Modeling the Re-identification Risk per Record in Microdata

Elsayed A.H. Elamir and Chris J. Skinner
University of Southampton, Department of Social Statistics
Southampton, SO17 1BJ, U.K.
E-mail:eahe@socsci.soton.ac.uk and cjs@socsci.soton.ac.uk

1 Introduction

A measure of re-identification risk at the record level has a variety of potential uses in statistical disclosure control for microdata. It might be used to detect records for the application of record-level methods of disclosure control. It might be used to judge whether the whole microdata file is safe, if safety is defined as the absence of any records which are likely to identify an individual. We extend the work of Skinner and Holmes (1998) in two ways. First, we formulate a new definition of re-identification risk per record by extending the file-level definition of Skinner and Elliot (2002). Second, we investigate how these measures can be extended to accommodate measurement error in identifying variables.

We suppose that the risk of re-identification arises from the possibility that an intruder may attempt to use a set of (categorical) identifying variables to match a record in the microdata with a known individual in the population. Skinner and Elliot (2002) define a file-level measure of risk Θ as the probability that such an attempt leads to a correct match given that the record is sample unique, that is it is unique with respect to the identifying variables in the microdata sample (the worst case). Probability is defined here with respect to a random draw of the known individual from the population so that Θ may be expressed as

$$\Theta = \sum_j \mathrm{I}\left(f_j = 1\right) \Big/ \sum_j F_j \mathrm{I}\left(f_j = 1\right) \tag{1}$$

where j denotes a combination of values of the identify variables and f_j and F_j are the numbers of individuals with combination j in the sample and population, respectively. It is assumed that the identifying variables are measured identically in the microdata and for the known individual. A limitation of this measure is that it does not differentiate between possible varying risks for different records in the microdata.

In this paper, we extend this definition to the record-level, by letting Θ_j be the probability that the match is correct for a record with a fixed combination of values j. Probability is defined here with respect to a model generating the population. If we restrict attention to sample unique records, all with different values of j, all these records will in general have different values of Θ_j. Assuming again that the identifying variables are measured identically in the microdata and for the known individual, we may write

$$\Theta_j = \mathrm{E}\left(1/F_j\right) \tag{2}$$

We suppose that it is reasonable to assume that F_j will be unknown to the intruder and that inference about Θ_j must be made using the sample microdata.

2 Per-record Level Model

Suppose that $F_j \sim \text{Poisson}(\lambda_j)$, that sampling is Bernoulli with sampling fraction π and that attention is restricted to sample unique records. In this case, treating λ_j as fixed,

$$\Theta_j = \text{E}\left[\frac{1}{(1-\pi)\,\lambda_j}\left(1 - e^{-(1-\pi)\lambda_j}\right)\right] \tag{3}$$

Inference about λ_j and hence Θ_j (π is known) uses the observed $f_j \sim \text{Poisson}(\mu_j)$, where $\mu_j = \pi\lambda_j$. We consider a compound log-linear model, where

$$\mu_j = \exp(x_j\beta)\, w_j \tag{4}$$

where x_j is a vector of indicator variables, representing specified main effects and interactions of the identifying variables and $w_j = \exp(\varepsilon_j)$ is a possible random effect, independent of x_j; see, for example, Agresti (1996) and Cameron and Trivedi (1998).

A simple measure is obtained by fitting a standard log-linear model (setting $w_j = 1$), letting $\widehat{\mu}_j$ be the fitted value in the model and setting $\widehat{\lambda}_j = \widehat{\mu}_j/\pi$. Alternative, measures are obtained by allowing w_j to follow gamma and inverse Gaussian distributions and defining Θ_j as the expectation over the conditional distribution of w_j given $f_j = 1$.

We present the results of a simulation study in which w_j is simulated from either a uniform, gamma or lognormal distributions and the estimated values $\widehat{\Theta}_j$ are compared with the actual values Θ_j.

3 Extension to Misclassification

We briefly refer to an extension of measures (1) and (3) to the case where the identifying variables are not measured identically in the microdata and for the known individual, that is there is misclassification.

REFERENCES

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

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

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 (2002). Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, 14, 361-372.

RÉSUMÉ

Une mesure de risque de réidentification au niveau record a une variété d'utilités potentielles pour des microdonnees. Ces measures pourraient fournir l'évidence utile pour soutenir des décisions au sujet de divulgation. Nous proposons une nouvelle mesure de risque au niveau record qui est la probabilité qui une correspondence unique entre un fichier de micodonnées et une unité de population est correcte. Pour des variables discrétes sujet á aucune erreur de mesure, nous étudions cette mesure pour un modéle Poisson et un modéle Poisson-gamma et prolongeons ces modéles au cas de la classification fausse des variables. D'ailleurs, nous présentons une étude de simulation basée sur distribution Poisson-gamma.