

MULTIPLE IMPUTATION FOR SURVEY DATA ANALYSIS

Yeats Ye, University of Maryland at College Park, MD

ABSTRACT

Missing values in survey data may lead to biased parameter estimates that do not accurately represent the population.

There are a variety of techniques to deal with survey missing data. For instance, weighting adjustments are used in some studies to compensate for biased estimators produced by survey non-response (1). Some conventional imputation methods such as model-based imputation (mean, ratio and regression imputation) are still used to replace missing values with acceptable values (2).

However, adjusted weights need to be created for each variable with item non-response. This requires a great deal of time and effort (2). Conventional imputation methods focus on estimating mean parameter values and ignore the between-imputation component of variability. Such imputation methods result in smaller standard errors than if variability were taken into account, thus increasing the likelihood of statistical significance. Instead of filling in a single value for each missing value in survey data, more advanced multiple imputation (MI) can replace each missing value with a set of plausible values that represent the uncertainty about the right value to impute.

Methods for multiple imputation are described in detail by Dr. Allison in his book as well as the materials for his course on multiple imputation. These methods do not describe how to incorporate the fact that many of our large scale data sets use complex clustered sampling designs. As such, simple random sample methods, even with sample weights, may be inappropriate.

This paper will demonstrate and discuss how we use the procedures MI, MIanalysis and Surveylogistic in SAS(r) to perform analyses with these three cumulative logit models using complex survey data.

INTRODUCTION

Proc Surveylogistic fits linear logistic regression models for discrete response survey data by the maximum likelihood method and incorporates the sample design into the analysis.

By default, an observation is excluded in the analysis if it has a missing value for any variable, including the following variables: Weight, STRATA and CLUSTER.

Proc MI and Proc MIanalysis are multiple imputation procedures. When Proc MI is used, missing values are imputed multiple times to generate multiple complete data sets. Each imputed data set can be analyzed by a SAS standard procedure to estimate a desired model. The results from the complete data sets are combined to produce inferential results by Proc MIanalysis. This method reflects the uncertainty about the predictions of the unknown missing values, and the resulting estimated variances of the parameter estimates will be unbiased.

DATA SOURCE

The data set is the Early Childhood Longitudinal Survey, Kindergarten Cohort - Third Grade Longitudinal Public Use data file. Almost 20,000 (19,684) children were included in the data. The children were initially interviewed in kindergarten during the 1998-99 school year (Wave 1 and Wave 2). A 30% subsample of children was re-interviewed in the first grade (Wave 3) but all were followed up in the Spring of kindergarten (Wave 4) and the Spring of third grade (Wave 5).

The purpose of this paper is to test three cumulative logit models to see how child financial support from and contact with a nonresidential biological father are associated with children's physical health following the separation of their parents.

A subset of 1,765 children who fit the following criteria was selected: the child was living with his/her biological mother and without his/her biological father in kindergarten (K); the child's biological father was still alive when he/she was in the third grade.

Dependent variables:

health (Child's health was reported by the mother at K)
support (Degree of contact with child at K)
contact (Degree of support to child at K)

Independent variables:

meduc2 (Maternal education levels at W2)
gender (Child's sex)
rblack (Race black)
rhisp (Race hisp)
rother (Race other)
lnincome (Logged, (income+1))
i2ins (Health insurance coverage at K)
marital (Mother's marital status at K.)
mempl2 (Maternal employment at K)
depres2 (Mother feels depressed at K)
bwp (Birth weight in pounds)

STEPS FOR MI

1. Check data using PROC MI with option nimpute = 0.

```
proc mi data = sasuser.missing
(keep = health1 support2 contact2 meduc2 gender rblack rhisp rother lnincome i2ins marital mempl2 depres2 bwp
YRDADLIV )
Nimpute = 0;
```

Before we perform logistic regressions, we need to check missing patterns using Proc MI with nimpute = 0. The variable Yrdadliv (Number of years biodad lived with child) is an auxiliary variable. It is only for imputation and not in the analytical model. Missing values were found as i2ins(1), marital(2), mempl2(17), depres2(34), bwp(38), minaw(349) and Yrdadliv (8). There are no missing values on dependent variables.

2. Impute data using PROC MI

```
proc mi data = sasuser.missing out =sasuser.imped seed = 123 nimpute = 3;
var health1 support2 contact2 meduc2 gender rblack rhisp rother lnincome i2ins marital mempl2 depres2 bwp
minaw c1_5fp0 Yrdadliv ;
run;
```

Three dependent variables, all predictors, an auxiliary variable, and a sample weight variable were included in the PROC MI procedure because leaving out the dependent variables would cause bias (3).

The MCMC method in Proc MI, by default, uses a single chain to produce five imputations. It also completes 200 burn-in iterations before the first imputation and 100 iterations between imputations (4). The number of iterations above can be changed by the user.

Proc MI was forced to produce the same results every time by the option Seed = n. Three complete data sets were generated after the imputation process since the option of the number of imputed data sets was changed to 3 in our code.

3. Analyze data using Proc Surveylogistic with by _imputation_;

```
*run model 1 using proc surveylogistic;
proc surveylogistic data = sasuser.imped;
stratum c15fpstr;
cluster c15fpps;
weight c1_5fp0;
model health1(descending) = meduc2 gender rblack rhisp rother lnincome i2ins marital depres2 bwp / link =
clogit covb expb; *EXPB: displays the exponentiated coefficients (i.e., the odds ratios);
*COVB displays the covariance matrix of the parameter estimates);
by _imputation_;
ods output parameterestimates = model_1imp covb = covmat1;
run;

*modify the SAS data of Model_1imp to avoid error message. Because variable of Intercept is not on the COVB =
data set;
data model_1imp;
set model_1imp(rename = (variable = var));
if classval0 ne . then variable = var||'_'||classval0;
```

```

else variable = var;
drop var;
run;

*combine results into a data set of parameter estimates, standard errors and so on;
proc mianalyze parms = model_1imp covb = covmat1;
modeleffects intercept_4 intercept_3 intercept_2 meduc2 gender rblack rhisp rother lnincome i2ins marital depres2
bwp ;
ods output parameterestimates = comb_health ;
run;

*****
;

*run model 2 (cumulative logit model with a ordinal response variable );
proc surveylogistic data = sasuser.imped;
stratum c15fpstr;
cluster c15fpps;
weight c1_5fp0;
model contact2(descending) =
meduc2 gender rblack rhisp rother lnincome i2ins marital / link = clogit covb expb;
by _imputation_;
ods output parameterestimates = model_2imp covb = covmat2;
run;

*modify the SAS data of Model_2imp to avoid error message. Because variable of Intercept is not on the COVB =
data set;
data model_2imp;
set model_2imp(rename = (variable = var));
if classval0 ne . then variable = var||'_'||classval0;
else variable = var;
drop var;
run;

*combine results into a data set of parameter estimates, standard errors and so on;
proc mianalyze parms = model_2imp covb = covmat2 mult;
modeleffects intercept_3 intercept_2 meduc2 gender rblack rhisp rother lnincome i2ins marital minaw;
ods output parameterestimates = comb_contact ;
run;

*****
;

*run model 3 (generalized logits mode with a nominal response variable );
proc surveylogistic data = sasuser.imped;
stratum c15fpstr;
cluster c15fpps;
weight c1_5fp0;
model support2 (descending) =
meduc2 gender rblack rhisp rother lnincome i2ins marital minaw mempl2/ link = clogit covb expb; /*Specifies link
function*/
by _imputation_;
ods output parameterestimates = model_3imp covb = covmat3;
run;

*modify the SAS data of Model_1imp to avoid error message. Because variable of Intercept is not on the COVB =
data set;
data model_3imp;
set model_3imp(rename = (variable = var));
if classval0 ne . then variable = var||'_'||classval0;

```

```

else variable = var;
drop var;
run;

```

**combine results into a data set of parameter estimates, standard errors and so on;*

```

proc mianalyze parms = model_3imp covb = covmat3 mult;
modeleffects intercept_2 intercept_1 meduc2 gender rblack rhisp rother lnincome i2ins marital mempl2;
ods output parameterestimates = comb_support;
run;

```

ods output close;

Analysis of three cumulative logit models was performed with an ordinal response variable in each model using the By _imputation_ statement in the Surveylogistic procedure. Variable Health of the 1st model is children's health, which was reported by the mothers of the children while they were in kindergarten [K(1 = Reg/Poor, 2 = Very Good, 3 = Good and 4 = Excellent)]; variable Contact of the 2nd model is Biodad Contact at W2 (1 = Contact during last month, 2 = Last contact between 1 year and 1 month, and 3 = No contact during last year); and variable Support of the 3rd model is Biodad Supports at W2 (0 = No support, 1 = Awarded support & 2 = Reg support). After the above analysis was completed, two input data, i.e. parameter estimates listed in the PARMS = and covariance matrices listed in the COVB =, were used in Procedure MAnalyze. Output data set listed in ODS output was a combined result for model predictors.

3. Display results

Table 1		Model 1: Health							
variable	estimateOrg	estimateImp	oddsOrg	oddsImp	p_valueOrg	p_valueImp	StdErrOrg	StdErrImp	
intercept_4	-2.112636	-1.973146	0.120919	0.139019	0.0005	0.0008	0.605161	0.588582	
intercept_3	-0.599295	-0.496348	0.549199	0.60875	0.316	0.3938	0.597628	0.582055	
intercept_2	1.205236	1.38037	3.337545	3.976371	0.0496	0.0214	0.613731	0.599827	
meduc2	0.273599	0.279034	1.314687	1.321852	0.0273	0.0191	0.123923	0.119016	
gender	0.277191	0.25349	1.319418	1.288515	0.0209	0.0288	0.119964	0.115969	
rblack	-0.038541	-0.073046	0.962192	0.929558	0.8123	0.6338	0.162295	0.153354	
rhisp	-0.319446	-0.289385	0.726552	0.748724	0.0393	0.0582	0.155005	0.15278	
rother	0.071855	0.044421	1.074499	1.045422	0.7438	0.8286	0.219826	0.205173	
lnincome	0.052052	0.045414	1.053431	1.046461	0.2185	0.2509	0.042302	0.039553	
i2ins	0.040681	0.015293	1.041519	1.015411	0.8375	0.9356	0.198368	0.189403	
marital	0.27316	0.275603	1.314111	1.317325	0.1312	0.1142	0.180971	0.174458	
depres2	-0.162165	-0.175016	0.850301	0.839444	0.109	0.0886	0.101186	0.102491	
bwp	0.164909	0.159929	1.179285	1.173428	0.0002	0.0003	0.044855	0.044653	

Table 2		Model 2: Contact							
variable	estimateOrg	estimateImp	oddsOrg	oddsImp	p_valueOrg	p_valueImp	StdErrOrg	StdErrImp	
intercept_3	-1.568187	-1.560654	0.208423	0.209999	0.0041	0.0042	0.546231	0.545725	
intercept_2	-0.86041	-0.852168	0.422989	0.426489	0.1144	0.1178	0.54493	0.544881	
meduc2	0.291572	0.287957	1.33853	1.3337	0.0393	0.0412	0.141484	0.141028	
gender	0.185681	0.186946	1.204038	1.205562	0.0746	0.0716	0.104162	0.103776	
rblack	-0.117594	-0.118333	0.889057	0.8884	0.3804	0.3758	0.134049	0.133601	
rhisp	-0.722657	-0.725395	0.485461	0.484133	0.0001	0.0001	0.18233	0.182107	
rother	-0.504334	-0.506539	0.603908	0.602578	0.0068	0.0065	0.186445	0.186177	
lnincome	0.181725	0.181008	1.199285	1.198425	0.0005	0.0005	0.052218	0.052279	
i2ins	0.077679	0.080548	1.080776	1.08388	0.6338	0.6207	0.163045	0.162754	
marital	-0.823934	-0.825536	0.438702	0.438	0	0	0.132539	0.132582	

Table 3

Model 3: Support

variable	estimateOrg	estimateImp	oddsOrg	oddsImp	p_valueOrg	p_valueImp	StdErrOrg	StdErrImp
intercept_2	-2.60378	-2.556361	0.073993	0.077587	0	0	0.507045	0.476675
intercept_1	-1.241161	-1.184404	0.289048	0.305928	0.0142	0.0126	0.506343	0.474948
meduc2	0.262398	0.244243	1.300044	1.276655	0.0106	0.0166	0.102677	0.101952
gender	-0.009269	-0.015959	0.990774	0.984167	0.9383	0.8896	0.119813	0.114975
rblack	-0.929435	-0.935619	0.394777	0.392343	0	0	0.150251	0.149121
rhisp	-1.202831	-1.188859	0.300343	0.304568	0	0	0.166197	0.165025
rother	-0.894341	-0.904335	0.408877	0.404811	0.0001	0.0001	0.226146	0.231384
lnincome	0.11383	0.108904	1.120562	1.115056	0.0205	0.0177	0.049114	0.045896
i2ins	0.555965	0.549793	1.743623	1.732895	0.0013	0.0017	0.172823	0.174889
marital	0.071463	0.068944	1.074078	1.071376	0.6618	0.6717	0.163387	0.162655
mempl2	0.430604	0.444683	1.538186	1.559996	0.0018	0.0013	0.13815	0.138627

Two data sets were finally combined together as the results of the analysis of cumulative logit models using data steps for each model, of which one data set was produced from unimputed data while the other from imputed data.

```

data temp(keep = model variable estimate p_value stderr df lclmean uclmean waldchisq parm);
retain model "Original Data";
set model_1 comb_1;
if ProbChiSq = " " then p_value = round(Probt, 0.0001);
else p_value = round(ProbChiSq, 0.0001);
if variable = " " then model = "Imputed Data" ;
run;
data sasuser.compare_health;
retain model variable estimate odds p_value;
set temp;
if variable = " " then variable = parm ;
odds = exp(estimate) ;
drop parm;
run;

```

CONCLUSION

Multiple imputation can be used with any data and model. Proc MI and Proc MIanalysis are easy to learn and to use with SAS standard procedures such as Proc REG, Proc Logistic, and Proc GLM. Attention should be given to the fact that the data set listed in Covb = , which was created by the Surveylogistic procedure using the imputed data sets, may not include intercept variables. One needs to modify it in order to avoid an error message when Proc MIanalyze with option Covb = is used.

The results from studies reported here, which were analyzed by Proc Surveylogistic with imputed data, make much more sense than those with unimputed data.

TRADEMARKS

SAS(r) and all other SAS(r) Institute Inc. products or service names are registered trademarks or trademarks of SAS® institute Inc. in the USA and other countries. (r) indicates USA registration. Other brand and product names are registered trademarks or trademarks of their respective companies.

REFERENCES

1. Pike, G.R. (2007). Using Weighting Adjustments to Compensate for Survey Nonresponse (Paper presented at the annual meeting of the Association for Institutional Research, Kansas City, Missouri, June 2007)
2. Smit, J. (2009). Imputation of business survey (Data Workshop on Informal Employment and Informal Sector Data Analysis, Tabulations and Country Reports, Bangkok)
3. Allison, P.D. (2001) "Missing Data." *Sage University Papers in Quantitative Applications in the Social Sciences*, 07136. Thousand Oaks, CA: Sage and materials from Dr. Allison's course, Missing Data (2008)
4. SAS OnlineDoc documentation for SAS version 9.1.

ACKNOWLEDGEMENTS

I would like to thank Professor Sandra Hofferth, Director of the Maryland Population Research Center, for her help. I would also like to thank Mr. Rob Agnelli from SAS Technical Support for his assistance.

CONTACT INFORMATION

All comments, questions, and inquiries can be sent to:

Yeats Ye
Statistical Analysis Coordinator
Maryland Population Research Center
University of Maryland at College Park
0124K Building#162
College Park, Maryland 20742
Tel: 301.405.1858
Fax: 301.405.5743
Email: hye@umd.edu