

Coping with Missing Data

Martha Storandt

Research Methods Series

Cohen et al. 2003, Chapter 10

- Basic Issue: Minimize missing data

Basic Issue

- Minimize missing data
 - Know what information you will need for data analysis before you begin

Basic Issue

- Minimize missing data
 - Know what information you will need
 - Don't ask questions people frequently won't be able to answer

Basic Issue

- Minimize missing data
 - Know what information you will need
 - Don't ask questions people won't answer
 - Review data before it's too late to go back and obtain missing information when it is detected

In choosing a method, consider

- How much data are missing?
 - Choice of method is less critical if only a small percentage of data are missing.

In choosing a method, consider

- How much data are missing?
- How large is the sample?
 - Some methods only work for large sample
($N > 200$)

In choosing a method, consider

- How much data are missing?
- How big is the sample?
- Why are the data missing?

Why are data missing?

- “Missingness” of X depends on value of X
 - doesn't want to tell you
- Value of X is out of range of test
 - test is too hard for person
- Contingent on response to other questions
 - Are you sad? If so, why?
- Completely at random

In choosing a method, consider

- How much data are missing?
- How big is the sample?
- Why are the data missing?
- Who will be using the data set?

Traditional approaches

- Drop variable

Traditional approaches

- Drop variable
- Drop subjects
 - listwise

Dropping subjects

- If number dropped is small and N is large, usually not a problem, but...

Dropping subjects

- If number is large, may
 - Bias sample
 - Observed values may be higher or lower
 - Observed values may be less variable
 - Bias random assignment to conditions
 - Lose power

Traditional approaches

- Drop variable
- Drop subjects
 - Listwise deletion
 - Pairwise deletion

Illustrative Missing-Date Score Matrix

Subject	Y	X ₁	X ₂	X ₃
1	72	38	6	92
2	84	52	12	114
3	63	47		108
4	81	63	8	
5	47			
6	62		7	110

Pairwise deletion

- Uses as many cases as have values for pair of variables
- Thus correlations in matrix based on different samples sizes
- Use only if data are missing completely at random

Treatment studies: Dropouts and nonadherence

Two analytic strategies:

- As treated
- As randomized (intent to treat)

(Petkova & Teresi, Psychosomatic Medicine,
2002)

As randomized (intent to treat)

- Usually failure to complete therapy is partially an outcome
- Is a nonignorable bias
- Include noncompliance in the analytic model

Treatment of obsessive-compulsive behavior

- Exposure-based behavioral therapy
- High dropout rate (e.g., 50-60%); can't comply with exposure
- But high success (e.g., 90%) in those who complete

Traditional approaches

- Drop variable
- Drop subjects
- Average available items

use if scale has high reliability and
measures a single, well-defined domain

Traditional approaches

- Drop variable
- Drop subjects
- Average available items
- Substitute means

Substitute means

- Advantages
 - Simple
 - Doesn't change estimate of population mean
 - If use in regression, doesn't change B for variable

Substitute means

- Disadvantages
 - Artificially deflates standard error of B because of larger N (cheating?)
 - Missing cases may differ systematically

Traditional approaches

- Drop variable
- Drop subjects
- Average items
- Substitute means
- Code missingness

Missing Data on a Nominal Scale

- Can use any of the usual procedures for coding a nominal scale
 - Dummy coding
 - Effects coding
 - Contrast coding

Example: Religion

- Protestant
- Catholic
- Jewish
- Other/none
- Refused to answer

Maximum Likelihood Estimation

- Estimates the statistics for the full sample
- Assumes data are missing completely at random
- Expectation-maximization (EM) algorithm: alternately estimate and maximize
- Can be used for missing data in either independent or dependent variables

Total Sample

	Y = 0	Y = 1	Missing Y	Total
X = 0	250	50	80	380
X = 1	350	50	20	520
Missing X	30	70		100
Total	630	270	100	1000

Cases with complete data

	Y = 0	Y = 1	Total
X = 0	250	50	300
X = 1	350	150	500
Total	600	200	800

Cases (%) with complete data

	Y = 0	Y = 1	Total	
X = 0	.3125	.0625	.375	
X = 1	.4375	.1875	.625	
Total	.75	.25	1.00	(N = 800)

Allocate missing cases

	Y = 0	Y = 1	Missing Y	Total
X = 0	250 + 80 (250/300) + 30 (250/600)	50	(80)	380
X = 1	350	150	20	520
Miss. X	(30)	70		100
Total	630	270	100	1000

Allocate missing cases

	Y = 0	Y = 1	Missing Y	Total
X = 0	250 +80(.8333) +30(.4167)	50	(80)	380
X = 1	350	150	20	520
Miss. X	(30)	70		100
Total	630	270	100	1000

Allocate missing cases

	Y = 0	Y = 1	Missing Y	Total
X = 0	250 +66.67 +12.50	50	(80)	380
X = 1	350	150	20	520
Miss. X	(30)	70		100
Total	630	270	100	1000

Allocate missing cases

	Y = 0	Y = 1	Missing Y	Total
X = 0	250	50	(80)	380
	+66.67	+13.33		
	+12.50	+17.50		
X = 1	350	150	(20)	520
	+14.00	+6.00		
	+17.50	+52.50		
Miss. X	(30)	70		100
Total	630	270	100	1000

Allocate missing cases

	Y = 0	Y = 1	Total
X = 0	329.17	80.83	410
X = 1	381.50	208.50	590
Total	710.67	289.33	1000

Estimated proportions including missing cases
(estimates based on complete cases alone)

	Y = 0	Y = 1	Total
X = 0	.3292 (.3125)	.0808 (.0625)	.41 (.375)
X = 1	.3815 (.4375)	.2085 (.1875)	.59 (.625)
Total	.7107 (.75)	.2893 (.25)	1.00

Estimated proportions after 3 iterations
(estimates based on complete cases alone)

	Y = 0	Y = 1	Total
X = 0	.3271 (.3125)	.0875 (.0625)	.4146 (.375)
X = 1	.3790 (.4375)	.2063 (.1875)	.5854 (.625)
Total	.7062 (.75)	.2938 (.25)	1.00

Imputation

- Hot deck

Imputation

- Hot deck
- Multiple imputation and representation of uncertainty
 - Free software: www.multiple-imputation.com

Imputation

- Hot deck
- Multiple imputation and representation of uncertainty
- Ordinary least squares
 - Use other variable(s) to predict missing value

Imputation

- Hot deck
- Multiple imputation and representation of uncertainty
- Ordinary least squares
- Maximum likelihood estimates
 - Typically assume data missing completely at random

Comparison of methods

- Cohen et al's academic salary example
- Dependent variable: Salary
- Independent variables
 - Time since PhD
 - Sex
 - Number of publications
 - Citations

Comparison of methods

- 62 cases with complete data
- 7 missing citations
- 8 missing time since PhD
- 10 missing publications and citations

Summary of estimates of effects (Bs)
using four different methods

Method	Time	Sex	Pubs	Citations
Listwise	857	918	93	202
Pair-wise	1055	1247	51	144
Means	769	1145	128	161
Means/dummy	736	502	134	196
EM algorithm	869	1515	118	154
Imputation	743	1143	131	214