A Basic Introduction to Missing Data

John Fox

Sociology 740

Winter 2014

Outline

Why Missing Data Arise

- *Global* or *unit non-response*.
  - In a survey, certain respondents may be unreachable or may refuse to participate.
- *Item non-response*.
  - Some respondents may not know the answers to specific questions, or may refuse to respond to them.
- Errors in data collection or processing.
  - An interviewer may fail to ask a question of a survey respondent.

Why Missing Data Arise

- Missing data may be built into the design of a study.
  - Particular questions may be asked only of a random subset of respondents.
- Some data values in a study may be *censored*.
  - In survival (event-history) analysis, the focal event (e.g., death) may not occur for some subjects before the end of the period of observation.
- Missing data should be distinguished from data that are *conditionally undefined*.
  - A survey respondent who has no children cannot report their ages.

Missing Data Basics

- There is no magic cure for missing data
  - It is generally impossible to proceed in a principled manner without making at least partly unverifiable assumptions about the process that gives rise to the missing information.
- Rubin (1976) introduced some key distinctions concerning missing data.
  - Let the matrix $\underset{(n\times p)}{\mathbf{X}}$ represent the complete data for a sample of n observations on p variables.
  - Some of the entries of $\mathbf{X}$, denoted by $\mathbf{X}_{\text{mis}}$ are missing, and the remaining entries, $\mathbf{X}_{\text{obs}}$ are observed.

Missing Data Basics

Kinds of Missing Data: Missing Completely at Random

- Missing data are *missing completely at random* (*MCAR*) if the missing data (and hence the observed data) can be regarded as a simple random sample of the complete data.
  - The probability that a data value is missing (*missingness*) is unrelated to the data value itself or to any other value, missing or observed, in the data set.

Missing Data Basics

Kinds of Missing Data: Missing at Random

- If missingness is related to the observed data but—conditioning on the observed data—not to the missing data, then data are *missing at random* (*MAR*).
  - Certain individuals may refuse to report their income, and may differ systematically in income from the sample as a whole.
  - If one respondent's decision to withhold information about income is independent of others' responses and *conditional on* the information provided (e.g., education, occupation), failure to provide information on income is independent of income, then the data are MAR.
  - MCAR is a stronger condition—and a special case—of MAR.

Missing Data Basics

Kinds of Missing Data: Missing Not at Random

- If missingness is related to the missing values themselves even when the information in the observed data is taken into account, then missing data are *missing not at random* (*MNAR*).
  - If conditional on all of the observed data, individuals with higher incomes are more likely than others to withhold information about their incomes, then the missing income data are MNAR.

Missing Data Basics

Ignorable vs. Non-Ignorable Missing Data

- If the data are MCAR or MAR then it is not necessary to model the process that generates the missing data in order to accommodate the missing data.
  - The "mechanism" that produces the missing data is *ignorable*.
- When data are MNAR, the missing-data mechanism is *non-ignorable*.
  - It is necessary to model this mechanism to deal with the missing data in a valid manner.

Missing Data Basics

Can We Establish Whether Missing Data are Ignorable?

- Except in some special situations (e.g., when missing data are missing by design), it is not possible to know whether data are MCAR, MAR, or MNAR.
- Showing that missingness on some variable in a data set is related to observed data on other variables, proves that the missing data are not MCAR.
- Demonstrating that missingness in a variable is not related to observed data in other variables does not *prove* that the missing data are MCAR.
  - Non-respondents may be differentiated from respondents in some *unobserved* manner.

Missing Data Basics

Preliminary Examples

- 250 observations are sampled from a bivariate-normal distribution with means $\mu_1 = 10$, $\mu_2 = 20$, variances $\sigma_1^2 = 9$, $\sigma_2^2 = 16$, and covariance $\sigma_{12} = 8$.
  - The population correlation between X_1 and X_2 is $\rho_{12} = 8/\sqrt{9 \times 16} = 2/3$.
  - The slope for the regression of X_1 on X_2 is $\beta_{12} = 8/16 = 1/2$
  - The slope for the regression of X_2 on X_1 is $\beta_{21} = 8/9 \approx 0.889$.
- The variable X_1 is completely observed, but missing data on X_2 will be generated in different ways.
  - This pattern—where one variable has missing data and all others are completely observed—is called *univariate missing data*.

Missing Data Basics

Preliminary Examples: Three Mechanisms for Generating Missing Data

- MCAR: 100 of the observations on X_2 are selected at random and set to missing.
- MAR: an observation's missingness on X_2 is related to its (observed) value of X_1:

$$\Pr(X_{i2} \text{ is missing}) = \frac{1}{1 + \exp\left[\frac{1}{2} + \frac{2}{3}(X_{i1} - 10)\right]}$$

  - The logistic regression coefficients were calibrated so that approximately 100 observations will have missing data on X_2, with the probability that X_2 is missing declining as X_1 grows.
  - Because X_1 and X_2 are positively correlated, there are relatively few small values of X_2.

Missing Data Basics

Preliminary Examples: Three Mechanisms for Generating Missing Data

- MNAR: an observation's missingness on X_2 is related to the (potentially unobserved) value of X_2 itself:

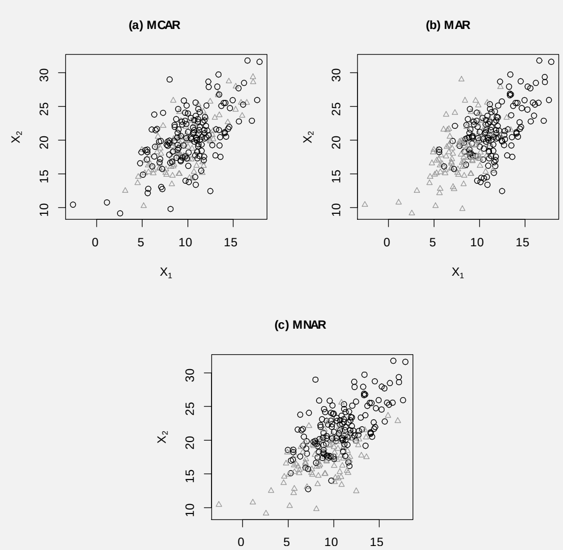
$$\Pr(X_{i2} \text{ is missing}) = \frac{1}{1 + \exp\left[\frac{1}{2} + \frac{1}{2}(X_{i2} - 20)\right]}$$

  - Here too there are relatively few small values of X_2.

Outline Why Missing Data Arise Basics Traditional Approaches Principled Approaches

Missing Data Basics

Preliminary Examples: Three Mechanisms for Generating Missing Data



Triangles are missing observations.

John Fox Sociology 740
A Basic Introduction to Missing Data

Some Traditional Approaches to Missing Data

Questions to answer

1. Does the method provide *consistent estimates* of population parameters, or does it introduce systematic biases?
2. Does the method provide *valid statistical inferences*, or are confidence intervals and p-values distorted?
3. Does the method use the observed data *efficiently* or does it discard information?

Some Traditional Approaches to Missing Data

Complete-Case Analysis

- Also called *list-wise* or *case-wise deletion* of missing data.
- Probably the most widely used approach.
- Ignores observations with any missing data on the variables included in the analysis.
- Advantages:
  - Simple to implement.
  - Provides consistent estimates and valid inferences when the missing data are MCAR.
  - Provides consistent estimates of regression coefficients and valid inferences when missingness does not depend on the response variable (even if data are not MCAR).

Some Traditional Approaches to Missing Data

Complete-Case Analysis

- Disadvantages:
  - Because it discards some valid data, complete-case analysis generally is not efficient.
    - With many variables, a great deal of valid data may be discarded.
  - When data are MAR or MNAR, complete-case analysis usually provides biased results and invalid inferences.

Some Traditional Approaches to Missing Data

Available-Case Analysis

- Also called *pair-wise deletion* of missing data.
- Uses all non-missing observations to compute each statistic of interest.
  - In a least-squares regression, the regression coefficients can be calculated from the means, variances, and covariances of the variables.

Some Traditional Approaches to Missing Data

Available-Case Analysis

- Problems:
  - Available case analysis appears to use more information than complete-case analysis, but estimators based on available cases can be *less* efficient than those based on complete cases.
  - By basing different statistics on different subsets of the data, available-case analysis can lead to nonsensical results, such as correlations outside the range from -1 to $+1$.
  - Except in simple cases, such as linear least-squares regression, it is not obvious how to apply the available-case approach.
  - Available-case analysis generally provides biased estimates and invalid inferences when data are MAR or MNAR.

Some Traditional Approaches to Missing Data

Imputation

- Replacing missing values with plausible *imputed* values.
  - The resulting completed data set is then analyzed using standard methods.

Some Traditional Approaches to Missing Data

Unconditional Mean Imputation

- *Unconditional mean imputation* (or *mean substitution*) replaces each missing value with the mean of the observed data for the variable.
  - Mean imputation preserves the means of variables, but it makes their distributions less variable and tends to weaken relationships between variables.
  - Mean imputation generally yields biased regression coefficients and invalid inferences even when data are MCAR.
  - By treating the missing data as if they were observed, mean imputation exaggerates the effective size of the data set, further distorting statistical inference—a deficiency that it shares with other simple imputation methods.

Some Traditional Approaches to Missing Data

Conditional Mean Imputation

- *Conditional-mean imputation* replaces missing data with predicted values, obtained, for example, from a regression equation (*regression imputation*).
  - The imputed observations tend to be less variable than real data, because they lack residual variation.
  - Another problem is that we have failed to account for uncertainty in the estimation of the regression coefficients used to obtain the imputed values.
  - Regression imputation improves on unconditional mean imputation, but it generally provides biased estimates and invalid inferences even for missing data that are MCAR.

Some Traditional Approaches to Missing Data

Conditional Mean Imputation

- The first of problem with regression imputation (removal of residual variation) can be addressed by adding a randomly sampled residual to each filled-in value.
- The second problem (uncertainty in the regression coefficients used for prediction of missing data) leads naturally to Bayesian multiple imputation of missing values.

Some Traditional Approaches to Missing Data

Application to the Illustrative MAR Data: Based on A Simulation With 1000 Samples

Parameter	*Complete Cases*	*Mean Imput.*	*Regr. Imput.*	*Multiple Imput.*
	Mean Parameter Estimate (RMSE)			
$\mu_1 = 10$	11.476	10.001	10.001	10.001
	(1.489)	(0.189)	(0.189)	(0.189)
$\mu_2 = 20$	21.222	21.322	20.008	20.008
	(1.355)	(1.355)	(0.326)	(0.344)
$\beta_{12} = 0.5$	0.391	0.391	0.645	0.498
	(0.117)	(0.117)	(0.151)	(0.041)
$\beta_{21} = 0.889$	0.891	0.353	0.891	0.890
	(0.100)	(0.538)	(0.100)	(0.106)

Some Traditional Approaches to Missing Data

Application to the Illustrative MAR Data: Based on A Simulation With 1000 Samples

Parameter	*Complete Cases*	*Mean Imput.*	*Regr. Imput.*	*Multiple Imput.*
	Conf.-Interval Coverage (Mean Interval Width)			
μ_1	0	.951	.951	.951
	(0.792)	(0.750)	(0.750)	(0.746)
μ_2	.005	0	.823	.947
	(1.194)	(0.711)	(0.881)	(1.451)
β_{12}	.304	.629	.037	.955
	(0.174)	(0.246)	(0.140)	(0.175)
β_{21}	.953	0	.661	.939
	(0.396)	(0.220)	(0.191)	(0.463)

Principled Approaches to Missing Data

Maximum-Likelihood Estimation

- The method of maximum likelihood can be applied to parameter estimation in the presence of missing data.
  - Doing so requires making assumptions about the distribution of the complete data and about the process producing missing data.
  - If the assumptions hold, then the resulting maximum-likelihood estimates have their usual optimal properties, such as consistency and asymptotic efficiency.
- Let $p(\mathbf{X}; \boldsymbol{\theta}) = p(\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}; \boldsymbol{\theta})$ represent the joint probability-density for the complete data $\mathbf{X}$.
  - Rubin (1976) showed that the maximum-likelihood estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ can be obtained from the marginal distribution of the observed data, *if data are MAR*.

Principled Approaches to Missing Data

Maximum-Likelihood Estimation

- The *expectation-maximization* (*EM*) algorithm, due to Dempster, Laird, and Rubin (1977), is a general iterative method for finding maximum-likelihood estimates in the presence of arbitrary patterns of missing data.
  - The EM algorithm is broadly applicable, generally easy to implement, and effective
  - A disadvantage of the EM algorithm is that it does not produce standard errors for the estimated parameters.

Principled Approaches to Missing Data

Bayesian Multiple Imputation

- *Bayesian multiple imputation* (*MI*) is a flexible and general method for dealing with data that are missing at random.
- Like maximum-likelihood estimation, MI begins with a specification of the distribution of the complete data (assumed to be known except for a set of parameters to be estimated from the data).

Principled Approaches to Missing Data

Bayesian Multiple Imputation

- The essential idea of multiple imputation is to reflect the uncertainty associated with missing data by imputing *several* values for each missing value, each imputed value drawn from the *predictive distribution* of the missing data, and therefore producing not one but several completed data sets.
  - Standard methods of statistical analysis are then applied in parallel to the completed data sets.
  - Parameters of interest are estimated along with their standard errors for each imputed data set.
  - Estimated parameters are averaged across completed data sets.
  - Standard errors are combined across imputed data sets, taking into account the variation among the estimates in the several data sets, thereby capturing the added uncertainty due to having to impute the missing data.

Principled Approaches to Missing Data

Bayesian Multiple Imputation

- A multivariate-normal model for the complete data is both relatively simple and useful in applications.
  - Because the model assumed to describe the complete data is used just to obtain imputed values for the missing data, results produced by MI are usually not sensitive to the assumption of multivariate normality.
  - But there are some pitfalls to be avoided (discussed below).
- The details of methods for drawing multiple imputations from the multivariate-normal model are beyond the scope of this presentation.

Principled Approaches to Missing Data

Bayesian Multiple Imputation: Rubin's Rules

- Having obtained g completed data sets, we produce g sets of regression coefficients, $B_0^{(l)}, B_1^{(l)}, \ldots . B_k^{(l)}$, and coefficient standard errors, $\text{SE}\left(B_0^{(l)}\right), \text{SE}\left(B_1^{(l)}\right), \ldots, \text{SE}\left(B_k^{(l)}\right)$, for $l = 1, \ldots, g$.
- Rubin (1987) provides simple rules for combining information across multiple imputations, valid as long as the sample size is sufficiently large for the separate estimates to be approximately normally distributed.
  - Point estimates of the population regression coefficients are obtained by averaging across imputations:

$$\widetilde{\beta}_j \equiv \frac{\sum_{l=1}^{g} B_j^{(l)}}{g}$$

Principled Approaches to Missing Data

Bayesian Multiple Imputation: Rubin's Rules

- Standard errors of the estimated coefficients are obtained by combining information about within and between-imputation variation in the coefficients:

$$\widetilde{\text{SE}}(\widetilde{\beta}_j) \equiv \sqrt{V_j^{(W)} + \frac{g+1}{g} V_j^{(B)}}$$

where the within-imputation component is $V_j^{(W)} \equiv \frac{\sum_{l=1}^{g} \text{SE}^2\left(B_j^{(l)}\right)}{g}$ and the between-imputation component is $V_j^{(B)} \equiv \frac{\sum_{l=1}^{g}\left(B_j^{(l)} - \widetilde{\beta}_j\right)^2}{g-1}$.

Principled Approaches to Missing Data

Bayesian Multiple Imputation: Rubin's Rules

- Inference based on $\widetilde{\beta}_j$ and $\widetilde{\mathrm{SE}}(\widetilde{\beta}_j)$ uses the t-distribution, with degrees of freedom

$$df_j = (g-1)\left(1 + \frac{g}{g+1} \times \frac{V_j^{(W)}}{V_j^{(B)}}\right)^2$$

  - For example, to construct a 95-percent confidence interval for β_j,

$$\beta_j = \widetilde{\beta}_j \pm t_{.025, df_j} \widetilde{\mathrm{SE}}(\widetilde{\beta}_j)$$

Principled Approaches to Missing Data

Bayesian Multiple Imputation: Missing Information

- The *estimated rate of missing information* about the parameter β_j is

$$\widehat{\gamma}_j = \frac{R_j}{R_j + 1}$$

where

$$R_j \equiv \frac{g+1}{g} \times \frac{V_j^{(B)}}{V_j^{(W)}}$$

Principled Approaches to Missing Data

Bayesian Multiple Imputation: Efficiency

- The efficiency of the MI estimator relative to the maximally efficient maximum-likelihood estimator is

$$\mathrm{RE}(\widetilde{\beta}_j) = \frac{g}{g + \gamma_j}$$

  - If the number of imputations g is very large, MI is as efficient as ML.
  - Even when the rate of missing information is high and the number of imputations modest, the relative efficiency of the MI estimator hardly suffers.
    - When $\gamma_j = 0.5$ and $g = 5$, then $\mathrm{RE}(\widetilde{\beta}_j) = 5/(5 + 0.5) = 0.91$, and $\sqrt{\mathrm{RE}(\widetilde{\beta}_j)} = 0.95$.

Principled Approaches to Missing Data

Bayesian Multiple Imputation: Practical Considerations

- Multiple imputation cannot preserve features of the data that are not represented in the imputation model.
  - It is important to insure that the imputation model is consistent with the intended analysis.
- Try to include variables in the imputation model that make the assumption of ignorable missingness reasonable.
  - Think of imputation as a pure prediction problem.
  - Finding variables that are highly correlated with a variable that has missing data will likely improve the quality of imputations, as will variables that are related to missingness.
  - Use all relevant variables, even ones not used in the substantive analysis (an “inclusive” strategy).
  - There is nothing wrong in using the response variable to help impute missing values of explanatory variables.

Principled Approaches to Missing Data

Bayesian Multiple Imputation: Practical Considerations

- If possible, transform variables to approximate normality.
- Adjust the imputed data to resemble the original data.
  - For example, imputed values of an integer-valued variable can be rounded to the nearest integer.
  - Imputed values of a 0/1 dummy variable can be set to 0 if less than or equal to 0.5 and to 1 if greater than 0.5.
- Make sure that the imputation model captures relevant features of the data.
  - Using the multivariate-normal distribution for imputations will not preserve *nonlinear* relationships and *interactions* among the variables, unless we make special provision for these features of the data.

Outline Why Missing Data Arise Basics Traditional Approaches Principled Approaches

Principled Approaches to Missing Data

Bayesian Multiple Imputation: Example

- The regression of infant mortality on GDP per capita, the percentage of married women practicing contraception, and the average number of years of education for women.
  - Extension of an earlier example.
  - To linearize the regression, I log-transformed both infant mortality and GDP.
- A complete-case analysis includes only 62 of the 207 countries, with missing data for the individual variables as follows:

Infant Mortality	GDP	Contraception	Female Education
6	10	63	131

John Fox Sociology 740
A Basic Introduction to Missing Data

Principled Approaches to Missing Data

Bayesian Multiple Imputation: Example

- I based imputations on a multivariate-normal model with the four variables in the regression plus the total fertility rate, the expectation of life for women, the percentage of women engaged in economic activity outside the home, and the illiteracy rate for women.
  - Preliminary examination of the data suggested that the multivariate-normal model could be made more appropriate for the data by transforming several of these variables.

Principled Approaches to Missing Data

Bayesian Multiple Imputation: Example

- Means and standard deviations of the variables are as follows, based on complete cases and on ML assuming ignorable missing data:

	$\log_e$Inf.Mort.	$\log_e$GDP	Contra.	Female Ed.
Estimates based on Complete Cases				
Mean	3.041	8.151	50.90	11.30
SD	(1.051)	(1.703)	(23.17)	(3.55)
Maximum-Likelihood Estimates				
Mean	3.300	7.586	44.36	10.16
SD	(1.022)	(1.682)	(24.01)	(3.51)

Principled Approaches to Missing Data

Bayesian Multiple Imputation: Example

- Using Schafer's (1997) data-augmentation method, and employing the multivariate-normal model, I obtained imputations for 10 completed data sets.
- Results:

	Intercept	$\log_e$GDP	Contra.	Female Ed.
		Complete-Case Analysis		
B_j	6.88	-0.294	-0.0113	-0.0770
$\text{SE}(B_j)$	(0.29)	(0.058)	(0.0042)	(0.0338)
		Multiple-Imputation Analysis		
$\widetilde{\beta}_j$	6.57	-0.234	-0.00953	-0.105
$\widetilde{\text{SE}}(\widetilde{\beta}_j)$	(0.18)	(0.049)	(0.00294)	(0.033)
Miss. Inf. $\widehat{\gamma}_j$	0.20	0.61	0.41	0.69