Influence diagnostic in survey sampling:
Estimating the conditional bias

J. L. Moreno-Rebollo, A. Muñoz-Reyes, M. D. Jiménez-Gamero and J. Muñoz-Pichardo

Departamento de Estadística e Investigación Operativa, Facultad de Matemáticas, Universidad de Sevilla, C/Tarifa s.n., 41012 Sevilla, Spain.
(e-mail: moreno@cica.es anamun@cica.es dolores@cica.es pichardo@cica.es)

Received October 2000

Abstract. The conditional bias has been proposed by Moreno Rebollo et al. (1999) as an influence diagnostic in survey sampling, when the inference is based on the randomization distribution generated by a random sampling. The conditional bias is a population parameter. So, from an applied point of view, it must be estimated. In this paper, we propose an estimator of the conditional bias and we study conditions that guarantee its unbiasedness. The results are applied in a Simple Random Sampling and in a Proportional Probability Aggregated Size Sampling, when the ratio estimator is used.

Key words: Influence diagnostic; Conditional bias; Survey Sampling

1 Introduction

Although in the literature many influence diagnostics have been proposed for assessing the impact that individual observations have on statistical conclusions, there are no many references to influence diagnostics in sample survey, specially within a design-based approach. In this context, we can cite the papers by Smith (1987), Gwet and Rivest (1992), and Hulliger (1995).

Moreno-Rebollo et al. (1999) have proposed the conditional bias as an influence measure that it can be used for arbitrary estimators and sampling designs. Let \( U = \{u_1, \ldots, u_N\} \) be a finite population and \( D = \{S, p(\cdot)\} \) a sampling design defined on \( U \), being \( S \) the sample space and \( p(\cdot) \) the probability distribution on \( S \). Let \( Y \) be a characteristic of the population, \( Y = \{Y_1, \ldots, Y_N\} \), \( \theta = \theta(Y) \) the parameter of interest and \( \hat{\theta} = \hat{\theta}(s) \) an estimator of \( \theta \), \( s \in S \). Denoting by \( I_i(s) \), \( i = 1, \ldots, N \), the random variables, \( I_i(s) = 1 \) if \( u_i \in s \); \( I_i(s) = 0 \) otherwise, and by \( \pi_i \) the first-order inclusion probabilities, \( \pi_i = p(I_i = 1) \), \( i = 1, \ldots, N \), the conditional bias of \( \theta \) caused by the presence of \( u_i \), \( 0 < \pi_i < 1 \), in the sample, \( \mathcal{I}(I_i = 1; \theta) \), is given by
\[ \mathcal{J}(I_{1} = 1; \hat{\theta}) = E(\hat{\theta}/I_{1} = 1) - E(\hat{\theta}) \\
= (1 - \pi_{1})(E(\hat{\theta}/I_{1} = 1) - E(\hat{\theta}/I_{1} = 0)). \tag{1} \]

We will consider, for convenience in notation, the conditional bias due to the presence of unit \( u_{1} \).

Obviously, the conditional bias is an unknown population parameter, depending on the sampling design, \( D \), and the considered estimator, \( \hat{\theta} \). Moreno-Rebollo et al. (1999) have proposed to estimate \( \mathcal{J}(I_{1} = 1; \hat{\theta}) \) based on the conditional sampling design, \( D_{1} = \{S_{1}, p_{1}(\cdot)\} \), with \( S_{1} = \{s \in S; u_{1} \not\in s\} \) and probability distribution

\[ p_{1}(s) = \frac{p(s)}{\pi_{1}}, \quad \forall s \in S_{1}, \]

since the interest is the assessment of the influence of a population unit, \( u_{1} \), that belongs to the sample. In that article this subject is treated with Horvitz-Thompson estimator.

In this paper, we propose an estimator of the conditional bias in a more general setup, studying conditions under which it is unbiased. In order to illustrate the obtained results, we first determine the proposed estimator in a simple random sampling design, obtaining a case deletion diagnostic. In particular, when the ratio estimator is used, we compare the results with the ones obtained by Gwet and Rivest (1992) in the same framework. Both influence measures are expressed in terms of the residuals with respect to a linear model through the origin, which is the implicit basis of the ratio estimator. The proposed estimator is also obtained in a PPAS (Probability Proportional Aggregated Size) design. In this sampling design, the estimator of the conditional bias is not a case deletion statistic. When the ratio estimator is used, the conditional bias and the proposed estimator, that in this case is unbiased, are also expressed in terms of a residual with respect to a linear model through the origin.

2 Estimation: A general approach

The aim of this Section is to propose an estimator of the conditional bias in a more general setup that the one considered in Moreno-Rebollo et al. (1999) and to study its unbiasedness. We will suppose that the sampling design is of fixed size \( n \),

\[ D^{(n)}(U) = \{S^{(n)}(U), p^{(n)}_{U}(\cdot)\}, \]

and that all samples of size \( n \) belong to \( S^{(n)}(U) \), although the probability of some of them could be null. From \( D^{(n)}(U) \), we consider the conditional sampling designs \( D^{(n)}_{1}(U) = \{S^{(n)}_{1}(U), p^{(n)}_{1U}(\cdot)\} \), defined in Section 1, and \( D^{(n)}_{(1)}(U) = \{S^{(n)}_{(1)}(U), p^{(n)}_{(1)U}(\cdot)\} \), with sample space \( S^{(n)}_{(1)}(U) = \{s \in S^{(n)}(U); u_{1} \not\in s\} \), and probability distribution given by