INFO 7470/ILRLE 7400
Statistical Tools:
Missing Data Methods

John M. Abowd and Lars Vilhuber

March 15, 2011

Outline

- Missing data overview
- Missing records
  - Frame or census
  - Survey
- Missing items
- Overview of different products
- Overview of methods
- Formal multiple imputation methods

Missing Data Overview

- Missing data are a constant feature of both sampling frames (derived from censuses) and surveys
- Two important types are distinguished
  - Missing record (frame) or interview (survey)
  - Missing item (in either context)
- Methods differ depending upon type

Missing Records: Frame or Census

- The problem of missing records in a census or sampling frame is detection
- By definition in these contexts the problem requires external information to solve

Census of Population and Housing

- Dress rehearsal Census
- Pre-census housing list review
- Census processing of housing units found on a block not present on the initial list
- Post-census evaluation survey
- Post-census coverage studies

Economic Censuses and the Business Register

- Discussed in lecture 2
- Start with tax records
- Unduplication in the Business Register
- Weekly updates
- Multi-units updated with Company Organization Survey
- Multi-units discovered during the inter-censal surveys are added to the BR

Missing Records: Survey

- Non-response in a survey is normally handled within the sample design
- Follow-up (up to a limit) to obtain interview/data
- Assessment of non-response within sample strata
- Adjustment of design weights to reflect non-responses

Missing Items

- Imputation based on the other data in the interview/case (relational imputation)
- Imputation based on related information on the same respondent (longitudinal imputation)
- Imputation based on statistical modeling
  - Hot deck
  - Cold deck
  - Multiple imputation

Census 2000 PUMS Missing Data

- *Pre-edit:* When the original entry was rejected because it fell outside the range of acceptable values.
- *Consistency:* Imputed missing characteristics based on other information recorded for the person or housing unit.
- *Hot Deck:* Supplied the missing information from the record of another person or housing unit.
- *Cold Deck:* Supplied missing information from a predetermined distribution.

CPS Missing Data

- *Relational imputation:* use other information in the record to infer value
- *Longitudinal edits:* use values from the previous month if present in sample
- *Hot deck:* use values from actual respondents whose data are complete for the, relatively few, conditioning variables

County Business Patterns

- The County and Zipcode Business Patterns data are published from the Employer Business Register
- This is important because variables used in these publications are edited to publication standards
- The primary imputation method is a longitudinal edit
- http://www.census.gov/epcd/cbp/view/cbpmethodology.htm

Economic Censuses

- Like demographic products, there are usually both edited and unedited versions of the publication variables in these files
- Publication variables (*e.g.*, payroll, employment, sales, geography, ownership) have been edited
- Most recent files include allocation flags to indicate that a publication variable has been edited or imputed
- Many historical files include variables that have been edited or imputed but do not include the flags

QWI Missing Data Procedures

- Individual data
  - Multiple imputation
- Employer data
  - Relational edit
  - Bi-directional longitudinal edit
  - Single-value imputation
- Job data
  - Use multiple imputation of individual data
  - Multiple imputation of place of work
    - Use data for each place of work

BLS National Longitudinal Surveys

- Non-responses to the first wave never enter the data
- Non-responses to subsequent waves are coded as “interview missing”
- Respondent are not dropped for missing an interview. Special procedures are used to fill critical items from missed interviews when the respondent is interviewed again
- Item non-response is coded as such

Federal Reserve Survey of Consumer Finances (SCF)

- General information on the Survey of Consumer Finances: http://www.federalreserve.gov/pubs/oss/oss2/scfindex.html
- Missing data and confidentiality protection are handled with the same multiple imputation procedure

SCF Details

- Survey collects detailed wealth information from an over-sample of wealthy households
- Item refusals and item non-response are rampant (see Kennickell article)
- When there is item refusal, interview instrument attempts to get an interval
- The reported interval is used in the missing data imputation
- When the response is deemed sensitive for confidentiality protection, the response is treated as an item missing (using the same interval model as above)
- First major survey released with multiple imputation.

Relational Imputation

- Uses information from the same respondent
- *Example:* respondent provided age but not birth date. Use age to impute birth date.
- *Example:* some members of household have missing race/ethnicity data. Use other members of same household to impute race/ethnicity

Longitudinal Imputation

- Look at the respondent's history in the data to get the value
- *Example:* respondent's employment information missing this month. Impute employment information from previous month
- *Example:* establishment industry code missing this quarter. Impute industry code from most recently reported code

Cross Walks and Other Imputations

- In business data, converting an activity code (*e.g.*, SIC) to a different activity code (*e.g.*, NAICS) is a form of missing data
- In general, the two activity codes are not done simultaneously for the same entity
- Often these imputations are treated as 1-1 when they are, in fact, many-to-many

Probabilistic Methods for Cross Walks

- Inputs:
  - original codes
  - new codes
  - information for computing
    - Pr(new code | original code, other data)
- Processing
  - Randomly assign a new code from the appropriate conditional distribution
- See Lab 8

The Theory of Missing Data Models

- General principles
- Missing at random
- Weighting procedures
- Imputation procedures
- Hot decks
- Introduction to model-based procedures

General Principles

- Most of today's lecture is taken from *Statistical Analysis with Missing Data*, 2nd edition, Roderick J. A. Little and Donald B. Rubin (New York: John Wiley & Sons, 2002).
- The basic insight is that missing data should be modeled using the same probability and statistical tools that are the basis of all data analysis.
- Missing data are not an anomaly to be swept under the carpet.
- They are an integral part of very analysis.

Missing Data Patterns

- Univariate non-response
- Multivariate non-response
- Monotone
- General
- File matching
- Latent factors, Bayesian parameters

Missing Data Mechanisms

- The complete data are defined as the matrix Y ($n \times K$).
- The pattern of missing data is summarized by a matrix of indicator variables M ($n \times K$).
- The data generating mechanism is summarized by the joint distribution of Y and M.

$$m_{ij} = \begin{cases} 0, \text{if } y_{ij} \text{ is observed} \\ 1, \text{if } y_{ij} \text{ is missing} \end{cases}$$

$$p\left(Y, M \middle| \theta, \phi\right)$$

Missing Completely at Random

- In this case the missing data mechanism does not depend upon the data *Y*.
- This case is called MCAR.

$$p(M|Y,\theta,\phi) = p(M|\phi)$$

Missing at Random

- Partition Y into observed and unobserved parts.
- Missing at random means that the distribution of M depends only on the observed parts of Y.
- Called MAR.

$$Y = \left(Y_{\text{obs}}, Y_{\text{mis}}\right)$$

$$p(M|Y,\theta,\phi) = p(M|Y_{\text{obs}},\phi)$$

Not Missing at Random

- If the condition for MAR fails, then we say that the data are not missing at random, NMAR.
- Censoring and more elaborate behavioral models often fall into this category.

The Rubin and Little Taxonomy

- Analysis of the complete records only
- Weighting procedures
- Imputation-based procedures
- Model-based procedures

Analysis of Complete Records Only

- Assumes that the data are MCAR.
- Only appropriate for small amounts of missing data.
- Used to be common in economics, less so in sociology.
- Now very rare.

Weighting Procedures

- Modify the design weights to correct for missing records.
- Provide an item weight (*e.g.*, earnings and income weights in the CPS) that corrects for missing data on that variable. See Bollinger and Hirsch discussion later in lecture.
- See complete case and weighted complete case discussion in Rubin and Little.

Imputation-based Procedures

- Missing values are filled-in and the resulting "Completed" data are analyzed
  - Hot deck
  - Mean imputation
  - Regression imputation
- Some imputation procedures (*e.g.*, Rubin's multiple imputation) are really model-based procedures.

Imputation Based on Statistical Modeling

- Hot deck: use the data from related cases in the same survey to impute missing items (usually as a group)
- Cold deck: use a fixed probability model to impute the missing items
- Multiple imputation: use the posterior predictive distribution of the missing item, given all the other items, to impute the missing data

Current Population Survey

- Census Bureau Imputation Procedures:
  - Relational Imputation
  - Longitudinal Edit
  - Hot Deck Allocation Procedure

“Hot Deck” Allocation

- Labor Force Status
  - Employed
  - Unemployed
  - Not in the Labor Force

(Thanks to Warren Brown)

“Hot Deck” Allocation

	Black	Non-Black
Male		
16 – 24		
25+		ID #0062
Female		
16-24		
25+		

“Hot Deck” Allocation

	Black	Non-Black
Male		
16 – 24	ID #3502	ID #1241
25+	ID #8177	ID #0062
Female		
16-24	ID #9923	ID #5923
25+	ID #4396	ID #2271

CPS Example

- Effects of hot-deck imputation of labor force status.

Public Use Statistics

	Total AXLFSR	No change	Allocated
Total A_LFSR	220,284,576		
Working	131,704,236		
W/job,not at work	4,572,653		
Unemp,looking for work	7,967,976		
Unemp,on layoff	1,371,469		
Not in labor force	74,668,242		
Total A_AGE	220,284,576		
Average A_AGE	44.1		
Std Err A_AGE	0.15		
Total A_SEX	220,284,576		
Male	105,972,746		
Female	114,311,831		

Allocated v. Unallocated

	Total AXLFSR	No change	Allocated
Total A_LFSR	220,284,576	219,529,643	754,933
Working	131,704,236	131,294,888	409,348
W/job,not at work	4,572,653	4,564,589	8,063
Unemp,looking for work	7,967,976	7,919,562	48,414
Unemp,on layoff	1,371,469	1,367,766	3,703
Not in labor force	74,668,242	74,382,838	285,405
Total A_AGE	220,284,576	219,529,643	754,933
Average A_AGE	44.1	44.2	35.2
Std Err A_AGE	0.15	0.15	1.96
Total A_SEX	220,284,576	219,529,643	754,933
Male	105,972,746	105,603,454	369,292
Female	114,311,831	113,926,189	385,641

Bollinger and Hirsch CPS Missing Data

- Studies the particular assumptions in the CPS hot deck imputer on wage regressions
- Census Bureau uses too few variables in its hot deck model
- Inclusion of additional variables improves the accuracy of the missing data models
- See Bollinger and Hirsch

Model-based Procedures

- A probability model based on p(Y, M) forms the basis for the analysis.
- This probability model is used as the basis for estimation of parameters or effects of interest.
- Some general-purpose model-based procedures are designed to be combined with likelihood functions that are not specified in advance.

Little and Rubin's Principles

- Imputations should be
  - Conditioned on observed variables
  - Multivariate
  - Draws from a predictive distribution
- Single imputation methods do not provide a means to correct standard errors for estimation error.

Rest of lecture (PDF format)

Applications to Complicated Data

- Computational formulas for MI data
- Examples of building Multiply-imputed data files

Computational Formulas

- Assume that you want to estimate something as a function of the data Q(*Y*)
- Formulas account for missing data contribution to variance

$$Q_m\left(Y^m\right) = \text{estimand from the m}^{\text{th}}\text{ implicate}$$

$$\overline{Q} = \sum_{m=1}^{M} Q_m\left(Y^m\right) \Big/ M$$

$$\overline{Q} = \text{average estimand}$$

$$V_m\left(Y^m\right) = \text{covariance matrix of } Q_m\left(Y^m\right) \text{from the m}^{\text{th}}\text{ implicate}$$

$$\overline{V} = \sum_{m=1}^{M} V_m\left(Y^m\right) \Big/ M$$

$$\overline{V} = \text{average covariance matrix}$$

$$B = \left[\sum_{m=1}^{M} \left(Q_m\left(Y^m\right) - \overline{Q}\right)\left(Q_m\left(Y^m\right) - \overline{Q}\right)^T\right] \Big/ M$$

$$B = \text{between implicate variation of } Q_m\left(Y^m\right)$$

$$T = \overline{V} + \left(1 + \frac{1}{M}\right) B$$

$$T = \text{total variance matrix of } Q(Y)$$

Examples

- Survey of Consumer Finances
- Quarterly Workforce Indicators

Survey of Consumer Finances

- Codebook description of missing data procedures

How are the QWIs Built?

- *Raw input files:*
  - UI wage records
  - QCEW/ES-202 EQUI report
  - Census Numident/Personal Characteristics File
  - Census Place of Residence
  - LEHD geo-coding system
- *Processed data files:*
  - Individual characteristics
  - Employer characteristics
  - Employment history with earnings

Flow Chart

Processing the Input Files

- Each quarter the complete history of every individual, every establishment, and every job is processed through the production system
- Missing data on the individual and employment history records are multiply imputed
- Missing data on the employer characteristics are singly-imputed (explanation to follow)

Garden Variety Problems

- Missing demographic data on the individual file (birth date, sex, education, place of residence)
  - Multiple imputations using information from the individual, establishment, and employment history files
  - Model estimation component updated every quarter
- This process was used to create the current snapshot (S2004/S2008) but not the current public-use data (updated individual data imputation procedure uses much more information)

The Mother of all Missing Data Problems

- The employment history records only code employer to the UI account level
- Establishment characteristics (industry, geo-codes) are missing for multi-unit establishments
- The establishment (within UI account) is multiply imputed using a dynamic multi-stage probability model
- Estimation of the posterior predictive distribution depends on the existence of a state with establishments coded on the UI wage record (MN)

Can It Be Done?

- Every quarter the QWI processes over 6 billion employment histories (unique person-employer pair) covering 1990 to 2010
- Approximately 30% of these histories require multiple employer imputations
- So, the system does more than 25 billion full information imputations every quarter
- The information used for the imputations is current, it includes all of the historical information for the person and every establishment associated with that person's UI account

Does It Work?

- Full assessment using the state that codes both
- Summary slide follows

MN Known Unit vs. MN Imputed Unit Weighted