Computational Statistics and Data Analysis 56 (2012) 680–697

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Finite population estimation under generalized linear model assistance

Luz Marina Rondon [a,*], Luis Hernando Vanegas [a], Cristiano Ferraz [b]

[a] Universidad Nacional de Colombia, Facultad de Ciencias, Departamento de Estadística, Carrera 30 No. 45-03, Bogotá, Colombia

[b] Departamento de Estatística, CCEN-UFPE, Cidade Universitária, Recife, PE 50740-540, Brazil

ARTICLE INFO

Article history:
Received 29 January 2011
Received in revised form 31 August 2011
Accepted 14 September 2011
Available online 24 September 2011

Keywords:
Auxiliary information
Generalized linear models
Model-assisted estimation
Pseudo-maximum likelihood

ABSTRACT

Finite population estimation is the overall goal of sample surveys. When information regarding auxiliary variables are available, one may take advantage of general regression estimators (GREG) to improve sample estimates precision. GREG estimators may be derived when the relationship between interest and auxiliary variables is represented by a normal linear model. However, in some cases, such as when estimating class frequencies or counting processes means, Bernoulli or Poisson models are more suitable than linear normal ones. This paper focuses on building regression type estimators under a model-assisted approach, for the general case in which the relationship between interest and auxiliary variables may be suitably described by a generalized linear model. The finite population distribution of the variable of interest is viewed as if generated by a member of the exponential family, which includes Bernoulli, Poisson, gamma and inverse Gaussian distributions, among others. The resulting estimator is a generalized linear model regression estimator (GEREG). Its general form and basic statistical properties are presented and studied analytically and empirically, using Monte Carlo simulation experiments. Three applications are presented in which the GEREG estimator shows better performance than the GREG one.

1. Introduction

Unbiased estimation of finite population means, totals and percentages, based on sample surveys using probability sampling, can be accomplished by the Horvitz–Thompson estimator. When information regarding auxiliary variables is available, however, one may take advantage of general regression estimators (GREG) to improve sample estimates' precision. Such GREG estimators may be derived under a model assisted estimation approach (see, for example, Särndal et al., 2003). They also can be derived when the relationship between interest and auxiliary variables may be represented by a normal linear model, under a prediction approach for finite populations. Several authors, such as Bolfarine and Zacks (1992), Isaki and Fuller (1982), Wright (1983) and Fuller (2002), have been working on the subject of regression estimators. Nevertheless, since the efficiency of the GREG methodology lies on the ability of the formulated regression model to describe the concomitant behavior of the interest variable and the auxiliary ones, sometimes it is necessary to consider using models to deal appropriately with cases where the scale measurement of the interest variable is not continuous or it is continuous but highly skewed. Considering broader classes of models is also needed for cases in which the relationship between the interest variable and the auxiliary ones is clearly not linear. Lehtonen and Veijanen (1998), for example, considered deriving a Logistic General Regression Estimator (LGREG) using a multinomial response model to describe such relationship, for estimating class frequencies. Breidt and Opsomer (2000) focused on a nonparametric regression modeling assisted approach,

* Corresponding author. Tel.: +57 1 3165000.
E-mail address: lmrondonp@unal.edu.co (L.M. Rondon).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.09.017

Table 1
Principal distributions belonging to exponential family.

Distribution	$b(\theta)$	θ	ϕ	$V(\mu)$
Normal	$\theta^2/2$	μ	$1/\sigma^2$	1
Poisson	e^{θ}	$\log\mu$	1	μ
Bernoulli	$\log(1+e^{\theta})$	$\log\{\mu/(1-\mu)\}$	1	$\mu(1-\mu)$
Gamma	$-\log(-\theta)$	$-1/\mu$	$1/(\mathrm{CV})^2$	μ^2
Inverse G.	$-\sqrt{-2\theta}$	$-1/2\mu^2$	ϕ	μ^3

proposing and studying the properties of local polynomial regression estimators. Estevao and Särndal (2004) studied several applications of domain estimation using calibration. Duchesne (2003) compared the efficiency of an LGREG and a GREG based on a normal linear model using Monte Carlo simulation. Lehtonen et al. (2003) compared the performance of different estimators, including those assisted by the logistic regression model. Lehtonen et al. (2005) studied the importance of model specification for estimating the total of a polytomous interest variable for a number of large or small domains. Myrskylä (2007) investigated variance estimation for the LGREG estimator for domains. Li (2008) introduced the Box–Cox transformation into the generalized regression estimator, which can be especially suitable to deal with highly skewed continuous and positive interest variables, where normal response models may not be appropriate. Using a model-assisted estimation approach, this paper focuses on proposing regression estimators for the general case in which the relationship between interest and auxiliary variables may be appropriately described by a generalized linear model. The finite population distribution of the interest variable is viewed as if generated by a member of the exponential family distribution, which includes normal (even with variance heterogeneity), Bernoulli, Poisson, gamma and inverse Gaussian distributions. The resulting estimator is a generalized linear model regression estimator (GEREG). The GEREG setup is flexible enough to accommodate classical estimators, such as the ratio estimator, as particular cases. Therefore, it is expected that a GEREG estimator would perform at least as better as a GREG estimator on those situations where linear models would not fit the data as well as generalized linear models. This paper introduces the GEREG estimator in its general form and their statistical properties, taking into account a broad class of sampling designs. We consider the usual inference setup for finite population, based on a sample S, of size n, selected from a population, of size N ($n < N$), according to a probability sample with first and second-order inclusion probabilities given by $\pi_k = P(k \in S)$ and $\pi_{kl} = P(k, l \in S)$, respectively. y_k is defined to be the value of the interest variable, measured at element $k \in U$, for direct element sampling schemes. When considering a one-stage cluster design, we regard U as the set of clusters listed on a sampling frame, and y_k as the aggregated value of the variable of interest, for cluster $k \in U$. In any case, we denote by $\mathbf{x}_k = (x_{k1}, \ldots, x_{kJ})^T$ the auxiliary information vector for the element (or cluster) $k \in U$. The paper is organized in four sections: Section 1 is an introduction; Section 2 is a brief review of generalized linear models for finite population; in Section 3 we present the GEREG estimator and its main statistical properties. We also present a simulation study that illustrates the performance of the GEREG estimator. Section 4 presents three applications of the proposed estimator.

2. Generalized linear model in finite population

The literature concerning generalized linear models (GLM) is vast. McCullagh and Nelder (1989), Dobson (2001) and Fox (2008) are only a few examples of books presenting the subject. In a GLM setup, $Y_1, \ldots, Y_n$ are considered the values of an interest variable Y measured in n elements. The Y's are supposed to be independent random variables with probability distribution belonging to the exponential family. Let Y_k be the random variable Y for element k. Its density function may be expressed as

$$f(y; \theta_k, \phi_k) = \exp\{\phi_k[y\theta_k - b(\theta_k)] + c(y, \phi_k)\}, \tag{1}$$

where $c(\cdot)$ is a known function, θ_k is the canonical parameter, $\mathrm{E}(Y_k) = \mu_k = b'(\theta_k)$ is the expected value, $\mathrm{Var}(Y_k) = \phi_k^{-1}V(\mu_k)$, with $V_k = V(\mu_k) = \partial\mu_k/\partial\theta_k$ is the variance function, and $\phi_k^{-1} > 0$ is the dispersion parameter. Generalized linear models are defined by (1) and by the following systematic component

$$g(\mu_k) = \eta_k = \sum_{j=1}^{J} \beta_j x_{kj} = \mathbf{x}_k^T\boldsymbol{\beta}, \tag{2}$$

where $\mathbf{x}_k = (x_{k1}, \ldots, x_{kJ})^T$ is a vector of J explanatory variables measured at element k, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)^T$ a vector of unknown parameters, and $g(\cdot)$, a monotone differentiable function, called the link function. When $g(\cdot)$ is defined in such a way that $\theta_k = \eta_k$ for every k, then $g(\cdot)$ is called the canonic link function. It has been assumed in this work that if k and l are such that $\phi_k \neq \phi_l$, then $\phi_k \propto \delta_k$, for $k = 1, \ldots, n$, with δ_k being a known quantity for every k. Particular cases of such setup include normal, Bernoulli, Poisson, gamma and the inverse Gaussian response models. Table 1 shows $b(\theta)$, θ, ϕ and $V(\mu)$ forms for main exponential family distributions.

The maximum-likelihood method may be used to estimate GLM parameters. The estimation method can use sampling weights in the finite population context (see Nordberg, 1989). For instance, $\hat{\mu}_k^U$ and $\hat{\mu}_k^S$ are estimators of μ_k; however, the