Statistical Disclosure Limitation and Edit Imputation

Hang Kim, Alan Karr, Jerry Reiter

Department of Statistical Science, Duke University

Supported by the NSF NCRN grant to Duke/NISS: NSF-SES-11-31897

November 1, 2013

Outline of Talk

How should one integrate statistical disclosure limitation and edit-imputation?

- Background
  - Statistical disclosure limitation (SDL)
  - Editing and imputation
- Two broad strategies
  - Editing after SDL
  - Edit-preserving SDL
- Empirical illustration with manufacturing data

SDL Setting

- Agency seeks to disseminate microdata on individual records.
- We work with data that are all continuous, although similar issues apply when data include categorical variables.
- Exemplary SDL strategies for continuous data:
  - Noise addition
  - Microaggregation
  - Microaggregation followed by noise addition
  - Rank swapping
  - Synthetic data

Edit and Imputation Setting

- Values must satisfy certain logical constraints.
- Continuous data: constraints include range restrictions (e.g., $y_j > 0$) and ratio edits (e.g., $0 < y_j/y_k < 1000$).
- Typical process includes
  - identify records that fail the constraints,
  - select set of fields that could be changed to create a record that satisfies constraints,
  - change those fields in a way that satisfies constraints.
- First two talks of this session offer examples of this process.

SDL and Edit Imputation

- Some SDL processes can create edit rule violations.
- What should one do?
  - Ignore it, option 1: release data with violations. Not desirable.
  - Ignore it, option 2: delete records with violations. Bias inducing.
  - Run usual SDL first, fix up any violations that result by blanking and imputing.
  - Modify SDL procedure so that it automatically generates data that satisfy constraints.
- Discuss and illustrate these with empirical example.

Empirical Example: 1991 Columbia Manufacturing Survey

Variable	Label	Range restriction
Skilled labor	SL	0.9–400
Unskilled labor	UL	0.9–1,000
Wages paid to skill labor	SW	300–3,000,000
Wages paid to unskilled labor	UW	600–4,000,000
Real value added	VA	50–1,000,000
Real material used in products	MU	10–1,000,000
Capital	CP	5–1,000,000

- 6521 observations, 7 variables.
- Hypothetical, data-derived range restrictions.

Empirical Example: Hypothetical Ratio Edits

	V_2						
V_1	SL	UL	SW	UW	VA	MU	CP
SL	1	20	0.01	0.01	0.1	0.3	2
UL	50	1	0.1	0.005	0.3	5	5
SW	20000	100000	1	50	300	500	1000
UW	66666.7	10000	100	1	200	5000	5000
VA	10000	20000	10	10	1	200	700
MU	50000	100000	33.3	100	100	1	1000
CP	20000	10000	10	16.7	100	100	1

Data-derived ratio edits ($V_1/V_2 \leq b$) for the 1991 Colombia Manufacturing Survey.

Empirical Example: SDL then edit

- Mask number of skilled employees, number of unskilled employees, and capital. Leave the remaining variables unaltered.
- Don't worry about edit violations when doing SDL.
- Work with the natural logarithms of all variables.
- SDL techniques
  - Add noise from $N(0, c\Sigma)$, where $c = 0.16$.
  - Rank swapping separately for each variable with interval of 10%.
  - Microaggregation with 3 establishments per cluster based on principal components clustering.
  - Microaggregation followed by adding noise.
- Edits done by blanking all three variables and imputing using the mixture normal engine of Kim *et al.* (2013).

Empirical Example: Edit-preserving SDL

- Rank swapping and two noise addition methods: use rejection sampling approach (keep trying until you get dataset that satisfies constraints).
- Partially synthetic data generated by
  - Estimating joint distribution of all 7 variables using the mixture normal distribution of Kim *et al.* (2013).
  - Deriving conditional distributions from this model.
  - Imputing replacement values from the conditional distributions.
- These approaches guaranteed to generate values that satisfy all constraints.

Empirical Example: Measures of Risk

- We use the *percentage of linked* criterion (Domingo Ferrer *et al.* 2001).
- First, compute the distances

$$d_{i,j} = \sqrt{\sum_k (y_{ik} - \tilde{y}_{jk})^2}, \qquad \forall\, i,j = 1,\ldots,n,$$

where $k \in$(SL, UL, CP) and $\tilde{y}_{jk}$ is the perturbed version of y_{jk}.
- For each i, find the record j that achieves the minimum value of $d_{i,j}$.
- Let $t_i = 1$ when the index of i and j belong to the same record, i.e., the record in D^{rel} is linked correctly to D based on matching the available variables; let $t_i = 0$ otherwise.
- The risk measure is $PL = \sum_{i=1}^n t_i/n$.

Empirical Example: KL Measure of Utility

- Approximate Kullback-Leibler (KL) divergence of released data D^{rel} from original data D.
- Use a closed-form expression based on a normality assumption,

$$KL = \frac{1}{2}\left[\mathrm{tr}\left\{(\Sigma^{rel})^{-1}\Sigma\right\} + \left(\bar{y}^{rel} - \bar{y}\right)^T (\Sigma^{rel})^{-1}\left(\bar{y}^{rel} - \bar{y}\right) - p - \log\left(\frac{|\Sigma^{rel}|}{|\Sigma|}\right)\right],$$

- $\bar{y}$ and Σ are the sample mean and the sample covariance in D.
- $\bar{y}^{rel}$ and Σ^{rel} are the sample mean and the sample covariance in D^{rel}.

Empirical Example: Propensity Score Measure of Utility

- Propensity score (U) utility measure (Woo *et al.* 2009).
- Concatenate D^{rel} and D, and add an indicator variable whose values equal one for all records in D^{rel} and equal zero for all records in D.
- Use indicator variable as outcome in the logistic regression,

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{a=1}^{7} \beta_a \log Y_{ia} + \sum_{a,b} \log Y_{ia} \log Y_{ib} + \sum_{a,b,c} \beta_{abc} \log Y_{ia} \log Y_{ib} \log Y_{ic}.$$

- For $i = 1, \ldots, 2n$, compute the set of predicted probabilities $\hat{p}_i$.
- The risk measure is

$$\mathrm{U} = \frac{1}{2n} \sum_{i=1}^{2n} \left(\hat{p}_i - \frac{1}{2}\right)^2.$$

Empirical Example: SDL Causes Edit Violations

Numbers of records that violate edit rules across 20 replications after implementing perturbative SDL methods.

Methods	Mean (%)	SD
`Noise`	157.8 (2.45)	10.1
`Swap`	134.2 (2.09)	6.6
`Mic`	5.0 (0.08)	–
`MicN`	84.1 (1.31)	6.7

Empirical Example: Results

Measured data utility and disclosure risk. Entries include the averages of KL, U_{prop} and PL from 20 replications of each method.

	Approach	Noise	Swap	Mic	MicN	Synt
KL	I	.34	.24	1.34	.64	–
	II	.35	–	–	.66	.02
U_{prop}	I	.0225	.0013	.0463	.0406	–
	II	.0225	–	–	.0425	.0007
PL	I	2.05	1.12	.78	.45	-
	II	2.26	–	–	.45	.70

Concluding Remarks

- Differences in risk-utility profiles from SDL-then-edit versus edit-preserving SDL minor, especially compared to differences across SDL methods.
- Partially synthetic data: dominates on utility with one of lowest risk values. Microaggregation plus noise also on the frontier of R-U map.
- One could use partial synthesis to impute missing data and simultaneously do edit-preserving SDL. Appropriate inference methods should be identical to those in Reiter (2004).