CASC

CASC PROJECT
Computational Aspects of Statistical Confidentiality
March, 2004

Microdata: New Disclosure Risk Assessment Methodology

Elsayed Elamir and Chris Skinner
Southampton Statistical Sciences Research Institute,
University of Southampton, SO17 1BJ, U.K.

Deliverable No: 1.2-D4

Contents

1 Introduction

This report describes research undertaken to develop methodology for the assessment of disclosure risk for microdata. The focus is on the case of unperturbed microdata, that is, where no disclosure limitation methods have been employed. There are many ways to conceive of disclosure in the case of microdata; see Duncan and Lambert (1989). We focus on the case of identity disclosure and conceive of disclosure risk in rough terms as the probability that a microdata record may be identified, that is, that it may be linked to some known individual with some confidence using the information released in the microdata file.

We suppose that all data collected are subject to confidentiality restrictions and that the released microdata will have direct identification, such as name and address, removed. Like many authors, we shall suppose that disclosure might arise from the possibility of identifying a respondent indirectly on the basis of *identifying* or *key variables*; see, for example, Bethlehem et al. (1990), Blien et al. (1992), Duncan and Lambert (1989), Paass (1988) and Skinner et al. (1994). These are variables with values assumed known both for individuals in the microdata sample and for certain identifiable individuals in the population. We shall assume that the relevant units are individuals, but other units, such as households, are possible.

For a sample survey, the microdata file usually contains records for all units in the sample, each record contains two disjoint forms of information, identifying variables and sensitive variables. The identifying variables are assumed to be categorical, a realistic assumption in many censuses and social surveys.

2 Measures of Disclosure Risk

Bethlehem et al. (1990) and Skinner and Holmes (1993), among others, have developed a statistical framework for the estimation of disclosure risk. The

structure of the statistical framework can be described as follows. Suppose the microdata to be released by the statistical office consist of a standard rectangular data matrix

$$Z = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{pmatrix}$$

The rows of the matrix correspond to n sample units and represent records, the units may be individuals, households, businesses or other entities. The columns correspond to p variables so that z_{ij} is the value of the jth variable for the ith unit. We shall suppose that the file has already been anonymized by removing direct identifiers such as name and address, but geographical variables such as area of residence may remain. Also, we assume that $x_1, ..., x_m$ represent the key variables, present in both the microdata and the prior information. Typically, the x variables will form a subset of the z variables with $m < p$. The combinations of values of $x_1, ..., x_m$ define a series of cells. The number of cells, J, is equal to the number of cells with a positive count in the population, i.e., the product of the numbers of categories of each of the x variables less the number of cells with zero counts. A sample of size n is taken from the population of size N. Note that, not all population cells need be observed. Measures of disclosure risk will be defined in terms of the population and sample counts in the cells.

The recording of the key variables in the microdata and the prior information may be different, for example because of measurement error. Thus, it may often be necessary to conceive of two sets of key variables; $x_1, ..., x_m$, as recorded in the microdata and $\widetilde{x}_1, ..., \widetilde{x}_m$ as recorded in the prior information. Most measures of disclosure risk so far assume $x_i = \widetilde{x}_i$, In this research, we develop a simple measure of disclosure risk in the case of $x_i \neq \widetilde{x}_i$ which can happen because of, for example, measurement error in either x_i or $\widetilde{x}_i$, see,

for example, Fuller (1993) and Kuha and Skinner (1997).

2.1 File-level Measures of Risk

We define four simple measures of disclosure risk. The first two measures have established uses with 100% census data. The first measure of disclosure risk is the proportion of units in the population which have unique combinations of values of identifying variables, i.e. they are population unique (PU) and it has been used for disclosure risk assessment of census microdata in the U.S.A. and U.K.; see, for example, Greenberg and Voshell (1990) and Marsh et al. (1991). The measure is defined as

$$\Pr(\text{PU}) = \frac{N_1}{N}, \tag{1}$$

where N_1 is the number of values of X which are unique in the population and X is a categorical variable. Taking values $1, ..., J$ corresponding to the cells defined by combinations of values of $x_1, ..., x_m$.

Since only records which are sample unique (i.e. unique in the sample with respect to the identifying variables) can be population unique, it may be argued that a more realistic measure of disclosure risk is of the proportion of sample unique records which are population unique, this is the second measure; see, for example Skinner and Holmes (1993), which is defined as

$$\Pr(\text{PU} \mid \text{SU}) = \sum_j \text{I}(f_j = 1, F_j = 1) \Big/ \sum_j \text{I}(f_j = 1) \tag{2}$$

where

$$F_j = \sum_{i \in U} \text{I}(X_i = j) \qquad j = 1, 2, ..., J \tag{3}$$

and

$$f_j = \sum_{i \in s} \text{I}(X_i = j) \qquad j = 1, 2, ..., J \tag{4}$$

are the population and sample frequencies, respectively, X_i is the value of X for the ith unit and I(.) is the indicator function: $\mathrm{I}(\mathrm{A}) = 1$ if A is true and $\mathrm{I}(\mathrm{A}) = 0$ otherwise.

This is the conditional probability that, for a unit randomly drawn from the population, the unit is population unique given that the unit is sample unique.

The third measure was proposed by Skinner and Elliot (2002). The measure is defined as the probability that a unique match between a microdata record and a population unit is correct,

$$\theta = \Pr(\text{correct match} \mid \text{unique match}) \tag{5}$$

which they showed could be expressed as

$$\theta = \frac{\sum_j \mathrm{I}(f_j = 1)}{\sum_j F_j(f_j = 1)}, \tag{6}$$

if the unit is drawn at random from the population. If instead the unit is drawn at random from the sample, them this conditional probability may be expressed as

$$\theta_s = \frac{\sum_j F_j^{-1}\mathrm{I}(f_j = 1)}{\sum_j (f_j = 1)}.$$

This is the fourth measure.

To estimate θ we define the population frequencies of frequencies N_r as

$$N_r = \sum_{j=1}^{J} \mathrm{I}(F_j = r) \qquad r = 1, 2, ... \tag{7}$$

For example, N_1 is the number of population uniques. The sample quantities

n_r are defined analogously N_r. The sample frequencies of frequencies are

$$n_r = \sum_{j=1}^{J} \mathrm{I}\left(f_j = r\right) \qquad r = 0, 1, 2, ... \tag{8}$$

Skinner and Elliot (2002) showed that under simple random sampling θ may be estimated consistently by a simple point estimator as

$$\widehat{\theta} = \frac{\pi n_1}{\pi n_1 + 2\left(1-\pi\right) n_2} \tag{9}$$

where $\pi = n/N$ is the sampling fraction.

As we note from (6) and (9), θ and $\widehat{\theta}$ are sample dependent. They derived the variance of $\widehat{\theta} - \theta$ as

$$\nu = c^2 \sum_{j=1}^{J} F_j \left(F_j - 1\right) \left(\pi F_j - 3\pi + 2\right) \left(1-\pi\right)^{F_j - 1},$$

where

$$c = \left[\sum F_j \pi \left(1-\pi\right)^{F_j-1}\right] \Big/ \left[\sum F_j^2 \pi \left(1-\pi\right)^{F_j-1}\right]^2.$$

An estimate of ν is given by

$$\widehat{\nu} = \frac{2\left(1-\pi\right)\left[3\left(1-\pi\right) n_3 + \left(2-\pi\right) n_2\right]}{\left[\pi n_1 + 2\left(1-\pi\right) n_2\right]^2} \widehat{\theta}^2$$

In summary, the four measures are:

$$\Pr\left(\mathrm{PU}\right) = \sum \mathrm{I}(f_j = 1, F_j = 1)/n,$$

$$\Pr(PU|SU) = \sum \mathrm{I}(f_j = 1, F_j = 1) \Big/ \sum \mathrm{I}((f_j = 1),$$

$$\theta = \sum \mathrm{I}\left(f_j = 1\right) \Big/ \sum F_j \mathrm{I}\left(f_j = 1\right),$$

and

$$\theta_s = \sum F_j^{-1} \mathrm{I}\left(f_j = 1\right) \Big/ \sum \mathrm{I}\left(f_j = 1\right),$$

where all the summations are over $j = 1, ..., J$. We note that, the first two measures may be interpreted as the proportions of sample individuals or sample unique individuals, respectively, which are population unique; see, for example, Fienberg and Makov (1998) and Samuels (1998). Since only sample unique records can be population unique we must have $\Pr(PU) \leq \Pr(PU|SU)$ and the latter measure may be treated as more conservative. Skinner and Elliot (2002) argue, however, that both these measures may be over-optimistic, because they fail to reflect the risk arising from values of X which are twins ($F_j = 2$), triples ($F_j = 3$) and so forth, and they introduce the third and fourth measures. These may be interpreted as the probability that an observed match (on the key variables) between a sample unique individual and a known individual in the population is in fact correct, according to whether the individual is drawn at random (with equal probability) from the population, for θ, or from the sample unique cases, for θ_s. Whether θ or θ_s is a more realistic measure depends upon the assumed threat from the intruder, but it will always be the case that $\theta \leq \theta_s$.

3 Allowing for Misclassification in a File level

In this section we extend the definition of $\theta = \Pr$ (correct match | unique match) to accommodate misclassification and show that the Data Intrusion Simulation approach of Skinner and Elliot (2002) may be naturally extended to estimate θ consistently. To allow for misclassification, we now let X denote the combination of values of the identifying variables as recorded in the microdata and $\widetilde{X}$ denote the corresponding variable as measured by a potential intruder using external information. We say that misclassification occurs for unit i if $X_i \neq \widetilde{X}_i$. We do not assume that either X or $\widetilde{X}$ measures the truth. They simply reflects two ways of classifying the same quantity. In particular,

it is possible that either X or $\widetilde{X}$ is subject to measurement error and that X is subject to deliberate perturbation as a means of disclosure control.

By analogy with the definitions of F_j, f_j, N_r and n_r in Section 2.1 we let

$$\begin{aligned} \widetilde{F}_j &= \sum_{i \in U} \mathrm{I}\left(\widetilde{X}_i = j\right), \ \widetilde{f}_j = \sum_{i \in s} \mathrm{I}\left(\widetilde{X}_i = j\right) \\ \widetilde{N}_r &= \sum_{j=1}^{J} I\left(\widetilde{F}_j = r\right), \ \widetilde{n}_r = \sum_{j=1}^{J} \mathrm{I}\left(\widetilde{f}_j = r\right). \end{aligned}$$

Under misclassification, we may write θ_m =Pr(correct match| unique match) under attack method B given in Skinner and Elliot (2002) as

$$\theta_m = \frac{\sum_{i \in s} \mathrm{I}\left(f_{X_i} = 1, \widetilde{X}_i = X_i\right)}{\sum_j \widetilde{F}_j \mathrm{I}\left(\mathrm{f_j} = 1\right)} \tag{10}$$

In order to estimate θ_m we assume that misclassification takes place according to a random mechanism in which

$$\Pr\left(\widetilde{X}_i = j^* | X_i = j\right) = M_{jj^*} \quad j = 1, ..., J \quad j^* = 1, ..., J, \quad i \in U \tag{11}$$

where the matrix $M = [M_{jj^*}]$ is a $J \times J$ misclassification matrix and is assumed known.

In practice, an agency will not know M exactly, but may conduct a sensitivity analysis for plausible values of M; see, for example, Kuha and Skinner (1997). In order to obtain a point estimator of $\widehat{\theta}$ of θ we add a step to the Data Intrusion Simulation given in Skinner and Elliot (2002).

3.1 Data Intrusion Simulation under Misclassification

Repeat the following steps (independently) for $k = 1, ..., K$.

1. remove 1 unit at random from the sample;

2. determine the value of $\widetilde{X}$ for the unit randomly using M, that is set $\widetilde{X} = j^{\star}$ with probability $M_{jj^{\star}}$, where j is the unit's value of X;

3. copy the unit back into the sample with probability π(keeping its original value of X);

4. record whether the value $\widetilde{X}$ of the removed unit matches uniquely the value of x for a sample unit ($R_{uk} = 1$ if so) and, if so, whether this match is correct ($R_{ck} = 1$ if so).

The resulting estimator of θ_m is given by

$$\widehat{\theta}_m(K) = \sum_{k=1}^{K} R_{ck} R_{uk} \Big/ \sum_{k=1}^{K} R_{uk} \tag{12}$$

As in Skinner and Elliot (2002), we may obtain a closed form expression for $\widehat{\theta}_m$ as the limit of $\widehat{\theta}_m(K)$ as $K \longrightarrow \infty$. In addition to the two events considered by Skinner and Elliot (2002), a third possible event when a unique match arises must also be considered, that in step 2 X is misclassified to a value j which corresponds to a sample unique in the microdata at step 3. This event occurs with probability

$$A = \sum_{i \in s} \sum_{j=1}^{J} M_{X_i j} I(X_i \neq j) I(f_j = 1) \Big/ n \tag{13}$$

Following an analogous argument to the proof of Proposition 1 in Skinner and Elliot (2002), we obtain

$$\widehat{\theta}_m = \frac{\pi \sum_j I(f_j = 1) M_{jj}}{\pi \sum_j \mathrm{I}(f_j = 1) M_{jj} + 2(1-\pi) \sum_j \mathrm{I}(f_j = 2) M_{jj} + nA} \tag{14}$$

Note that, in the absence of misclassification, $M_{jj} = 1$, $A = 0$ and $\widehat{\theta}_m$ reduces to the expression (9). The expression for A reduces in general to

$$A = \sum_j \left[\mathrm{E}_M \left(\widetilde{f}_j \right) - M_{jj} \right] \mathrm{I}(f_j = 1) \Big/ n \tag{15}$$

where $\mathrm{E}_M(.)$ denoted the expected value with respect to the misclassification mechanism.

The consistency of $\widehat{\theta}_m$ for θ_m is outlined in the next proposition.

Proposition 1: $\widehat{\theta}_m - \theta_m = o_p(1)$, under the probability distribution induced by both the Bernoulli sampling and the misclassification mechanism in (11), where $\widehat{\theta}_m$ and θ_m are defined in (14) and (10), and assuming that F_j are bounded above.

Proof: Note first that by taking expectations of the numerator and denominators of (10) with respect to the misclassification mechanism and using independence between j we have

$$\theta_m = \sum_j I(f_j = 1) M_{jj} \Big/ \sum_j I(f_j = 1) E_M\left(\widetilde{F}_j\right) + o_p(1) \tag{16}$$

By comparing expressions (10) and (16) it is sufficient to show that

$$\begin{aligned}&\sum_j \mathrm{I}(f_j = 1) M_{jj}/n + [2(1-\pi)/\pi] \sum_j \mathrm{I}(f_j = 2) M_{jj}/n + A/\pi \\ &= \sum_j \mathrm{I}(f_j = 1) \mathrm{E}_M\left(\widetilde{F}_j\right)/n + o_p(1)\end{aligned}$$

This may be shown using the following results

$$\mathrm{E}[\mathrm{I}(f_j = 1) M_{jj}] = F_j \pi (1-\pi)^{F_j - 1} M_{jj}$$

$$\mathrm{E}[\mathrm{I}(f_j = 2) Mjj] = 0.5 F_j (F_j - 1) \pi^2 (1-\pi)^{F_j - 2} M_{jj}$$

$$\mathrm{E}(A) = \mathrm{E}\left[\sum_j \left(\sum_{j^\star \neq j} f_{j^\star} M_{j^\star j}\right) \mathrm{I}(f_j = 1)\right]/n$$

$$= \sum_j \pi \left[\mathrm{E}_M \left(\widetilde{F}_j \right) - F_j M_{jj} \right] F_j \pi \left(1-\pi\right)^{F_j - 1} / n$$

$$\mathrm{E}\left[\mathrm{I}\left(f_j = 1\right) \mathrm{E}_M \left(\widetilde{F}_j \right) \right] = F_j \pi \left(1-\pi\right)^{F_j - 1} \mathrm{E}\left(\widetilde{F}_j \right)$$

3.2 Simplified Measure

The expression for $\widehat{\theta}_m$ may be complex to compute because of its dependence on terms of $M_{jj^\star}$ for $j \neq j$. We now consider putting θ_m and $\widehat{\theta}_m$ into simpler form, depending only on the diagonal elements of the misclassification matrix. Let us define the sample misclassification factor as

$$\rho = \sum_{i \in s} \mathrm{I}\left(f_{x_i} = 1, x_i = \widetilde{x}_i\right) \Big/ \sum_j \mathrm{I}\left(f_j = 1\right) \tag{17}$$

Our simplified measure of disclosure risk in the case of misclassification is given by

$$\theta_{mm} = \Pr\left(\text{correct match} | \text{unique match}\right) \times \text{Misclass.factor} = \frac{\sum_j \mathrm{I}\left(\mathrm{f}_j = 1\right)}{\sum_j F_j \mathrm{I}\left(f_j = 1\right)} \times \rho \tag{18}$$

Using (17), this simplifies to

$$\theta_{mm} = \frac{\sum_{i \in s} \mathrm{I}\left(\mathrm{f}_{x_i} = 1, X_i = \widetilde{X}_i\right)}{\sum_j F_j \left(f_j = 1\right)}. \tag{19}$$

When there is no misclassification ($X_i = \widetilde{X}_i$) we have $\rho = 1$, that is

$$\sum_{i \in s} \mathrm{I}\left(f_{x_i} = 1, X_i = X_i\right) = \sum_{j=1}^{J} \mathrm{I}\left(f_j = 1\right) = n_1. \tag{20}$$

When there is complete misclassification ($X_i \neq \widetilde{X}_i$) we have $\rho = 0$, that is

$$\sum_{i \in s} \mathrm{I}\left(f_{x_i} = 1, X_i \neq X_i\right) = 0$$

Following (Skinner and Elliot (2002)), an estimator of θ_{mm} is given by

$$\widehat{\theta}_{mm} = \frac{\pi \sum_j \mathrm{I}\left(f_j = 1\right) M_{jj}}{\pi n_1 + 2\left(1 - \pi\right) n_2} \tag{21}$$

If we assume the M_{jj} are known and equal to all, we could obtain the variance of θ_{mm} and $\widehat{\theta}_{mm}$ as follows

$$\nu_{mm} = \rho^2 z^2 \sum_{j=1}^{J} F_j\left(F_j - 1\right)\left(\pi F_j - 3\pi + 2\right)\left(1 - \pi\right)^{F_j - 1} \tag{22}$$

where

$$z = \left[\sum F_j \pi \left(1 - \pi\right)^{F_j - 1}\right] \Big/ \left[\sum F_j^2 \pi \left(1 - \pi\right)^{F_j - 1}\right]^2 \tag{23}$$

and the estimate of ν_{mm} is given by

$$\widehat{\nu}_{mm} = \frac{2\left(1 - \pi\right)\left[3\left(1 - \pi\right) n_3 + \left(2 - \pi\right) n_2\right]}{\left[\pi n_1 + 2\left(1 - \pi\right) n_2\right]^2} \widehat{\theta}^2 M_{jj}^2 \tag{24}$$

See Skinner and Elliot (2002) for these expressions in the no misclassification case.

In the next Section we study the statistical properties of θ, $\widehat{\theta}$, $\widehat{\theta}_m$ and $\widehat{\theta}_{mm}$ using a simulation study.

4 Simulation Procedures

We set up a simulation study, allowing for misclassification, as follows:

1. We assume m binary key variables where m is a pre-specified value e.g. $m = 12$ and there are $J = 2^m$ possible combinations of key variables.

2. Let Y_k denote the value either 0 or 1 of key variable k $(k = 1, ..., m)$, let $j = 1..., J$ denote the cells formed by cross-classifying $Y_1.....Y_m$ and let Y_{kj} denote the value of Y_k for cell j.

3. Let the probability for cell j is given by

$$\pi_j = \prod_{k=1}^{m} p^{y_{kj}} (1-p)^{1-y_{kj}} \qquad j = 1, 2, ..., J$$

where p is a pre-specified value $0 < p < 1$. Hence, it assumed that the variables $Y_1, ..., Y_m$ an independent random variables with $Pr(Y_k = 1) = p$ for each k.

4. Generate λ_j from the Lognormal distribution or gamma distribution with chosen parameters, as an example for some choices for these parameters see Tables (1) and (4).

5. Generate population frequencies as Poisson random variables by

$$F_j \sim \text{poisson}\,(\lambda_j) \qquad j = 1, 2, ..., J$$

6. Generate $\widetilde{y}_{kji}$ for $k = 1, 2, ..., m$, $j = 1, 2, ..., J$ and $i = 1, ..., F_j$ as follows

   (a) For each $j = 1, .., J$ and each $i = 1, ..., F_j$ generate $\widetilde{y}_{1ji}, \widetilde{y}_{2ji}, ..., \widetilde{y}_{mji}$ independently as

$$\widetilde{y}_{kji} = \begin{cases} 1 & \text{with prob.} \quad 1-\alpha_1 \quad \text{if} \quad y_{kj} = 1 \\ 0 & \text{with prob.} \quad \alpha_1 \quad \text{if} \quad y_{kj} = 1 \\ & \\ 1 & \text{with prob.} \quad \alpha_0 \quad \text{if} \quad y_{kj} = 0 \\ 0 & \text{with prob.} \quad 1-\alpha_0 \quad \text{if} \quad y_{kj} = 0 \end{cases}$$

   where α_0 and α_1 are a pre-specified misclassification probabilities.

(b) Compute $\widetilde{F}_j$ as

$$\widetilde{F}_j = \sum_{j=1}^{J} \sum_{i=1}^{F_j} \mathrm{I}\left(j_{ji}^* = j\right) \quad j = 1, 2, ..., J$$

where

$$j_{ji}^* = 1 + \sum_{k=1}^{m} 2^{(k-1)} \widetilde{y}_{kji}$$

7. Generate A_{ji}, $j = 1, 2, ..., J$ and $i = 1, 2, ..., F_j$ by (note that, A is sampling indicator)

$$A_{ji} = \begin{cases} 1 & \text{with} \quad \text{prob.} \quad \pi \\ 0 & \text{with} \quad \text{prob.} \quad (1-\pi) \end{cases}$$

unit ji included i-sample

8. compute sample frequencies as

$$f_j = \sum_{i=1}^{F_j} A_{ji}$$

9. Compute θ_m, $\widehat{\theta}_m$,θ_{mm}, $\widehat{\theta}_{mm}$,θ and $\widehat{\theta}$ and their variances.

10. Repeat from (6) to (9) R times. In the simulation we take $R = 100$ and the mean and variances of different measures are evaluated over R replications in the usual way.

Note that to compute $\sum_{i \in s} M_{x_i j}$ for given j we use

$$\begin{aligned} \sum_{i \in s} M_{x_i j} &= \sum_{j_o}^{J} \sum_{i=1}^{F_{j_o}} \mathrm{I}\left(A_{j_o i} = 1\right) M_{j_o j} \\ &= \sum_{j_o=1}^{J} f_{j_o} M_{j_o j} \end{aligned}$$

		$\alpha_0 = \alpha_1$				
		0.0	0.02	0.05	0.10	0.20
Mean of	Risk measure, θ_m	0.02471	0.01876	0.01284	0.00754	0.00247
	Estimator, $\widehat{\theta}_m$	0.02545	0.01896	0.01264	0.00726	0.00248
	Bias, $\widehat{\theta}_m - \theta_m$	0.00074	0.0002	-.0002	-.0002	0.00001
	Standard error(θ_m)	0.00372	0.00301	0.00202	0.00146	0.00096
	Standard error($\widehat{\theta}_m$)	0.00708	0.00390	0.00190	0.00077	0.00017
	Standard error($\widehat{\theta}_m$-θ_m)	0.00867	0.00534	0.00285	0.00131	0.00095
Mean of	Risk Measure, θ_{mm}	0.02471	0.02099	0.01636	0.01103	0.00400
	Estimator, $\widehat{\theta}_{mm}$	0.02545	0.02165	0.01635	0.01083	0.00407
	Bias, $\widehat{\theta}_{mm} - \theta_{mm}$	0.00074	0.00066	-.00001	-.00020	0.00007
	Standard error(θ_{mm})	0.00372	0.00374	0.00319	0.00220	0.00158
	Standard error($\widehat{\theta}_{mm}$)	0.00708	0.00602	0.00437	0.00349	0.00104
	Standard error($\widehat{\theta}_{mm}$-θ_{mm})	0.00867	0.00786	0.00605	0.00387	0.00203

Table 1: Simulation results for lognormal distribution with $m = 12$, $p = 0.35$, $J = 2^m$, $\mu = 0$, $\sigma = 1$, $N = 17411$ and $\pi = 0.02$.

where

$$M_{j_o j} = \prod_{k=1}^{m} (1-\alpha_1)^{y_{kj_o} y_{kj}} \, \alpha_1^{y_{kj_o}(1-y_{kj})} \times \alpha_0^{(1-y_{kj_o})y_{kj}} \, (1-\alpha_0)^{(1-y_{kj_o})(1-y_{kj})}$$

4.1 Results of Simulation Study

The simulation results for the three measures θ, $\widehat{\theta}$, θ_m, $\widehat{\theta}_m$, θ_{mm} and $\widehat{\theta}_{mm}$ are given in Tables 1, 2, 3, 4, 5 and 6. Also, box plots and quantile plots for θ, $\widehat{\theta}$, θ_m, $\widehat{\theta}_m$, θ_{mm} and $\widehat{\theta}_{mm}$ are given in Figures 1, 2, 3 and 4.

Notes on the Tables

1. θ, $\widehat{\theta}$, θ_m, $\widehat{\theta}_m$, θ_{mm} and $\widehat{\theta}_{mm}$ are random quantities (it is not a fixed population parameters but are sample dependent).

2. The estimators tend to increase as the sampling fraction increases and

		$\alpha_0 = \alpha_1$				
		0.0	0.02	0.05	0.10	0.20
Mean of	Risk measure, θ_m	0.04801	0.03232	0.01993	0.01012	0.00307
	Estimator, $\widehat{\theta}_m$	0.05055	0.03308	0.02045	0.01013	0.00306
	Bias, $\widehat{\theta}_m - \theta_m$	0.00254	0.00076	0.00052	0.00001	−.00001
	standard error(θ_m)	0.00527	0.00376	0.00261	0.00204	0.00090
	standard error($\widehat{\theta}_m$)	0.01231	0.00512	0.00242	0.00102	0.00021
	standard error($\widehat{\theta}_m$-θ_m)	0.01353	0.00583	0.00299	0.00189	0.00089
Mean of	Risk Measure, θ_{mm}	0.04801	0.04075	0.03181	0.02086	0.00834
	Estimator, $\widehat{\theta}_{mm}$	0.05055	0.04300	0.03420	0.02145	0.00825
	Bias, $\widehat{\theta}_{mm} - \theta_{mm}$	0.00254	0.00225	0.00239	0.00059	−.00009
	standard error(θ_{mm})	0.00527	0.00534	0.00485	0.00424	0.00248
	standard error($\widehat{\theta}_{mm}$)	0.01231	0.01048	0.00813	0.00537	0.00195
	standard error($\widehat{\theta}_{mm}$-θ_{mm})	0.01353	0.01149	0.00963	0.00676	0.00321

Table 2: Simulation results for lognormal distribution with $m = 12$, $p = 0.35$, $J = 2^m$, $\mu = 0$, $\sigma = 1$, $N = 17411$ and $\pi = 0.05$.

		$\alpha_0 = \alpha_1$				
		0.0	0.02	0.05	0.10	0.20
Mean of	Risk measure, θ_m	0.08410	0.05026	0.02834	0.01295	0.00354
	Estimator, $\widehat{\theta}_m$	0.08612	0.04968	0.02701	0.01268	0.00350
	Bias, $\widehat{\theta}_m - \theta_m$	0.00202	−.00058	−.00133	−.00027	−.00004
	standard error(θ_m)	0.00899	0.00639	0.00433	0.00279	0.00114
	standard error($\widehat{\theta}_m$)	0.01702	0.00614	0.00272	0.00112	0.00028
	standard error($\widehat{\theta}_m$-θ_m)	0.01960	0.00771	0.00373	0.00256	0.00111
Mean of	Risk Measure, θ_{mm}	0.08410	0.07230	0.05804	0.03712	0.01399
	Estimator, $\widehat{\theta}_{mm}$	0.08612	0.07326	0.05453	0.03544	0.01371
	Bias, $\widehat{\theta}_{mm} - \theta_{mm}$	0.00202	0.00096	−.00351	−.00168	−.00028
	standard error(θ_{mm})	0.00899	0.00943	0.00920	0.00787	0.00476
	standard error($\widehat{\theta}_{mm}$)	0.01702	0.01448	0.01023	0.00696	0.00318
	standard error($\widehat{\theta}_{mm}$-θ_{mm})	0.01960	0.01797	0.01473	0.01085	0.00549

Table 3: Simulation results for lognormal distribution of $m = 12$, $p = 0.35$, $J = 2^m$, $\mu = 0$, $\sigma = 1$, $N = 17411$ and $\pi = 0.10$.

		$\alpha_0 = \alpha_1$				
		0.0	0.02	0.05	0.10	0.20
Mean of	Risk measure, θ_m	0.04836	0.04209	0.03372	0.02207	0.00916
	Estimator, $\widehat{\theta}_m$	0.05139	0.04391	0.03474	0.02290	0.00928
	Bias, $\widehat{\theta}_m - \theta_m$	0.00303	0.00182	0.00102	0.00083	0.00012
	standard error(θ_m)	0.00259	0.00305	0.00329	0.00364	0.00236
	standard error($\widehat{\theta}_m$)	0.01706	0.01245	0.00775	0.00359	0.00116
	standard error($\widehat{\theta}_m$-θ_m)	0.01718	0.01272	0.00909	0.00543	0.00264
Mean of	Risk Measure, θ_{mm}	0.04836	0.04122	0.03224	0.02043	0.00815
	Estimator, $\widehat{\theta}_{mm}$	0.05139	0.04372	0.03452	0.02213	0.00872
	Bias, $\widehat{\theta}_{mm} - \theta_{mm}$	0.00303	0.00250	0.00228	0.0017	0.00057
	standard error(θ_{mm})	0.00259	0.00316	0.00357	0.00344	0.00213
	standard error($\widehat{\theta}_{mm}$)	0.01706	0.01451	0.01144	0.00661	0.00285
	standard error($\widehat{\theta}_{mm}$-θ_{mm})	0.01718	0.01474	0.01256	0.00791	0.00390

Table 4: Simulation results for gamma distribution of $m = 12$, $p = 0.35$, $J = 2^m$, $\alpha = 3.61$, $\beta = 4.99$, $N = 17411$ and $\pi = 0.02$.

		$\alpha_0 = \alpha_1$				
		0.0	0.02	0.05	0.10	0.20
Mean of	Risk measure, θ_m	0.05520	0.04675	0.03660	0.02434	0.00909
	Estimator, $\widehat{\theta}_m$	0.05605	0.04742	0.03657	0.02364	0.00928
	Bias, $\widehat{\theta}_m - \theta_m$	0.00085	0.00067	−.00003	−.0007	0.00019
	standard error(θ_m)	0.00244	0.00253	0.00292	0.00258	0.00216
	standard error($\widehat{\theta}_m$)	0.01065	0.00774	0.00459	0.00219	0.00061
	standard error($\widehat{\theta}_m$-θ_m)	0.01133	0.00834	0.00558	0.00358	0.00232
Mean of	Risk Measure, θ_{mm}	0.05520	0.04673	0.03680	0.02435	0.00909
	Estimator, $\widehat{\theta}_{mm}$	0.05605	0.04768	0.03699	0.02375	0.00933
	Bias, $\widehat{\theta}_{mm} - \theta_{mm}$	0.00085	0.00095	0.00019	−.00060	0.00024
	standard error(θ_{mm})	0.00244	0.00277	0.00326	0.00281	0.00221
	standard error($\widehat{\theta}_{mm}$)	0.01065	0.00906	0.00659	0.00417	0.00169
	standard error($\widehat{\theta}_{mm}$-θ_{mm})	0.01133	0.00969	0.00769	0.00521	0.00291

Table 5: Simulation results for gamma distribution with $m = 12$, $p = 0.35$, $J = 2^m$, $\alpha = 3.61$, $\beta = 4.99$, $N = 17411$ and $\pi = 0.05$.

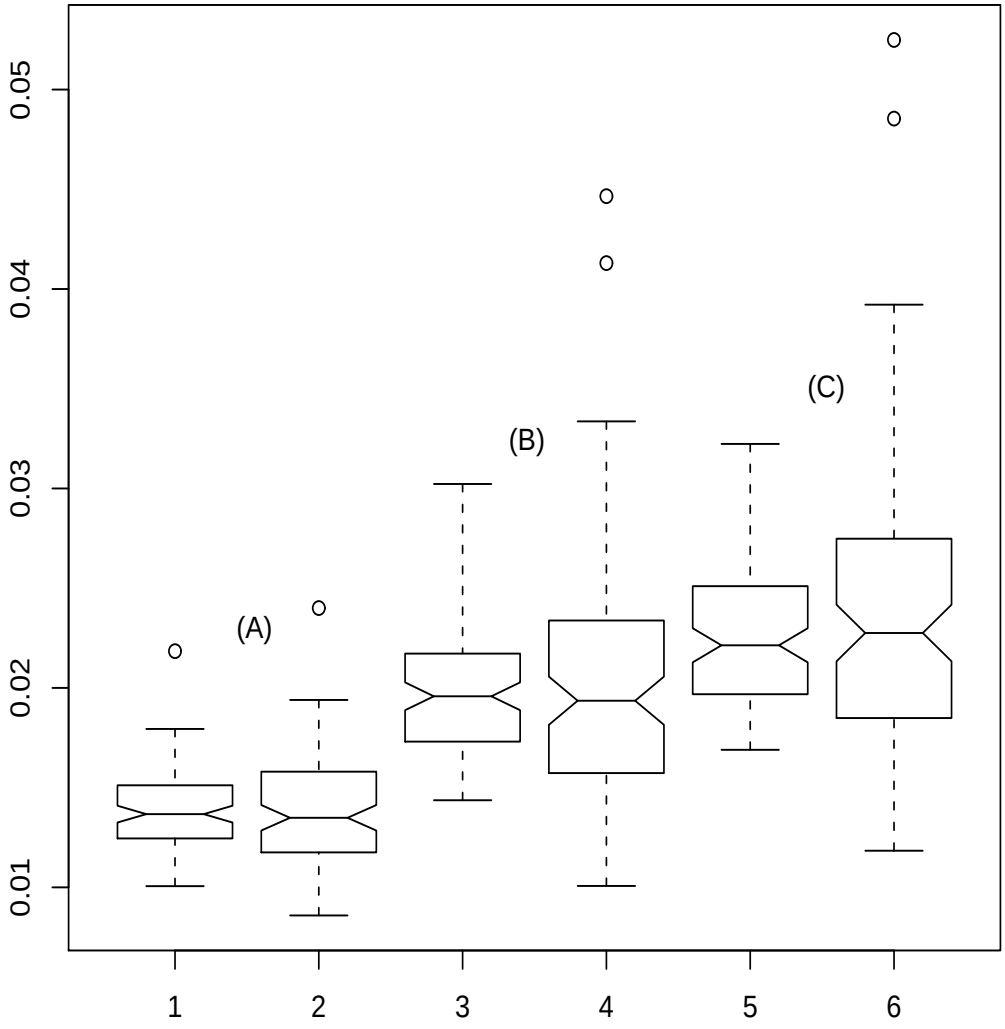

Figure 1: Boxplot for lognormal distribution with $m = 12$, $p = 0.35$, $J = 2^m$, $\mu = 0$, $\sigma = 1$ and $N = 17411$ for (A) θ_m and $\widehat{\theta}_m$, (B) θ_{mm} and $\widehat{\theta}_{mm}$, and (C) θ and $\widehat{\theta}$ for values of $\pi = 0.02$ and $\alpha_0 = \alpha_1 = 0.02$.

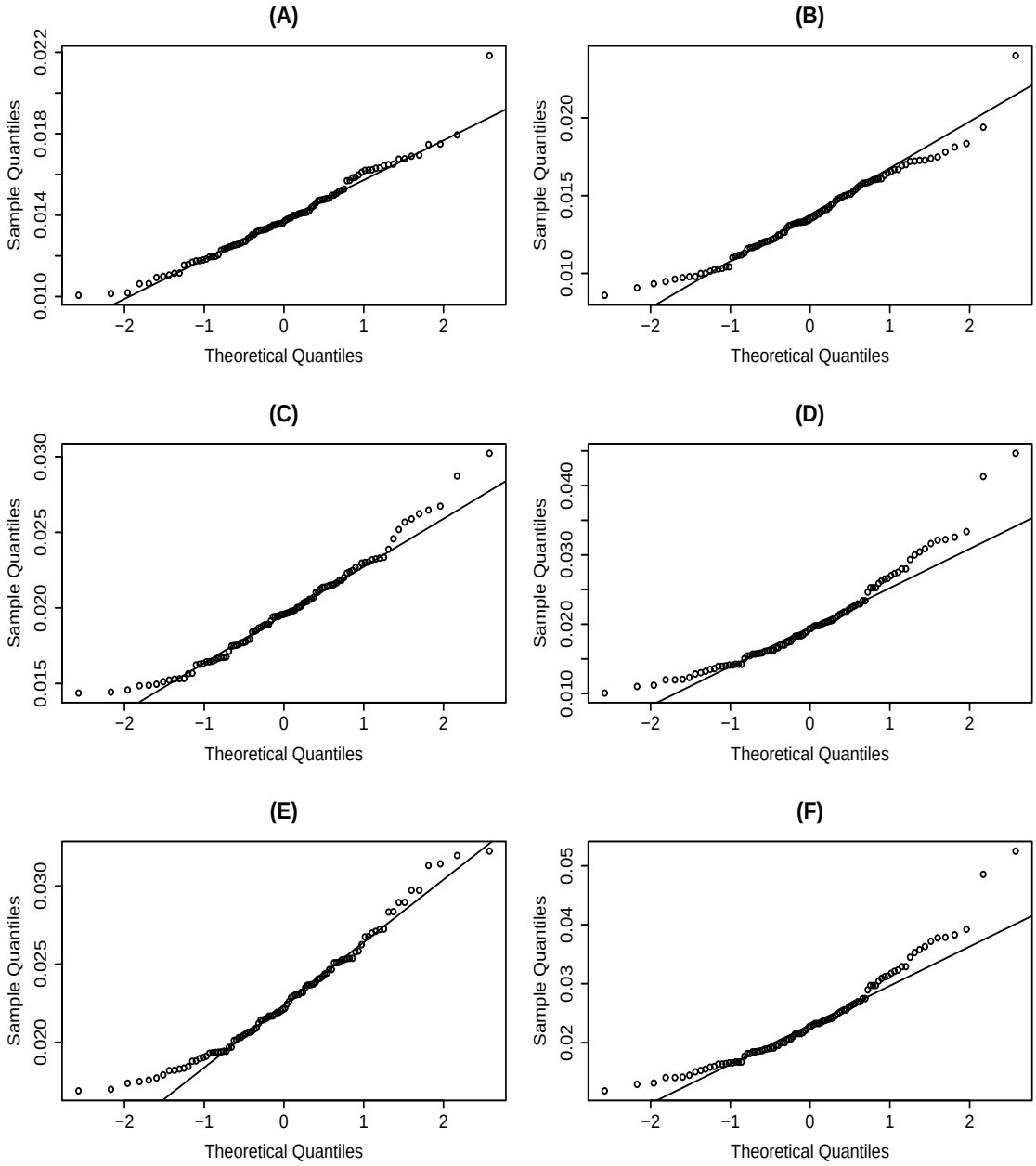

Figure 2: Quantile quantile plot for lognormal distribution with $m = 12$, $p = 0.35$, $J = 2^m$, $\mu = 0$, $\sigma = 1$ and $N = 17411$ for (A) θ_m and $\widehat{\theta}_m$, (B) θ_{mm} and $\widehat{\theta}_{mm}$, and (C) θ and $\widehat{\theta}$ for values of $\pi = 0.02$ and $\alpha_0 = \alpha_1 = 0.02$.

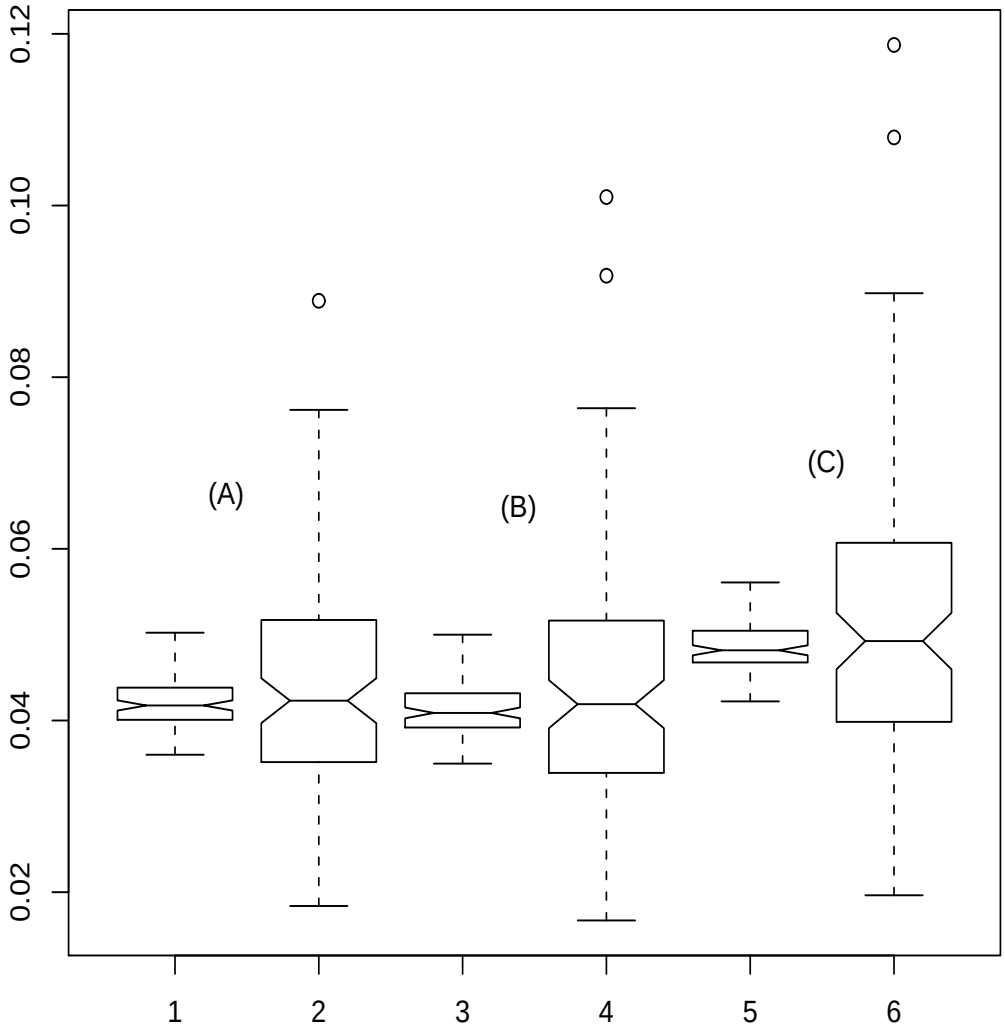

Figure 3: Box plot for gamma distribution with $m = 12$, $p = 0.35$, $J = 2^m$, $\alpha = 3.61$, $\beta = 4.99$ and $N =$ for (A) θ_m and $\widehat{\theta}_m$, (B) θ_{mm} and $\widehat{\theta}_{mm}$, and (C) θ and $\widehat{\theta}$ for values of $\pi = 0.02$ and $\alpha_0 = \alpha_1 = 0.02$.

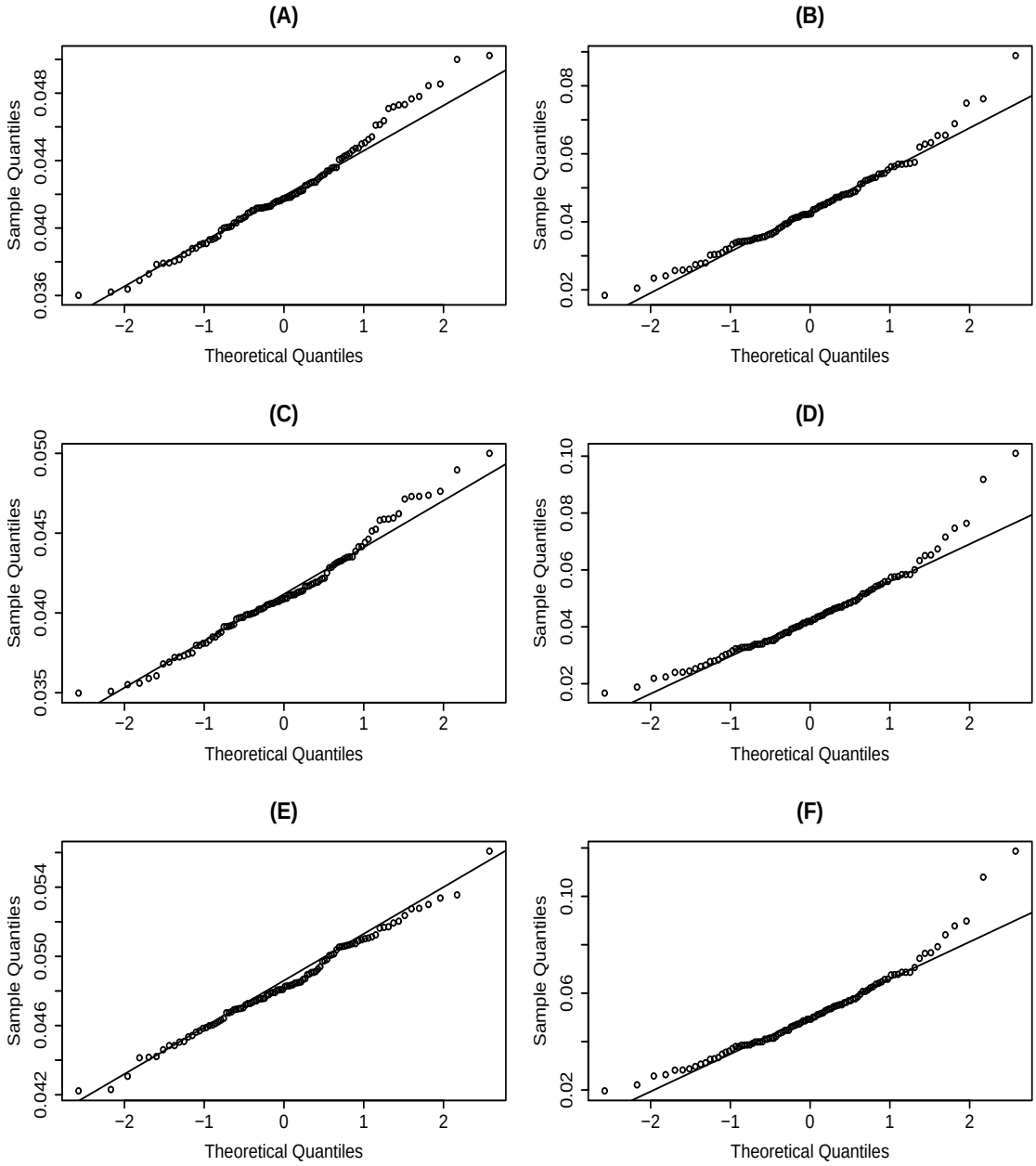

Figure 4: Quantile quantile plot for gamma distribution with $m = 12$, $p = 0.35$, $J = 2^m$, $\alpha = 3.61$, $\beta = 4.99$ and $N =$ for (A) θ_m and $\widehat{\theta}_m$, (B) θ_{mm} and $\widehat{\theta}_{mm}$, and (C) θ and $\widehat{\theta}$ for values of $\pi = 0.02$ and $\alpha_0 = \alpha_1 = 0.02$.

		$\alpha_0 = \alpha_1$				
		0.0	0.02	0.05	0.10	0.20
Mean of	Risk measure, θ_m	0.06650	0.05472	0.04135	0.02538	0.00971
	Estimator, $\widehat{\theta}_m$	0.06616	0.05444	0.04098	0.02565	0.00961
	Bias, $\widehat{\theta}_m - \theta_m$	$-.00034$	$-.00028$	$-.00037$	.00027	$-.00010$
	standard error(θ_m)	0.00305	0.00352	0.00372	0.00345	0.00249
	standard error($\widehat{\theta}_m$)	0.01003	0.00694	0.00426	0.00179	0.00047
	standard error($\widehat{\theta}_m$-θ_m)	0.01099	0.00800	0.00546	0.00382	0.00253
Mean of	Risk Measure, θ_{mm}	0.06650	0.05647	0.04422	0.02832	0.01136
	Estimator, $\widehat{\theta}_{mm}$	0.06616	0.05628	0.04409	0.02896	0.01145
	Bias, $\widehat{\theta}_{mm} - \theta_{mm}$	$-.00034$	$-.00019$	$-.00013$	0.00064	0.00009
	standard error(θ_{mm})	0.00305	0.00395	0.00449	0.00401	0.00302
	standard error($\widehat{\theta}_{mm}$)	0.01003	0.00853	0.00687	0.00449	0.00187
	standard error($\widehat{\theta}_{mm}$-θ_{mm})	0.01099	0.00958	0.00785	0.00613	0.00369

Table 6: Simulation results for gamma distribution with $m = 12$, $p = 0.35$, $J = 2^m$, $\alpha = 3.61$, $\beta = 4.99$, $N = 17411$ and $\pi = 0.10$.

tend to decrease as the misclassification probabilities increase.

3. In terms of bias, we see that for each sampling fraction the absolute bias is small.

4. In terms of coefficient of variation, the estimators are fairly stable for different misclassification probabilities and sampling fraction as well.

5. The estimator $\widehat{v}^{0.5}$ does appear to be approximately unbiased although we can see there is a slight upwards bias.

6. The quantile and box plots do not show series departure from normal distribution.

5 Disclosure Risk at the Record Level

File-level measures such as 'the proportion of individuals in the microdata file at risk of disclosure' may be problematic, if it is considered unacceptable for disclosure to arise for *any* individual in the file. In this case, even if one individual out of $10,000$ in the microdata sample is seriously 'at risk' then this might be unacceptable, despite the small value (0.0001) of the measure. The basic problem here is that the measure is a 'file-level' measure which 'averages the risk' across the whole microdata sample and thus may conceal small parts of the sample where the risk is high.

To address such concerns, it is natural to consider a record-level measure, i.e. a measure which may take a different value for each record in the microdata; see, Elliot (2001). Such a measure may help identify those parts of the sample where disclosure risk is high and more protection is needed and may be aggregated in different ways to a file level measure if desired; see, Lambert (1993). While record-level measures may provide greater flexibility and insight when assessing whether specified forms of microdata output are 'disclosive', they are potentially more difficult to estimate than file-level measures.

Skinner and Holmes (1998) propose one approach to the estimation of record-level measures. They restrict attention to *sample unique* records, i.e. records with combinations of values of the key variables which are unique in the microdata sample, on the grounds that these are the records most at risk. They define their measure as the probability of population uniqueness, with probability interpreted with respect to a model. Like Bethlehem et al. (1990), they assume a compound Poisson model for the generation of the frequencies of the values of the key variables, but with a log-normal distribution for the compound error rather than a gamma distribution. Like Fienberg and Makov (1998), they use a log-linear model to capture the dependence on the key variables. After estimating the model parameters, they use numerical integration to compute the measure.

We investigate an alternative approach. We propose a different measure, replacing the probability of population uniqueness by the probability that an observed match between a microdata record and an identifiable unit in the population is correct. This parallels the approach to file-level measures developed by Skinner and Elliot (2002). The estimation of this new measure is discussed in Section 2.1, with particular consideration of how the computations in Skinner and Holmes (1998) can be simplified.

In order to define record-level measures of disclosure risk we make use of the X information available for each record. The file level measures could all be interpreted as probabilities with respect to sampling mechanisms which draw individuals from the population or sample with equal probability. These probabilities are effectively unconditional on the value of X. To obtain record-level measures we propose to condition these probabilities on the values of the key variables defining X. This implies that any two records with the same value of X will have the same measure of disclosure risk. In fact, we shall only consider sample records and restrict attention to records which are sample unique, since these may be expected to be the most risky following Skinner and Holmes (1998), so that all records of interest will have different values of X.

We assume there is no measurement error in X (which could lead to false matches). In this case, there will be F_j individuals in the population which match a specified record with $X = j$. Assuming symmetry of the sampling scheme, as for example for simple random sampling or Bernoulli sampling, the probability that an observed match between this specified record and an individual in the population is correct, conditional on $X = j$ and F_j, is

$$\Pr\left(\text{correct match}|\,\text{unique match},\ X = j, F_j\right) = \frac{1}{F_j}.$$

In practice, F_j will generally be unknown. We therefore consider specifying a model which generates the F_j, $j = 1, ..., J$, and define the record-level

measure of risk for a specified sample unique record with $X = j$ as

$$\begin{aligned} \theta_j &= \Pr\left(\text{correct match} \middle| \text{unique match}, X = j\right) \\ &= \mathrm{E}\left(\frac{1}{F_j}\middle| f_j = 1\right) \end{aligned} \tag{25}$$

This expectation is with respect to both the model generating the F_j and the sampling scheme. The measure θ_j has the same form as the file-level measures θ and θ_s defined in Section 2.1, if the expectation in (25) is replaced by a mean of F_j^{-1} across sample unique records, either with weights proportional to F_j for θ or with equal weights for θ_s. In particular, we may expect that the (unweighted) average of the record-level measures θ_j will approximately equal θ_s. Since $\theta_s \geq \theta$, it follows that if θ is used as a file-level measure, e.g. for the reasons of simplicity of estimation discussed in Skinner and Elliot (2002), this measure will tend to understate the (unweighted) average of the record-level measures of risk θ_j.

To implement the definition of θ_j in practice, we need to specify the model generating the F_j. Following Bethlehem et al. (1990) and other authors, we assume that the F_j are independently Poisson distributed with means λ_j, treated initially as fixed parameters. We assume further, like Skinner and Holmes (1998) that the sampling scheme is such that f_j and $z_j = F_j - f_j$ are independently Poisson distributed as

$$f_j \mid \lambda_j \sim \mathrm{Po}\left(\pi\lambda_j\right) \text{ and } z_j \mid \lambda_j \sim \mathrm{Po}\left[(1-\pi)\,\lambda_j\right]. \tag{26}$$

This is the case, for example, under Bernoulli sampling with selection probability π. It follows that

$$\begin{aligned} \theta_j &= \mathrm{E}\left[\frac{1}{f_j + z_j} \mid f_j = 1, \text{data}\right] \\ &= \mathrm{E}\left[\mathrm{E}\left(\frac{1}{1+z_j} \mid \lambda_j\right) \mid f_j = 1, \text{data}\right]. \end{aligned} \tag{27}$$

It follows from (26) that

$$\begin{aligned} \mathrm{E}\left(\frac{1}{1+z_j} \mid \lambda_j\right) &= \sum_{z=0}^{\infty} \frac{1}{1+z} \frac{\exp\left[-\left(1-\pi\right)\lambda_j\right]\left(\left(1-\pi\right)\lambda_j\right)^z}{z!} \\ &= \frac{1}{\left(1-\pi\right)\lambda_j}\left\{1-\exp\left[-\left(1-\pi\right)\lambda_j\right]\right\}. \end{aligned} \tag{28}$$

If λ_j is fixed then (28) provide an expression for θ_j. If λ_j is random, we obtain from (27) and (28) that

$$\begin{aligned} \theta_j &= \mathrm{E}\left[\frac{1}{\left(1-\pi\right)\lambda_j}\left\{1-\exp\left[-\left(1-\pi\right)\lambda_j\right]\right\} \middle| f_j = 1, \mathrm{data}\right] \\ &= \int \frac{1}{\left(1-\pi\right)\lambda_j}\left\{1-\exp\left[-\left(1-\pi\right)\lambda_j\right]\right\} g\left(\lambda_j \mid f_j = 1\right) d\lambda_j \end{aligned}$$

where $g\left(\lambda_j \middle| f_j = 1\right)$ is the conditional density of λ_j given that $f_j = 1$. We now consider the estimation of θ_j from sample data.

5.1 Estimation of θ_j - Fixed λ_j

We assume that the F_j are unobserved and that the data available to to estimate θ_j consist of the sample frequencies f_j. From (26), these are assumed to be independently Poisson distributed, $f_j \sim \mathrm{Po}\left(\mu_j\right)$, where $\mu_j = \pi\lambda_j$. A log-linear model for the μ_j may be expressed as

$$\log\mu_j = x_j'\beta \tag{29}$$

where x_j is a vector containing specified main effects and interactions for $X_1, ..., X_m$. Such a model may be fitted using standard procedures; see, Agresti (1996), to give an estimated vector $\widehat{\beta}$ and fitted values

$$\widehat{\mu}_j = \exp\left(x_j'\widehat{\beta}\right).$$

From (28) the estimated disclosure risk is

$$\begin{aligned}\widehat{\theta}_j &= \frac{1}{(1-\pi)\,\widehat{\lambda}_j}\left\{1-\exp\left[-(1-\pi)\,\widehat{\lambda}_j\right]\right\} \\ &= \frac{1}{(1-\pi)\,\pi^{-1}\widehat{\mu}_j}\left\{1-\exp\left[-(1-\pi)\,\pi^{-1}\widehat{\mu}_j\right]\right\} \qquad (30)\end{aligned}$$

If a very complex log-linear model is chosen then the resulting $\widehat{\theta}_j$ may either be unstable or not very informative. In the extreme case, if a saturated model is employed, $\widehat{\mu}_j = 1$ for all j and the $\widehat{\theta}_j$ fail to discriminate at all between the sample unique cases. This suggests selecting a simpler log-linear model. The problem then is that, if the model is 'too' simple, the specified u_j may fail to capture all the variation between the μ_j, that is there may be overdispersion. Making allowance for overdispersion in $\widehat{\theta}_j$ is discussed in the next section.

5.2 Estimation of θ_j - Random λ_j (Gamma Distribution)

A common approach to allowing for overdispersion is by introducing a multiplicative error term; see, for example, Cameron and Trivedi (1998) and Agresti (1996). Suppose the distribution of a random count $y = f_j$ is conditionally Poisson, that is

$$y \mid \mu_j \sim \mathrm{Po}\left(\mu_j\right),$$

where now

$$\begin{aligned}\log \mu_j &= x_j'\beta + \varepsilon_j \\ \mu_j &= \exp\left(x_j'\beta + \varepsilon_j\right)\end{aligned}$$

For simplicity, we specify a gamma distribution for $w_j = \exp\left(\varepsilon_j\right)$ as

$$g\left(w; v, b\right) = \frac{b^v}{\Gamma\left(v\right)} w^{v-1}\exp\left(-bw\right), \quad v, b > 0,$$

where $\mathrm{E}(w) = v/b$ and $\mathrm{var}(w) = v/b^2$. To center the distribution of ε_j, the gamma mean is assumed to be one, $v = b$; that is

$$g(w_j \mid v) = \frac{v^v}{\Gamma(v)} w_j^{v-1} \exp(-v w_j). \tag{31}$$

The measure of disclosure risk θ_j is then given by

$$\theta_j = \int_0^\infty \frac{1}{(1-\pi)\pi^{-1} w \phi_j} \left\{1 - \exp\left[-(1-\pi)\pi^{-1} w \phi_j\right]\right\} g(w \mid f_j = 1)\, dw, \tag{32}$$

where $\phi_j = \exp\left(x_j' \beta\right)$.

From Skinner and Holmes (1998) we find that

$$g(w_j \mid f_j = 1) = \frac{\mu_j \exp(-\mu_j)\, g(w_j)}{\int \mu_j \exp(-\mu_j)\, g(w_j)\, dw_j}. \tag{33}$$

Under the gamma model given in (31), we find that the conditional distribution of w_j give $f_j = 1$ is also gamma with parameters $v+1$ and $v+\phi_j$. It follows from (32) and (33) that

$$\theta_j = \frac{\pi(\phi_j + v)}{(1-\pi)\phi_j v}\left[1 - \left(\frac{\phi_j + v}{\pi^{-1}\phi_j + v}\right)^v\right].$$

Suppose now that the Poisson-gamma (negative binomial) model is fitted to the f_j giving estimates $\widehat{v}$ and $\widehat{\beta}$ of the parameters. Let $\widehat{\mu}_j = \widehat{\phi}_j = \exp\left(x_j' \widehat{\beta}_j\right)$ then the estimated value of θ_j is given by

$$\widehat{\theta}_j = \frac{\pi\left(\widehat{\phi}_j + \widehat{v}\right)}{(1-\pi)\widehat{\phi}_j \widehat{v}}\left[1 - \left(\frac{\widehat{\phi}_j + \widehat{v}}{\pi^{-1}\widehat{\phi}_j + \widehat{v}}\right)^{\widehat{v}}\right]$$

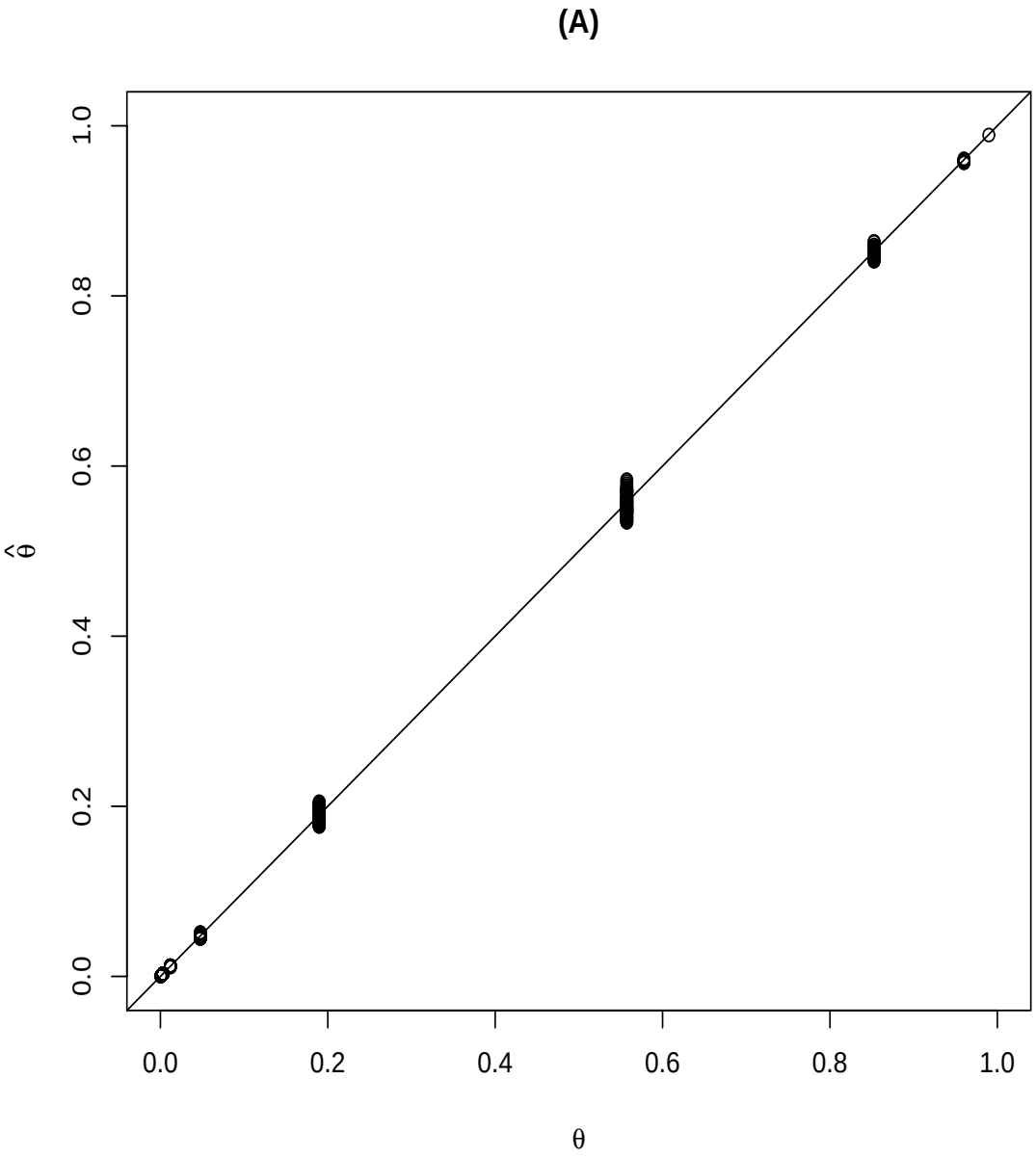

Figure 5: The true θ_j against the estimated $\widehat{\theta}_j$ (Poisson) when $\varepsilon_j = 0$, $N = 50000$, $\pi = 0.20$ and key values $J = 512$ using main effect model.

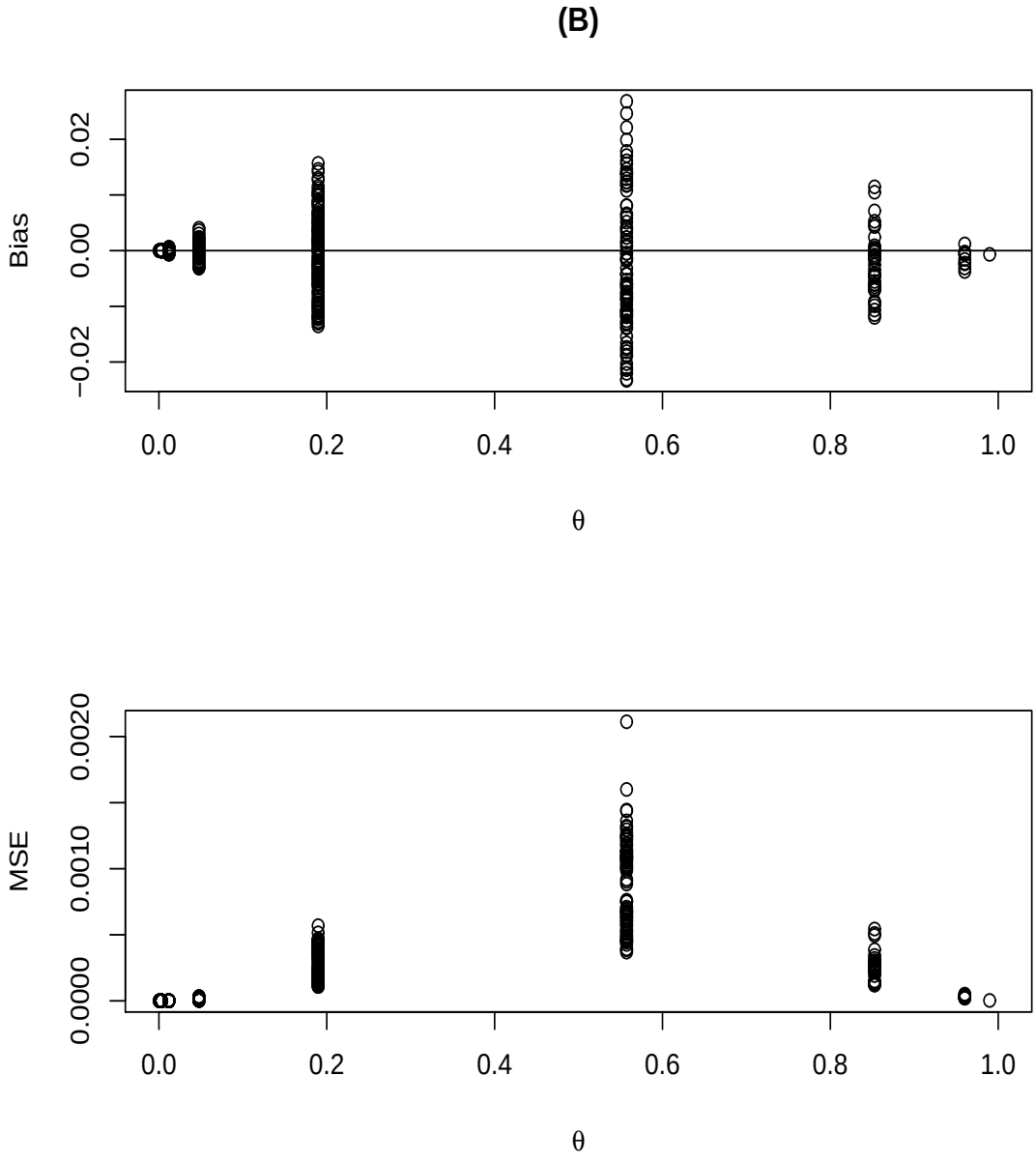

Figure 6: Bias and MSE of the estimated $\widehat{\theta}_j$ (Poisson) when $\varepsilon_j = 0$, $N = 50000$, $\pi = 0.20$ and key values $J = 512$ using main effect model.

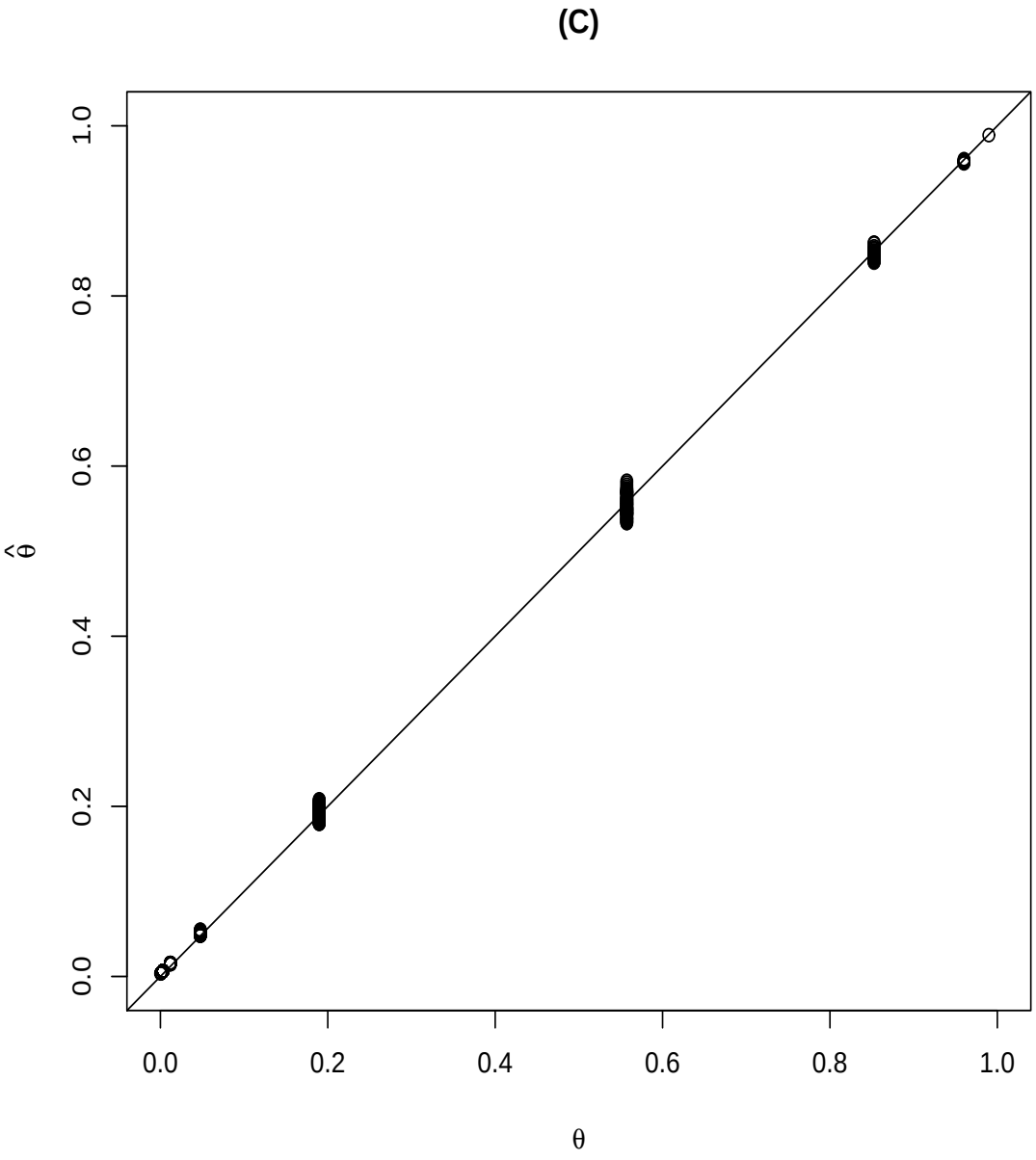

Figure 7: The true θ_j against the estimated $\widehat{\theta}_j$ (Poisson-gamma) when $\varepsilon_j = 0$, $N = 50000$, $\pi = 0.20$ and key values $J = 512$ using main effect model.

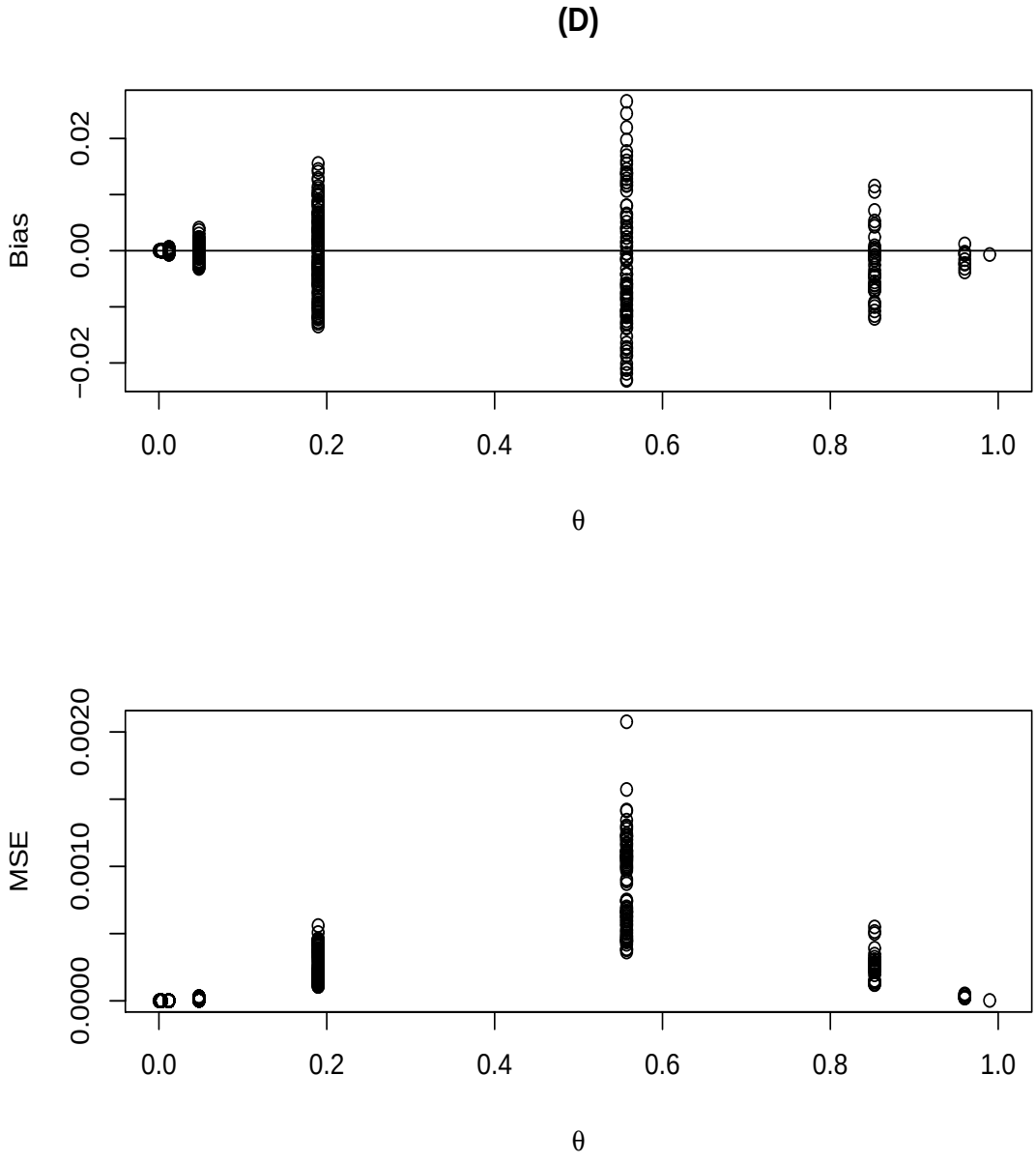

Figure 8: Bias and MSE of the estimated $\widehat{\theta}_j$ when $\varepsilon_j = 0$ (Poisson-gamma), $N = 50000$, $\pi = 0.20$ and key values $J = 512$ using main effect model.

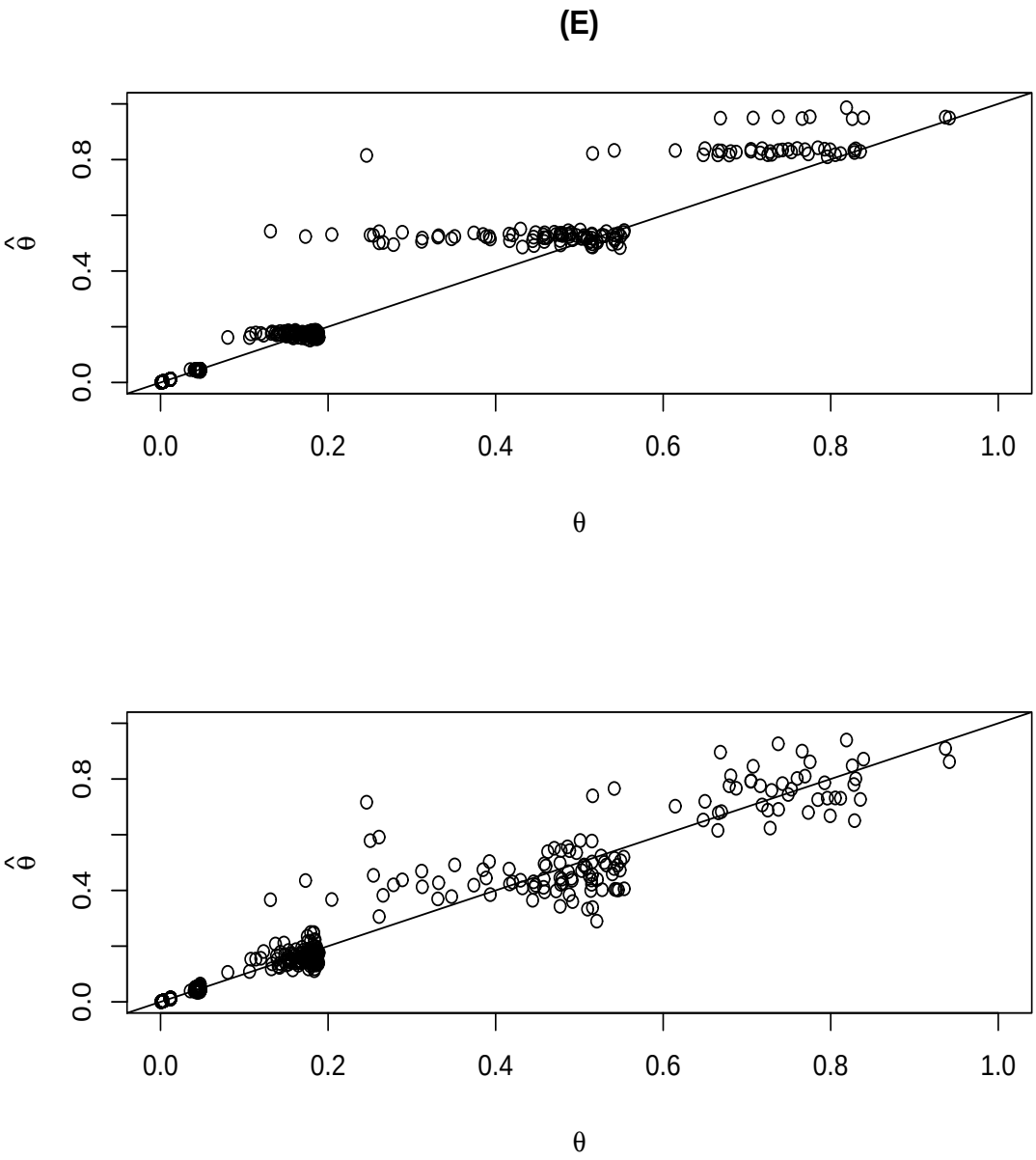

Figure 9: The true θ_j against the estimated $\widehat{\theta}_j$ (Poisson) when ε_j are lognormal ($\mu = 1, \sigma^2 = 3$), $N = 50000$, $\pi = 0.20$ and key values $J = 512$ using (a) main effect and (b) main effect and 2-interaction models.

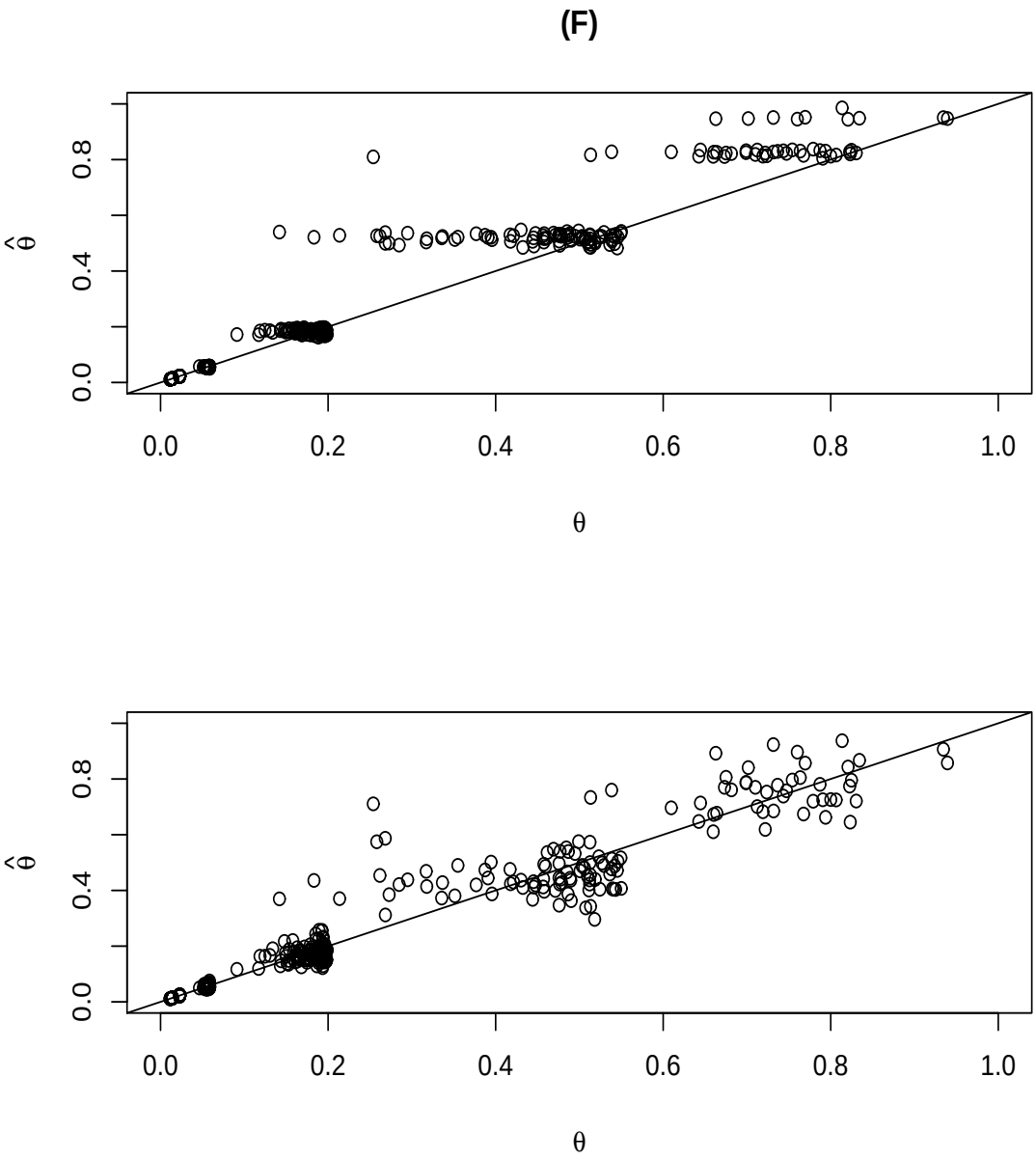

Figure 10: The true θ_j against the estimated $\widehat{\theta}_j$ (Poisson-gamma) when ε_j are lognormal ($\mu = 1, \sigma^2 = 3$), $N = 50000$, $\pi = 0.20$ and key values $J = 512$ using (a) main effect and (b) main effect and 2-interaction models.

5.2.1 Simulation Results

We follow the simulation procedure in Section 4. The results are given in Figures 5, 6, 7, 8, 9 and 10. We note from these Figures the following

1. When the assuming model is correct, $\widehat{\theta}_j$ is a good estimator of θ_j.
2. When the assuming model is not correct, $\widehat{\theta}_j$ is not a good estimator of θ_j although it gives some idea about the risk.
3. In terms of bias, when the assume model is correct, we find that for each sampling fraction the absolute bias is small.

5.3 Estimation of θ_j- Random λ_j (Inverse-Gaussian Distribution)

We now consider an alternative to the gamma distribution. Suppose that w_j follows, IG (γ, δ), the inverse Gaussian distribution with parameters γ and δ and probability density function given by

$$g(w; \delta, \gamma) = \frac{\delta w^{-3/2}}{\sqrt{2\pi}} \exp\left[\delta\gamma - \frac{1}{2}\left(\frac{\delta^2}{w} + \gamma^2 w\right)\right].$$

Then by using the IG (γ, δ) distribution as a mixing density the Poisson-inverse Gaussian distribution arises with mean δ/γ and variance $\delta/\gamma + \delta/\gamma^3$. The distribution has been examined by many authors; see, for example, Sichel (1974), Sichel (1982).

Assuming that E $(w) = 1$ as in Section 5.2 implies that $\gamma = \delta$ so that we could rewrite as

$$g(w; \gamma) = \frac{\gamma w^{-3/2}}{\sqrt{2\pi}} \exp\left[\gamma^2 - \frac{1}{2}\left(\frac{\gamma^2}{w} + \gamma^2 w\right)\right]$$

The ML estimate for the parameter γ of the IG distribution can be obtained given the restriction that $\gamma = \delta$ from a sample of observations y_i as

$$\widehat{\gamma} = n\left(\sum_{i=1}^{n} Y_i^{-1} + \sum_{i=1}^{n} Y_i - 2n\right)^{-1}$$

Then as in Section 5.2 the measure of disclosure risk is

$$\theta_j = \int_0^{\infty} \frac{1}{(1-\pi)\,\pi^{-1} w_j \phi_j} \left[1 - e^{-(1-\pi)\pi^{-1} w_j \phi_j}\right] g\left(w_j \mid f_j = 1\right) dw_j,$$

where

$$\widehat{\phi}_j = \exp\left(x_j' \widehat{\beta}\right)$$

and

$$g\left(w_j \mid f_j = 1\right) = \frac{\phi_j w_j e^{-\pi \phi_j w_j} g\left(w_j\right)}{\int \phi_j w_j e^{-\pi \phi_j w_j} g\left(w_j\right) dw_j}.$$

Under the inverse-Gaussian model we obtain

$$g\left(w_j \mid f_j = 1\right) = \frac{w_j \sqrt{2\pi\phi_j w_j + \gamma^2}}{\gamma \exp\left(\gamma^2 - \gamma\sqrt{2\pi\phi_j w_j + \gamma^2}\right)} g\left(w_j\right).$$

Hence,

$$\theta_j = \int_0^{\infty} \frac{\pi}{(1-\pi)\,\phi_j w_j} \left[1 - e^{-\frac{(1-\pi)}{\pi}\phi_j w_j}\right] \frac{w_j \sqrt{2\pi\phi_j + \gamma^2}}{\gamma \exp\left(\gamma^2 - \gamma\sqrt{2\pi\phi_j + \gamma^2}\right)} g\left(w_j\right) dw_j.$$

The measure of disclosure risk is given by

$$\theta_j = \frac{\sqrt{2\pi\phi_j + \gamma^2}\left[1 - \exp\left[\gamma\left(\sqrt{2\pi\phi_j + \gamma^2} - \sqrt{2\phi_j + \gamma^2}\right)\right]\right]}{\gamma\left(1-\pi\right)\phi_j}$$

which allows that the mean of y_j is given by ϕ_j (parameterized using the exponential function to restrict the outcomes to positive counts) . Then we have

$$\widehat{\theta}_j = \frac{\sqrt{2\pi\widehat{\phi}_j + \widehat{\gamma}^2}\left[1 - \exp\left[\widehat{\gamma}\left(\sqrt{2\pi\widehat{\phi}_j + \widehat{\gamma}^2} - \sqrt{2\widehat{\phi}_j + \widehat{\gamma}^2}\right)\right]\right]}{\widehat{\gamma}\left(1-\pi\right)\widehat{\phi}_j}$$

where $\widehat{\gamma}$ is to be estimated from the data as above.

5.3.1 Suggested Algorithm

We suggest the following algorithm for Poisson-inverse Gaussian regression

1. Find estimates of $\left(\gamma_{(k)}, \beta_{(k)}\right)$ where γ is the parameter of the mixing distribution and β are the regression coefficient (Poisson fitting).

2. Calculate
$$\phi_j = \exp\left(\beta_{(k)} x_j\right)$$
by using these values, obtain the pseudovalues
$$h_i = \mathrm{E}\left(\theta_i \mid y_i, x_i, \gamma_{(k)}, \beta_{(k)}\right) = \frac{\left(y_i + 1\right) P\left(y_i + 1\right)}{\phi_i P\left(y_i\right)}$$
and
$$s_i = \mathrm{E}\left(\theta_i^{-1} \mid y_i, x_i, \gamma_{(k)}, \beta_{(k)}\right) = \begin{cases} \frac{\frac{\phi_i P(y_i - 1)}{y_i P(y_i)}}{\gamma^{(m)}\sqrt{\gamma_{(k)}^2 + 2\phi_i + 1}} & y_i \neq 0 \\ \frac{\gamma^{(m)}\sqrt{\gamma_{(k)}^2 + 2\phi_i + 1}}{\gamma_{(k)}^2} & y_i = 0 \end{cases}$$
where $P\left(y_i\right)$ denotes the $P - IG\left(\phi_j, \gamma_{(k)}, \gamma_{(k)}\right)$ probability function

3. Update the regression parameters β using the pseudovalues h_i as the offset values.

4. Update γ with
$$\gamma_{(k+1)} = n\left(\sum s_i + \sum w_i - 2n\right)^{-1}$$

5. If the criterion is satisfied then stop iterating, else go back to step 1.

6 Record level measure with misclassification

As in Section 3, let $\widetilde{X}$ denote the combination of values as measured by a potential intruder using external information and X as measured in the microdata. Take a unit i in the microdata with key value $X(i) = j$. Suppose $f_{X(i)} (= f_j) = 1$ and suppose there are $\widetilde{F}_j$ units with $\widetilde{X} = j$ in the population. Suppose we find a unit at random in the population with $\widetilde{X} = j$. Let A_j be the set of units in the population with $\widetilde{X} = j$. So the number of units in A_j is $\widetilde{F}_j$. Then

$$\Pr(\text{match correct}) = \Pr(i \in A_j) \Pr(\text{match correct} | \, i \in A_j)$$

Since

$$\Pr(\text{match correct} | \, i \notin A_j) = 0$$

We have

$$\Pr(i \in A_j) = M_{jj} \quad \Pr\ (i \text{ classified correctly as } j)$$

Then, if $\widetilde{F}_j$ is known and treated as fixed

$$\Pr(\text{match correct} | \, i \in A_j) = \frac{1}{\widetilde{F}_j} \quad \text{(sample units drawn with equal prob.)}$$

Hence

$$\theta_j = \Pr(\text{match correct}) = \frac{M_{jj}}{\widetilde{F}_j}$$

If $\widetilde{F}_j$ is unknown then we write

$$\theta_j = \mathrm{E}\left(\frac{M_{jj}}{\widetilde{F}_j} \middle| f_j = 1, \text{data} \right) = M_{jj} \mathrm{E}\left(\frac{1}{\widetilde{F}_j} \middle| f_j = 1, \text{data} \right)$$

As an approximation, when the elements in the main diagonal of misclassification matrix are high, we have

$$\mathrm{E}\left(\frac{1}{\widetilde{F}_j} \mid f_j = 1, ...\right) \simeq \mathrm{E}\left(\frac{1}{F_j} \mid f_j = 1, ...\right)$$

Based on this approximation, we defined the following measures

$$\theta_j^M \simeq M_{jj}\mathrm{E}\left(\frac{1}{F_j}\middle| f_j = 1, \mathrm{data}\right)$$

Using this approximation we have for fixed λ_j case in Section 5.1

$$\theta_j^M = \frac{M_{jj}}{(1-\pi)\,\lambda_j}\left(1 - e^{-(1-\pi)\lambda_j}\right)$$

For the negative binomial model in Section 5.2, we have

$$\theta_j^M = \frac{M_{jj}\,(\pi\lambda_j + v)}{(1-\pi)\,v\lambda_j}\left[1 - \left(\frac{\pi\lambda_j + v}{\lambda_j + v}\right)^v\right]$$

and for the Inverse-Gaussian distribution in Section 5.3, we have

$$\theta_j^M = \frac{M_{jj}\sqrt{2\pi\lambda_j + \gamma^2}\left[1 - \mathrm{e}^{\gamma\left(\sqrt{2\pi\lambda_j+\gamma^2} - \sqrt{2\lambda_j+\gamma^2}\right)}\right]}{\gamma\,(1-\pi)\,\lambda_j}$$

Figure 11 shows the relation between the measure $M_{jj}/\widetilde{F}_j$ and the approximation M_{jj}/F_j from the Poisson-gamma model. We can see that when there is no misclassification both measures have the same form $1/F_j$. When the misclassification is not too serious, M_{jj}/F_j gives a good approximation to $M_{jj}/\widetilde{F}_j$, but when the misclassification is high we find that M_{jj}/F_j is not a good approximation to $M_{jj}/\widetilde{F}_j$.

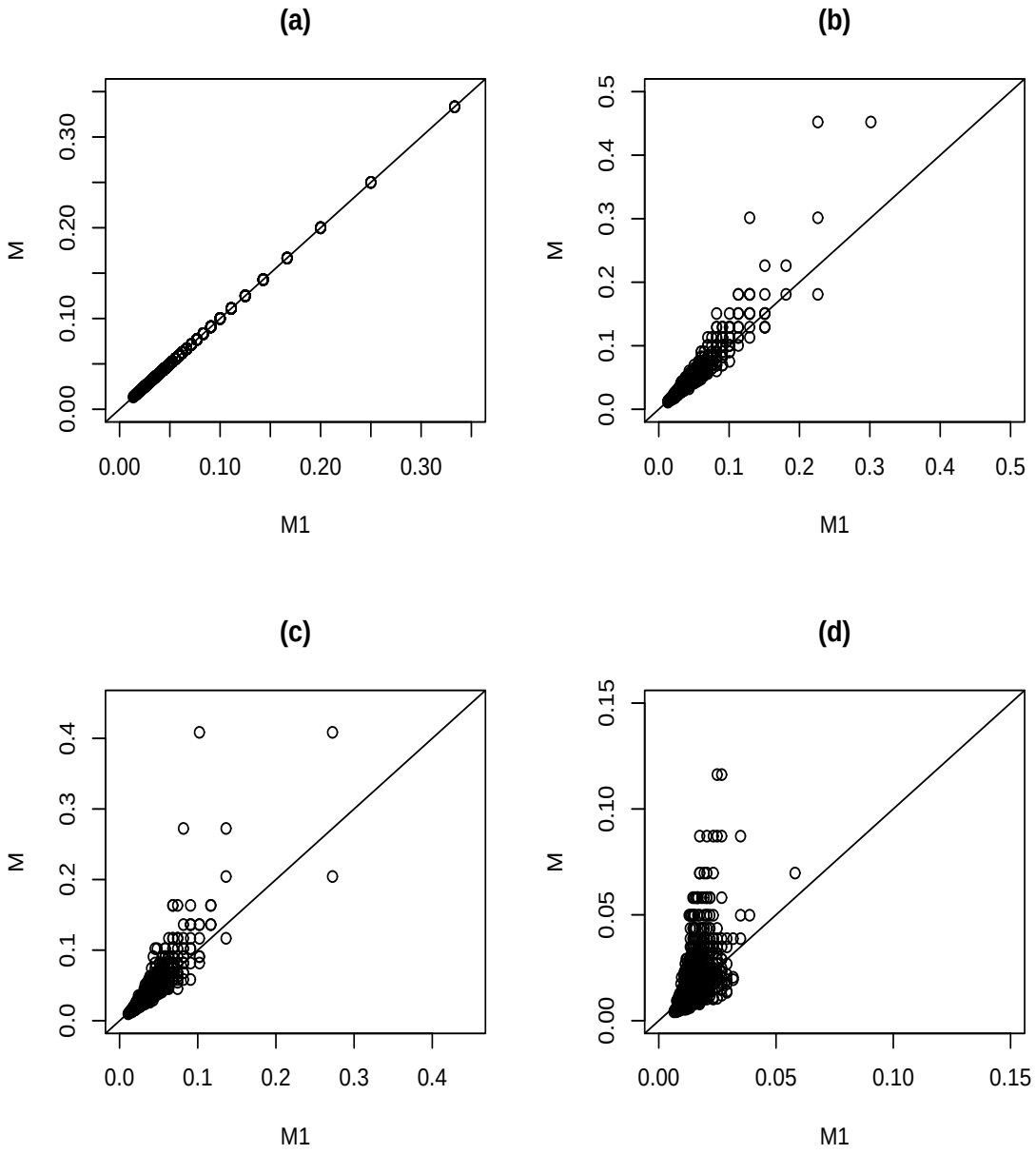

Figure 11: The true measure $M_1 = M_{jj}/\widetilde{F}_j$ and the approximation $M = M_{jj}/F_j$ from Poisson-gamma model $\alpha = 3.61$, $\beta = 4.99$ when (a) $M_{jj} = 1$, (b) $M_{jj} = 0.90$, (c) $M_{jj} = 0.80$ and (d) $M_{jj} = 0.35$

7 Conclusion

In this report, we have discussed measures of disclosure risk for microdata at both the file level and the record level. We have shown how the record-level measure of disclosure risk of Skinner and Holmes (1998), defined in terms of the probability of population uniqueness, may be extended in a parallel way to the development in Skinner and Elliot (2002) a record-level measure of the probability that an observed match is correct. Both measures depend on the specification of a log-linear model for an assumed set of key variables. In an empirical evaluation of different versions of the new record-level measure using real survey data, we found evidence of discrimination by the measure between records of different levels of risk, in particular records which are very likely to be population unique could be identified by consideration of records with high values of the measure. We have also extended both file-level and record-level measures to the case of misclassification.

References

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. New York: Wiley.

Bethlehem, J. G., W. J. Keller, and J. Pannekoek (1990). Disclosure control of microdata. *Journal of the American Statistical Association 85*, 38–45.

Blien, U., H. Wirth, and M. Muller (1992). Disclosure risk for microdata stemming from official statistics. *Statistica Neerlandica 46*, 69–82.

Cameron, C. A. and P. K. Trivedi (1998). *Regression Analysis of Count Data*. Cambridge.

Duncan, G. and D. Lambert (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statstics 7*, 207–217.

Elliot, M. (2001). Disclosure risk assessment. P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz eds. In *"Confidentiality Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies"*, North-Holland, pp. 75–90.

Fienberg, S. and U. Makov (1998). Confidentiality, uniqueness and disclosure limitation for categorical data. *Journal of Official Statistics 14*, 385–397.

Fuller, W. A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics 9*, 383–406.

Greenberg, B. and L. Voshell (1990). Relating risk of disclosure for microdata and geographic area size. *Proc.Sect.Survey Res.Meth., Am.Statist.Ass.*, 450–455.

Kuha, J. and C. Skinner (1997). Categorical data analysis and misclassification. In *L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz and D. Trewin eds. "Survey Measurement and Process Quality"*, New York: Wiley, pp. 633–669.

Lambert, D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics 9*, 313–331.

Marsh, C., C. Skinner, S. Arber, P. B., S. Openshaw, J. Hobcraft, D. Lievesley, and J. Walford (1991). The case for a sample of anonymized records from the 1991 census. *Journal of the Royal Statistical Society, Ser. A 154*, 305–340.

Paass, G. (1988). Disclosure risk and disclosure avoidance for microdata. *Journal of Business and Economic Statistics 6*, 487–500.

Samuels, S. (1998). A Bayesian, species-sampling-inspired approach to the uniques problems in microdata disclosure risk assessment. *Journal of Official Statistics 14*, 373–383.

Sichel, H. S. (1974). On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society, Series A 137*, 52–34.

Sichel, H. S. (1982). Repeat-buying and the generalized inverse gaussian-poisson distribution. *Applied Statistics 31*, 193–204.

Skinner, C. and M. Elliot (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society, Series B 64*, 855–867.

Skinner, C. and D. Holmes (1993). Modelling population uniqueness. In *Proceedings of the International Seminar on Confidentiality*, Dublin, pp. 175–199.

Skinner, C. and D. Holmes (1998). Estimating the re-identification risk per record in microdata. *Journal of Official Statistics 14*, 361–372.

Skinner, C., C. Marsh, S. Openshaw, and C. Wymer (1994). Disclosure control for census microdata. *Journal of Official Statistics 10*, 31–51.