How should we handle missing data?

Scott Oatley
soatley@ed.ac.uk

What is Missing Data?

What is Missing Data? (Theory)

- MCAR
- MAR
- MNAR

Why Should we care about Missing Data?

- ‘Flipping’ – where missingness flips the substantive significance of a finding from positive to negative or vice versa

- ‘Flopping’ – where missingness minimises or over-empahsises the size of the substrative finding

- ‘Flip-Flopping’ – where missingness flips the substantive significance and minimises/over emphasises the result

What does this all mean?

- We can’t ignore missing data
- And yet most studies do
- ‘’I’ve looked at the missingness in my data and confirmed there will be no bias…’’

How to handle missing data?

- Several approaches
- Some good
- Some bad
- Some ugly

https://github.com/Scott0atley/YouthTransitions/tree/main/Q-Step

The Bad

- Listwise Deletion

- This just ignores the issue

The Ugly

- Recoding Missingness to a single value

- Say you have a binary independent variable where all missingness occurs in model
  - Code all missingness = 0 in that variable
  - Code all missingness = 1 in that variable

The Ugly

- Single mean/modal imputation

The Ugly

- Multiple Imputation with zero auxiliary variables

The Good

- Full Information Maximum Likelihood (FIML)
  - (Or MLMV in stata)
  - Uses SEM framework
  - Can’t use for non-linear models in Stata (Can in MPLUS)

The Good

- Multiple Imputation with auxiliary variables

Multiple good ways to handle missing data?

- Multiple Imputation versus FIML

Table 1: Simulation Regression Models Using a MCAR Principle

	Complete Records 'God Model'	Complete SEM	Missingness Introduced at Independent Variable 3	All Missingness coded as =0	All Missingness coded as =1	Single Use Modal Imputation	FIML	Imputed with no auxiliary variables and 10 imputations	Imputed with 10 imputations	Imputed with 100 imputations
Independent Variable 1	-0.18 ***	-0.18 ***	-0.18 ***	-0.26 ***	-0.26 ***	-0.18 ***	-0.18 ***	-0.17 ***	-0.18 ***	-0.18 ***
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Independent Variable 2	-0.19 ***	-0.19 ***	-0.20 ***	-0.26 ***	-0.26 ***	-0.20 ***	-0.19 ***	-0.19 ***	-0.20 ***	-0.20 ***
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Independent Variable 3	-0.19 ***	-0.19 ***	-0.20 ***	-0.06 ***	-0.06 ***	-0.20 ***	-0.20 ***	-0.20 ***	-0.19 ***	-0.19 ***
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Intercept	1.15 ***	1.15 ***	1.16 ***	1.29 ***	1.31 ***	1.16 ***	1.15 ***	1.15 ***	1.16 ***	1.16 ***
	(0.02)	(0.02)	(0.03)	(0.02)	(0.01)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
Number of observations	1000	1000	512	1000	1000	512	1000	1000	1000	1000

***** p<.001, ** p<.01, * p<.05**

Data Source: Simulation using a MCAR principle. 51 per cent missingness introduced.

Table 2: Simulation Regression Models Using a MAR Principle

	Complete Records 'God Model'	Complete SEM	Missingness Introduced at Independent Variable 3	All Missingness coded as =0	All Missingness coded as =1	Single Use Modal Imputation	FIML	Imputed with no auxiliary variables and 10 imputations	**Imputed with 10 imputations**	Imputed with 100 imputations
Independent Variable 1	[-0.19,-0.19]	[-0.19,-0.19]	[-0.10,-0,10]	[-0.28,-0.27]	[-0.19,-0.19]	[-0.28,-0.27]	[-0.12,-0.12]	[-0.20,-0.20]	**[-0.19,-0.18]**	[-0.20,-0.20]
	[(0.02,0.02)]	[(0.02,0.02)]	[(0.01,0.01)]	[(0.02,0.02)]	[(0.02,0.02)]	[(0.02,0.02)]	[(0.02,0.02)]	[(0.02,0.02)]	**[(0.02,0.02)]**	[(0.02,0.02)]
Independent Variable 2	[-0.19,-0.19]	[-0.19,-0.19]	[-0.10,-0,10]	[-0.28,-0.28]	[-0.19,-0.19]	[-0.28,-0.28]	[-0.12,-0.12]	[-0.18,-0.18]	**[-0.19,-0.19]**	[-0.19,-0.19]
	[(0.02,0.02)]	[(0.02,0.02)]	[(0.01,0.01)]	[(0.02,0.02)]	[(0.02,0.02)]	[(0.02,0.02)]	[(0.02,0.02)]	[(0.02,0.02)]	**[(0.02,0.02)]**	[(0.02,0.02)]
Independent Variable 3	[-0.19,-0.19]	[-0.19,-0.19]	[-0.10,-0,10]	[0.07,0.07]	[-0.19,-0.19]	[0.07,0.07]	[-0.25,-0.25]	[-0.20,-0.20]	**[-0.19,-0.19]**	[-0.18,-0.18]
	[(0.02,0.02)]	[(0.02,0.02)]	[(0.01,0.01)]	[(0.02,0.02)]	[(0.02,0.02)]	[(0.02,0.02)]	[(0.01.0.01)]	[(0.02,0.02)]	**[(0.02,0.02)]**	[(0.02,0.02)]
Number of observations	1000	1000	513	1000	1000	1000	1000	1000	**1000**	1000

***** p<.001, ** p<.01, * p<.05**

Data Source: Simulation using a MAR principle. 51 per cent missingness introduced.

Thank You

- Any Questions?