Bayesian Data Editing for Continuous Microdata

Alan F. Karr
RTI International

(with L. H. Cox, H. Kim, J. P. Reiter, Q. Wang)

ITSEW 2014, Washington
October 3, 2014
Problem Statement

**Setting**  Numerical microdata that may be

- Missing
- Erroneous

**Dataset of Interest**  U.S. Census Bureau’s every-five-years Census of Manufactures (CM)

**Goal**  Simultaneously (and multiply) impute edit constraint-satisfying replacements for *both* missing values and erroneous values

**Impact**

- Improve data quality
- Reduce cost: editing is estimated to consume 20–40% of survey costs
**Notation**

- $i =$ subject
- $j =$ numerical attribute
- $X_i(j) =$ “true” value of attribute $j$ for subject $i$
- $Y_i(j) =$ reported value of attribute $j$ for subject $i$
- $S_i(j) =$ binary error indicator for attribute $j$ for subject $i$
  - *Conceptually,* $S_i(j) = 1(Y_i(j) \neq X_i(j))$
  - *Operationally,* $S_i(j) = 1$ means that a replacement will be imputed for $Y_i(j)$
Classes of Edit Constraints

**Range Constraints** \( L(j) \leq Y_i(j) \leq U(j) \)

**Ratio Constraints** \( Y_i(j) / Y_i(\ell) \leq \alpha_{j,\ell} \) (better as \( Y_i(j) \leq \alpha_{j,\ell} Y_i(\ell) \))

**Balance Constraints** \( Y_i(j_1) + Y_i(j_2) + \cdots + Y_i(j_\ell) = Y_i(j_m) \)

**Compatibility Constraints** (usually only for categorical data):
\( Y_i(j_1) = y_1 \) and \( Y_i(j_2) = y_2 \) are incompatible
Two Steps in Automated Data Editing

**Error Localization**  Determine (estimate) $S_i(j)$
- Multiple approaches, discussed momentarily

**Error Correction**  Determine (calculate) replacement values for those $Y_i(j)$ for which $S_i(j) = 1$
- Generally, some form of imputation
- Violations of balance edits sometimes resolved by definition (not always a good idea)
This Talk: Compare Three Methods

**Fellegi-Holt (FH)** *(JASA, 1976)*
- Error Localization: Use optimization algorithm to determine [weighted] minimum number of attributes to impute
- Error Correction: Historically, hot deck or . . . . In this talk, constraint-preserving imputation algorithm of Kim, et al. *(JBES, 2014, to appear)*

**Flag All Active Items (AAI)**
- Error Localization: Flag every $Y_i(j)$ that is involved in an edit violation
- Error Correction: Constraint-preserving imputation algorithm of Kim, et al.

**Bayesian Editing (BE)** Integrate localization and correction
What’s Wrong with Fellegi-Holt

1. Have to enumerate all implied constraints (otherwise can’t be sure that minimization has been achieved)

2. Diagram showing edit-passing and edit-failing reported values.
Structure of the BE Model

More Notation

- $\mathcal{X} =$ feasible region defined by range and ratio constraints
- $T =$ set of variables that are not “sums” in balance constraints
- $A_i \in \{0, 1, 2, 3\} =$ “nature of errors” indicator for subject $i$

Model for $\{X_i(j) : j \in T\}$ Mixed multivariate normal restricted to $\mathcal{X}$: parameters $K, \mu_k, \Sigma_k, \pi$

Model for $\pi$ Dirichlet process (stick-breaking representation)

Model for $\{X_i(j) : j \notin T\}$ Equal to sum of components
Model Structure—2

**Model for** $A_i \mid X_i$  May involve parameters $\psi$, but $f(a \mid x, \psi) \propto 1$

**Model for** $S_i \mid (X_i, A_i)$  May involve parameters $\psi$, but $f(s \mid x, a, \psi) \propto 1$

**Model for** $Y_i \mid (X_i, S_i)$  $E_i = \{ j : S_i(j) = 1 \}$ (erroneous components)

- $S_i(j) = 0 \Rightarrow Y_i(j) = X_i(j)$
- $Y_i(E_i)$ uniform on (subset of bounding hypercube) \ $\mathcal{X}$

**Model for Missingness**  At the moment, MAR

- $Y_i(j)$ missing $\Rightarrow S_i(j) = *$

**Priors**  The standard noninformative choices
BIG Inference Assumptions

**AAI and BE** \( Y_i \in \mathcal{X} \Rightarrow S_i = 0 \)

- Tempting interpretation: \( Y_i \in \mathcal{X} \Rightarrow X_i = Y_i \)
- Safer interpretation: If \( Y_i \in \mathcal{X} \), no basis for changing it

**AAI** \( Y_i(j) \) involved in an edit violation \( \Rightarrow S_i(j) = 1 \)
The MCMC

- Gibbs update for all but a few steps
- Data augmentation techniques to ease estimation of truncated normal distributions (O’Malley and Zaslavsky, JASA, 2008)
- Simultaneously draw imputed values $X$ and editing indicators $S$
  1. Propose $S^*$ from neighbors of current $S$ using birth-death process
  2. Generate $X^*$ given $S^*$ from constrained mixture of normals
  3. Accept/reject $(X^*, S^*)$ by Metropolis-Hastings
Structure

- 9 variables
  - Range constraints for every variable
  - Ratio constraints for some pairs of variables
  - Two balance constraints: \( X(4) = X(1) + X(2) + X(3) \) and \( X(7) = X(5) + X(6) \)
- \( n = 2000 \) error-free values of 
  \[
  \left( X_i(1), X_i(2), X_i(3), X_i(5), X_i(6), X_i(8), X_i(9) \right)
  \]
  from mixture of normals; calculate \( X_i(4) \) and \( X_i(7) \) from balance constraints
- For 1000 out of 2000 records, introduce edit-failing records using model (so no mis-specification)
- 5% missingness, CAR
- 500 simulations
Pictorial Results: Data
Pictorial Results: Correlations
### Numerical Results: 95% CI Coverage for Population Means

<table>
<thead>
<tr>
<th>Variable</th>
<th>True X</th>
<th>E-P X</th>
<th>True S</th>
<th>FH</th>
<th>AAI</th>
<th>BE</th>
</tr>
</thead>
<tbody>
<tr>
<td>X (1)</td>
<td>95.2</td>
<td>95.4</td>
<td>96.2</td>
<td>90.0</td>
<td>96.2</td>
<td>95.8</td>
</tr>
<tr>
<td>X (2)</td>
<td>93.0</td>
<td>95.4</td>
<td>95.6</td>
<td>6.4</td>
<td>97.0</td>
<td>95.4</td>
</tr>
<tr>
<td>X (3)</td>
<td>94.4</td>
<td>95.6</td>
<td>94.0</td>
<td>95.2</td>
<td>97.6</td>
<td>96.2</td>
</tr>
<tr>
<td>X (4)</td>
<td>93.4</td>
<td>93.0</td>
<td>94.6</td>
<td>96.6</td>
<td>94.8</td>
<td>95.2</td>
</tr>
<tr>
<td>X (5)</td>
<td>93.8</td>
<td>94.0</td>
<td>94.4</td>
<td>0.0</td>
<td>93.4</td>
<td>92.4</td>
</tr>
<tr>
<td>X (6)</td>
<td>94.8</td>
<td>94.2</td>
<td>93.8</td>
<td>0.8</td>
<td>97.8</td>
<td>93.0</td>
</tr>
<tr>
<td>X (7)</td>
<td>94.8</td>
<td>94.4</td>
<td>94.2</td>
<td>10.8</td>
<td>94.4</td>
<td>92.2</td>
</tr>
<tr>
<td>X (8)</td>
<td>95.0</td>
<td>95.6</td>
<td>94.6</td>
<td>96.6</td>
<td>95.8</td>
<td>93.8</td>
</tr>
<tr>
<td>X (9)</td>
<td>95.6</td>
<td>92.2</td>
<td>96.4</td>
<td>67.0</td>
<td>94.0</td>
<td>95.4</td>
</tr>
</tbody>
</table>
Numerical Results: Relative Bias for Regression Coefficients

Model $X_i(9) = \beta(0) + \beta(1)X_i(1) + \beta(5)X_i(5) + \beta(9)X_i(9) + \epsilon_i$

<table>
<thead>
<tr>
<th>Variable</th>
<th>True $X$</th>
<th>E-P $X$</th>
<th>True $S$</th>
<th>FH</th>
<th>AAI</th>
<th>BE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta(0)$</td>
<td>0.2</td>
<td>0.1</td>
<td>0.3</td>
<td>-2.6</td>
<td>-1.8</td>
<td>0.9</td>
</tr>
<tr>
<td>$\beta(1)$</td>
<td>-0.8</td>
<td>-1.6</td>
<td>-0.3</td>
<td>51.7</td>
<td>10.3</td>
<td>-2.9</td>
</tr>
<tr>
<td>$\beta(5)$</td>
<td>0.0</td>
<td>0.4</td>
<td>0.3</td>
<td>-41.6</td>
<td>-3.3</td>
<td>1.7</td>
</tr>
<tr>
<td>$\beta(9)$</td>
<td>0.2</td>
<td>0.5</td>
<td>-0.3</td>
<td>-0.4</td>
<td>-2.2</td>
<td>-0.4</td>
</tr>
</tbody>
</table>

Relative Bias $= \frac{1}{|Q|} \left( \frac{1}{R} \sum_{r=1}^{R} \hat{Q}_r - Q \right)$
Basics

- Part of Economic Census (most recent data: 2007)
- Example attributes: (logs of) cost of materials, total employment, total value of shipments, ... (so linear regressions are Cobb-Douglas production functions)
- Industry-specific ratio and balance constraints
- Current method: combination of manual and FH + hot deck (SPEER), labeled FC (Final Census)

**Our Study** One NAICS code, 1869 establishments, 27 variables, Title 13-protected (so worked in RDC)
Pictorial Results: Correlations
## AAI or BE?

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Winner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specification of constraints</td>
<td>Tie</td>
</tr>
<tr>
<td>Intellectual appeal</td>
<td>BE: borrows more strength</td>
</tr>
<tr>
<td>“Right” amount of imputation</td>
<td>BE</td>
</tr>
<tr>
<td>Incorporate domain knowledge of errors</td>
<td>BE: prior on $S$</td>
</tr>
<tr>
<td>Estimated distribution of $S$</td>
<td>BE: posterior distribution</td>
</tr>
<tr>
<td>Bayes “shock factor”</td>
<td>AAI</td>
</tr>
<tr>
<td>Computational burden</td>
<td>AAI: $10\times$ speed</td>
</tr>
<tr>
<td>Information about measurement error</td>
<td>Neither</td>
</tr>
</tbody>
</table>
Unresolved Issues: Specific

1. What are the effects of model mis-specification?
2. What are the tradeoffs between record-level correctness and inferential correctness?
3. Should the same imputation model be used for both missing and erroneous data?
4. What about weights?
Unresolved Issues: Broad

1. What if administrative data are available?
2. Do we need a taxonomy for erroneousness: erroneous completely at random, at random, non-ignorably?
3. What difference would it make to have a (good) measurement error model?
4. Can we integrate edit, imputation and disclosure limitation?
Acknowledgements and More Information

**Support**  NSF grant SES–1131897


**Contact Information**  karr@rti.org