Analyzing Frames and Samples with Missing Data

John M. Abowd
March 2007
Outline

• Missing data overview
• Missing records
  – Frame or census
  – Survey
• Missing items
• Overview of different products
• Overview of methods

© John M. Abowd 2007, all rights reserved
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 last week’s lecture
• Start with tax records
• Unduplication in the Business Register
• Weekly updates
• Multiunits updated with Company Organization Survey
• Multiunits discovered during the inter-censal surveys are added to the BR
Missing Records: Survey

• Nonresponse 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 nonresponse
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

• a. Pre-edit. When the original entry was rejected because it fell outside the range of acceptable values.

• b. Consistency. Imputed missing characteristics based on other information recorded for the person or housing unit.

• c. Hot Deck. Supplied the missing information from the record of another person or housing unit.

• d. 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
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](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:  

• 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
    \[ \text{Pr(new code | original code, other data)} \]

• Processing
  – Randomly assign a new code from the appropriate conditional distribution

• See Lab 7