HANDLING MISSING DATA: A PRACTICAL INTRODUCTION TO TECHNIQUES AND METHODS

Scott Oatley

TIMELINE

10:00 – 10:15 – Welcome/Introductions
10:15 – 11:00 – First Session
11:00 – 11:15 – Coffee Break
11:15 – 12:00 – Second Session
12:00 – 1:00 – Lunch
1:00 – 1:45 – Third Session
1:45 – 2:00 – Coffee Break
2:00 – 2:45 – Fourth Session
2:45 – 3:00 – Coffee Break
3:00 – 3:45 – Fifth Session
3:45 – 4:00 – Coffee Break
4:00 – 5:00 – Q&A Session

INTRODUCTIONS

NOTES FOR TODAY

- Lunch NOT provided
  - Do feel free to use this space for the lunch hour!

- Resources are all available via my website: https://scott0atley.github.io/Scott0atley/training/
  - This will include all resources; .do/.R files, other resources, this powerpoint etc

- Reading list for this course based on specific topics we will cover
  - Also on the website

SESSIONS

SESSION ONE — Introduction to Missing Data

SESSION TWO — Ways to Handle Missing Data

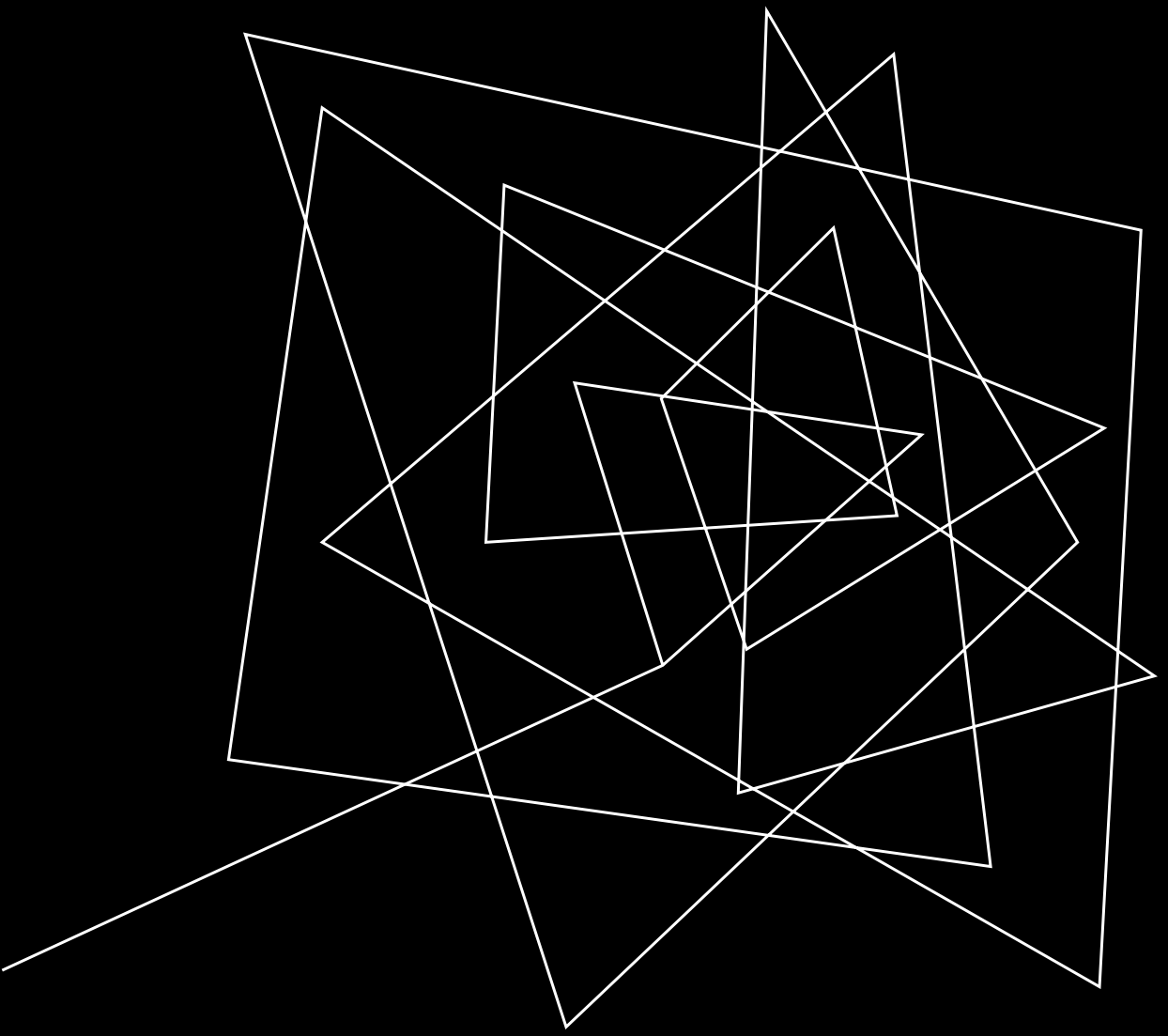
SESSION THREE — Full Information Maximum Likelihood

SESSION FOUR — Multiple Imputation

SESSION FIVE — Comparison of Methods

BASIC SETUP

- Training Materials & Working Examples are provided
- You can use either Stata .do files OR .R files
  - There is also a very limited amount of Python and MPLUS material
- Most working examples will be using Stata
- My knowledge of software:
  - Stata > R > Python > MPLUS
- Make sure files are downloaded with either Stata or Rstudio running on your machine!



SESSION ONE – INTRODUCTION TO MISSING DATA

SESSION ONE – INTRODUCTION TO MISSING DATA

- Introduction into forms of missingness
- What is the problem with missingness?
- Some examples of missingness mechanisms
- How much of missingness is an issue?
- How to know what kind of missingness you are dealing with?
- Practical

MISSING DATA

- Missing data pervades social science research
- Just as qualitative research suffers from potential interviewees cancelling, or important members of a focus group not turning up, so too does quantitative research suffer from missingness
- Missing data when present has the possibility of obfuscating the full story being told by the data at large
- Most* of the time, a simple complete records analysis (CRA) suffices for much of the statistical analysis we conduct as quantitative researchers
    - We can never be certain if all we do is run a CRA

MISSING DATA

- Missing data is a term that describes an instance whereby either an item (variable) or unit (person/entire observation) is missing from the sample
- Unit Missing – entire observation/individual has refused to participate or can no longer participate in the survey (quite common in longitudinal contexts)
- Item Missing – whereby a unit has taken part in the survey but has not answered a particular item or question
- Missingness attributed to death or emigration is considered ‘natural’ from the original sample
- Those, however, that either refuse continued survey participation or complete surveys partially may present problems of biased estimates

MISSINGNESS MECHANISMS

- The primary concern surrounding missing data is that dependent upon the type of missingness, there is a potential to affect inferences made by the analysis of studies (Hawkes and Plewis, 2006: 479; Silverwood *et al.*, 2021).

- The first step for any analyst seeking to ‘handle’ this missing data is to assess if that missingness is hiding a particular set of values that is meaningful for their analysis (Little, Carpenter and Lee, 2022).

- If there is an identified pattern within the missing data present within a given analysis, this has the potential to influence results and alter substantive conclusions had this missingness not occurred – or been appropriately accounted for

MISSINGNESS MECHANISMS

- Three missingness mechanisms:
- MCAR – Missing Completely At Random
- MAR – Missing At Random
- MNAR – Missing Not At Random

MISSINGNESS MECHANISMS - 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

- Evaluating the assumption that missingness on Y depends on some observed variable in X is straightforward. Allison (2012) uses the example of income depending on gender by testing whether the proportions of men and women who report their income differ – a logistic regression in which the dependent variable is the response indicator could be estimated, and significant coefficients would suggest a violation of the MCAR mechanism (ibid)

- Testing whether missingness on Y does not depend on Y itself is much more complicated. Unless we have existing linked data such as tax records in the income example, it is almost impossible to evaluate this assumption.

MISSINGNESS MECHANISMS - MCAR

- The upside of an MCAR mechanism is that estimated coefficients will not provide biased results in the presence of data Missing Completely At Random, however their estimation may be less precise in the form of inflated standard errors (Kang, 2013).
- MCAR = inflated standard errors
- Complete Records Analysis is perfectly fine under a MCAR mechanism

MISSINGNESS MECHANISMS - 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
- Again, Y on Y dependence is VERY difficult to find evidence for

MISSINGNESS MECHANISMS – MNAR

- MNAR depends on unobserved values (Silverwood et al. 2021)
- 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

INVESTIGATION OF MISSINGNESS

- There should always be a section detailing missingness in descriptive statistics

- Do NOT use the line “complete records were used and missingness was looked at” when all you did was run a tabulate command

- A few ways to assess the missingness on: Y is missing does depend on X
  - Generate binary variable where 1=missing and 0=non-missing observations and run a logit
  - Similar method with a t-test or chi2 if y = category
  - Use custom built commands like mcartest in Stata (Little 1998; Li 2013) or mcar_test via njtierney’s naniar package in R

- You would need linked administrative records to assess Y missing dependent on Y assumption
  - Sometimes exists
  - Sometimes key variables like housing tenure provide easy access admin records

HOW MUCH MISSINGNESS IS AN ISSUE?

- Impossible question to know unless missingness mechanism is confirmed (Hyuk Lee and Huber Jr 2021)
- Some literature uses the 10% (Bennet 2001) or even 5% rule (Schafer 1999)
- The proportion of missing data should not be used to guide decisions on handling missing data (Madley-Dowd et al 2019)

INVESTIGATION OF MISSINGNESS - PRACTICAL

- Three datasets are provided to you:
  - Each dataset has a different missingness mechanism
  - All variables in each dataset are planned to be used in a hypothetical analytical model

- Using the tools provided, assess these mechanisms

- In this poll tell me which dataset goes with which missingness mechanism

- A few tips:
  - Always create some descriptive statistics:
    - Tables or figures
  - Try a t-test,chi2, logistic regression, or in Stata or R the custom-built command mcartest

INVESTIGATION OF MISSINGNESS – PRACTICAL

Table 1: Regression Models

	God Model	
x1	25.53	***
	(3.74)	
x2	500.05	***
	(92.17)	
x3	64.42	***
	(9.37)	
Intercept	2948.32	*
	(1403.80)	
Number of observations	1000	
AIC	17390.56	
BIC	17410.19	
Adjusted R-squared	0.12	

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

Data Source: Simulated

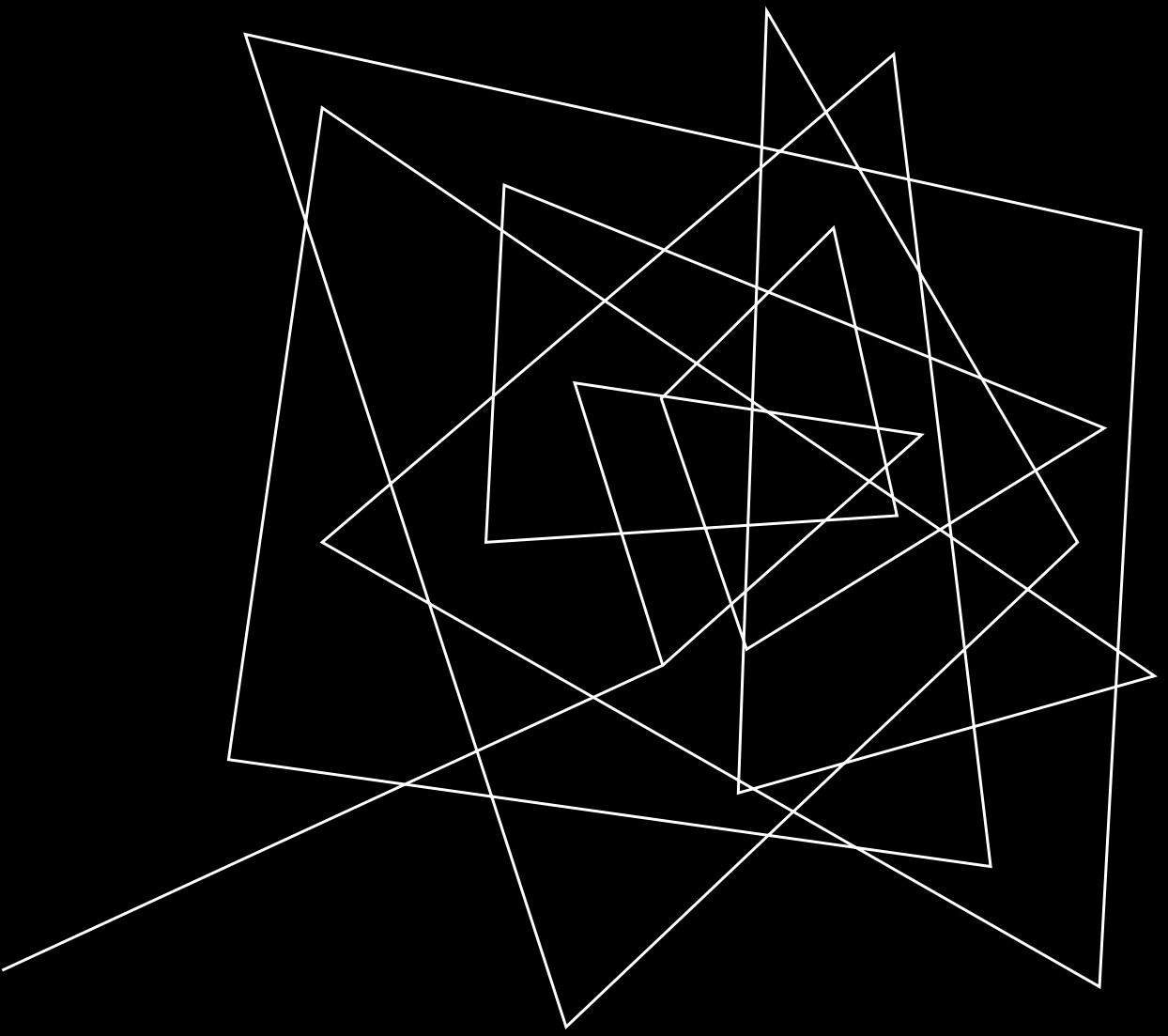
INVESTIGATION OF MISSINGNESS – PRACTICAL

Table 2: Regression Models

	God Model		MCAR Model		MAR Model		MNAR Model	
x1	25.53	***	26.08	***	6.75	*	19.17	***
	(3.74)		(5.17)		(3.21)		(5.55)	
x2	500.05	***	633.89	***	239.07	**	588.31	*
	(92.17)		(127.96)		(78.69)		(256.85)	
x3	64.42	***	69.95	***	11.99		71.13	***
	(9.37)		(12.92)		(8.18)		(13.87)	
Intercept	2948.32	*	2080.68		10252.38	***	2201.78	
	(1403.80)		(1934.89)		(1221.25)		(2067.43)	
Number of observations	1000		531		389		474	
AIC	17390.56		9242.44		6280.20		8268.96	
BIC	17410.19		9259.54		6296.05		8285.61	
Adjusted R-squared	0.12		0.13		0.03		0.09	

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

Data Source: Simulated



SESSION TWO – WAYS TO HANDLE MISSING DATA

SESSION TWO – WAYS TO HANDLE MISSING DATA

- Inadequate Standard
- Less than gold standard
- Gold standard
- Practical
- (Assumption under a MAR mechanism going forward)

INADEQUATE STANDARD – LISTWISE DELETION

- The CRA approach is unpredictable; there is no way to know the consequences of this loss of information if data is found to be MAR (Carpenter and Kenward, 2012)

- Only viable under two conditions:
  - There is no missingness
  - Missingness is MCAR

INADEQUATE STANDARD – LISTWISE DELETION

Table 2: Regression Models

	God Model		MCAR Model		MAR Model		MNAR Model	
x1	25.53	***	26.08	***	6.75	*	19.17	***
	(3.74)		(5.17)		(3.21)		(5.55)	
x2	500.05	***	633.89	***	239.07	**	588.31	*
	(92.17)		(127.96)		(78.69)		(256.85)	
x3	64.42	***	69.95	***	11.99		71.13	***
	(9.37)		(12.92)		(8.18)		(13.87)	
Intercept	2948.32	*	2080.68		10252.38	***	2201.78	
	(1403.80)		(1934.89)		(1221.25)		(2067.43)	
Number of observations	1000		531		389		474	
AIC	17390.56		9242.44		6280.20		8268.96	
BIC	17410.19		9259.54		6296.05		8285.61	
Adjusted R-squared	0.12		0.13		0.03		0.09	

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

Data Source: Simulated

INADEQUATE STANDARD – SINGLE IMPUTATION

- Single use comes in many forms:
  - Mean
  - Modal
  - Median

- In the example of a categorical variable, takes the mode of the value in said variable and imputes that modal value across all missing values in the data.

- Single imputation ignores all uncertainty and always underestimates the variance in each model

- Advocates of this approach argue that whilst not perfect this approach doesn’t delete a single case and incorporates all available information into a given model

INADEQUATE STANDARD – SINGLE IMPUTATION

- There is a possibility that the estimates from this method may fall close to the true range; of course, the exact opposite is equally likely

- Conclusively shown to perform poorly except under exceptionally special conditions (Collins, Schafer and Kam, 2001; Little and Rubin, 2019)

INADEQUATE STANDARD – SINGLE MEAN IMPUTATION

Table 3: Regression Models

	God Model		MAR Model		Single Mean Imp	
x1	29.69	***	15.10	***	30.78	***
	(3.83)		(4.45)		(6.08)	
x2	38.91	***	25.16	***	25.71	***
	(0.93)		(1.34)		(2.68)	
x3	49.10	***	21.94	*	44.91	**
	(9.64)		(10.88)		(15.26)	
Intercept	5293.88	***	11609.07	***	9343.46	***
	(1464.51)		(1661.95)		(2349.47)	
Number of observations	1000		505		1000	
AIC	17443.09		8571.30		18362.90	
BIC	17462.72		8588.19		18382.53	
Adjusted R-squared	0.64		0.41		0.11	

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

Data Source: Simulated

INADEQUATE STANDARD – SINGLE MODE IMPUTATION

Table 4: Regression Models

	God Model		MAR Model		Single Modal Imp	
x1	25.53	***	6.75	*	17.65	***
	(3.74)		(3.21)		(3.49)	
x2						
0	0.00		0.00		0.00	
	(0.00)		(0.00)		(0.00)	
1	500.05	***	239.07	**	-1572.50	***
	(92.17)		(78.69)		(109.82)	
x3	64.42	***	11.99		56.31	***
	(9.37)		(8.18)		(8.68)	
Intercept	2948.32	*	10252.38	***	5057.75	***
	(1403.80)		(1221.25)		(1306.51)	
Number of observations	1000		389		1000	
AIC	17390.56		6280.20		17232.49	
BIC	17410.19		6296.05		17252.13	
Adjusted R-squared	0.12		0.03		0.24	

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

Data Source: Simulated

INADEQUATE STANDARD – SINGLE MEDIAN IMPUTATION

Table 5: Regression Models

	God Model		MAR Model		Single Median Imp	
x1	25.54	***	6.78	*	17.19	***
	(3.74)		(3.40)		(2.88)	
x2						
1	0.00		0.00		0.00	
	(0.00)		(0.00)		(0.00)	
2	467.51	**	96.80		-86.40	
	(145.27)		(112.71)		(172.89)	
3	782.81	***	182.41		1890.74	***
	(143.52)		(118.67)		(129.45)	
4	1095.82	***	287.00	*	-268.02	
	(147.55)		(129.57)		(175.75)	
5	1441.84	***	396.85	*	-139.67	
	(148.16)		(172.64)		(172.40)	
x3	63.96	***	6.17		36.05	***
	(9.39)		(8.97)		(7.26)	
Intercept	3513.22	*	11432.78	***	6476.20	***
	(1412.78)		(1345.57)		(1090.10)	
Number of observations	1000		204		1000	
AIC	17394.72		3193.57		16863.43	
BIC	17429.07		3216.80		16897.79	
Adjusted R-squared	0.17		0.03		0.46	

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

Data Source: Simulated

LESS THAN GOLD STANDARD – DUMMY VARIABLE ADJUSTMENT

- Dummy variable adjustment is where all missingness at the given variable is coded to a value within the model. In the example of a binary dummy variable, all missingness is coded to either equal zero or equal one
- Incorporates all information into the regression model
- For the simple model of data missing at Y variable, a dummy variable adjustment will not provide the ‘true’ estimates but if the complete records analysis is compared to a model where all missingness equals zero and another model where all missingness equals one, then the range of the estimates can be located
- Jones (1996) demonstrated that dummy variable adjustment yields biased parameter estimates even when the data is MCAR
- The ability to provide a range of the estimates does provide some utility to this technique

LESS THAN GOLD STANDARD – DUMMY VARIABLE ADJUSTMENT

- iff the complete case analysis and both dummy variable adjustment models present a beta coefficient that is throughout all models positive, one can present those results similar to how we ought to interpret log odds

- The results would present evidence for a positive coefficient – though the exact size is unknown, some information can be gathered and reported

- This technique has the most utility in scenarios where missingness is so great that it begins to stretch the abilities of even gold-standard techniques

LESS THAN GOLD STANDARD – DUMMY VARIABLE ADJUSTMENT

Table 6: Regression Models

	God Model		MAR Model		Dummy = 0		Dummy = 1	
x1	25.53	***	6.75	*	17.65	***	21.65	***
	(3.74)		(3.21)		(3.49)		(3.23)	
x2								
0	0.00		0.00		0.00		0.00	
	(0.00)		(0.00)		(0.00)		(0.00)	
1	500.05	***	239.07	**	-1572.50	***	1926.59	***
	(92.17)		(78.69)		(109.82)		(99.20)	
x3	64.42	***	11.99		56.31	***	47.21	***
	(9.37)		(8.18)		(8.68)		(8.15)	
Intercept	2948.32	*	10252.38	***	5057.75	***	4432.72	***
	(1403.80)		(1221.25)		(1306.51)		(1215.79)	
Number of observations	1000		389		1000		1000	
AIC	17390.56		6280.20		17232.49		17098.55	
BIC	17410.19		6296.05		17252.13		17118.18	
Adjusted R-squared	0.12		0.03		0.24		0.34	

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

Data Source: Simulated

LESS THAN GOLD STANDARD – SURVEY WEIGHTS

- Inverse Probability Weighting (IPW)
- Creates weighted copies of complete records to remove selection bias introduced by missing data
- IPW only determines weights from incomplete cases and partially observed cases are discarded in the weighted analysis
- Weighted estimates can have unacceptably high variance (Seaman et al., 2012; Seaman and White, 2013; Little, Carpenter and Lee, 2022)
- Weights are a great addition to, though not a substitution for gold standard approaches
- Won’t be discussed more here…

GOLD STANDARD – FULL INFORMATION MAXIMUM LIKELIHOOD

- Uses maximum likelihood function based on the observed data for each individual, utilising all available information possible
- Aims to find parameter values that best explain observed data, given a specific model structure
- FIML does not use imputation; it directly uses the probability of observing the data given the model's parameters
- Exceptionally easy to implement compared to the other estimation procedures discussed (Enders, 2001)

GOLD STANDARD – FULL INFORMATION MAXIMUM LIKELIHOOD

- One downside of the FIML approach is all variables within a FIML based model are assumed to have multivariate normality
- In theory, FIML can be extended to non-linear outcomes, by treating categorical data under a multinominal sampling distribution (Vermunt, 1997; Edwards, Berzofsky and Biemer, 2018)
- The majority of common place statistical software does not accommodate this, Latent Gold, and Mplus are two exceptions (Muthen and Muthen, 2017)

GOLD STANDARD – FULL INFORMATION MAXIMUM LIKELIHOOD

Table 7: Regression Models

	God Model		MAR Model		FIML	
x1	29.69	***	15.10	***	27.17	***
	(3.83)		(4.45)		(5.01)	
x2	38.91	***	25.16	***	39.34	***
	(0.93)		(1.34)		(1.24)	
x3	49.10	***	21.94	*	32.90	**
	(9.64)		(10.88)		(12.29)	
Intercept	5293.88	***	11609.07	***	7726.20	***
	(1464.51)		(1661.95)		(1864.23)	
mean(x1)					40.17	***
					(0.39)	
mean(x2)					203.17	***
					(2.49)	
mean(x3)					150.10	***
					(0.15)	
var(e.y)					2.1e+06	
					(1.7e+05)	
var(x1)					149.86	
					(6.70)	
var(x2)					2503.52	
					(194.45)	
var(x3)					23.67	
					(1.06)	
cov(x1,x2)					-1.30	
					(23.13)	
cov(x1,x3)					2.44	
					(1.88)	
cov(x2,x3)					6.26	
					(9.04)	
Number of observations	1000		505		1000	
AIC	17443.09		8571.30		37178.94	
BIC	17462.72		8588.19		37247.65	
Adjusted R-squared	0.64		0.41			

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

Data Source: Simulated

GOLD STANDARD – MULTIPLE IMPUTATION

- Multiple imputation generates replacement values or imputations for the missing data values and repeats this procedure over many iterations to produce a ‘semi-Bayesian’ framework for the most appropriate fit of estimates

- For multiple imputation models to be compared to a complete records analysis, the former needs to be ‘‘congenial’’ (White, Royston and Wood, 2011) with the latter
  - variables in the complete record analysis are identical to those included in multiple imputation

- The correct variance/covariance matrix will not be estimated, and a substantive comparison between the two will become impossible and impracticable due to a loss of statistical power (Von Hippel, 2009; Lynch and Von Hippel, 2013)

- Multiple Imputation can be implemented easily and readily across software platforms, unlike FIML

GOLD STANDARD – MULTIPLE IMPUTATION

- It can be a lengthy procedure that has the potential to induce human error due to the need to select auxiliary variables, set the correct data for imputation, and set the correct seed for replication etc

- There is also a time efficiency argument, whereby for multiple imputation, if the dataset is large, or there is large amounts of missingness, then the time to impute the model of interest can take a large amount of time.

- MI is an attractive method because it is practical and widely applicable (Carpenter and Kenward, 2012)

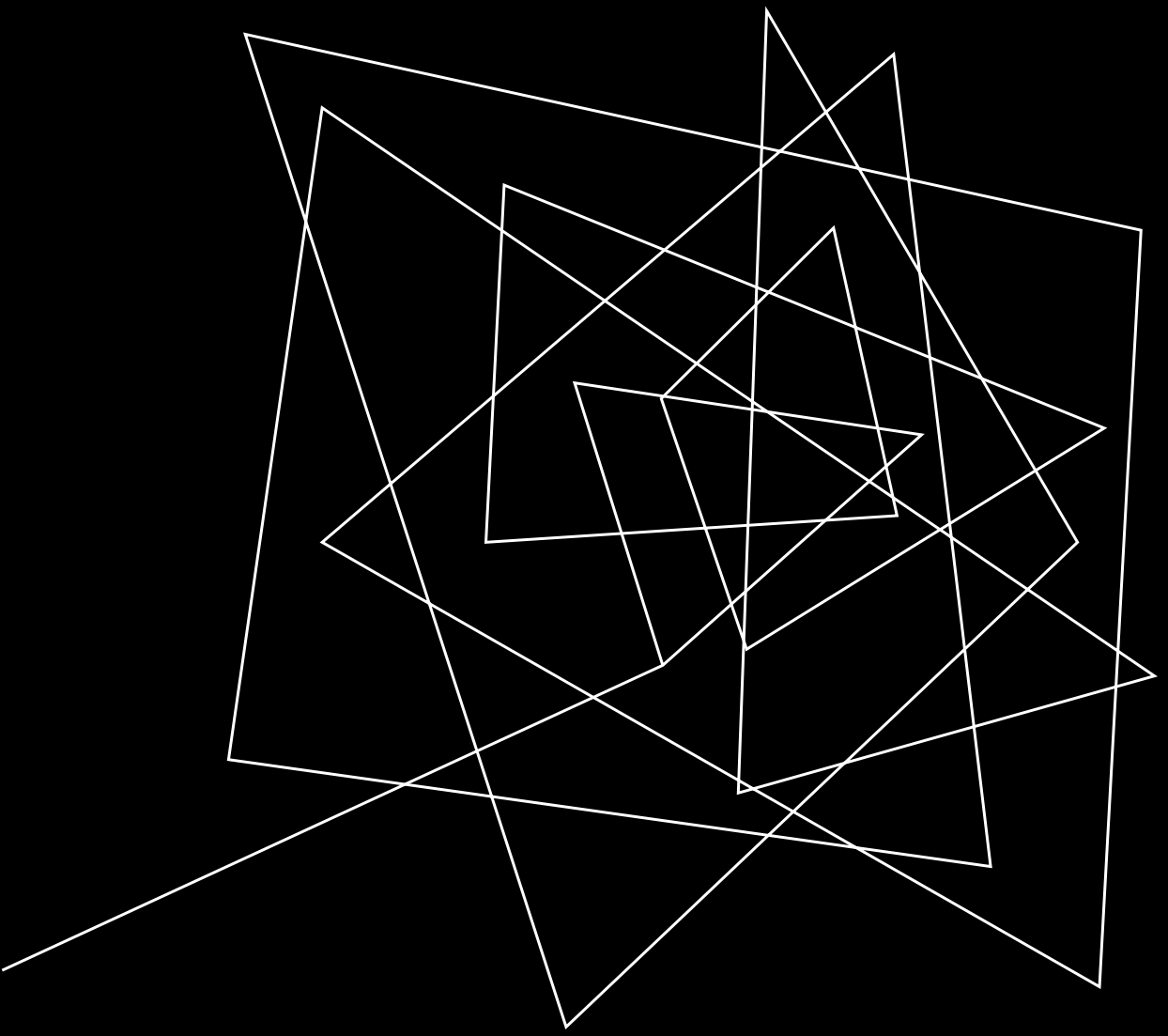
GOLD STANDARD – MULTIPLE IMPUTATION

Table 8: Regression Models

	God Model		MAR Model		MI	
x1	29.69	***	15.10	***	29.75	***
	(3.83)		(4.45)		(4.72)	
x2	38.91	***	25.16	***	39.25	***
	(0.93)		(1.34)		(1.33)	
x3	49.10	***	21.94	*	34.95	**
	(9.64)		(10.88)		(12.82)	
Intercept	5293.88	***	11609.07	***	7335.05	***
	(1464.51)		(1661.95)		(1932.89)	
Number of observations	1000		505		1000	
AIC	17443.09		8571.30			
BIC	17462.72		8588.19			
Adjusted R-squared	0.64		0.41		.	

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

Data Source: Simulated



SESSION THREE – MAXIMUM LIKELIHOOD

SESSION THREE –MAXIMUM LIKELIHOOD

- What is Maximum likelihood?
- How to run?
- Different types
- Practical

SESSION THREE –WHAT IS ML?

- Process of estimating the parameters of a distribution that maximises the likelihood of the observed data belonging to that distribution
- How likely a distribution with certain values for its parameters fits our data
- Very efficient

SESSION THREE –WHAT IS ML?

- The likelihood function:

$$L(\theta)=P(x1,x2,...,xn|\theta)$$

- This is the probability density of the observed data given the unknown parameter θ
- Maximum likelihood chooses the value of θ that maximises the likelihood function
- Often times we maximise the log-likelihood as it is statistically more efficient
- MLE for linear regression gives the same estimates as OLS
- Also forms the foundation of a lot of Bayesian inference

SESSION THREE –HOW TO RUN ML?

- Maximum Likelihood mainly runs via sem in Stata (lavaan in R)
- SEM has three estimation methods to use
  - ml – maximum likelihood
  - mlmv – maximum likelihood with missing values (FIML)
  - adf – asymptotic distribution free – non-linear

- MLE for linear regression gives the same estimates as OLS

SESSION THREE –HOW TO RUN ML?

Table 8: Regression Models

	OLS God Model		ML God Model		FIML God Model		ADF God Model	
x1	29.69	***	29.69	***	29.69	***	29.69	***
	(3.83)		(3.82)		(3.82)		(3.78)	
x2	38.91	***	38.91	***	38.91	***	38.91	***
	(0.93)		(0.93)		(0.93)		(1.00)	
x3	49.10	***	49.10	***	49.10	***	49.10	***
	(9.64)		(9.62)		(9.62)		(9.32)	
Intercept	5293.88	***	5293.88	***	5293.88	***	5293.88	***
	(1464.51)		(1461.58)		(1461.58)		(1451.35)	
var(e.y)			2.2e+06		2.2e+06		2.2e+06	
			(97724.76)		(97724.76)		(95324.90)	
Number of observations	1000		1000		1000		1000	
AIC	17443.09		41960.12		41960.12			
BIC	17462.72		41984.66		41984.66			
Adjusted R-squared	0.64							

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

Data Source: Simulated

SESSION THREE –MAXIMUM LIKELIHOOD

- There are currently three ML estimation algorithms for use when missing data is present with either an MCAR or MAR mechanism
  - Multi-group
  - Expectation-maximisation
  - Full-information Maximum Likelihood

SESSION THREE –MULTI-GROUP

- The sample is divided into subgroups which each share the same pattern of missing data
- A likelihood function is computed for each of the subgroups and the groupwise likelihood functions are accumulated across the entire sample and maximised
- There are some practical issues of implementing this multiple-group based ML approach (Enders, 2001)
- The major drawback of this approach is that it is a group level, rather than individual level ML estimation

SESSION THREE –EXPECTATION-MAXIMISATION

- This estimation uses a two-step iterative procedure where missing observations are filled in or imputed and the unknown parameters are estimated using maximum likelihood missing data algorithms

- The EM approach can only be used to obtain ML estimates of a mean vector and covariance matrix and as a result standard errors will be negatively biased and bootstrapping is recommended (Enders, 2001)

SESSION THREE –FULL-INFORMATION MAXIMUM LIKELIHOOD

- It has also been called the raw maximum likelihood estimation for its likelihood function being calculated at the individual
- It is also exceptionally easy to implement compared to the other estimation procedures discussed (Enders, 2001)
- One downside of the FIML approach is all variables within a FIML based model are assumed to have multivariate normality
- In theory FIML can be extended to non-linear outcomes, by treating categorical data under a multinominal sampling distribution (Vermunt, 1997; Edwards, Berzofsky and Biemer, 2018)
  - We shall test this...

SESSION THREE –FULL-INFORMATION MAXIMUM LIKELIHOOD

Table 10: FIML Metric Regression Models

	ML God Model		FIML Model	
x1	29.69	***	27.17	***
	(3.82)		(5.01)	
x2	38.91	***	39.34	***
	(0.93)		(1.24)	
x3	49.10	***	32.90	**
	(9.62)		(12.29)	
Intercept	5293.88	***	7726.20	***
	(1461.58)		(1864.23)	
var(e.y)	2.2e+06		2.1e+06	
	(97724.76)		(1.7e+05)	
mean(x1)			40.17	***
			(0.39)	
mean(x2)			203.17	***
			(2.49)	
mean(x3)			150.10	***
			(0.15)	
var(x1)			149.86	
			(6.70)	
var(x2)			2503.52	
			(194.45)	
var(x3)			23.67	
			(1.06)	
cov(x1,x2)			-1.30	
			(23.13)	
cov(x1,x3)			2.44	
			(1.88)	
cov(x2,x3)			6.26	
			(9.04)	
Number of observations	1000		1000	
AIC	41960.12		37178.94	
BIC	41984.66		37247.65	

**** p<.001, ** p<.01, * p<.05*
Data Source: Simulated

SESSION THREE –FULL-INFORMATION MAXIMUM LIKELIHOOD

Table 11: FIML Binary Regression Models

	ML God Model		FIML Model	
x1	25.53	***	26.59	***
	(3.73)		(3.85)	
x2	500.05	***	798.78	***
	(91.99)		(229.24)	
x3	64.42	***	60.69	***
	(9.35)		(10.23)	
Intercept	2948.32	*	3251.25	*
	(1400.99)		(1485.47)	
var(e.y)	2.1e+06		2.0e+06	
	(92723.28)		(1.3e+05)	
mean(x1)			40.17	***
			(0.39)	
mean(x2)			0.62	***
			(0.05)	
mean(x3)			150.16	***
			(0.15)	
var(x1)			149.86	
			(6.70)	
var(x2)			0.26	
			(0.02)	
var(x3)			23.89	
			(1.07)	
cov(x1,x2)			-0.28	
			(0.33)	
cov(x1,x3)			4.03	*
			(1.90)	
cov(x2,x3)			0.20	
			(0.13)	
Number of observations	1000		1000	
AIC	32678.53		31844.96	
BIC	32703.07		31913.67	

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

Data Source: Simulated

SESSION THREE –FULL-INFORMATION MAXIMUM LIKELIHOOD

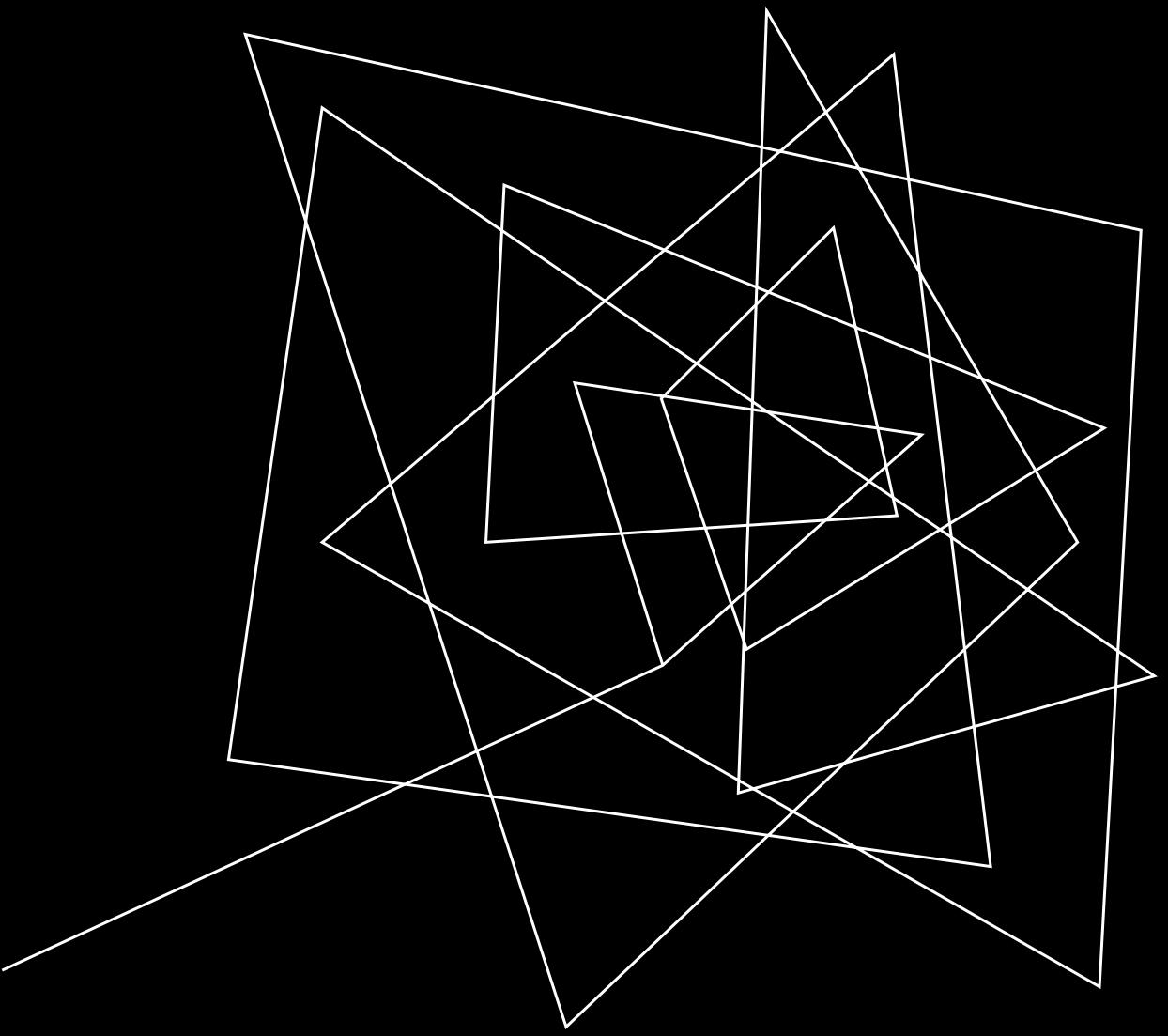
Table 12: FIML Nominal Category Regression Models

	ML God Model		FIML Model	
x1	25.54	***	24.20	***
	(3.73)		(5.26)	
x2== 2.0000	467.51	**	506.79	
	(144.76)		(564.26)	
x2== 3.0000	782.81	***	954.98	
	(143.02)		(556.12)	
x2== 4.0000	1095.82	***	1502.55	**
	(147.03)		(536.02)	
x2== 5.0000	1441.84	***	2077.64	**
	(147.64)		(698.95)	
x3	63.96	***	54.69	***
	(9.36)		(14.11)	
Intercept	3513.22	*	4551.44	*
	(1407.82)		(2082.15)	
var(e.y)	2.1e+06		1.8e+06	
	(92553.08)		(2.7e+05)	
Number of observations	1000		1000	
AIC	34540.03		31942.91	
BIC	34579.29		32060.70	

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

Data Source: Simulated

Note: var cut off for visibility



SESSION FOUR – MULTIPLE IMPUTATION

SESSION FOUR – MULTIPLE IMPUTATION

- Types of Multiple Imputation
- Auxiliary Variables
- MI Preparation
- Setting/Register Data
- Imputation Procedure
  - Burnin
  - Seed
  - Iterations
- Diagnostics

SESSION FOUR – MULTIPLE IMPUTATION

- Is it making up data?
- What is passive Imputation?
- How to treat transformations?
- What variables need to be included?
- How much missingness is possible?
- How many imputations?

SESSION FOUR – MULTIPLE IMPUTATION

- Multiple imputation generates replacement values or imputations for the missing data values and repeats this procedure over many iterations to produce a ‘semi-Bayesian’ framework for the most appropriate fit of estimates

- For multiple imputation models to be compared to a complete records analysis, the former needs to be ‘‘congenial’’ (White, Royston and Wood, 2011) with the latter

- Congeniality or consistency in this respect means that the same variables in the complete record analysis are identical to those included in multiple imputation

SESSION FOUR – MULTIPLE IMPUTATION

- There are two main types of MI in statistical analysis
  - Multivariate normal Model (MVN)
  - Multivariate imputation by Chained Equations (MICE)

- The main advantage of MI approaches over other gold-standard methods is that it offers an alternative that allows non-linear dependent and individual variables within models

- We will be focusing on MICE for the most part

SESSION FOUR – MULTIPLE IMPUTATION

- Multiple imputation does, however, have some drawbacks
- It can be a lengthy procedure that has the potential to induce human error due to the need to select auxiliary variables, set the correct data for imputation, and set the correct seed for replication etc
- There is also a time efficiency argument, whereby for multiple imputation, if the dataset is large, or there is large amounts of missingness, then the time to impute the model of interest can take a large amount of time

SESSION FOUR – LANGUAGE

- Setting
- Registering
- Imputation procedure
- Iterations/Imputations
- Burn-in
- Seed
- Relative Increases in Variance (RVI) - Proportional increase in total sampling variance that is due to missing information
- Fraction of Missing Information (FMI) - Proportion of the total sampling variance that is due to missing data

SESSION FOUR – MICE

- MICE is a form of multiple imputation that fills in or imputes missing data within a given dataset through iterative predictive models or k imputations
- This specification is required when imputing a variable that must only take on specific values, such as the categorical nature of sex for example
- Using MICE, each imputation k is drawn from the posterior distribution of the parameters in the given imputation model, and then the model itself is imputed (Carpenter and Kenward, 2012)
- To create the kth imputation, new parameters are drawn from the posterior distribution
- An essential advantage of MI is that it can be applied for data missing at the response variable or its covariates (Carpenter and Kenward, 2012)

SESSION FOUR – AUXILIARY VARAIBLES

- MI uses auxiliary variables – variables not included in the main model but are used when setting the data to be imputed

- The auxiliary variables' main function is to improve the predictive ability of the imputation model over and above the information recovered from just using the information provided by the analytical variables in the model (Collins, Schafer and Kam, 2001)

- Auxiliary variables are essential when there are high levels of missingness upon a given variable (Johnson and Young, 2011; Young and Johnson, 2011)

- There is no strict threshold for what an auxiliary variable needs to be included within the imputation
  - some have recommended an $r > 0.4$ on at least one of the analytical variables within the model (Allison, 2012a)
    - Though this is disputed (Enders, 2010)

SESSION FOUR – AUXILIARY VARAIBLES

- There is no strict threshold for what an auxiliary variable needs to be included within the imputation
  - some have recommended an r > 0.4 on at least one of the analytical variables within the model (Allison, 2012a)
    - Though this is disputed (Enders, 2010)

  - Others, such as Silverwood et al. (2021), argue that if an auxiliary variable is predictive of the outcome variable, it makes them suitable for inclusion within the imputation model

- An auxiliary variable does not have the requirement that the given variable has to have complete information to be valuable – auxiliary variables can still be influential when they have missingness (Enders, 2010)

SESSION FOUR – IS IT MAKING UP DATA?

- This argument can be used for single imputation approaches
- Multiple imputation builds on uncertainty associated with missing data and as such never presents a single value but multiple iterations produce possibilities of values that are subsequently averaged over the imputation procedure

SESSION FOUR – WHAT IS PASSIVE IMPUTATION?/HOW DO I TREAT TRANSFORMATIONS?

- Passive variables are functions of imputed variables
- Impute then transform?
- Just Another Variable (JAV)
- If working in linear context – JAV, if non-linear, conduct sensitivity analysis

SESSION FOUR – WHAT VARIABLES NEED TO BE INCLUDED?

- All analytical variables!
- Including dependent variable!
- And any auxiliary variables

SESSION FOUR – HOW MUCH MISSINGNESS?

- Contemporary studies and simulations have increasingly stretched and stress-tested the limits of MI (Hardt et al., 2013)
  - 50% was once thought to be the limit of many gold standard procedures

- A simulation by Hardt et al (2013) demonstrated that large amounts of missingness can be present within a model without breaking down MI or FIML mechanisms (ibid).

- Imputation-based models are consistently found to outperform a CRA in both absolute bias and Root Mean Squared Error (RMSE) with increasing levels of missingness (Hyuk Lee and Huber Jr., 2021)

- The most extreme case from Madely-Down et al (2019) demonstrates that so long as the imputation model is properly specified and data are MAR then unbiased results can be obtained even with up to 90 per cent missingness
  - An imputation model compared to a CRA can achieve a reduction in 99.97 per cent bias when missingness is at 90 per cent (ibid)

SESSION FOUR – HOW MANY IMPUTATIONS?

- Traditional literature suggests anywhere from 5-10 iterations is sufficient (White, Royston and Wood, 2011)

- With more contemporary studies suggesting upwards of 50 (Silverwood et al. 2021)

- White et al. (2010) suggests using the Fraction of Missing Information (FMI) statistics as a baseline
  - The maximum FMI suggests the threshold for the number of imputations
  - This is shown to reduce standard errors and stabilise p-values

- Bodner (2008) has provided one of the first concrete procedures

SESSION FOUR – HOW MUCH MISSINGNESS? – BODNER (2008)

TABLE 2
Values of 95% Interpercentile Ranges of 5,000 Simulated 95% Confidence Interval Half-Widths, Null Hypothesis Significance Test *p* Values, and Fractions of Missing Information as a Function of the True Fraction of Missing Information and the Number of Imputations

		λ						
	m	*.05*	*.10*	*.20*	*.30*	*.50*	*.70*	*.90*
$\hat{h}_m$	3	.04	.08	.17	.30	.62	1.24	3.12
	5	.02	.04	.09	.15	.30	.57	1.48
	10	.01	.02	.05	.08	.17	.31	.82
	20	.01	.02	.03	.05	.10	.19	.53
	30	.01	.01	.03	.04	.08	.15	.43
	40	.01	.01	.02	.04	.07	.13	.36
	50	.01	.01	.02	.03	.06	.12	.33
	75	.00	.01	.02	.03	.05	.09	.27
	100	.00	.01	.01	.02	.04	.08	.23
	∞	.00	.00	.00	.00	.00	.00	.00
$\hat{p}_m$	3	.08	.15	.26	.39	.62	.89	.96
	5	.06	.10	.18	.28	.49	.74	.87
	10	.04	.06	.12	.18	.34	.55	.75
	20	.03	.04	.08	.13	.24	.40	.62
	30	.02	.03	.06	.10	.19	.32	.50
	40	.02	.03	.06	.08	.16	.29	.44
	50	.02	.03	.05	.07	.15	.25	.38
	75	.01	.02	.04	.06	.12	.21	.32
	100	.01	.02	.03	.05	.11	.18	.28
	∞	.00	.00	.00	.00	.00	.00	.00
$\hat{\lambda}_m$	3	.24	.42	.64	.76	.86	.88	.72
	5	.15	.28	.46	.58	.69	.65	.35
	10	.10	.17	.32	.40	.48	.41	.20
	20	.06	.12	.21	.28	.32	.28	.13
	30	.05	.10	.17	.22	.27	.22	.11
	40	.04	.08	.15	.19	.23	.19	.09
	50	.04	.07	.13	.17	.20	.17	.08
	75	.03	.06	.11	.14	.16	.14	.07
	100	.03	.05	.09	.12	.14	.12	.06
	∞	.00	.00	.00	.00	.00	.00	.00

Note. The 95% interpercentile range (IPR) is the difference in outcome values at the 97.5th and 2.5th percentiles.

SESSION FOUR – HOW MUCH MISSINGNESS? – BODNER (2008)

TABLE 3
Minimum Number of Imputations Needed for Estimated 95% Confidence Interval Half-Widths and Fractions of Missing Information to Achieve Specified Precision at Two Levels of Confidence

	$\hat{h}_m \in [\bar{h}_m \pm .1\bar{h}_m]$		$\hat{\lambda}_m \in [\bar{\lambda}_m \pm .1]$	
λ	*80% Confidence*	*95% Confidence*	*80% Confidence*	*95% Confidence*
.05	2	3	2	4
.10	3	6	5	9
.20	7	12	11	23
.30	12	24	18	36
.50	27	59	23	50
.70	50	114	16	36
.90	108	258	4	10

SESSION FOUR – MI PREPARATION

- Isolate valid missingness
  - Drop dead/migration units
- Missing table summaries and patterns to isolate problematic variables
- Identify missingness mechanisms if possible
- Identify auxiliaries via past papers or via regression, t-tests, correlations of missingness

SESSION FOUR – MI SETTING/REGISTER (STATA ONLY)

- mi set - used to set a regular Stata dataset to be an mi dataset
  - The data are recorded in a style: wide, mlong, flong, or flongsep
    - I suggest flong*
    - Wide - Stacks imputed values side-by-side, creating new variables for each imputed dataset, fast but memory-intensive for large datasets
    - Mlong - Combines data into a single dataset, but only includes observations with missing values in the original data. This is more memory-efficient than flong, but less efficient than wide for large datasets.
    - Flong - Combines data into a single dataset, including all original and imputed observations. It's a good balance between memory usage and data management, but can be slower for large datasets.
    - Flongsep - Each imputed dataset is stored as a separate .dta file. This is the most flexible for data management but can be inefficient

- Mi register - Variables are registered as imputed, passive, or regular, or they are left unregistered

SESSION FOUR – MI SETTING/REGISTER (STATA ONLY)

- Conduct your imputation via mi impute chained
- For each type of variable you will need to specify and augment the imputation:
- (logit, augment) etc

- Working example

```
mi impute chained ///
///
(logit, augment) obin sex tenure maw5 aconnn512 genability11 toilet itoilet cooking water dconnn2492 econbin ///
///
(mlogit, augment) nssec ///
///
, rseed(12346) dots force add(40) burnin(20) savetrace(MI_test_trace, replace)
```

- Follow up by now running your imputation regression:

```
mi estimate, saving(miestfile1, replace) esample(misample1) post dots: logit econbin i.obin i.sex i.tenure ib(3).nssec
```

SESSION FOUR – MI (R ONLY)

- Actually, a lot simpler than Stata version
- No need to set or register data, simply create a dataframe and parse that through mice
- Example:
- Imp <- mice(data, m = 10, method = “pmm”, seed = 12345)
- Diagnostics:
- Summary(imp) – summary model statistics
- Plot(imp) – trace plots
- Densityplot(imp) – compare distributions

SESSION FOUR – MI (R ONLY)

- Warning: pmm = predictive mean matching
- If we have multiple different types of variables in R we need to change the method =
- To do this we create a ‘method vector’
- Each analytical variable is assigned its own specific method:

Var Type	Method Name	Description
Continuous	Pmm	Predictive mean matching
	Norm	Linear regression
Binary	Logreg	Logistic regression
Ordinal	Polr	Proportional odds
Nominal	Polyreg	multinomial
Count	Poission	Poisson regression

SESSION FOUR – MI (R ONLY)

- Auxiliaries are also a bit of a pain in R
- Have to make sure we use quickpred to include them as predictors in the imputation
- Also have to make sure we don’t impute them via the method

SESSION FOUR – MI STATA VERSUS R

- Both have complexities
- Stata is a lot of MI based setup regarding setting up data and registering it
- R is a lot of dataframe setup regarding methods and auxiliaries
- The mice package in R does come with a lot of nice visuals and diagnostics that in Stata you have to do via other commands
  - Nice but not essential...

SESSION FOUR – REPORTING

- What software was used?
- Type of imputation – MVN or MICE
- Justification for the imputation method
- Number of iterations and burn-in
- Proportion of missing observations for each imputed variable
- Variables used in your imputation model
- Your imputation seed

SESSION FOUR – PRACTICAL

- Example simulated dataset with a variety of variables
- Take your time to parse through the code and deduce what is what
- We will come back and go line by line together
- See what happens if you start altering key aspects of MI

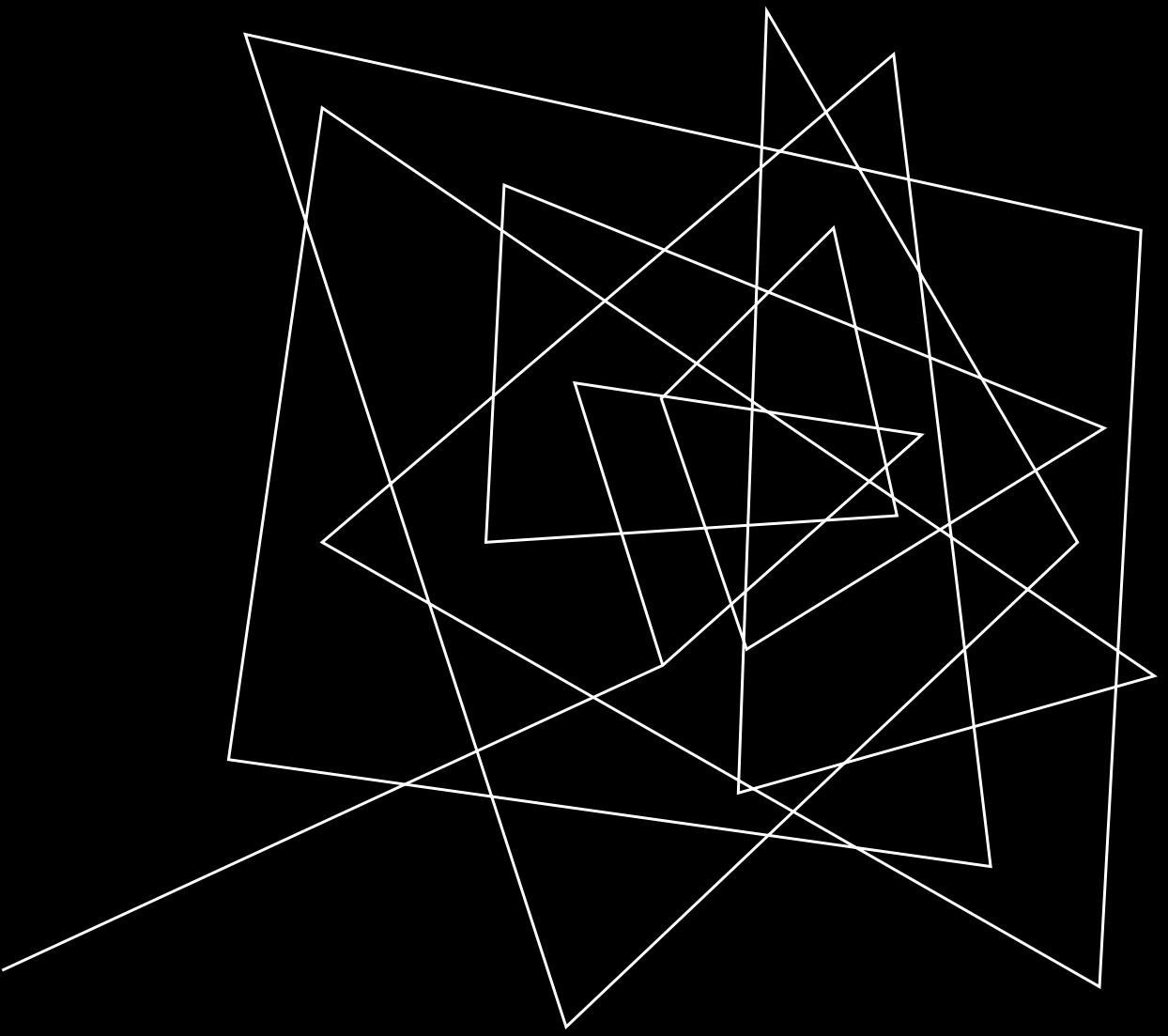
SESSION FOUR – FURTHER MATERIALS

- We have only scratched the surface on MI

- There are so many different possibilities for MI that now you have the basics, you can explore on your own

- Two powerful resources I've found:
- For Stata:
- https://stats.oarc.ucla.edu/stata/seminars/mi_in_stata_pt1_new/

- For R:
- https://www.jstatsoft.org/article/view/v045i03



SESSION FIVE – COMPARISON OF METHODS

SESSION FIVE – COMPARISON OF METHODS

- Comparison of gold standard techniques
- Basic side-by-side view
- A little simulation pre-print
  - Absolute Bias
  - % of SE Bias
  - Root Mean Square Error
  - CI Length
  - CI Coverage
  - Time of One Model

SESSION FIVE – COMPARISON OF METHODS

Table 13: Simulation Regression Models Using a MAR Principle

	Complete Records 'God Model'		**Complete SEM 'God Model'**		**MCAR**		**MAR**		**Single Mean Imputation**		**FIML**		**Imputed with no auxiliary variables and 10 imputations**		**Imputed with 10 imputations**		**Imputed with 100 imputations**	
	Coef.	CIs	Coef.	CIs	Coef.	CIs	Coef.	CIs	Coef.	CIs	Coef.	CIs	Coef.	CIs	Coef.	CIs	Coef.	CIs
$\beta_1 X_1$	30.01	[22.25, 37.77]	30.01	[22.27, 37.75]	29.97	[18.95, 40.99]	18.88	[10.21, 27.55]	33.93	[21.51, 46.34]	30.00	[20.07, 39.94]	23.26	[13.24, 33.29]	32.75	[22.14, 43.36]	31.45	[21.54, 30.51]
	(3.96)		(3.95)		(5.62)		(4.42)		(6.34)		(5.07)		(5.12)		(5.41)		(5.06)	
$\beta_2 X_2$	40.02	[38.15, 41.88]	40.02	[38.16, 41.88]	40.03	[37.39, 42.67]	25.14	[22.49, 27.80]	25.76	[20.43, 31.08]	40.03	[37.56, 42.50]	41.92	[39.40, 44.45]	40.52	[38.25, 42.78]	40.03	[37.40, 41.36]
	(0.95)		(0.95)		(1.35)		(1.35)		(2.72)		(1.26)		(1.29)		(1.16)		(1.34)	
$\beta_3 X_3$	49.88	[31.23, 68.53]	49.88	[31.27, 68.50]	51.30	[24.88, 77.72]	31.70	[11.00, 52.39]	56.64	[26.91, 86.38]	50.24	[26.57, 73.92]	44.37	[20.76, 67.97]	58.83	[35.73, 81.93]	55.42	[30.51, 80.33]
	(9.52)		(9.50)		(13.48)		(10.56)		(15.17)		(12.08)		(12.04)		(11.79)		(12.71)	
Number of observations	1000		1000		499		518		1000		1000		1000		1000		1000	

Note: Monte Carlo Simulation using a MAR mechanism. 51 per cent missingness introduced.

SESSION FIVE – COMPARISON OF METHODS

Table 14: Model Error/Efficiency Measures

	MAR			Single Mean			FIML			MI Zero Auxiliaries 10 iterations			MI 10 iterations			MI 100 iterations		
	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$
Absolute Bias	-37.06	-37.15	-36.60	13.09	-35.61	13.29	0.01	0.07	0.49	-22.46	4.81	-11.27	9.17	1.30	17.66	4.83	0.07	10.84
% SE Bias	-252.11	-1100.19	-174.01	62.08	-524.98	43.79	-0.54	-0.72	1.33	-130.78	149.96	-46.28	50.80	44.82	74.95	30.85	-2.15	44.28
Coverage	0.29	0.00	0.58	0.92	0.00	0.93	0.95	0.95	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1	1	0.80
Length	17.34	5.31	41.38	24.84	10.66	59.47	19.87	4.94	47.35	20.05	5.06	47.21	21.22	4.53	46.20	19.82	5.25	49.81
RMSE	1183.40			2392.80			1242.75			985.91			1046.00			1099.01		
Time of one model (seconds)	0.011			0.011			0.046			0.55			1.368			61.177		
Number of Observations	518			1000			1000			1000			1000			1000		

Note: Mote Carlo Simulation using a MAR mechanism. 51 per cent missingness introduced.

SESSION FIVE – ABSOLUTE BIAS

- Absolute Percentage Bias (APB) measures the absolute bias as a percentage of the true value
- The magnitude of this bias informs us how far the estimate from the model is from the true value
- $\pm 20\%$ as a threshold of unacceptable absolute bias
- Both the MAR model and the single mean imputation model perform relatively badly in this statistic

SESSION FIVE – COMPARISON OF METHODS

Table 14: Model Error/Efficiency Measures

	MAR			Single Mean			FIML			MI Zero Auxiliaries 10 iterations			MI 10 iterations			MI 100 iterations		
	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$
Absolute Bias	-37.06	-37.15	-36.60	13.09	-35.61	13.29	0.01	0.07	0.49	-22.46	4.81	-11.27	9.17	1.30	17.66	4.83	0.07	10.84
% SE Bias	-252.11	-1100.19	-174.01	62.08	-524.98	43.79	-0.54	-0.72	1.33	-130.78	149.96	-46.28	50.80	44.82	74.95	30.85	-2.15	44.28
Coverage	0.29	0.00	0.58	0.92	0.00	0.93	0.95	0.95	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1	1	0.80
Length	17.34	5.31	41.38	24.84	10.66	59.47	19.87	4.94	47.35	20.05	5.06	47.21	21.22	4.53	46.20	19.82	5.25	49.81
RMSE	1183.40			2392.80			1242.75			985.91			1046.00			1099.01		
Time of one model (seconds)	0.011			0.011			0.046			0.55			1.368			61.177		
Number of Observations	518			1000			1000			1000			1000			1000		

Note: Mote Carlo Simulation using a MAR mechanism. 51 per cent missingness introduced.

SESSION FIVE – BIAS % OF STANDARD ERROR

- measures the bias relative to the standard error of the estimator
- measures the magnitude of bias compared to its precision
- Following the work of Collins, Schafer and Kam (2001) we set our threshold at $\pm 50\%$
- MAR model, $\beta_2 X_2$ has a bias % standard error score of -1100%. This means that the bias is around 11 times lower than the standard error of the estimate
- The single mean imputation model also reaches troubling levels around $\beta_2 X_2$, with a -525% reduction, meaning that the bias is around 5.25 times lower than the standard error of the estimate

SESSION FIVE – COMPARISON OF METHODS

Table 14: Model Error/Efficiency Measures

	MAR			Single Mean			FIML			MI Zero Auxiliaries 10 iterations			MI 10 iterations			MI 100 iterations		
	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$
Absolute Bias	-37.06	-37.15	-36.60	13.09	-35.61	13.29	0.01	0.07	0.49	-22.46	4.81	-11.27	9.17	1.30	17.66	4.83	0.07	10.84
% SE Bias	-252.11	-1100.19	-174.01	62.08	-524.98	43.79	-0.54	-0.72	1.33	-130.78	149.96	-46.28	50.80	44.82	74.95	30.85	-2.15	44.28
Coverage	0.29	0.00	0.58	0.92	0.00	0.93	0.95	0.95	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1	1	0.80
Length	17.34	5.31	41.38	24.84	10.66	59.47	19.87	4.94	47.35	20.05	5.06	47.21	21.22	4.53	46.20	19.82	5.25	49.81
RMSE	1183.40			2392.80			1242.75			985.91			1046.00			1099.01		
Time of one model (seconds)	0.011			0.011			0.046			0.55			1.368			61.177		
Number of Observations	518			1000			1000			1000			1000			1000		

Note: Mote Carlo Simulation using a MAR mechanism. 51 per cent missingness introduced.

SESSION FIVE – RMSE

- reports the overall accuracy of a model and is the root mean square error between the actual estimate and its predicted value

- The single mean imputation model documents a very high RMSE which demonstrates its inefficiency once more

SESSION FIVE – COMPARISON OF METHODS

Table 14: Model Error/Efficiency Measures

	MAR			Single Mean			FIML			MI Zero Auxiliaries 10 iterations			MI 10 iterations			MI 100 iterations		
	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$
Absolute Bias	-37.06	-37.15	-36.60	13.09	-35.61	13.29	0.01	0.07	0.49	-22.46	4.81	-11.27	9.17	1.30	17.66	4.83	0.07	10.84
% SE Bias	-252.11	-1100.19	-174.01	62.08	-524.98	43.79	-0.54	-0.72	1.33	-130.78	149.96	-46.28	50.80	44.82	74.95	30.85	-2.15	44.28
Coverage	0.29	0.00	0.58	0.92	0.00	0.93	0.95	0.95	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1	1	0.80
Length	17.34	5.31	41.38	24.84	10.66	59.47	19.87	4.94	47.35	20.05	5.06	47.21	21.22	4.53	46.20	19.82	5.25	49.81
RMSE	1183.40			2392.80			1242.75			985.91			1046.00			1099.01		
Time of one model (seconds)	0.011			0.011			0.046			0.55			1.368			61.177		
Number of Observations	518			1000			1000			1000			1000			1000		

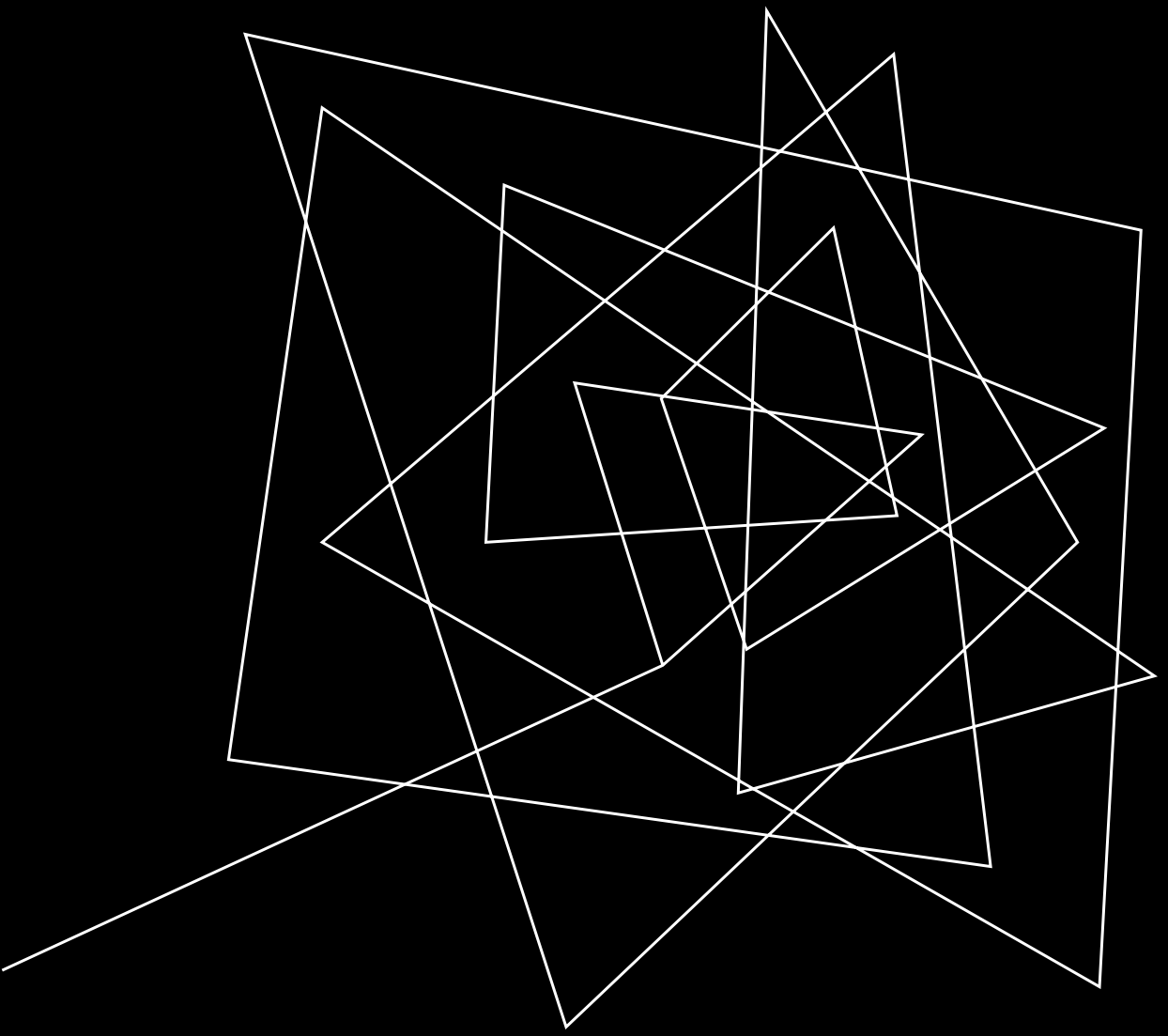
Note: Mote Carlo Simulation using a MAR mechanism. 51 per cent missingness introduced.

SESSION FIVE – TIMINGS

- Timings are provided to assess the cost-benefit of each handling missing data method
- FIML takes less than a second to perform one model, which is relatively like a MI approach with zero auxiliary variables and 10 iterations. By far the longest approach in terms of raw time is the MI 100 iterations model, which takes 62 seconds to perform.

SESSION FIVE – FULL SIMULATION

- Full Simulation is provided for your benefit/edification
- Jupyter Notebook found on website



Q & A