National Center for Health Statistics
Data Linkage

On dealing with “incompletely linked” data in linked survey/administrative databases: An empirical comparison of alternative methods

Dean H. Judson
Jennifer D. Parker
National Center for Health Statistics

Outline

- The problem of incompletely linked data
- Data
- Methods
- Toy model
- Results
- Conclusion and future work

Objectives

- When data are incompletely linked, an unique “missing data” problem emerges
- Two goals:
  - Determine if inferential models’ coefficients are biased due to incomplete linkage
  - Determine if individual subgroups are more affected than others

Definitions

- “Incompletely linked data”: Data sets which, by design or because of lacking linkage. information, are linked at a rate less than 100%.
- “Administrative longitudinal data”: Linked data sets which contain administrative data over time.
- “Linkage ineligible”: Survey respondents who are ineligible to be linked.
- “Program ineligible”: Respondents who are not part of the administrative program.

The problem at hand

"Standard" missing data:

"Linkage ineligible" missing data:

Q: Does the missingness pattern impact inferences greatly, and if so, can we fix the situation?

Data

- 1997-2005 National Health Interview Survey with Medicare match flags:

  1 Eligible, link was found;

  2 Eligible, link was not found;

  3 Ineligible
- Percent ineligible peaked in 2006 at 57% (Miller et al., 2011)
- Treat potential “nonresponse bias”

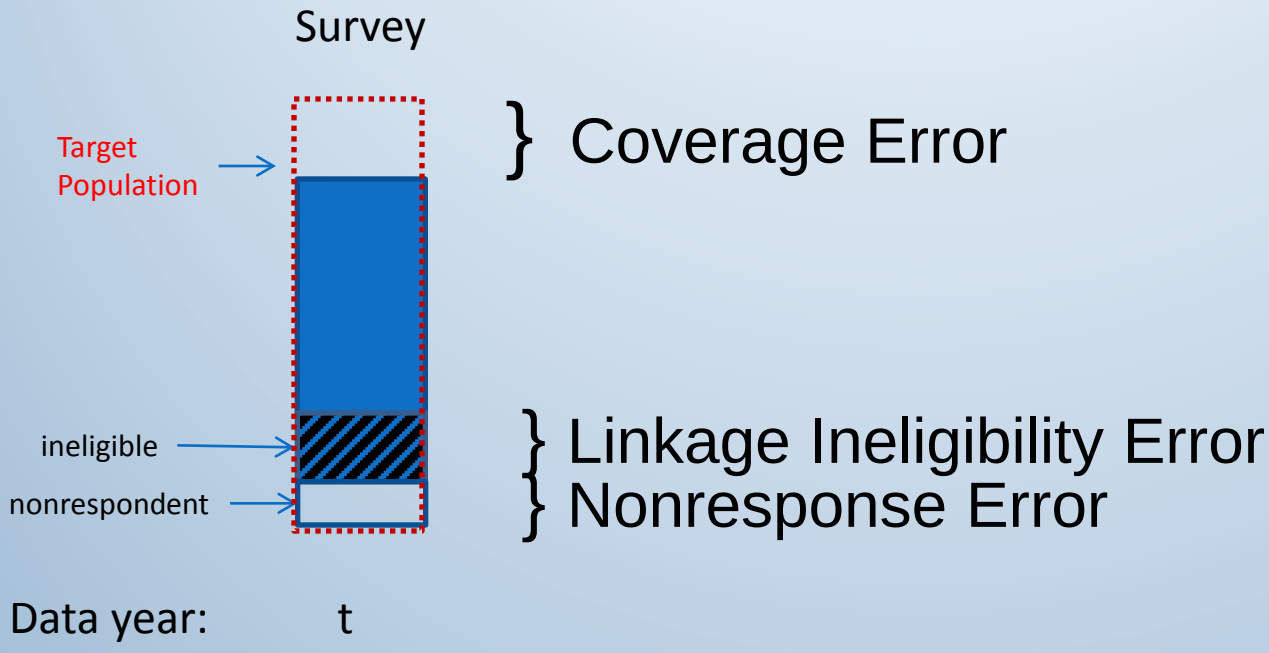
Survey weighting 101 (stylized)

- Typically, final weights are the product of weighting adjustments, e.g.
- Sampling weight (1/P[selection]) *times*
- Nonresponse adjustment by weighting class (region, race of householder) *times*
- Coverage adjustment to housing unit control totals (by class) *times*
- Coverage adjustment to person control totals (by age/race/sex/Hispanic origin)

Reweighting

- Reweighting is a standard technique for correcting for linkage ineligibility
- Mirel and Parker, 2011, describe only modest impacts of reweighting for NHANES
- Conceptually, reweighting occurs after the post-stratification controls, and simply represents another coverage adjustment, following similar principles

(No linkage, cross-sectional, single survey year, loss only due to survey nonresponse and linkage ineligibility)



Methods

- Step one: Logistic regression of fair/poor health (1/0) on:
  - Continuous age and age-squared;
  - Indicators of marital statuses (Married, Div/Sep, Widowed)
  - Indicators of race/ethnicity (NHW, NHB, NHO);
  - Indicators of educational attainment (HS, College+);
  - Indicator of uninsured status; and
  - Indicator of survey year, INTERACTED WITH:
  - Indicator of Nonlinkability
- Want to see interaction effects of 1.0 (i.e., no effect)
- Step two: Remove linkage ineligibles, test various reweighting strategies
- Step three: Remove linkage ineligibles, test various reweighting strategies with key subgroups

Step One: Run Toy Model

- Based on model presented in Zheng and Schimmele, AJPH, 2005 (and others):
- "Natural Experiment": Compare coefficients estimated on entire survey respondent population vs. those estimated only on *linkage eligible*
- First step: Baseline model vs. model interacted with (nonlinkage) dummy

Logistic Regression Estimates Using Whole Sample, (Non)Linkage dummy included

	Odds ratio (relative to baseline category)	t statistic
(Base model above)		
Notlinkable	0.791***	(-5.99)
Married, notlinkable	0.945*	(-2.25)
Div/Sep, notlinkable	0.841***	(-5.25)
Widowed, notlinkable	1.066	(1.89)
WNH, notlinkable	1.068**	(2.59)
BNH, notlinkable	0.985	(-0.51)
ONH, notlinkable	1.086	(1.61)
HS education, notlinkable	1.062**	(2.64)
College+, notlinkable	1.075**	(3.05)
uninsured, notlinkable	0.867***	(-5.41)
(Survey year dummies omitted)		
Observations	778905	

Exponentiated coefficients (odds ratios are relative to the omitted first category for indicator variables); *t* statistics in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; age and age squared are treated as continuous; Div/Sep refers to Divorced or Separated; NHW is Non Hispanic White; BNH is Non Hispanic Black; NHO is Non Hispanic Other race; HS is high school degree attained; College+ refers to some college or more; indicator variables take the value one if the record is in the named class, zero otherwise.

Step Two: Remove Linkage Ineligibles

- One toy model
- Full sample (“truth” deck)

  vs.
- Eligible-only w/ different reweighting strategies
- Do coefficients change? Are inferences “at risk” of damage due to choice of reweighting model?
- Thus, focus on *bias* relative to the *known* (full survey) model

Example reweighting strategies

- PROC WTADJUST (Sudaan)
- Create cross-classification table of characteristics relevant to linkage ineligibility (e.g., age, race/ethnicity, sex, region, education)
- Estimate proportion ineligible by class using a model/strategy
- Linkage ineligible receive final weight of *zero* (will not contribute to analysis)
- Linkage eligible receive final weight of (approximately) original weight * (1/proportion ineligible in their class)
- Collapse classes if class size n “too small” (e.g., <30) for reliable estimation of the adjustment factor 1/p

Example reweighting strategies, cont.

- Margin-only model:
  - Age || race/ethnicity || sex; no interaction effects
- Saturated model:
  - Age * race/ethnicity * sex; all one-, two-, three-way interactions
- Continuous age model:
  - Age, Age-squared treated continuous; race/ethnicity*sex
- Region/SES model:

  Any of above, PLUS (or interacted with) region, education

Example Sudaan code

```
* MARGIN ONLY MODEL W/REGION AND EDUCATION;
proc wtadjust data=local.merged_nhis_1997_2005_d design=wr
        adjust=nonresponse notsorted;
   nest stratum psu;
   weight wtfa;
   reflevel age_cat=2 raceeth=2 sex=1 region=1 educ=2;
   class age_cat raceeth2 sex region education / include=missing;
   model linkable=age_cat raceeth2 sex region education;
   idvar linkable age_cat raceeth2 sex region education id;
   print beta sebeta p_beta margadj / betafmt=f10.4
   sebetafmt=f10.4;
   output /predicted=all filename=match1 filetype=sas replace;
run;

* SATURATED (AGE/RACE/SEX) MODEL INDEPENDENT OF REGION AND
   EDUCATION;
   model linkable=age_cat*raceeth2*sex region education;
```

-
```
   CONTINUOUS AGE;
   model linkable=age_p age_p2 raceeth2*sex*region*education;
```

Diagnostics

- Check marginal adjustment factors (there should not be any “large” differences)
- Check sums, means, variance, kurtosis of reweights against original weights
- Correlate and plot different reweights against each other
- Plot reweights against original weights, omitting zero reweights

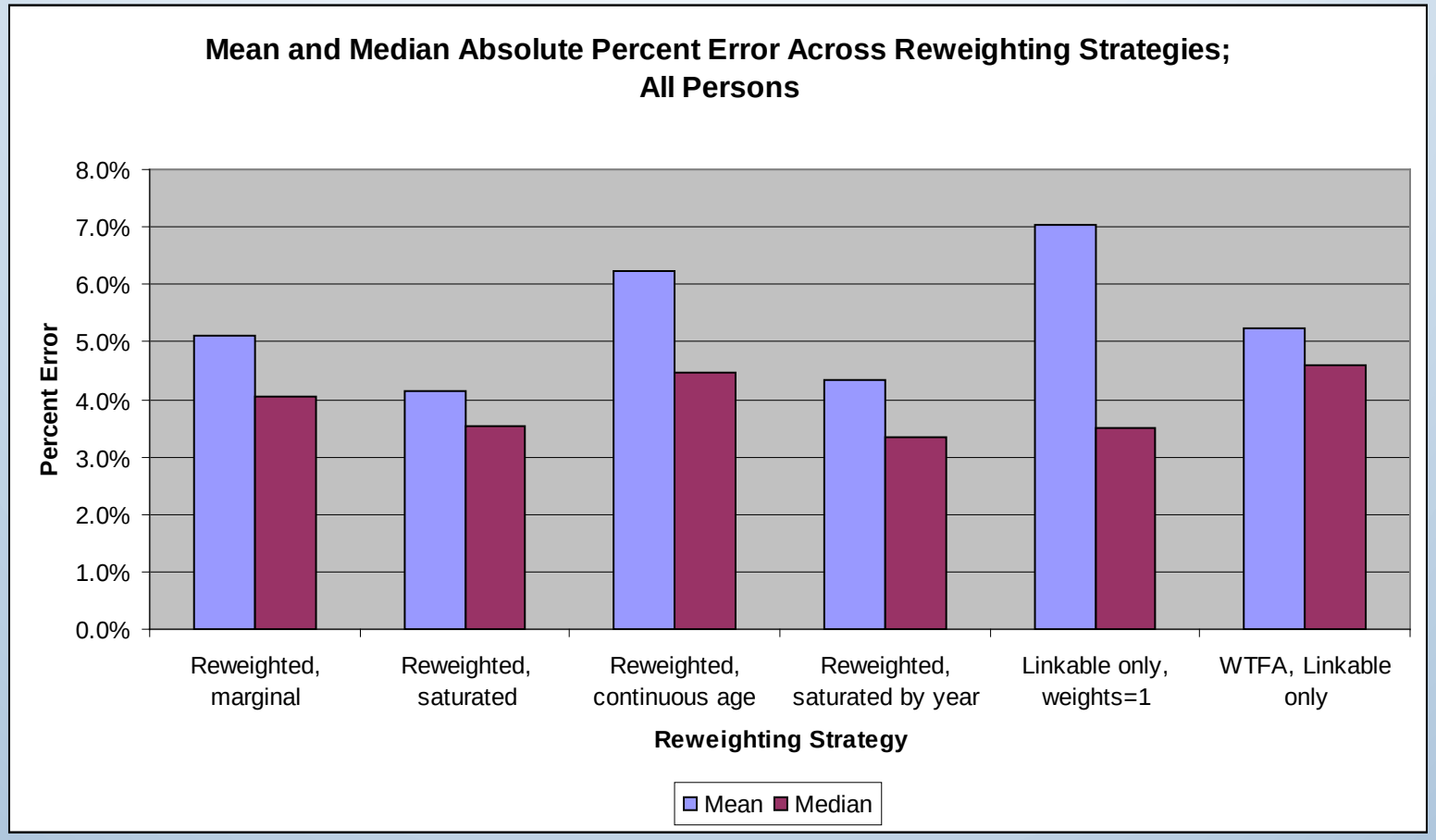
Basic output: Logit coefficients (sample)

	Original estimates	Reweighted, marginal	Reweighted, saturated	Reweighted, continuous age	Reweighted, saturated by year	Linkable only, weights=1	WTFA, Linkable only
Age	0.113	0.118	0.118	0.120	0.118	0.117	0.118
Age squared	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
Marital Status: Not Married	0.303	0.298	0.298	0.312	0.294	0.307	0.287
Marital Status: Married	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Marital Status: Div/Sep	0.450	0.493	0.495	0.502	0.490	0.489	0.495
Marital Status: Widowed	0.095	0.097	0.097	0.114	0.093	0.118	0.088
Race/Ethnicity: Hispanic	0.323	0.340	0.337	0.329	0.349	0.400	0.348
Race/Ethnicity: NHW	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Race/Ethnicity: NHB	0.582	0.614	0.617	0.608	0.621	0.632	0.619
Race/Ethnicity: NHO	0.188	0.179	0.182	0.174	0.182	0.187	0.181
Education level: <HS	0.697	0.725	0.727	0.735	0.724	0.698	0.723
Education level: HS+	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Education level: College+	-0.587	-0.604	-0.603	-0.604	-0.593	-0.588	-0.598
Insured	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Uninsured	0.217	0.224	0.225	0.221	0.237	0.160	0.255
Not Foreign Born	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Foreign Born	-0.433	-0.427	-0.436	-0.430	-0.442	-0.440	-0.433

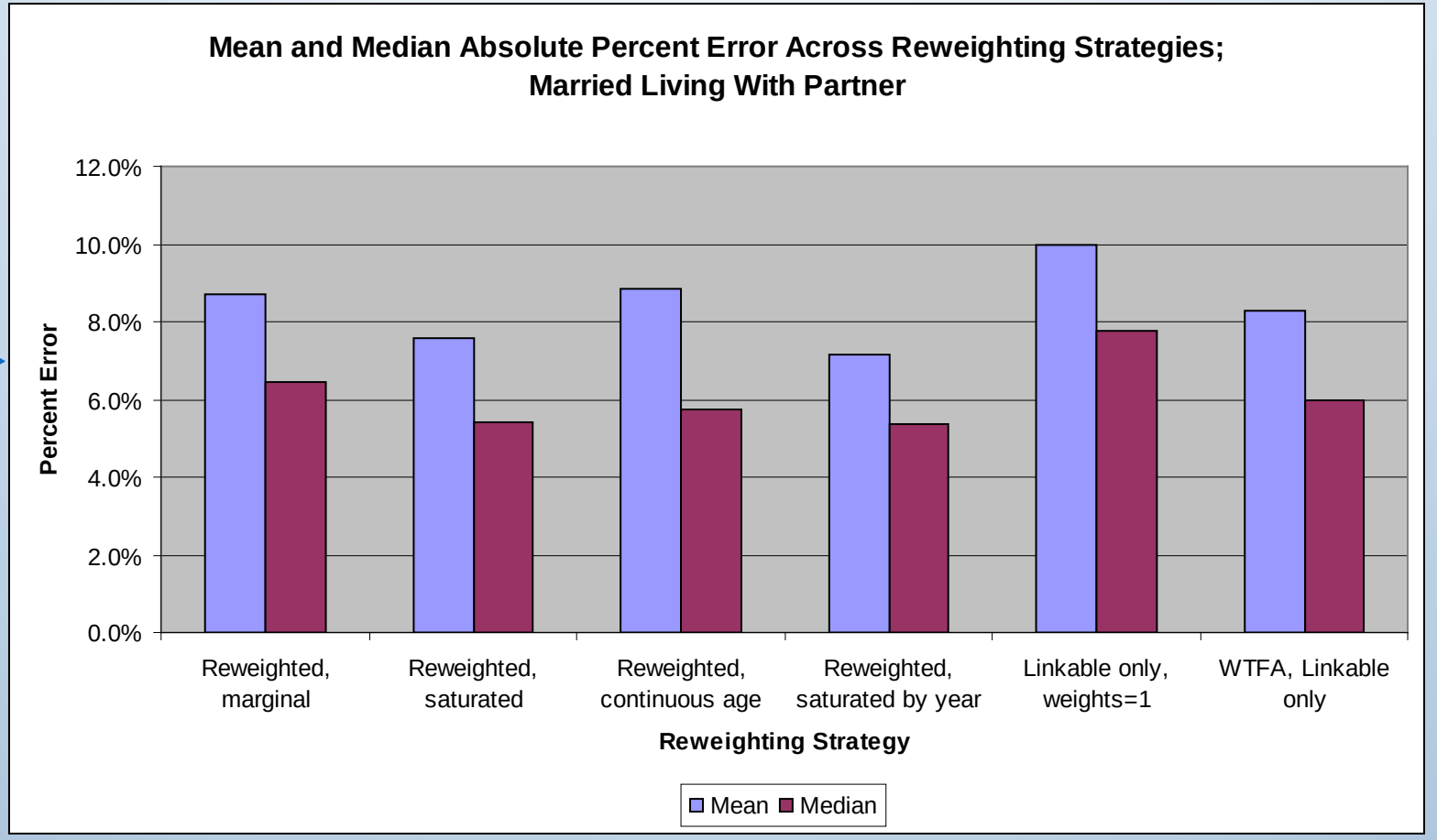
Step Three: Summary Measures of Error for All Persons and Select Subgroups

- Criteria:
  - Absolute percent error (one coefficient)
  - Mean absolute percent error (across all coefficients)
  - Median absolute percent error (across all coefficients)
- Error *relative to original full-sample model coefficients*
- All persons and several subgroups tested:
  - Age 65+, age <19, Hispanic/non, Married/non, educational attainment groups, foreign born

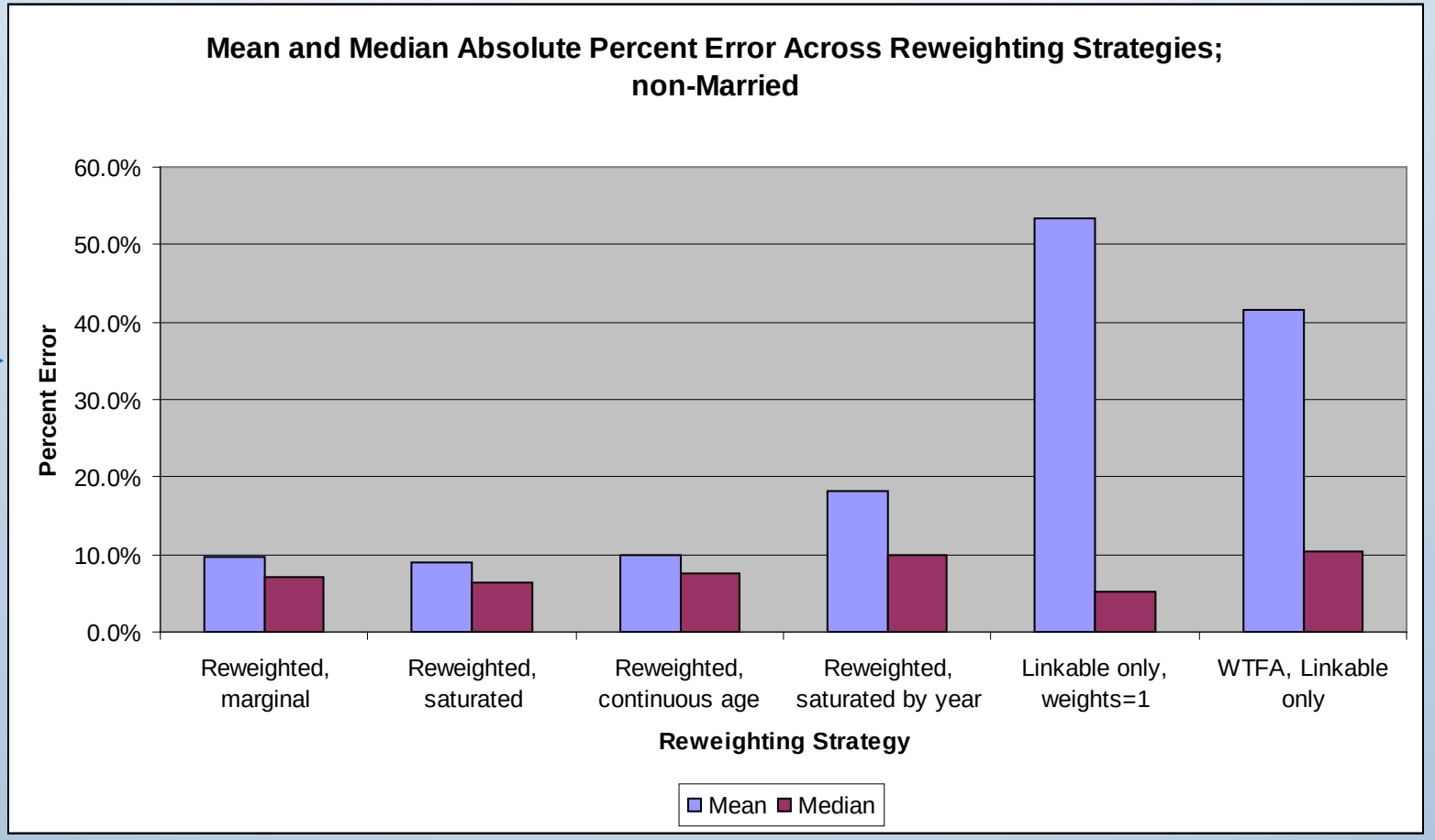
Note Scale →


Mean and Median Absolute Percent Error Across Reweighting Strategies; All Persons
Percent Error
8.0%
7.0%
6.0%
5.0%
4.0%
3.0%
2.0%
1.0%
0.0%
Reweighted, marginal
Reweighted, saturated
Reweighted, continuous age
Reweighted, saturated by year
Linkable only, weights=1
WTFA, Linkable only
Reweighting Strategy
Mean
Median

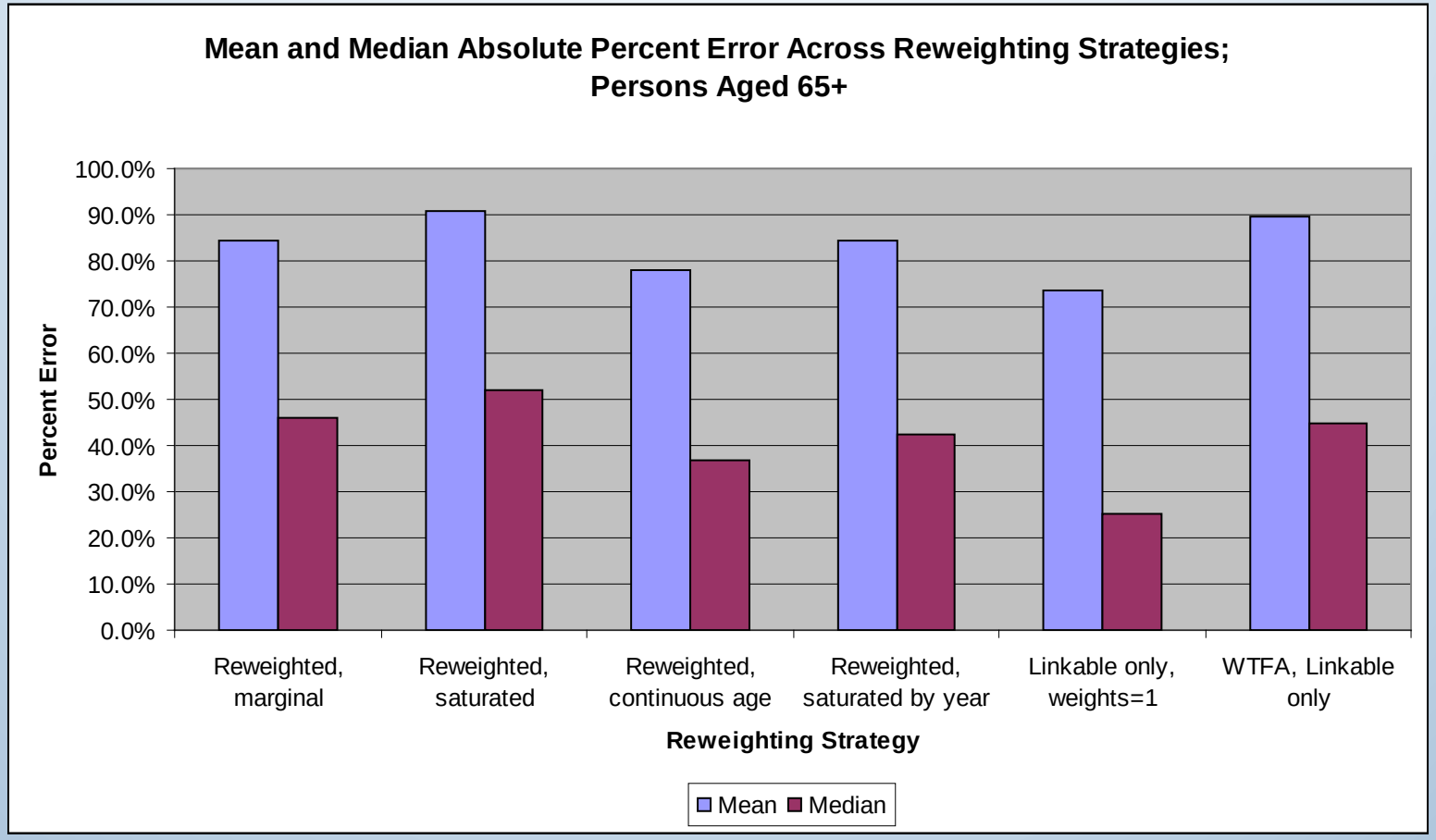
Note Scale →


Mean and Median Absolute Percent Error Across Reweighting Strategies; Married Living With Partner
12.0%
10.0%
8.0%
6.0%
4.0%
2.0%
0.0%
Percent Error
Reweighted, marginal
Reweighted, saturated
Reweighted, continuous age
Reweighted, saturated by year
Linkable only, weights=1
WTFA, Linkable only
Reweighting Strategy
Mean
Median

Note Scale →


Mean and Median Absolute Percent Error Across Reweighting Strategies; non-Married
Percent Error
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
Reweighted, marginal
Reweighted, saturated
Reweighted, continuous age
Reweighted, saturated by year
Linkable only, weights=1
WTFA, Linkable only
Reweighting Strategy
Mean
Median

Note Scale →


Mean and Median Absolute Percent Error Across Reweighting Strategies; Persons Aged 65+
Percent Error
100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
Reweighted, marginal
Reweighted, saturated
Reweighted, continuous age
Reweighted, saturated by year
Linkable only, weights=1
WTFA, Linkable only
Reweighting Strategy
Mean
Median

Conclusions From this Exercise

- Omitting linkage ineligibles (especially with naïve weights) results in notable biases in coefficients.
- Reweighting usually reduces, but does not entirely eliminate, these biases.
- Reweighting strategies have comparable effects.
- Some subgroups appear especially ‘at risk’.

Next Steps

- Other approaches we'd like to test:
  - Mass multiple imputation (multiple imputation of individual values using chained equations)
  - Statistical matching (finding donors and imputing entire missing record)
  - Simultaneous estimation of ineligibility probability and substantive model (in WTADJUST it's a two-step procedure)