CENTER FOR LIFE COURSE HEALTH RESEARCH

RESEARCH UNIT OF MEDICAL IMAGING, PHYSICS AND TECHNOLOGY

OYS OULU UNIVERSITY HOSPITAL

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

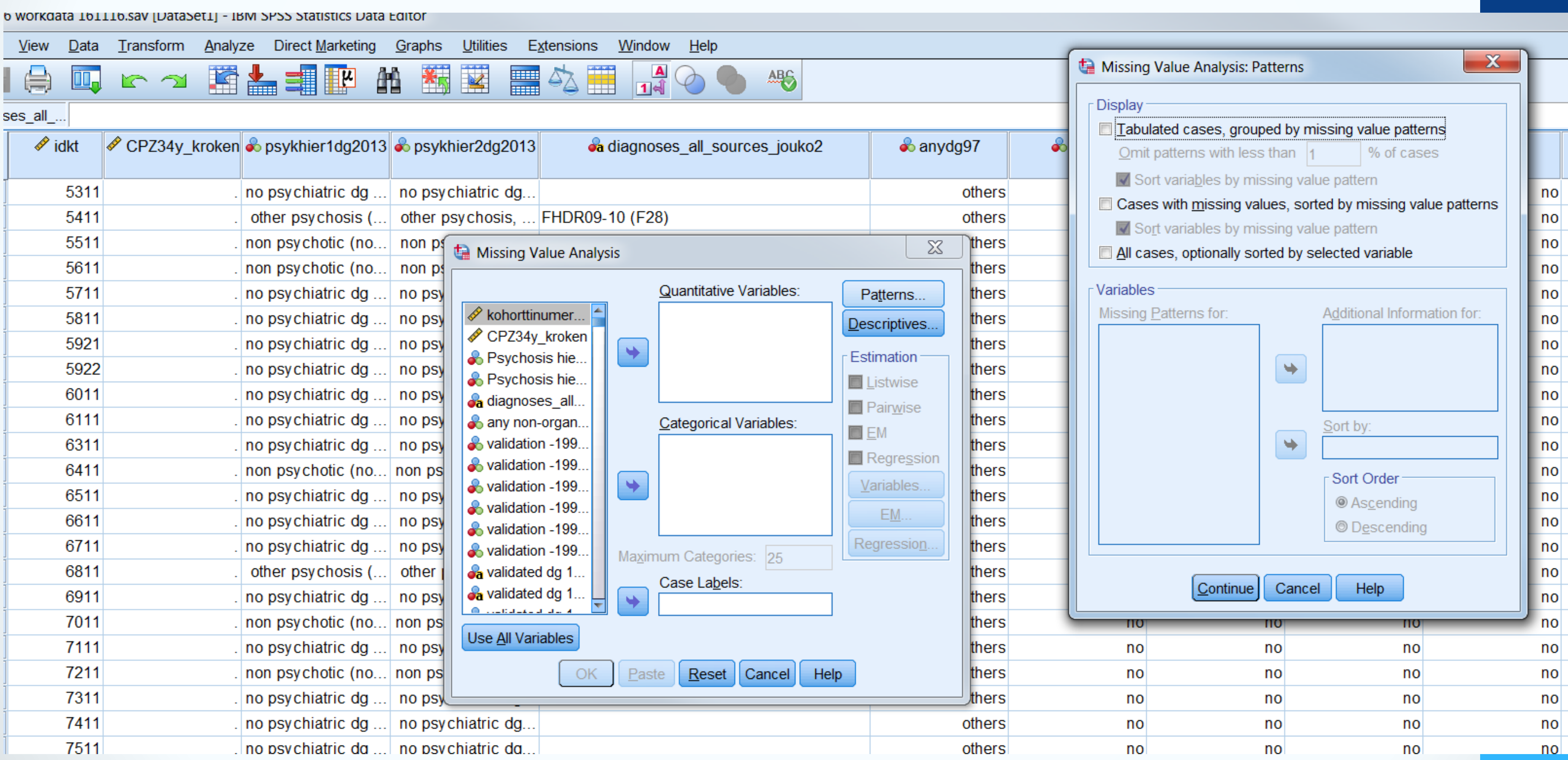
Advantages of multiple imputation

- 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



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	tutkittu
Original data	35	57	tutkittu
Original data	60	80	tutkittu
Original data	90	82	
Original data	81	.	
Original data	40	29	
Original data	50	.	
Original data	35	46	
Original data	50	60	
Original data	51	.	

Imputation_	X1	X2
1	35	57
1	60	80
1	90	82
1	81	44
1	40	29
1	50	79
1	35	46
1	50	60
1	51	69

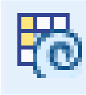
Imputation_	X1	X2
		57
		80
		82
		77
		29
		48
		46
		60
2	51	58

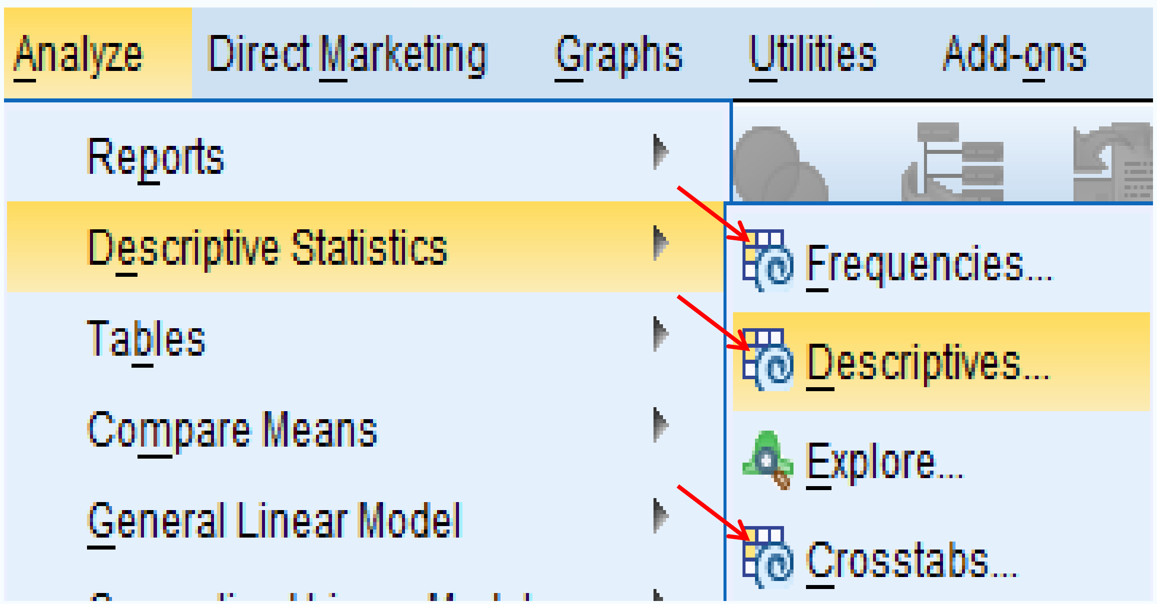
Imputation_	X1	X2
		57
		80
		82
		65
		29
		55
		46
		60
3	51	71

...

Imputation_	X1	X2	kato43y
10	35	57	tutkittu
10	60	80	tutkittu
10	90	82	tutkittu
10	81	68	kato
10	40	29	tutkittu
10	50	31	kato
10	35	46	tutkittu
10	50	60	tutkittu
10	51	34	kato

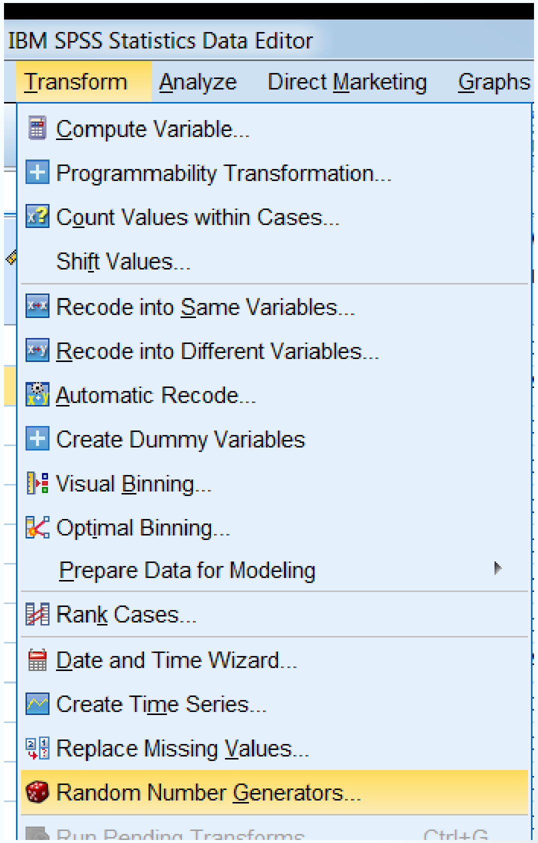
Using the multiple imputation

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



Replication of the MI

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



Comparing means in two independent sample

Group Statistics

Imputation Number		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

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

Independent Samples Test

Imputation Number			Levene's Test for Equality of Variances 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

Imputation Number			B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for 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 and delete all related files before you leave the room. Please ensure that you do not save or send any of the data provided today.**</u>