Empirical Bayesian Misclassification Analysis on Categorical Data

Han Seung Lim
Researcher, RMS Dept, Credit Ratings Division,
Korea Management Consulting and Credit Rating Corporation,
26-4 Youido Dong, Yungdungpo Ku,
SEOUL 150-737, KOREA.
haslim@kmcc.com

Chong Sun Hong
Professor, Department of Statistics, SungKyunKwan University,
3-53 Myongryun Dong, Chongro Ku,
SEOUL 110-745, KOREA.
cshong@skku.ac.kr

1. Introduction

Currently, lots of large-scaled social survey researches are performed by using many survey methods. For large-scaled survey researches spending low cost, however, it may be easy to obtain nonresponse or missing data and hard to expect exact answers. There are two kinds of errors on survey researches which are sampling errors and non-sampling errors. Non-sampling errors could contain both nonresponse errors and response errors. A nonresponse error occurs when it does not obtain a response about a certain unit in sample, and a response error does when one observation of data is classified into wrong categories, which is so called as misclassification.

If misclassified data itself is analyzed, then biases of estimators may be occured and standard Pearson X^2 tests may have inflated true type I error rates (Bross 1954). Therefore, it is necessary and important that one should decide whether categorical data is misclassified before the analysis. Bross (1954), Barrons (1977), Hochberg (1977), Fuchs (1998) and many others have been researched to solve misclassification problems on two-way contingency table. Tenenbein (1970, 1971), among others, proposed the double sampling scheme which is a method to analyze misclassification errors under given survey cost: A subsample of small sample size is drawn from a misclassified fallible sample and explore the degree of misclassification to decrease biases of estimated cell probabilities.

Nonresponse or missing data in categorical data can be divided in a context of the missing data mechanism. If the missing data mechanism is either missing at random (MAR) or missing complete at random (MCAR), then missing data is ignorable nonresponse (IN) so that this

is estimated with observations only. On the other hand, if the mechanism is neither MAR nor observed at random (OAR), this data is called as nonignorable nonresponse (NIN). When this nonresponse data is ignored and analyzed, the randomness of sample may be affected and estimates can be biased (Kalton 1983). Recently while kinds of Bayesian methods are developed to estimate nonignorable nonresponse values, Park and Brown (1994) propose MLEs by using EM algorithms assuming that nonresponse cell count has a prior probability distribution. In Bayesian estimation process, Sebastiani and Ramoni (1997) induce an interval estimator and a point estimator which are named by Bound and Collapse. Tebaldi and West (1998) estimate missing data by using Markov Chain Monte Carlo (MCMC) methods such as Metropolis-Hasting algorithm.

In this paper, when data on $I \times J$ contingency table has misclassification errors at only one of two categorical variables and marginal sums of a well-classified variable are fixed, methods to estimate nonresponse data (missing data) are extended to decide whether the data is misclassified. We regard marginal sums of a misclassified categorical variable as missing observations and estimate missing cell probabilities by Bayesian method via the concepts of Bound and Collapse of Sebastiani and Ramoni (1997). Some useful informations for Bayesian estimating are collected from the double sampling scheme of Tenenbein (1970), sampling cost, external-informations and pre-informations which are results surveyed few years ago or results of similar projects surveyed by other research institutes. We propose MC-statistic (Misclassification statistic) to measure the degree of misclassification errors for given categorical data and MC_i-statistics ($i = 1, \ldots, I$) to measure at i-th level of a fallible variable. Distributions of these statistics have been researched with lots of misclassified data obtained by various simulation methods. This simulation study is considered for two cases which are ignorable nonresponse (IN) and nonignorable nonresponse (NIN), and this researches are investigated with respect to of initial sample size (N), double sampling rates, misclassification rates, values of marginal probabilities and conditional probabilities.

REFERENCES

Barron, B. A. (1977). The Effects of misclassification on the estimation of relative risk. *Biometrics.*, **33**, 414-418.

Bross, I. (1954). Misclassification in tables. *Biometrics.*, **10**, 478-486.

Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. *Journal of the American Statistical Association.*, **77**, 270-278.

Good, I. J. (1968). *The Estimation of probability: An essay on modern bayesian methods.* MIT press, Cambridge.

Hochberg, Y. (1977). On the use of double sampling schemes in analyzing categorical data with misclassification errors. *Journal of the American Statistical Association.*, **72**, 914-921.

Kalton, G. (1983). *Compensating for missing survey data.* Research report series, Institute for social research.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data.* John Wiley & Sons, New York.

Park, T. S. and Brown, M. B. (1994). Models for categorical data with nonignorable nonresponse. *Journal of the American Statistical Association.*, **89**, 44-52.

Sebastiani, P. and Ramoni, M. (1997). Bayesian inference with missing data using bound and collapse. *KMI-TR. Open University.*, **58**,

Tebaldi, C. and West, M. (1998). Reconstruction of contingency tables with missing data. *Duke technical report. Duke University.*

Tenenbein, A. (1970). A double sampling scheme for estimation from binomial data with misclassification. *Journal of the American Statistical Association.*, **65**, 1350-1361.

Tenenbein. A. (1971). A double sampling scheme for estimation from binomial data with misclassification: sample size determination. *Biometrics.*, **27**, 935-944.

RESUME

En ce qui concerne la donnee categorielle, l'erreure peut provenir de la demarche de la collection de cette donnee. Si on analyse la donneee mal classifiee en la prenant pour bien construite, cela va entrainer l'erreur au resultat d'estimation. Au contraire, si on prend la donne mal classee pour bien construite, on doit depenser le cout et le temps inutiles pour la corriger en vain. Ainsi s'interroger si la donnee est bien construite ; c'est une etape tres importante avant d'analyser celle-la . Dans le cas ou la donnee est mal classifiee a un seul niveau de deux variables dans le tableau contingent bi-dimentionnel, on a fixe la somme marginale d'une variable bien construite pour verifier s'il y a une classification mal menee, et a reclassifie la donne eventuellement erronnee au moyen des concepts de Bound et de Collapse proposes par Sebastiani et Ramoni (1997). Le schema de double echqntillonage (Tenenbien 1970) est utilise pour obtenir des informations sur la classification erronee. Nous proposons le teste statistique afin de resoudre les problemes de la classification mal construite et d'exqminer les behaviors de la statistique a partir des etudes de simulation.