

Missing Data

(material drawn from Little & Schenker 1995)

A. Introduction

1. Social scientists commonly encounter problems of missing, incomplete, and non-comparable data
2. Several specific problems
 - a. unit non-response – subject fails to participate in an interview
 - i. subject may be unavailable
 - ii. subject may refuse to participate
 - iii. in this case, no interview data are available
 - iv. some other data may be available, such as information from the sampling frame or information from other interviews
 - b. item non-response – subject fails to provide a usable answer to a particular survey question
 - i. subject may not know answer
 - ii. subject may refuse
 - iii. question may have been skipped by interviewer
 - iv. some questions depend on interview mode (sensitive questions might be asked in person or through a mail survey but not over the phone)
 - v. responses may be partly or completely masked to prevent inadvertent disclosure (e.g., MSA status is purposefully missing for some respondents in the Current Population Survey)
 - vi. in this case, we have other information (responses by the subject) from the interview itself
 - c. non-comparable data – in repeated surveys or panels,

questions, response categories and definitions can change over time

- i. occupational and industrial codes change every decade or so
- ii. questions are revised
- iii. response categories are expanded or collapsed

3. These problems, in turn, lead to analytical concerns
 - a. the subjects with missing observations or responses may be systematically different from subjects with full information, leading to a form of selectivity bias
 - b. even if the data are randomly missing, there is still a loss of information
 - c. problems of missing data may interact with other survey design issues

B. First steps – after inputting a data set, we first check the extent of missing data and data quality problems

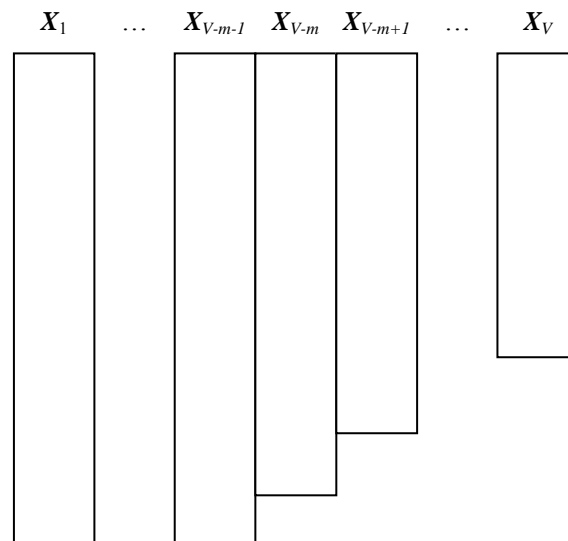
1. We first look for unit non-response (if indicators are available in the data)
 - a. for example, the PSID has “why non-response” variables that indicate why a person or household did not participate in a particular interview wave
 - b. non-response indicators often appear in longitudinal and other complex data sets where subjects (or parts of observational units) may participate in one portion of the study but not in others (e.g., attrition, responses from some family members but not others)
2. We next look at item non-response and the consistency of responses

- a. examine how many subjects have usable responses for the items of interest
 - b. examine how many subjects are missing different combinations of items
3. We also look for “hidden” data problems
- a. many surveys take steps to “fix” data problems (we will be discussing these below)
 - b. we want to know which data may have been affected or altered through this process
 - c. problems indicated by “allocation” flags
4. The potential for bias usually increases with the extent of missing data
- a. if only a small portion of the data for an analysis population are missing, say ten percent or less, the biases and other analytical problems are likely to be negligible
 - b. note the operative term here is “analysis population”
 - i. the overall extent of missing data might be small
 - ii. but the problems may be concentrated in a particular analysis sub-group

C. More formal discussion of patterns of missing data

- 1. descriptions of data
 - a. let \mathbf{X} be an $(N \times V)$ data matrix with elements x_{ij} , where i indexes the respondent (observational unit) and j indexes the type of data item
 - b. let \mathbf{M} be a similarly dimensioned matrix containing binary indicators m_{ij} that equal one if x_{ij} is missing
- 2. monotone and non-monotone patterns of missing data

- a. consider the columns of \mathbf{X} , denote these $\mathbf{X}_1, \dots, \mathbf{X}_V$
- b. suppose that \mathbf{X}_V is the only column with missing data; we would call this a *univariate* pattern
- c. suppose now that several columns (\mathbf{X}_{V-m} through \mathbf{X}_V) have missing data
 - i. arrange the columns from fewest missing elements (\mathbf{X}_{V-m}) to most (\mathbf{X}_V)
 - ii. consider all adjacent columns, \mathbf{X}_{V-j} and \mathbf{X}_{V-j+1} , where $m \leq j \leq 1$
 - iii. if every non-missing element in \mathbf{X}_{V-j+1} has a corresponding non-missing element in \mathbf{X}_{V-j} the missing data follow a *monotone pattern*



- d. if there are multiple columns of missing data that do not follow the above pattern, the data are *non-monotone*
3. also want to consider whether \mathbf{X} and \mathbf{M} are related
 4. let $p(\mathbf{M} | \mathbf{X}, \theta)$ be the conditional distribution of \mathbf{M} given \mathbf{X} and a set of parameters θ

5. data are said to be “missing completely at random” (MCAR) if $p(\mathbf{M} | \mathbf{X}, \theta) = p(\mathbf{M} | \theta)$ for all \mathbf{X}
6. we can assess these relationships by examining distributions of the available \mathbf{X} for people with and without missing data
7. suppose that differences exist (or that they cannot be adequately examined); a next consideration is whether differences between respondents with and without missing data can be accounted for using the available observed variables
 - a. divide \mathbf{X} into two components:
 - i. \mathbf{X}_{obs} observed portion
 - ii. \mathbf{X}_{mis} portion with missing data
 - b. data are said to be “missing at random” (MAR) if $p(\mathbf{M} | \mathbf{X}, \theta) = p(\mathbf{M} | \mathbf{X}_{\text{obs}}, \theta)$ for all \mathbf{X}_{mis}
 - c. example,
 - i. suppose that \mathbf{X} consists of two variables, education and age, and that education is missing for some individuals but age is available for everyone
 - ii. education is MAR if the pattern of missing responses only depends on age
 - iii. education is not MAR if the pattern of missing responses depends at least partly on education itself

D. Naïve approaches for addressing missing data

1. Suppose that data are missing; what do we do next? Economists typically apply the following naïve approaches (listed by frequency of use)

2. Complete-case analysis
 - a. in a complete-case analysis, we discard observations that are missing any of the items that we are interested in for our general analysis
 - b. we have used this approach so far in our homework and example assignments
 - c. strengths
 - i. approach is very easy to implement; all we need to do is check the data for missing observations
 - ii. analyses that adopt this approach are valid (unbiased) if data are MCAR
 - d. weaknesses
 - i. the approach may drop usable data and therefore be inefficient
 - ii. analyses are biased if data are not MCAR
 - iii. bias depends on the strength of the association between M and X and on the extent of missing data (standard selectivity results)

3. Available-case analysis
 - a. use the largest sets of available information for estimating individual parameters
 - b. for example in a 2SLS analysis,
 - i. we might be missing data on the dependent variable in the second-stage equation
 - ii. we could run the first-stage equation without accounting for the patterns of missing data for this variable (the first-stage parameters can all be estimated without this variable)
 - iii. we would then run the second-stage equation on the restricted sample

- c. a more general example is a method-of-moments estimator
 - i. we could estimate each moment (means, variances, covariances) using the data available for that particular moment
 - ii. we would combine these moments in our final calculation of estimates
 - d. the available-case approach makes more use of the existing data
 - e. however, there are potential problems
 - i. estimates may be biased if the data are not MCAR
 - ii. there may be other computational problems, such as estimated variance/covariance matrices not being positive definite
 - iii. standard error calculations are usually very specialized and difficult
4. Unconditional mean imputation
- a. in this approach, we replace the missing values of \mathbf{X} with the unconditional means from the non-missing values
 - b. this sometimes produces reasonable results for simple descriptive statistics
 - c. however, it does not respect the associations between the different elements of \mathbf{X} , which can lead to problems in multivariate analyses
 - d. a modification for multivariate analyses is to use unconditional mean imputation but to also include the elements of \mathbf{M} as explanatory variables

E. Reweighting

1. Let's reconsider the complete-case approach
 - a. a problem with this approach is that it may be biased if the data are not MCAR
 - b. the reweighting approach uses complete cases but uses weights to adjust the sample to reflect the distribution of the analysis population
 - c. reweighting is especially attractive for cases of unit non-response
2. basic approach
 - a. form cells based on the X variables that are available for all respondents (i.e., don't suffer from any item non-response problems)
 - i. note the list of available variables may be thin
 - ii. variables from the sampling frame may be used
 - b. calculate the proportion of respondents with complete cases; this is an estimate of the probability of response
 - c. use the inverse of the response rate as a weight
 - d. instead of a cell-based approach, could estimate a binary choice model of response and use predictions as the response probabilities
3. advantages
 - a. approach is relatively easy to implement
 - b. potentially reduces bias
 - c. does not require a model of the joint distribution of the data
4. disadvantages
 - a. does not address biases from variables that do not

- appear in the weight calculations
- b. can have very high variance
- c. some weights can be unreasonably high, especially if the data were weighted to begin with; informal “trimming” procedures are sometimes needed
- d. can be difficult to calculate some other statistics

F. Overview of more formal imputation procedures

1. Many survey data sets, especially survey data sets prepared by the government, use *imputation* procedures when responses are missing (these are also sometimes referred to as *allocation* procedures)
 - a. imputation refers to assigning a response when none is given
 - b. alternative is to simply identify a response as missing (i.e., leave the solution to the user)
 - c. how do you use the available information to impute a response?
2. Several advantages of imputation
 - a. recreates the original rectangular data set
 - b. when carried out by the researchers who created the data, imputation can use their special knowledge of the data; it may also use variables that are masked from public release versions
 - c. another advantage when carried out by the original researchers is that it produces a data set that is consistent for different users
3. Another use for imputation is to assign values for variables that were not originally included in a survey,

- i.e., to add variables to the data set
 - a. example, Census Bureau's experimental poverty measures
 - i. primary data set is Current Population Survey
 - ii. key variables for the new poverty measures such as work and child care expenses are not recorded in the CPS
 - iii. use imputation to add these measures to the CPS
 - b. example, adding price and local employment information into cross-section data sets
- 4. Basic principles
 - a. imputations should be based on predictive distributions of the missing values given the observed values, preferably using as much of the observed data as possible
 - b. imputations should also preserve as many parts of the distributions of the missing data as possible
- 5. Conditional mean imputation
 - a. replace the missing values of X with the conditional means from the non-missing values
 - i. means would be conditioned on a small number of values
 - ii. example, in the Three-City Study, missing data on earnings is conditioned on work status; use the observed mean for working respondents for workers with missing earnings
 - b. more generally, divide data set into cells based on the other observable variables and then use the means of the observed data within those cells to impute
 - c. this procedure is nearly as easy to use as unconditional

mean imputation, but it uses some of the predictive power from the available data

d. disadvantages

- i. uses only a small slice of the available data
- ii. does not preserve other parts of the distribution

6. Simple model-based approaches (explicit models)

a. description

- i. estimate a regression (or other) model using the observed data; obtain coefficients
- ii. combine coefficients with available information on the other variables to predict values of the missing variable

b. critique

- i. straightforward to implement; also easy to determine the statistical properties of imputations
- ii. may be sensitive to specification issues (especially if the outcome variable is discrete or limited)
- iii. simple predictions do not capture the full distribution (only have the variation associated with the predictors)

c. stochastic regression

- i. follow the same procedure as above, but also include a pseudo-random error with the same variance as the standard error in the regression
- ii. the variance of the imputations will follow that of the original data more closely
- iii. want to be careful to provide specific seeds to the pseudo-random functions so that results can be exactly reproduced

7. Random (categorical) matching (implicit model)

- a. description
 - i. randomly draw an outcome from observations that share characteristics with the problem observation
 - ii. use this draw as an imputation
 - b. critique
 - i. usually preserves the characteristics of the marginal distribution (e.g., the unconditional mean and variance)
 - ii. can view this as a non-parametric procedure with weaker assumptions than the explicit modeling approach
 - iii. harder to capture conditional distributions (may align well for matching variables but not for other variables)
 - iv. which characteristics should be used to form the match? how should they be used?
 - c. example, hot deck procedures in CPS and other Census Bureau data sets
 - i. procedure tries to group observations into cells with many criteria
 - ii. if no matches are found, the criteria are broadened
 - iii. Lillard, Smith & Welch (1986) examined hot-deck procedures in CPS data for wages
 - iv. more recently, Bollinger & Hirsch (2006) found that CPS imputed earnings data did a poor job in analyses involving variables, like union status, that were not included in the imputation
8. Approximate matching
- a. problem with the exact matching approach is that you may have cells with no matches; this can happen if

- i. \mathbf{X} is continuous, or
- ii. \mathbf{X} has many dimensions
- b. may instead have to form close matches
- c. categorical matching
 - i. could construct categories based on the \mathbf{X} variables
 - ii. hard or impossible to find exact matches
 - iii. somewhat arbitrary, but may be important and useful if there is some strong conceptual or empirical basis for stratifying
 - iv. still limited
- d. simple distance metrics
 - i. if there is only one X variable or one continuous variable after stratifying, could look at a simple distance metric
 - ii. problem is harder if there are multiple variables
 - iii. for multiple variables can use the Mahalanobis metric

$$d = (\mathbf{X}_A - \mathbf{X}_B)' \text{Cov}(\mathbf{X}_B)^{-1} (\mathbf{X}_A - \mathbf{X}_B)$$

9. Multiple Imputation

- a. problem with all of the imputation procedures listed above is that they understate the uncertainty associated with a given imputation
- b. idea behind multiple imputation is to impute several values and then reweight the data
- c. procedure for multiple imputation, PROC MI, is available in SAS¹

¹ See http://support.sas.com/documentation/cdl/en/statug/59654/HTML/default/mi_toc.htm.

References:

Bollinger, Christopher, and Barry T. Hirsch. "Match Bias from Earnings Imputation in the Current Population Survey: The Case of Imperfect Matching." *Journal of Labor Economics* 24:3 (July 2006), 483-519.

Lillard, Lee, James P. Smith and Finis Welch. "What Do We Really Know About Wages: The Importance of Nonreporting and Census Imputation." *Journal of Political Economy* 94:3 part 1 (June 1986), 489-506.

Little, Roderick J., and Nathaniel Schenker. "Missing Data." In G. Arminger, C. Clogg and M. Sobel (eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. New York: Plenum, 1995.

Schafer, Joseph L., and John W. Graham. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7:2 (June 2002), 147-177.