

MI DATA ANALYSIS

Imputing missing data from the Australian Longitudinal Study of Health and Relationships using (m)ICE

Jason Ferris

Focus:

- To illustrate the 'ease' of using Stata and two user written programs - **ICE** (Royston, 2005a; 2005b) - & - **MIM** (Carlin et al, 2008). Data on men's Chronic Pelvic Pain (CPP; Pitts et al., 2007) from the Australian Longitudinal Study of Health and Relationships (ALSHR) was used to demonstrate multiple imputation (MI) for cross-sectional data and MI for longitudinal data
- The outcome of interest in this example is non-specific CPP
- ICE** - Uses *Multivariate Imputation by Chain Equations* and imputation method proposed by van Buuren et al (1999) to impute missing data for *m* datasets.
- MIM** - allows manipulation and analysis of *m* imputed datasets that are *stacked* in one data file.

Missing Data and Imputation:

- The presence of missing data is likely to lead to bias
- Is likely to lead to a reduction of power (due to reduce sample size)
- The treatment of missing data is often avoided and analysis is typically complete-case only (i.e., a 'case' is dropped if any variable in a model has missing data)
- single imputation methods (e.g., mean or mode) is considered inappropriate unless the amount of missing data is notably small - however why use it if MI methods are apt and available
- Most statistical software now have facilities to deal with missing data: i.e., hot decking, Markov chain Monte Carlo, Multiple Imputation
- To use MI techniques it is assumed missing data is at least missing at random (MAR) - although this is often more difficult to assume with longitudinal data
- The goal of MI is not to *accurately* infer the missing values but to *accurately* infer population estimands

Cross-sectional (data from panel 1 only):

Stata code:

```
ice qhem13 experience exp2 exp3 qst13 expXuti eXu2 eXu3 qhem07 qhem10 rage using cpp_single, m(5) p
assive(expXuti:experience*qst13 \exp2:(experience==2) \exp3:(experience==3) \eXu2:(expXuti==2) \
eXu3:(expXuti==3)) sub(experience: exp2 exp3, expXuti: eXu2 eXu3) boot() seed(190608) replace
```

- User intervention is necessary to make sure that the code above adequately reflects what imputations are required
- It is recommended that the variables used for imputation 'match-up' with the model under analysis. Although it is also suggested to include variables that are known to be well correlated with the variables containing missing values
- Indicator expansion for interaction terms or categorical variables (that is i.) can not be used as such terms need to be manually created (e.g. gen expXuti=experience*qst13; tab expXuti, gen(eXu)) and explicitly defined in the options (see passive and substitution).
- Other options such as specifying the type of predictor variable or prediction models are also available
- Bootstrapping - boot() - is used to compensate for the possibility that the joint distribution of variables are not multivariate normal.

OR for CPP - using non-imputed raw data (complete case) and imputed data

	Complete Case: n=4235					Imputed data: m=5, n=4290				
	OR	SE	p	CI _{low}	CI _{upp}	OR	SE	p	CI _{low}	CI _{upp}
opposite sex (OS)	0.58	0.122	0.010	0.39	0.88	0.61	0.128	-2.340	0.02	0.41
neither sex (NS)	0.35	0.191	0.055	0.12	1.02	0.38	0.203	-1.810	0.07	0.13
UTI	0.73	0.368	0.528	0.27	1.96	0.75	0.379	-0.570	0.57	0.28
OSxUTI	2.58	1.364	0.074	0.91	7.27	2.46	1.297	1.710	0.09	0.88
NSxUTI	10.56	15.56	0.110	0.59	190.03	10.25	15.13	1.580	0.12	0.56
urination pain	3.18	0.591	0.000	2.21	4.58	3.11	0.565	6.230	0.00	2.17
dyspareunia	2.58	0.510	0.000	1.75	3.80	2.57	0.506	4.800	0.00	1.75
age	1.07	0.029	0.009	1.02	1.13	1.08	0.029	2.750	0.01	1.02
age ²	1.00	0.000	0.010	1.00	1.00	1.00	0.000	-2.670	0.01	1.00

Complete case: xi3: svy: logistic qhem13 i.experience*qst13 qhem07 qhem10 rage ragesq if _mj==0

Imputed data: xi3: mim: svy: logistic qhem13 i.experience*qst13 qhem07 qhem10 rage ragesq

Cross-sectional results:

- Data was imputed for 55 missing cells
- In both datasets survey weights were used and accounted for
- Differences between the two approaches are minimal; given the few missing cases across the covariates at intake panel.
- The imputed results (ORs and SE's) are based on aggregated results across the 5 imputed datasets (each with n=4290)

Data:

- ALSHR began in 2005
- Panel 1: 4290 men where interviewed
- Panel 2: 3148 interviewed (LTFU: 23%); 153 were not available at panel 2 but were at panel 3
- Panel 3: 2567 interviewed (LTFU: 22%)
- Among other variables univariate analysis suggested the following were associated with CPP: sexual experience (exp), sexual identity, a history of sexual coercion, a history of genital warts, Candida (thrush) or urinary tract infection (UTI), smoking and drinking. Due to the strong association between CPP and both urination pain and dyspareunia (sex pain) these are also included in the final model
- After model building the following logit model is used to demonstrate MI and missing data for the outcome variable non-specific CPP:

$$\ln(\text{CPP}) = b_0 + b_1 \text{exp}_1 + b_2 \text{exp}_2 + b_3 \text{uti} + b_4 \text{exp}_1 \text{Xuti} + b_5 \text{exp}_2 \text{Xuti} + b_6 \text{urine} + b_7 \text{dysp} + b_8 \text{age} + b_9 \text{age}^2$$

References:

- Carlin, J. B., Galati, J. C., & Royston, P. (2008). A new framework for managing and analyzing multiply imputed data in Stata. *Stata Journal*, 8(1), 49-67.
- Pitts, M., Ferris, J., Smith, A., Shelley, J., & Richters, J. (2008). Prevalence and Correlates of Three Types of Pelvic Pain in a Nationally Representative Sample of Australian Men. *The Journal of Sexual Medicine*, 5(5), 1223-1229.
- Royston, P. (2005a). Multiple imputation of missing values: update. *Stata Journal*, 5(2), 188-201.
- Royston, P. (2005b). Multiple imputation of missing values: update of ice. *Stata Journal*, 5(4), 527-536.

Longitudinal (data from panels 1 - 3):

Stata code:

```
(all information): ice qhem13* qst13* qhem07* qhem10* rage* if panels==1 | panels==3 using cpp_
multiple_p2, m(5) boot() seed(190608) replace
```

```
(complete cases): ice qhem13* qst13* qhem07* qhem10* rage* if panels==1 | panels==3 using cpp_
multiple_p2, m(5) boot() seed(190608) dropmissing replace
```

- To accommodate for within subject correlation between variables the dataset (usually in long form) is **reshaped** to wide form. Reshaping allows for responses on a particular variable across panels (i.e., response to sexual experience at times 1, 2 and 3) along with other variables of interest to attribute to the imputation of missing data for a particular time point for a particular variable
- Missing cases across panels - where a case (person under observation) is missing data for one panel but present for the bounding panels the information from the two panels was used to impute values for the missing variables from the missed panel. That is, missing panel data for panel 2 is imputed from variable information given at panel 1 and panel 3
- Once the imputation datasets are complete it is necessary to **reshape** the dataset back to long form
- The **MIM** options are now able to be implemented. For example generating a new variable age² can be done for all imputed datasets directly by **gen age=age^2**

OR for CPP - using all information across waves (all information) and complete case information (for CPP)

	All information (n=4290)						Complete cases (p1-p3; n=2567)					
	Non-imputed			Imputed			Non-imputed			Imputed		
	OR	SE	p	OR	SE	p	OR	SE	p	OR	SE	p
opposite sex (OS)	0.55	0.132	0.01	0.72	0.202	0.237	0.70	0.199	0.206	0.71	0.200	0.225
neither sex (NS)	0.33	0.164	0.03	0.68	0.371	0.475	0.68	0.383	0.495	0.67	0.369	0.468
UTI	1.82	1.077	0.32	1.90	1.254	0.329	2.00	1.334	0.300	1.88	1.238	0.338
OSxUTI	1.62	0.981	0.43	1.39	0.935	0.624	1.43	0.972	0.603	1.40	0.942	0.615
NSxUTI	6.91	15.02	0.37	6.13	12.17	0.361	6.46	13.12	0.358	6.20	12.29	0.358
urination pain	4.45	0.795	0.00	4.34	0.814	0.000	4.04	0.784	0.000	4.41	0.828	0.000
dyspareunia	3.33	0.627	0.00	3.13	0.626	0.000	3.13	0.642	0.000	3.11	0.631	0.000
age	1.07	0.031	0.02	1.06	0.037	0.122	1.06	0.038	0.114	1.06	0.037	0.121
age ²	1.00	0.000	0.03	1.00	0.000	0.123	1.00	0.000	0.112	1.00	0.000	0.123
ln(σ ²)	1.49	0.096		3.85	0.355		1.41	0.104		3.83	0.355	

All information: xi3: xtlogit qhem13 i.experience*qst13 qhem07 qhem10 rage ragesq if _mj==0, i(pid) or

Complete case: xi3: mim: xtlogit qhem13 i.experience*qst13 qhem07 qhem10 rage ragesq, i(pid) or

- The prefix **MIM** can be added to a range of commands to incorporate the information from the imputed datasets eg:
mim: tab qhem13
mim: mean age
mim: testparm _lexperienc_1 _lexperienc_2 - F(2,1000) = 0.75, p= 0.4722
- If separate imputed data files are required for other software packages (e.g. R) **misplit/mijoin** can be used

Longitudinal results:

- When imputation methods are incorporated there are notable changes in both the estimand coefficients (i.e. the odds ratios) and the standard errors - particularly when **ICE** is used to attempt to impute the 'whole' data - that is 'all information' compared to 'complete cases'
- The differences between the non-imputed and imputed results are only marginal when **ICE** is used only on complete cases across the three waves (this includes those cases imputed for panel 2)
- Where information is more-complete (i.e., respondents' age) the impact of imputing missing data between the dependent variable and independent variable is less pronounced

Australian
Research Centre
in Sex, Health
& Society



Acknowledgment to the ALSHR chief investigators

Prof. Anthony Smith

Prof. Marian Pitts

Dr Julia Shelley

Assoc. Prof. Juliet Richters

