Statistics and Probability Letters 105 (2015) 1–5

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Variance estimation in ranked set sampling using a concomitant variable

Ehsan Zamanzade [a,*], Michael Vock [b]

[a] Department of Statistics, University of Isfahan, Isfahan 81746-73441, Iran

[b] Institute of Mathematical Statistics and Actuarial Science, University of Bern, Sidlerstrasse 5, CH-3012 Bern, Switzerland

ARTICLE INFO

Article history:
Received 20 December 2014
Received in revised form 27 April 2015
Accepted 30 April 2015
Available online 9 June 2015

MSC:
primary 62D05
secondary 62G08
tertiary 62G05

Keywords:
Ranked set sampling
Judgment post stratification
Concomitant variable
Nonparametric regression

ABSTRACT

We propose a nonparametric variance estimator when ranked set sampling (RSS) and judgment post stratification (JPS) are applied by measuring a concomitant variable. Our proposed estimator is obtained by conditioning on observed concomitant values and using nonparametric kernel regression.

1. Introduction

Ranked set sampling (RSS), proposed by McIntyre (1952, 2005), is a sampling strategy which uses ranking information to give more efficient statistical inference than simple random sampling (SRS). To collect a balanced ranked set sample using set size k, one first draws a sample of size k^2 and then divides it into k samples of size k and ranks them in an increasing magnitude (without actually measuring them, i.e., by eye inspection or using a concomitant variable). One then selects for measurement the observation with rank r from the rth sample, for $r = 1, \ldots, k$. This process is repeated n times in order to obtain a sample of measured units of size $N = nk$. Therefore, a balanced ranked set sample consists of n independent measured units with judgment rank one, n independent measured units with judgment rank two, and so on. An unbalanced ranked set sample differs from the balanced one by the number of measured units with rank r not necessarily being the same for all ranks. In fact, if n_r is the number of observations with rank r, then $N = \sum_{r=1}^{k} n_r$ is the total sample size.

Another variation of RSS, proposed by MacEachern et al. (2004), is judgment post stratification (JPS). To collect a JPS sample of size N, using set size k, one first draws a simple random sample of size N and measures all N units. Then, for each measured unit, one draws a supplemental random sample of size $(k-1)$ from the population and finds the rank of the measured unit when it is added to this sample. Therefore, a JPS sample of size N consists of a simple random sample of size N and their corresponding ranks. The similarity of JPS sampling scheme and unbalanced RSS is that the number of units with rank r (n_r) is not constant. However, JPS differs from unbalanced RSS by the fact that the n_r's are not fixed

* Corresponding author.
E-mail addresses: Ehsanzamanzadeh@yahoo.com, E.Zamanzade@sci.ui.ac.ir (E. Zamanzade), Michael.vock@stat.unibe.ch (M. Vock).

http://dx.doi.org/10.1016/j.spl.2015.04.034
0167-7152/

in advance. In fact, if $(n_1, n_2, \ldots, n_k)$ is the vector of the numbers of units with rank i, then one can simply show that $(n_1, n_2, \ldots, n_k) \sim \text{Multinomial}\,(N, (1/k, 1/k, \ldots, 1/k))$.

A lot of research has been done on RSS and JPS since their introductions. In the RSS scheme, Takahasi and Wakimoto (1968) were the first who proved that the mean estimator from RSS is more efficient than that from SRS. Stokes (1980), MacEachern et al. (2002) and Perron and Sinha (2004) proposed different variance estimators. The problem of estimating a distribution function has been considered by Stokes and Sager (1988), Kvam and Samaniego (1994) and Duembgen and Zamanzade (2013). Frey et al. (2007) and Li and Balakrishnan (2008) proposed some tests for assessing the assumption of prefect rankings, followed by Vock and Balakrishnan (2011), Zamanzade et al. (2012), Frey and Wang (2013), and Zamanzade et al. (2014).

In the JPS sampling scheme, Wang et al. (2008) and Frey and Feeman (2012) proposed some mean estimators. The problem of estimating the population variance has been considered by Frey and Feeman (2013). Frey and Ozturk (2011) and Wang et al. (2012) proposed some distribution function estimators.

Frey (2011) proposed some mean estimators in RSS and JPS based on measuring a concomitant variable, showing how the values of the concomitant variable can be used more efficiently than just for ranking. A good review on existing literature on RSS and its variations is given by Wolfe (2012).

The rest of this paper is organized as follows: in Section 2, we propose a nonparametric variance estimator for ranked set samples and judgment post stratification based on a concomitant variable. Then, in Section 3, we compare the proposed estimator with its leading competitors in the literature. Some concluding remarks are provided in Section 4.

2. Introduction of the variance estimator

In this section, we introduce a new variance estimator for ranked set samples and judgment post stratified data in the case that the ranking is based on the measurement of a concomitant variable. Let Y, X be the variable of interest and the concomitant variable, respectively, let $\mathbf{X} = \{X_1, X_2, \ldots, X_{kN}\}$ be the full set of X values which is used in rankings, and let $X^m_{(1)} < X^m_{(2)} < \cdots < X^m_{(N)}$ be the X values corresponding to the measured units ($X^m_{(i)}$ to Y_i).

The concomitant variable information can be incorporated into the estimation of the population variance by using the identity $\text{Var}\,(Y) = E\left(\mu^2_X\right) - E^2\left(\mu^1_X\right)$, where $\mu^l_X = E\left(Y^l \mid X\right)$, for $l = 1, 2$. By the identities $E\left(Y^l\right) = E\left(\mu^l_X\right)$ and $\text{Var}\left(Y^l\right) = E\left(\text{Var}\left(Y^l \mid X\right)\right) + \text{Var}\left(\mu^l_X\right)$, for $l = 1, 2$, the estimates of $E\left(\mu^l_X\right)$ $(l = 1, 2)$ can be used for the estimation of $E\left(Y^l\right)$ $(l = 1, 2)$, and the estimates based on this conditioning are expected to have less variance than direct estimates.

The parameters $E\left(\mu^l_X\right)$ $(l = 1, 2)$ can be estimated by taking the average over the kN estimates of $E\left(Y^l \mid X_{(i)}\right),\ i = 1, 2, \ldots, kN,\ l = 1, 2$. We propose to estimate the quantities $E\left(Y^l \mid X = x\right)$ $(l = 1, 2)$ by using nonparametric kernel regression. Let $k\,(.)$ be a kernel function and $h > 0$ the bandwidth, then these quantities can be estimated using the weighted average:

$$m\,(x, h_l) = \frac{\sum_{i=1}^{N} Y_i^l k\left(\frac{x - X^m_{(i)}}{h_l}\right)}{\sum_{i=1}^{N} k\left(\frac{x - X^m_{(i)}}{h_l}\right)}, \quad l = 1, 2.$$

We use the standard normal density function as the kernel function, and since we want to use the regression equation for prediction, we propose to select the value for the bandwidth based on the cross-validation (CV) method. I.e., we select the value of the bandwidth (h_l) to minimize $CV\,(h_l) = \frac{1}{N}\sum_{i=1}^{N}\left(Y^l_i - m^{-i}\left(X^m_{(i)}, h_l\right)\right)^2$, where $m^{-i}\,(x, h_l)$ is the estimate of the regression equation without using the observation $\left(Y^l_i, X^m_{(i)}\right)$ $(i = 1, 2, \ldots, N)$. Therefore, we can consider the value of h^{cv}_l $(l = 1, 2)$ that is selected by the CV method as the best "predictive" estimate for h_l $(l = 1, 2)$. So, we propose to estimate the population variance when RSS and JPS are applied by using a concomitant variable by

$$\hat{\sigma}^2_N = \frac{1}{kN}\sum_{i=1}^{kN} m\left(X_i, h^{cv}_2\right) - \left(\frac{1}{kN}\sum_{i=1}^{kN} m\left(X_i, h^{cv}_1\right)\right)^2.$$

Remark 1. It can be shown (see for example Takezawa, 2006, p. 117) that $CV\,(h_l)$ can be computed efficiently by using the relation $CV\,(h_l) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y^l_i - m(X_i, h_l)}{1 - H^{h_l}_{ii}}\right)^2$, where the $H^{h_l}_{ii}$'s are the diagonal members of the hat matrix. In the simulation study in the next section, h_l is selected from a sequence of values in the interval $[N^{-\frac{1}{4}}/3, 3N^{-\frac{1}{4}}]$ with steps of 0.01, where $N^{-\frac{1}{4}}$ is the bandwidth that is used by Frey (2011).

3. Monte Carlo comparisons

In this section, we compare the proposed estimator with its leading competitors in RSS and JPS settings. For this purpose, we use the imperfect ranking model proposed by Dell and Clutter (1972), assuming (Y, X) follows a standard bivariate nor-