A Simulation Study on the Effect of Matching Error on Triple System Estimation

Richard Griffin

U.S. Census Bureau

FCSM Research Conference

November 5, 2013

Topics

- Background & Motivation
- Matching error model for Dual System Estimation (DSE)
- Matching error model for Triple System Estimation (TSE)
- Simulations
- Results
- Analysis and Summary

Background

- Goal is to estimate true population count
- Census (C); Administrative list (A); Post-Enumeration Survey (P)
- Bias in DSE (using C and P) caused by causal dependence or variation in capture probabilities
- TSE, adding A to C and P may reduce bias
- Assume all enumerations in all lists are correct

Background

- No sampling or movers
- Matching error is modeled
- Purpose is to determine effect of matching error on estimation accuracy
- Analysis of estimators with no matching error in Griffin (2014)

DSE Observed Counts After Matching

	In PES	Out of PES	Total
In Census	N_{11}	N_{10}	N_{1+}
Out of Census	N_{01}	N_{00} = ?	
Total	N_{+1}		

DSE for Unobserved Cell

$$\hat{N}_{00} = \frac{N_{10}N_{01}}{N_{11}}$$

DSE Matching Error Model Biemer (1988)

For all persons in the PES, define indicators

- y = 1 for matched to Census

  y = 0 for not matched to Census

- z = 1 for truly in Census

  z = 0 for truly not in Census

DSE Matching Error Model Biemer (1988)

- Taylor Series expansion used to approximate E(DSE) under the matching error model and Var(DSE)

DSE Matching Error Rates

- False Positive Error Rate

  $P[\, y = 1 \mid z = 0 \,] = \Phi$

- False Negative Error Rate

  $P[\, y = 0 \mid z = 1 \,] = \alpha$

Triple System Observed Counts After Matching

	In A		Out of A	
	In PES	Out of PES	In PES	Out of PES
In Census	N_{111}	N_{101}	N_{110}	N_{100}
Out of Census	N_{011}	N_{001}	N_{010}	$N_{000} = ?$

NSOI – No Second Order Interaction Model

- Assume that the dependence in the 2x2 table for C and P using only persons on A is the same as the dependence for C and P using only persons not on A (same for other pairs of sources)
- All pairs of sources can exhibit dependence but the amount of dependence in each pair is assumed to be unaffected by conditioning on the third source

NSOI for Unobserved cell

$$\hat{\theta}_1 = \frac{N_{111}/N_{011}}{N_{101}/N_{001}} = \frac{N_{111}N_{001}}{N_{011}N_{101}}$$

$$\hat{N}_{000} = \hat{\theta}_1 \frac{N_{100}N_{010}}{N_{110}}$$

Other TSE Models

- Fienberg (1972)
- There is more than one potential log linear model useful for TSE, each producing a different estimator
- All these models are more restrictive than the no second order interaction model
- If any of these models is valid, so is the no second order interaction model

CI - Conditionally Independent Model

- At each level of C, P and A are independent

$$\hat{N}_{000} = \frac{N_{001} N_{100}}{N_{101}}$$

JI – Jointly Independent Model

- One source jointly independent other two
- Ordinary two-way independence between A and a combined C and P list

$$\hat{N}_{000} = \frac{N_{001} N_{++0}}{N_{++1} - N_{001}}$$

$$N_{++0} = N_{110} + N_{100} + N_{010}$$

$$N_{++1} = N_{111} + N_{101} + N_{011} + N_{001}$$

Triple System Matching Procedure

- Step 1: First match P to C and then P to A to determine N_{111}, N_{110}, N_{011}, and N_{010}
- Step 2: Next match A to C to determine N_{101} and N_{001}
- Step 3: Finally, match C to A and P to determine N_{100}

Triple System Matching Error Model

- For all persons in either of two lists (C&P,C&A, or P&A) define indicators for match or not matched to the third list
- For all persons, define indicators for truly in or not in each of the three lists

Triple System Matching Error Model

- Assume all false positive rates = φ
- Assume all false negative rates = α
- This is a limitation since error rates could vary for matching between different pairs of lists

Triple System Matching Error Model

- Use indicator functions and Taylor Series expansion to approximate expected value under the matching error model and variance of NSOI, JI, and CI
- See paper; much more complicated algebra and Taylor approximation than for DSE

Generating a 1000 Person Population

- For k = 1 to 1000, independently generate X_k ~N(0,1) and calculate needed probabilities for a specified set of parameters
- Each probability is from a simple logistic regression with an intercept and slope parameter
- Details in paper

Generating a 1000 Person Population

- Θ_1 odds ratio for C and P from 2 x 2 table, given in A
- Θ_2 odds ratio for C and P from 2 x 2 table, given not in A
- Odds ratios are given but do not vary by k
- Randomly assign each person k to one of the eight cells based on the probabilities, Θ_1, and Θ_2

Simulations

- 1000 independent replications of the population generation just described.
- Tabulate the counts in each of the 8 cells for each replication
- Calculate each of the alternative estimators for each replication

Average Expected Value of Estimates (True POP = 1000) $\Phi = .062; \alpha = .023$

Estimation Alternatives	$\Theta_1=1.5; \Theta_2=1.2$	$\Theta_1=.75; \Theta_2=.85$	$\Theta_1=.75; \Theta_2=.75$	$\Theta_1=1.5; \Theta_2=1.5$
NSOI	1124	**1015**	1047	**1052**
JI	1088	1080	1089	1079
CI	**963**	955	**974**	946
DSE	1002	1283	1332	934

Average Standard Error of Estimates

Estimation Alternatives	Θ_1=1.5; Θ_2=1.2	Θ_1=.75; Θ_2=.85	Θ_1=.75; Θ_2=.75	Θ_1=1.5; Θ_2=1.5
NSOI	106	93	100	90
JI	25	24	24	24
CI	22	21	21	21
DSE	32	36	36	31

Alternative Error Rates Average Expected Value of Estimates (True POP = 1000)

Estimation Alternatives	$\Theta_1 = 1.75$ $\Theta_2 = 1.25$			
False Positive Rate	.062	.12	.03	0
False Negative Rate	.023	.05	.01	0
NSOI	1183	402	1220	1125
JI	1093	**410**	1106	1098
CI	**967**	400	**967**	**965**
DSE	958	445	967	952

Conclusion and Summary

- Results would vary with different parameters and odds ratios.
- Conclusions serve as an illustration of potential results and cannot be used to make definitive generalizations.
- Matching error can mitigate the theoretical advantages of triple system modeling.

Conclusion and Summary

- Due to matching error, the DSE may be more accurate than any of the triple system estimators.
- For the triple system estimators, CI or JI may be more accurate than NSOI due to using fewer of the seven observed cells in the triple system set up.

Conclusion and Summary

- Increasing the matching error from the levels used by Biemer (1988) has a great effect on the accuracy of all the estimators.
- The best estimators, CI and DSE, for the Biemer level matching error had about a 4 percent undercount and doubling the matching error resulted in about a 60 percent undercount.

Conclusion and Summary

- Decreasing the matching error level produced little change in accuracy.
- The relative accuracy among the estimators did not change much with varying matching error levels.
- CVs of estimators average around .03 or less for JI, CI, and DSE but about .10 for NSOI

References

- Biemer, P.P. (1988), Using Information from Demographic Analysis in Post-Enumeration survey Estimation, Survey Methodology, 14,117-134
- Fienberg, S. (1972), “The Multiple Recapture Census for Closed Populations and Incomplete 2k Contingency Tables, ”Biometrica, Vol. 59, No.3, 591-603.
- Griffin, R. (2014), “Potential Uses of Administrative Records for Triple System Modeling for Estimation of Census Coverage Error in 2020”, Journal of Official Statistics, to be published in special addition

- richard.a.griffin@census.gov