Missing data - a review and empirical application

Robin Mitra

April 25, 2022

Introduction

- Missing data are a common unavoidable problem in many fields
- The presence of missing values can complicate typical analyses that analysts wish to perform on the data
- Complete case analysis is the simplest method to handle missing data (throw away any unit/row in the data that has missing values)
- However, ad hoc methods to deal with missing data can severely compromise the validity of any analyses performed

Missing data mechanism

- Not all missing data are the same
- We need to consider the process or mechanism that causes the missing data, this can be broken down into three categories
  - Missing completely at random (MCAR): the missing values arise independently of any variables of interest
  - Missing at random (MAR): the missing data mechanism depends on variables of interest, but these are all observed
  - Not missing at random (NMAR): the missing data mechanism depends on variables of interest, some of which may be unobserved or missing

Missing mechanism - illustration

- Suppose we are conducting a survey of people's income and some income values are missing
  - If we think that people with higher incomes are less likely to report their income then we have NMAR
  - If we think that people's job category (observed) determines whether missing data arise or not then we have MAR
  - If we think missing incomes values arise purely as a random sample within the survey then we have MCAR
- There is not much we can do if missingness is NMAR so normally assume MCAR or MAR
- If missing data are MCAR then complete case analysis results in unbiased estimates
- However even if MCAR is a reasonable assumption complete case analysis can be inefficient

Complete case illustration

- What is the problem with using complete case analysis here?

BMI – x1	Race – x2	Sex – x3	Age – x4	Renal disease - x5
23.6	Asian	M	38	2
?	White	F	32	4
21.5	?	F	40	1
29.1	Black	?	63	6
21.4	?	M	58	4
30.0	Other	?	45	3
22.3	Asian	M	?	2
19.5	White	M	47	5
28.7	White	F	48	?

Figure: Complete case Illustration

Complete case illustration

- Ideally want to address the missing data problem appropriately without significantly increasing the burden on analysts
- This is what motivates multiple imputation

BMI – x1	Race – x2	Sex – x3	Age – x4	Renal disease - x5
23.6	Asian	M	38	2
?	White	F	32	4
21.5	?	F	40	1
29.1	Black	?	63	6
21.4	?	M	58	4
30.0	Other	?	45	3
22.3	Asian	M	?	2
19.5	White	M	47	5
28.7	White	F	48	?

Multiple imputation

- In multiple imputation the treatment of the missing data and analysis of the data is done in two distinct stages
- First an imputer deals with the missing data
- To do this the imputer forms a model for the complete data (X)
- Using this model the imputer imputes the missing data (X_{mis}) conditional on the observed data (X_{obs}) i.e. from $p(X_{mis}|X_{obs})$.
- The imputer does this m times to generate m multiply imputed data sets $X_{com}^{(k)}, \quad k = 1, \ldots, m$
- The imputed data sets are then released to analysts who can treat each imputed data set as if it were the original fully observed data and apply their standard statistical analyses to the data
- In this way the burden of dealing with the missing data largely falls on the imputer and not the analysts

Multiple imputation - illustration

Observed data

X_1	X_2
✓	✓
✓	✓
✓	✓
✓	✓
✓	✓
✓	?
✓	?

Figure: Multiple Imputation Illustration

Multiple imputation - illustration

Figure: Multiple Imputation Illustration

Multiple imputation - combining rules

- We assume an analyst wishes to infer about some quantity: Q
- In each imputed data set $X_{com}^{(k)}$ $(k = 1, \ldots, m)$ the analyst obtains a point and a variance estimate, q_k and u_k in the usual way
- These estimates are combined across the imputed data sets:

$$\bar{q}_m = \frac{\sum_{k=1}^{m} q_k}{m} \qquad b_m = \frac{\sum_{k=1}^{m} (q_k - \bar{q}_m)^2}{(m-1)}$$

$$\bar{u}_m = \frac{\sum_{k=1}^{m} u_k}{m}$$

Inference with multiply imputed data

- Estimate of Q: $\bar{q}_m$
- Estimate of variance is:

$$T_m = \bar{u}_m + \left(1 + \frac{1}{m}\right) b_m$$

- Analysts can then base inferences on a $t-$distribution
- For example analysts can construct a 95% confidence interval for Q:

$$\bar{q}_m \pm t_{\nu_m}(0.025)\sqrt{T_m} \quad \text{where}$$

$$\nu_m = (m-1)\left(1 + \frac{1}{m+1}\frac{\bar{u}_m}{b_m}\right)^2$$

Survival study

- We consider a study examining survival times of patients after undergoing a kidney transplant
- Patients in the study underwent a transplant between 2001 and 2008, after this they were followed up to see how long they survived
- Additional background/covariate information was collected, including demographic, health and social-economic variables, for both the recipient and donor, e.g. gender, BMI, ACORN index etc.
- Analysis of interest would be to determine the effect of these variables on survival

Missing data by design problem

- There is a special type of missing data by design problem that arises here
- Up until 2003 recipient BMI was not recorded, and in subsequent years was missing for a large proportion of patients
- In the data there are approximately 64% of patients missing in this variable
- Given how the missing data arise, complete case analysis should result in unbiased estimates
- Further there is often scepticism about imputing such large amounts of data to be used in analysis
- However, a complete case analysis here results in only 27.6% of the data

Analysis

- We compare the performance of multiple imputation versus complete cases here
- The analysis model fitted to the data is a Cox proportional hazards regression model, this allows us to see the relationships between the different variables and survival
- Not all the variables will be significantly associated with survival
- We first find a subset of variables that are significant through stepwise regression using the method of Wood et al. (2008)
- We then compare point and interval estimates obtained from multiple imputation and complete cases
- The MICE package in R was used to impute missing values here (imputes using full conditional regression models fit to the data)

Partial results

covariate	complete cases		multiple imputation	
	est	95% CI	est	95% CI
Recipient sex	0.203	(0.002, 0.403)	0.155	**(0.064, 0.246)**
Serum creatinine	0.006	(0.005, 0.007)	0.006	**(0.006, 0.006)**
Donor age	0.005	(-0.002, 0.013))	0.004	**(0.000, 0.007)**
Recipient age	0.026	(0.018, 0.034)	0.027	**(0.024, 0.031)**
Recipient BMI	-0.017	(-0.037, 0.003)	-0.026	**(-0.042, -0.009)**
Donor BMI	0.014	(-0.004, 0.031)	0.009	**(0.000, 0.018)**
Donor CMV Status positive	0.314	(0.118, 0.509)	0.134	**(0.046, 0.223)**

Simulation

- As we do not know the true values of the missing data we also run some simulations based on the complete case subsample
- Specifically we subset on the complete cases and randomly re-introduce missing patterns back into the data (consistent with how they appeared in the real data)
- We then obtain estimates from using multiple imputation and complete cases applied to this subsample and compare these to the estimates obtained from the (fully observed) complete case data
- We then repeat this process 100 times

Simulation - results

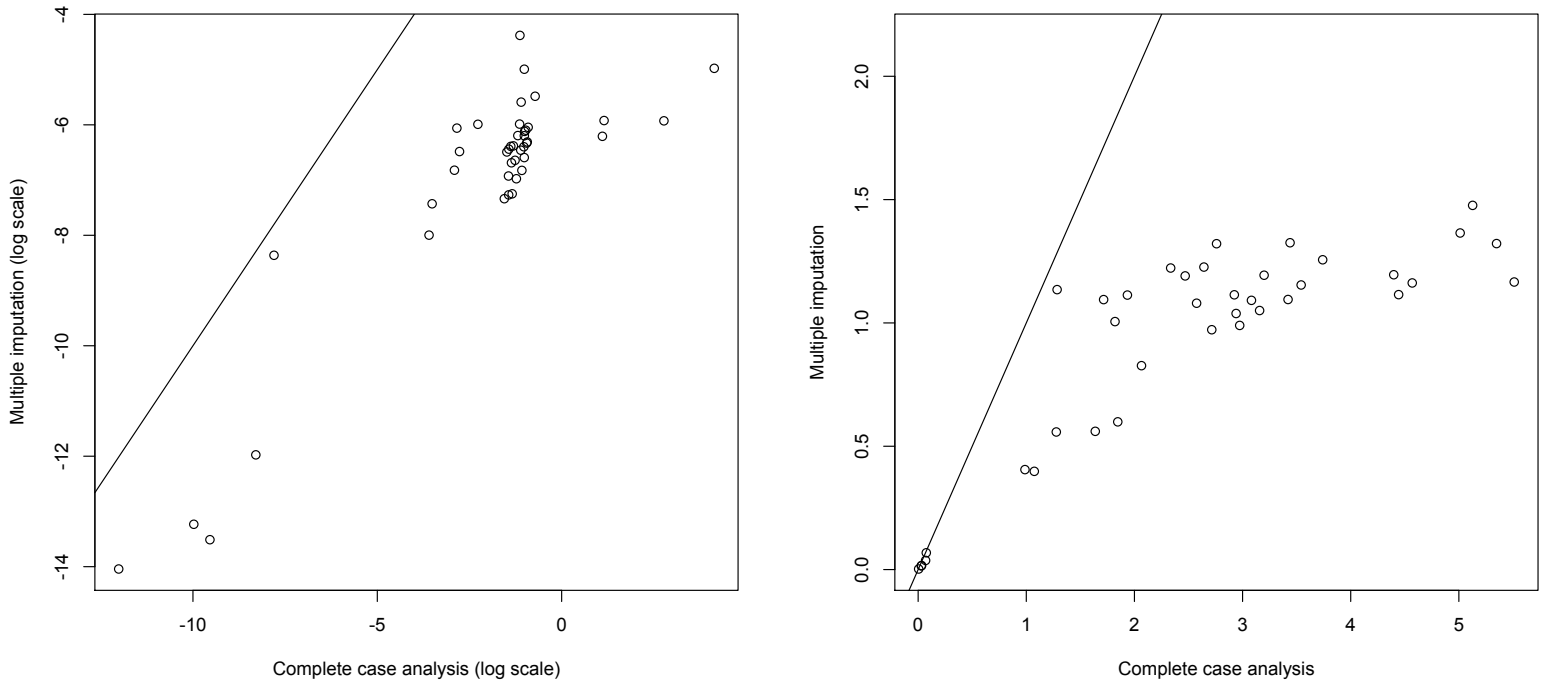

Figure: Plots of MSE (left) and average confidence interval length (right) for coefficient estimates obtained from using multiple imputation and complete case analysis. Points below the y=x line indicate a larger MSE or interval length for a complete case analysis.

Stepwise simulation

- We also apply stepwise regression procedures to each of the 100 simulated data sets using both MI (based on Wood et al. (2008)) and complete cases
- We compare the final selected variable sets from each approach to the final selected variable set obtained from the fully observed data

Stepwise simulation - results

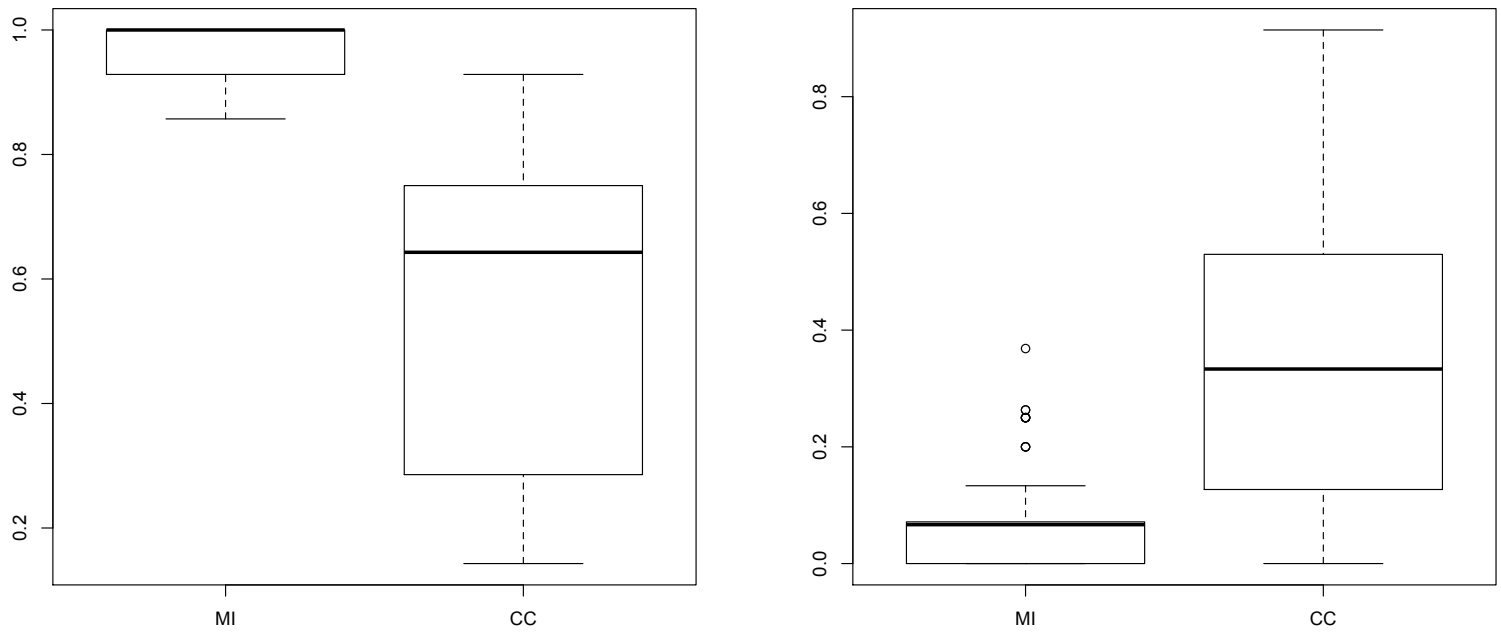

Figure: Proportion of correct covariates included (left) and proportion of covariates selected that should not be included (right) across the 100 replications

Conclusion

- We see that there are advantages in using multiple imputation when missingness is MCAR, even when there is a large proportion of missing data
- It would be interesting to explore what other areas this missing by design problem could be an issue, e.g. meta analysis using several clinical trial data sets
- This leads to a more general problem where we assume an underlying "structure" to the missing data
- Structured missingness is likely in large complex datasets, particularly when it is constructed from combining different different sources and presents some unique challenges

References

Wood, A. M., White, I. R., and Royston, P. (2008). How should variable selection be performed with multiply imputed data? Statistics in Medicine 27, 17, 3227-3246.

Mitra, R., Pankhurst, L., Kimber, A., and Collett, D. (2019). Multiply imputing missing values arising by design Multiply imputing missing values arising by design in transplant survival data. Biometrical Journal

Acknowledgements

- Laura Pankhurst, David Collett (NHS BT) and Alan Kimber (University of Southampton).
- The authors undertook this work under NIHR grant RMOFS2012/03.