# 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*, $S_i(j) = \mathbf{1}\big(Y_i(j) \neq X_i(j)\big)$
  - *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)

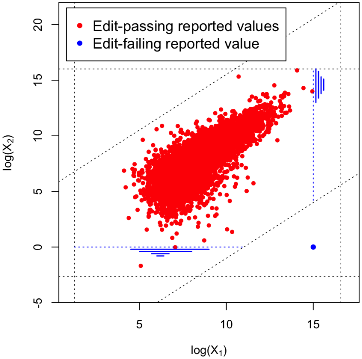
**Flag All Active Items (AAI)**

- 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. 



# 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 $A_i|X_i$** May involve parameters $\psi$, but $f(a|x,\psi) \propto 1$

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

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

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

**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

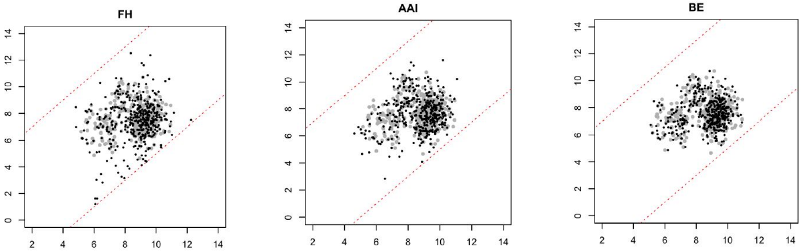
# Structure

- 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

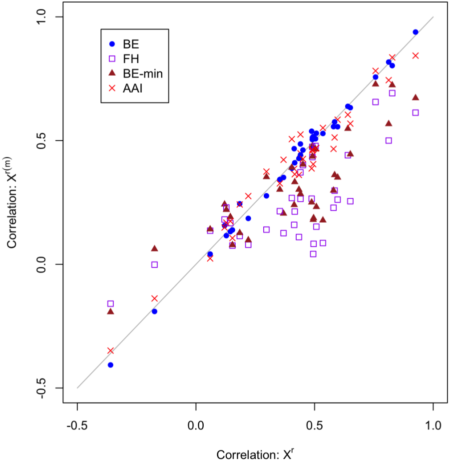
$$\big(X_i(1), X_i(2), X_i(3), X_i(5), X_i(6), X_i(8), X_i(9)\big)$$

  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

| Variable | True $X$ | E-P $X$ | True $S$ | FH | AAI | BE |
|---|---|---|---|---|---|---|
| $X(1)$ | 95.2 | 95.4 | 96.2 | 90.0 | 96.2 | 95.8 |
| $X(2)$ | 93.0 | 95.4 | 95.6 | 6.4 | 97.0 | 95.4 |
| $X(3)$ | 94.4 | 95.6 | 94.0 | 95.2 | 97.6 | 96.2 |
| $X(4)$ | 93.4 | 93.0 | 94.6 | 96.6 | 94.8 | 95.2 |
| $X(5)$ | 93.8 | 94.0 | 94.4 | 0.0 | 93.4 | 92.4 |
| $X(6)$ | 94.8 | 94.2 | 93.8 | 0.8 | 97.8 | 93.0 |
| $X(7)$ | 94.8 | 94.4 | 94.2 | 10.8 | 94.4 | 92.2 |
| $X(8)$ | 95.0 | 95.6 | 94.6 | 96.6 | 95.8 | 93.8 |
| $X(9)$ | 95.6 | 92.2 | 96.4 | 67.0 | 94.0 | 95.4 |

# Numerical Results: Relative Bias for Regression Coefficients

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

| Variable | True $X$ | E-P $X$ | True $S$ | FH | AAI | BE |
|---|---|---|---|---|---|---|
| $\beta(0)$ | 0.2 | 0.1 | 0.3 | -2.6 | -1.8 | 0.9 |
| $\beta(1)$ | -0.8 | -1.6 | -0.3 | 51.7 | 10.3 | -2.9 |
| $\beta(5)$ | 0.0 | 0.4 | 0.3 | -41.6 | -3.3 | 1.7 |
| $\beta(9)$ | 0.2 | 0.5 | -0.3 | -0.4 | -2.2 | -0.4 |

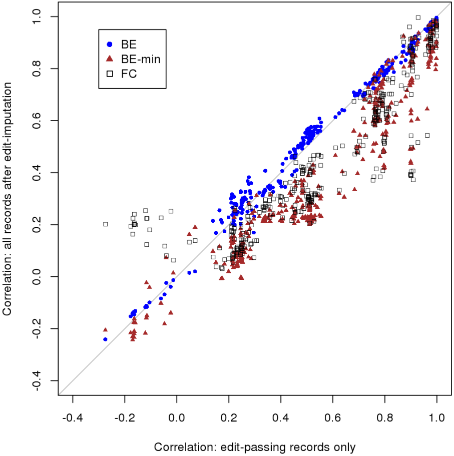
**Relative Bias =** $\frac{1}{|Q|}\left(\frac{1}{R}\sum_{r=1}^{R}\hat{Q}_r - Q\right)$

# 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?

| **Criterion** | **Winner** |
|---|---|
| Specification of constraints | Tie |
| Intellectual appeal | BE: borrows more strength |
| "Right" amount of imputation | BE |
| Incorporate domain knowledge of errors | BE: prior on $S$ |
| Estimated distribution of $S$ | BE: posterior distribution |
| Bayes "shock factor" | AAI |
| Computational burden | AAI: 10$\times$ speed |
| Information about measurement error | Neither |

# 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

**Technical Report** Kim, Cox, Karr, Reiter, Wang, "Simultaneous Edit-Imputation for Continuous Microdata," NISS Technical Report 189: http://www.niss.org/sites/default/files/tr189.pdf (submitted to *JASA*)

**Contact Information** karr@rti.org