*J. R. Statist. Soc.* A (2007)
**170**, *Part* 4, *pp.* 923–940

# Multiple imputation: an alternative to top coding for statistical disclosure control

Di An and Roderick J. A. Little

*University of Michigan, Ann Arbor, USA*

[Received May 2006. Final revision February 2007]

**Summary.** Top coding of extreme values of variables like income is a common method of statistical disclosure control, but it creates problems for the data analyst. The paper proposes two alternative methods to top coding for statistical disclosure control that are based on multiple imputation. We show in simulation studies that the multiple-imputation methods provide better inferences of the publicly released data than top coding, using straightforward multiple-imputation methods of analysis, while maintaining good statistical disclosure control properties. We illustrate the methods on data from the 1995 Chinese household income project.

*Keywords*: Confidentiality; Disclosure protection; Multiple imputation

## 1. Introduction

Statistical disclosure control (SDC) is a class of procedures that deliberately alter data that are collected by statistical agencies before release to the public, to prevent the identity of survey respondents from being revealed. These methods have increased in importance, with the extensive use of computers and the Internet. The goal of SDC methods is to reduce the risk of disclosure to acceptable levels, while releasing a data set that provides as much useful information as possible for researchers. One aspect of this is the ability to draw valid statistical inferences from the altered data.

Top coding is a simple and common SDC method that seeks to prevent disclosure on the basis of extreme values of a variable, by censoring values above a prechosen 'top code'. For example, in surveys that include income, extremely high income values are considered to be sensitive and have the potential to reveal the identity of respondents. By recoding income values that are greater than a selected top-code value to that value, respondents with very high income have reduced risk of disclosure.

It is left to the analyst to decide how top-coded data are analysed. One approach is to categorize the variable so that top-coded cases all fall in one category—this is sensible but precludes analyses that treat the variable as continuous. Another approach is to ignore the fact of top coding and to treat the top-coded values as the truth. This method is straightforward, but clearly the data distribution is distorted and biased estimates will be obtained. A better method is to treat the extreme values as censored. Under an assumed statistical model, maximum likelihood (ML) estimates can be obtained by using algorithms such as the expectation–maximization (EM) algorithm (Dempster *et al.*, 1977). This method is model based and should yield good inferences if the model is correctly specified. But we expect this method to be quite sensitive to

*Address for correspondence*: Roderick J. A. Little, Department of Biostatistics, University of Michigan, 1420 Washington Heights M4045, Ann Arbor, MI 48109-2029, USA.
E-mail: rlittle@umich.edu

 0964–1998/07/170923

model misspecification, especially when the upper tail of the assumed distribution differs markedly from that of the true distribution. The data users can also apply an imputation method to the top-coded data set and fill in the censored values. A limitation is that the imputed data fail to reflect imputation uncertainty, and imputations are sensitive to assumptions about the right tail of the distribution. We propose alternatives to top coding that allow better inferences for the data user by using simple multiple-imputation (MI) combining rules, while preserving the SDC benefits of top coding.

MI has been proposed as a method of SDC (Little, 1993; Rubin, 1993; Little *et al.*, 2004; Reiter, 2003, 2005a, b). An imputation model is built from the original data and observed values are replaced by draws from the predictive distribution based on the model. The imputation process is repeated several times and the imputed data sets are then released to the public. Applying this approach to our problem, we delete the data values that are greater than a cut-off point, which is chosen to be smaller than the top code to achieve a mixing of sensitive and non-sensitive values, and apply MI to fill in these values. We then release multiple-imputed data sets to the public. Data users can apply MI combining rules (Reiter, 2003) to obtain valid inferences, as described in Section 3.

We propose non-parametric and parametric MI methods. The non-parametric method is a hot deck procedure, where we replace the deleted values with values that are randomly drawn with replacement from the set of deleted values. The parametric method is Bayesian and assumes a model for the data, draws model parameters from their posterior distribution and then imputes the deleted values with random draws from the posterior predictive distribution.

We compare estimates of the mean of the data from our methods with two estimates from top-coded data. The first, as described previously, is to treat the top-coded values as the true values. The second is to treat those values that are greater than the top code as censored and to apply ML estimation under an assumed model.

We also investigate situations where covariates are present. We use the proposed MI methods to fill in for deleted values without conditioning on covariates. We then perform regression analysis on the imputed data set and compare regression coefficients with those from original and top-coded data. Extensions of our methods that condition on covariate data are also outlined.

The rest of this paper is organized as follows. Section 2 presents SDC approaches, and Section 3 describes corresponding methods of inference for a population mean. Section 4 describes a simulation study to evaluate the approaches in Section 3, and Section 5 applies the methods to data from the 1995 Chinese household income project. Section 6 considers estimates of regression coefficients for a regression where the outcome is subject to our disclosure control methods. Section 7 gives conclusions and discusses future work.

## 2. Methods of statistical disclosure control

Let $Y$ denote a survey variable (e.g. income) and suppose that values of $Y$ that are greater than a particular value $y_T$ are considered too sensitive for release to the public. We consider the following approaches to SDC.

(a) *Top coding*: treat $y_T$ as a top-code value, i.e. replace values of $Y$ that are greater than $y_T$ by $y_T$. The resulting sample is referred to as 'top coded'.
(b) *Hot deck MI (HDMI)*: choose a value $y_I$ that is smaller than $y_T$. Delete the values of $Y$ that are greater than $y_I$ and replace them with random draws from the set of deleted values. We choose $y_I < y_T$ to achieve a mixing of sensitive and non-sensitive values. We refer to $y_I$ as the cut-off point.

(c) *Parametric MI (PMI)*: the HDMI method provides disclosure protection by scrambling sensitive and non-sensitive values, but it is arguably limited from the point of view of SDC, since actual sensitive data values are released. The PMI methods address this concern by releasing data that have been simulated from a parametric model. First, values that are greater than $y_I$ are deleted, as with HDMI. The model—we consider the log-normal model and power-transformed normal model (the power normal model for short)—is fitted to the data. Parameters are drawn from their posterior distribution under the assumed model, and deleted values are imputed with draws from their predictive distribution. See Appendix A for details.

Write the complete data as $Y = (Y_{ret}, Y_{del})$, where $Y_{ret}$ denotes the retained values and $Y_{del}$ denotes the deleted values beyond the cut-off. We consider two versions of PMI, labelled PMIC and PMID. For PMIC, we draw the parameter $\phi$ of the model for the data $Y$ from its posterior distribution given the complete data $Y$, i.e.

$$\text{PMIC}: \phi^* \sim P(\phi|Y).$$

For PMID, we apply the parametric model to the deleted data $Y_{del}$ and draw $\phi$ from its posterior distribution given $Y_{del}$:

$$\text{PMID}: \phi^* \sim P(\phi|Y_{del}).$$

The other steps are the same for these two methods, i.e. we draw deleted values from the truncated predictive distribution

$$Y^*_{del} \sim P(Y|Y > y_I, \phi^*).$$

PMID is less efficient than PMIC since it models the deleted data and fails to exploit fully the information in $Y$ when drawing values of parameters. However, modelling the deleted data only as in PMID provides useful robustness to model misspecification, as we shall see below.

## 3. Methods of inference for the mean

We first consider the properties of these SDC methods for inferences about the mean of a variable $Y$ subject to top coding. Some comments concerning inference for other parameters are provided in Sections 6 and 7. The following estimates and associated standard errors are considered:

(a) *before deletion (method BD)*—the sample mean of original data $(y_1, y_2, \ldots, y_n)$ before SDC is

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} y_i; \tag{1}$$

this estimate is used as a bench-mark for comparing SDC methods;

(b) *top coding (method TC)*—the sample mean of the top-coded data set, namely

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n} y_{it}, \tag{2}$$

where $y_{it} = y_i$ when $y_i < y_T$ and $y_{it} = y_T$ when $y_i \geqslant y_T$; this approach is obviously biased, and our objective is to improve on it with other methods;

(c) *log-normal ML (method LNML)*—the ML estimate based on the log-normal model, computed by the EM algorithm (Appendix B). The log-normal is chosen as a convenient model for right-skewed data, but we emphasize that other models could be considered.

The standard errors for methods (a)–(c) are computed by the bootstrap, with $B = 100$ bootstrap samples.

The five remaining methods are all based on MI and create $D$ sets of imputations for values that are beyond the chosen cut point $y_I$; $D$ imputed data sets are thus created. For the $d$th imputed data set $Y^{(d)} = (y_1^{(d)}, y_2^{(d)}, \ldots, y_n^{(d)})$, where $y_i^{(d)} = y_i$ if $y_i < y_I$ and $y_i^{(d)}$ is the $d$th MI draw if $y_i \geqslant y_I$. The MI estimate is then

$$\hat{\theta}_{MI} = \frac{1}{D} \sum_{d=1}^{D} \hat{\theta}^{(d)}, \tag{3}$$

where $\hat{\theta}^{(d)}$ is the sample mean of the $d$th data set. The MI estimate of variance is

$$T_{MI} = \text{var}(\hat{\theta}_{MI}) = \bar{W} + B/D, \tag{4}$$

where $\bar{W} = \Sigma_{d=1}^{D} W^{(d)}/D$ is the average of the within-imputation variances $W^{(d)}$ for imputed data set $d$ and $B = \Sigma_{d=1}^{D} (\hat{\theta}^{(d)} - \hat{\theta}_{MI})^2/(D-1)$ is the between-imputation variance. Formula (4) differs from the original MI formula for missing data (where $B$ is multiplied by a factor $(D+1)/D$; see for example Little and Rubin (2002), page 86), for reasons that were discussed in Reiter (2003). Imputations for these MI methods are created as follows.

(d) *Hot deck MI (method HDMI)*—imputations are drawn randomly with replacement from the set of values beyond the cut-off $y_I$.
(e) *Log-normal MIC (method LNMIC)*—imputations are posterior predictions from a log-normal model fitted to the complete data before deletion.
(f) *Log-normal MID (method LNMID)*—imputations are posterior predictions from a log-normal model fitted to the deleted data beyond the cut-off.
(g) *Power normal MIC (method PNMIC)*—imputations are posterior predictions from the power normal model, the power-transformed normal distribution fitted to the full data before deletion. For convenience the power transformation is estimated by ML, and parameters are drawn from the full data posterior distribution, treating the power transformation as known. An alternative approach is to draw the power from its posterior distribution as well, but we made use of the widely available ML routine `box.cox.powers()` in R (R Project, 2007) in our calculations.
(h) *Power normal MID (method PNMID)*—imputations are posterior predictions from the power normal model, fitted to the deleted data beyond the cut-off.

## 4. Simulation study

A simulation study was carried out to evaluate and compare the SDC methods in Section 3. We computed point estimates of means and the corresponding variances and confidence intervals from the imputed data sets and compared them with those calculated from the original data set before SDC.

### 4.1. Study design

Data sets were generated from the following four distributions, all with mean 1: exponential(1), gamma(1.25, 0.8), log-normal(−0.2, 0.4) and square-root normal(0.9, 0.19) (the variances of these distributions are 1, 0.8, 0.49 and 0.69 respectively). Fig. 1 shows the form of these distributions beyond their approximate upper 10th percentile. For each simulated data set, we calculated the eight mean estimates and their corresponding variances as discussed in Section 3. To assess the validity of inferences, we calculated the 95% confidence intervals (CIs) based on

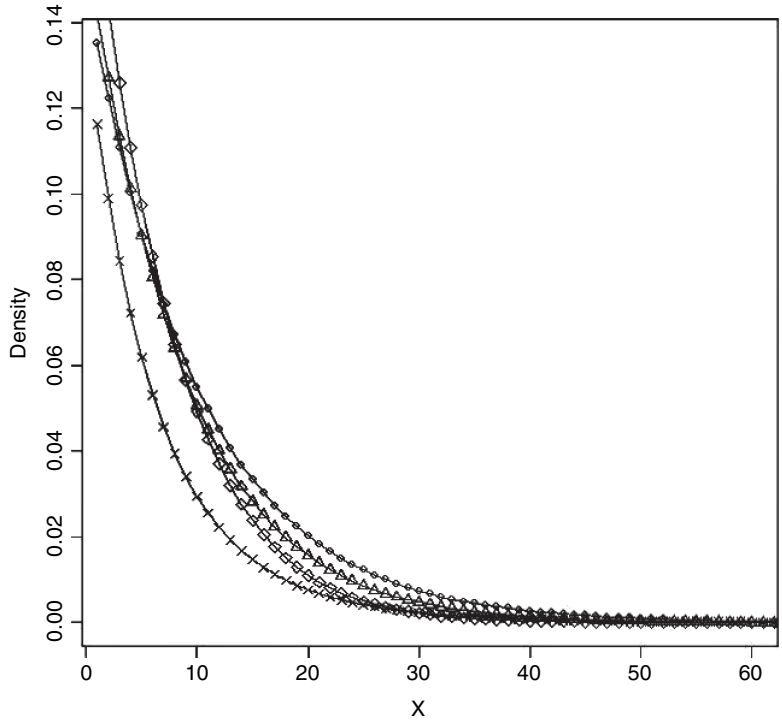


**Fig. 1.** Tails of the data distributions in the simulation study: ○, exponential(1); △, gamma(1.25, 0.8); ×, log-normal(−0.2, 0.4); ◇, square-root normal(0.9, 0.19)

the usual normal approximation and computed the proportion of CIs that contain the true mean. For parametric estimates in Section 3, the simulated data distributions are allowed to differ from those assumed in the statistical models, to provide an assessment of sensitivity to model misspecification.

In our simulations we chose the 95th percentile of the population distribution as the top-code value $y_T$. Denote by $n_S$ the number of sensitive sample values that are greater than $y_T$. We studied two alternative values for the cut-off point $y_I$: $y_{I90}$, the value with $2n_S$ larger values in the sample, and $y_{I80}$, the value with $4n_S$ larger values in the sample. These values correspond approximately to the 90th- and 80th-percentile values of the distribution, and for this reason we label the version of a method with an asterisk that uses cut-off $y_{I90}$ '*90' and the version that uses cut-off $y_{I80}$ '*80'.

Clearly the disclosure risk is reduced by increasing the fraction of non-sensitive values that are imputed. A simple measure of the risk of disclosure is the proportion of multiple-imputed values beyond the top-code value $y_T$. For all the MI methods, this is approximately 50% when the cut-off point is $y_{I90}$, and approximately 25% when the cut-off point is $y_{I80}$.

### *4.2. Results*

Tables 1 and 2 present simulation results for sample sizes 2000 and 200 respectively. Results are based on 500 data sets for each model. We set $B = 100$ for the number of bootstrap samples. For both NPMI and PMI methods, we created $D = 5$ imputed data sets. As expected, method TC underestimates the mean and has poor confidence coverage, particularly for the $n = 2000$ sample size where bias is a relatively large component of the RMSE. The HDMI methods (HDMI90

**Table 1.** Inferences about the mean from the simulation study, sample size 2000†

| *Method* | *Results for exponential data* | | | | *Results for gamma data* | | | | *Results for log-normal data* | | | | *Results for square-root normal data* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Bias* ($\times 10^3$) | *RMSE* ($\times 10^3$) | *Relative width* | *Coverage* (%) | *Bias* ($\times 10^3$) | *RMSE* ($\times 10^3$) | *Relative width* | *Coverage* (%) | *Bias* ($\times 10^3$) | *RMSE* ($\times 10^3$) | *Relative width* | *Coverage* (%) | *Bias* ($\times 10^3$) | *RMSE* ($\times 10^3$) | *Relative width* | *Coverage* (%) |
| BD | −2 | 24 | 1.00 | 93.8 | −0 | 19 | 1.00 | 96.2 | 1 | 16 | 1.00 | 94.0 | −0 | 18 | 1.00 | 94.4 |
| TC | −51 | 55 | 0.84 | 23.2 | −42 | 45 | 0.85 | 30.0 | −39 | 41 | 0.80 | 13.6 | −33 | 37 | 0.89 | 45.6 |
| LNML | 359 | 363 | 2.40 | 0 | 213 | 216 | 1.81 | 0 | 1 | 16 | 1.01 | 93.8 | 823 | 836 | 7.99 | 0 |
| HDMI90 | −2 | 24 | 1.05 | 94.8 | −0 | 19 | 1.05 | 97.4 | 1 | 16 | 1.09 | 96.6 | −0 | 19 | 1.04 | 95.4 |
| HDMI80 | −2 | 24 | 1.12 | 95.8 | −0 | 19 | 1.10 | 98.2 | 1 | 17 | 1.14 | 96.2 | −0 | 18 | 1.08 | 96.8 |
| LNMIC90 | 206 | 212 | 2.41 | 1.0 | 130 | 134 | 1.85 | 1.0 | 0 | 17 | 1.02 | 94.8 | 354 | 362 | 4.19 | 0.6 |
| LNMIC80 | 317 | 322 | 2.80 | 0 | 202 | 206 | 2.09 | 0 | 1 | 17 | 1.04 | 94.4 | 594 | 606 | 5.24 | 0.2 |
| LNMID90 | −2 | 24 | 1.00 | 93.8 | −1 | 19 | 1.01 | 95.8 | −0 | 16 | 1.00 | 94.4 | −1 | 19 | 1.01 | 93.8 |
| LNMID80 | −4 | 24 | 1.00 | 93.4 | −2 | 19 | 1.01 | 95.8 | −1 | 17 | 0.99 | 93.2 | −1 | 19 | 1.01 | 94.4 |
| PNMIC90 | 11 | 27 | 1.08 | 89.6 | 7 | 21 | 1.05 | 95.2 | 0 | 17 | 1.02 | 95.0 | 9 | 21 | 1.05 | 93.0 |
| PNMIC80 | 14 | 29 | 1.10 | 89.0 | 9 | 22 | 1.07 | 93.8 | 1 | 17 | 1.03 | 94.6 | 15 | 24 | 1.07 | 88.6 |
| PNMID90 | 2 | 27 | 1.18 | 95.0 | 2 | 21 | 1.15 | 97.2 | 0 | 17 | 1.15 | 94.0 | 1 | 19 | 1.08 | 95.2 |
| PNMID80 | 21 | 61 | 2.29 | 97.4 | 14 | 34 | 1.72 | 98.0 | 5 | 27 | 1.65 | 96.2 | 8 | 24 | 1.40 | 96.8 |

†BD, before deletion; TC, top coded; LNML, censored ML for the log-normal model; HDMI, hot deck MI; LNMIC, log-normal MI fitted to complete data; LNMID, log-normal MI fitted to deleted data; PNMIC, power normal MI fitted to complete data; PNMID, power normal MI fitted to deleted data; RMSE refers to the root-mean-squared error; 'relative width' is the fraction of the 95% CI width comparing with the BD estimate 'coverage' refers to the 95% CI coverage.

**Table 2.** Inferences about the mean from the simulation study, sample size 200

| *Method* | *Results for exponential data* | | | | *Results for gamma data* | | | | *Results for log-normal data* | | | | *Results for square-root normal data* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Bias* ($\times 10^3$) | *RMSE* ($\times 10^3$) | *Relative width* | *Coverage (%)* | *Bias* ($\times 10^3$) | *RMSE* ($\times 10^3$) | *Relative width* | *Coverage (%)* | *Bias* ($\times 10^3$) | *RMSE* ($\times 10^3$) | *Relative width* | *Coverage (%)* | *Bias* ($\times 10^3$) | *RMSE* ($\times 10^3$) | *Relative width* | *Coverage (%)* |
| BD | 5 | 71 | 1.00 | 94.2 | −5 | 60 | 1.00 | 95.2 | −6 | 50 | 1.00 | 93.2 | −1 | 55 | 1.00 | 94.8 |
| TC | −45 | 75 | 0.84 | 84.6 | −47 | 69 | 0.86 | 86.4 | −45 | 60 | 0.81 | 77.2 | −34 | 59 | 0.89 | 90.4 |
| LNML | 384 | 424 | 2.56 | 38.0 | 207 | 232 | 1.80 | 58.2 | −5 | 50 | 1.02 | 95.0 | 833 | 961 | 9.39 | 40.8 |
| HDMI90 | 5 | 72 | 1.06 | 96.0 | −5 | 60 | 1.06 | 95.4 | −6 | 51 | 1.08 | 94.4 | −1 | 55 | 1.04 | 96.2 |
| HDMI80 | 5 | 71 | 1.11 | 96.4 | −5 | 62 | 1.11 | 95.8 | −5 | 52 | 1.14 | 95.8 | −2 | 55 | 1.08 | 97.6 |
| LNMIC90 | 227 | 277 | 2.42 | 87.4 | 126 | 165 | 1.84 | 92.2 | −7 | 52 | 1.03 | 93.4 | 364 | 447 | 4.17 | 80.8 |
| LNMIC80 | 338 | 395 | 2.85 | 70.2 | 192 | 232 | 2.08 | 78.0 | −5 | 53 | 1.06 | 93.6 | 608 | 732 | 5.22 | 46.8 |
| LNMID90 | 8 | 73 | 1.03 | 94.8 | −4 | 61 | 1.03 | 94.8 | −4 | 51 | 1.03 | 95.6 | −0 | 57 | 1.02 | 94.4 |
| LNMID80 | 6 | 73 | 1.02 | 94.4 | −6 | 62 | 1.02 | 95.2 | −7 | 52 | 1.01 | 94.4 | −1 | 57 | 1.03 | 95.6 |
| PNMIC90 | 18 | 79 | 1.09 | 94.8 | 0 | 65 | 1.05 | 95.8 | −6 | 51 | 1.05 | 95.0 | 8 | 58 | 1.06 | 95.6 |
| PNMIC80 | 23 | 83 | 1.12 | 94.6 | 5 | 65 | 1.08 | 95.8 | −4 | 53 | 1.07 | 95.4 | 17 | 63 | 1.09 | 95.4 |
| PNMID90 | 15 | 94 | 1.22 | 94.8 | 4 | 69 | 1.23 | 96.4 | −2 | 55 | 1.19 | 95.2 | 3 | 60 | 1.10 | 96.6 |
| PNMID80 | 73 | 407 | 2.83 | 96.0 | 23 | 222 | 1.85 | 95.8 | 3 | 67 | 1.40 | 95.0 | 16 | 112 | 1.57 | 96.4 |

and HDMI80) have minimal bias and close to nominal coverage for all the simulated populations, with small increases in root-mean-squared error RMSE and CI width compared with the BD estimate. LNML dominates other methods for log-normal data but has serious bias and very poor confidence coverage for the other data sets, suggesting marked sensitivity to model specification. The LNMIC methods have similar properties, although they are less biased and have somewhat better confidence coverage than LNML when the model is misspecified. The LNMID methods are much more robust than their LNMIC counterparts, yielding minimal bias and good confidence coverage for all problems that were simulated.

The PNMIC methods do consistently well in terms of RMSE. Confidence coverage is close to the nominal value, except for exponential data with $n = 2000$ where coverage is a little low. This suggests that the power normal model yields good fits to the range of models that were simulated. The PNMID methods also perform well in terms of bias and confidence coverage, but they are less efficient than the PNMIC methods.

When lowering the cut-off point from $y_{I90}$ to $y_{I80}$, we observe minor increases in RMSE for HDMI, LNMID and PNMIC, and LNMIC when correctly specified. More substantial increases in RMSE are seen for PNMID, and LNMIC when misspecified. The losses in efficiency for HDMI80, LNMID80 and PNMIC80 may be acceptable given the increase in disclosure protection.

To provide a visual illustration of the imputation methods under potentially misspecified models, Fig. 2 shows the original deleted data values and the imputed values from the HDMI and four PMI methods, with cut-off $y_{I90}$, for one of the simulated square-root normal data sets with $n = 2000$. Note that the mean of the deleted values is 2.78. The HDMI predictions look similar to the deleted values and have a similar mean, 2.80.

The LNMIC predictions are too severely skewed and have some extreme predictions, reflecting the damaging effect on predictions in the tail of applying a misspecified model to the full data set. The LNMIC predictions average 5.68, which is a marked overestimate. In contrast, when the log-normal model is correctly specified, the predictions track the deleted values well (the data are not shown). The LNMID predictions have the shape of a normal distribution, reflecting effects of model misspecification, but their mean, 2.72, matches the mean of the deleted values well.

The PNMIC predictions match the deleted values quite well and have a similar mean (2.87), reflecting that this model is correctly specified, since the power normal model includes the square-root normal model as a particular case. The PNMID predictions are more skewed than the deleted values, a reflection that the power normal model does not fit that well when applied to the deleted values; however, these predictions average 2.79, which is very close to the mean of the deleted values.

In summary, we see that, for inference about the mean, the HDMI method performs best overall but has the limitations in terms of SDC that were noted above. Among the parametric imputations, LNMID has the best performance and it works almost as well as HDMI. In particular it gives good estimates of the mean even when the log-normal model is misspecified and LNMIC is biased, reflecting the fact that the effect of misspecification on the mean is limited when the model is fitted to the deleted data. (In contrast this method will work less well for large percentiles under misspecification, since the imputed distribution in the upper tail is distorted.) PNMIC also does quite well, reflecting that the power normal model fits the simulated distributions well. The PNMID method is satisfactory in terms of bias and confidence coverage, but it is considerably less efficient than PNMIC or LNMIC since it is fitting the larger power normal model to the small set of deleted values. The risk of disclosure is reduced when we increase the set of values being mixed with the sensitive cases, at the expense of some loss of efficiency of the estimate.

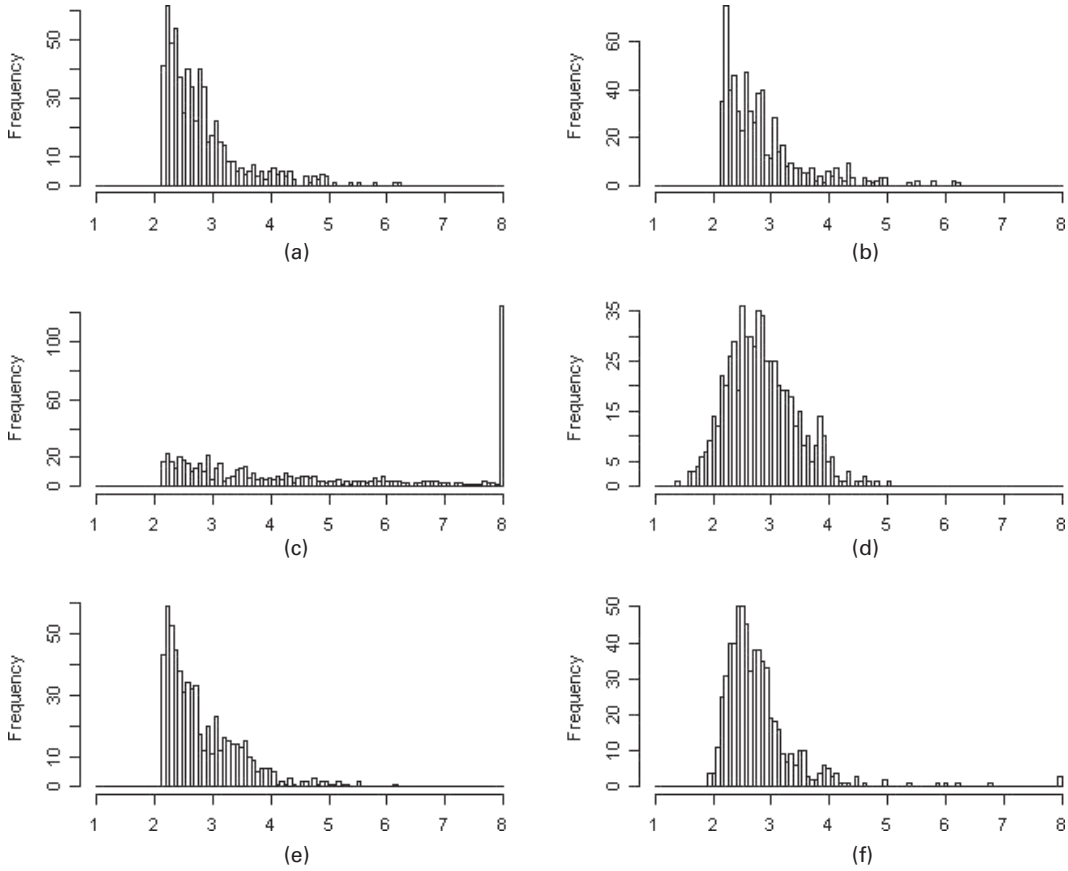


**Fig. 2.** Deleted and imputed values for square-root normal data ($n = 2000$) (values greater than 8 are pooled into one category): (a) deleted values, mean = 2.78; (b) imputed (HDMI), mean = 2.80; (c) imputed (LNMIC), mean = 5.68; (d) imputed (LNMID), mean = 2.72; (e) imputed (PNMIC), mean = 2.87; (f) imputed (PNMID), mean = 2.79

## 5. Application

We applied the above SDC methods to a subset of data from the 1995 Chinese household income project, which was conducted by the Inter-university Consortium for Political and Social Research at the University of Michigan (Riskin *et al.*, 2000). This project was designed to measure the personal income distribution in the People's Republic of China in 1995. Income information on both households and individuals was recorded for rural and urban areas. Since SDC was not applied to the released data set, the effectiveness of the various SDC methods can be readily assessed.

### *5.1. Data analysis*

We illustrated application of the SDC methods to both urban and rural individual income values. After deletion of missing and zero income values, the urban data set included 15983 individuals and the rural data set had 6296 individuals. We applied the top coding, HDMI and PMI methods to the data and computed the estimates (a)–(h) that were described in Section 3. The power transformation parameter that was estimated by the R function was 0.13 for the rural data and 0.45 for the urban data.

**Table 3.** Comparison of mean estimates, 1995 Chinese household income project, urban and rural data

| *Method* | *Results for the urban data* | | | | *Results for the rural data* | | | |
|---|---|---|---|---|---|---|---|---|
| | *Estimate* | *Fractional deviation from BD mean (%)* | *Standard error of estimate* | *Relative 95% CI width* | *Estimate* | *Fractional deviation from BD mean (%)* | *Standard error of estimate* | *Relative 95% CI width* |
| BD | 6196 | 0 | 36 | 1.0 | 2196 | 0 | 339 | 1.0 |
| TC | 5895 | −4.86 | 25 | 0.70 | 1969 | −10.36 | 25 | 0.65 |
| LNML | 7732 | 25.8 | 85 | 2.38 | 2675 | 21.8 | 59 | 1.53 |
| HDMI90 | 6196 | −0 | 41 | 1.16 | 2196 | 0 | 45 | 1.16 |
| HDMI80 | 6196 | −0 | 43 | 1.19 | 2197 | 0.01 | 47 | 1.22 |
| LNMIC90 | 6760 | 9.10 | 58 | 1.61 | 2512 | 14.39 | 70 | 1.80 |
| LNMIC80 | 7320 | 18.14 | 69 | 1.92 | 2653 | 20.80 | 77 | 1.98 |
| LNMID90 | 6174 | −0.35 | 33 | 0.92 | 2179 | −0.81 | 36 | 0.93 |
| LNMID80 | 6162 | −0.55 | 32 | 0.90 | 2164 | −1.46 | 35 | 0.90 |
| PNMIC90 | 6035 | −2.60 | 29 | 0.80 | 2205 | 0.39 | 39 | 1.01 |
| PNMIC80 | 6089 | −1.73 | 30 | 0.83 | 2223 | 1.21 | 41 | 1.05 |
| PNMID90 | 6135 | −1.98 | 37 | 1.03 | 2196 | −0.02 | 70 | 1.80 |
| PNMID80 | 6108 | −1.41 | 39 | 1.09 | 2378 | 8.26 | 338 | 8.74 |

### 5.2. Results

Table 3 displays the results from the data analysis. We plot the original deleted data values and the imputed values from the PMI and HDMI methods using cut-off point $y_{I90}$ in Figs 3 and 4 for the urban and rural data respectively.

Predictably, in both urban and rural cases, method TC underestimates the mean and yields an underestimate of the standard error because of the reduction in standard deviation from top coding. HDMI90 provides the estimate of the mean that is closest to the BD mean, with a 16% increase in standard error. LNML has a large positive bias, indicating sensitivity to the lack of fit of the log-normal model for these data. LNMIC90 is also quite biased, although it performs better than LNML. LNMID90 has negligible bias and a slightly smaller standard error than BD in both the urban and the rural data. The power normal model estimates PNMIC90 and PNMID90 also have small bias. For the urban data, PNMIC90 has relative CI widths that are less than that from BD, which seems anticonservative; for the rural data it has a standard error that is very similar to that of BD. PNMID90 shows a slight increase in CI width for urban data but a large increase in CI width for rural data, reflecting difficulties in fitting this complex model to the deleted data. Changing the cut-off point to $y_{I80}$ results in some increases in bias and standard error for LNMIC80 estimates. Estimates from HDMI80, LNMID80 and PNMIC80 are still acceptable, as are PNMID80 estimates in the urban sample. For the rural data, PNMID80 yields an estimate with strikingly large bias and standard error, the result of some very extreme outliers from imputation. It is important to check that the method is not creating extreme outliers as in this illustration.

## 6. Study of statistical disclosure control methods with covariates

To make the situation more complicated and realistic, we now introduce covariates into our analysis. We use the previous MI methods to impute deleted values, apply a linear regression model to the imputed data set, calculate estimates of regression coefficients and compare them

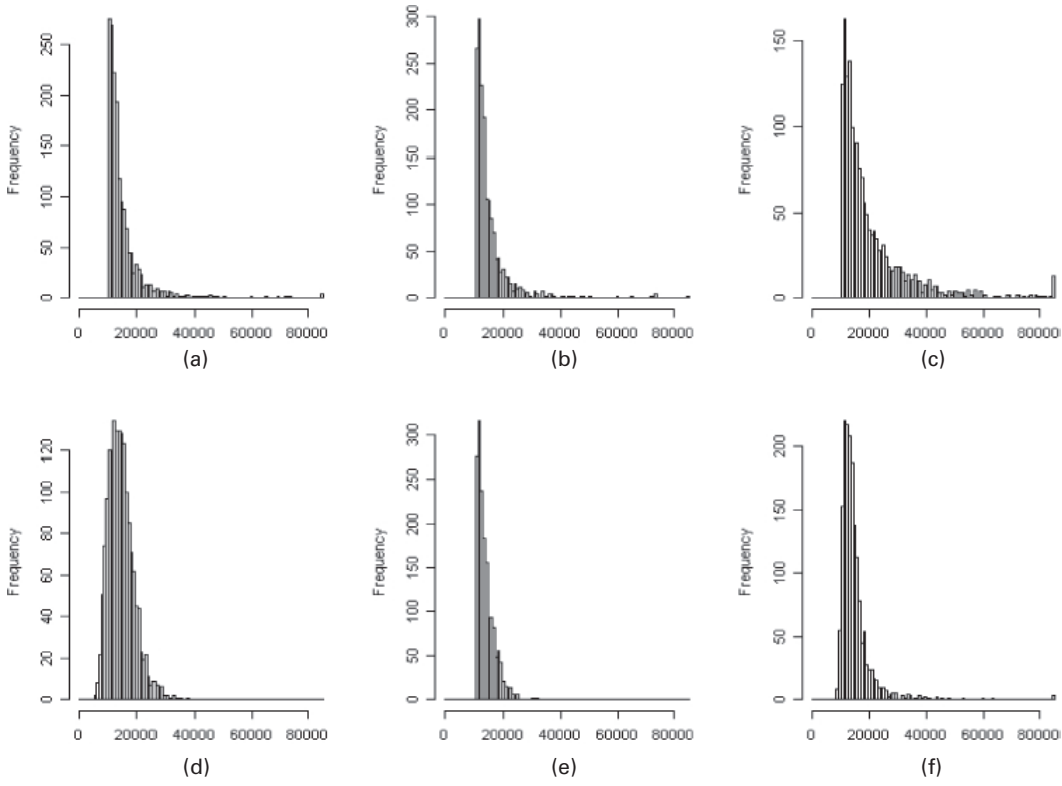


**Fig. 3.** Deleted and imputed values for the 1995 Chinese household income project, urban data (values greater than 85 000 are pooled into one category): (a) deleted values, mean = 15 164.56; (b) imputed (HDMI), mean = 14 921.94; (c) imputed (LNMIC), mean = 20 787.62; (d) imputed (LNMID), mean = 14 606.76; (e) imputed (PNMIC), mean = 13 712.23; (f) imputed (PNMID), mean = 14 081.03

with those from the original data. Since the MI methods do not condition on the covariates, we expect some bias from this procedure; our interest is in the size of the bias and resulting distortions in confidence coverage.

### *6.1. Simulation study*

Data sets were generated from the following two distributions:

(a) high correlation distribution,

$$\begin{pmatrix} X \\ \log(Y) \end{pmatrix} \sim \text{bivariate normal}\left\{\begin{pmatrix} 38 \\ 8.6 \end{pmatrix}, \begin{pmatrix} 93 & 5 \\ 5 & 0.38 \end{pmatrix}\right\};$$

(b) low correlation distribution,

$$\begin{pmatrix} X \\ \log(Y) \end{pmatrix} \sim \text{bivariate normal}\left\{\begin{pmatrix} 38 \\ 8.6 \end{pmatrix}, \begin{pmatrix} 93 & 2.5 \\ 2.5 & 0.38 \end{pmatrix}\right\}.$$

Here $X$ is considered as the independent variable and $Y$ is the dependent variable. For each simulated data set, we applied the SDC methods to impute for deleted values and performed linear regression of $\log(Y)$ on $X$. We then calculated the estimates of the regression coefficient,

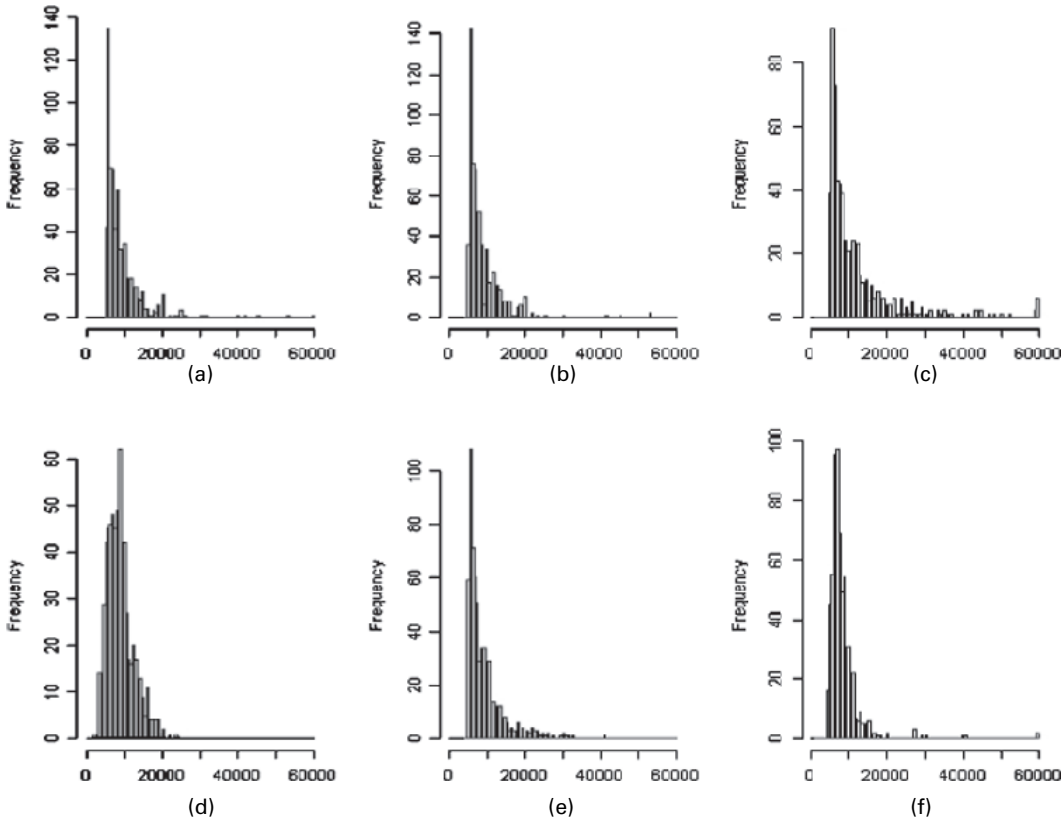


**Fig. 4.** Deleted and imputed values for the 1995 Chinese household income project, rural data (values greater than 60 000 are pooled into one category): (a) deleted values, mean = 8971.90; (b) imputed (HDMI), mean = 8780.45; (c) imputed (LNMIC), mean = 11 777.76; (d) imputed (LNMID), mean = 8698.32; (e) imputed (PNMIC), mean = 9103.11; (f) imputed (PNMID), mean = 8408.86

their corresponding variances and confidence coverage, as we did for the estimates of the mean in Section 4.

Table 4 displays results for sample sizes 2000 and 200. For data from the high correlation distribution, method TC underestimates the regression coefficient, with large RMSE and very poor confidence coverage. HDMI90 also underestimates the coefficient, as is to be expected since the relationship between the outcome and covariate is attenuated by randomly 'shuffling' the values that are beyond the top code. Nevertheless it is less biased and has better coverage than method TC. The other PMI90 methods yield almost the same result as HDMI90. When changing the cut-off point to $y_{I80}$, all MI methods yield estimates with more bias and RMSE, reduced efficiency and worse confidence coverage. When the data are from the low correlation distribution, all the MI methods have satisfactory properties. This suggests that, for more moderately correlated data, the attenuating effect from imputing without conditioning on $X$ is relatively minor. For the smaller sample size of 200, all the methods are improved in terms of RMSE and coverage, since bias is less of an issue.

### *6.2. Application in Inter-university Consortium for Political and Social Research data*

We also consider the effect of the SDC methods on a multiple regression, estimated on a subset of the urban data in the 1995 Chinese household income project. Our sample included 10 752

**Table 4.** Inference for regression coefficients from the simulation study

| *Method* | *Results for sample size 2000* | | | | | | | | *Results for sample size 200* | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *High correlation* | | | | *Low correlation* | | | | *High correlation* | | | | *Low correlation* | | | |
| | *Estimate* ($\times 10^4$) | *RMSE* ($\times 10^4$) | *Relative width* | *Coverage (%)* | *Estimate* ($\times 10^4$) | *RMSE* ($\times 10^4$) | *Relative width* | *Coverage (%)* | *Estimate* ($\times 10^4$) | *RMSE* ($\times 10^4$) | *Relative width* | *Coverage (%)* | *Estimate* ($\times 10^4$) | *RMSE* ($\times 10^4$) | *Relative width* | *Coverage (%)* |
| BD | 537 | 8 | 1.00 | 94.0 | 268 | 13 | 1.00 | 93.8 | 536 | 24 | 1.00 | 94.6 | 266 | 41 | 1.00 | 94.6 |
| TC | 510 | 28 | 0.95 | 5.6 | 255 | 19 | 0.94 | 76.6 | 508 | 39 | 0.95 | 73.8 | 253 | 43 | 0.94 | 92.8 |
| HDMI90 | 528 | 12 | 1.13 | 82.6 | 263 | 14 | 1.04 | 94.6 | 526 | 27 | 1.14 | 95.2 | 262 | 41 | 1.04 | 95.4 |
| HDMI80 | 514 | 25 | 1.25 | 26.6 | 256 | 18 | 1.06 | 86.2 | 511 | 37 | 1.26 | 91.2 | 255 | 43 | 1.06 | 95.2 |
| LNMIC90 | 528 | 13 | 1.07 | 78.2 | 264 | 14 | 1.02 | 95.0 | 526 | 29 | 1.08 | 91.2 | 261 | 42 | 1.02 | 94. |
| LNMIC80 | 514 | 26 | 1.14 | 22.2 | 256 | 18 | 1.03 | 84.2 | 511 | 39 | 1.15 | 83.4 | 254 | 43 | 1.03 | 95.2 |
| LNMID90 | 528 | 13 | 1.08 | 78.2 | 263 | 14 | 1.02 | 94.2 | 526 | 28 | 1.10 | 93.2 | 262 | 41 | 1.03 | 95.0 |
| LNMID80 | 514 | 26 | 1.16 | 21.2 | 256 | 18 | 1.03 | 84.6 | 512 | 37 | 1.18 | 87.2 | 255 | 43 | 1.04 | 95.6 |
| PNMIC90 | 528 | 13 | 1.06 | 77.2 | 264 | 14 | 1.01 | 93.2 | 526 | 28 | 1.08 | 93.7 | 262 | 41 | 1.02 | 94.8 |
| PNMIC80 | 514 | 25 | 1.13 | 23.0 | 257 | 18 | 1.02 | 85.6 | 512 | 38 | 1.16 | 86.0 | 255 | 43 | 1.03 | 94.9 |
| PNMID90 | 527 | 13 | 1.06 | 72.8 | 263 | 15 | 1.01 | 91.6 | 525 | 28 | 1.08 | 93.9 | 261 | 41 | 1.02 | 95.2 |
| PNMID80 | 512 | 27 | 1.13 | 14.8 | 256 | 18 | 1.02 | 83.4 | 510 | 38 | 1.15 | 83.6 | 254 | 43 | 1.03 | 94.9 |

individuals and 10 variables, with the logarithm of income treated as the dependent variable. The covariates were age, gender, marital status, level of education, occupation, work environment, work intensity, years of work experience and logarithm of hours worked per week. To simplify the analysis, we investigate only the scenario where the covariates are complete. We applied the top coding, HDMI and PMI methods to the data, where the PMI methods were applied to the marginal distribution of the dependent variable. We again computed estimates of regression coefficients.

We plot standardized regression coefficients after imputation against those from the original data set in Fig. 5. We choose HDMI, LNMID and PNMIC as representations of the MI methods and use the 90th, 80th, 60th and 40th percentiles of the outcome variable as cut-off points, to assess the effect of increasingly severe imputation. We observe that, with $y_{I90}$, the regression coefficients from the imputed data set are very close to those from the data set before imputation, and imputation with $y_{I80}$ also has a minor effect on the coefficients. This particular case is similar to the low correlation scenario from the simulation study. We conclude that, in a situation where the outcome and covariates are not strongly associated, the MI methods proposed are robust to the failure of the imputation model to condition on covariates. Lowering cut-off points results in larger deviation from the original coefficients, leading to greater attenuation of the relationship between outcome and covariates.

## 7. Discussion

Why should the secondary data analyst prefer our proposed MI methods for SDC to top coding? First, appropriate treatment of the top-coded data, using methods like ML for censored data, requires custom algorithms that are not widely available in standard statistical software; as a result we believe that analysts often treat the top codes as true values and assume that the bias that is introduced by this will be small. In contrast, MI inferences require only complete-data methods and simple MI combining rules. Second, the MI methods tend to be less sensitive than top coding to model misspecification, as seen in our simulation studies. There are two reasons for this—the random draws from the predictive distribution provide variability even if the model is wrong, and the MIs are based on parameter estimates that use information in the original data that is not available in the top-coded data. The data producer is also better able to assess and limit model misspecification, since she or he can compare analyses that are based on the MI data with analyses that are based on the original data. In particular, the imputations from the model can be compared with the true values.

For the data producer, MI has the advantage that the balance between disclosure protection and loss of information can be controlled by the choice of cut-off and number of MIs that are released. The use of MI allows uncertainty of imputation to be propagated, and the MIs of a particular value enhance disclosure protection by making clear to a potential snooper that these values are not real.

For inference about the mean, the HDMI, PNMIC and LNMID methods were decisively superior to top coding in our simulations. It is clear that treating the top-coded data as the observed data yields bias, the size of which depends on the fraction of cases that were top coded and the extremity of the top code. The ML methods that are based on top-coded data are more difficult to implement for the data user and are vulnerable to model misspecification. Of our preferred MI methods, the HDMI method produces excellent inferences but has limitations as an SDC method, since original values in the data set are retained. The PNMIC and LNMID methods both yielded good inferences for the mean, with the PNMIC methods yielding imputations that match well the distribution of the deleted values. The LNMIC method was vulnerable to

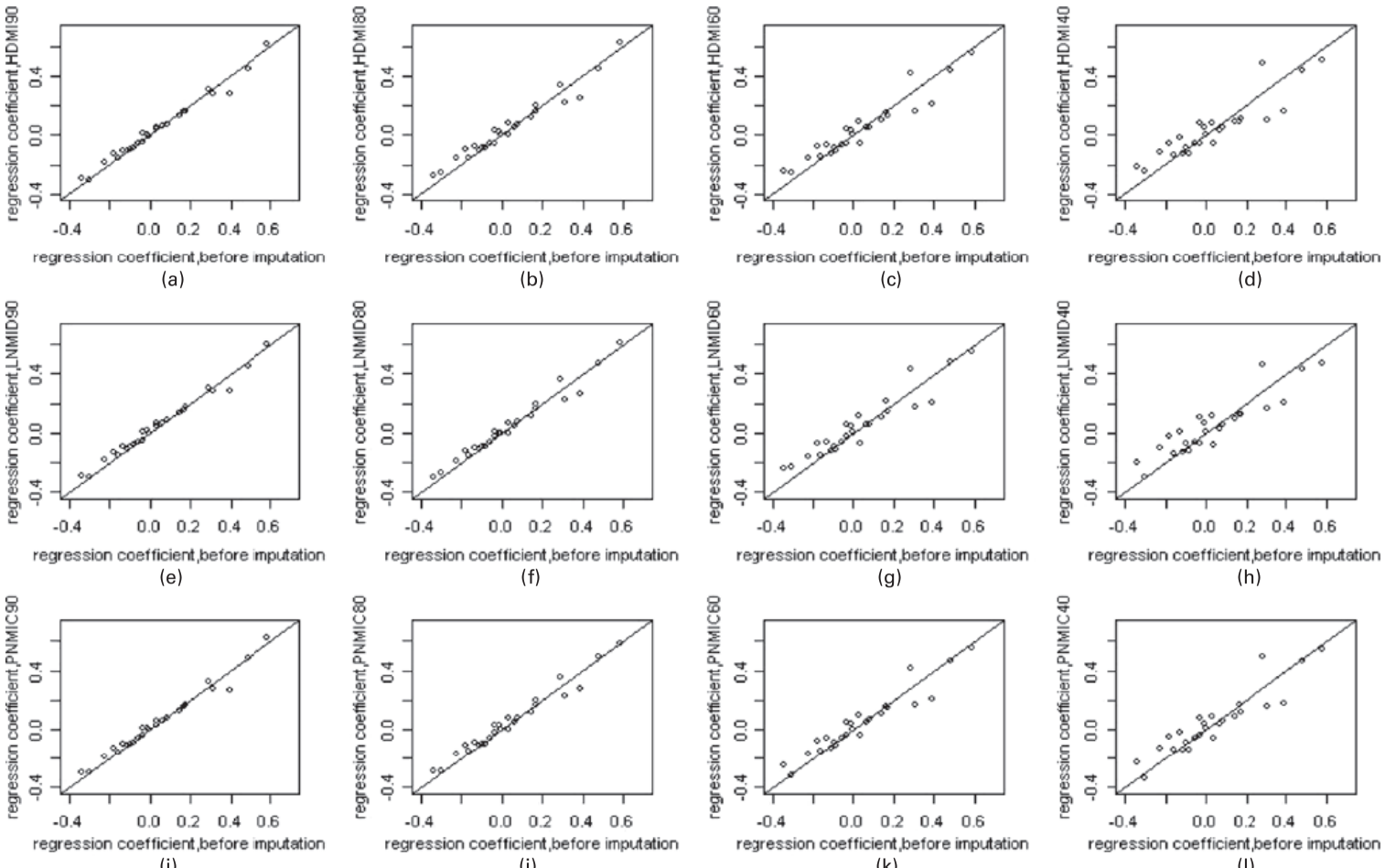


**Fig. 5.** Standardized regression coefficients, after *versus* before imputation, 1995 Chinese household income project, urban data (———, $y = x$): (a) HDMI, cut-off point 90th percentile; (b) HDMI, cut-off point 80th percentile; (c) HDMI, cut-off point 60th percentile; (d) HDMI, cut-off point 40th percentile; (e) LNMID, cut-off point 90th percentile; (f) LNMID, cut-off point 80th percentile; (g) LNMID, cut-off point 60th percentile; (h) LNMID, cut-off point 40th percentile; (i) PNMIC, cut-off point 90th percentile; (j) PNMIC, cut-off point 80th percentile; (k) PNMIC, cut-off point 60th percentile; (l) PNMIC, cut-off point 40th percentile

misspecification, and the PNMID method yielded good confidence coverage but tended to be less efficient than LNMID and PNMIC.

We chose the log-normal and power normal models to illustrate parametric MI, since they are commonly used to model skewed data; they are not universal, and the MI approach could be applied by the data producer with other models that are more suitable for the data at hand. MI that is based on a model fit to all the data (as in the LNMIC and PNMIC methods) is efficient, but vulnerable to model misspecification. Hence, if this approach is adopted, attention to good model specification is needed—in particular, it is important to check that the distribution of the imputed values in the tail is similar to the distribution of the deleted values.

MI that is based on a model fitted to the deleted values alone (the LNMID and PNMID methods) involves some loss of efficiency but is more robust to model misspecification, since the model is being fitted to the data that are being deleted. Here simpler models worked well for the mean, but more refined models may still be needed to get the shape of the distribution in the tail right. We note that, although method TC is generally inferior, it is better than MI when estimating percentiles below the top code but above the cut-off point, since the MI methods delete values in this range that are retained by method TC.

Our results clearly demonstrate the trade-off between reducing the risk of disclosure by allowing a larger pool of non-sensitive values for mixing with the sensitive cases, and reduced efficiency of the estimates. The MI technology is very helpful in propagating the increased uncertainty from the disclosure control method, resulting in good confidence coverage.

MI of deleted values should in principle condition on the observed information, and hence a refinement of the methods proposed is to condition the predictive distribution of the deleted values on observed covariates. Our preliminary assessment of inferences for regression coefficients in Section 6 confirms that a failure to condition on covariates leads to an attenuation of relationships between these covariates and $Y$. The bias was serious for highly correlated covariates and large samples, but in other situations it was surprisingly minor. This suggests that, when applying the MI method to multivariate data, it may suffice to condition on a relatively small set of covariates that are strongly associated with the variable that is subject to SDC. A simple way of doing this for a small set of categorical covariates is to apply the methods that were presented here within strata defined by the covariates, as in the urban and rural strata in the application in Section 5. More generally, regression-based extensions of the methods PNMIC and PNMID can be readily defined by including the key covariates in the mean function. We plan to develop and assess these refinements in future work.

We have confined attention here to inferences from top coding and MI methods; other alternatives to top coding are also of interest. One such alternative is to add random noise (e.g. normal noise as in Fuller (1993)) to the values beyond the top code. This method may yield satisfactory (if less efficient) inferences for the mean, but noise with substantial variance needs to be added to yield reductions of disclosure risk that are comparable with those of MI, and adding such noise potentially distorts the distribution. Also custom adjustments are needed for inferences about other parameters, such as regression coefficients. If multiple imputes are created by adding noise to the true value, the average of these imputations converges to the true value as the number of imputations increases, which is an undesirable property from the perspective of disclosure protection. Our MI methods do not have this property: the average of the MI imputed values converges to the conditional mean of the predictive distribution, not to the true deleted value. Thus increasing the number of MIs improves efficiency of inferences without compromising gains in disclosure protection. This is a major attraction of MI as an SDC method.

## Acknowledgements

This work was supported by National Institute of Child and Human Development grant P01 HD045753. The authors thank Trivellore Raghunathan and three referees for useful comments.

## Appendix A: Parametric multiple-imputation method for log-normal model and power-transformed normal model

For $X$ from a log-normal$(\mu, \sigma^2)$ distribution, $Y = \log(X) \sim N(\mu, \sigma^2)$. If $X$ is from the power-transformed normal$(\mu, \sigma^2, \lambda)$ distribution with $\lambda \neq 0$, $Y = (X^\lambda - 1)/\lambda \sim N(\mu, \sigma^2)$. To apply the PMI method we estimate $\lambda$ by its ML estimate $\hat{\lambda}$ by using the widely available routine `box.cox.powers( )` in R (see R Project (2007)) and then assume that $Y = (X^{\hat{\lambda}} - 1)/\hat{\lambda} \sim N(\mu, \sigma^2)$. (A more principled approach would also simulate $\lambda$ from its posterior distribution.)

Given data $Y = (y_1, \ldots y_n)$ from the $N(\mu, \sigma^2)$ distribution, the posterior distribution of parameters is

$$\sigma^2 | Y \sim \frac{(n-1)S^2}{\chi^2_{n-1}}, \qquad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{5}$$

and

$$\mu | \sigma^2, Y \sim N(\bar{y}, \sigma^2/n). \tag{6}$$

We draw parameters $\mu^*$ and $\sigma^{*2}$ from their posterior distribution and then draw deleted values for normal data from the predictive distribution

$$Y^*_{\text{del}} \sim N\{\mu^*, \sigma^{*2} | Y > \log(y_1)\}. \tag{7}$$

We then transform the draws of normal data back to log-normal,

$$\text{log-normal: } X^*_{\text{del}} = \exp(Y^*_{\text{del}}), \tag{8}$$

and power-transformed normal data,

$$X^*_{\text{del}} = \sqrt[\hat{\lambda}]{(\hat{\lambda} Y^*_{\text{del}} + 1)}. \tag{9}$$

## Appendix B: EM algorithm for log-normal model

If $X$ is log-normal$(\mu, \sigma^2)$, then $Y = \log(X)$ is $N(\mu, \sigma^2)$ and $\mu' = E(X) = \exp(\mu + \sigma^2/2)$. Let $Y = (y_1, \ldots, y_n)$ be a random sample from $N(\mu, \sigma^2)$, and suppose that $y_i$ is treated as missing if and only if $y_i > c$, where $c$ is a known censored value. Without loss of generality, we assume that $y_i$ is observed for $i = 1, 2, \ldots, r$ and missing for $i = r+1, \ldots, n$. The complete-data likelihood is

$$L(\mu, \sigma | Y) \propto \exp\left\{ -n \log(\sigma) - \sum_{i=1}^{n} \frac{y_i^2}{2\sigma^2} - \frac{n\mu^2}{2\sigma^2} + \mu \sum_{i=1}^{n} \frac{y_i}{\sigma^2} \right\}. \tag{10}$$

The complete-data sufficient statistics are

$$S(Y) = \left( \sum_{i=1}^{n} y_i, \sum_{i=1}^{n} y_i^2 \right). \tag{11}$$

We write $Y = (Y_{\text{obs}}, Y_{\text{del}})$, where $Y_{\text{obs}}$ denotes the observed values and $Y_{\text{mis}}$ denotes the missing values. Given parameter estimates $\theta^{(t)} = (\mu^{(t)}, \sigma^{(t)})$, the $(t+1)$th iteration of the EM method is as follows:

(a) E-step,

$$\begin{aligned} s_0^{(t+1)} &= E\left( \sum_{i=1}^{n} y_i | Y_{\text{obs}}, \theta^{(t)} \right) \\ &= \sum_{i=1}^{r} y_i + \sum_{i=r+1}^{n} E(y_i | y_i > c, \theta^{(t)}) \\ &= \sum_{i=1}^{r} y_i + (n-r) \int_c^\infty y \frac{1}{\sqrt{(2\pi\sigma^{(t)2})}} \exp\left\{ \frac{-(y-\mu^{(t)})^2}{2\sigma^{(t)2}} \right\} \mathrm{d}y \left[ \int_c^\infty \frac{1}{\sqrt{(2\pi\sigma^{(t)2})}} \exp\left\{ \frac{-(y-\mu^{(t)})^2}{2\sigma^{(t)2}} \right\} \mathrm{d}y \right]^{-1} \end{aligned} \tag{12}$$

$$s_1^{(t+1)} = E\left(\sum_{i=1}^{n} y_i^2 | Y_{\text{obs}}, \theta^{(t)}\right)$$
$$= \sum_{i=1}^{r} y_i^2 + (n-r)\int_c^{\infty} y^2 \frac{1}{\sqrt{(2\pi\sigma^{(t)2})}} \exp\left\{\frac{-(y-\mu^{(t)})^2}{2\sigma^{(t)2}}\right\} \mathrm{d}y \left[\int_c^{\infty} \frac{1}{\sqrt{(2\pi\sigma^{(t)2})}} \exp\left\{\frac{-(y-\mu^{(t)})^2}{2\sigma^{(t)2}}\right\} \mathrm{d}y\right]^{-1}; \tag{13}$$

(b) M-step,

$$\begin{aligned} \mu^{(t+1)} &= s_0^{(t+1)}/n, \\ \sigma^{(t+1)2} &= s_1^{(t+1)}/n - s_0^{(t+1)2}/n^2. \end{aligned} \tag{14}$$

Once the sequence of $\theta^{(t)}$ has converged to a stable value $(\tilde{\mu}, \tilde{\sigma})$, we calculate the ML estimate of $\mu'$ as

$$\hat{\theta}_4 = \tilde{\mu}' = \exp(\tilde{\mu} + \tilde{\sigma}^2/2). \tag{15}$$

## References


Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

Fuller, W. A. (1993) Masking procedures for microdata disclosure limitation. *J. Off. Statist.*, **2**, 383–406.

Little, R. J. A. (1993) Statistical analysis of masked data. *J. Off. Statist.*, **9**, 407–426.

Little, R. J., Liu, F. and Raghunathan, T. (2004) Statistical disclosure techniques based on multiple imputation. In *Applied Bayesian Modeling and Causal Inference from Incomplete-data Perspectives* (eds A. Gelman and X.-L. Meng), pp. 141–152. New York: Wiley.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*. New York: Wiley.

Reiter, J. P. (2003) Inference for partially synthetic, public use microdata sets. *Surv. Methodol.*, **29**, 181–188.

Reiter, J. P. (2005a) Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. *J. R. Statist. Soc.* A, **168**, 185–205.

Reiter, J. P. (2005b) Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *J. Statist. Planng Inf.*, **131**, 365–377.

Riskin, C., Zhao, R. and Li, S. (2000) Chinese Household Income Project, 1995. Political Economy Research Institute, University of Massachusetts, Amherst. (Available from `http://webapp.icpsr.umich.edu/cocoon/ICPSR-STUDY/03012.xml`.)

R Project (2007) *The R Project for Statistical Computing*. (See `http://www.r-project.org/`.)

Rubin, D. B. (1993) Satisfying confidentiality constraints through use of synthetic multiply-imputed microdata. *J. Off. Statist.*, **9**, 461–468.