Computational Statistics and Data Analysis 78 (2014) 159–175

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Small area prediction for a unit-level lognormal model

Emily Berg [a,*], Hukum Chandra [b]

[a] Department of Statistics, Iowa State University, United States
[b] Indian Agricultural Statistics Research Institute, India

ARTICLE INFO

Article history:
Received 31 October 2012
Received in revised form 11 March 2014
Accepted 11 March 2014
Available online 4 April 2014

Keywords:
Lognormal
Mean squared error
Small area estimation

ABSTRACT

Many variables of interest in business and agricultural surveys have skewed distributions. Small area estimation methods are investigated under an assumption that the lognormal model is a reasonable approximation for the distribution of the response given covariates. Closed form expressions for an empirical Bayes (EB) predictor and for the associated mean squared error estimator are derived. In simulation studies, the EB predictors are more efficient than model-based direct and synthetic estimators previously proposed for lognormal data. Also, coverage of confidence intervals for the lognormal predictions approximate the nominal coverage. The simulations also demonstrate that the suggested predictor is robust to departures from the assumptions of the lognormal model. The methodology is successfully applied to estimate erosion rates for hydrologic units using data from the Conservation Effects Assessment Project.

1. Introduction

Statistical agencies are often asked to produce small area estimates for skewed variables. For example, the United States National Agricultural Statistics Service (NASS) publishes estimates of the acres harvested in a variety of crops (Bellow and Lahiri, 2011). The United States Natural Resources Conservation Service (NRCS) is interested in estimating the acres in roads at the county level (Wang and Fuller, 2003). Chandra and Chambers (2011) consider estimation of the total expenditures of Australian farms using data from the Australian Agricultural and Grazing Industries Survey (AAGIS). Standard estimators of the mean of a skewed variable can be inefficient. Chambers (1986) discusses robust estimation of a population mean.

When domain sample sizes are too small to support reliable direct estimators, effects of skewness and outliers can be more pronounced, and appropriately accounting for the distribution of the response given available auxiliary information is particularly important. A common approach to small area estimation is to use model-based estimators instead of design-based estimators, and efficiency gains are realized if the models contain information about the structure of the domain means. Rao (2003), Jiang and Lahiri (2006), and Pfeffermann (2013) provide comprehensive reviews of small area estimation.

Small area methods for non-normal, continuous response variables using ideas of robustness or non-normal distributions have been proposed. Chambers and Tzavidis (2006), Ghosh et al. (2008), Gershunkaya (2010), and Sinha and Rao (2009) develop robust small area estimation procedures to mitigate the effects of outliers. The theory underlying these methods assumes that mean models are linear and that outliers have symmetric distributions, assumptions not necessarily appropriate for skewed data. Slud and Maiti (2006) construct an EB predictor for a small area mean under an assumption that the area-level direct estimators have lognormal distributions. The EB predictor proposed in this paper differs from the Slud and Maiti (2006) predictor because we work with unit-level data instead of area-level data. Ghosh and Maiti (2004) consider

* Correspondence to: Snedecor Hall, Ames, IA, 50011, United States.
E-mail addresses: emily.j.berg@gmail.com, emilyb@iastate.edu (E. Berg), hchandra12@gmail.com (H. Chandra).

http://dx.doi.org/10.1016/j.csda.2014.03.007
0167-9473/

small area estimation for unit-level models based on natural exponential quadratic variance function families. The Ghosh and Maiti (2004) approach assumes that the covariates are the same across units in a single small area. In contrast, we desire a procedure that is applicable for a situation (such as the framework of Chandra and Chambers (2011)) where unit-level covariates are available. Dagne (2001) develops hierarchical Bayes predictors of small area means based on unit-level models where the variables are assumed to have normal distributions after an appropriate Box–Cox transformation. The unit-level lognormal model is a special case where the parameter governing the Box–Cox transformation is zero.

Chandra and Chambers (2011) use a lognormal distribution as a basis for constructing a model-based direct estimator for a small area mean. The model-based direct estimator of Chandra and Chambers (2011) is a weighted sum of sampled units, where the weights are defined to give the minimum mean squared error linear predictor of the population mean if the parameters of the lognormal distribution were known. Chandra and Chambers (2011) compare the model-based direct estimator to an estimator based on an approach of Karlberg (2000). The estimator based on ideas of Karlberg (2000) is a type of synthetic estimator because the estimator only accounts for between-area variability through the covariates.

We derive an empirical Bayes, or empirical best, (EB) predictor and an approximately unbiased mean squared error estimator under a unit-level lognormal model. The EB predictor is an estimator of the conditional expectation of an area mean given observed data with respect to a unit-level lognormal model. The suggested methods are computationally simple because the predictor and mean squared error estimator have closed form expressions.

Our interest in the unit-level lognormal model was motivated by several studies. One is the study where the objective was to estimate average expenditures of Australian farms in domains defined by Australian agricultural regions using AAGIS data. This study provides the framework for the simulation of Section 4.5, where we consider a finite population representative of the AAGIS data. Another application that fits the general data structure considered in this paper is estimation of the acres harvested at the county level using data from the NASS County Estimates Survey (Bellow and Lahiri, 2011). The response variable, the harvested acres, has a skewed distribution, and the covariate, a size measure from the NASS list frame, is known for all farms in the population.

The first two motivating applications are related to agriculture, but the methodology developed in this paper has potential application in business surveys. Fuller (2009, pg. 190) contains an example where the variable of interest, total payroll, has a log-linear relationship with total employment.

The application study for this paper is the Conservation Effects Assessment Project (CEAP), a collection of surveys designed to monitor processes such as soil loss and chemical runoff. We apply the proposed methodology to estimate mean erosion rates in domains defined by small watersheds using unit-level auxiliary information from an administrative soils database.

The rest of this paper is organized as follows. In Section 2, we present the unit-level lognormal model and derive a closed-form expression for the EB predictor of the population mean. In Section 3, we derive an approximately unbiased estimator of the MSE of the EB predictor. We evaluate the properties of the EB predictor and the proposed MSE estimator through simulation in Section 4. We compare the EB predictor to the synthetic estimator based on Karlberg (2000) and the model-based direct estimator of Chandra and Chambers (2011) in Section 4. In Section 5, we apply the lognormal model to obtain estimates of mean erosion rates in small domains defined by hydrologic units using data from the Conservation Effects Assessment Project (CEAP). Section 6 contains concluding remarks.

2. Empirical Bayes predictors for a unit-level lognormal model

We assume the units in the population have lognormal distributions and write the loglinear mixed model for the variable of interest, y_{ij}, as

$$\log(y_{ij}) := l_{ij} = \beta_0 + \boldsymbol{z}_{ij}\boldsymbol{\beta}_1 + u_i + e_{ij}, \tag{1}$$

where $(u_i, e_{ij}) \sim N(\mathbf{0}, \mathrm{diag}(\sigma_u^2, \sigma_e^2))$, and $\boldsymbol{z}_{ij}$ is a vector of covariates. Let the observations $\{(y_{ij}, \boldsymbol{z}_{ij}) : i = 1, \ldots, D; j \in s_i\}$ be available, where s_i denotes the set of j in the sample for area i, and $\mid s_i \mid = n_i$. Let U_i denote the set of N_i indexes in the population for area i, and let $\bar{s}_i$ denote the set of elements in area i that are not in the sample. Assume that $\boldsymbol{z}_{ij}$ is available for the population of N_i values in area i, and let $\{y_{ij}; i = 1, \ldots, D, j \in s_i\} \cup \{\boldsymbol{z}_{ij} : i = 1, \ldots, D, j \in U_i\}$ be the available data. The quantity of interest is the area mean,

$$\bar{y}_{N_i} = \frac{1}{N_i} \sum_{j \in U_i} y_{ij}. \tag{2}$$

2.1. Minimum MSE predictor

The minimum mean squared error (MMSE) predictor of $\bar{y}_{N_i}$ under model (1) is

$$E[\bar{y}_{N_i} \mid (\boldsymbol{y}, \boldsymbol{z})] = \frac{1}{N_i} \left\{ \sum_{j \in s_i} y_{ij} + \sum_{j \in \bar{s}_i} E[y_{ij} \mid (\boldsymbol{y}, \boldsymbol{z})] \right\}, \tag{3}$$