STRATOS
TG1: Regression modelling with missing data: principles, methods, software and examples

Carpenter J, Lee KJ, Goetghebeur E, Little RJA, Tilling K, White IR, Rotnitzky A, and Hogan JW

ISCB, August 2015

Outline

- Aim of the paper
- What the audience want to know
- Proposed structure
- Datasets for illustration
- Presentation of recommendations
- Challenges

The STRATOS Initiative

- ***Objective:*** *to provide accessible and accurate guidance in the design and analysis of observational studies for applied statisticians and data analysts with varying levels of statistical education, and experience*.

- Guidance documents aimed at 3 levels

  Level 1: low statistical knowledge

  Level 2: experienced statistical knowledge

  Level 3: expert statistical knowledge

Aims

- When is a complete case analysis is likely to be ‘good enough’?
- Review, illustration and critique of the established methodology when more complex analyses are required
- Provide worked examples and guidance on methods and software

What the audience want to know…

- “A lot of people arrive at doing MI the way I did, i.e. borrow a do-file from someone who has done MI on a similar dataset, tinker with the variables in the MI command, run it, see that the imputed estimates aren't so different, write-up and publish. This means that incorrect approaches are likely to propagate virally…Therefore, to the hands-on MSc student, **it boils down to "what should I put into my imputation?" "what should I leave out of my imputation?".**

What the audience want to know…

- "Some things I would like to see in any paper would be:
  - **the use of MI in software other than Stata**
  - how to determine how many imputations are necessary
  - pros and cons of using MI"

- "A few things which might be useful:
  - **table of missing data methods by software**
  - chart/checklist for deciding type of missing data (MNAR -> MAR -> MCAR)
  - dealing with non-Gaussian data
  - dealing with multilevel missing data"

Proposed structure

- **When might complete records be 'good enough'?**
  - Including discussion of descriptive statistics
- **Beyond complete records: what is available?**
  - Methods: Direct likelihood, inverse probability weighting, multiple imputation, structural equation modelling, (full) Bayesian analysis
  - Study types: cohort, case-control, case series
  - Software
- **Methods in action**
  - Apply each method to the same example: provide code
- **Critique of methods**
  - Strengths and weakness of each methods
  - Recommendations
- **Summary and Discussion**

Datasets for illustration

- Publically available datasets (reproducibility)

- Contain common problems
  - Cross-sectional data
  - Longitudinal data
  - Missing baseline and/or repeated measures
  - Missing outcome data

Dataset 1: The Youth Cohort Study

- Began in 1984
- UK Government funded representative survey of pupils in England and Wales at school leaving age (school year 11, age 16-17)
- Sequence of cohort studies collecting data on the young people's experience of education, qualifications, employment and training
- To date the study covers 13 cohorts and over 40 surveys.
- Publically available
- Used in Carpenter and Kenward (2013)

Dataset 1: The Youth Cohort Study

- Structure and timing of data collection has varied over the life-cycle of the cohort but Carpenter uses a harmonised dataset of YSC cohorts from 1984-2002 (5 cohorts – 55,145 participants)
- Explored the relationship between year 11 education attainment and key measures of social stratification
  - A number of items of measures are only partially observed e.g. parental occupation missing in 12%, GCSE scores (outcome) 1%, ethnicity 1%.
  - Complete case analysis loses 8,934 observations

Dataset 1: The Youth Cohort Study

- Importantly GCSE score is substantially higher among those with parental occupation observed

→ Data are missing at random

- How would you handle the missing data if you were faced with this analysis?

Dataset 2…

- Ideally want a longitudinal dataset
- Missing covariates
  - Potentially time dependent
- Time to event outcome
  - Completely observed?

Presentation of recommendations

- Start with cross-sectional data
- Focus on estimation of outcome regression model (potentially a survival model) $E(Y \mid \mathbf{X})$
- Assume data are missing at random (MAR)

- How best to present recommendations in a clear and concise manner…

Example 1

Setting:	Analysis method:					
	Direct ML	MI	IPW	Full Bayes	AIPW	EM
No auxiliary variables	A= 0 E = 1 R = 0	A=2 E=1 R=0	A=2 E=0 R=0	A=1 E=1 R=0	A=1 E=1 R=1	A=1 E=1 R=0
Auxiliary variables	A= 0 E = 1 R = 0	A=2 E=1 R=0	A=1 E=0 R=0	A=1 E=1 R=0	A=0 E=1 R=1	A=1 E=1 R=0

Applicability**:** 0 – not without major programming
1 – possible with level 2 programming
2 – relatively straightforward within existing commands

Efficiency: 0 – some loss of information
1 – efficient

Robustness to miss-specification (R): 0 – none
1 – some

* Courtesy of Els Goetghebeur / James Carpenter

Example 2: Potential validity

Method	For missing covariate		For missing outcome	
	RCT	observ'l	single	repeated
LOCF	Not applicable		Valid under LOCF assumption	
Complete cases	Valid under $M_X \perp\!\!\!\perp Y \mid X$		Valid under MAR	Valid under CD-MCAR
Missing = failure	Valid	Valid if missing = failure	Valid if missing = failure	
Mean imputation		Fails to control confounding	Bias, SE ↓↓↓	
Missing indicator			Not applicable	
Regression imputation	Valid under MAR (imp. model = other X's only)		SE ↓↓	
Multiple imputation	Valid under MAR		Valid under MAR	

* Courtesy of Ian White

Example 2: Efficiency

Method	For missing covariate		For missing outcome	
	RCT	observ'l	single	repeated
LOCF			Over-efficient?	
Complete cases	Inefficient		Efficient	Inefficient
Missing = failure	Efficient only if M=F	Efficient	Efficient	
Mean imputation	Efficient ?			
Missing indicator	Efficient *			
Regression imputation	Efficient			
Multiple imputation	Efficient		Efficient	

? if missingness not predictive
* if weighted

* Courtesy of Ian White

Example 3

Method	Assumption	Advantages	Disadvantages	When It May Be Useful
Complete case	completeness independent of outcome given covariates	easy to do	may be inefficient	high % complete cases, most incomplete cases have missing outcome, and little auxiliary Info. for outcomes
IPW	completeness independent of outcome and covariates given missingness predictors	fairly easy to do	may be inefficient, especially if weights very variable; limited use with non-monotone missingness	monotone missingness, e.g. wave dropout, and little auxiliary Info. for outcomes
MI	MAR	can be easier than full likelihood, especially if auxiliary. info.	potential for being used incorrectly	many incomplete cases have observed outcome, or auxiliary Info. available
…				

* Courtesy of Sean Seaman

Example 4

Factors to consider when choosing a method of analysis:

1. Fraction of missing values for each variable
2. Fraction of incomplete cases
3. Fraction of incomplete cases among those with observed outcome and exposure (FICO)
4. Availability of auxiliary variables
5. Distribution of number of missing values
6. Patterns of jointly missing data
7. Reasons for missing data
8. Plausible missingness mechanisms
9. Clustering of data

low FICO & no AVs → CCA?

simple pattern → IPW?

possible departures from MAR

* Courtesy of Ian White

Example 4

1. Very little missing data
2. Missing data only in the outcome
3. Other patterns with low FICO
4. Multilevel data
5. Interactions in the model
6. Mis-specified model
7. Simple missing data patterns
8. Too much missing data

1–3: low FICO & no AVs → CCA?

4 → REALCOM etc.?

5 → care

6–8: IPW?

→ sensitivity analysis

* Courtesy of Ian White

Challenges

- Scope
  - Cross-sectional and longitudinal studies in a single paper?
  - Restrict focus to regression modelling?
  - Which missing data methods to include?

- Presentation of results?

References

- Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd edition, New York: John Wiley.
- White, I. R., & Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Statistics in Medicine, 29(28), 2920-2931.
- Carpenter, J., & Kenward, M. G. (2013). Multiple Imputation and its Application. Chichester: Wiley.
- Hogan JW, Roy J, Korkontzelou C. (2004) Tutorial in Biostatistics: Handling dropout in longitudinal studies. Stat Med. 23: 1455-1497.
- Horton NJ and Kleinman KP. (2007) Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. American Statistician. 61 (1) 1-12.