Survey Methodology, December 1992
Vol. 18, No. 2, pp. 179–185
Statistics Canada

Robustness and Optimal Design Under Prediction Models for Finite Populations

RICHARD M. ROYALL[1]

ABSTRACT

In many finite population sampling problems the design that is optimal in the sense of minimizing the variance of the best linear unbiased estimator under a particular working model is bad in the sense of robustness - it leaves the estimator extremely vulnerable to bias if the working model is incorrect. However there are some important models under which one design provides both efficiency and robustness. We present a theorem that identifies such models and their optimal designs.

KEY WORDS: Balanced sample; Bias protection; Model failure; Working model.

1. INTRODUCTION

The "ratio estimator" of a finite population total $T = y_i + \ldots + y_N$ is $\hat{T} = N\bar{x}\bar{y}_s/\bar{x}_s$, where $\bar{x} = (x_1 + \ldots + x_N)/N$ is the known population mean of an auxiliary variable and $\bar{x}_s$ and $\bar{y}_s$ are sample means. This is the best linear unbiased (BLU) estimator of T under the model M:

$$E(Y_i) = \beta x_i,$$

$$\operatorname{cov}(Y_i, Y_j) = \begin{cases} \sigma^2 x_i & i = j \\ 0 & \text{else.} \end{cases}$$

This estimator is biased under alternative models having different regression functions, in general, but protection against bias under specific alternatives can be assured by careful choice of the sample, as will be described below.

Throughout this paper we will be concerned with populations for which a particular model, such as M, is believed to apply, at least to a satisfactory degree of approximation. Our inferences will be made with reference to this model. For example, we will call an estimator $\hat{T}$ unbiased only if $E_M(\hat{T} - T) = 0$. On the other hand, we recognize that the model is an approximation and that it might be seriously wrong. Thus we describe it as a **working model**, and seek sampling and estimation procedures that are robust in the sense of performing well, not only under that working model, but also under alternative models that might better describe the relationships between variables in our population.

We denote by $M(\delta_0,\delta_1, \ldots \delta_J : v)$ the general polynomial regression model:

$$E(Y_i) = \sum_{j=0}^{J} \delta_j \beta_j x_i^j$$

[1] Richard M. Royall, Johns Hopkins University, Baltimore, MD 21205 U.S.A.

$$\operatorname{cov}(Y_i, Y_j) = \begin{cases} v_i \sigma^2 & i = j, \\ 0 & \text{else} \end{cases}$$

where δ_j is a zero-one indicator of whether the regressor x^j is included in the model. The best linear unbiased estimator under this model is denoted by $\hat{T}(\delta_0, \ldots, \delta_J : v)$. Thus our first model was $M(0, 1 : x)$, and $\hat{T}(0, 1 : x)$ is the ratio estimator.

Royall and Herson (1973) showed that $\hat{T}(0, 1 : x)$ remains unbiased under $M(\delta_0, \ldots, \delta_J : v)$ for any vector $(\delta_0, \ldots, \delta_J)$ of zeroes and ones, and any $v_1, \ldots, v_N$, if the sample is **balanced** on $x, x^2, \ldots, x^J$:

$$\sum_s x_i^j/n = \sum_1^N x_i^j/n \quad j = 1, 2, \ldots, J.$$

This means that in a balanced sample $\hat{T}(0, 1 : x)$ is robust in the sense that it remains unbiased under regression models that are much more general than the working model $M(0, 1 : x)$. Royall and Herson (1973, sec. 4.5) also detailed how approximate balance ensures the approximate unbiasedness of $\hat{T}(0, 1 : x)$. Furthermore they showed that in a balanced sample this estimator retains not only its unbiasedness but also its **optimality** under a wide variety of polynomial regression models, including $M(1 : 1)$, $M(1, 1 : x)$, and $M(0, 1, 1 : x^2)$. Specifically, the estimator is optimal under any polynomial regression model of degree J or less, provided only that the model's variance function is expressible as a linear combination of the regressors.

The robustness of the ratio estimator in balanced samples is achieved at a high cost in efficiency under the original working model $M(0, 1 : x)$. Under this model the sample that minimizes the variance consists of the n units whose x-values are largest, and the efficiency of a balanced sample is only $\bar{x}/\max_s(\bar{x}_s)$. (Royall and Herson 1973).

For the linear regression estimator, theoretical results have been established that are quite analogous to those sketched above for the ratio estimator, but with one important difference. The estimator is $\hat{T}(1, 1 : 1) = N[\bar{y}_s + b(\bar{x} - \bar{x}_s)]$, where $b = \sum_s(x_i - \bar{x}_s)y_i/\sum_s(x_i - \bar{x}_s)^2$. It is the optimal (BLU) estimator under the constant variance linear regression model, $M(1,1 : 1)$. When the sample is balanced, this estimator is robust, remaining unbiased (and optimal) under the same broad class of polynomial regression models as the ratio estimator. But unlike the ratio estimator, the regression estimator achieves robustness in balanced samples at **no cost in efficiency** - the variance under the working model $M(1, 1 : 1)$ is minimized in balanced samples, where $\bar{x}_s = \bar{x}$. This phenomenon occurs because the error variance $E(\hat{T} - T)^2$ is the sum of a constant and a term proportional to $(\bar{x} - \bar{x}_s)^2 \operatorname{var}(b)$. Minimizing $\operatorname{var}(b)$ requires maximizing $\sum_s(x_i - \bar{x}_s)^2$, but this term is eliminated altogether in samples with $\bar{x}_s = \bar{x}$.

Are there other models under which the same sample that minimizes the variance of the BLU estimator can also protect against bias under a wide range of alternative models? In particular, are there such models for problems requiring non-constant variance functions? We show that the answer is positive, giving a theorem that characterizes a family of models with the desired property and identifies the corresponding optimal samples. The results in this paper integrate and generalize those of Kott (1984) and Tallis (1986). They are also closely related to the work of Pereira and Rodrigues (1983) and Tam (1986), as well as that of Isaki and Fuller (1982).

2. BASIC RESULTS

It is convenient to shift to vector and matrix notation, in which Y is the population vector $(Y_1, Y_2, \ldots, Y_N)'$ and the model $M(X : V)$ specifies that $E(Y) = X\beta$ and $\mathrm{var}(Y) = V\sigma^2$, where X is an $N \times p$ matrix of regressors, V is diagonal, and the vector β and the scalar σ^2 are unknown. For a given sample s of n units we list the sample units first, so that

$$Y = \begin{pmatrix} Y_s \\ Y_r \end{pmatrix}, \quad X = \begin{pmatrix} X_s \\ X_r \end{pmatrix}, \quad V = \begin{pmatrix} V_s & 0 \\ 0 & V_r \end{pmatrix},$$

where Y_r is the $(N - n)$-vector corresponding to the non-sample units, *etc.* We let 1_s and 1_r denote vectors $(1, \ldots 1)'$ of lengths n and $(N - n)$.

The population total is $T = 1_s'Y_s + 1_r'Y_r$. After the sample s is observed, the first component, $1_s'Y_s$, is known. The BLU estimator of T is obtained by adding to this known quantity the BLU predictor of $1_r'Y_r$:

$$\hat{T}(X : V) = 1_s'Y_s + 1_r'X_r\hat{\beta}(X : V),$$

where $\hat{\beta}(X : V) = (X_s'V_s^{-1}X_s)^{-1}X_s'V_s^{-1}Y_s$. The error variance is

$$\mathrm{var}(\hat{T}(X : V) - T) = 1_r'(X_r'A_s^{-1}X_r + V_r)1_r\sigma^2,$$

where $A_s = X_s'V_s^{-1}X_s$. These formulas simplify when the vector $V1$ is in the linear manifold generated by the columns of X, which we denote by $\mathfrak{M}(X)$.

Lemma 1. If $V1 \in \mathfrak{M}(X)$ then

$$\hat{T}(X : V) = 1'X\hat{\beta}(X : V)$$

and under $M(X : V)$

$$\mathrm{var}(\hat{T}(X : V) - T) = (1'XA_s^{-1}X'1 - 1'V1)\sigma^2.$$

Proof: The estimator simplifies because $V1 \in \mathfrak{M}(X)$ means that $V1 = Xc$ for some vector c, so that $X_s'1_s = X_s'V_s^{-1}X_sc$, from which we have $1_s'X_s\hat{\beta} = c'X_s'V_s^{-1}Y_s = 1_s'Y_s$. The variance formula follows from $\mathrm{cov}(\hat{T}, T) = \mathrm{cov}(1'X\hat{\beta}, 1_s'Y_s) = 1'XA_s^{-1}X_s'1_s = 1'Xc = 1'V1$.

Lemma 1 shows that for models with $V1 \in \mathfrak{M}(X)$, the sample affects the variance only through A_s^{-1}. This simplifies both the study of how the variance depends on the sample and the search for efficient samples.

The collection of samples that satisfy

$$1_s'W_s^{-1/2}X_s/n = 1'X/1'W^{1/2}1,$$

where W is an $N \times N$ matrix, will be denoted by $B(X : W)$. When W is the identity matrix, I, $B(X : I)$ is the collection of samples that are balanced on the columns of X. Royall and Herson (1973) proved that BLU estimators under a wide family of polynomial regression models are greatly simplified in balanced samples:

Theorem 1. Under $M(X : V)$ with $V1 \in \mathrm{M}(X)$, if $s \in B(X : I)$ then

$$\hat{T}(X : V) = (N/n) 1_s' Y_s$$
$$\operatorname{var}(\hat{T}(X : V)) = [(N/n) - 1] 1' V 1 \sigma^2. \tag{1}$$

The next theorem shows that if $V = I$ then the variance in (1) is the minimum possible, *i.e.* balanced samples $B(X : I)$, are optimal if $I1 \in \mathcal{M}(X)$; it also identifies optimal samples for a class of models with more general variance structure.

Theorem 2. Under $M(X : V)$ if both $V1$ and $V^{1/2}1 \in \mathcal{M}(X)$, then

$$\operatorname{var}(\hat{T}(X : V) - T) \geq \left[(1' V^{1/2} 1)^2 / n - 1' V 1\right] \sigma^2;$$

the bound is achieved if and only if $s \in B(X : V)$, in which case

$$\hat{T}(X : V) = (1' V^{1/2} 1)(1_s' V_s^{-1/2} Y_s)/n.$$

Proof: Since $V1 \in \mathcal{M}(X)$, the quantity to be minimized is $a' A_s^{-1} a$, where $a = X'1$ (Lemma 1). Now $V^{1/2}1 \in \mathcal{M}(X)$ implies that there is a p-vector c_1 for which $V^{1/2}1 = Xc_1$ and, since V is diagonal, this ensures that $V_s^{1/2} 1_s = X_s c_1$ for every sample s. From this it follows that $c_1' A_s c_1 = n$, and the desired inequality then follows from Schwarz's:

$$(a' A_s^{-1} a)(c_1' A_s \, c_1) = (a' A_s^{-1} a) \cdot n \geq (a' c_1)^2.$$

The necessary and sufficient condition for equality is $a' = k c_1' A_s$, where $k = 1' V^{1/2} 1/n$. This is equivalent to $s \in B(X : V)$ because $c_1' A_s = 1_s' V_s^{-1/2} X_s$. The simple forms for the estimator $\hat{T}(X : V)$ and its variance are then easily obtained algebraically.

The formulas in Theorem 2 are familiar in conventional (randomization-based) sampling theory. The BLU estimator $\hat{T}(X : V)$ takes the simple form of the Horvitz-Thompson estimator $\hat{T}_{HT} = \sum_s y_i / \pi_i$, when π_i, the inclusion probability for unit i, is proportional to $v_i^{1/2}$. And the variance bound is the one established by Godambe and Joshi (1965, Theorem 6.1) for the model-based expectation of the random sampling variance.

Suppose that we have, for a working model $M(X : V)$ that satisfies the conditions of Theorem 2, an optimal sample s and BLU estimator $\hat{T}$. If we now consider a more general model $M(X, Z : V)$ with additional regressor(s) Z, the results of Theorem 2 continue to apply so long as the sample belongs to $B(Z : V)$ as well as to $B(X : V)$. Our sample and estimator remain optimal under the more general model, and the variance is unchanged. That is, we can maintain optimality under our working model (minimum variance sample and BLU estimator) and also protect against bias caused by the additional regressor(s) Z by imposing the additional constraint $B(Z : V)$ on the sample. This procedure not only protects our estimator from bias under $M(X, Z : V)$, it ensures that our sample and estimator both remain **optimal** under the more general model. Of course unbiasedness is ensured under the even more general model $M(X, Z : W)$, where W is any covariance matrix.

3. EXAMPLES

Four models have been particularly prominant in finite population sampling theory. In the polynomial regression model notation of section 1 these are $M(1 : 1)$, $M(1{,}1 : 1)$, $M(0{,}1 : x)$, and $M(0{,}1 : x^2)$. Optimal estimators under the first three models are the expansion, regression and ratio estimators, respectively. The optimal estimator under the fourth model,

$\hat{T}(0,1 : x^2) = \sum_s y_i + (N - n)\bar{x}_r \sum_s (y_i/nx_i)$, is approximated by the mean-of-ratios estimator $\hat{T}_{HT} = N\bar{x}\sum_s (y_i/nx_i)$ when the sampling fraction n/N is small.

One approach to finding a practical sampling and estimation strategy under one of these four working models is to use the best linear unbiased estimator under the model, while ensuring robustness by choosing a sample in which the estimator remains unbiased under more general polynomial regression models. For the first two models, $M(1 : 1)$ and $M(1, 1 : 1)$, we have seen that this strategy produces bias-robustness for free, at no cost in efficiency under the working model. Under both of these models bias protection requires simple (unweighted) balance; but the models satisfy the conditions of Theorem 2 with $V = I$, which implies that simple balance is optimal.

For the other two models, however, there is tension between robustness and efficiency. In section 1 we noted that under $M(0, 1 : x)$ the ratio estimator is optimal, and while the optimal sample consists of the n units maximizing $\bar{x}_s$, protection from bias under $M(1, 1 : x)$ requires a sample where $\bar{x}_s$ is not maximized but set equal to the population mean, $\bar{x}$. The situation under $M(0, 1 : x^2)$ is similar: the optimal sample is again the one where the sample mean $\bar{x}_s$ is maximized, but protection of the optimal estimator against bias under polynomial regression models requires an "overbalanced" sample, in which the sample mean equals $\sum_r x_i^2 / \sum_r x_i$ (Scott, Brewer and Ho 1978).

Under both of these models, $M(0, 1 : x)$ and $M(0, 1 : x^2)$, robustness can be achieved at a smaller cost in efficiency by starting with a more general working model. Theorem 2 shows the way. Consider first the model $M(0, 1 : x^2)$. If we use $\hat{T}(0, 1 : x^2)$ in an over-balanced sample, the error variance is $\{(N\bar{x})^2/n - \sum x_i^2 + \sum_s (x_i - \bar{x}_s)^2\}\sigma^2$. But if we use the more general working model $M(0, 1, 1 : x^2)$ and estimator $\hat{T}(0, 1, 1 : x^2)$, the theorem shows that any sample in which $\bar{x}_s = \sum x_i^2 / \sum x_i$ is optimal, yielding the minimum variance $\{(N\bar{x})^2/n - \sum x_1^2\}\sigma^2$. Now bias protection against even more general polynomial regression models can be obtained at no cost in efficiency by imposing the additional constraints of Condition $B(X : V)$ *i.e.* $\sum_s x_i^{j-1}/n = \sum_1^N x_i^j / \sum_1^N x_i \; j = 0, 3, \ldots, J$. Under these constraints on the sample, collectively called π-balance, $T(0, 1, 1 : x^2)$ is the mean-of-ratios estimator (Kott 1984). This sample and estimator remain optimal under all models of the form $M(\delta_0, 1, 1, \delta_3, \ldots, \delta_J : x^2)$.

Balanced samples $B(X : V)$ do not always exist. The above example illustrates this; when n becomes so large that $n/N > N(\bar{x}^2)/\sum x_i^2$ there can be no π-balanced sample, because otherwise the variance formula would become negative. Note that the condition $n/N > N(\bar{x}^2)/\sum x_i^2$ implies that $\max(x_i) > N\bar{x}/n$, so that in such populations there is no probability sampling plan with inclusion probability proportional to x.

To generalize the other model, $M(0.1 : x)$, so that the theorem will apply we can add a regressor, $x^{1/2}$:

$$E(Y_i) = \beta_{1/2} x_i^{1/2} + \beta_1 x_i$$

$$\operatorname{var}(Y_i) = \sigma^2 x_i.$$

According to Theorem 2 any sample satisfying

$$\sum_s x_i^{1/2} \Big/ n = \sum_1^N x_i \Big/ \sum_1^N x_i^{1/2} \tag{2}$$

is optimal under this model, yielding the best linear unbiased estimator $\sum x_i^{1/2} \sum_s x_i^{-1/2} y_i / n$ and the minimum variance, $\{(\sum x_i^{1/2})^2/n - N\bar{x}\}\sigma^2$. This variance compares favorably with that of the ratio estimator in a balanced sample, $N\bar{x}(N/n - 1)\sigma^2$. Now optimality of the sample and the estimator if in fact $E(Y_i) = \beta_0 + \beta_{1/2}x_i^{1/2} + \beta_1 x_i + \beta_2 x_i^2$ can be maintained (with no increase in variance) by imposing the additional conditions on the sample:

$$\sum_s x_i^{-1/2} \Big/ n = N \Big/ \sum_1^N x_i^{1/2}$$
$$\sum_s x_i^{3/2} \Big/ n = \sum_1^N x_i^2 \Big/ \sum_1^N x_i^{1/2}. \tag{3}$$

These conditions, (2) and (3), give the BLU estimator the simple form:

$$\sum_1^N x_i^{1/2} \sum_s (y_i/x_i^{1/2})/n,$$

which is of course the Horvitz-Thompson estimator for a probability-proportional-to-$x^{1/2}$ sampling plan.

4. PROBABILITY SAMPLING

The results in Section 2 are important in relation to an unobserved regressor Z. If Z were, like X, known for all population units, then we could use $M(X, Z : V)$ as the working model and $\hat{T}(X, Z : V)$ as the estimator in the first place. But suppose that we are unaware of the importance of Z and are using the working model $M(X : V)$ and the estimator $\hat{T}(X : V)$ when in fact $M(X, Z : V)$ applies. In this context we will refer to a sample from $B(X : V)$ as "balanced on X." Although we can choose a sample that is balanced on X, we cannot ensure that it will be balanced on Z, and if it is not, then our estimator is biased:

$$E(\hat{T}(X : V) - T) = \left[(1/n)(1'V^{1/2}1)(1_s'V_s^{-1/2}Z_s) - 1'Z\right]\gamma.$$

where γ is the Z-coefficient: $EY = X\beta + Z\gamma$.

Random sampling can help to provide protection against biases like this. If we use a probability sampling plan with inclusion probabilities, $\pi_i = nv_i^{1/2}/1'V^{1/2}1$, $i = 1, 2, \ldots, N$, then we will have balance on Z in expectation:

$$E_\pi 1_s'V_s^{-1/2}Z_s/n = 1'Z/1'V^{1/2}1,$$

the subscript π indicating that the expectation is with respect to the random sampling plan, not a prediction model. Furthermore, if our sampling plan is one under which $\text{var}_\pi(1'V_s^{-1/2}Z_s/n)$ approaches zero as n grows, then the probability that we will draw a sample that is badly unbalanced, say one in which $|1_s'V_s^{-1/2}Z_s/n - 1'Z/1'V^{1/2}1| > \delta$, can be made small by taking a large enough sample, n. That is, probability sampling can provide balance on Z "in probability."

The strength of this result is in its scope–it applies for any matrix Z of regressors whatsoever. In particular it applies for the matrix X of regressors in our working model, as well as for overlooked regressors. The weakness of course is that it applies to the sample selection process, not to a result of that process. The sample actually drawn will, with predictable frequency, be badly unbalanced on the known regressors X. If balance on X is important in a particular study, it should not be left to chance (This was documented empirically by Royall and Cumberland 1981). Restricted random sampling plans which guarantee that the selected sample will be balanced on X, such as Wallenius's "basket method" (1980), might represent a reasonable compromise strategy.

It sometimes happens that a regressor Z that is ignored when the sample is selected becomes available afterwards, as in the case of post-stratification for example. If it is determined that the selected sample is badly balanced on Z, then probability sampling has failed to provide the expected protection against bias under $M(X, Z : V)$; if it is too late to draw another sample, then to protect against the bias we must use an estimator that is unbiased under this model. That is, probability sampling does not guarantee approximate balance on Z; it only ensures that we have a good chance at approximate balance. It justifies confidence that a given sample is reasonably well balanced, in the absence of evidence to the contrary. It does not justify ignoring evidence of imbalance when it occurs.

Note that under the above probability sampling plan the estimator $(1'V^{1/2}1)(1_s'V_s^{-1/2}Y_s)/n$, which is $\hat{T}(X : V)$ if both $V1$ and $V^{1/2}1$ belong to $\mathfrak{M}(X)$ and s is in $B(X : V)$, is unbiased with respect to the probability distribution generated by the sampling plan. But if the sample actually selected is not balanced on X (*i.e.* if s is not in $B(X : V)$) then this estimator is not unbiased under $M(X : V)$.

REFERENCES

GODAMBE, V.P., and JOSHI, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations - I. *Annals of Mathematical Statistics*, 36, 1707-1723.

ISAKI, C.T., and FULLER, W.A. (1987). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

KOTT, P.S. (1984). A fresh look at bias-robust estimation in a finite population. In *Proceedings of the Section Survey Research Methods, American Statistical Association*, 176-178.

PEREIRA, C.A., and RODRIGUES, J. (1983). Robust linear prediction in finite populations. *International Statistical Review*, 51, 293-300.

ROYALL, R.M., and HERSON, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68, 880-889.

ROYALL, R.M., and CUMBERLAND, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 73, 66-77.

SCOTT, A.J., BREWER, K.R.W., and HO, W.W.H. (1978). Finite population sampling and robust estimation. *Journal of the American Statistical Association*, 73, 359-361.

TALLIS, G.W. (1986). On the optimality of balanced sampling. *Statistics and Probability*, 4, 141-144.

TAM, S.M. (1986). Characterization of best model-based predictors in survey sampling. *Biometrika*, 73, 232-235.

WALLENIUS, K.T. (1980). Statistical methods in sole source contract negotiation. *Journal of Undergraduate Mathematics and Applications*, 0, 35-47.