Small area estimation in business information technology

Militino, A.F., Ugarte, M.D., Goicoa, T.
Departamento de Estadística e Investigación Operativa, Universidad Pública de Navarra
Campus de Arrosadía, 31006 Pamplona, Spain
E-mail*: militino@unavarra.es

Abstract

In this paper, alternative logistic model based estimators are suggested to derive estimates from information and communication technology of businesses (ICTB) surveys. Final estimates are benchmarked to match with direct estimates at provincial level and standard errors are given by means of bootstrap techniques. A simulation study is conducted to compare the performance of the estimators. Results are illustrated with the 2010 ICTB survey of the Basque Country (Spain).

1 Introduction

There is an increasing demand of information for smaller administrative divisions not originally planned in the sample design, known as small areas (Jiang and Lahiri, 2006). The extensive research in small area estimation in the last few years contribute to extend its use in statistical offices (Militino et al., 2012). Model-based estimators are mainly based on mixed and generalized linear mixed models, but other alternatives are also possible (Ugarte et al., 2008).

The aim of this paper is to obtain model-based small area estimators for providing estimates of binary variables in unplanned domains of information and communication technology of businesses (ICTB) surveys. For illustration purposes, the 2010 ICTB survey of the Basque Country, Spain, will be considered. The Basque Country is an Autonomous Region located in the North of Spain, and it is divided into three provinces: Araba, Gipuzkoa, and Bizkaia, which in turn are divided into twenty small areas or counties. The region has 7,234 km^2 and 2,185,000 inhabitants in 2010. However, it is one of the most industrialized regions in Spain with a high level of entrepreneurship and it has its own Statistical Office.

Assuming that the ICTB survey is designed for breakdowns defined as employment size, economic activity and region (NUT-2), logistic model-based estimators are going to be derived at NUT-3 or a more disaggregated levels. Auxiliary information comes from the registers that collect all of the enterprises in the whole region classified by the aforementioned breakdowns. A sample of 7725 establishments is drawn using a stratified random sampling design with three strata defined by economic activities, employment size breakdowns (less than 10 employees, between 11 and 99, and more than 100 employees) and provinces.

The target population consists of all establishments with more than one employee in any economic activity developed in the region, except for the primary sector and domestic help. The establishment is defined as a single business entity operating in the region either as a legally constituted body, such as a company, trust, local or central government trading organization, incorporated society, or self-employed individual. In 2010, the number of establishments was 195222, irregularly distributed in the twenty counties of the three provinces.

The survey was designed to provide direct estimates at provincial level, employment size, and economic activity breakdowns. However, new demands require to provide these estimates at county level. Dismissing the possibility of increasing the sample size, small-area estimators are proposed.

2 Logistic Model Based Estimators and Mean Squared Error

For estimating in unplanned domains (NUT-3) using the 2010 ICTB survey of the Basque Country, we define the so called double logistic model consisting of two logistic models of fixed effects. In the first model the response variable is the proportion of enterprises satisfying an specific item of the survey and the explanatory variables are employment size, economic activity, and belonging county. This model is used when there are sample in all counties, all employment size breakdowns, and all economic activities because in this case, the model coefficients could be estimated and then, it is possible to project estimates in the unplanned domains. However, the county is not an stratification variable, and then, it is not possible to guarantee sample in all counties and as a consequence, it will not be possible to compute unplanned domain estimates. As alternative we propose a logistic model where the county indicator variable is substituted by its province. For comparison purposes we also propose a projective estimator. It gives estimates for out of sample areas whenever all counties, employment size breakdowns, and economic activities have sampled elements. Otherwise, the projective estimator is not useful.
A simple bootstrap procedure for estimating the MSE of the small area logistic estimators and their benchmarked counterparts are also proposed. This technique avoids approximation methods for calculating the analytical variance of the logistic estimators. Moreover, it is appropriate for the benchmarked estimators as closed expressions are usually very difficult to obtain. To assess the performance of double logistic estimator and projective estimator, as well as its benchmarked versions in terms of relative bias and error a simulation study has been conducted. Results in our application and the simulation study reveal a better performance of the double logistic estimator in terms of relative bias and error.
Three items of the ICTB survey are chosen to illustrate results: e-commerce, electronic data interchange (EDI), and internet connection. E-commerce is defined as the set of commercial transactions (to buy or sell various products or services) via internet demand. The EDI variable comprises electronic communication usually used to place orders, send invoices, and make economic transactions among different institutions and organizations through internet or other networks. Internet connection refers to have internet access through land line and modem or through ADSL. The main reason for choosing these three variables is because of its different incidence (10-20% for e-commerce, 15-30% for EDI, and 60-90% for internet connection). The small domains for this survey are defined as the combination of 20 counties, 27 economic activities, and 3 employment sizes giving rise to $20 \times 27 \times 3 = 1620$ small areas.
In the population register of 2010 there were not enterprises in all the small areas, but in 1147, where only 950 are represented in the sample. Therefore, a reminder of 197 small domains are out of sample areas. Finally, proportions and totals of establishments using e-commerce, doing EDI, and having internet connection are aggregated to provide estimates at county level.
Alternative estimators have also been studied in the particular survey considered here. For example, an estimator based on a logistic mixed model with employment size and economic activity as fixed effects and county as a random effect. However, bootstrap tests have dismissed the statistical significance of the random effects in most variables. A logistic synthetic estimator, that is, an estimator derived from a logistic model with province, employment size and economic activity as explanatory variables has been also considered, but this estimator provides very similar estimates in all counties, indicating that it might not be appropriate as it does not take into account county specificities. This result agrees with other results showing that synthetic estimates are not sufficiently variable to be plausible (Särndal, 1984; Malec et al., 1997).

In summary, this logistic estimator seems to be an attractive solution because of its simplicity, unbiasedness, and precision. It is also appropriate for out of sample areas because the use of stratification variables as explanatory variables guarantees the availability of sampled elements to fit the model. Finally, in this paper we have focused on county estimates because these domains are rarely planned in most ICTB surveys, yet other domains could also be used.

Acknowledgments: This work has been supported by the Spanish Ministry of Science and Innovation (project MTM 2011-22664 which is co-funded by FEDER). We would like to thank the Basque Statistical Office, EUSTAT, for providing the data.

References

Jiang, J., and Lahiri, P. (2006b). Mixed model prediction and small area estimation. *Test*, **15**, 1-96.

Militino, A. F., Goicoa, T., and Ugarte, M. D. (2012). Estimating the percentage of food expenditure in small areas using bias-corrected P-spline estimators. *Computational Statistics and Data Analysis*, **56**, 2934-2948.

Särndal, C. E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Series in Statistics.

Ugarte, M. D., Militino, A. F., and Goicoa, T. (2008). Adjusting Economic Estimates in Business Surveys. *Journal of Applied Statistics*, **35**, 1253-1265.