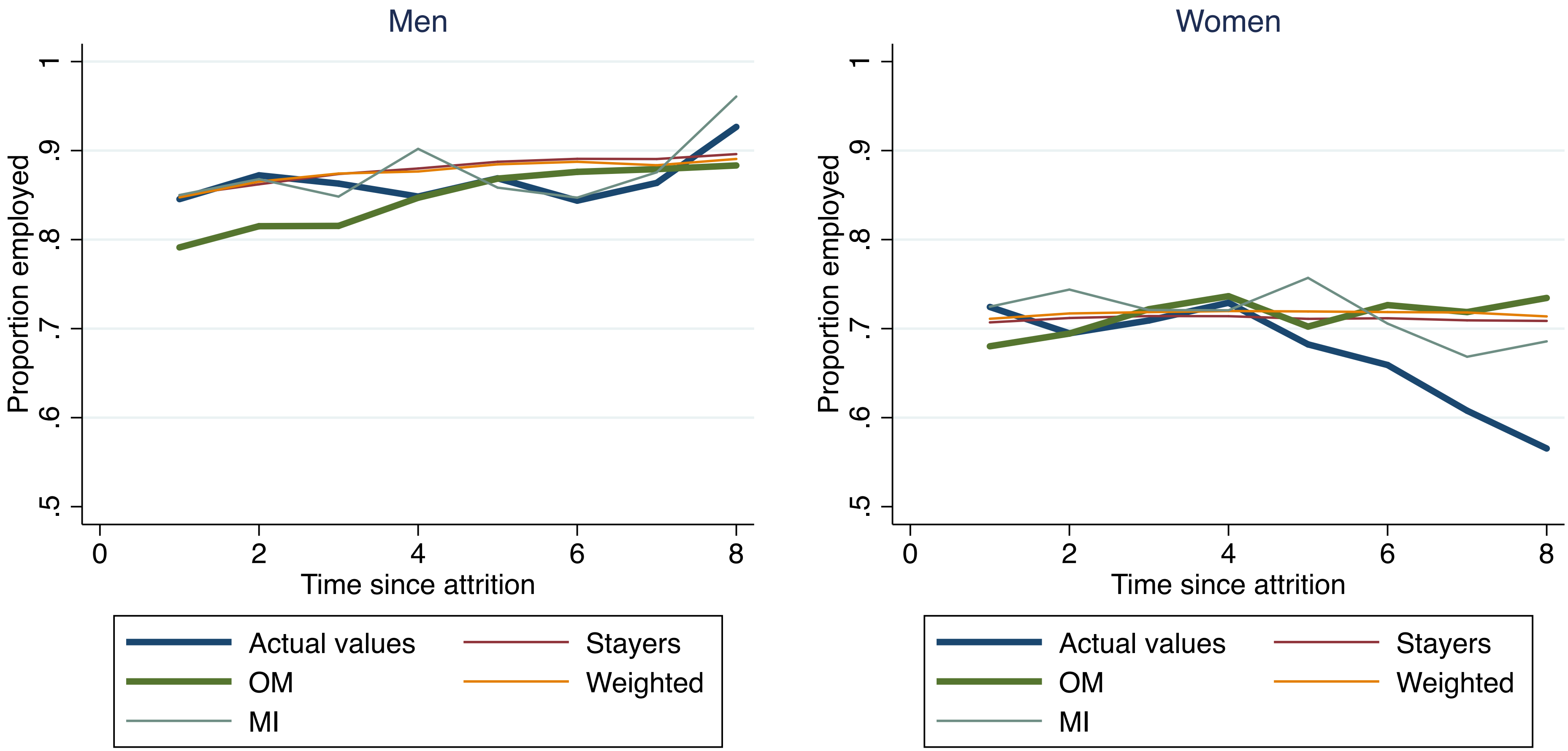
Optimal matching analysis

- For each treatment individual, distance to all control sequences. Length based upon treatment sequence length.
- Identify nearest match. In case of multiple matches take the average across matches

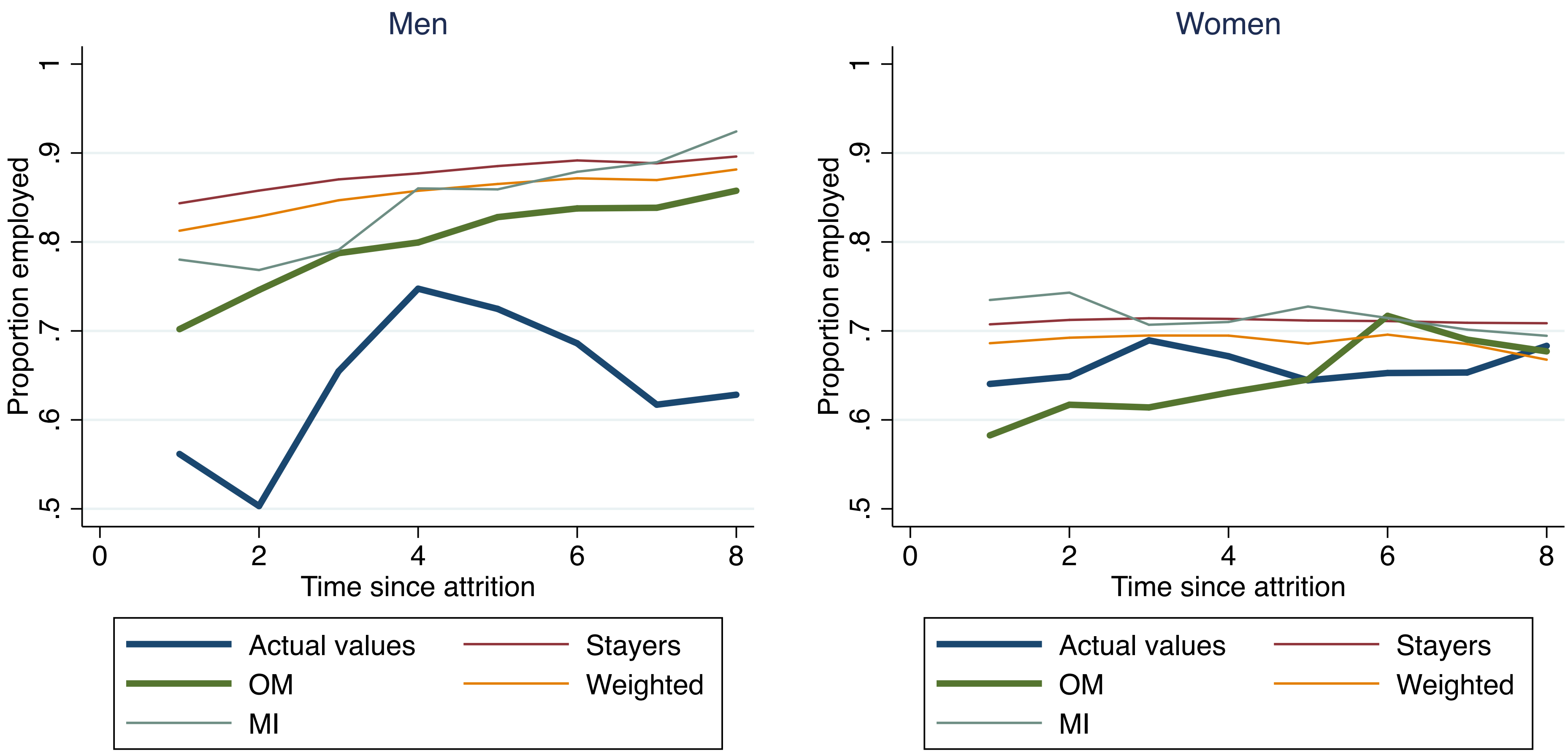
3) Multiple imputation

- MI difficulties:
  - Convergence for nominal variables
- Outcomes: employed
- Technique: logit
- Independent variables
  - Last two years before attrition
  - Employed dummy, occupational prestige (not employed=0), employer change (not employed, employed no change/ newly employed, employer change)

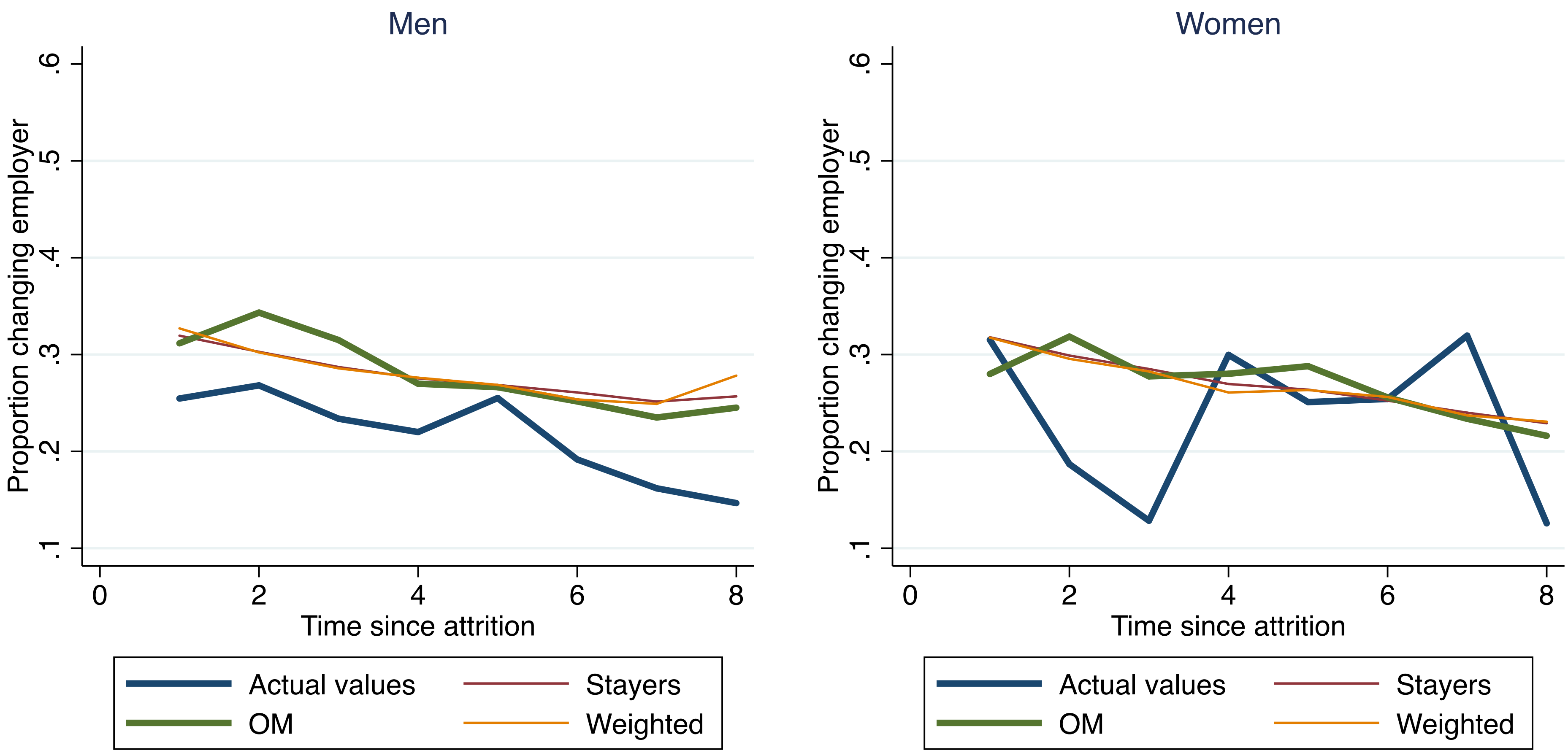
Results: Employment, Refused



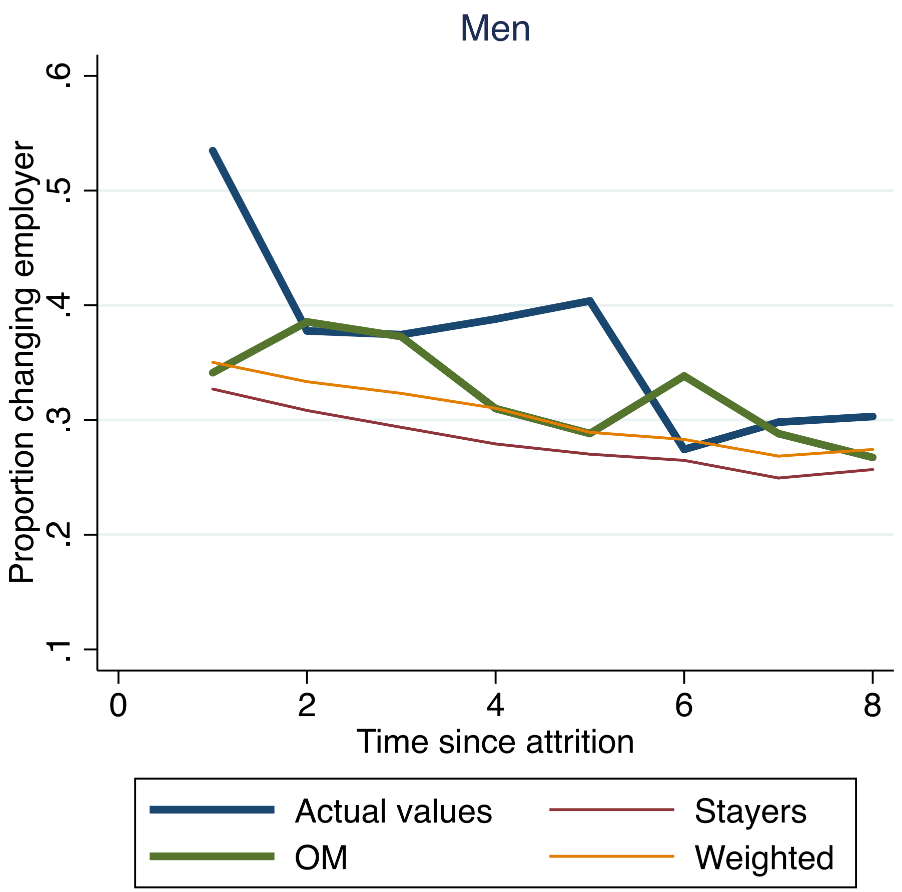
Results: Employment, Double Miss

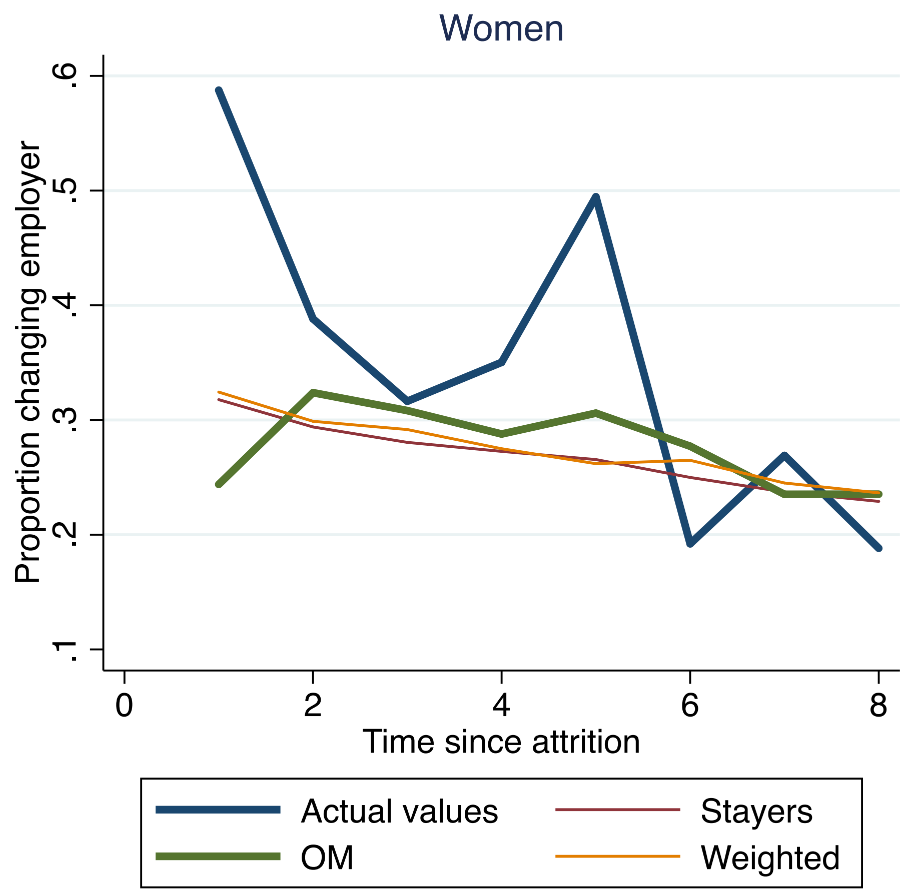


Results: Emp Change, Refused



Results: Emp Change, Double Miss





Summary

- Individuals who exit a survey are different
- Type of attrition matters
- Weighting
  - Only marginal improvement over ignoring missing values (relying on stayers)

  Future
  - May be unfair case, use better variables

Summary (cont)

- MI
  - Difficult to implement for nominal variables
  - In some cases better, some worse

Future
  - Add other (non-sequence) variables
  - Other software besides Stata

Summary (cont)

- OM
  - In some cases better
  - May be less successful if observed sequence is very short
    - If running a longitudinal study, collect some retrospective data at first survey

Future

  - Add in other life-course measures (esp for women)
    - Ways to deal with multi-dimensionality
  - Other matching methods besides nearest neighbor
  - Combine with Mahalanobis or some other technique to include non-sequence data? (background variables)
  - Greater weight on time periods just before treatment?

Conclusions (cont.)

- Explore counterfactual applications as well
  - May observe more treatment selection factors
    - Ashenfelter dip (1978)
  - Existing randomized experiment, compare to synthetic controls

	Men		Women	
Samples	Refused	2-in-a-row	Refused	2-in-a-row
Employed first year	-0.037*	0.198***	-0.030	0.084*
	(0.018)	(0.023)	(0.019)	(0.037)
Employed both years	-0.033	0.159***	-0.028	0.106*
	(0.019)	(0.026)	(0.020)	(0.042)
Separate estimates by education group				
Employed first year				
Less than high school	-0.058	0.139***	-0.009	0.123
	(0.048)	(0.037)	(0.061)	(0.086)
High school	-0.055*	0.263***	-0.046	0.124
	(0.027)	(0.033)	(0.028)	(0.076)
Some college	-0.018	0.148*	0.000	0.145*
	(0.039)	(0.061)	(0.034)	(0.058)
College or more	0.005	0.146	-0.038	-0.017
	(0.044)	(0.107)	(0.047)	(0.070)
Employed both years				
Less than high school	-0.061	0.082	-0.083	0.177
	(0.049)	(0.048)	(0.090)	(0.137)
High school	-0.044	0.219***	-0.022	0.161
	(0.027)	(0.037)	(0.029)	(0.087)
Some college	-0.019	0.129*	-0.012	0.173**
	(0.040)	(0.062)	(0.037)	(0.060)
College or more	0.001	0.163	-0.047	-0.019
	(0.045)	(0.112)	(0.046)	(0.067)

Notes: Each pair of coefficients represents the result of a separate estimation under different sample restrictions. The coefficients represent the employer separation rate compared to the omitted group of non-attritters. All models control for a quadratic of age and the first two rows control for dummy variables representing the four education groups. Standard errors in parentheses.

* p<0.05, ** p<0.01, *** p<0.001, two-tailed tests

Samples	Full Sample	Adjusted	Full Sample	Adjusted
Employed first year	0.095***	0.076***	0.016	-0.005
	(0.010)	(0.010)	(0.010)	(0.010)
Employed both years	0.094***	0.069***	0.061***	0.033**
	(0.010)	(0.011)	(0.011)	(0.012)
Separate estimates by education group				
Employed first year				
Less than high school	0.116***	0.092**	0.058	0.034
	(0.026)	(0.030)	(0.032)	(0.036)
High school	0.095***	0.070***	0.010	-0.018
	(0.016)	(0.017)	(0.016)	(0.017)
Some college	0.098***	0.072***	-0.006	-0.029
	(0.020)	(0.022)	(0.019)	(0.021)
College or more	0.078***	0.076***	0.023	0.017
	(0.018)	(0.019)	(0.019)	(0.020)
Employed both years				
Less than high school	0.101***	0.084**	0.079	0.040
	(0.027)	(0.031)	(0.042)	(0.048)
High school	0.075***	0.041*	0.063***	0.024
	(0.016)	(0.017)	(0.018)	(0.019)
Some college	0.108***	0.077***	0.041	0.015
	(0.020)	(0.022)	(0.021)	(0.023)
College or more	0.098***	0.090***	0.064**	0.050*
	(0.019)	(0.020)	(0.020)	(0.022)

Notes: All values in table are slope coefficients (and standard errors) for the *cohort* variable and each represents the result of a separate estimation using different sample restrictions. All models control for a quadratic of age and the first two rows of models control for dummy variables representing the four education groups. Standard errors in parentheses.

[1] 1966 cohort and men from the 1979 cohort

[2] 1968 cohort and women from the 1979 cohort

* $p<0.05$, ** $p<0.01$, *** $p<0.001$, two-tailed tests