Re-Examining File-Level Re-Identification Risk Assessment for Survey Microdata

Lin Li[1], Jianzhu Li[1], Tom Krenzke[1], Natalie Shlomo[2]
[1]Westat, 1600 Research Blvd., Rockville, MD 20850
[2]University of Manchester, Manchester, United Kingdom

Abstract
In this paper, we discuss some practical issues encountered when estimating file-level disclosure risk measures of re-identification in survey microdata. We typically use the log-linear modeling approach (Skinner and Shlomo, 2008) to estimate disclosure risk in survey microdata files. Several challenges emerge that relate to satisfying goodness-of-fit (GOF) criteria of the log-linear models in the presence of model assumption violations, and handling large numbers of variables. In the former, we ran simulations to explore the accuracy of estimating risk based on the GOF criteria particularly for the case of complex survey designs and differential survey weights. For the latter, we provide guidance for variable selection with insights on how to proceed with the risk assessment and provide meaningful results. We used the National Science Foundation's Survey of Doctorate Recipients data as a case study. The results of evaluating the disclosure risk estimates under several approaches lead to guidance for a sensitivity analysis that helps to provide for a better estimate of file-level re-identification risk in survey microdata.

Key Words: data confidentiality, disclosure, log-linear models, goodness of fit, sensitivity analysis

1. Introduction

Statistical agencies are obligated to protect the respondents' identities when they release survey microdata to the public. The survey microdata often go through statistical disclosure control treatment (e.g., recoding) before being released. To determine whether the treated data is safe enough to be released, agencies need to assess the re-identification risk of the data file. Skinner and Shlomo (2008) investigated the use of the log-linear modeling method in risk assessment. Westat has adopted the log-linear modeling method to assess re-identification risk.

Several challenges emerge that relate to satisfying goodness-of-fit (GOF) criteria of the log-linear models in the presence of model assumption violations, and handling large numbers of variables. The purpose of this paper is to discuss the practical issues encountered and investigate ways to address the challenges. Section 2 provides the theoretical background on the log-linear modeling method. Section 3 uses a case study to illustrate the practical issues encountered when using the log-linear modeling approach, and provides practical guidance for variable selection as well. In Section 4, we explore the accuracy of risk estimates based on the GOF criteria through a simulation study. Finally, Section 5 summarizes the findings from the simulation study and provides guidance to achieve a better estimate of file-level re-identification risk in survey microdata.

2. Re-Identification Risk Assessment with Log-Linear Models

Hundepool, et al. (2012) discuss in a microdata context that a re-identification operation is achieved by an intruder when comparing a target individual in a sample with an available list of units (external file) that contains individual identifiers (e.g., name and address), plus a set of identifying variables. Re-identification occurs when the unit in the released file and a unit in the external file are linked and belong to the same individual in the population. The risk also exists since a "nosy neighbor" may know a handful of facts about a person and could search the file to find the person.

Individual re-identification risk is the probability that the microdata record indeed belongs to a known unit. Many cases are uniquely identified (referred to as "sample uniques") by a relatively small number of variables having specific values, variables (referred to as "key variables" hereafter) that may exist in external files or easily known about an individual. Skinner and Shlomo (2008) proposed two risk measures as follows.

- Expected number of sample uniques that are population uniques:

$$\tau_1^* = \sum_k p(F_k = 1 | f_k = 1)$$

- Expected number of correct matches for sample uniques:

$$\tau_2^* = \sum_k E(1/F_k | f_k = 1)$$

where k refers to cells formed by key variables, f_k is the sample frequency in cell k, and F_k is the population frequency in cell k.

In practice, F_k needs to be estimated. Researchers have investigated the use of models to provide more stable estimates of risk. Skinner and Shlomo (2008) provide an improved risk measure using log-linear models. They assume that the population count in cell k is a realization of independent Poisson random variables with mean λ_k. That is, $F_k \sim Poisson(\lambda_k)$, and the sample is drawn by Bernoulli sampling where individuals in cell k have the same known inclusion probability π_k so that the sample counts f_k are independent Poisson random variables $f_k \sim Poisson(\pi_k \lambda_k)$. Under the Bernoulli sampling assumption, they have $F_k \mid f_k \sim Poisson[\lambda_k(1-\pi_k)] + f_k$. Both $\lambda_k$and π_k are needed in order to estimate F_k. To obtain estimates of λ_k, log-linear models are fit on the observed sample counts in cells formed by key variables and their interactions. π_k is assumed known. However, π_k often needs to be estimated using sampling weights (e.g., π_k can be estimated by the overall sampling rate $\hat{\pi} = \sum_k f_k / \sum_k \sum_i w_{ki}$, or the cell sampling rate $\hat{\pi}_k = f_k / \sum_i w_{ki}$, where w_{ki} is the sampling weight for case i in cell k).

Since risk estimates may be sensitive to the model specification, Skinner and Shlomo (2008) proposed a GOF criterion to check the adequacy of the specification. They provided four types of GOF criteria, depending on the choice of risk measure and the choice of variance estimator. They found that models that "work" for one risk measure (τ_1^* or τ_2^*) also tend to work for the other risk measure. In our paper, we focus on the risk measure τ_1^* and the GOF criteria ($\hat{B}_1 / \sqrt{V_R}$) for simplicity. The GOF statistic tends to decrease as more

interaction terms are added to the model. If the model is underfit, GOF tends to be positive and the risk is overestimated. Skinner and Shlomo proposed using the closeness of GOF to zero as evidence of an absence of underfitting. In many empirical experiments that they undertook, they found that the independence log-linear model tends to underfit and leads to overestimation of risk measures, and the all three-way interaction model tends to overfit and leads to underestimation of risk measures. Skinner and Shlomo found that the all two-way interaction log-linear model often leads to good estimates of the risk measures. As a practical approach, they suggested first computing the GOF for the independence model and the all two-way interaction model. If the latter model shows no sign of underfitting, they suggested starting with the independence model and adding the two-way interaction terms for different pairs of key variables, chosen sequentially to reduce GOF, until a model is identified that shows no evidence of underfitting. We have adopted their log-linear modeling approach to estimate disclosure risk in survey microdata files. In the following section, we discuss some practical issues we encountered through a case study on the Survey of Doctorate Recipients (SDR) data.

3. Case Study on the SDR Data

We conducted a re-identification risk assessment on the 2017 SDR Public Use File (PUF). The SDR is a longitudinal survey conducted approximately every 2 years that is designed to provide demographic and career history information about individuals who earned a research doctoral degree in a science, engineering, or health (SEH) field from a U.S. academic institution. SDR follows a sample of individuals with SEH doctorates throughout their careers from the year of their degree award until age 76. The panel is refreshed each survey cycle with a sample of new SEH doctoral degree earners. For the 2017 SDR, all 2015 sample members who remained age eligible for the survey were retained, and a sample of new graduates were added. The new graduates sample was selected using a stratified sample design, where the strata were defined by fine fields of study. Within each stratum, a random sample was selected systematically with probability proportional to size to oversample underrepresented racial and ethnic minorities in the SDR population. The resulting 2017 SDR sample consists of 124,580 people and 85,739 respondents. The overall sampling rate was about 11 percent, although sampling rates varied greatly across strata. Consequently, the sampling weights had a large variation with a coefficient of variation of over 100. We note the risk assessment in this illustration treats the dataset as cross-sectional and does not account for the risk from the longitudinal nature of the data.

3.1 Variable Selection

The first step in estimating risk is to determine the key variables to be included in the log-linear models. There are a large number of indirect identifying variables available in the 2017 SDR PUF. We selected eight variables from the large pool to be included in the model. Some of the categories were combined since they may not be distinguishable or may not convey useful information (e.g., “unknown”, “other”, etc.). The average cell size is 1.44[1] by crossing the eight key variables.

In the following, we offer some practical recommendations on how to select variables in order to provide meaningful risk estimates. First, since risk estimates can be sensitive to

[1] The average cell size was computed based on all cells, including those with no samples.

the number of variables and levels of each variable used in the model, a reasonable assumption is needed for the level of information available to data intruders. Include indirect identifiers that can be relatively easily obtained by intruders from external sources, such as geographical variables, demographic variables (age, sex, race/ethnicity, …) and sensitive attributes (disability, income, …). Review the indirect identifiers and combine the categories that may not be distinguishable or may not convey useful information (e.g., combine the categories such as "unknown" and "others", recode continuous age into 2-year intervals, etc.). If there are design variables that lead to large variations in selection probabilities, they should be included in the key variables as well.

Since including too many variables may cause overestimation of the risk, the number of variables can be reduced by choosing one variable to represent a group of similar variables (e.g., pick one out of year of highest degree, year of most recent degree, and year of first degree). If, after risk assessment, it is decided that the risk is too high and some treatment needs to be applied to a chosen variable, the same treatment needs to be applied to similar variables as well.

After deciding on the key variables, check the average cell size in a cross-tabulation of all the selected key variables. Make sure the cell size is not too small. Collapse cell categories if necessary and if it still reflects what will be released from a risk assessment point of view.

3.2 Model Fitting

As mentioned earlier, the SDR 2017 has a complex sample design that violates the log-linear model assumptions on independent selection and equal inclusion probabilities within cells. If the design strata were available to be used in the key variables, it would help mitigate the violations. However, since the SDR 2017 PUF consists of samples from multiple panels across years that were not selected with the same stratification scheme, it is not possible to include such a variable. The within-cell selection rate π_k is also unknown and difficult to estimate due to the large variation in sampling weights.

We fit the log-linear models with eight key variables (e.g., age, sex, race/ethnicity, etc.). An unusually high level of interaction terms were needed to satisfy the GOF criteria of the log-linear models, although other research and our experience showed that all two-way interactions are often sufficient. As shown in Figure 1, the GOF statistic is close to six even with all five-way interactions in the model. This raises a questions, "Is the GOF statistic still a good guide in model fitting when there are violations on assumptions?" In the following section, we present a simulation study as an effort to answer this question.

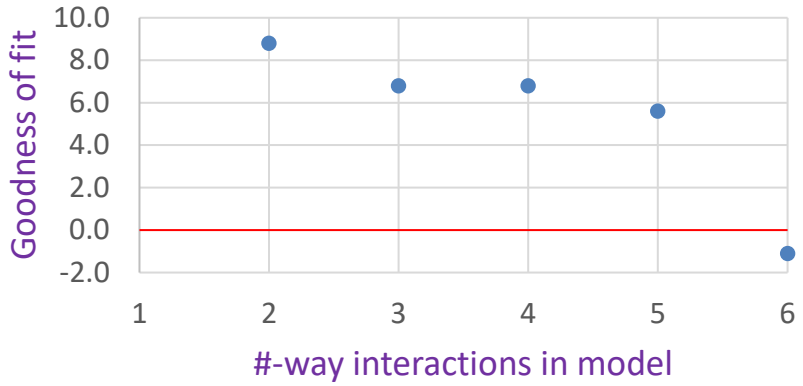

Figure 1: The goodness-of-fit statistic by the level of interaction terms in the log-linear model for the Survey of Doctorate Recipients 2017 Public Use File

4. Simulation Study

The purpose of the simulation study is to explore the accuracy of risk estimates based on the GOF criteria particularly for the case of complex survey designs and differential survey weights. We describe the simulation settings in Section 4.1, and discuss the results for the nine simulation scenarios in turn in Sections 4.2.1 – 4.2.3.

4.1 Simulation Setup

We took the SDR 2017 PUF as the population and drew a 1 percent sample (sample size=840) for 1,000 samples by simple random sampling (SRS) and stratified sampling. The SRS sample aligned well with the log-linear model assumptions and was included for verification. The stratified sample was selected in two different fashions: one defined strata using two of the key variables, the other defined strata without using any key variables. The sampling weights range from about 2 to over 300 with a coefficient of variation of about 100 percent for each of the samples. We fit log-linear models with the counts in cells formed by cross-classifying eight key variables.[2] The models are fit in two ways: using sample counts or weighted counts. For the outcome, we examined the GOF statistic against bias of risk estimates. The GOF statistic is the $\hat{B}_1/\sqrt{V_R}$ proposed in Skinner and Shlomo (2008). The risk is measured as the number of sample uniques that are also population uniques (τ_1^*). Bias is the difference between the risk estimated from the fitted log-linear model and the true risk based on the information from the simulation sample and population, which is the 2017 SDR PUF. A positive value for bias indicates overestimation of risk, and vice versa. The risk estimates and GOF were computed in two ways: estimating π_k with the overall sampling rate $\hat{\pi}$ or the cell sampling rate $\hat{\pi}_k$ as discussed in Section 2. In total, there are nine simulation scenarios as listed in Table 1 below.

[2] The eight key variables are the same as those used in the risk assessment in Section 3.

Table 1: Simulation Scenarios

Scenario	*Selection method*	*Stratifier in key variables*	*Weighted counts*	*Compute outcome statistics with $\hat{\pi}$ or $\widehat{\pi}_k$*
0	SRS	N/A	Yes	$\hat{\pi} = \widehat{\pi}_k$
1a	Stratified sampling	Yes	Yes	$\hat{\pi}$
1b		Yes	Yes	$\widehat{\pi}_k$
1c		Yes	No	$\hat{\pi}$
1d		Yes	No	$\widehat{\pi}_k$
2a		No	Yes	$\hat{\pi}$
2b		No	Yes	$\widehat{\pi}_k$
2c		No	No	$\hat{\pi}$
2d		No	No	$\widehat{\pi}_k$

4.2 Simulation Results

For simulation results, we examine the GOF statistic against the bias of risk estimates for the 1,000 samples in each scenario. The simulation results for the SRS sample (Scenario 0), stratified sample with stratifiers in the key variables (Scenarios 1a – 1d), and stratified sample without stratifiers in the key variables (Scenarios 2a – 2d) are discussed in the following three subsections in turn.

4.2.1 Scenario 0 results

For Scenario 0 (SRS sample), the simulation results for models with main effects only and all two-way interactions are shown in Figure 2. It shows that for the independence model, all samples have both positive GOF and positive bias. The value of bias ranges roughly from 20 to 60 (i.e., the estimated number of sample uniques being also population uniques is about 20 to 60 more than the actual number of such people). For the all two-way interaction model, the majority of samples have both negative GOF and negative bias, although a small proportion of the samples have positive GOF and negative bias. Since SRS sample meets the assumptions of the log-linear models, the GOF performs as it should in general. The positive GOF for the independence model and negative GOF in most of the samples for all two-way-interaction models suggest that the best model lies between the two extremes. If we were to conduct the simulation to select significant two-way interaction terms, we suspect the GOF and bias would both move closer to zero.

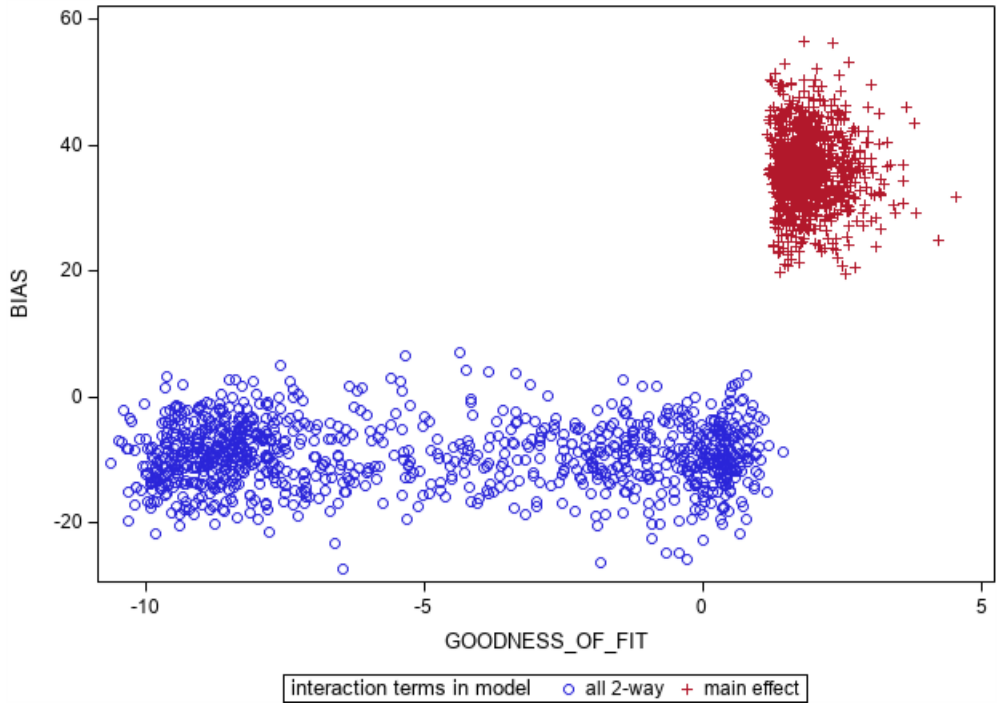

Figure 2: Bias of risk estimates by goodness-of-fit statistic for simple random sampling sample

4.2.2 Scenarios 1a-1d results

For scenarios 1a-1d we selected stratified samples and included two stratifiers in the eight key variables for log-linear models. We looked at the independence model, which showed clear signs of underfitting. Figure 3 shows the simulation results for the all two-way interaction models for the four scenarios. As can be seen, for scenarios 1a (using $\hat{\pi}$ and weights) and 1b (using $\widehat{\pi}_k$ and weights), almost all of samples have both positive GOF and positive bias. For these two scenarios, because most of the samples have GOF close to or greater than 2, the models are underfit and we will look at models with all three-way interactions later. For scenarios 1c (using $\hat{\pi}$ and sample counts) and 1d (using $\widehat{\pi}_k$ and sample counts), about a quarter of the samples have positive GOF (indicating positive bias) although their bias is actually negative, which indicates that GOF may be misleading. The magnitude of underestimation is more severe for scenario 1c (using $\hat{\pi}$) than scenario 1d (using $\widehat{\pi}_k$).

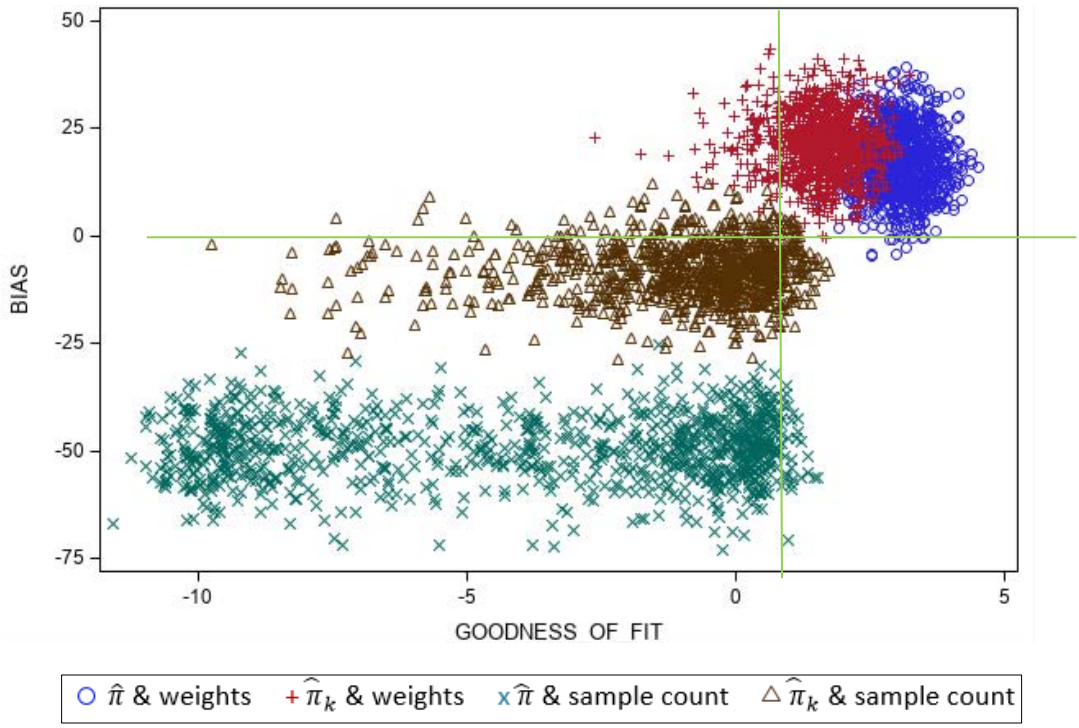

Figure 3: Bias of risk estimates by goodness of fit statistic for stratified sample with models including stratifiers in key variables – all two-way interactions

Since the all two-way interaction models are still underfitted for scenarios 1a and 1b, we fit all three-way interaction models and show the results in Figure 4. As can be seen in the plot, although nearly all of the bias is negative, the GOF is positive (indicating positive bias) for all samples in scenario 1a (using $\hat{\pi}$ and weights) and about a quarter of samples in scenario 1b (using $\widehat{\pi}_k$ and weights). This shows again that GOF may be misleading when the model assumptions are violated. We are most concerned about the situation where the GOF is positive while the risk is actually underestimated. We recommend to be conservative in risk estimation and conduct sensitivity analysis on the risk estimates (e.g., making risk estimates using some of the scenarios in this simulation). In addition, similar to Figure 3, the scenario using $\hat{\pi}$ (1a) performs worse than the scenario using $\widehat{\pi}_k$ (1b) (i.e., many more samples in scenario 1a have misleading signs for GOF than scenario 1b). This is not surprising since $\widehat{\pi}_k$ is a more accurate estimator of cell sampling rate than $\hat{\pi}$ when stratifiers are used in the models. Also for Scenario 1b, the majority of GOF is negative, indicating the all three-way interaction model is overfitted. Satisfactory results are more likely if we were to use only the significant three-way interaction terms for this scenario.

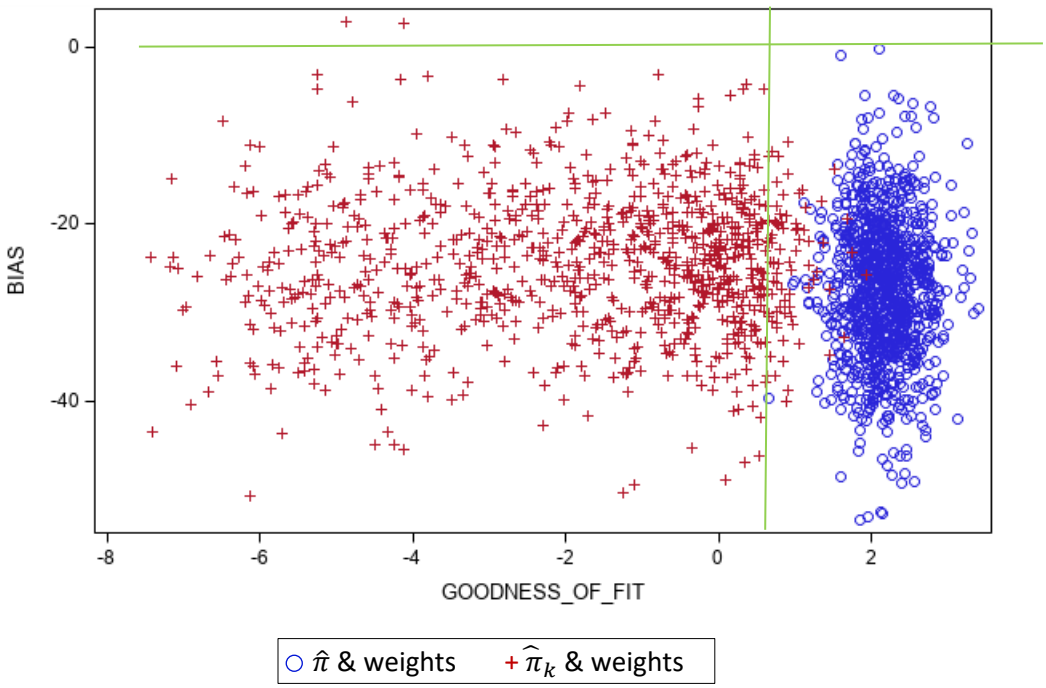

Figure 4: Bias of risk estimates by goodness-of-fit statistic for stratified sample with models including stratifiers in key variables – using weights and all three-way interactions

4.2.2 Scenarios 2a-2d results

For scenarios 2a-2d, we selected stratified samples but did not include the stratifiers in the eight key variables for log-linear models. Figure 5 shows the simulation results for models with all two-way interactions for them. In all four scenarios, the majority of samples have positive GOF corresponding to positive bias, which is good. Since it is unclear from the plot which scenario performs better, we computed the percentage of samples with GOF consistent with bias (i.e., positive GOF corresponds to positive bias, vice versa) as shown in Table 2. The percentage for scenarios using $\widehat{\pi}_k$ (2b and 2d) is smaller than those using $\hat{\pi}$ (scenarios 2a and 2c), which is the opposite from what we saw in scenarios 1a-1d. Because these scenarios did not include stratifiers in the key variables, $\widehat{\pi}_k$ is not a reliable estimate of cell sampling rate due to differential weights and small cell size.

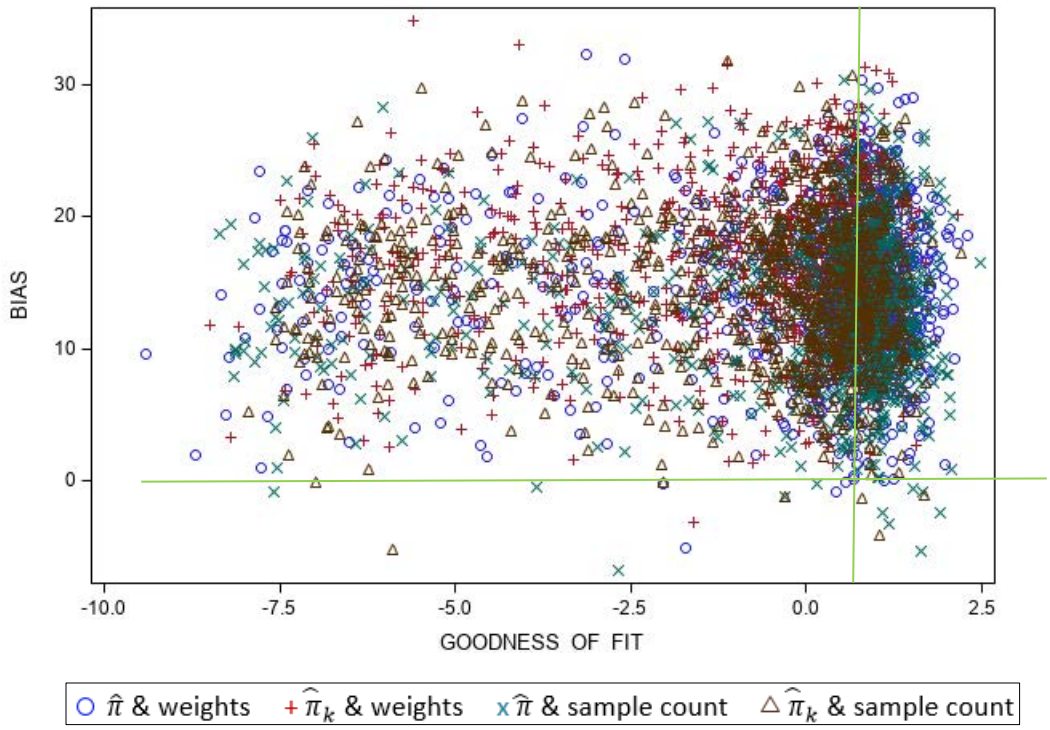

Figure 5: Bias of risk estimates by goodness-of-fit statistic for stratified sample without including stratifiers in key variables – all two-way interactions

Table 2: Scenarios 2a-2d: Percentage of Samples with Goodness of Fit Consistent with Bias for Models with All Two-Way Interactions

Scenario	*Stratifier in key variables*	*Weighted counts*	$\hat{\pi}$ *or* $\widehat{\pi}_k$	*Percent of samples with GOF consistent with bias*
2a	No	Yes	$\hat{\pi}$	73%
2b	No	Yes	$\widehat{\pi}_k$	50%
2c	No	No	$\hat{\pi}$	73%
2d	No	No	$\widehat{\pi}_k$	52%

5. Summary

We have discussed the practical issues encountered when assessing the file-level re-identification risk using the log-linear modelling approach. We provided practical guidance for selecting key variables for the log-linear models. We also explored the accuracy of risk estimates based on the GOF criteria through a simulation study particularly for the case of a complex survey design and differential survey weights.

The simulation results showed that when model assumptions are violated, the GOF criteria may be misleading, and it may lead to underestimation of risk. It would be helpful to check the robustness of the risk estimates through sensitivity analysis. The simulation study itself illustrated a possible sensitivity analysis by using four different approaches to estimate cell

sampling rate and fit the model, and including more or fewer interaction terms in the model. Given the importance of protecting respondents' identities and complex sample designs that violate the risk measure assumptions, we recommend conducting sensitivity checks and being conservative in risk assessment.

The simulation results confirmed that when stratifiers were included in the log-linear models, $\widehat{\pi}_k$ estimated cell sampling rate more reliably than $\hat{\pi}$. However, when stratifiers were not included in the model, $\widehat{\pi}_k$ might perform worse than $\hat{\pi}$ due to large variation in sampling weights and small cell size. For stratified samples, stratifiers should be included in the log-linear models whenever it is available, as it would help to achieve a better estimate of cell sampling rate and satisfy the model assumption on equal sampling rate within cells.

As mentioned in Section 2, this paper is limited to one of the two risk measures and one of the four GOF statistics. It would be interesting to explore whether the other risk measure and the other three GOF statistics follow the same pattern when using GOF to guide the model selection to obtain accurate risk estimates. In addition, our simulation did not search for significant interaction terms for the log-linear models due to the limit of time. The selection of significant terms, although computer intensive, could be explored further to find the best-fitted model.

References

Skinner, C. J., and Shlomo, N. 2008. Assessing identification risk in survey micro-data using log linear models. *Journal of American Statistical Association, 103*(483), 989-1001.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., and de Wolf, P.-P. 2012. *Statistical disclosure control*. Chichester, UK: John Wiley & Sons.