Blending Data Through Statistical Matching, Modeling, and Imputation

Jerry Reiter
Department of Statistical Science, Duke University
and Census Bureau

Acknowledgments

- Research ideas in this talk supported by
  - National Science Foundation
    - SES 1131897, SES 1733835

- Any views expressed are those of the author and not necessarily of NSF or the Census Bureau

Goal of Presentation

- Outline some general methods for blending data
  - Statistical matching (also known as data fusion)
  - Imputation strategies with auxiliary data
- Present my opinions on challenges and opportunities for different methods
- No technical details, no record linkage (thanks Beka!)
- Ignore privacy concerns for time; not intended to minimize their importance

Statistical matching

- Instructive to work with a two file setting
  - File A has variables X and Y
  - File B has variables X and Z
  - Files have disjoint sets of records, so that Y and Z are never observed simultaneously

- Goal is to learn about associations between Y and Z, possibly given elements in X

Fundamental problem

- We cannot estimate the joint distribution of (Y, Z) from the data alone
- Need some form of external information to proceed with statistical matching
  - Assumptions about association between Y and Z given X
  - Another dataset with Y and Z (and ideally X) observed simultaneously
  - Constraints on associations from other sources

Assumptions in statistical matching

- Most common assumption is conditional independence: Y is independent of Z given X
- Typical methods used for statistical matching implicitly assume this, including
  - Nearest neighbor hot deck: for each record in File B, find record in File A with most similar value of x, and use its observed y as an imputation for the missing Y
  - Regression modeling: estimate a model that predicts Y from X, and use it to impute the missing values of Y in File B
  - Joint modeling: use a flexible joint distribution to the data, such as a mixture model, to impute missing items

Nearest Neighbor Hot Deck: Pros, Cons, and Quality Concerns

- Pros
  - Easy to explain to others
  - Hot deck familiar to statistical agencies
  - Can generate realistic multivariate imputations
- Cons
  - Conditional independence is a strong assumption that is difficult to evaluate– if not true, matching could be unreliable
  - Have to select distance function and subset of X, which can be tricky with many X of different types and multivariate (Y, Z)
  - Single imputation underestimates uncertainty
  - Can cause difficulties with edits

Pros, Cons, and Quality Concerns

- Quality concerns
  - Are X variables defined similarly?
  - Are data files contemporaneous?
  - What should we do with complex designs?
    - Concatenate files and re-weight so that the concatenated file represents some target population?
    - Use only one file for analysis/dissemination?
  - How to propagate uncertainty?
    - Multiple imputation? (May be challenging with hot deck and rich X)
  - How to do sensitivity analysis?
    - Alternative matching algorithms or distances?

Facilitating sensitivity analysis with regression modeling approaches

- Regression approach can be viewed as specifying a model for Y, such as

$$Y = X\beta + Z\alpha + \varepsilon$$

- With conditional independence, we set $\alpha = 0$.
- For sensitivity tests, could choose other values of α, for example, by fixing the partial correlation of $(Y, Z \mid X)$
- Generate imputed Ys under such multiple plausible models, and assess sensitivity of results

Pros, Cons, and Quality Concerns

- Pros
  - Regression modeling more flexible than hot deck, e.g., use predictive engines from machine learning
  - Can specify models so that imputations satisfy edits
  - Can check quality of regression model
  - Prescriptive and flexible approach to sensitivity analysis
  - Naturally leads to multiple imputation for uncertainty propagation (given value of α)
- Cons
  - Still have to make unverifiable assumptions about α
  - Have to select model
- Many of the same quality concerns as with hot deck

Auxiliary Data with Y, Z Observed

- Subsets of Y and Z may be observed simultaneously, along with a subset of X, in other data files
- Use that information to reduce reliance on conditional independence (or other unverifiable) assumptions
  - All variables in (Y, Z) observed for all variables in X
  - Arbitrary subsets observed in one file
  - Multiple subsets observed across different files

First case: All observed

- Regress Y on (X, Z), and use model to (multiply) impute missing Y in File B, likewise for Z in File A
- Overarching quality concern
  - Conditional distribution in auxiliary data must be valid in File B
  - Similar time periods, populations, sampling designs (account for differences if possible)
  - Specify good fitting model in auxiliary data
- This concern holds for other cases to follow

Second case: One auxiliary file

- Only some variables in (Y, Z) observed jointly, possibly with some variables in X

- For some multivariate distributions, possible to estimate subsets of parameters and fix remainder
  - Multivariate normal: use auxiliary data to estimate elements covariance matrix, and fix others at feasible values

Second case: One auxiliary file

- General strategy for arbitrary joint models
  - Append auxiliary data to File B, and estimate joint model using the incomplete data
  - Construct appended data so as not to distort the marginal distributions of (X, Y) and (X, Z)
  - See Fosdick, De Yoreo, and Reiter (2016, *Annals of Applied Statistics*) for an example of this approach

Third case: Multiple auxiliary files

- Pieces of the joint distribution of (X, Y, Z) available in multiple datasets
- Again, for specific joint models like MVN it is straightforward to estimate parameters corresponding to the known marginal and conditionals
- For arbitrary joint distributions, conceptually one could use the augmented cases approach
  - This has not been tried, at least to my knowledge

Pros, Cons, and Quality Concerns

- Pros
  - Use of auxiliary information reduces reliance on unverifiable assumptions
  - Can specify models so that imputations satisfy edits
  - Can check quality of auxiliary data models for predicting marginal distributions of observed variables
  - Naturally leads to multiple imputation for uncertainty propagation
- Cons
  - Still have to choose model and make some unverifiable assumptions about not observed marginal and conditionals
  - Have to be careful how one constructs auxiliary data, especially when using joint models
  - Can be difficult to do sensitivity analysis with flexible joint models

Thoughts on what to report

- Agencies performing statistical matching should be transparent about
  - Meta-data for files used in the matching
  - Steps taken to harmonize X variables and other edits
  - Assumptions and models used in matching
  - Assessments of quality of fit of regression models
  - Results of sensitivity analyses
- In addition to above, agencies using auxiliary data should be transparent about
  - Potential selection biases in auxiliary data
  - Specification of conditional distributions in auxiliary data
  - Combinations of variables that were not observed jointly

Thoughts on research directions

- How useful are convenient, non-representative auxiliary data?
  - Fosdick et a. (2016) use data from CivicScience, a rapid response internet polling company, to get simultaneous measurements of Y, Z in a marketing data fusion
  - Data clearly not representative jointly (more older people in CivicScience data than in surveys to be fused) but perhaps reasonable to assume $Y \mid X, Z$ is valid in CivicScience data

- How do we implement the "piecewise" conditional distribution approach? How do we inform users what they can expect to estimate well and what they cannot for their specific queries?

- How do we propagate uncertainty in this context?
  - Initial simulation studies suggest existing multiple imputation combining rules are not quite right