optimStrat: An R package for assisting the choice of the sampling strategy

Keywords: *survey sampling, stratified sampling, probability propotional-to-size sampling, sampling strategy, R, Shiny*

1 Introduction

We are interested in the estimation of the total of a study variable. One auxiliary variable is available and it can be used for obtaining an efficient strategy, where efficiency will be understood in terms of design-based variance.
The strategy that couples proportional-to-size sampling with the regression estimator (denoted πps–reg) has sometimes been called optimal (Särndal et al. [1], Brewer [2], Isaki and Fuller [3]). This optimality, however, relies on a superpopulation model (section 2) which might not (and most certainly will not) hold exactly in practice. Using the same model, Wright [4] proposed strong model-based stratification, a strategy that couples stratified simple random sampling with the regression estimator. The aim of this paper is to compare these strategies and show that if the model is misspecified, πps–reg is not optimal anymore.
We then propose some statistics that can be computed at the design stage of the survey and can be used for making the decision about which strategy to employ. These statistics are approximations to the anticipated variance obtained by assuming that the finite population size tends to infinity. However, this assumption can be dropped and the approximations can be obtained by simulation, instead. This is the approach implemented in *optimStrat*, a package developed for the statistical software environment R [5].

2 Framework

The aim is to estimate the total $t_y = \sum_U y_k$ of one study variable $\boldsymbol{y}' = (y_1, y_2, \cdots, y_N)$ in a population U with unit labels $\{1, 2, \cdots, N\}$ where N is known. It is assumed that there is one auxiliary variable $\boldsymbol{x}' = (x_1, x_2, \cdots, x_N)$, $x_k > 0$, known for each element in U. A without-replacement sample s of size n is selected and y_k is observed for all units $k \in s$.
Six strategies will be described in this section. We will assume that when defining the sampling strategy, the statistician is willing to admit that the following model *adequately describes* the relation between the study variable, $\boldsymbol{y}$, and the auxiliary variable, $\boldsymbol{x}$. The values of the study variable $\boldsymbol{y}$ are realizations of the model ξ_0

$$Y_k = \delta_0 + \delta_1 x_k^{\delta_2} + \epsilon_k \tag{1}$$

The error terms ϵ_k are random variables satisfying

$$\mathrm{E}_{\xi_0}[\epsilon_k] = 0 \qquad \mathrm{V}_{\xi_0}[\epsilon_k] = \delta_3^2 x_k^{2\delta_4} \qquad \mathrm{E}_{\xi_0}[\epsilon_k \epsilon_l] = 0 \ (k \neq l)$$

where the moments are taken with respect to the model ξ_0, and δ_i are constant parameters.
Model ξ_0 as defined above is then used for assisting the definition of the sampling strategy as follows.

Strategy 1, πps(δ_4)–reg(δ_2) At the design stage consider πps with $\pi_k = n\frac{x_k^{\delta_4}}{t_{x^{\delta_4}}}$. At the estimation stage consider the reg-estimator with $\boldsymbol{x}_k = (1, x_k^{\delta_2})$.
Strategy 2, STSI(δ_4)–reg(δ_2) At the design stage consider STSI with strata defined by using the cum$\sqrt{f}$-rule on $x_k^{\delta_4}$ and Neyman allocation. At the estimation stage consider the reg-estimator with $\boldsymbol{x}_k = (1, x_k^{\delta_2})$.
Strategy 3, STSI(δ_2)–HT At the design stage consider STSI with strata defined by using the cum$\sqrt{f}$-rule on $x_k^{\delta_2}$ and Neyman allocation. At the estimation stage consider the HT estimator.
Strategy 4, πps(δ_4)–pos(δ_2) At the design stage consider πps with $\pi_k = n\frac{x_k^{\delta_4}}{t_{x^{\delta_4}}}$. At the estimation stage consider the pos-estimator with poststrata defined by using the cum$\sqrt{f}$-rule on $x_k^{\delta_2}$.
Strategy 5, STSI(δ_4)–pos(δ_2) At the design stage consider STSI with strata defined by using the cum$\sqrt{f}$-rule on $x_k^{\delta_4}$ and Neyman allocation. At the estimation stage consider the pos-estimator with poststrata defined by using the cum$\sqrt{f}$-rule on $x_k^{\delta_2}$.

3 The case of a misspecified model

First we will define how "misspecification"shall be understood in this paper. ξ_0 (which from now on will be called *working model*) reflects the knowledge or beliefs the statistician has about the relation between $\boldsymbol{x}$ and $\boldsymbol{y}$ at the design stage. Nevertheless, one hardly believes that this is the true generating model. We will assume that this true model exists but it is unknown to the statistician. It will be denoted by ξ. Any deviation of ξ_0 with respect to ξ is a misspecification of the model. As this definition is too wide and in order to keep the analysis tractable, we will limit ourselves to a very simple type of misspecification, which is when the working model is of the form (1) and the true model, ξ, is

$$Y_k = \beta_0 + \beta_1 x_k^{\beta_2} + \epsilon_k \qquad \text{with } \mathrm{E}_\xi[\epsilon_k] = 0 \qquad \mathrm{V}_\xi[\epsilon_k] = \beta_3^2 x_k^{2\beta_4} \qquad \mathrm{E}_\xi[\epsilon_k \epsilon_l] = 0 \ (k \neq l) \tag{2}$$

with $\beta_2 \neq \delta_2$ or $\beta_4 \neq \delta_4$.
The following result establishes an approximation to the anticipated variance for each of the five strategies defined in section 2.

Result 1 *If ξ_0 is assumed when ξ is the true model, then:*

1) the anticipated variance of $\pi ps(\delta_4)$–reg(δ_2) can be approximated by

$$AA_{\xi,\pi ps}\left[\hat{t}_{reg}\right] \equiv \beta_1^2 \frac{N^2 \overline{x^{\delta_4}}}{n} \left(\left(S_{\beta_2,\beta_2-\delta_4} - \overline{x^{\beta_2}} S_{\beta_2,-\delta_4} \right) - 2\frac{S_{\beta_2,\delta_2}}{S_{\delta_2,\delta_2}} (S_{\beta_2,\delta_2-\delta_4} - \overline{x^{\delta_2}} S_{\beta_2,-\delta_4}) + \frac{S_{\beta_2,\delta_2}^2}{S_{\delta_2,\delta_2}^2} (S_{\delta_2,\delta_2-\delta_4} - \overline{x^{\delta_2}} S_{\delta_2,-\delta_4}) + F_0 \overline{x^{2\beta_4-\delta_4}} \right) \tag{3}$$

2) the anticipated variance of STSI(δ_4)–reg(δ_2) can be approximated by

$$AA_{\xi,STSI}\left[\hat{t}_{reg}\right] \equiv \beta_1^2 \sum_h \frac{N_h^2}{n_h} \left(\left(S_{\beta_2,\beta_2,U_h} - 2\frac{S_{\beta_2,\delta_2}}{S_{\delta_2,\delta_2}} S_{\beta_2,\delta_2,U_h} + \frac{S_{\beta_2,\delta_2}^2}{S_{\delta_2,\delta_2}^2} S_{\delta_2,\delta_2,U_h} \right) + F_0 \overline{x_{U_h}^{2\beta_4}} \right) \tag{4}$$

3) the anticipated variance of STSI(δ_2)–HT can be approximated by

$$AA_{\xi,STSI}\left[\hat{t}_{HT}\right] \equiv \beta_1^2 \sum_h \frac{N_h^2}{n_h} \left(S_{\beta_2,\beta_2,U_h} + F_0 \overline{x_{U_h}^{2\beta_4}} \right) \tag{5}$$

4) the anticipated variance of $\pi ps(\delta_4)$–$pos(\delta_2)$ can be approximated by

$$AA_{\xi,\pi ps}\left[\hat{t}_{pos}\right] \equiv \beta_1^2 \frac{N\overline{x^{\delta_4}}}{n}\left(\sum_g N_g\left(S_{\beta_2,\beta_2-\delta_4,U_g} - \overline{x_{U_g}^{\beta_2}} S_{\beta_2,-\delta_4,U_g}\right) + NF_0\overline{x^{2\beta_4-\delta_4}}\right) \tag{6}$$

5) Let $U_{hg} = U_h \cap U'_g$ be the intersection between the hth stratum and the gth poststratum. The anticipated variance of STSI(δ_4)–pos(δ_2) can be approximated by

$$AA_{\xi,STSI}\left[\hat{t}_{pos}\right] \equiv \beta_1^2 \sum_h \frac{N_h^2}{n_h}\left(\frac{1}{N_h}\sum_g N_{hg}\left(\overline{x_{U_{hg}}^{2\beta_2}} - 2\overline{x_{U'_g}^{\beta_2}}\,\overline{x_{U_{hg}}^{\beta_2}} + \overline{x_{U'_g}^{\beta_2}}^2\right) - \frac{1}{N_h^2}\left(\sum_g N_{hg}\left(\overline{x_{U_{hg}}^{\beta_2}} - \overline{x_{U'_g}^{\beta_2}}\right)\right)^2 + F_0\overline{x_{U_h}^{2\beta_4}}\right) \tag{7}$$

It can be seen that even under this simple misspecification of the model, πps(δ_4)–reg(δ_2) does not minimize the approximation to the anticipated variance. Let us compare, for example, (3) and (4) in the case where $\delta_2 = \beta_2$. In that case we get

$$\mathrm{AA}_{\xi,\pi\mathrm{ps}}\left[\hat{t}_{\mathrm{reg}}\right] = \beta_1^2 \frac{N^2}{n} F_0 \overline{x^{\delta_4}}\,\overline{x^{2\beta_4-\delta_4}} \qquad \text{and} \qquad \mathrm{AA}_{\xi,\mathrm{STSI}}\left[\hat{t}_{\mathrm{reg}}\right] = \beta_1^2 F_0 \sum_h \frac{N_h^2}{n_h}\overline{x_h^{2\beta_4}}$$

If we allow $S_{\delta_4,\delta_4,h}$ constant, and take into account that Neyman optimal allocation was used, we get that $\mathrm{AA}_{\xi,\pi\mathrm{ps}}\left[\hat{t}_{\mathrm{reg}}\right] > \mathrm{AA}_{\xi,\mathrm{STSI}}\left[\hat{t}_{\mathrm{reg}}\right]$ when $2\beta_4 < \delta_4$ and $\delta_4 > 0$.
The assumption of $N \to \infty$ in result 2 is actually required only for getting closed and nice expressions that can be printed in a paper. Alternatively, one could resort to simulations in order to approximate the anticipated variance. We have developed an R package that performs these type of simulations. It will be described in the next section.

4 optimStrat

Table 1 briefly describes some of the functions in optimStrat. The main functions, stratvar and optimApp, are described below. The package is available from CRAN at https://CRAN.R-project.org/package=optimStrat.

4.1 stratvar

stratvar simulates values of a study variable using $\boldsymbol{x}$ and the superpopulation model ξ via simulatey. Then, the variance of the five strategies defined at the end of section 2 are computed assuming model ξ_0 instead. The process is iterated it times. Alternatively, a positive integer can be given for x instead of a vector. In that case, x observations from a gamma distribution with skewness equal to sk and mean equal to 48, plus one unit, are generated as auxiliary variable.
The output of the function is a data frame with dimension it×17 where each row corresponds to the results of each iteration. The first eleven columns are the arguments, followed by the correlation between $\boldsymbol{x}$ and the simulated y-values. The last five columns show the variances of the five sampling strategies.

Function	Description
simulatey	Simulate values for the study variable based on the auxiliary variable x and the parameters of the superpopulation model.
stratify	Stratify the auxiliary variable, x, into H strata using the cum-sqrt-rule.
varstsi	Compute the design variance of the HT estimator of the total of y under STSI using Neyman allocation with respect to x.
varpips	Compute the design variance of the HT estimator of the total of y under πps proportional to x.
stratvar	Simulate a study variable using simulatey. Then compute the design variance of five sampling strategies. The process is iterated it times.
optimApp	Call shiny to run a web-based application of stratvar.

Table 1: Main functions in package optimStrat.

4.2 optimApp

This function calls shiny [6] to run a web-based application of stratvar. An online version can be found at https://embuenoc.shinyapps.io/180426_shinyapp/.

5 Conclusions

The strategy that couples πps with the regression estimator is optimal when the superpopulation model exists and some of its parameters are known.
Taking into account how strong these assumptions are, it was shown in section 3 that this optimality breaks down when there is a misspecification of the model. Approximations to the anticipated variance of five strategies were obtained for a simple type of misspecification. They were used to verify that πps–reg is not necessarily optimal anymore. In fact its use may lead to variances many times bigger than some other strategies, e.g. STSI–reg, that seem to be more robust.
Package optimStrat provides tools that can be used at the design stage of a survey, e.g. simulatey, stratify, varstsi and varpips. In particular, stratvar –and its interactive web-based application, optimApp– provides a simple approach for choosing the sampling strategy to implement in a survey when auxiliary information is available. All that is required from the user is the auxiliary variable itself and some prior "knowledge" about its association with the unknown study variable. This prior knowledge is defined through the parameters δ_2 and δ_4 in (1). If the model was correct, πps–reg would be the optimal strategy to implement. However, taking into account the uncertainty about this prior knowledge, the package allows to compare five sampling strategies assuming that the true model that relates the auxiliary variable with the study variable is (2) instead.
The method is easy to implement, even if the user is not familiar with programming in R, thanks to the interactive application optimApp, which can also be found at https://embuenoc.shinyapps.io/180426_shinyapp/.

References

[1] C.E. Särndal, B. Swensson, and J. Wretman. *Model assisted survey sampling.* Springer, 1992.

[2] K.R.W. Brewer. A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, 5(1):5–13, 1963. DOI: 10.1111/j.1467-842x.1963.tb00132.x.

[3] C.T. Isaki and W.A. Fuller. Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96, 1982. DOI: 10.2307/2287773.

[4] R.L. Wright. Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78(384):879–884, 1983. DOI: 10.1080/01621459.1983.10477035.

[5] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL https://www.R-project.org/.

[6] W. Chang, J. Cheng, JJ Allaire, Y. Xie, and J. McPherson. *shiny: Web Application Framework for R*, 2018. URL https://CRAN.R-project.org/package=shiny. R package version 1.1.0.