

Missing Values:

What should we teach our students?

What should we do ourselves?

Harry BG Ganzeboom

SILC Seminar VU University

September 12, 2023

Motivation

- It is now almost 20 years ago that the SILC Seminar started, under the name of Missing Values Seminar (participants: Harry Ganzeboom, Maarten Buis, Adriaan Hoogendoorn).
- I see 'complete cases analysis' (= 'listwise deletion of missing values') all over in student research and professional research, without any reflection on its consequences.

Earlier

- Mariska van der Horst (2019). Onderzoekslab Missende Waarden
- Harry Ganzeboom (2020). Onderzoekslab Missende Waarden.
- Earlier presentation: May 12 2022 (??).

Major points

- Missing values are endemic to survey data. Learning about proper MV treatment is more important than learning about regression diagnostics, bootstrapping, indirect effects or logistic regression.
- There are two different consequences of using (only) complete cases analysis:
 - A. Loss of power – because you are analysing fewer cases than you actually have.
 - B. Bias – because you may omit a non-random part of your data.
- MV treatment is about using the data that you DO HAVE, not about imputing the data you DO NOT HAVE.
- MI and ML are (asymptotically) equivalent methods of MV treatment. Students should learn both, but it is easier (and more intuitive) to learn ML.

The two consequences of MV

- Missing values in data lead to two problems, in particular with complete cases ('listwise') analysis:
 - #1 Loss of statistical power by omitting all incomplete cases.
 - #2 Bias when the incomplete cases (this includes non-response = completely missing cases) are different from the complete cases.
- These are two different things:
 - Loss of power **always** occurs
 - Bias **may** occur.

I think loss of power is in practice the more important consequence. It is important to make make clear how many data are lost.
- Complete case analysis leads to **the least powerful and potentially most biased selection** of your data.
- Reports often focus on the amount of data preserved (complete cases) and hardly ever report on how much could have been preserved (available cases). I propose to make the difference more explicit by reporting about incomplete cases.

Which of the two consequences is more consequential?

- Loss of power and bias are two different problems, and it is hard to say which is the more important one.
- Many people think that #2 (bias) is the main or only problem. I disagree: the loss of power is the more important issue.
- There is a parallel with measurement: sampling / measurement unreliability vs. sampling / measurement validity.
- These insights I received from Donald Rubin.

Missing values scenario's

- MCAR = missings arise completely at random;
- MAR = missings arise randomly, conditional upon the distributions of non-missings;
- MNAR = missings arise systematically, but the sources are not known and not measured.
- Some literature says that MNAR can be diagnosed and repaired using Heckman's selection model – but I do not understand how and how it can be applied (Heckman's selectivity model requires that you can predict the missing value mechanism – which would turn MNAR \rightarrow MAR).
- The only scenario's we can address are MCAR and MAR. Notice that the choice between these is problem-specific, i.e. specific to the variables / model analyzed.
- My problem is: I understand the difference between MCAR and MAR, but I do not see how they influence analytical choices.

MCAR (in linear models)

- Missings are randomly generated:
- → available-cases ('pairwise') correlation / covariance matrix is the same as the complete cases ('listwise') version; the difference is only in the N of Cases.
- → available cases estimation of structural model (factor and regression analysis) leads to the same coefficients as complete cases analysis.
- This comparison (complete cases vs available cases; complete cases vs incomplete cases) is always a useful start of the analysis.
- In MCAR only problem #1 of missing values applies.

MAR

- Missings are randomly generated, given the distribution of measured variables in the model.
- → Covariances / correlations will be different between complete cases / available cases / incomplete cases, but the coefficients of structural model will be the same.
- This may apply to part of data / model.
- Example: missing values arise by low literacy and low socio-economic status. If MAR applies, the structural model can be equalized between complete / incomplete cases, adjusted for these sources of missingness.
- In MAR both problem #1 and #2 of missing values arise.

Complete and incomplete cases

- Split the data in the complete cases and incomplete cases.
- Complete cases + incomplete cases = available cases.
- Report descriptives (M, SD, CORR, N) for both parts of the data.
 - This report will make you aware of about how many valid data are lost by complete cases analysis.
 - You can test the differences in M en SD using T-Test (in SPSS). I do not think that significance testing is of great interest here, but it is helpful.
 - I do not know a test of inequality of a covariance / correlation matrix in SPSS. In Stata SEM it is fairly simple:

Stappenplan Missing Values

- Step 1: Run off frequencies of all variables in the analysis to see whether the MV are properly labelled and marked.
- Step 2: Create a variable NMISS that counts the number of MV in your active variables; alternatively, create a variable PATMISS that shows the occurrence of MV in each case.
 - Crosstabulate NMIS * PATMIS.
- Step 3a: using NMISS, split the data in complete cases and incomplete cases. Compare the MEANs and SDs to see to what extent the situation is MCAR or MAR.
- Step 3b: using NMISS, split the data in complete and incomplete cases. Compare the (pairwise) correlations between the two splits.

In praise of SPSS

- SPSS distinguishes between user-defined and system missings. User-defined missings can have labels. These are two very useful features.
- SPSS can use pairwise and listwise correlations / covariances to do regression and factor analyses.
- Comparing listwise and pairwise models is very useful to diagnose the nature of the missingness mechanism: MCAR or MAR?
- However: ***the statistics (SE) in pairwise REGR in SPSS are wrong.***
- The only equivalent (in fact: equivalent but an improvement) is MLMV / FIML in SEM. This is what our students should learn.

Step 1: Describing the data

- Run frequencies of all your (selected) variables.
- Check whether the MV are appropriately labelled and that you understand how they have come about.
- Substitute missings that have very plausible ('logical') values. E.g. DK may be coded to the midpoint of answering scale; people who do not work have zero labour income.
- Run the DESC and CORR for all available cases.

Some simple ways to keep your data intact

- Before or after moving to any complicated method (MI or ML) to treat missing values, try the following:
 - Impute 'logical' MV with their most plausible values. E.g. people without a job have zero labour income.
 - Impute MV in batteries by using 'available information' averaging.
 - Substitute MV in control variables by their means, while controlling the substitution by an indicator variable.
- None of these three simple procedures are correct or optimal, but they are easy to do and often create little harm.
- They can be motivated by the rule: ***use as many of the data that you DO HAVE.***
- Compare results before and after MV treatment.

AVAILABLE CASES

	N	Min	Max	Mean	SD
age	1273	20	79	45.37	14.591
female	1273	0	1	.57	.495
educyr	1241	1.0	17.0	7.953	4.4053
fasei	791	11	88	30.74	18.348
masei	522	11	88	27.67	20.216
asei	953	11	88	36.82	18.753
Valid N (listwise)	342				

COMPLETE CASES

	N	Min	Max	Mean	SD
age	342	20	75	45.62	14.202
female	342	0	1	.56	.497
educyr	342	1.0	17.0	8.993	4.7376
fasei	342	11	88	31.69	19.120
masei	342	11	88	28.77	20.857
asei	342	11	88	37.54	19.054
Valid N (listwise)	342				

INCOMPLETE CASES

	N	Min	Max	Mean	SD
age	931	21	79	45.27	14.738
female	931	0	1	.58	.494
educyr	899	1.0	17.0	7.558	4.2079
fasei	449	11	88	30.01	17.725
masei	180	11	82	25.59	18.820
asei	611	11	85	36.42	18.586
Valid N (listwise)	0				

Step 3: Comparing complete and incomplete cases

- Describe and test the difference in Means and SD using:
T-TEST group=complete(0,1) /var=var1 to vark
This provides a test of equality of means (Student) and a test of equality of SDs (Levene)
- Describe and test the differences in correlations. I know of no easy way in SPSS, but in SEM it is super-easy:
sem (var1) (var2) .. (vark) , group(complete) ginvariant(covex)

COMPARING COMPLETE AND INCOMPLETE DATA

<u>Group Statistics</u>					<u>Equal SD</u>	<u>Equal M</u>
		N	Mean	SD	Levene F	Student t
ZAGE	COMPLETE	342	.018	.973	.8	.4
	INCOMPLETE	931	-.006	1.010		
ZFEMALE	COMPLETE	342	-.023	1.004	.9	-.5
	INCOMPLETE	931	.008	.999		
ZEDUC	COMPLETE	342	.236	1.075	22.5	4.9
	INCOMPLETE	899	-.090	.955		
ZFASEI	COMPLETE	342	.052	1.042	6.3	1.3
	INCOMPLETE	449	-.040	.966		
ZMASEI	COMPLETE	342	.054	1.032	5.6	1.8
	INCOMPLETE	180	-.103	.931		
ZASEI	COMPLETE	342	.038	1.016	1.0	.9
	INCOMPLETE	611	-.021	.991		

Step 2: Describing the missingness.

- Generate a count variable NMISS that counts the number of missing values per case.
- NMISS can be recoded into a 0/1 variable that indicates the incomplete and complete data.
 - Count NMISS = var1 to vark (missing, sysmiss).**
 - Recode NMISS (0=1)(1 thru hi=0) into COMPLETE.**
- Report not only the N of the complete cases, but also the N of the incomplete cases.

COUNT OF THE MISSINGS

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 0	342	26.9	26.9	26.9
1	440	34.6	34.6	61.4
2	340	26.7	26.7	88.1
3	139	10.9	10.9	99.1
4	12	.9	.9	100.0
Total	1273	100.0	100.0	

COMPLETE CASES

		age	female	educyr	fasei	masei	asei
Pearson Correlation	age	1.000	-.109	-.273	-.121	-.204	-.084
	female	-.109	1.000	.162	.085	.149	.169
	educyr	-.273	.162	1.000	.471	.499	.686
	fasei	-.121	.085	.471	1.000	.518	.369
	masei	-.204	.149	.499	.518	1.000	.381
	asei	-.084	.169	.686	.369	.381	1.000
N	age	342	342	342	342	342	342
	female	342	342	342	342	342	342
	educyr	342	342	342	342	342	342
	fasei	342	342	342	342	342	342
	masei	342	342	342	342	342	342
	asei	342	342	342	342	342	342

INCOMPLETE CASES

		age	female	educyr	fasei	masei	asei
Pearson Correlation	age	1.000	.010	-.331	-.092	-.330	-.043
	female	.010	1.000	-.062	-.091	-.125	.073
	educyr	-.331	-.062	1.000	.377	.519	.608
	fasei	-.092	-.091	.377	1.000	.698	.215
	masei	-.330	-.125	.519	.698	1.000	.262
	asei	-.043	.073	.608	.215	.262	1.000
N	age	931	931	899	449	180	611
	female	931	931	899	449	180	611
	educyr	899	899	899	439	170	594
	fasei	449	449	439	449	80	288
	masei	180	180	170	80	180	89
	asei	611	611	594	288	89	611

SEM TEST OF EQUALITY OF CORRELATIONS

**sem (zage) (zfemale) (zeducyr) (zasei) (zfasei) (zmasei) ,
method(mlmv) group(complete) ginvariant(covex)**

**LR test of model vs. saturated: chi2(21) =
25.38, Prob > chi2 = 0.2311**

N=953

DETERMINANTS OF ZASEI (OCCUPATION)

		<u>Complete</u>	<u>Incomplete</u>	<u>Available</u>
ZAGE	B	0.124	0.169	0.142
	SE	0.042	0.039	0.027
FEMALE	B	0.130	0.012	0.049
	SE	0.080	0.067	0.049
ZEDUCYR	B	0.631	0.646	0.647
	SE	0.045	0.052	0.032
ZFASEI	B	0.041	-0.038	0.008
	SE	0.046	0.069	0.039
ZMASEI	B	0.040	0.138	0.056
	SE	0.048	0.098	0.046
N		342	611	953