Logistic Regression Type Small Area Estimations Based on Relative Error

Hee-Jin Hwang\(^1\) · Key-Il Shin\(^2\)

\(^1\)Data Information Center, NHN; \(^2\)Department of Statistics, Hankuk University of Foreign Studies

(Received January 2011; accepted March 2011)

Abstract

Almost all small area estimations are obtained by minimizing the mean squared error. Recently relative error prediction methods have been developed and adapted to small area estimation. Usually the estimators obtained by using relative error prediction is called a shrinkage estimator. Especially when data set consists of large range values, the shrinkage estimator is known as having good statistical properties and an easy interpretation. In this paper we study the shrinkage estimators based on logistic regression type estimators for small area estimation. Some simulation studies are performed and the Economically Active Population Survey data of 2005 is used for comparison.

Keywords: Shrinkage estimator, mean squared error, logistic mixed model, logistic regression model.

1. Introduction

Most sample surveys are designed to produce estimates for a whole geographical area. The sample sizes of small areas such as cities and counties would be small (even zeros for some areas) because the overall sample size in a survey is usually determined to provide a specific accuracy at a higher level of aggregation than that of a small area. In that case, sample proportions such as unemployment rates in counties may be poorly estimated. Usually survey estimates based on such small sample sizes could provide formidable standard errors leading to unacceptable confidence intervals. However, a heavy burden of time and cost occurs in obtaining an acceptable accuracy for the statistic that of a small area. Therefore, instead of doing the extra survey overcoming the unacceptable accuracy caused by small sample size, reliable official statistics for the small areas can be produced efficiently by applying the small area estimation methods.

The small area estimation methods can be classified as design-based and model-based. In general, when the auxiliary information is available, model-based methods are known better. Especially, for binary data such as unemployment data, the logistic regression estimator and logistic mixed model estimator are known as having good statistical properties. In addition, the random effects models that treat each county as a cluster can provide improved estimates. Studies on the logistic regression...
type estimators to estimate unemployment rates can be found in Kim and Choi (2004) and Yeo et al. (2008).

In this study, we suggest estimators obtained by minimizing relative error (RE) or mean squared percentage error (MSPE) instead of minimizing MSE. This estimator has the advantage of an easy calculation obtained by multiplying a certain constant term to the original estimator obtained by minimizing MSE. The estimator known as a shrinkage estimator was studied by Hwang and Shin (2008, 2009); however, in those papers, the application of logistic regression estimator and logistic mixed estimator to small area estimation was not performed in their papers. Hence the applicability of a shrinkage estimation method to logistic regression type estimators should be studied and the superiority of shrinkage small area estimators in an aspect of relative error criterion should be examined.

In this paper we compare the efficiency of each estimator such as direct estimator, logistic regression estimator, logistic mixed estimator, and the shrinkage type estimators derived from logistic type estimators. In Section 2, we briefly explain logistic regression type small area estimators which are practically and widely used for binary data. In addition, the shrinkage estimators made from the original small area estimators are explained. Section 3 performs the analysis of data using the estimators mentioned before and the results for comparison of the estimators are shown. For comparison, we use the unemployment data from the Monthly Report on the Economically Active Population Survey in Korea (2005). Section 4 includes some concluding remarks.

2. Suggested Methods for Small Area Estimations

In this section, we briefly summarize some small area estimators mentioned in Section 1. Some widely used estimators for binary data obtained by minimizing MSE are explained and we study the new shrinkage estimators obtained by using RE.

2.1. Small area estimators using MSE

Some small area estimation methods have been suggested as design-based estimation such as direct estimation, synthetic estimation, and composite estimation. In addition, the well-known model based estimations have been suggested such as regression estimation. Empirical Bayes estimation (EB), Hierarchical Bayes estimation (HB). For binary response data, logistic regression type estimators are widely used. Therefore, in this study, we consider two logistic regression type estimators specially used for binary data: the logistic regression estimator and the logistic mixed estimator. Details on these estimators can be found in Agresti (2002) and Rao (2003).

2.1.1. Direct estimator

The direct estimator, \( \hat{Y}_{DE} \) is defined as \( \hat{Y}_{DE} = \sum_j w_{ij} y_{ij} \), where \( \hat{Y}_i \) is the estimate of the interesting variable in \( i^{th} \) small area, \( w_{ij}, i = 1, \ldots, n; j = 1, \ldots, n_i \) is the sampling weight and \( y_{ij} \) is the value of \( j^{th} \) element in \( i^{th} \) small area. Usually the sampling weight \( w_{ij} \) have the same value in the same stratum. Hence for simplicity, we use \( w_{ij} = 1 \) for all \( i \) and \( j \) throughout this paper. This constant sampling weight does not affect the comparison result of the estimators’ superiority.

2.1.2. Logistic regression estimator

The logistic regression model is defined by following.

\[
\log \left( \frac{p_{ij}(x)}{1 - p_{ij}(x)} \right) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_p x_{pjij}.
\] (2.1)