Statistica Sinica **22** (2012), 777-794
doi:http://dx.doi.org/10.5705/ss.2010.238

BIAS-ROBUSTNESS AND EFFICIENCY OF MODEL-BASED INFERENCE IN SURVEY SAMPLING

Desislava Nedyalkova and Yves Tillé

University of Neuchâtel

Abstract: In model-based inference, the selection of balanced samples has been considered to give protection against misspecification of the model. A recent development in finite population sampling is that balanced samples can be randomly selected. There are several possible strategies that use balanced samples. We give a definition of balanced sample that embodies overbalanced, mean-balanced, and π-balanced samples, and we derive strategies in order to equalize a d-weighted estimator with the best linear unbiased estimator. We show the value of selecting a balanced sample with inclusion probabilities proportional to the standard deviations of the errors with the Horvitz-Thompson estimator. This is a strategy that is design-robust and efficient. We show its superiority compared to other strategies that use balanced samples in the model-based framework. In particular, we show that this strategy is preferable to the use of overbalanced samples in the polynomial model. The problem of bias-robustness is also discussed, and we show how overspecifying the model can protect against misspecification.

Key words and phrases: Balanced sampling, finite population sampling, polynomial model, ratio model, robust estimation.

1. Introduction

The principal difference between the model-based and the classical design-based approach for estimating finite population totals lies in the source of randomness they use (Särndal (1978)). In design-based sampling, the inference is based on the stochastic structure induced by the sampling design. In the model-based, or prediction approach, the inference depends on the validity of the model used to describe the data. In this case, the randomness is due to the population model and not to the sampling design.

The model-based approach was developed, among others, by Royall (1976, 1992), Royall and Cumberland (1981), and Chambers (1996). When the data are assumed to follow a linear model, Royall (1976) proposed the use of the best linear unbiased predictor. The model-based approach has been criticized due to the fact that it may lead to severe bias if the model assumptions are violated. In contrast to model-based inference, design-based inference is design-robust by

definition. Brewer and Särndal (1983) point out that, since the inference is not based on a model, there is no need to worry about model misspecification.

Much of the work in model-based research has been devoted to the construction of robust strategies. More specifically, in order to protect the inference against a misspecified model, Royall and Herson (1973a,b) and Scott, Brewer, and Ho (1978) point out the importance of *balanced* samples, where balance is achieved by equalizing the sample moments of the independent variables with those in the population. They came to the conclusion that the sample must be *balanced*, but not necessarily random.

Another way to accomplish design-robustness in the model-based approach is to choose an appropriate sampling design. Since Deville and Tillé (2004)'s paper, it is now possible to randomly select balanced samples using a procedure called the cube method. Nedyalkova and Tillé (2008) have shown that under a random balanced sampling design, with inclusion probabilities proportional to the standard deviations of the errors of the model, and under certain conditions defined as 'fully explainable heteroscedasticity', the best linear unbiased estimator is the Horvitz-Thompson estimator. This is an optimal strategy that reconciles the two approaches.

Scott, Brewer, and Ho (1978) and Royall and Herson (1973a) recommend the use of balanced samples in order to protect against a misspecification of the model, while Nedyalkova and Tillé (2008) recommend using balanced sampling for minimizing the anticipated variance under a linear model. An important particular case is the polynomial model. Scott, Brewer, and Ho (1978) suggest using overbalanced samples with an ad-hoc weighting system.

We investigate the different strategies leading to design-robust estimations under the model-based framework, and consider the polynomial model, overbalancing and robustness under misspecification in light of Nedyalkova and Tillé (2008). The notation and definitions are given in Section 2. The model-based framework is briefly introduced in Section 3. In Section 4, we consider a large class of balanced designs, called d-balanced designs, that embody balanced, π-balanced, mean-balanced, and overbalanced samples. We also consider a class of weighted estimators, called d-weighted estimators, and discuss several strategies for which the Best Linear Unbiased (BLU) estimator is the d-estimator. These results generalize some results of Nedyalkova and Tillé (2008) for the Horvitz-Thompson estimator. In Section 5, we give a definition of design-robustness and show that an appropriate strategy to protect against misspecification of the model consists of overspecifying the model and then balancing on the independent variables of the model. In Section 6, we revisit the polynomial model to show that the overbalancing strategy of Scott, Brewer, and Ho (1978) is suboptimal, and to give an alternative strategy that minimizes the anticipated variance.

In Section 7, we draw some general conclusions on selecting a sample when one reasonably believes in a linear model but seeks protection against misspecification.

2. Notation and Definitions

Consider a population U of size N. Each unit of the population is identified by a label $k = 1, \ldots, N$. Suppose that a register is available, and that the values of p auxiliary variables are known for each unit of the population. Let y_k be the value taken by the variable of interest y on the kth unit of the population. The values y_k are unknown. We are interested in estimating the population total $Y = \sum_{k \in U} y_k$. The total Y is estimated by a sample s of size n, where s is a subset of U. A sample is only a subset of the population and is not necessarily randomly selected.

A sampling design is a tool to randomly select a sample. It is defined by assigning to each sample s a probability $p(s)$ of being selected. Let S denote the random sample such that $\Pr(S = s) = p(s)$. The inclusion probability π_k is then the probability that unit k is selected in the sample. We denote by $\mathrm{E}_p(\cdot)$ and $\mathrm{var}_p(\cdot)$, respectively, the expectation and variance under the sampling design $p(\cdot)$, and by $\bar{S}$ the set of units of the population which are not in S.

Definition 1. An estimator $\widehat{Y}$ is design-unbiased if $\mathrm{E}_p(\widehat{Y}) = Y$.

Definition 2. A sample s is d-balanced on a set of variables $\mathbf{x}_k' = (x_{k1} \cdots x_{kp})$ if and only if

$$\sum_{k \in s} d_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k,$$

where $d_1, \ldots, d_k, \ldots, d_N$ is a set of weights that do not depend on the sample s.

The weights $d_1, \ldots, d_k, \ldots, d_N$ are positive values. An important example is given by the $d_k = 1/\pi_k$ used in the Horvitz-Thompson estimator.

When the sample is randomly selected, an inclusion probability can be assigned to each statistical unit. If $\pi_k > 0$, for all $k \in U$, the Horvitz-Thompson estimator of Y,

$$\widehat{Y}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k},$$

is design-unbiased, i.e. $\mathrm{E}_p(\widehat{Y}_\pi) = Y$.

For a balanced sample, the inclusion probabilities are $\pi_k = 1/d_k$, and the procedure that randomly selects a balanced sample is called a balanced sampling design. According to Deville and Tillé (2004), a sampling design $p(\cdot)$ is balanced on the auxiliary variables $x_1, \ldots, x_p$ if and only if

$$\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k. \tag{2.1}$$

Authors such as Cumberland and Royall (1981) and Kott (1986) would call this a 'π-balanced sampling', as opposed to a mean-balanced sampling defined by the equation

$$\frac{1}{n}\sum_{k\in S} x_k = \frac{1}{N}\sum_{k\in U} x_k.$$

We use the expression 'balanced sampling' to denote a sampling design that satisfies (2.1) for one or more auxiliary variables, a mean-balanced sampling being a particular case of this balanced sampling when the sample is selected with inclusion probabilities n/N. If the population size is small, a balanced sampling design can be implemented by a linear program. For larger population sizes, the cube method may be used (see Deville and Tillé (2004) or Tillé (2006)). Whatever the algorithm used for selecting a balanced sample, an exact balanced sample cannot generally be found because it does not exist. It is however, always possible to select a sample that is almost balanced. We assume that the balancing error can be neglected. The selection of a balanced sample requires the use of a register that contains the values of the auxiliary variables for each unit of the population.

3. Model-based Strategy and BLU Estimator

We assume that the population follows a linear model M,

$$y_k = \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_k, \tag{3.1}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients and $\boldsymbol{\varepsilon}$ is a vector of random variables ε_k such that

$$\mathrm{E_M}(\varepsilon_k) = 0, \mathrm{var_M}(\varepsilon_k) = \sigma^2\nu_k^2, \mathrm{cov_M}(\varepsilon_k, \varepsilon_\ell) = 0 \text{ if } k \neq \ell,$$

Suppose $\nu_k, k \in U$, are known. For simplicity, we scale them so that $\sum_{k\in U}\nu_k = N$.

Model (3.1) includes the possibility of heteroscedasticity. The heteroscedasticity of ν_k or ν_k^2 may or may not be proportional to some auxiliary variables included in $\mathbf{x}_k$. Nedyalkova and Tillé (2008) have shown the importance of using auxiliary variables whose linear combination is equal to ν_k or ν_k^2. Indeed, in this case, the optimal strategies for the model-based and design-based frameworks are the same.

An important and common hypothesis is that the random sample S and the errors ε_k of the model are independent. The symbols $\mathrm{E_M}(\cdot)$, $\mathrm{var_M}(\cdot)$, $\mathrm{cov_M}(\cdot)$ denote, respectively, the expected value, the variance, and the covariance under M.

Definition 3. An estimator $\widehat{Y}$ is model-unbiased if $\mathrm{E_M}(\widehat{Y} - Y) = 0$.

Definition 4. The model mean squared error of an estimator $\widehat{Y}$ is $\mathrm{MSE}_M(\widehat{Y}) = \mathrm{E_M}\left(\widehat{Y} - Y\right)^2$.

When $\widehat{Y}$ is model-unbiased, the model mean squared error is also called the error variance, for instance in Royall and Cumberland (1981).

Definition 5 (Isaki and Fuller (1982))**.** The anticipated mean squared error of an estimator $\widehat{Y}$ is $\mathrm{MSE}_{pM}(\widehat{Y}) = \mathrm{E}_p\mathrm{E_M}(\widehat{Y} - Y)^2$.

When $\widehat{Y}$ is design-unbiased, the anticipated mean squared error is also called the anticipated variance.

Royall (1976) showed that, in the framework of model-based inference, the Best Linear Unbiased (BLU) estimator is

$$\begin{aligned}\widehat{Y}_{\mathrm{BLU}} &= \sum_{k\in S} y_k + \sum_{k\in \bar{S}} \mathbf{x}_k'\widehat{\boldsymbol{\beta}}_{\mathrm{BLU}} = \sum_{k\in U} \mathbf{x}_k'\widehat{\boldsymbol{\beta}}_{\mathrm{BLU}} + \sum_{k\in S}(y_k - \mathbf{x}_k'\widehat{\boldsymbol{\beta}}_{\mathrm{BLU}}) \\ &= \sum_{k\in U} \mathbf{x}_k'\widehat{\boldsymbol{\beta}}_{\mathrm{BLU}} + \sum_{k\in S} e_k, \qquad (3.2)\end{aligned}$$

$$e_k = y_k - \mathbf{x}_k'\widehat{\boldsymbol{\beta}}_{\mathrm{BLU}}, \qquad (3.3)$$

where

$$\widehat{\boldsymbol{\beta}}_{\mathrm{BLU}} = \mathbf{A}^{-1}\sum_{k\in S}\frac{\mathbf{x}_k y_k}{\nu_k^2},$$

$$\mathbf{A} = \sum_{k\in S}\frac{\mathbf{x}_k\mathbf{x}_k'}{\nu_k^2}.$$

The error variance of the best linear unbiased estimator is

$$\mathrm{E_M}(\widehat{Y}_{\mathrm{BLU}} - Y)^2 = \sigma^2\left(\sum_{k\in\bar{S}}\mathbf{x}_k'\mathbf{A}^{-1}\sum_{\ell\in\bar{S}}\mathbf{x}_\ell + \sum_{k\in\bar{S}}\nu_k^2\right).$$

We refer to the usual definition given, for instance, in Hájek (1959, 1981), Ramakrishnan (1975), and Joshi (1979).

Definition 6. A strategy is a pair $(p(\cdot), \widehat{Y})$ consisting of a sampling design and an estimator.

Strategy 1. *Use the best linear unbiased estimator and choose a sample of size n that minimizes* $\mathrm{E}_M(\widehat{Y}_{\mathrm{BLU}} - Y)^2$.

This is an optimal, purely unbiased strategy that is not design-robust because, in some cases, it can lead to the choice of a very extreme sample, as in the

following example. Suppose the model has no intercept and only one regressor (see for instance, Royall and Herson (1973a)):

$$y_k = x_k\beta + \varepsilon_k, \tag{3.4}$$

with $\mathrm{var}_{\mathrm{M}}(\varepsilon_k) = \sigma^2\nu_k^2$ and $\nu_k^2 \propto x_k$. Then the BLU estimator is

$$\widehat{Y}_{\mathrm{BLU}} = \frac{\sum_{k\in U} x_k}{\sum_{k\in S} x_k} \sum_{k\in S} y_k,$$

which, in this case, is the ordinary ratio estimator $\widehat{Y}_{\mathrm{R}}$. As $\nu_k^2 \propto x_k$ and $\sum_{k\in U}\nu_k = N$, we have that

$$\nu_k^2 = \frac{N^2 x_k}{\left(\sum_{k\in U}\sqrt{x_k}\right)^2}.$$

The error variance of $\widehat{Y}_{\mathrm{R}}$ is given by the expression

$$\mathrm{E}_{\mathrm{M}}(\widehat{Y}_{\mathrm{R}} - Y)^2 = \sigma^2 \frac{N^2}{\left(\sum_{k\in U}\sqrt{x_k}\right)^2} \left(\frac{\sum_{k\in\bar{S}} x_k}{\sum_{k\in S} x_k} \sum_{k\in U} x_k\right).$$

Thus, in this case, the optimal purely model-based strategy consists of choosing the units with the n largest values of variable x (Royall (1970)), a very extreme sample. The strategy can be dangerous if the model is wrong. It is thus reasonable to opt for a strategy that guarantees correct estimation when the model is misspecified, and that leads to design-unbiased inference. For these reasons, Strategy 1 is rarely used (see Hansen, Madow, and Tepping (1983)).

4. Balanced Sample under a Linear Model

In this section we generalize the concept of balanced samples. Our results are thus more general than those in Nedyalkova and Tillé (2008). Consider the d-weighted estimator

$$\widehat{Y}_d = \sum_{k\in S} d_k y_k = \sum_{k\in S} d_k \mathbf{x}_k'\widehat{\boldsymbol{\beta}}_{\mathrm{BLU}} + \sum_{k\in S} d_k e_k, \tag{4.1}$$

where e_k is defined in (3.3). Under d-balanced sampling, $\widehat{Y}_d$ is model-unbiased and its error variance is

$$\mathrm{E}_{\mathrm{M}}(\widehat{Y}_d - Y)^2 = \sigma^2 \left[\sum_{k\in S}(d_k - 1)^2\nu_k^2 + \sum_{k\in\bar{S}}\nu_k^2\right]. \tag{4.2}$$

By comparing (4.1) with (3.2), we generalize Result 7 of Nedyalkova and Tillé (2008).

Result 1. *A sufficient condition that* $\widehat{Y}_{\mathrm{BLU}} = \widehat{Y}_d$ *is that*

- *the sampling design is d-balanced on* $\mathbf{x}_k$,
- $\sum_{k\in S} e_k(d_k - 1) = 0.$

A particular case of Result 1 is given below.

Corollary 1. *A sufficient condition that* $\widehat{Y}_{\mathrm{BLU}} = \widehat{Y}_d$ *is that*

- *the sampling design is d-balanced on* $\mathbf{x}_k$,
- *there exists a vector* $\boldsymbol{\lambda}$ *such that* $\boldsymbol{\lambda}'\mathbf{x}_k = \nu_k^2(d_k - 1)$, *for all* $k \in U$.

Proof. If

$$\frac{\boldsymbol{\lambda}'\mathbf{x}_k}{\nu_k^2(d_k - 1)} = 1,$$

then

$$\sum_{k\in S} e_k(d_k - 1) = \sum_{k\in S} \frac{\boldsymbol{\lambda}'\mathbf{x}_k}{\nu_k^2(d_k - 1)} e_k(d_k - 1) = \sum_{k\in S} \frac{\boldsymbol{\lambda}'\mathbf{x}_k}{\nu_k^2} e_k = 0.$$

Consider a strategy that meets the conditions of Result 1.

Strategy 2.

- *Use a d-balanced sampling design on* $\mathbf{x}_k$, *with* d_k *chosen so that* $d_k = (\nu_k^2 + \boldsymbol{\lambda}'\mathbf{x}_k)/\nu_k^2$, *for all* $k \in U$,
- *use the d-weighted estimator.*

A particular case of Strategy 2 was recommended by Scott, Brewer, and Ho (1978) in the case of a polynomial model. Since the conditions of Result 1 are met, we have that $\widehat{Y}_{\mathrm{BLU}} = \widehat{Y}_d$. The strategy judiciously chooses the d_k's to equalize the d-weighted estimator and the BLU estimator.

For a sample size n, this strategy can be implemented by using the inclusion probabilities

$$\pi_k = \frac{\nu_k^2}{\nu_k^2 + \boldsymbol{\lambda}'\mathbf{x}_k}.$$

The value of $\boldsymbol{\lambda}$ can be chosen freely subject to

$$\sum_{k\in U} \pi_k = \sum_{k\in U} \frac{\nu_k^2}{\nu_k^2 + \boldsymbol{\lambda}'\mathbf{x}_k} = n.$$

After some algebra, it is possible to show that

$$\mathrm{E_M}(\widehat{Y}_d - Y)^2 = \sigma^2 \sum_{k\in \bar{S}} d_k \nu_k^2 = \sigma^2 \left(\sum_{k\in \bar{S}} \nu_k^2 + \sum_{k\in \bar{S}} \boldsymbol{\lambda}'\mathbf{x}_k \right).$$

Moreover,

$$\mathrm{E}_p\mathrm{E}_\mathrm{M}(\widehat{Y}_d - Y)^2 = \sigma^2 \sum_{k\in U} \boldsymbol{\lambda}'\mathbf{x}_k.$$

Nevertheless, we see below that this is not necessarily the best strategy.

Eventually, consider a strategy proposed by Nedyalkova and Tillé (2008) that also meets the conditions of Result 1.

Strategy 3.

- *Use a d-balanced sampling design on* $\mathbf{x}_k$*, with* $d_k \propto \nu_k^{-1}$ *where* $\mathbf{x}_k$ *is complemented so that there exist two vectors* $\boldsymbol{\alpha}$ *and* $\boldsymbol{\gamma}$ *such that* $\boldsymbol{\alpha}'\mathbf{x}_k = \nu_k^2$ *and* $\boldsymbol{\gamma}'\mathbf{x}_k = \nu_k$*, for all* $k \in U$*, a 'fully explainable heteroscedasticity',*
- *use the d-weighted estimator.*

Strategy 3 was recommended by Nedyalkova and Tillé (2008) because it minimizes the anticipated variance among the strategies that are design-unbiased (see also Fuller (2009, p.187)). It is thus better than Strategy in the class of design-unbiased strategies.

With Strategy 3, we have $\widehat{Y}_\mathrm{BLU} = \widehat{Y}_d$. The condition of 'fully explainable heteroscedasticity' can be obtained by adding ν_k^2 and $d_k\nu_k^2$ to the list of balancing variables. Thus,

$$\sum_{k\in S} d_k\nu_k^2 = \sum_{k\in U} \nu_k^2, \quad \sum_{k\in S} d_k^2\nu_k^2 = \sum_{k\in U} d_k\nu_k^2,$$

and the error variance of the d-weighted estimator given in (4.2) simplifies to

$$\mathrm{E}_\mathrm{M}(\widehat{Y}_d - Y)^2 = \sigma^2 \sum_{k\in U} (d_k - 1)\nu_k^2.$$

For a sample size n, we must take $d_k = N/(\nu_k n)$, and $\pi_k = 1/d_k$. The error variance of the d-weighted estimator, given in (4.2), simplifies to

$$\mathrm{E}_\mathrm{M}(\widehat{Y}_d - Y)^2 = \sigma^2 \sum_{k\in U} \left(\frac{N\nu_k}{n} - \nu_k^2\right) = \sigma^2 \left(\frac{N^2}{n} - \sum_{k\in U} \nu_k^2\right).$$

5. Bias-robustness in Submodels

A large part of the model-based inference is dedicated to the bias-robustness of the BLU estimator in the case of misspecification of the model. In this section, we propose a formalization of the question of bias-robustness and an analysis of the consequences of model misspecification. We assume that a model M is used to conceive the strategy, but that the true underlying model is M^*.

Definition 7. A strategy is bias-robust for a model M^* if $\mathrm{E}_{\mathrm{M}^*}(\widehat{Y} - Y) = 0$.

Consider a model

$$M : y_k = \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_k$$

and an alternative model

$$M^* : y_k = \mathbf{z}_k'\boldsymbol{\gamma} + \eta_k,$$

where $\mathrm{E}_{\mathrm{M}^*}(\eta_k) = 0$. There is no assumption on the covariance matrix of the vector of η_k.

Definition 8. Model M^* is a submodel of M if there exists a matrix $\mathbf{A}$ such that $\mathbf{A}\mathbf{x}_k = \mathbf{z}_k$, not necessarily of full rank.

The basic fact needed to show that a strategy is robust is the following.

Result 2. *The strategy that consists of using a d-balanced sampling design on* $\mathbf{x}_k$ (*with any vector of* d_k) *and the d-weighted estimator is bias-robust for any submodel of* M.

Proof. Under model M^*, and using a d-balanced sample,

$$\widehat{Y}_d = \sum_{k\in S} d_k y_k = \sum_{k\in S} d_k(\mathbf{z}_k'\boldsymbol{\gamma} + \eta_k) = \sum_{k\in S} d_k(\mathbf{A}'\mathbf{x}_k'\boldsymbol{\gamma} + \eta_k) = \mathbf{A}'\sum_{k\in U}\mathbf{x}_k'\boldsymbol{\gamma} + \sum_{k\in S} d_k\eta_k.$$

Thus,

$$\mathrm{E}_{\mathrm{M}^*}(\widehat{Y}_d - Y) = \mathrm{E}_{M^*}\left(\sum_{k\in S} d_k\eta_k - \sum_{k\in U}\eta_k\right) = 0.$$

Result 2 encourages the statistician to overspecify the model, i.e., to introduce additional variables into model M in order to ensure that M^* is really a submodel of M. Indeed, if model M* is true, the variance under the model of the d-weighted estimator is the same irrespective of whether the sample is balanced on $\mathbf{x}_k$ or on $\mathbf{z}_k$. An over-specification of the model does not increase the variance under the model of the estimator because the independent variables of the model are only used for balancing the sample. Moreover, the d-weighted estimator does not depend on an estimated coefficient that relies on the number of auxiliary variables.

Suppose now that the model is misspecified, i.e., that model M^* is not a submodel of M. If the sampling design is d-balanced on the independent variables of model M, then

$$\mathrm{E}_{\mathrm{M}^*}(\widehat{Y}_d - Y) = \left(\sum_{k\in S} d_k\mathbf{z}_k' - \sum_{k\in U}\mathbf{z}_k'\right)\boldsymbol{\gamma}.$$

Let $f_k = \mathbf{z}_k'\boldsymbol{\gamma} - \mathbf{x}_k'\boldsymbol{\phi}$ be the residuals of a linear regression of $(\mathbf{z}_k'\boldsymbol{\gamma})$ on $\mathbf{x}_k$ and $\boldsymbol{\phi}$ the regression coefficient vector. Then, if the sampling design is d-balanced on $\mathbf{x}_k$,

$$\mathrm{E}_{\mathrm{M}^*}(\widehat{Y}_d - Y) = \left[\sum_{k\in S} d_k(\mathbf{x}_k'\boldsymbol{\phi} + f_k) - \sum_{k\in U}(\mathbf{x}_k'\boldsymbol{\phi} + f_k)\right] = \sum_{k\in S} d_k f_k - \sum_{k\in U} f_k, \quad (5.1)$$

for any value $\boldsymbol{\phi}$, in particular when

$$\boldsymbol{\phi} = \left(\sum_{k\in U} \frac{\mathbf{x}_k\mathbf{x}_k'}{\mathrm{var}(\eta_k)}\right)^{-1} \sum_{k\in U} \frac{\mathbf{x}_k\mathbf{z}_k'\boldsymbol{\gamma}}{\mathrm{var}(\eta_k)}.$$

The model variance thus only depends on the residuals of the regression, i.e., on the part of the model that remains misspecified.

If we do not possess information about the $\mathbf{z}_k$, we select a random sample with inclusion probabilities $\pi_k = 1/d_k$, because in selecting the sample randomly, the expected value under the sampling design of (5.1) is zero. Moreover, we can expect that, under reasonable asymptotic assumptions, the quantity

$$\frac{\mathrm{E}_{\mathrm{M}^*}(\widehat{Y}_d - Y)}{N} = \frac{1}{N}\left(\sum_{k\in S}\frac{f_k}{\pi_k} - \sum_{k\in U} f_k\right),$$

remains bounded in probability with respect to the sampling design when multiplied by $\sqrt{n}$. The random selection of a sample and the use of a design-unbiased estimator give an ultimate bias protection in the case where it is not possible to overspecify the model. This guarantees a negligible bias under M^* when n is large.

6. Application to the Polynomial Model

6.1. Presentation of the model

The polynomial model was studied, among others, by Royall and Herson (1973a), Scott, Brewer, and Ho (1978), and Valliant, Dorfman, and Royall (2000). The model is

$$y_k = \sum_{j=0}^{J} \delta_j\beta_j x_k^j + \varepsilon_k, \quad (6.1)$$

where x_k is the only independent variable, β_j is the jth regression coefficient, δ_j is equal to 1 or 0 as the term $\beta_j x_k^k$ appears or not in the regression, $\mathrm{E}_{\mathrm{M}}(\varepsilon_k) = 0$, $\mathrm{var}_{\mathrm{M}}(\varepsilon_k) = \sigma^2\nu_k^2$, and $\mathrm{cov}_{\mathrm{M}}(\varepsilon_k, \varepsilon_\ell) = 0$, when $k \neq \ell$. We assume $\sum_{k\in U}\nu_k = N$.

Now, from Result 2, for any set of vectors of d_k, the d-weighted estimator is bias-robust under any submodel of (6.1) provided

$$\sum_{k\in S} d_k x_k^j = \sum_{k\in U} x_k^j, \text{ for } j = 0, \ldots, J. \tag{6.2}$$

This implies several results on the polynomial model.

6.2. A first suboptimal strategy

Let $S^*(J)$ be a particular sample for which

$$\frac{\sum_{k\in\bar{S}} x_k^j}{\sum_{k\in\bar{S}} x_k} = \frac{\sum_{k\in S} x_k^{j+1}/\nu_k^2}{\sum_{k\in S} x_k^2/\nu_k^2}, \text{ for } j = 0, \ldots, J. \tag{6.3}$$

With a sample satisfying (6.3), Scott, Brewer, and Ho (1978) showed that the estimator

$$\widehat{Y}_0 = \sum_{k\in S} y_k + \sum_{k\in\bar{S}} x_k \frac{\sum_{k\in S} y_k x_k/\nu_k^2}{\sum_{k\in S} x_k^2/\nu_k^2},$$

is BLU for any polynomial Model (6.1), and any value of ν_k. This simple condition on the sample implies that the estimator is bias-robust for a large class of polynomial models.

It can easily be shown that a sufficient condition for a sample to satisfy (6.3) is

$$\sum_{k\in S} x_k^j \left(1 + \frac{\lambda x_k}{\nu_k^2}\right) = \sum_{k\in U} x_k^j, \text{ for } j = 0, \ldots, J, \tag{6.4}$$

where λ is a scalar that does not depend on j.

Equality (6.4) can be satisfied by using Strategy 2. In this particular case, we have $\pi_k = 1/d_k$ and $d_k = 1 + \lambda x_k/\nu_k^2$.

Strategy 2. (for polynomial model) *Select a balanced sample such that*

$$\sum_{k\in S} \frac{x_k^j}{\pi_k} = \sum_{k\in U} x_k^j, j = 0, \ldots, J,$$

with the unequal inclusion probabilities

$$\pi_k = \frac{1}{1 + \lambda x_k/\nu_k^2}, k \in U.$$

The constant λ is chosen as a function of the desired sample size, where

$$\sum_{k\in U}\pi_k=\sum_{k\in U}\frac{1}{1+\lambda x_k/\nu_k^2}=n. \tag{6.5}$$

The solution of (6.5) in λ is unique. Strategy 2 is recommended by Scott, Brewer, and Ho (1978). The inclusion probabilities are, however, chosen so that the BLU estimator is equal to the Horvitz-Thompson estimator and is thus not optimized.

Two particular cases of (6.3) are as follows.

(a) When $\nu_k^2\propto x_k$, under the condition $\sum_{k\in U}\nu_k=N$,

$$\nu_k^2=\frac{N^2x_k}{\left(\sum_{k\in U}\sqrt{x_k}\right)^2}.$$

In this case, (6.3) reduces to

$$\frac{\sum_{k\in\bar{S}}x_k^j}{\sum_{k\in\bar{S}}x_k}=\frac{\sum_{k\in S}x_k^j}{\sum_{k\in S}x_k},\ \text{for } j=0,\ldots,J.$$

Thus, the sample should satisfy the condition

$$\frac{1}{n}\sum_{k\in S}x_k^j=\frac{1}{N-n}\sum_{k\in\bar{S}}x_k^j=\frac{1}{N}\sum_{k\in U}x_k^j,\ \text{for } j=0,\ldots,J.$$

Royall and Herson (1973a) call such samples *balanced*. In this case, $\widehat{Y}_0$ reduces to the ordinary ratio estimator

$$\widehat{Y}_{\mathrm{R}}=\sum_{k\in U}x_k\frac{\sum_{k\in S}y_k}{\sum_{k\in S}x_k}.$$

(b) When $\nu_k^2\propto x_k^2$, it is easily shown that $\nu_k^2=N^2x_k^2/\left(\sum_{k\in U}x_k\right)^2$. In this case, (6.3) reduces to

$$\sum_{k\in S}\frac{x_k^{j-1}}{n}=\frac{\sum_{k\in\bar{S}}x_k^j}{\sum_{k\in\bar{S}}x_k},\quad\text{for } j=0,\ldots,J.$$

The sample $S^*(J)$ is called *overbalanced* (Scott, Brewer, and Ho (1978)), and $\widehat{Y}_0$ reduces to

$$\widehat{Y}_{\mathrm{OB}}=\sum_{k\in S}y_k+\left(\frac{1}{n}\sum_{k\in S}\frac{y_k}{x_k}\right)\sum_{k\in\bar{S}}x_k.$$

6.3. An alternative strategy for the polynomial model

Strategy 3. (for polynomial model)

- *Use inclusion probabilities that are proportional to ν_k, subject to*

$$\sum_{k\in U} \pi_k = n,\ 0 \le \pi_k \le 1.$$

- *Select a balanced sample according to the balancing equations*

$$\sum_{k\in S} \frac{x_k^j}{\pi_k} = \sum_{k\in U} x_k^j, j = 0, \ldots, J, \tag{6.6}$$

$$\sum_{k\in S} \frac{\nu_k^2}{\pi_k} = \sum_{k\in U} \nu_k^2,$$

$$\sum_{k\in S} \frac{\nu_k}{\pi_k} = \sum_{k\in U} \nu_k. \tag{6.7}$$

- *Use the Horvitz-Thompson estimator.*

Note that, since $\pi_k \propto n\nu_k/N$, (6.7) becomes $\frac{N}{n}\sum_{k\in S} 1 = N$ and only means that the sample size must be fixed. Nedyalkova and Tillé (2008) proved that this strategy minimizes the anticipated variances in the class of design-unbiased strategies. Strategy 3 is thus better than Strategy 2 in the sense that its anticipated variance is always smaller. With Strategy 3, the inclusion probabilities are chosen to minimize the variance, while with Strategy 2 the inclusion probabilities are chosen to meet the technical condition given in Result 1.

Two particular cases are as follows.

(a) When $\nu_k^2 \propto x_k$ with $\sum_{k\in U} \nu_k = N$, then

$$\nu_k^2 = \frac{N^2 x_k}{\left(\sum_{k\in U} \sqrt{x_k}\right)^2}.$$

As $\pi_k \propto \nu_k$ with $\sum_{k\in U} \pi_k = n$, it follows that $\pi_k = (n/N)\nu_k$. In this case, if the sample has a fixed sample size and is balanced on x_k, it is automatically balanced on ν_k and ν_k^2.

(b) When $\nu_k^2 \propto x_k^2$, with $\sum_{k\in U} \nu_k = N$, then

$$\nu_k^2 = \frac{N^2 x_k^2}{\left(\sum_{k\in U} x_k\right)^2}.$$

Here too, we have $\pi_k = (n/N)\nu_k$. In this case, if the sample has a fixed sample size and is balanced on x_k^2, it is automatically balanced on ν_k and ν_k^2.

6.4. A particular case: the ratio model

Consider again the model without intercept and only one regressor with $\nu_k^2 \propto x_k$, as at (3.4), a particular case of the polynomial model. We have seen that the BLU estimator under this model is the ordinary ratio estimator

$$\widehat{Y}_{\mathrm{R}} = \sum_{k\in U} x_k \frac{\sum_{k\in S} y_k}{\sum_{k\in S} x_k},$$

with error variance

$$\mathrm{E_M}(\widehat{Y}_{\mathrm{R}} - Y)^2 = \sigma^2 \frac{N^2}{\left(\sum_{k\in U} \sqrt{x_k}\right)^2} \left(\frac{\sum_{k\in \bar{S}} x_k}{\sum_{k\in S} x_k} \sum_{k\in U} x_k \right).$$

Here Strategy 2 is endorsed by Royall and Herson (1973a).

Strategy 2. (for the ratio model) *Select a mean-balanced sample of size n and use the ratio estimator.*

With a mean-balanced sample on x_k, satisfying the condition

$$\frac{1}{n}\sum_{k\in S} x_k = \frac{1}{N-n}\sum_{k\in \bar{S}} x_k,$$

the ratio estimator reduces to the sample mean

$$\widehat{Y}_{\mathrm{R}} = \sum_{k\in U} x_k \frac{\sum_{k\in S} y_k}{\sum_{k\in S} x_k} = \frac{1}{n}\sum_{k\in S} y_k.$$

Moreover,

$$\mathrm{E_M}(\widehat{Y}_{\mathrm{R}} - Y)^2 = \sigma^2 \frac{N^2(N-n)}{n} \frac{\sum_{k\in U} x_k}{\left(\sum_{k\in U} \sqrt{x_k}\right)^2} = \mathrm{E}_p \mathrm{E_M}(\widehat{Y}_{\mathrm{R}} - Y)^2.$$

Strategy 3. (for the ratio model) *Select a balanced sample on x_k with inclusion probabilities $\pi_k \propto \nu_k \propto \sqrt{x_k}$ and use the Horvitz-Thompson estimator.*

In order to show that Strategy 3 is better than Strategy 2, we compare the anticipated and model-variances, $\mathrm{E}_p\mathrm{E_M}(\widehat{Y} - Y)^2$ and $\mathrm{E_M}(\widehat{Y} - Y)^2$, of the ratio and Horvitz-Thompson estimators under (3.4).

Nedyalkova and Tillé (2008) have shown that, under Strategy 3,

$$\mathrm{E_M}(\widehat{Y}_\pi - Y)^2 = \mathrm{E}_p \mathrm{E_M}(\widehat{Y}_\pi - Y)^2 = \sigma^2 \left(\frac{N^2}{n} - \sum_{k\in U} \nu_k^2 \right).$$

After replacing ν_k with $N\sqrt{x_k}/\sum_{k\in U}\sqrt{x_k}$, we obtain

$$\mathrm{E}_p\mathrm{E}_\mathrm{M}(\widehat{Y}_\pi - Y)^2 = \sigma^2\left[\frac{N^2}{n} - \frac{N^2\sum_{k\in U}x_k}{\left(\sum_{k\in U}\sqrt{x_k}\right)^2}\right].$$

With

$$D = \mathrm{E}_p\mathrm{E}_\mathrm{M}(\widehat{Y}_\mathrm{R} - Y)^2 - \mathrm{E}_p\mathrm{E}_\mathrm{M}(\widehat{Y}_\pi - Y)^2 = \mathrm{E}_\mathrm{M}(\widehat{Y}_\mathrm{R} - Y)^2 - \mathrm{E}_\mathrm{M}(\widehat{Y}_\pi - Y)^2,$$

simplification yields

$$D = \frac{N^2}{n}\left[\frac{N\sum_{k\in U}x_k}{\left(\sum_{k\in U}\sqrt{x_k}\right)^2} - 1\right] \geq 0.$$

Thus, Strategy 3 is better than Strategy 2 under (3.4).

7. Discussion

Under a linear model, the use of a purely model-based strategy (Strategy 1) can be dangerous. Balanced samples offer good protection against model misspecification. A d-balanced sampling design with the d-weighted estimator is a bias-robust strategy that assures protection against misspecification of the model. There however exist several ways of selecting balanced samples. The d-weighted estimator can be equivalent to the BLU estimator if some technical conditions are met. These conditions can be met by either choosing the *ad hoc* inclusion probabilities (Strategy 2) or by adding ν_k and ν_k^2 to the list of balancing variables and choosing the optimal inclusion probability (Strategy 3).

For the polynomial model, Royall and Herson (1973a) and Scott, Brewer, and Ho (1978) used ad hoc inclusion probabilities. They showed that, in this case, an unweighted ratio estimator is BLU for a large class of polynomial models. Nevertheless, this strategy consists of choosing the inclusion probabilities in such a way that a technical property is satisfied. Strategy 3 is more appropriate, because the technical condition is met by adding two balancing variables and the inclusion probabilities can be chosen to minimize the anticipated variance.

Strategy 2 is thus not admissible in the sense that it is always possible to obtain a smaller anticipated variance by selecting the units with inclusion probabilities proportional to the standard deviations of the errors of the model and using the Horvitz-Thompson estimator. Strategy 3 has the advantage of providing a bias-robust estimator, even if the model is misspecified. Indeed, in both cases, the estimator does not depend on the auxiliary variables and is always design-unbiased. In case of misspecification of the model or of measurement errors the estimators of totals thus remain unbiased but a part of the efficiency can

be lost. The importance of this loss depends on the way in which the correlation between the independent variables of the model and the variables of interest is hindered by the measurement errors.

Eventually, the best protection against a misspecification of the model consists of extending the list of balancing variables. Indeed, the addition of balancing variables that could be correlated to the variables of interest does not increase the model variance, but protects against bias under the model. If the model cannot really be specified, the ultimate protection against misspecification is always the random selection of the sample.

We hope that these results give a general guideline on the way of planning a sample survey under a realistic linear model. In practice, designing a survey is often a more complex exercise for the following reasons:

- It is difficult to specify a model without knowing the variable(s) of interest. The sampling design is necessarily conceived before knowing the dependent variable. Thus the validation of the model is difficult.
- A survey is generally done for multiple purposes. Arbitration must thus be done between the variables of interest, the areas of interest, and the parameters to estimate, and thus between the models.
- Nonrepsonse implies that the set of respondents is never exactly balanced.

Our results however give a general and ideal framework on how to plan and estimate a total under a linear model.

The result of Neyman (1934) concerning optimal stratification suffers the same problems. Optimality depends on the variable of interest, on the parameters to estimate, on the areas of interest. The variances within the strata are never really known, thus they must be estimated using another survey, or must be derived from a hypothesis of the existence of a size effect. Indeed, optimal stratification is a particular case of our result. The generalization of our results to cluster sampling, two-stage, and two-phase sampling remains a challenging topic of research.

Acknowledgement

The authors thank two anonymous reviewers and the Editor for their constructive comments that helped us improve the quality of this paper. This research is supported by the Swiss National Science Foundation (grant no. FN 205121-105187/1).

References

Brewer, K. and Särndal, C.-E. (1983). Six approaches to enumerative survey sampling. In *Incomplete Data in Sample Surveys* (Edited by W. G. Madow, I. Olkin, and D. B. Rubin), 363-405. Academic Press, N.Y.

Chambers, R. L. (1996). Robust case-weighting for multipurpose establishment surveys. *J. Off. Statist.* **12**, 3-32.

Cumberland, W. and Royall, R. (1981). Prediction models in unequal probability sampling. *J. Roy. Statist. Soc. Ser. B* **43**, 353-367.

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika* **91**, 893-912.

Fuller, W. A. (2009). *Sampling Statistics*. John Wiley, New Jersey.

Hájek, J. (1959). Optimum strategy and other problems in probability sampling. *Cosopis. Pest. Mat.* **84**, 387-423.

Hájek, J. (1981). *Sampling from a Finite Population*. Marcel Dekker, New York.

Hansen, M. H., Madow, W. G. and Tepping, B. J. (1983). An Evaluation of model-dependent and probability-sampling inferences in sample surveys. *J. Amer. Statist. Assoc.* **78**, 776-793.

Isaki, C. and Fuller, W. A. (1982). Survey design under a regression population model. *J. Amer. Statist. Assoc.* **77**, 89-96.

Joshi, V. M. (1979). The best strategy for estimating the mean of a finite population. *Ann. Statist.* **7**, 531-536.

Kott, P. S. (1986). When a mean-of-ratios is the best linear unbiased estimator under a model. *Amer. Statist.* **40**, 202-204.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection, *J. Roy. Statist. Soc.* **97**, 558-606.

Nedyalkova, D. and Tillé, Y. (2008). Optimal sampling and estimation strategies under the linear model. *Biometrika* **95**, 521-537.

Ramakrishnan, M. (1975). Choice of an optimum sampling strategy. *Ann. Statist.* **3**, 669-679.

Royall, R. (1970). On finite population sampling theory under certain linear regression models. *Biometrika* **57**, 377-387.

Royall, R. (1976). The linear least squares prediction approach to two-stage sampling. *J. Amer. Statist. Assoc.* **71**, 657-664.

Royall, R. (1992). Robustness and optimal design under prediction models for finite populations. *Surv. Methodol.* **18**, 179-185.

Royall, R. and Cumberland, W. (1981). The finite population linear regression estimator and estimators of its variance. An empirical study. *J. Amer. Stat. Assoc.* **76**, 924-930.

Royall, R. and Herson, J. (1973a). Robust estimation in finite populations I. *J. Amer. Statist. Assoc.* **68**, 880-889.

Royall, R. and Herson, J. (1973b). Robust estimation in finite populations II: Stratification on a size variable. *J. Amer. Statist. Assoc.* **68**, 891-893.

Särndal, C.-E. (1978). Design-based and model-based inference in survey sampling. *Scand. J. Statist.* **5**, 27-52.

Scott, A., Brewer, K. and Ho, E. (1978). Finite population sampling and robust estimation. *J. Amer. Statist. Assoc.* **73**, 359-361.

Tillé, Y. (2006). *Sampling Algorithms*. Springer-Verlag, New York.

Valliant, R., Dorfman, A. and Royall, R. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley, New York.

Institute of statistics, University of Neuchâtel, Pierre à Mazel 7, 2000 Neuchâtel, Switzerland.
E-mail: desislava.nedyalkova@unine.ch

Institute of statistics, University of Neuchâtel, Pierre à Mazel 7, 2000 Neuchâtel, Switzerland.
E-mail: yves.tille@unine.ch

(Received October 2010; accepted may 2011)