Multiple Imputation II

John M. Abowd
March 2007

Outline

- 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

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

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

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

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

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

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

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

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

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

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

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
  - ES-202 EQUI report
  - Census Numident
  - 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

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 2.5 billion employment histories (unique person-employer pair) covering 1990 to 2006

- Approximately 30% of these histories require multiple employer imputations

- So, the system does more than 10 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

Legend:
- ☐ Earnings, Full-Quarter Accessions
- ☐ Beginning-of-Period Employment
- ☐ End-of-Period Employment
- ☐ Full-Quarter Employment
- ☐ Separations
- ☐ Accessions

Percent Discrepancy (vertical axis): 20%, 15%, 10%, 5%, 0%, -5%, -10%, -15%

Horizontal axis: 10th Percentile, Median, 90th Percentile