Three popular missing data methods compared: complete records; inverse probability weighting; multiple imputation[1]

Rod Little, James R. Carpenter, Kate Lee

(on behalf of Topic Group 1)

James.Carpenter@lshtm.ac.uk · J.Carpenter@ucl.ac.uk

STRATOS INITIATIVE

Newcastle, Thu 25th August 2022

Acknowledgements

Katherine J Lee (Melbourne)

Kate Tilling (Bristol)

Rosie Cornish (Bristol)

Roderick J A Little (Michigan)

Melanie L Bell (Arizona)

Els Goetghebeur (Ghent)

Joseph W Hogan (Brown)

The STRATOS initiative

Overview

- Motivation
- Data: Youth Cohort Study
- Brief review of
  1. Complete Records (CR);
  2. Inverse Probability Weighting (IPW);
  3. Multiple Imputation (MI)
- Choosing the tools: CR or IPW or MI?
- Analysis
- Discussion

Motivation: level 1 (low/interemediate statistical knowledge)

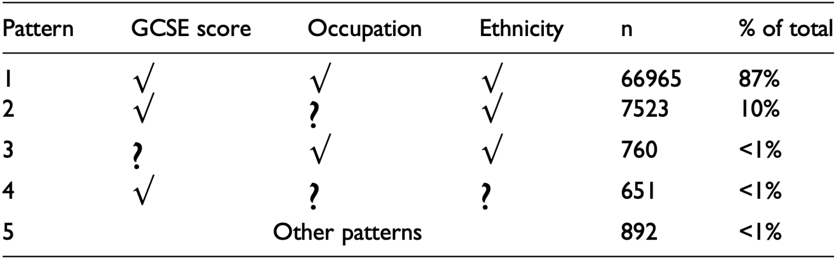
Despite the fact that Complete Records/Complete Cases, Inverse Probability Weighting (IPW) and Multiple Imputation (MI) are common in practice, our experience is that the principles underlying the choice between these methods are not as well understood as they might be. For example:

1. Inference for sample mean: in many settings where the auxiliary variables are strongly related to the propensity to respond and weakly related to values of the variable of interest, IPW actually leads to worse inferences than CR;
2. In the statistics literature, CR is widely criticized and seen as inferior to methods like MI that use all available data. However, for some regression problems CC is optimal, and MI is actually less, not more, efficient;
3. CR is often thought of as biased unless data are missing completely at random, and IPW/MI thought to reduce the bias; but this is not always the case.

Example: Youth Cohort Study

Variable name	Description
Educational Achievement Score	GCSE points score. Each pupil sits up to 15 GCSE exams. The results for each are converted into a score from 7 (highest grade) to 0 (fail). These are summed across a pupil's exams and capped at 84 (equivalent to 12 GCSEs at the top grade).
Cohort	year of data collection: 1990, 93, 95, 97, 99
Boy	indicator variable for boys
Occupation	parental occupation, categorized as managerial, intermediate or working
Ethnicity	Categorized as Bangladeshi, Black, Indian, other Asian, Other, Pakistani or White

Missing data patterns



Pattern	GCSE score	Occupation	Ethnicity	n	% of total
1	√	√	√	66965	87%
2	√	?	√	7523	10%
3	?	√	√	760	<1%
4	√	?	?	651	<1%
5		Other patterns		892	<1%

- All variables are statistically significantly associated with the chance of parental occupation being missing;
- But ethnicity and GCSE score are also strong predictors of parental occupation.

Brief review: Complete Records / Complete Case analysis

CR for a set of variables simply discards units where any of these variables are missing.

- Key advantage is simplicity: it is the default analysis in most statistical software packages.
- Drawback 1: the complete cases are not a random subsample of the original sample unless the data are MCAR (typically unrealistic). If data not MCAR, CR is biased for summary measures.
- Drawback 2: CR discards information in the incomplete cases, which has typically cost non-trivial resources to collect, and which will often contain information for reducing bias and/or increasing the efficiency of CC estimates.

Inverse Probability Weighting

- IPW weights complete units by the inverse of an estimate of the probability of response (i.e. observing the data) [2]

Inverse Probability Weighting

- IPW weights complete units by the inverse of an estimate of the probability of response (i.e. observing the data) [2]
- A simple approach for creating weights is to form adjustment cells based on background variables measured for both respondents and non-respondents. All nonrespondents are assigned zero weight and the nonresponse weight for all respondents in an adjustment cell is then the inverse of the estimated response rate in that cell. For more details see [3], Example 3.6.

Inverse Probability Weighting

- IPW weights complete units by the inverse of an estimate of the probability of response (i.e. observing the data) [2]
- A simple approach for creating weights is to form adjustment cells based on background variables measured for both respondents and non-respondents. All nonrespondents are assigned zero weight and the nonresponse weight for all respondents in an adjustment cell is then the inverse of the estimated response rate in that cell. For more details see [3], Example 3.6.
- With more extensive information, a generalisation is response propensity stratification, where (i) the indicator for a CR is regressed on the background variables, using the combined data for respondents and nonrespondents, typically using logistic regression (ii) a predicted response probability is computed, and (iii) the weighted substantive model is fitted.

Inverse Probability Weighting

- IPW weights complete units by the inverse of an estimate of the probability of response (i.e. observing the data) [2]
- A simple approach for creating weights is to form adjustment cells based on background variables measured for both respondents and non-respondents. All nonrespondents are assigned zero weight and the nonresponse weight for all respondents in an adjustment cell is then the inverse of the estimated response rate in that cell. For more details see [3], Example 3.6.
- With more extensive information, a generalisation is response propensity stratification, where (i) the indicator for a CR is regressed on the background variables, using the combined data for respondents and nonrespondents, typically using logistic regression (ii) a predicted response probability is computed, and (iii) the weighted substantive model is fitted.
- So, when estimating a population mean, the sample mean is replaced by the weighted mean.

Notes on IPW

Drawbacks of IPW:

- only complete records are re-weighted: hence IPW is inefficient when (as is often the case) there is substantial information in the partially observed records;
- weight variables must be fully observed.
- extreme weights can inflate the variance of weighted estimates.
- Variance estimation for weighted estimates will ideally take into account uncertainty in the estimated weights, otherwise standard errors will be over- estimated so inferences will be conservative, see [3], Ch. 5.

Nevertheless, theory [4] suggests that this is an effective method for removing bias when data are MAR. Among many, [5] give suggestions for modifying the method to improve efficiency.

Multiple Imputation

Imputing the missing values has the advantage that, unlike CC or IPW, observed values in the incomplete cases are retained to make full use of them in the analysis.

Multiple Imputation

Imputing the missing values has the advantage that, unlike CC or IPW, observed values in the incomplete cases are retained to make full use of them in the analysis.

Because we can never recover the actual missing value, a single imputation for each missing value cannot reflect the imputation uncertainty, and as a result standard errors of estimates based on analysis of a single filled-in data tend to be underestimated.

Multiple Imputation

Imputing the missing values has the advantage that, unlike CC or IPW, observed values in the incomplete cases are retained to make full use of them in the analysis.

Because we can never recover the actual missing value, a single imputation for each missing value cannot reflect the imputation uncertainty, and as a result standard errors of estimates based on analysis of a single filled-in data tend to be underestimated.

In fact, the primary goal of MI may be viewed as preserving the information in the observed values for inference, not to get the absolute best predictions of the missing values.

Basic MI steps

(a) estimate a predictive distribution for the missing values given the observed values in the data set — for example using joint modelling or full conditional specification [6];

(b) fill in, or impute, the missing values with draws from this predictive distribution (note that the imputations are draws, that is random selections from the predictive distribution, not means of the predictive distribution);

(c) repeat step (b) $M > 1$ times (where, say, $M = 10$ or 20) to create M datasets, each containing different sets of draws of the missing values.

(d) Fit our intended model to each of the M datasets, obtaining point estimates and standard errors, and

(e) Combine the M results for inference using Rubin's rules [7, 6].

Advantages and drawbacks of MI

Key advantages:

- includes information in partially observed records;
- includes information in variables which are not part of the substantive model.

Key challenges:

- Relatively computationally complex — but good software widely available
- Imputation must be consistent with the substantive model, and this may require some care.

Sample mean

We wish to estimate the sample mean of a variable Y which has a non-trivial proportion of missing values.

Suppose we have variables $\mathbf{X}$ which

(a) are associated with the probability of observing Y (the propensity to respond), and

(b) are associated with the values of Y.

Suppose also

(c) we have variables $\mathbf{Z}$ which, given $\mathbf{X}$ are *not* associated with the propensity to respond, but *may* be additionally associated with the values of Y.

For each of (a), (b), (c), we classify the associations as 'low' or 'high'.

Sample mean: choosing between IPW/MI

Association of *X* with Response *R* (ie. strength of propensity to respond)	Association of outcome *Y* with (i) propensity to respond and (ii) Z (as defined in the text)			
	Propensity: Low Z: Low	Propensity: Low Z: High	Propensity: High Z: Low	Propensity: High Z: High
Low	Cell LLL IPW MI Bias: --- --- Var: --- ---	Cell LLH IPW MI Bias --- --- Var --- ↓	Cell LHL IPW MI Bias: --- --- Var: ↓ ↓	Cell LHH IPW MI Bias: --- --- Var: ↓ ↓↓
High	Cell HLL IPW MI Bias: --- --- Var: ↑ ---	Cell HLH IPW MI Bias: --- --- Var: ↑ ↓	Cell HHL IPW MI Bias: ↓ ↓ Var: ↓ ↓	Cell HHH IPW MI Bias: ↓ ↓ Var: ↓ ↓↓

Example: distribution of parental occupation in YCS

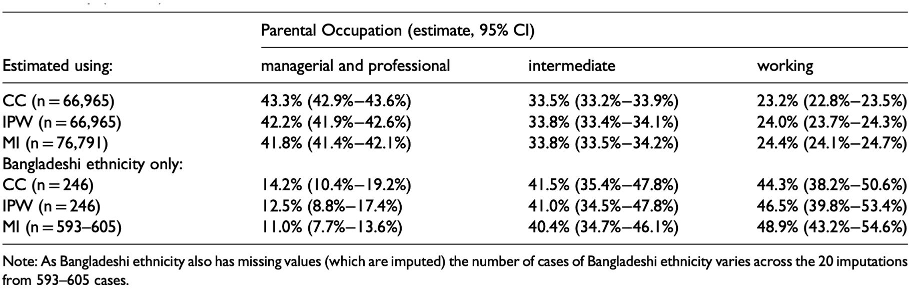
Recall that $\sim$11% of parental occupation data are missing.

All the other variables are statistically significant predictors of observing parental occupation (the propensity to respond) — not least because of the large sample size.

Further, the GCSE score and ethnicity are strong predictors of parental occupation.

Therefore, we are likely somewhere between the ‘HHL’ and the ‘HHH’ cells of the previous table.

Results: distribution of parental occupation



	Parental Occupation (estimate, 95% CI)		
Estimated using:	managerial and professional	intermediate	working
CC (n = 66,965)	43.3% (42.9%–43.6%)	33.5% (33.2%–33.9%)	23.2% (22.8%–23.5%)
IPW (n = 66,965)	42.2% (41.9%–42.6%)	33.8% (33.4%–34.1%)	24.0% (23.7%–24.3%)
MI (n = 76,791)	41.8% (41.4%–42.1%)	33.8% (33.5%–34.2%)	24.4% (24.1%–24.7%)
Bangladeshi ethnicity only:			
CC (n = 246)	14.2% (10.4%–19.2%)	41.5% (35.4%–47.8%)	44.3% (38.2%–50.6%)
IPW (n = 246)	12.5% (8.8%–17.4%)	41.0% (34.5%–47.8%)	46.5% (39.8%–53.4%)
MI (n = 593–605)	11.0% (7.7%–13.6%)	40.4% (34.7%–46.1%)	48.9% (43.2%–54.6%)

Note: As Bangladeshi ethnicity also has missing values (which are imputed) the number of cases of Bangladeshi ethnicity varies across the 20 imputations from 593–605 cases.

When CR will be efficient & unbiased

CR will be fully efficient (and more efficient than MI) when[1]:

(a) The substantive model is a regression of Y on $\mathbf{X}$, only Y has missing values and these are MAR given $\mathbf{X}$.

(b) The substantive model is a regression of *longitudinal* $\mathbf{Y}$ on $\mathbf{X}$, with (i) the focus on the regression of the final Y value on $\mathbf{X}$, and (ii) when $\mathbf{Y}$ has missing values (typically due to attrition) and these are MAR given observed Y values and $\mathbf{X}$.

In case (b), it is important to take care over the choice of the correlation model; a relatively unstructured approach will reduce bias typically at minimum cost to power [8], Ch. 3.

[1] Assuming the substantive model is correctly specified

When CR will be unbiased

Consider the regression of Y on $\mathbf{X}$, and suppose that there are missing data in all variables.

(a) CR is unbiased (but may be inefficient) when the chance of a complete record depends on any combination of $\mathbf{X}$, but given these not on Y.
— in particular CR can be *unbiased* when data are MNAR, when IPW and MI would be biased.

[2] See [9] for discussion of logistic regression.

When CR will be unbiased

Consider the regression of Y on $\mathbf{X}$, and suppose that there are missing data in all variables.

(a) CR is unbiased (but may be inefficient) when the chance of a complete record depends on any combination of $\mathbf{X}$, but given these not on Y.
— in particular CR can be *unbiased* when data are MNAR, when IPW and MI would be biased.

(b) CR is *biased* when missingness depends on Y (given $\mathbf{X}$).

[2] See [9] for discussion of logistic regression.

When CR will be unbiased

Consider the regression of Y on $\mathbf{X}$, and suppose that there are missing data in all variables.

(a) CR is unbiased (but may be inefficient) when the chance of a complete record depends on any combination of $\mathbf{X}$, but given these not on Y.
— in particular CR can be *unbiased* when data are MNAR, when IPW and MI would be biased.

(b) CR is *biased* when missingness depends on Y (given $\mathbf{X}$).

Because MI is more efficient than CR and IPW, it may be preferred from a mean square error perspective even if a moderate MNAR mechanism is suspected.

This is especially the case if there are a large number of covariates, and missing data are scattered over them in a haphazard way, so that the fraction of complete cases is relatively small.[2]

[2] See [9] for discussion of logistic regression.

YCS analysis: what to expect

The substantive model regresses GCSE score on all the covariates (ethnicity, parental occupation, sex, cohort).

GCSE score is strongly predictive of missing parental occupation
$\Longrightarrow$ CC is likely to be biased if data are MAR.

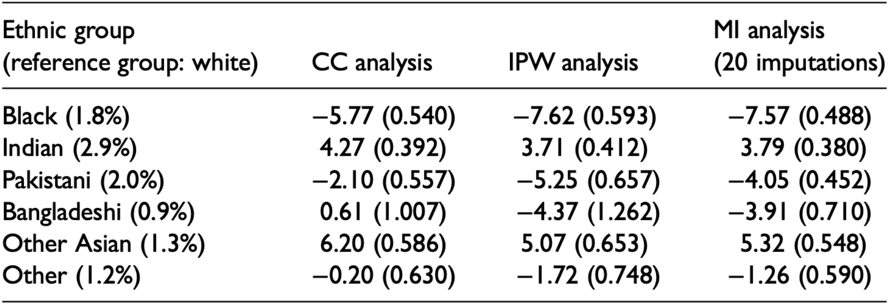
Parental occupation is more likely to be missing for some of the ethnic groups (e.g. Bangladeshi)
$\Longrightarrow$ MI is likely to be more efficient for ethnicity coefficients.

Ethnic group has relatively few missing values, but is a strong predictor of

- missing parental occupation,
- parental occupation values, and
- GCSE score

$\Longrightarrow$ CC estimates for this variable are likely biased.

YCS analysis: results



Ethnic group			MI analysis
(reference group: white)	CC analysis	IPW analysis	(20 imputations)
Black (1.8%)	−5.77 (0.540)	−7.62 (0.593)	−7.57 (0.488)
Indian (2.9%)	4.27 (0.392)	3.71 (0.412)	3.79 (0.380)
Pakistani (2.0%)	−2.10 (0.557)	−5.25 (0.657)	−4.05 (0.452)
Bangladeshi (0.9%)	0.61 (1.007)	−4.37 (1.262)	−3.91 (0.710)
Other Asian (1.3%)	6.20 (0.586)	5.07 (0.653)	5.32 (0.548)
Other (1.2%)	−0.20 (0.630)	−1.72 (0.748)	−1.26 (0.590)

Choosing the tool[1]

Meth.	When to use	When to avoid
CC	• when the probability that a case is complete depends on the covariates but, given these, not the outcome — unbiased (though not fully efficient).	• when estimating the mean of an incomplete outcome if data are not MCAR • when there are good auxiliary variables (MI or IPW more efficient)
IPW	• when there are useful auxiliary variables (more efficient than CC); • Under MAR	• when MI is also valid, because IPW is generally less efficient as (i) only reweights complete cases and (ii) cannot use incomplete auxiliary variables.
MI	• when useful auxiliary variables & MAR holds (more efficient than a CC analysis); • when MAR holds	• when not confident the imputation model is (i) consistent with the scientific model and (ii) correctly specified (risk of bias)

Note on auxiliary variables

- Useful auxiliary variables need to be good predictors of the missing values.
- If they are additionally good predictors of the propensity of data to be complete, they correct for bias (if data are MAR).
- If they *only* predict the propensity of data to complete, they add noise, and may induce bias.

See Spratt *et al* [10].

Discussion

- This talk, based on the recent paper by TG 1[1], has given some practical guidance for choosing between the three most common approaches for handling missing data.
- In order to choose between methods, and plan the analysis, exploring the pattern of missing data, and the predictors of a complete record (both from the data and in discusion with collaborators) is crucial — alongside a clear idea of the scientific model.
  - causal graphs can be useful for this [11, 12].
- Sensitivity analysis will often be needed. This sounds scary, but is actually quite simple with MI [13], [14].
- Reporting remains a challenge — often it is not clear how to reproduce published analyses which use missing data methods — TG1 is reflecting on how to move this forward.

References I

[1] R J A Little, J R Carpenter, and K Lee on behalf of the STRATOS initiative.
A comparison of three popular methods for handling missing data: complete case analysis, wieghting and multiple imputation *sociological methods and research (e-pub ahead of print) doi: 10.1177/00491241221113873*, 2022.

[2] S Seaman, I R White, A J Copas, and L Li.
Combining multiple imputation and inverse-probability weighting.
Biometrics, 68:129—137, 2012.

[3] R J A Little and D B Rubin.
Statistical analysis with missing data (third edition).
Chichester: Wiley, 2019.

[4] P R Rosenbaum and D B Rubin.
The central role of the propensity score in observational studies for causal effects.
Biometrika, 70:41–55, 1983.

[5] A A Tsiatis, M Davidian, and W Cao.
Improved doubly robust estimation when data are monotonely coarsened, with application to longitudinal studies with dropout.
Biometrics, 67:536–545, 2011.

[6] J R Carpenter and M G Kenward.
Multiple Imputation and its Application.
Chichester, Wiley, 2013.

[7] D B Rubin.
Multiple imputation for nonresponse in surveys.
New York: Wiley, 1987.

[8] James R Carpenter and Michael G Kenward.
Missing data in clinical trials — a practical guide.
Birmingham: National Health Service Co-ordinating Centre for Research Methodology, 2008.

References II

[9] J W Bartlett, O Harel, and J R Carpenter.
Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression.
American Journal of Epidemiology, 182:730–736, 2015.

[10] M Spratt, J R Carpenter, J A C Sterne, B Carlin, J Heron, J Henderson, and K Tilling.
Strategies for Multiple Imputation in Longitudinal Studies.
American Journal of Epidemiology, 172:478–487, 2010.

[11] Katherine J. Lee, Kate M. Tilling, Rosie P. Cornish, Roderick J.A. Little, Melanie L. Bell, Els Goetghebeur, Joseph W. Hogan, and James R. Carpenter.
Framework for the treatment and reporting of missing data in observational studies: The treatment and reporting of missing data in observational studies framework.
Journal of Clinical Epidemiology, 134:79–88, 2021.

[12] Rhian M Daniel, Michael G Kenward, Simon N Cousens, and Bianca L De Stavola.
Using causal diagrams to guide analysis in missing data problems.
Statistical Methods in Medical Research, ?:?, 2011.

[13] James R. Carpenter and Melanie Smuk.
Missing data: A statistical framework for practice.
Biometrical Journal, 63(5):915–947, 2021.

[14] J R Carpenter.
Multiple Imputation-Based Sensitivity Analysis, in W iley *Statistics Reference Online*, ISBN: 9781118445112; doi:10.1002/9781118445112.stat07852, 2019.