Faculty of Science
CHARLES UNIVERSITY IN PRAGUE

Handling missing data when addressing differential cohort fertility by education

Jitka Rychtaříková

13.2. 2014

5th Demographic Conference of "Young Demographers„
Actual Demographic Research of Young Demographers (not only) in Europe

Faculty of Science
CHARLES UNIVERSITY IN PRAGUE

IMPUTATION

Single imputation substitutes a value for each missing value. For example, each missing value can be **imputed from the variable mean** of the complete cases.

Multiple imputation replaces each missing value with a set of plausible values. Multiple imputation does not attempt to estimate each missing value through simulated values, but rather to represent a **random sample of the missing values.** This process results in valid statistical inferences that properly reflect the uncertainty due to missing values; for example, valid confidence intervals for parameters.

Missing Completely at Random **(MCAR):** ***no patterns*** **in missingness**; are not related (or correlated) to any variables in the study.

*Missing at Random **(MAR):** missingness on a variable is related to other variables in your analysis. Missingness on a variable **can be predicted by other variables.**

Not Missing at Random **(NMAR or MNAR):** Missingness is related to the variable itself. **Must create a separate model that accounts for missing data.**

Multiple imputation inference involves three distinct phases:

The missing data are filled in m times to generate m complete data sets. **Proc MI**

The m complete data sets are analyzed by using standard procedures. **Proc GENMOD** by _imputation_

The results from the m complete data sets are combined for the inference. **Proc MIANALYZE**

Faculty of Science
CHARLES UNIVERSITY IN PRAGUE

Missing Data Patterns

A data set with variables is said to have a ***monotone missing pattern*** when the event that a variable Y_j is missing for a particular individual implies that all subsequent variables Y_k (k>j), are missing for that individual.

Monotone Missing Data Patterns

Group	Y1	Y2	Y3
1	X	X	X
2	X	X	.
3	X	.	.

Non-monotone Missing Data Patterns

Group	Y1	Y2	Y3
1	X	X	X
2	X	.	X
3.	X	X	.
4.	X	.	X

A data set with an *arbitrary missing pattern* is a data set with either a monotone missing pattern or a non-monotone missing pattern.

Faculty of Science
CHARLES UNIVERSITY IN PRAGUE

Data: census of the Czech Republic 2011

female birth cohorts 1926-1976: 3 164 617

children	Frequency	Percent	Cumulative Frequency
0	205 614	6,6	205 614
1	584 268	18,8	789 882
2	1 670 355	53,7	2 460 237
3	505 820	16,3	2 966 057
4+	146 704	4,7	3 112 761
Frequency Missing = 51 856			1,64 %

education	Frequency	Percent	Cumulative Frequency
basic	663 194	21.68	663 194
vocational	1 018 590	33.29	3 059 560
secondary	1 049 283	34.30	1 712 477
tertiary	328 493	10.74	2 040 970
Frequency Missing = 10 5057			3,32 %

marital_status	Frequency	Percent	Cumulative Frequency
divorced	554 399	17.54	554 399
married	1 866 917	59.07	2 421 316
single	196 418	6.22	2 617 734
widowed	542 639	17.17	3 160 373
Frequency Missing = 4 244			0,13 %

community size	Frequency	Percent	Cumulative Frequency
1999-	817 536	25.83	817 536
2000-4999	368 028	11.63	1 185 564
5000-9999	285 676	9.3	1 471 240
10000-19999	295 214	9.33	1 766 454
20000-49999	417 765	13.20	2 184 219
50000-99999	276 359	8.73	2 460 578
100000+	704 039	22.25	3 164 617

Faculty of Science
CHARLES UNIVERSITY IN PRAGUE

Missing Data Patterns: arbitrary

	generation	com_size	children	education	marital_status
1	X	X	X	X	X
2	X	X	X	X	.
3	X	X	X	.	X
4	X	X	X	.	.
5	X	X	.	X	X
6	X	X	.	X	.
7	X	X	.	.	X
8	X	X	.	.	.

```
proc mi data=z nimpute=0;
class children education marital_status;
fcs;
var generation com_size children education marital_status;
run;
```

Faculty of Science
CHARLES UNIVERSITY IN PRAGUE

Imputation Methods PROC MI

The imputation method of choice depends on the patterns of missingness in the data and the type of the imputed variable.

FCS: fully conditional specification method that assumes the existence of a joint distribution for all variables.

Pattern of	Type of	Type of	Available Methods
Missingness	Imputed Variable	Covariates	
Monotone	Continuous	Arbitrary	Monotone regression
			Monotone predicted mean matching
			Monotone propensity score
Monotone	Classification (ordinal)	Arbitrary	Monotone logistic regression
Monotone	Classification (nominal)	Arbitrary	Monotone discriminant function
Arbitrary	Continuous	Continuous	MCMC full-data imputation
			MCMC monotone-data imputation
Arbitrary	Continuous	Arbitrary	FCS regression
			FCS predicted mean matching
Arbitrary	Classification (ordinal)	Arbitrary	FCS logistic regression
Arbitrary	**Classification (nominal)**	**Arbitrary**	**FCS discriminant function**

There are two major iterative methods for doing multiple imputation for general missing data patterns:
the **Markov chain Monte Carlo (MCMC)** method and
the **fully conditional specification (FCS)** method.

The first "burn-in" iterations are designed to ensure that the algorithm has converged to the correct posterior distribution.

```
proc mi data=z seed=120214 out=zz;
   class children education marital_status;
   fcs nbiter=10 discrim(children/details)
discrim(education/details) discrim(marital_status/details);
   var generation com_size children education marital_status;
run;
```

The MIANALYZE procedure reads parameter estimates and associated standard errors or covariance matrix that are computed by the standard statistical procedure for each imputed data set.

The MIANALYZE procedure then derives valid univariate inference for these parameters. With an additional assumption about the population between and within imputation covariance matrices, multivariate inference based on Wald tests can also be derived.

The MODELEFFECTS statement lists the effects to be analyzed, and the CLASS statement lists the classification variables in the MODELEFFECTS statement.

Faculty of Science
CHARLES UNIVERSITY IN PRAGUE

Poisson regression *is a form of regression analysis used to model count data and contingency tables. Poisson regression assumes the response variable Y has a Poisson distribution, and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters. A Poisson regression model is sometimes known as a log-linear model.*

```
proc genmod data=zz;
class generation (ref="1950") com_size(ref="19999") education
(ref="secondary") marital_status (ref="married");
model child= generation com_size education
marital_status/dist=poisson covb;
by _imputation_;
ods output ParameterEstimates=gmparms;
run;
```

```
proc mianalyze parms(classvar=level)=gmparms ;
class generation com_size education marital_status;
   modeleffects Intercept generation com_size education
marital_status;
run;
```

Faculty of Science
CHARLES UNIVERSITY IN PRAGUE

Parameter Estimates						
Parameter		**Estimate**	**Std Error**	**95% Confidence Lim**		**Pr > |t|**
Intercept		0.681030	0.002902	0.67534	0.68672	<.0001
generation	1926	-0.089135	0.005277	-0.09948	-0.07879	<.0001
generation	1927	-0.087296	0.005096	-0.09728	-0.07731	<.0001
generation	1928	-0.076288	0.004880	-0.08585	-0.06672	<.0001
generation	1929	-0.075685	0.004764	-0.08502	-0.06635	<.0001
generation	1930	-0.074153	0.004575	-0.08312	-0.06519	<.0001
generation	1931	-0.075368	0.004537	-0.08426	-0.06648	<.0001
generation	1932	-0.072517	0.004468	-0.08127	-0.06376	<.0001
generation	1933	-0.071463	0.004485	-0.08025	-0.06267	<.0001
generation	1934	-0.069671	0.004453	-0.07840	-0.06094	<.0001
generation	1935	-0.073993	0.004448	-0.08271	-0.06528	<.0001
generation	1936	-0.063657	0.004418	-0.07232	-0.05500	<.0001
generation	1937	-0.060431	0.004377	-0.06901	-0.05185	<.0001
generation	1938	-0.059635	0.004275	-0.06801	-0.05126	<.0001
generation	1939	-0.056141	0.004192	-0.06436	-0.04792	<.0001
generation	1940	-0.051447	0.004034	-0.05935	-0.04354	<.0001
generation	1941	-0.053786	0.004001	-0.06163	-0.04594	<.0001
generation	1942	-0.041891	0.003937	-0.04961	-0.03417	<.0001
generation	1943	-0.038683	0.003793	-0.04612	-0.03125	<.0001
generation	1944	-0.034706	0.003753	-0.04206	-0.02735	<.0001
generation	1945	-0.025360	0.003782	-0.03277	-0.01795	<.0001
generation	1946	-0.015208	0.003575	-0.02221	-0.00820	<.0001
generation	1947	-0.014193	0.003542	-0.02113	-0.00725	<.0001
generation	1948	-0.005386	0.003569	-0.01238	0.00161	0.1312
generation	1949	-0.002900	0.003589	-0.00993	0.00413	0.4191
generation	1950	0	0	.	.	.

Faculty of Science
CHARLES UNIVERSITY IN PRAGUE

Parameter		Estimate	Std Error	95% Confidence Lim		Pr > |t|
generation	1951	-0.002655	0.003548	-0.00961	0.00430	0.4543
generation	1952	-0.004004	0.003562	-0.01099	0.00298	0.2610
generation	1953	-0.004547	0.003586	-0.01158	0.00248	0.2048
generation	1954	-0.006066	0.003598	-0.01312	0.00099	0.0918
generation	1955	-0.006147	0.003610	-0.01322	0.00093	0.0886
generation	1956	-0.004078	0.003623	-0.01118	0.00302	0.2604
generation	1957	-0.005505	0.003667	-0.01269	0.00168	0.1333
generation	1958	-0.007291	0.003750	-0.01464	0.00006	0.0519
generation	1959	-0.004041	0.003852	-0.01159	0.00351	0.2942
generation	1960	-0.004809	0.003856	-0.01237	0.00275	0.2123
generation	1961	-0.005873	0.003833	-0.01339	0.00164	0.1255
generation	1962	-0.009705	0.003822	-0.01720	-0.00221	0.0111
generation	1963	-0.015968	0.003727	-0.02327	-0.00866	<.0001
generation	1964	-0.019917	0.003706	-0.02718	-0.01265	<.0001
generation	1965	-0.025663	0.003758	-0.03303	-0.01830	<.0001
generation	1966	-0.025494	0.003803	-0.03295	-0.01804	<.0001
generation	1967	-0.029344	0.003830	-0.03685	-0.02184	<.0001
generation	1968	-0.032460	0.003843	-0.03999	-0.02493	<.0001
generation	1969	-0.033717	0.003800	-0.04116	-0.02627	<.0001
generation	1970	-0.040043	0.003770	-0.04743	-0.03265	<.0001
generation	1971	-0.040719	0.003743	-0.04805	-0.03338	<.0001
generation	1972	-0.046934	0.003705	-0.05419	-0.03967	<.0001
generation	1973	-0.057153	0.003638	-0.06428	-0.05002	<.0001
generation	1974	-0.071997	0.003608	-0.07907	-0.06493	<.0001
generation	1975	-0.098544	0.003664	-0.10572	-0.09136	<.0001
generation	1976	-0.128092	0.003745	-0.13543	-0.12075	<.0001

Parameter		Estimate	Std Error	95% Confidence Lim		Pr > |t|
com_size	100000	-0.104478	0.001634	-0.10768	-0.10127	<.0001
com_size	1999	0.063288	0.001540	0.06027	0.06631	<.0001
com_size	19999	0	0	.	.	.
com_size	4999	0.033230	0.001768	0.02976	0.03669	<.0001
com_size	49999	-0.009640	0.001744	-0.01306	-0.00622	<.0001
com_size	9999	0.023046	0.001882	0.01936	0.02673	<.0001
com_size	99999	-0.029808	0.001933	-0.03360	-0.02602	<.0001
education	basic	0.170226	0.001191	0.16789	0.17256	<.0001
education	secondary	0	0	.	.	.
education	tertiary	-0.037367	0.001533	-0.04037	-0.03436	<.0001
education	vocational	0.075455	0.001018	0.07346	0.07745	<.0001
marital_status	divorced	-0.049648	0.001121	-0.05185	-0.04745	<.0001
marital_status	married	0	0	.	.	.
marital_status	single	-1.030322	0.002901	-1.03601	-1.02464	<.0001
marital_status	widowed	0.011248	0.001275	0.00875	0.01375	<.0001

Faculty of Science
CHARLES UNIVERSITY IN PRAGUE

Thank you for your attention

Jitka Rychtaříková

rychta@natur.cuni.cz