Evaluation of per-record identification risk by additive modeling of interaction for contingency table cell probabilities

Akimichi TAKEMURA
Department of Mathematical Informatics
Graduate School of Information Science and Technology
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN
takemura@stat.t.u-tokyo.ac.jp

Abstract

We propose to fit a Lancaster-type additive model of interaction terms for cell probabilities of contingency tables to evaluate the conditional probability of population uniqueness of sample unique cells in microdata sets. Moment estimation of the Lancaster-type additive model is straightforward and the proposed estimation procedure is intuitively appealing from the viewpoint of disclosure risk assessment. In order to increase flexibility of the procedure, we also consider Ridge type shrinkage of the Lancaster-type additive model towards the independence model. We demonstrate the usefulness of the proposed procedure by applying it to a test data set based on 1990 U.S. Census PUMS data.

1. Introduction

In evaluating the disclosure risk of a given microdata set, the number (or the proportion) of the population uniques among the sample unique records is an important overall measure of the disclosure risk. For estimating the number of population uniques, various models have been proposed, including Poisson-Gamma model, Ewens sampling formula, and more recently, Pitman model. These models treat the sample uniques exchangeably. Therefore under these models the estimated proportion of population uniques among the sample unique records is the common estimated conditional probability of population uniqueness for every sample unique record. However it is clear that some sample unique records are more likely to be population uniques than other records, depending on the intuitive "rareness" of the records. One way of evaluating the per-record identification risk is the modeling of cell probabilities of the contingency table corresponding to a microdata set, where all the key variables of the microdata set are categorized and the joint frequencies of the key variables are counted. Then the per-record identification risk can be evaluated in terms of the estimated probability of population uniqueness of sample uniques cells. This approach was investigated in [3] and [1]. They used the standard log-linear model of cell probabilities of contingency tables. In this paper we consider fitting the Lancaster-type additive model of interaction terms, because of its simple computation and interpretation.

2. Lancaster-type additive model and its shrinkage to independence

Here we describe the Lancaster-type additive model. For simplicity of notation we describe the model for 3-way contingency tables, although in actual applications we need to use m-way tables, where m (the number of key variables) is often around 10. Extension of the model to higher dimensional tables is trivial except for notational complication.

Let p_{ijk} denote the cell probability of an $I \times J \times K$ contingency table. Denote the one-dimensional marginal probabilities by $p_{i\cdot\cdot}, p_{\cdot j\cdot}, p_{\cdot\cdot k}$ and the two-dimensional marginal probabilities by $p_{ij\cdot}, p_{i\cdot k}, p_{\cdot jk}$. The Lancaster-type additive model without three variable interaction in Lancaster's sense ([2], [4]) is defined by

$$p_{ijk} = p_{i\cdot\cdot}p_{\cdot j\cdot}p_{\cdot\cdot k}\Big\{1 + \Big(\frac{p_{ij\cdot}}{p_{i\cdot\cdot}p_{\cdot j\cdot}} - 1\Big) + \Big(\frac{p_{i\cdot k}}{p_{i\cdot\cdot}p_{\cdot\cdot k}} - 1\Big) + \Big(\frac{p_{\cdot jk}}{p_{\cdot j\cdot}p_{\cdot\cdot k}} - 1\Big)\Big\}. \tag{1}$$

The moment estimation of (1) is straightforward. Let n denote the sample size and let $n_{i\cdot\cdot}, n_{ij\cdot}, \ldots,$ denote the sample marginal frequencies. Then the cell probability is estimated as

$$\begin{aligned} \hat{p}_{ijk} &= \hat{p}_{i\cdot\cdot}\hat{p}_{\cdot j\cdot}\hat{p}_{\cdot\cdot k}\Big\{1 + \Big(\frac{\hat{p}_{ij\cdot}}{\hat{p}_{i\cdot\cdot}\hat{p}_{\cdot j\cdot}} - 1\Big) + \Big(\frac{\hat{p}_{i\cdot k}}{\hat{p}_{i\cdot\cdot}\hat{p}_{\cdot\cdot k}} - 1\Big) + \Big(\frac{\hat{p}_{\cdot jk}}{\hat{p}_{\cdot j\cdot}\hat{p}_{\cdot\cdot k}} - 1\Big)\Big\} \\ &= \frac{n_{i\cdot\cdot}}{n}\frac{n_{\cdot j\cdot}}{n}\frac{n_{\cdot\cdot k}}{n}\Big\{1 + \Big(\frac{nn_{ij\cdot}}{n_{i\cdot\cdot}n_{\cdot j\cdot}} - 1\Big) + \Big(\frac{nn_{i\cdot k}}{n_{i\cdot\cdot}n_{\cdot\cdot k}} - 1\Big) + \Big(\frac{nn_{\cdot jk}}{n_{\cdot j\cdot}n_{\cdot\cdot k}} - 1\Big)\Big\} \end{aligned} \tag{2}$$

Once the cell probability is estimated, the per-record disclosure risk for sample unique record (or cell) is given as follows. We assume that N individuals of the population fall into the cells of the contingency table according to the multinomial scheme and the sampling of n individuals is by simple random sampling without replacement. Let n_{ijk} and N_{ijk} denote the sample and the population cell frequencies. Then given that a cell is a sample unique ($n_{ijk} = 1$), the conditional probability of the cell being a population unique ($N_{ijk} = 1$) is written as

$$P(N_{ijk} = 1 \mid n_{ijk} = 1) = (1 - p_{ijk})^{N-n}. \tag{3}$$

If we replace p_{ijk} by $\hat{p}_{ijk}$ of (2) we obtain an estimated value of the conditional probability. The number of population unique in the microdata set is now estimated by

$$\sum_{(i,j,k)\,:\,n_{ijk}=1} \hat{P}(N_{ijk} = 1 \mid n_{ijk} = 1) = \sum_{(i,j,k)\,:\,n_{ijk}=1} (1 - \hat{p}_{ijk})^{N-n}.$$

Note that (3) is a decreasing function of p_{ijk} and this reflects an intuitively obvious fact that the disclosure risk of a sample unique is high if the estimated probability of the cell is small. $\hat{p}_{ijk}$ of (2) is small if univariate relative frequencies are small and the terms like $(nn_{ij\cdot})/(n_{i\cdot\cdot}n_{\cdot j\cdot})$ are small. This term is the ratio of the actual two-dimensional frequency n_{ij} to the estimated frequency $n_{i\cdot\cdot}n_{\cdot j\cdot}/n$ under the independence model. Therefore $\hat{p}_{ijk}$ of (2) combines one-dimensional rareness and the two-dimensional rareness, which are routinely considered in disclosure control practices.

We saw that the Lancaster-type additive model is simple to estimate and the estimated cell probability is easy to interpret from the viewpoint of the disclosure control. However in actual application of the model to microdata sets we encounter several difficulties and have to be careful about the interpretation of the model.

First, even though the model contains only up to the two-variable interactions, the number of estimated interaction terms may be large. In our applications the number of the key variables m is often around 10. Let $I_l,\ l = 1, \ldots, m$, denote the number of categories of l-th variable. Then total number of parameters estimated is roughly equal to

$$I_1 I_2 + I_1 I_3 + \cdots + I_{m-1} I_m.$$

For example, when $m = 10$ and $I_l = 10,\ l = 1, \ldots, m$, then $I_1 I_2 + I_1 I_3 + \cdots + I_{m-1} I_m = 4500$. Therefore there is a question of stability of the estimated cell probabilities. For this reason, we introduce a shrinkage factor $0 \leq \lambda \leq 1$ and consider shrinking $\hat{p}_{ijk}$ of (2) towards the independence model:

$$\hat{p}_{ijk}(\lambda) = \frac{n_{i..}}{n}\frac{n_{.j.}}{n}\frac{n_{..k}}{n}\Big\{1 + \lambda\Big(\frac{n n_{ij.}}{n_{i..}n_{.j.}} - 1\Big) + \lambda\Big(\frac{n n_{i.k}}{n_{i..}n_{..k}} - 1\Big) + \lambda\Big(\frac{n n_{.jk}}{n_{.j.}n_{..k}} - 1\Big)\Big\}. \tag{4}$$

Second, there is a problem of negative estimated cell probabilities. Note that (4) can be negative for $\lambda > 0$. Then a logical thing to do is to replace negative estimated value by 0 (i.e. $\hat{p}_{ijk}(\lambda) \mapsto \max(\hat{p}_{ijk}(\lambda), 0)$) and renormalize the cell probabilities. More precisely $\hat{p}_{ijk}(\lambda)$ of (4) should be replaced by

$$\frac{1}{c(\lambda)} \times \max(\hat{p}_{ijk}(\lambda), 0), \quad c(\lambda) = \sum_{\hat{p}_{i'j'k'}(\lambda) \geq 0} \hat{p}_{i'j'k'}(\lambda) = 1 - \sum_{\hat{p}_{i'j'k'}(\lambda) < 0} \hat{p}_{i'j'k'}(\lambda). \tag{5}$$

In the application of the next section we will see that $c(\lambda)$ can be substantially larger than 1. This indicates lack of fit of the additive model.

Third, there is the problem of the structural zeros. The estimator $\hat{p}_{ijk}$ of (2) has the simple form because of the assumption of no structural zeros in the $I \times J \times K$ contingency table. In actual contingency tables corresponding to microdata sets there are many structural zeros. The Lancaster-type additive model loses its simplicity when the structural zeros are incorporated in to the model. In this sense, we regard the model as convenient approximate model for quick evaluation of disclosure risk.

3. An example

We applied the Lancaster-type additive model to a test data set resampled from 1990 U.S. Census of Population and Housing Public Use Microdata Samples. We resampled $n = 9809$ individuals from the state of Washington and chose $m = 10$ variables for experimental purpose: 1. Relationship (14 categories), 2. Sex (2), 3. Age (91), 4. Marital status (5), 5. Place of birth (14), 6. Spouse present/absent (7), 7. Own child (2), 8. Age of own child (5), 9. Related child

(2), 10. Detailed relationship (10). The population size is $N = 4,867,000$. The dataset can be viewed as contingency table of the type

$$14 \times 2 \times 91 \times 5 \times 14 \times 7 \times 2 \times 5 \times 2 \times 10$$

with 249,704,000 cells. Note that fitting the log-linear model to contingency table of this size is computationally fairly difficult. The frequencies of the cell sizes (frequency of frequencies, size indices) of this data set is given as follows.

Cell size	1	2	3	4	5	6	7	8	9	10	11 $\leq$
Frequency	2249	521	275	132	104	60	59	34	46	19	124

The estimated number of population uqiques among the 2249 sample uniques for various values of λ is given as follows.

λ	0.0	0.2	0.4	0.6	0.8	1.0	Ewens	Pitman
no renormalization	885.02	480.27	372.82	315.93	278.88	252.13	5.88	213.96
renormalization	885.02	482.94	403.46	376.21	362.50	354.68		
sum of negative prob.	0	–0.013	–0.193	–0.465	–0.758	–1.065		

The numbers below "Ewens" and "Pitman" shows the estimated number of the population uniques under these models. Renormalization means that the renormalization in (5) was applied because of negative estimated cell probabilities. The sum of negatively estimated cell probabilities is substantial for $\lambda \geq 0.4$. This indicates that the fit of our additive model is not very good for this data set.

REFERENCES

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

[2] Lancaster, H. O. (1971). The multiplicative definition of interaction. *Austral. J. Statist.*, **13**, 36–44.

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

[4] Zentgraf, R. (1975). A note on Lancaster's definition of higher-order interactions. *Biometrika*, **62**, 375-378.

RESUME

Nous avons appliqué le modèle additif de Lancaster-type au problème de commande de révélation.