THE UNIVERSITY *of* EDINBURGH

How should we handle missing data?

Scott Oatley
soatley@ed.ac.uk

What is Missing Data?

- Item missing
- Unit missing

What is Missing Data? (Theory)

- MCAR
- MAR
- MNAR

Missing Completely At Random (MCAR)

- Suppose that only one variable Y has missing data, and that another set of variables represented by the vector X, is always observed (Marsden and Wright, 2010). The data is MCAR if the probability that Y is missing does not depend on either X or Y itself.

- Example: An example of MCAR is a weighing scale that ran out of batteries. Some of the data will be missing simply because of bad luck. (Van Buuren & Van Buuren 2012)

Missing At Random (MAR)

- Data on Y is considered MAR if the probability that Y is missing does not depend on Y, once we control for X. MAR allows for missingness on Y to depend on other variables so long as it does not depend on Y itself.

- Example: Women are less likely to report their incomes – regardless of what their income actually is.

Missing Not At Random (MNAR)

- Means missingness depends on unobserved values (Silverwood et al. 2021), and that the probability that Y is missing depends on Y itself, after adjusting for X (Marsden and Wright, 2010). For example, people who have been arrested may be less likely to report their arrest status.

- Example: People with low incomes do not answer the income question.

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 substantive finding

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

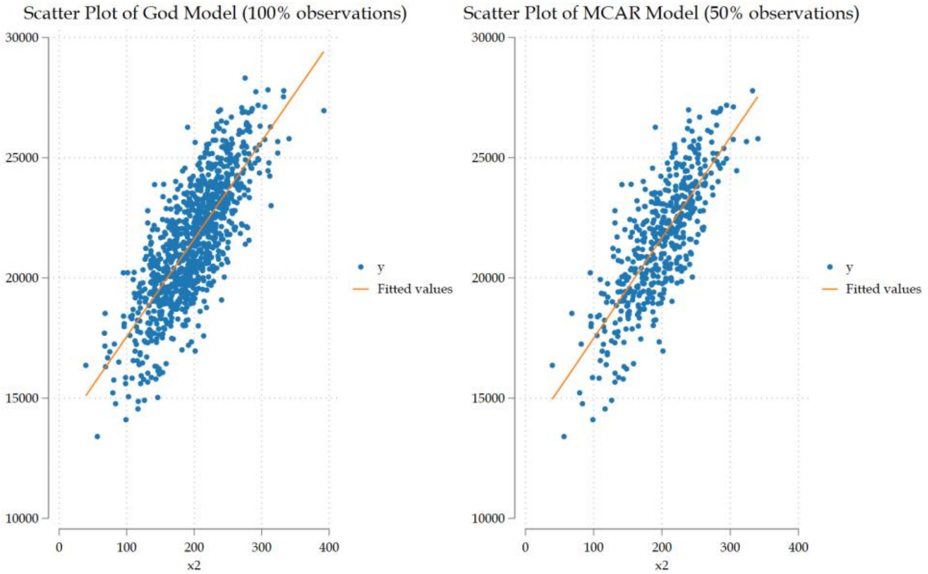
A quick working example

- A simple bivariate regression model is simulated

- The first model has a continuous dependent variable and a continuous independent variable with n=1000

- Model is injected with both a MCAR and a MAR mechanism to demonstrate potential issues

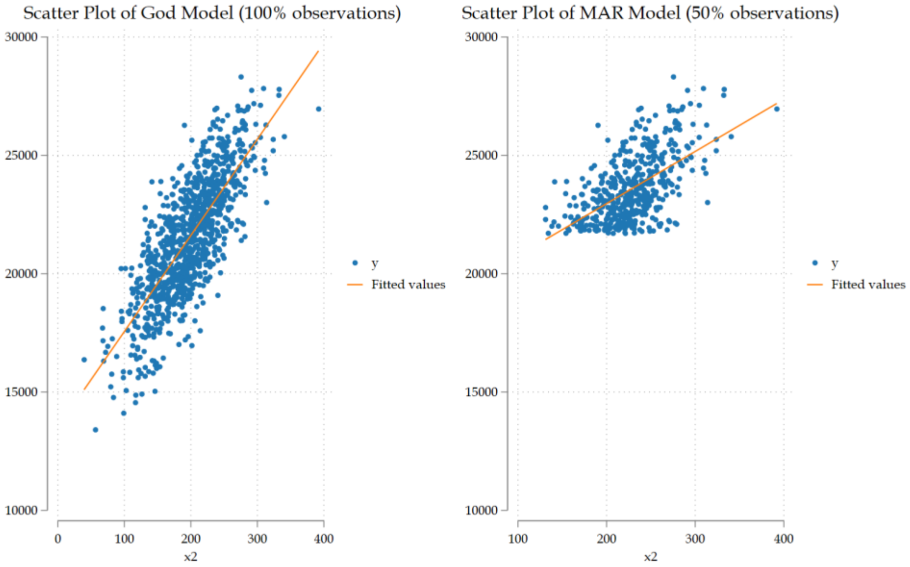
‘God’ Model versus MCAR



https://scott0atley.github.io/Scott0atley/presentations/

Influencing the world since 1583

‘God’ Model versus MCAR



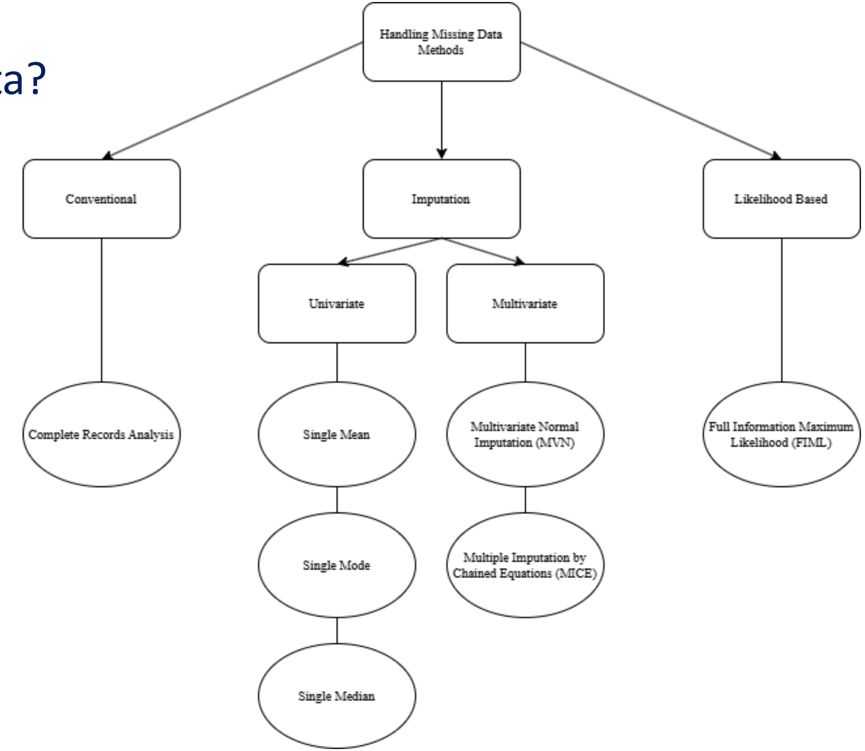
What does this all mean?

- 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

How to handle missing data?



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/median 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

Simulation Study

- N=1000
- 1 continuous dependent variable + 3 independent variables
- Missingness introduced into x2 variable
- Different handling missing data methods then assessed

Variables

- X1=(1000) means(40) sds(12)
- X2=n(1000) means(200) sds(50)
- X3=n(1000) means(150) sds(5)
- y = 30*x1 + 40*x2 + 50*x3 + rnormal(5000, 1500)

Missing Mechanisms

- MCAR =
- gen rmcar = rbinomial(1, 0.5) // MCAR: 50% chance of missingness (binary random)
- replace x2 = . if rmcar == 0 // Set x to missing where rmcar == 0

- MAR =
- gen prob_mar = logistic(y-21791)
- gen rmar = 0 if prob_mar==0
- replace x2 = . if rmar == 0 // Set x to missing where rmar == 0

Models

- 1) God Model
- 2) SEM God Model
- 3) MCAR Model
- 4) MAR Model
- 5) Single Mean Imputation Model
- 6) FIML Model
- 7) 10 imputation no auxiliary Model
- 8) 10 imputation auxiliary Model
- 9) 100 imputation auxiliary Model

Table 1: Simulation Regression Models Using a MAR Principle																		
	Complete Records 'God Model'		**Complete SEM**		**MCAR Introduced**		**MAR introduced**		**Single Use Mean Imputation**		**FIML**		**Imputed with no auxiliary variables and 10 imputations**		**Imputed with 10 imputations**		**Imputed with 100 imputations**	
	Coef.	***95% CIs***	***Coef.***	***95% CIs***	***Coef.***	***95% CIs***	***Coef.***	***95% CIs***	***Coef.***	***95% CIs***	***Coef.***	***95% CIs***	***Coef.***	***95% CIs***	***Coef.***	***95% CIs***	***Coef.***	***95% CIs***
Independent Variable 1	30.01	[22.25, 37.77]	30.01	[22.27, 37.75]	29.96	[18.95, 40.99]	18.44	[9.67, 27.20]	33.76	[21.31, 46.20]	29.92	[19.84, 40.00]	29.55	[18.62, 40.47]	31.36	[21.51, 41.21]	24.96	[15.55, 34.37]
	(3.96)		(3.95)		(5.62)		(4.47)		(6.35)		(5.14)		(5.57)		(5.03)		(4.80)	
Independent Variable 2	40.02	[38.15, 41.88]	40.02	[38.16, 41.88]	40.03	[37.39, 42.67]	24.76	[22.06, 27.45]	25.40	[19.94, 30.86]	40.03	[37.51, 42.54]	41.44	[38.89, 43.99]	41.51	[38.88, 44.13]	38.61	[36.20, 41.02]
	(0.95)		(0.95)		(1.35)		(1.38)		(2.78)		(1.28)		(1.30)		(1.34)		(1.23)	
Independent Variable 3	49.88	[31.23, 68.53]	49.88	[31.27, 68.49]	51.30	[24.88, 77.71[	30.23	[9.29, 51.18]	56.30	[26.41, 86.19]	49.55	[25.48, 73.61]	71.26	[50.46, 92.06]	44.77	[20.85, 68.69]	38.68	[16.16, 61.12]
	(9.52)		(9.50)		(13.48)		(10.69)		(15.25)		(12.28)		(10.61)		(12.21)		(11.49)	
Number of observations	1000		1000		499		489		1000		1000		1000		1000		1000	
R^2	0.68				0.66		0.43		0.12									
Data Source: Simulation using a MAR principle. 50 per cent missingness introduced.																		

What does this all mean?

- 1) God Model – perfect ideal model
- 2) SEM God Model – same as above
- 3) MCAR Model – inflated standard errors
- 4) MAR Model – big substantive issues
- 5) Single Mean Imputation Model – x2 issues, massive 95% CIs
- 6) FIML Model – Great!
- 7) 10 imputation no auxiliary Model – inflated x3 values
- 8) 10 imputation auxiliary Model – Great!
- 9) 100 imputation auxiliary Model – Great!

Conclusion

- No missing data is always the dream
- Dream is never reality
- Have to check for missing mechanisms
- If MCAR -> carry on, if MAR or MNAR -> look to handling missing data methods
- Using ‘bad’ methods is sometimes as bad as doing nothing!
- No difference in efficiency between FIML and MI approaches
- Use the method that best suits your data
    - FIML is very restricted in most software, MICE is ubiquitous and easy to implement

Thank You

- Any Questions?

- Van Buuren, S. and Van Buuren, S., 2012. *Flexible imputation of missing data* (Vol. 10, p. b1182). Boca Raton, FL: CRC press.