Nonrespondent subsample multiple imputation in two-phase random sampling for nonresponse

Nanhua Zhang
Division of Biostatistics & Epidemiology
Cincinnati Children's Hospital
Medical Center
(Joint work with Henian Chen & Michael Elliott)

Outline

- Overview
- Two-phase sampling for nonresponse
- A comparison of methods
- Nonresponse subsample multiple imputation
- Simulations
- Application
- Discussion and conclusion

Overview

- Methods for nonresponse
  - Complete-case analysis
  - Ignorable likelihood methods
  - Nonignorable modeling
- Limitations
  - All rely on untestable assumptions
  - Design to avoid missing data
  - Two-phase sampling helps

Two-phase sampling for nonresponse

- First proposed to reduce non-response bias in mail questionnaire
  - Hansen and Hurwitz (1946)
  - Weighting methods for estimate mean/total
- Survey setting
  - National Comorbidity Survey
  - Canadian National Household Surveys
- Clinical trials
  - NRC (2010)

Notation

Pattern	Observation, i	y_i	R_1	R_2	$R_{2\cdot 0}$
1	$i = 1,\ldots,m$	√	1	1	-
2	$i = m+1,\ldots,m+r$	?	0	1	1
3	$i = m+r+1,\ldots,n$	x	0	0	0

Key: √ denotes observed, x denotes at least one entry missing, ? denotes at least one entry missing in phase I but fully observed in phase II

$$\Pr(R_{2\bullet 0,i} = 1 \mid R_{1,i} = 0; z_i, y_i) = \pi$$

Multiple Imputation

- Nonresponse weighting (mean/total)
  - Unbiased
  - Not using auxiliary information
  - Large variance
- Multiple imputation
  - Uses auxiliary variables
  - More efficient when used properly
  - Three options for two-phase sampling

Multiple imputation

- Ignorable likelihood

$$L\left(\phi \mid Y_{obs}\right) \propto P\left(Y_{obs} \mid \phi\right)$$

- Multiple imputation

$$P\left(Y_{mis} \mid Y_{obs}\right) = \int P\left(Y_{mis} \mid Y_{obs}, \phi\right) P\left(\phi \mid Y_{obs}\right) d\phi$$

MI1: use only phase I data

Pattern	Observation, i	y_i	R_1	R_2	$R_{2\cdot 0}$
1	$i = 1,\ldots,m$	√	1	1	-
2	$i = m+1,\ldots,m+r$	~~?~~x	0	1	1
3	$i = m+r+1,\ldots,n$	x	0	0	0

Key: √ denotes observed, x denotes at least one entry missing, ? denotes at least one entry missing in phase I but fully observed in phase II

$$P(R_1 \mid Y_{obs,1}, Y_{obs,2}, Y_{mis}; \xi) = P(R_1 \mid Y_{obs,1}; \xi)$$

MI2: also use phase II data

Pattern	Observation, i	y_i	R_1	R_2	$R_{2\cdot 0}$
1	$i = 1,\ldots,m$	√	1	1	-
2	$i = m+1,\ldots,m+r$	~~?~~ √	0	1	1
3	$i = m+r+1,\ldots,n$	x	0	0	0

Key: √ denotes observed, x denotes at least one entry missing, ? denotes at least one entry missing in phase I but fully observed in phase II

$$P(R_2 \mid Y_{obs}, Y_{mis}; \xi) = P(R_2 \mid Y_{obs}; \xi)$$

CC1: use CCs from phase I

Pattern	Observation, i	y_i	R_1	R_2	$R_{2\cdot 0}$
1	$i = 1,\ldots,m$	√	1	1	-
2	$i = m+1,\ldots,m+r$	~~?~~x	0	1	1
3	$i = m+r+1,\ldots,n$	x	0	0	0

Key: √ denotes observed, x denotes at least one entry missing, ? denotes at least one entry missing in phase I but fully observed in phase II

$$P(R_1 \mid Y_{obs}, Y_{mis}; \xi) = P(R_1 \mid \xi)$$

CC2: also use phase II data

Pattern	Observation, i	y_i	R_1	R_2	$R_{2\cdot 0}$
1	$i = 1,\ldots,m$	√	1	1	-
2	$i = m+1,\ldots,m+r$	~~?~~ √	0	1	1
3	$i = m+r+1,\ldots,n$	x	0	0	0

Key: √ denotes observed, x denotes at least one entry missing, ? denotes at least one entry missing in phase I but fully observed in phase II

$$P(R_2 \mid Y_{obs}, Y_{mis}; \xi) = P(R_2 \mid \xi)$$

Nonrespondent subsample multiple imputation (NSMI)

Pattern	Observation, i	y_i	R_1	R_2	$R_{2\cdot 0}$
1	$i = 1,\ldots,m$	√	1	1	-
2	$i = m+1,\ldots,m+r$	~~?~~ √	0	1	1
3	$i = m+r+1,\ldots,n$	x	0	0	0

Key: √ denotes observed, x denotes at least one entry missing, ? denotes at least one entry missing in phase I but fully observed in phase II

$$\Pr(R_{2\cdot 0,i} = 1 \mid R_{1,i} = 0; z_i, y_i) = \pi$$

When and why is NSMI valid?

- Nonrespondent Subsample Missing at Random (NS-MAR)

$$\mathrm{P}(\boldsymbol{R}_{2\bullet 1}=1 \mid \boldsymbol{R}_1=0, \boldsymbol{Y}_{obs,2}, \boldsymbol{Y}_{mis}; \boldsymbol{\xi}, \boldsymbol{Z}) = \mathrm{P}(\boldsymbol{R}_{2\bullet 1}=1 \mid \boldsymbol{R}_1=0, \boldsymbol{Y}_{obs,2}; \boldsymbol{\xi}, \boldsymbol{Z})$$

- Rational
  - MAR is valid within nonrespondents in phase I but may be invalid if extended to whole sample

Simulation studies

- Goal:
  - Compare the performance to each method under different missing data mechanisms
  - Sample size consideration in phase II

Simulation studies

Pattern	Observation, i	z_i	x_i	y_i	R_1	$R_{2\bullet 0}$
1	$i = 1,\ldots,m$	√	√	√	1	-
2	$i = m+1,\ldots,m+r$	√	√	?	0	1
3	$i = m+r+1,\ldots,n$	√	√	x	0	0

$$(z,x)_i \sim N(0_{2\times 1}, \Sigma_{2\times 2} = \begin{pmatrix} 1 & .3 \\ .3 & 1 \end{pmatrix}), y_i \sim N(1+z_i+x_i, 1), i = 1,\ldots,1000$$

Simulation studies

- Missing data generation and two-phase sampling
- Phase I
  - MCAR: $\Pr\left(M_i = 1 \mid z_i, x_i, y_i\right) = \text{expit}\left(-1\right);$
  - MAR: $\Pr\left(M_i = 1 \mid z_i, x_i, y_i\right) = \text{expit}\left(-1 + z_i + x_i\right);$
  - MNAR: $\Pr\left(M_i = 1 \mid z_i, x_i, y_i\right) = \text{expit}\left(-y_i\right).$
- Phase II: $\Pr(R_{2\bullet 1,i} = 1 \mid M_i = 1; z_i, x_i, y_i) = 0.25.$

Simulation studies

- Six methods are applied to estimate the mean of *Y* and the regression coefficients:
  - CC1: complete-case analysis using respondents from phase I;
  - CC2: complete-case analysis using respondents from both phase I and II;
  - MI1: multiple imputation using data from phase I;
  - MI2: multiple imputation using data from both phase I and II;
  - NSMI: multiple imputation in the nonrespondent subsample in phase I using additional data from phase II
  - BD: before deletion
- Criterion: RMSE, empirical bias and coverage probability

Simulation studies

		MCAR					MAR					MNAR			
		μ	β_0	β_z	β_x		μ	β_0	β_z	β_x		μ	β_0	β_z	β_x
Bias*10,000	BD	8	3	3	4		-13	-2	4	-12		23	0	-5	3
	CC1	3	-4	1	2		-6063	-15	-14	-12		7983	2833	-1086	-1075
	CC2	2	-2	2	1		-4044	-4	-3	-7		5302	1667	-502	-507
	MI1	3	-2	-1	1		-25	-15	-12	-10		2856	2834	-1087	-1091
	MI2	-1	-4	1	-1		-14	-3	-2	-6		1699	1669	-504	-509
	NSMI	-1	-4	7	-2		3	16	7	2		52	20	-192	-181
RMSE*10000	BD	603	305	319	320		614	324	349	322		612	312	343	330
	CC1	701	355	388	388		6099	411	451	420		8008	2860	1163	1153
	CC2	670	341	368	365		4097	376	410	382		5343	1708	635	641
	MI1	633	361	395	395		680	426	462	435		2919	2863	1167	1152
	MI2	620	345	368	368		673	382	416	387		1818	1712	639	644
	NSMI	703	439	435	431		733	469	483	466		717	448	511	493
Coverage*100	BD	94.3	96.0	96.2	95.9		94.3	94.1	93.8	96.2		94.6	94.9	94.2	95.2
	CC1	95.5	96.3	95.3	94.8		0.0	95.4	93.5	95.4		0.0	0.0	25.6	25.0
	CC2	95.5	96.0	95.4	95.4		0.0	94.6	94.2	96.3		0.0	0.5	71.7	73.6
	MI1	95.2	96.4	95.1	94.6		94.6	95.7	93.5	94.9		0.6	0.0	28.4	28.0
	MI2	95.1	96.0	95.4	94.6		94.8	95.0	94.2	96.2		26.0	0.7	72.1	73.8
	NSMI	94.5	95.6	94.9	94.8		95.3	95.1	94.4	95.8		94.5	94.9	92.1	92.6

Simulation studies

- Missing data generation and two-phase sampling
- Phase I
  - MNAR: $\Pr\left(M_i = 1 \mid z_i, x_i, y_i\right) = \text{expit}\left(-y_i\right).$
- Phase II: $\Pr(R_{2\bullet 1,i} = 1 \mid M_i = 1; z_i, x_i, y_i) = \pi$

  π=0.05, 0.15, 0.25, 0.50

Simulation studies

		π=.05				π=.15				π=.25				π=.50			
		μ	β_0	β_z	β_x	μ	β_0	β_z	β_x	μ	β_0	β_z	β_x	μ	β_0	β_z	β_x
Bias	BD	48	-1	3	15	34	6	23	1	23	0	-5	3	-7	6	15	-15
	CC1	7980	2831	-1066	-1070	8019	2854	-1068	-1090	7983	2833	-1086	-1075	7965	2835	-1078	-1089
	CC2	7383	2537	-911	-914	6317	2066	-668	-689	5302	1667	-522	-507	3128	929	-221	-235
	MI1	2869	2824	-1066	-1066	2880	2854	-1069	-1090	2856	2834	-1087	-1071	2822	2832	-1078	-1085
	MI2	2582	2534	-910	-914	2080	2066	-666	-687	1699	1669	-524	-509	921	930	-221	-236
	NSMI	57	6	-207	-201	16	1	-165	-193	52	20	-192	-181	-2	6	-113	-125
RMSE	BD	616	327	334	336	612	320	339	340	612	312	343	330	612	312	323	321
	CC1	8005	2858	1145	1153	8045	2882	1145	1168	8008	2860	1163	1153	7991	2862	1155	1159
	CC2	7411	2567	1000	1009	6351	2103	776	799	5343	1708	655	641	3192	990	422	432
	MI1	2933	2853	1148	1153	2945	2885	1148	1171	2919	2863	1167	1152	2887	2861	1159	1159
	MI2	2664	2566	1001	1011	2179	2104	775	800	1818	1712	659	644	1118	992	424	433
	NSMI	1111	950	915	928	782	562	559	590	717	448	511	493	652	359	391	404
Coverage	BD	94.1	93.7	94.7	94.3	94.5	94.8	94.6	94.6	94.6	94.9	94.2	95.2	95.0	95.3	95.4	95.1
	CC1	0.0	0.0	25.7	25.9	0.0	0.0	25.3	24.3	0.0	0.0	25.6	25.0	0.0	0.0	23.7	23.1
	CC2	0.0	0.0	38.3	37.8	0.0	0.1	61.6	59.0	0.0	0.5	71.7	73.6	0.1	22.5	90.1	91.4
	MI1	0.2	0.0	29.8	30.6	0.6	0.0	27.9	27.9	0.6	0.0	28.4	28.0	0.1	0.0	27.3	26.8
	MI2	2.5	0.0	40.6	39.9	9.9	0.3	61.5	59.8	26.0	0.7	72.1	73.8	68.2	23.8	90.5	91.8
	NSMI	94.3	94.1	93.7	93.4	94.8	94.9	93.2	92.4	94.5	94.9	92.1	92.6	95.2	95.1	94.0	94.3

Application: Quality of Life

- Subjects: 750 young adults
- QOL assessed by the quality of life instrument for young adults (YAQOL)
  - Resources, relationship quality, and positive outlook
- Phase I: 603 out of 750 completed QOL survey
- Phase II: 39 out of the 147 nonrespondents were contacted and provided data on an abridged QOL instruments

Application

Quality of Life – Resources subscale

		CC1				CC2				IL1				IL2				NSMI			
Outcome		Est.	S.E.	LCL	UCL	Est.	S.E.	LCL	UCL	Est.	S.E.	LCL	UCL	Est.	S.E.	LCL	UCL	Est.	S.E.	LCL	UCL
	Mean	77.26	0.69	75.9	78.62	77.2	0.67	75.89	78.51	77.03	0.72	75.6	78.46	77.09	0.67	75.77	78.42	77.07	0.71	75.67	78.47
Regression																					
	Intercept	62.09	6.07	50.19	73.99	63.85	5.9	52.28	75.43	62.32	6.15	50.19	75.46	64.64	5.73	53.4	75.88	66.84	5.64	55.77	77.91
	Sex (male vs. female)	-2.48	1.38	-5.19	0.23	-2.7	1.33	-5.3	-0.1	-2.44	1.31	-5.02	0.14	-2.27	1.41	-5.07	0.52	-2.35	1.28	-4.87	0.16
	Race (White vs. non-White)	5.64	2.47	0.8	10.49	5.35	2.36	0.72	9.98	5	2.44	0.19	9.81	4.84	2.25	0.43	9.26	3.46	2.31	-1.07	7.99
	Education (≥HS* vs. < HS)	1.83	1.45	-1.02	4.68	1.61	1.4	-1.14	4.36	1.71	1.45	-1.14	4.56	2.04	1.33	-0.57	4.66	1.74	1.32	-0.85	4.34
	Age	0.46	0.26	-0.06	0.97	0.39	0.25	-0.1	0.89	0.47	0.28	-0.08	1.02	0.36	0.25	-0.13	0.84	0.33	0.24	-0.15	0.8

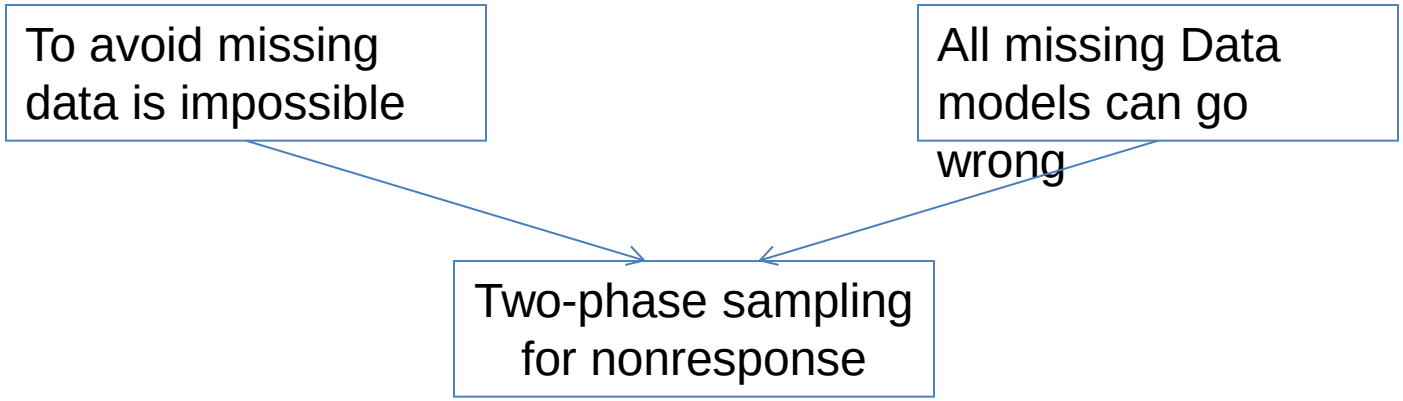
Discussion

- Two-phase sampling after 5 decades
  - Little research done to show benefit
- Traditional Methods fails to make full use of data
- Nonrespondent subsample multiple imputation
- Practical considerations
  - Abridged version, incentives, tailoring etc

Discussions

- Limitations
  - No gain if phase I is MCAR
  - Substantial bias if phase II is MNAR
- Cost-effectiveness
  - Trade-off between response rate and recruiting more subjects
- Repeated attempt design
  - Selection model
  - Pattern mixture model

Conclusion



Nonrespondent subsample multiple imputation