CENTER FOR LIFE COURSE HEALTH RESEARCH

RESEARCH UNIT OF MEDICAL IMAGING, PHYSICS AND TECHNOLOGY



Multiple imputation

Practical Statistics Club, 6th Oct 2017

How to analyse data?

- Complete-case analysis
- Available-case analysis
- Last value carried forward
- Using information from related observations
 - E.g. mother's SES if father's SES is not available
- Add an extra category for missing data?
- Impute with mean value or use regression analysis?
- Check that missingness is not due to questionnaire (e.g. participants may have been asked to skip questions)

Purpose of multiple imputation

- To generate possible values for missing values by creating several (e.g. 10)
 'complete' datasets
- As a result, output for each complete

dataset and pooled output

Advantages of multiple imputation

- Reduce bias
- Improve validity
- Increase precision
- Result to robust statistics

SPSS – missing value analysis

- Describes the pattern of missing data. Where are the missing values located? How extensive are they? Do pairs of variables tend to have values missing in multiple cases? Are data values extreme? Are values missing randomly?
- Estimates means, standard deviations, covariances, and correlations for different missing value methods: listwise, pairwise, regression, or EM (expectation-maximization). The pairwise method also displays counts of pairwise complete cases.
- Fills in (imputes) missing values with estimated values using regression or EM methods; however, multiple imputation is generally considered to provide more accurate results.

SPSS – missing value analysis

WOLKOATA TPTTTP'SAN [NATASELT] - IRW 2622 STATISTICS NATA FOILOL											
View Data Transform Analyze Direct Marketing Graphs Utilities Extensions Window Help											
🖨 🛄 🗠										S	
ses_all								Display			
🛷 idkt 🛷 CPZ	34y_kroken 🗞 psykhier1dg2013	🗞 psykhier2	2dg2013 🛛 🖓	diagnos	ses_all_sources_jouko2	🗞 anydg	<u>9</u> 97	Demit pa	ed cases, grouped by atterns with less than variables by missing		
5311	. no psychiatric dg	no psychiat	tric dg				others		_ , 0	sorted by missing value pat	no
5411	. other psychosis (nosis, FHDR09-		3)		others		variables by missing	, , ,	no
5511	. non psychotic (no	non pí 👝 🔊	Missing Value Analys	ric		X	thers		s, optionally sorted b		no
5611	. non psychotic (no	non p	without a straight with the straight st	515			thers	An case	s, optionally solice b	y selected variable	no
5711	. no psychiatric dg	no psy			Quantitative Variables:	Patterns	thers	Variables -			no
5811	. no psychiatric dg		kohorttinumer 🖆			Descriptives	thers	Missing Pa	atterns for:	Additional Information f	or: no
5921	. no psychiatric dg		CPZ34y_kroken	•	· · · · · · · · · · · · · · · · · · ·)	thers				no
5922	. no psychiatric dg	no nevi 1	Psychosis hie			Estimation	thers		•	•	no
6011	. no psychiatric dg		diagnoses all			Listwise	thers				no
6111	. no psychiatric dg		any non-organ		Categorical Variables:	Pair <u>w</u> ise	thers			Sort by:	no
6311	. no psychiatric dg	no psy 🛛 💰	validation -199	[<u>E</u> M	thers		•		no
6411	. non psychotic (no	non ps	validation -199			Regression	thers			Sort Order	no
6511	. no psychiatric dg		validation -199 validation -199	*		Variables	thers			 Ascending 	no
6611	. no psychiatric dg	no neu	validation -199			E <u>M</u>	thers			© Descending	no
6711	. no psychiatric dg		validation -199	Mavim	um Categories: 25	Regression	thers				no
6811	. other psychosis (other 👔 🎝	validated dg 1	_	Case Labels:		thers				no
6911	. no psychiatric dg		a validated dg 1 🖵	•	Case La <u>b</u> eis.		thers	1	<u>C</u> ontinue Car	ncel Help	no
7011	. non psychotic (no	non ps –	······································				thers	по	no	по	no
7111	. no psychiatric dg	no psy	se <u>A</u> ll Variables				thers	no	no	no	no
7211	. non psychotic (no	non ps	OK	Paste	e <u>R</u> eset Cancel Help		thers	no	no	no	no
7311	. no psychiatric dg	no psy	-				thers	no	no	no	no
7411	. no psychiatric dg	no psychiat	tric dg				others	no	no	no	no
7511	. no psychiatric da	no psychiat	tric da				others	no	no	no	no

Analyze patterns

Descriptive measures of the patterns of missing values in the data

- Monotone pattern
- Non-monotone pattern

Rubin D & Little RJA. Statistical Analysis with Missing Data. Wiley, 2002.

Impute Missing Data Values

- Generation of multiple complete datasets
- Important to find an appropriate model which incorporates random variation, e.g.
 - include the *outcome* to imputation model
 - include a wide range of variables to imputation model (all variables in the substantive analysis plus variables predictive of the missing values (if computationally feasible))
 - problems may arise with skewed or categorical variables

Multiply imputed datasets

 Both user- and system missing values are replaced when values are imputed

Imputation_	X1	X2	kato43y											
Original data	71	69	tutkitt	u										
Original data	35	57	tutkitt	u										
Original data	60	80	tutkitt	u										
Original data	90	82	Imp	utation	X1	X2	Ē							
Original data	81	\times .	_ mp	nation_	~1	~2								
Original data	40	29		1	35	57		X2						
Original data	50	\times .	1	1	60	80	5	57	X2					
Original data	35	46	-	1	90	00		80 :			Imputation_	X1	X2	kato43y
Original data	50	60	-	1	81	>> 44	-	82	57			25	67	1.11.11
Original data	51	\times .	1	1	40	00	-	77	80		10	35	57	tutkittu
			1	1	50	> 79	- 1	29	82		10	60	80	tutkittu
			1	1	35	46	-		65		10	90	82	tutkittu
			1	. 1	50	60	-	48)	29		10	81	68	kato
			1	. 1	51	> 69		46)	55	•••	10	40	29	tutkittu
			1			-	-	60 ;	46		10	50	31	kato
						2 5	1	58)	60		10	35	46	tutkittu
							3	51	71		10	50	60	tutkittu
							-				10	51	34	kato

Using the multiple imputation

- To activate the MI split the file by Imputation_
- Perform the desired analyses on each dataset by using standard methods (marked with Ro)

<u>A</u> nalyze	Direct <u>M</u> arketing	<u>G</u> raphs	<u>U</u> tilities /	\dd- <u>o</u> ns			
Re <u>p</u> o	rts	Þ					
D <u>e</u> sc	riptive Statistics	•	Frequencies				
Ta <u>b</u> le	S	*	Descriptives				
Co <u>m</u> p	oare Means	•					
<u>G</u> ene	ral Linear Model	• • • • • • • • • • • • • • • • • • •	Crosstal				

Replication of the MI

- If replication is needed, one needs to use
 - the same initialization value for the random number generator
 IBM SPSS Statistics Data Editor
 - the same data order
 - the same variable order
 - the same procedure settings

IBM SPSS Stat	istics Data	Editor						
<u>T</u> ransform	<u>A</u> nalyze	Direct Marketing	<u>G</u> raphs					
Compute	Variable							
🛨 Program	mability Tra	ansformation						
Count Values within Cases								
Shi <u>f</u> t Valu	es							
Recode into Same Variables								
🔤 <u>R</u> ecode in	nto Differer	nt Variables						
🛐 <u>A</u> utomatio	Recode							
🛨 Create D	ummy Vari	ables						
<mark>⊪</mark> ∦ Visual <u>B</u> ir	nning							
🔀 Opt <u>i</u> mal E	Binning							
Prepare l	Data for Mo	odeling	•					
🛃 Ran <u>k</u> Cas	ses							
🗎 Date and	☐ <u>D</u> ate and Time Wizard							
Create Time Series								
<table-of-contents> Replace I</table-of-contents>	Missing <u>V</u> a	lues						
🍘 Random	Number <u>G</u>	enerators						
	ling Tranef	forme (`trl+G					

Comparing means in two independent sample

Imputation Nu	mber	Life_Satisfaction	N	Mean	Std. Deviation	Std. Error Mean
Original data	HSCL_Depression_New	>= 3	48	1,7444	,37165	,05364
		< 3	625	1,2789	,25299	,01012
1	HSCL_Depression_New	>= 3	81	1,7695	,47153	,05239
		< 3	844	1,2998	,28657	,00986
2	HSCL_Depression_New	>= 3	81	1,7770	,47131	,05237
		< 3	844	1,2992	,28753	,00990
3	HSCL_Depression_New	>= 3	81	1,7761	,49073	,05453
		< 3	844	1,2996	,28676	,00987
4	HSCL_Depression_New	>= 3	81	1,7671	,46597	,05177
		< 3	844	1,3007	,28883	,00994
5	HSCL_Depression_New	>= 3	81	1,7663	,46105	,05123
		< 3	844	1,2979	,28150	,00969
Pooled	HSCL_Depression_New	>= 3	81	1,7712		,05276
		< 3	844	1,2994		,00992

Group Statistics

Life satisfaction: good <3 vs. poor >= 3

Independent Samples Test

Levene's Test for Equality of Variances

			_				
Imputation Nu	mber		F	Sig.	t	df	Sig. (2-tailed)
Original data	HSCL_Depression_New	Equal variances assumed	17,138	,000	11,815	671	,000
		Equal variances not assumed			8,528	50,400	,000
1	HSCL_Depression_New	Equal variances assumed	41,462	,000	13,153	923	,000
		Equal variances not assumed			8,812	85,762	,000
2	HSCL_Depression_New	Equal variances assumed	39,881	,000	13,342	923	,000
		Equal variances not assumed			8,964	85,807	,000
3	HSCL_Depression_New	Equal variances assumed	44,674	,000,	13,224	923	,000
		Equal variances not assumed			8,600	85,321	000,
4	HSCL_Depression_New	Equal variances assumed	37,472	,000	13,007	923	000,
		Equal variances not assumed			8,846	85,997	000,
5	HSCL_Depression_New	Equal variances assumed	41,289	,000	13,363	923	000,
		Equal variances not assumed			8,984	85,816	000,
Pooled	HSCL_Depression_New	Equal variances assumed			13,059	7149	,000
		Equal variances not assumed			8,789	34843,932	,000

LOGISTIC REGRESSION DICHOTOMIZED DEPRESSION AS AN OUTCOME

Variables in the Equation

									95% C.I.fo	r EXP(B)
Imputation Nur	mber		в	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Original data	Step 1 ^a	Gender(1)	,414	,200	4,268	1	,039	1,513	1,021	2,241
		Basic education(1)	,269	,207	1,689	1	,194	1,309	,872	1,965
		Constant	-1,834	,214	73,634	1	,000,	,160		
1	Step 1 ^a	Gender(1)	,464	,165	7,889	1	,005	1,590	1,150	2,197
		Basic education(1)	,321	,171	3,519	1	,061	1,379	,986	1,930
		Constant	-1,777	,178	100,237	1	,000,	,169		
2	Step 1 ^a	Gender(1)	,478	,166	8,321	1	,004	1,612	1,165	2,230
		Basic education(1)	,314	,172	3,351	1	,067	1,369	,978	1,916
		Constant	-1,787	,178	100,829	1	,000,	,168		
3	Step 1 ^a	Gender(1)	,491	,166	8,766	1	,003	1,635	1,181	2,263
		Basic education(1)	,307	,172	3,186	1	,074	1,359	,970	1,903
		Constant	-1,796	,178	101,424	1	,000,	,166		
4	Step 1 ^a	Gender(1)	,478	,166	8,321	1	,004	1,612	1,165	2,230
		Basic education(1)	,314	,172	3,351	1	,067	1,369	,978	1,916
		Constant	-1,787	,178	100,829	1	,000,	,168		
5	Step 1 ^a	Gender(1)	,431	,165	6,856	1	,009	1,539	1,115	2,126
		Basic education(1)	,287	,171	2,826	1	,093	1,333	,953	1,863
		Constant	-1,737	,176	97,150	1	,000	,176		
Pooled	Step 1 ^a	Gender(1)	,468	,167			,005	1,597	1,151	2,217
		Basic education(1)	,309	,172			,073	1,362	,972	1,908
		Constant	-1,777	,179			,000	,169	,119	,241

a. Variable(s) entered on step 1: Gender, Basic education.

EXAMPLE DATA SET

This sample data set is an **anonymised**, **randomly-selected** sample from the cohorts managed by the Northern Finland Birth Cohort Project Center. Any identifying information has been removed from the sample. Some variables have been recoded for purpose of this analysis and missingness of the data has been exaggerated.

<u>Please respect the confidentiality of the data</u> <u>and delete all related files before you leave</u> <u>the room. Please ensure that you do not save</u> <u>or send any of the data provided today.</u>