BIOS 6312: Modern Regression Analysis

Andrew J. Spieker, Ph.D.

Assistant Professor of Biostatistics
Vanderbilt University Medical Center

Set 14: Methods for Missing Data

Version: 04/07/2022

Table of Contents

Missing data

Reasons for missing data:

- Missing data *very* common in human studies.
- Why might data be missing?
  - Trivially.
  - By design.
  - By circumstance.

Missing data

Reasons for missing data: Trivial examples

- Everyone in the population *not* in your sample is missing.
- All variables *not* measured in your study are missing.
- In a randomized trial in which participants receive either treatment 0 or 1, the response to the treatment that they did *not* receive is missing.
  - This idea is *fundamental* to causal inference.

Missing data

Reasons for missing data: Missing by design

- Termination of study (i.e. *administrative censoring*)
  - Spent a unit learning methods to handle missing data of this type!
- Obtaining measures that are costly or invasive on a random *subset* of the original sample.

Missing data

Reasons for missing data: Missing by circumstance

- Error: illegible forms or clear data entry errors.
- Patient non-compliance.
- Patient withdrawal from a study (this is different from treatment discontinuation).
- Non-random loss to follow up

Missing data

Missing outcomes:

- Missing data can occur in the outcomes, the predictors, and/or any stratification variables.
- Missing data need to be understood, considered, and addressed.
- To best illustrate the principles, we'll focus on missing *outcomes* at first, and then we'll learn about multiple imputation by chained equations as a method to handle other types of missing data.

Table of Contents

Ad-hoc methods

Example: LDL study

- Suppose we recruit one-hundred individuals with hyperlipidemia to take part in a study to evaluate whether a new drug is successful in controlling LDL levels.
  - $N_0 = 50$ randomized to receive control ($X = 0$).
  - $N_1 = 50$ randomized to receive new drug ($X = 1$).
- After one month, we measure their LDL again to evaluate whether it is in some pre-specified "healthy" range.
  - This example is for illustration purposes only; ignore the simplicity induced by dichotomizing outcomes in this way.

Example: LDL study

X	Y	N
0	**0**	?
0	**1**	?
0	**?**	?
1	**0**	?
1	**1**	?
1	**?**	?

- There are six categories; each participant falls into exactly one.

LDL study: Observed data.

X	Y	N
0	**0**	20
0	**1**	20
0	?	10
1	**0**	20
1	**1**	20
1	?	10

Ad-hoc methods

Method #1: Complete case analysis

- On the basis of the observed data only, appears that both drugs are comparably effective, since:
  - $20/40 = 50\%$ in the *control* group had LDL control.
  - $20/40 = 50\%$ in the *experimental* group had LDL control.
- Problem: You don't know *why* your data are missing.

Method #2: Best/worst outcome (for patient)

			Make-believe	
X	Y	N	"Worst"	"Best"
0	**0**	20	30	20
0	**1**	20	20	30
0	**?**	10	×	×
1	**0**	20	30	20
1	**1**	20	20	30
1	**0**	10	×	×

- "Worst": all missing values assumed to be $Y = 0$.
- "Best": assumed to be $Y = 1$.
- Regardless of their treatment group, X.
- In each of these analyses, the treatment is *still* seen as comparably effective.

Method #3: Best/worst case (for researcher)

			Make-believe	
X	Y	N	"Worst"	"Best"
0	**0**	20	30	20
0	**1**	20	20	30
0	**?**	10	×	×
1	**0**	20	20	30
1	**1**	20	30	20
1	**0**	10	×	×

- Best and worst *case* require you to assume missing outcomes to be different between treatment groups.
- Worst case: treatment harmful; best case, treatment beneficial. Shatters misconception that missingness only matters if missing data rates differ between group.
- Sensible, simple sensitivity analysis approach.

Ad-hoc methods

Example: SBP study

- X: (0 = Control; 1 = Experimental treatment).
- Y: SBP.

ID	X	Y_1	Y_2	Y_3
1	0	150	130	140
2	0	160	150	?
⋮	⋮	⋮	⋮	⋮
99	1	140	120	110
100	1	100	?	?

- Outcomes measured repeatedly over time.
- Missing data occur at some times in some participants.

Method #4: Last observation carried forward (LOCF)

ID	X	Y_1	Y_2	Y_3
1	0	150	130	140
2	0	160	150	**150**
⋮	⋮	⋮	⋮	⋮
99	1	140	120	110
100	1	100	**100**	**100**

- What are some potential flaws in this approach?

Ad-hoc methods

Method #4: Last observation carried forward (LOCF)

- Doesn't tackle reason for missingness ...
- Ignores variation in individual's responses over time.
  - Another similar *ad hoc* method would impute the average of the patient's previous responses.
  - Yet another similar *ad hoc* method would impute the average of all observations at that time point.
  - These suffer from the same problems and are generally not appropriate.

Ad-hoc methods

Notes:

- *Ad-hoc* methods have almost no theoretical justification.
- Must formally characterize different kinds of missing data.

Ad-hoc methods

Kinds of missingness:

- Let M denote indicator of missingness in Y.
- Missing completely at random (MCAR): $Y \perp\!\!\!\perp M$.
  - Missingness independent of all variables in your study.
- Missing at random (MAR): $Y \perp\!\!\!\perp M|\mathbf{X}$.
  - Within strata of $\mathbf{X}$, missingness "completely at random."
- Missing not at random (MNAR): Other cases.
- **Key point**: Nothing in your data will tell you whether your data are missing not at random.
- Principled methods tend to take one of the following forms:
  - Develop and fit models that presume MAR.
  - Sensitivity analyses under hypothetical missingness patterns to form bounds on estimates.

Table of Contents

Weighting methods

Inverse probability weighting: A principled approach

- X: Some predictor of missingness.
- Y: Cumulative medical cost over one year ($\$ \times 100$).
- Seek to estimate E[Y]

ID	X	Y	M
1	0	5	0
2	0	20	0
3	0	?	1
4	0	30	0
5	0	24	0
6	1	12	0
7	1	8	0
8	1	?	1
9	1	11	0
10	1	17	0
⋮	⋮	⋮	⋮

Weighting methods

Inverse probability weighting:

- If data are not MCAR, cannot assume $\mathrm{E}[Y] = \mathrm{E}[Y|M = 0]$ (i.e., cannot rely on complete-case analysis).
- However, if data are MAR, then $\mathrm{E}[Y] = \mathrm{E}[\mathrm{E}[Y|M = 0, X]]$.
- The following estimator is *consistent* for $\mathrm{E}[Y]$:

$$\overline{Y}_{IPW} = \frac{1}{N} \sum_{i=1}^{N} \frac{Y_i(1 - M_i)}{\widehat{\mathrm{P}}(M = 0|X_i)}.$$

- How do you estimate $\mathrm{P}(M = 0|X = x_i)$?
  - Logistic regression, for instance!

Weighting methods

Inverse probability weighting:

- Upweight observations less likely to be observed.
  - Complete cases having values of Y that would suggest a high probability of missing Y.
  - Intuition: Weighting helps to correct the under-representation of this subgroup.
- Weighting methods can be generally unstable if any probabilities are too close to zero.
  - Methods such as truncation designed to accommodate that.
- Need to specify a missingness model, though not a model for the outcome given X.
- Inverse probability weighting is commonly used in all kinds of settings (not just to address missingness).

Table of Contents

Multiple imputation

Ideas:

- A multiple imputation procedure comprises three steps:
  1. Imputation (the goal is to get multiple "complete" data sets).
  2. Analysis (repeat the analysis on the complete data sets).
  3. Pooling (aggregate results).
- The easiest way to see how this works is through an example.

Multiple imputation

Example: Simple linear regression

- X: Medication adherence score (scale #1).
- Y: Medication adherence score (scale #2).
- Linear regression model: $\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x$.
- We want a principled approach to estimate β_1.
- Challenge: Some values for Y are missing!

Example: Simple linear regression

- Take the following imaginary data set.

ID	X	Y	M
1	2	8	0
2	2	1	0
3	5	?	1
4	6	19	0
5	11	10	0
6	12	24	0
7	15	24	0
8	16	?	1
9	18	24	0
10	18	31	0

Regression in Stata: Complete-case analysis by default

```
. regress Y X, robust

Linear regression                               Number of obs     =          8
                                                F(1, 6)           =      25.02
                                                Prob > F          =     0.0024
                                                R-squared         =     0.7289
                                                Root MSE          =     5.7427
```

Y	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
X	1.331667	.2662031	5.00	0.002	.6802911	1.983042
_cons	3.6425	3.869269	0.94	0.383	-5.82526	13.11026

Procedure: Regression-based approach

1. Using complete cases, estimate $\boldsymbol{\beta}$ and error variance σ^2.
2. Several times (for $k = 1, \ldots, K$):
   1. Impute missing values for Y based on random draws from, say, $\mathcal{N}(\hat{\beta}_0 + \hat{\beta}_1 X, \widehat{\sigma}^2)$.
   2. Obtain estimates $\hat{\beta}_1^{(k)}$, standard error $\widehat{\mathrm{SE}}(\hat{\beta}_1^{(k)})$.
3. Aggregate results from imputed data sets (Rubin):

$$\widehat{\beta}_1^A = \frac{1}{K}\sum_{k=1}^{K}\widehat{\beta}_1^{(k)}$$

$$\widehat{\mathrm{Var}}(\widehat{\beta}_1) = \frac{1}{K}\sum_{k=1}^{K}\widehat{\mathrm{Var}}(\widehat{\beta}_1^{(k)}) + \left(1+\frac{1}{K}\right)\sum_{k=1}^{K}\frac{\left(\widehat{\beta}_1^{(k)} - \widehat{\beta}_1^A\right)^2}{K-1}$$

Multiple imputation

Multiple imputation: Help from Stata!

```
mi set mlong
mi register imputed Y X
mi impute regress Y, add(100)
mi estimate: regress Y X, robust
```

Multiple imputation: Help from Stata!

```
. mi estimate: regress Y X, robust

Multiple-imputation estimates                   Imputations       =        100
Linear regression                               Number of obs     =         10
                                                Average RVI       =     0.4441
                                                Largest FMI       =     0.3767
                                                Complete DF       =          8
DF adjustment:   Small sample                   DF:     min       =       4.53
                                                        avg       =       4.55
                                                        max       =       4.57
Model F test:       Equal FMI                   F(   1,    4.5)   =       5.25
Within VCE type:        Robust                  Prob > F          =     0.0760
```

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
X	1.1037	.4815596	2.29	0.076	-.1741465	2.381547
_cons	6.043507	6.362988	0.95	0.390	-10.78504	22.87206

Multiple imputation

Notes:

- Presumes MAR.
- This is one kind of imputation approach (there are many, and we're about to learn another).
- Multiple imputation, which based on correctly specified models, can perform well even in the presence of large amounts of missing data.
- Generally do not need a particularly *large* number of imputations (though more does not hurt).
- The Rubin formula was designed (and is better equipped) for Bayesian imputation.
  - For whatever reason, it's gained enormous popularity in the frequentist world as well—possibly because it's intuitive, elegant, and not too complicated!

Multiple imputation

Chained equations:

- Suppose now that we have missing values on more than one variable.
- One popular approach is to use chained equations to impute.

Multiple imputation

Chained equations: To generating a complete data set

1. Step 1: Temporarily replace missing values with the (complete-case) sample mean of the corresponding variable as a placeholder.
2. Step 2: Set back to missing the placeholder imputations for the first variable having missing values.
3. Step 3: Regress that variable on any other variables you like via, for instance (complete-case) regression.
4. Step 4: Use the fitted model of Step 3 to generate random draws for missing values. (When this variable will be subsequently used as an independent variable in models for other variables, the predicted values will have filled in the missing values and should be used.)
5. Step 5: Repeat Steps 2–4 for each variable that has missing data.

Note: The rule for aggregating imputation-based estimates is the same!

Multiple imputation

Example: REACH

- Y: Six-month A1c.
- X: REACH (0 = Control; 1 = REACH).
- W: Baseline A1c.
- Z: Baseline SDSCA (Summary of Diabetes Self-Care Activities Measure).
- Regression model:
  $\mathrm{E}[Y|X = x, W = w, Z = z] = \beta_0 + \beta_1 x + \beta_2 w + \beta_3 z.$
  - Seek to estimate β_1.
- As you may know, missing data on many variables!

REACH: Characterizing missing data

```
. misstable summarize
```

Variable	Obs=.	Obs>.	Obs<.	Obs<. Unique values	Obs<. Min	Obs<. Max
educyears	9		496	29	4	28
dmdur	7		498	40	1	49
a1c0	10		495	84	4.7	15.7
a1c6	63		442	90	4.4	17.8
a1c12	62		443	92	4.5	18.1
sdsca0	7		498	61	0	8
sdsca6	80		425	58	0	8.3
sdsca12	82		423	55	0	8.1

REACH: Characterizing missing data

```
. misstable pattern
```

Missing-value patterns
(1 means complete)

Percent	Pattern 1	2	3	4	5	6	7	8
70%	1	1	1	1	1	1	1	1
4	1	1	1	1	0	0	0	0
3	1	1	1	1	1	1	1	0
3	1	1	1	1	1	0	0	1
2	1	1	1	1	0	1	1	0
2	1	1	1	1	0	1	1	1
2	1	1	1	1	1	0	1	1
2	1	1	1	1	1	1	0	1
2	1	1	1	1	0	1	0	0
2	1	1	1	1	1	0	0	0
2	1	1	1	1	1	1	0	0
<1	1	0	1	1	1	1	1	1

⋮

Variables are (1) dmdur (2) sdsca0 (3) educyears (4) a1c0 (5) a1c12 (6) a1c6 (7) sdsca6 (8) sdsca12

REACH: Complete-case analysis

```
. regress a1c6 a1c0 sdsca0 reach, robust

Linear regression                               Number of obs     =        429
                                                F(3, 425)         =      37.96
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2756
                                                Root MSE          =     1.7383
```

a1c6	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
a1c0	.5326862	.0516836	10.31	0.000	.4310989	.6342735
sdsca0	.0424605	.0571782	0.74	0.458	-.0699268	.1548477
reach	-.7257798	.1682805	-4.31	0.000	-1.056545	-.3950142
_cons	3.879427	.5803708	6.68	0.000	2.738672	5.020181

Note: $(505 - 429)/505 \approx 15.0\%$ missingness!

REACH: Chained equations (register)

```
. mi register imputed reach age gender educyears dmdur a1c0 a1c6 a1c12 sdsca0 sdsca6 sdsca12
(149 m=0 obs. now marked as incomplete)
```

REACH: Chained equations (impute)

```
. mi impute chained (regress) a1c6 a1c0 sdsca0, add(100) rseed(1)

Conditional models:
            sdsca0: regress sdsca0 a1c0 a1c6
              a1c0: regress a1c0 sdsca0 a1c6
              a1c6: regress a1c6 sdsca0 a1c0

Performing chained iterations ...

Multivariate imputation                     Imputations =      100
Chained equations                                 added =      100
Imputed: m=1 through m=100                      updated =        0

Initialization: monotone                     Iterations =     1000
                                                burn-in =       10

              a1c6: linear regression
              a1c0: linear regression
            sdsca0: linear regression
```

Variable	Observations per *m*			
	Complete	Incomplete	Imputed	Total
a1c6	442	63	63	505
a1c0	495	10	10	505
sdsca0	498	7	7	505

(complete + incomplete = total; imputed is the minimum across *m*
 of the number of filled-in observations.)

REACH: Chained equations (analyze and aggregate)

```
. mi estimate: regress a1c6 a1c0 sdsca0 reach, robust

Multiple-imputation estimates                   Imputations       =        100
Linear regression                               Number of obs     =        505
                                                Average RVI       =     0.1786
                                                Largest FMI       =     0.2203
                                                Complete DF       =        501
DF adjustment:   Small sample                   DF:     min       =     328.12
                                                        avg       =     381.99
                                                        max       =     434.24
Model F test:       Equal FMI                   F(   3,  479.6)   =      35.64
Within VCE type:        Robust                  Prob > F          =     0.0000
```

a1c6	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
a1c0	.5245264	.0507886	10.33	0.000	.4246928	.62436
sdsca0	.0381755	.0607274	0.63	0.530	-.0812887	.1576396
reach	-.6147974	.1637703	-3.75	0.000	-.9366785	-.2929163
_cons	3.907537	.603816	6.47	0.000	2.719955	5.095119

Multiple imputation

Example: REACH

- In these data, the imputation model does not result in a meaningful difference from the complete-case analysis.
- However, that will not always be so!

Table of Contents

Pragmatic studies

Missingness: Not always obvious

- It can sometimes be that you don't actually have as much missing data as you believe.
- Imagine a trial with an intervention that seeks to improve patient functionality in very sick subpopulations.
- The outcome is twenty-minute walk (meters, for instance).
- If someone dies on study prior to outcome measurement, what do we do?

Example:

- X: Treatment group.
- Y: Walking distance (meters).
- D: indicator of death.

ID	Tx	D	Y
1	0	0	40
2	1	0	20
3	0	1	?
4	0	0	35
5	1	0	4
6	1	1	?
7	0	0	65
8	1	1	?
9	0	0	120
⋮	⋮	⋮	⋮

Pragmatic studies

What are we estimating?

- Estimand of a complete-case analysis:
  - Mean difference among those who survive.
- If you try to fill in the values, what are you estimating?
  - Mean difference under the hypothetical (counterfactual) scenario in which the whole population survives long enough to walk at all.
- In a pragmatic trial, *neither* of these is of interest.
- Pragmatic approach: if someone dies, there is a somewhat high chance that they will walk approximately zero meters in twenty minutes. In a pragmatic sense, this is not truly the same thing as a missing data problem.

What are we estimating?

- Note: In the setting of many variables with many reasons for missingness, you can mix missing data methods with pragmatic questions.
- I do this in my research of cost outcomes.
  - I am often interested in mean cost under the hypothetical scenario in which the whole population receives a particular treatment, is not censored, and survive for as long as they would under that particular treatment strategy.
  - Note the three levels of missingness: Censoring, outcome under treatment not actually received, and death.

Pragmatic studies

Withdrawal from treatment vs. study: A key difference

- Missing data can not be avoided completely. However, there are some things you can do.
- Make a clear distinction between withdrawal from treatment and withdrawal from study.
- If withdrawn from treatment, should continue follow-up.
- Subject withdrawal from study should be:
  - Distinctly different from withdrawal from treatment.
  - Patient-initiated.
  - Done when the patient is gently made aware that he or she is compromising the integrity of the study by not authorizing you to follow them for any measurements or contact them.
- Distinction must be made *a priori*.

Recommendations

Prepare!

- Anticipate the challenges likely to occur and strategize how to mitigate those challenges.
- Collect information on variables that may predict missingness. Follow up with graphical and tabular methods.
- Distinguish between fundamentally unobservable vs. circumstantially unobserved.
- When confronted with missing data, *sensitivity analyses*!
- Recognize that there is *nothing* in your data that can tell you whether something is MAR, MCAR, MNAR.
- Avoid data *errors*!

Summary

Notes: Topics in this unit

- Missing data and its forms.
- *Ad hoc* methods
  - Appealing in that they address the problem of missing data in a way that is easy to implement an explain.
  - However, methods that are too simple come at the cost of validity.
- Weighting and imputation methods.
  - Fairly easy to implement in statistical software and have intuitive explanations associated with them.
- Let the scientific question inform your choices in how you view and address missing data!