Effects of missing data on statistical analysis

Leo Watson, Nathalie Moon ; Department of Statistical Sciences

UNIVERSITY OF TORONTO

Abstract

Analyzing **missing data mechanisms, modern approaches to handling missing data**. **Designing R simulations** to investigate hypotheses about imputation techniques.

Introduction

MOTIVATIONS

- Missing data arises everywhere in the real world but often troublesome & swept under the rug.
- Standard statistical analysis methods usually assume no missing data — gap between reality & common practice
- Look into modern imputation techniques & their pros/cons
  - Specifically, runtime and how to optimize it without sacrificing bias, error, and other performance measures.

DEFINITIONS

Missing Data Mechanisms

MCAR: Probability of missingness for data points in a dataset is constant.

- Each student's mark is stored in a spreadsheet but following a computer update 10% of the data is deleted at random.

MAR: Probability of missingness is dependent on some observed variable of the dataset.

- Most students joined a class from day 1, but some students joined late from the waitlist. 10% of students who joined on time missed submitting the first problem set, while 30% of late students missed the first problem set.

MNAR: Probability of missingness dependent on true value of the data point.

- Due to a system failure, the instructor loses all the students' marks. The instructor requests the students to calculate and share their final marks to the instructor. If they don't, the instructor will input that they got a B.
  - If a student's true mark is an A, they are 90% likely to state their true mark.
  - If a student's true mark is a C, they are 50% likely to state their true mark.

(A few) Imputation Techniques

Listwise Deletion:
(Complete Case Analysis)
Eliminates all observations containing ANY missing values in variables of interest

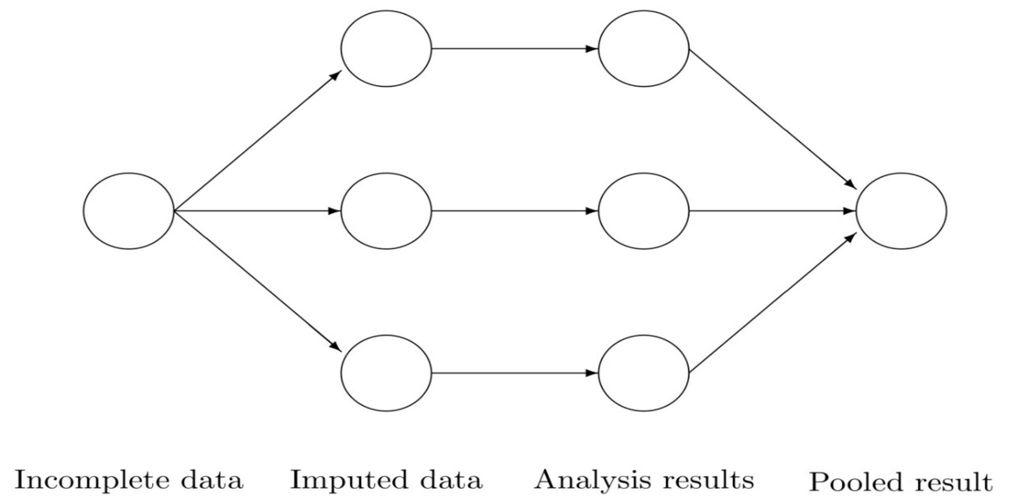
Y	X_1	X_2	X_3
4	0.2	1.2	20
3	NA	1.2	21
2	0.3	1.1	16
2	0.4	1.1	17
1	0.5	2	18
2	0.4	2.1	18
NA	0.2	1.4	19
2	0.1	1.2	22
2	0.1	NA	NA

→

Y	X_1	X_2	X_3
4	0.2	1.2	20
2	0.3	1.1	16
2	0.4	1.1	17
1	0.5	2	18
2	0.4	2.1	18
2	0.1	1.2	22
—	—	—	—
—	—	—	—
—	—	—	—

Multiple Imputation:

1. Takes incomplete dataset and creates multiple copies of it.
2. Impute incomplete columns with plausible values through an iterative predictive method for each copy
3. Obtain estimate for parameter of interest for each copy
4. Pool estimators together to create a single pooled estimate.



Investigation (1):
Multiple Imputation under varying degrees of MCAR, MAR, MNAR

OBJECTIVE

- Compare the effectiveness of multiple imputation under different missingness mechanisms.

METHODS

- Designed R simulations for each of MCAR, MAR, MNAR. Each simulation consists of
  1. Creating data from a complete data model (specified to the right)
  2. Removing some of it (*how* it's removed depends on the mechanism),
  3. Create multiple 'copies' of the data, imputing plausible values in each to make them complete
  4. For each copy in (3.):
     a) Obtain estimate and 95% confidence interval for $\beta_1 = 1$
     b) Measure the performance and statistical validity of the newly minted dataset
  5. Pool estimates from step 4b
  6. Repeat steps 1-5 many times (e.g. n = 1000) and calculate estimate, bias, etc.

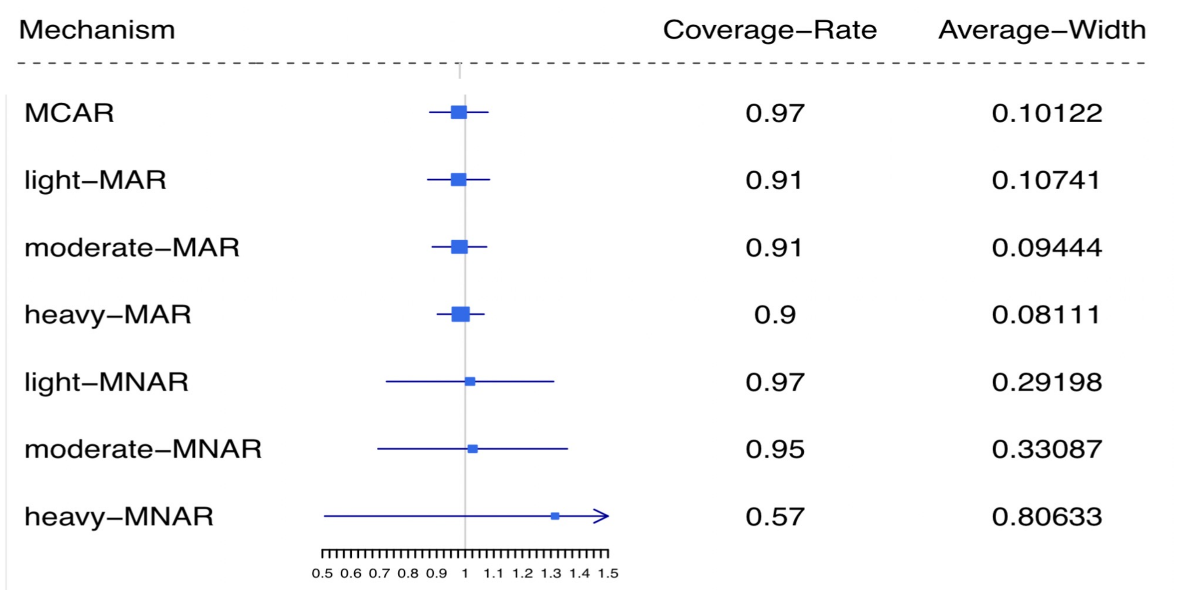
Complete Data Models:

1. MCAR: $y_i = x_i\beta_{1,} + \epsilon_i$
2. MAR: $y_i = x_{1,i}\beta_1 + x_{2,i}\beta_2 + \epsilon_i$
3. MNAR: $y_i = x_i\beta_{1,} + \epsilon_i$

RESULTS

	Estimate	Percent Bias	Coverage Rate	Average Width (of 95% CI)
MCAR	0.9779	2.209	0.97	0.102
MAR-light	0.9768	2.315	0.91	0.108
MAR-moderate	0.9799	2.011	0.91	0.095
MAR-heavy	0.9841	1.588	0.90	0.082
MNAR-light	1.0174	1.740	0.96	0.306
MNAR-moderate	1.0262	2.615	0.95	0.331
MNAR-heavy	1.3146	31.463	0.57	0.806

Narrower intervals that achieve a coverage rate as close as possible to 1 are preferred



DISCUSSION

- Depending on the missingness mechanism, the quality of your imputations will vary significantly.
- **Bias & average confidence interval width tended to increase as the mechanism's severity increased.**
- Although multiple imputation under MAR had the lowest average coverage rate, once we consider how MNAR had much larger confidence intervals, it is clear that **estimation is most impacted by MNAR mechanisms.**

Investigation (2):
When Listwise Deletion Outperforms Multiple Imputation

Case 1: Missing Data only in Response *Y*

HYPOTHESIS **If the missing data occurs only in *Y*, listwise deletion is preferred as it's faster and still provides unbiased estimators.**

METHODS

1. Create data from model $y_i = x_{1,i}\beta_1 + x_{2,i}\beta_2 + x_{3,i}\beta_3 + x_{4,i}\beta_4 + \epsilon_i$ **, where $\beta_1 = 1, \beta_2 = 2, \beta_3 = 3, \beta_4 = 4$.**
2. Remove data from *only* response *Y* (MCAR in this simulation example).
3. Get estimates using multiple imputation, measuring the runtime
4. Get estimates by applying listwise deletion to data, measuring the runtime
5. Repeat 1-4 for lots of iterations (*n = 1000*)
6. Compare the average runtime & pooled performance measures for multiple imputation and listwise deletion.

Y	X_1	X_2	X_3	X_4
30.42424	-0.6264538	1.5117812	0.9189774	1.3586796
NA	0.1836433	0.3898432	0.7821363	-0.1027877
NA	-0.8356286	-0.6212406	0.0745650	0.3876716
11.53927	1.5952808	-2.2146999	-1.9893517	-0.0538050

Listwise Deletion

	Percent_Bias	Coverage_Rate	Avg_Width	RMSE
Intercept	0.038	0.999	0.364	0.052
Wind	2.590	0.999	0.394	0.065
Temp	3.206	0.988	0.359	0.082
Month	0.726	1.000	0.340	0.049
Day	0.887	1.000	0.334	0.057

	average-runtime
Multiple_Imputation	0.0792553
ListwiseDeletion	0.0093304

Multiple Imputation

	Percent_Bias	Coverage_Rate	Avg_Width	RMSE
Intercept	0.033	0.996	0.414	0.066
Wind	2.185	0.986	0.453	0.082
Temp	0.136	0.990	0.408	0.078
Month	2.303	0.912	0.395	0.119
Day	0.723	0.986	0.386	0.075

Case 2: Missing Data independent of response *Y*

HYPOTHESIS **If missingness isn't dependent on *Y*, regression coefficients are free of bias.**

METHODS

- Replicating Case 1 Methods but in step (2), **create missingness in $X_{1,}, X_3, X_4$ dependent *on* X_2 value**

Listwise Deletion

	Percent_Bias	Coverage_Rate	Avg_Width	RMSE
Intercept	0.459	0.992	1.189	0.206
X_1	7.158	0.980	0.810	0.145
X_2	1.324	0.998	1.194	0.183
X_3	0.579	0.995	0.661	0.100
X_4	0.545	0.998	0.609	0.093

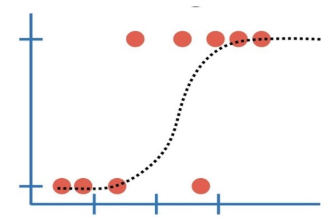
	average-runtime
Multiple_Imputation	0.2050613
ListwiseDeletion	0.0100341

Multiple Imputation

	Percent_Bias	Coverage_Rate	Avg_Width	RMSE
Intercept	0.522	0.951	1.189	0.217
X_1	8.813	0.968	0.810	0.163
X_2	0.216	0.973	1.194	0.193
X_3	2.863	0.956	0.661	0.142
X_4	0.924	0.974	0.609	0.110

Case 3: Logistic regression model & probability to be missing depends *only* on *Y*

HYPOTHESIS **If missingness is confined to predictors *X* and depends only on *Y* for a logistic regression model, listwise deletion regression coefficients are unbiased.**



METHODS

1. Create data from logistic regression model, where $\beta_1 = 1, \beta_2 = 2$.
2. Implement missingness where *Y = 0* observations have greater missingness probability in predictors than *Y = 1* observations.

Listwise Deletion

	Percent_Bias	Coverage_Rate	Avg_Width	RMSE
Intercept	NA	0.000	0.536	0.968
x1	0.490	0.988	0.627	0.127
x2	9.933	0.880	0.790	0.254

	avg-runtime
Multiple_Imputation	0.3609096
ListwiseDeletion	0.0142841

Multiple Imputation

	Percent_Bias	Coverage_Rate	Avg_Width	RMSE
Intercept	NA	0.994	1.046	0.161
x1	17.361	0.954	1.658	0.282
x2	3.643	0.985	1.873	0.264

DISCUSSION

- In each of the situations above, **listwise deletion is orders of magnitude faster and provides unbiased estimates of regression coefficients.**
- If in doubt, multiple imputation for your dataset is the safest approach
- There are far more imputation methods than just the two discussed here; it is vital to deliberately consider which imputation method is best for your dataset when performing statistical analyses.

	Runtime Ratio (LD/MI)
Case 1	~10x faster
Case 2	~20x faster
Case 3	~35x faster