NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

The University of Wollongong

Working Paper

11-13

POTENTIAL GAINS FROM USING UNIT LEVEL COST INFORMATION IN A MODEL-ASSISTED FRAMEWORK

David G. Steel and Robert Graham Clark

National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

POTENTIAL GAINS FROM USING UNIT LEVEL COST INFORMATION IN A MODEL-ASSISTED FRAMEWORK

David G. Steel and Robert Graham Clark [1]

Key Words: optimal allocation; optimal design; sample design; sampling variance; survey costs

Word Count: 2949

ABSTRACT

In developing the sample design for a survey we attempt to produce a good design for the funds available. Information on costs can be used to develop sample designs that minimise the sampling variance of an estimator of total for fixed costs. Improvements in survey management systems mean that it is now sometimes possible to estimate the cost of including each unit in the sample. This paper develops relatively simple approaches to determine whether the potential gains arising from using this unit level cost information are likely to be of practical use. It is shown that the key factor is the coefficient of variation of the costs relative to the coefficient of variation of the relative error on the estimated cost coefficients.

[1]National Institute for Applied Statistics Research Australia, University of Wollongong, NSW Australia 2522.

1. INTRODUCTION

Unequal unit costs have been reflected in sample designs by using simple linear cost models. In stratified sampling, a per-unit cost coefficient can sometimes be estimated for each stratum. The resulting allocation of sample to strata is proportional to the inverse of the square root of the stratum cost coefficients (Cochran 1977). In a multistage design the costs of including the units at the different stages of selection can be used to decide the number of units to select at each stage (Hansen et al. 1953).

While this theory is well established, unequal costs have not been used extensively in practice (Brewer and Gregoire 2009), perhaps because of a lack of good information on costs, and because of a focus on sample size rather than cost of enumeration. Groves (1989) argued that linear cost models are unrealistic, and that mathematical cost modelling can distract from more important decisions such as the mode of collection, the number of callbacks and how the survey interacts with other surveys conducted by the same organisation. Nevertheless, given the pressures on survey budgets, the final design should reflect costs and variance in a rational way, without being fixated on formal optimality.

Increasing use of computers in data collection is leading to more extensive and useful cost-related information on units on survey frames. In a programme of business surveys conducted by a national statistics institute, most medium and large enterprises will be selected in some surveys at least every year or two. This may provide information on costs for those businesses, for example some businesses may have required extensive follow-up or editing in a previous survey. Direct experience is less likely to be available for any given small business, but

datasets of costs could be modelled to give predictions of likely costs.

Similarly, computer assisted personal interviewing and computer-assisted telephone interviewing are widely used in surveys of people and households. Survey management software routinely collects detailed information on time spent by interviewers on different tasks, leading to a wealth of data relating to survey costs, which could then be used to model how cost relates to geography, age, sex and other variables.

One example of this trend is the creation of an Operations Research Unit by the Australian Bureau of Statistics (ABS 2007, p151). The focus of this unit is on the analysis of paradata about survey processes and costs to improve collection methodology, but a side effect may be more detailed cost data for use in sample design.

There are therefore increasing possibilities for compiling finely grained or even unit-specific cost estimates, which can potentially be used for sample design. However, unit cost estimates are likely to be far from perfect, and ignoring the uncertainty attached to them may lead to less efficient designs. Also, the cost of maintaining and modelling cost databases needs to be balanced against realistic evaluations of the improvement in sampling efficiency.

This paper develops relatively simple approximations to the gains arising from using unit level cost information in a model-assisted framework. Section 2 contains notation and some key expressions. Section 3 is concerned with the optimal design when cost parameters are known. Section 4 analyses the use of estimated unit costs, and Section 5 is a summary.

2. NOTATION AND OBJECTIVE CRITERION

Consider a finite population, U containing N units, consisting of values Y_i for $i \in U$. A sample $s \in U$ is to be selected using an unequal probability sampling scheme with positive probability of selection $\pi_i = P[i \in s]$ for all units $i \in U$. A vector of auxiliary variables $\boldsymbol{x_i}$ is assumed to be available either for the whole population, or for all units $i \in s$ with the population total, $\boldsymbol{t_x} = \sum_{i \in U} \boldsymbol{x_i}$, also known. The auxiliary variables could consist of, for example, industry, region and size in a business survey, or age, sex and region in a household survey.

In the model-assisted approach (see for example Särndal et al. 1992), the relationship between a variable of interest and the auxiliary variables is captured in a model, typically of the following form in single-stage surveys:

$$\left.\begin{array}{l} E_M\left[Y_i\right] = \boldsymbol{\beta^T x_i} \\ var_M\left[Y_i\right] = \sigma^2 z_i \\ Y_i \text{ independent of } Y_j \text{ for all } i \neq j \end{array}\right\} \tag{1}$$

where E_M and var_M denote expectation and variance under the model, $\boldsymbol{\beta}$ and σ^2 are unknown parameters and z_i are assumed to be known for all $i \in U$. Let E_p and var_p denote expectation and variance under repeated probability sampling with all population values held fixed.

The generalized regression estimator is a widely used model-assisted estimator of t_y:

$$\hat{t}_y = \sum_{i \in s} \left(y_i - \hat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{x_i}\right) + \hat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{t_x} \tag{2}$$

where $\hat{\boldsymbol{\beta}}$ may be a weighted or unweighted least squares regression coefficient of y_i on $\boldsymbol{x_i}$ using sample data.

The *anticipated variance* of $\hat{t}_y$ is defined by $E_M var_p\left[\hat{t}_y - t_y\right]$, and is

approximated by

$$E_M var_p\left[\hat{t}_y\right] \approx \sigma^2 \sum_{i \in U} \left(\pi_i^{-1} - 1\right) z_i \tag{3}$$

for large samples (Särndal et al. 1992, formula 12.2.12, p451) under model (1). Model-assisted designs and estimators should minimise $E_M var_p\left[\hat{t}_y\right]$ subject to approximate design unbiasedness, $E_p\left[\hat{t}_y\right] = t_y$. The anticipated variance has been used to motivate model-assisted sample designs in one stage (Särndal et al. 1992) and two stage sampling (Clark and Steel 2007; Clark 2009). One advantage of using the anticipated variance for this purpose is that it depends only on the selection probabilities and a small number of model parameters, which can be roughly estimated when designing the sample. In contrast, $var_p\left[\hat{t}_y\right]$ typically depends on the population values of y_i and on joint probabilities of selection, both of which are difficult to quantify in advance.

The cost of enumerating a sample is assumed to be $C = \sum_{i \in s} c_i$ where c_i is the cost of surveying a particular unit i. The values of c_i are usually assumed to be known. Typically c_i are also assumed to be constant for all units in the population, or constant within strata. With the generalization that c_i may be different for every unit i, the cost C depends on the particular sample s selected. The expected cost is $E_p[C] = \sum_{i \in U} \pi_i c_i$. The aim is to minimise the anticipated variance (3) subject to a constraint on the expected enumeration cost,

$$\sum_{i \in U} \pi_i c_i = C_f. \tag{4}$$

There will also be fixed costs that are not affected by the sample design and so do not have to be included here.

Some notation for population variances and covariances is needed. Consider the pairs (u_i, v_i), and let $S_{uv} = N^{-1} \sum_{i \in U} (u_i - \bar{u})(v_i - \bar{v})$ denote their population

covariance, and $S_u^2 = N^{-1}\sum_{i\in U}(u_i - \bar{u})^2$ denote the population variance of (u_i). Let $\bar{u}$ and $\bar{v}$ be the population means of u_i and v_i. The population coefficient of variation of (u_i) is $C_u = S_u/\bar{u}$. The population relative covariance of (u_i, v_i) is $C_{u,v} = S_{uv}/\bar{u}\bar{v}$. A useful result is

$$\sum_{i\in U} u_i v_i = N\bar{u}\bar{v}\left(1 + C_{u,v}\right). \tag{5}$$

3. OPTIMAL DESIGN WITH KNOWN COST AND VARIANCE PARAMETERS

3.1 Optimal Model-Assisted Design

The values of $(\pi_i : i \in U)$ which minimise (3) subject to (4) are

$$\pi_i = C_f \frac{z_i^{1/2} c_i^{-1/2}}{\sum_{j\in U} z_j^{1/2} c_j^{1/2}} \propto z_i^{1/2} c_i^{-1/2} \tag{6}$$

and the resulting anticipated variance is

$$AV_{opt} = E_M var_p\left[\hat{t}_y\right] = \sigma^2 C_f^{-1}\left(\sum_{i\in U} c_i^{1/2} z_i^{1/2}\right)^2 - \sigma^2 \sum_{i\in U} z_i. \tag{7}$$

This can be easily derived using Lagrange multipliers or the Cauchy-Schwarz Inequality, and generalizes Särndal et al. (1992, Result 12.2.1, p452) to allow for unequal costs. Higher probability of selection is given to units which have higher unit variance or lower cost. However the square roots of z_i and c_i in (6) means that probabilities of selection do not vary dramatically in many surveys.

It is assumed that the last term of (7), which represents the finite population correction, is negligible. Applying (5) gives:

$$AV_{opt} \approx \frac{\sigma^2 C_f^{-1} N^2 \bar{c}\bar{z}\left(1 + C_{\sqrt{c},\sqrt{z}}\right)^2}{\left(1 + C_{\sqrt{c}}^2\right)\left(1 + C_{\sqrt{z}}^2\right)} \tag{8}$$

where $C_{\sqrt{c}}$ and $C_{\sqrt{z}}$ refer to the population coefficients of variation of $\sqrt{c_i}$ and $\sqrt{z_i}$, respectively. To make our results interpretable, we will assume that unit costs c_i and variances σz_i are unrelated, so that $C_{\sqrt{c},\sqrt{z}} = 0$. This assumption may not always be satisfied in practice, but any relationship between c_i and z_i will be specific to the particular example, and could be either positive or negative. To identify general principles, it makes sense to ignore any such relationship. In practice, it is often reasonable to also assume that C_c^2 and C_z^2 are small. A Taylor Series expansion then shows that $C_c^2 \approx 4C_{\sqrt{c}}^2$ and $C_z^2 \approx 4C_{\sqrt{z}}^2$. Putting these approximations together, (8) becomes

$$AV_{opt} = \frac{\sigma^2 C_f^{-1} N^2 \bar{c}\bar{z}}{\left(1 + \frac{1}{4}C_c^2\right)\left(1 + \frac{1}{4}C_z^2\right)} \tag{9}$$

See the appendix for details of these derivations.

Ignoring Costs

If the costs are ignored, then (6) suggests that $\pi_i \propto z_i^{1/2}$. To make comparisons for the same expected cost, C_f,

$$\pi_i = C_f \frac{z_i^{1/2}}{\sum_{j \in U} z_j^{1/2} c_j}. \tag{10}$$

with resulting anticipated variance

$$AV_{nocosts} = \sigma^2 C_f^{-1} \left(\sum_{i \in U} z_i^{1/2}\right)\left(\sum_{i \in U} c_i z_i^{1/2}\right) - \sigma^2 \sum_{i \in U} z_i. \tag{11}$$

Applying derivations similar to those used in Section 3.1,

$$AV_{nocosts} \approx \frac{\sigma^2 C_f^{-1} N^2 \bar{c}\bar{z}}{\left(1 + \frac{1}{4}C_z^2\right)}. \tag{12}$$

(See the appendix for details.) Comparing (12) and (9), we see that taking costs into account in the design results in dividing the anticipated variance by $\left(1 + \frac{1}{4}C_c^2\right)$.

4. THE EFFECT OF USING ESTIMATED COST PARAMETERS

In practice, c_i are not known precisely. Suppose that estimates $\hat{c}_i = b_i c_i$ are used instead. Using the auxiliary variable and the estimated costs in the optimal probabilities implies $\pi_i \propto z_i^{1/2} \hat{c}_i^{-1/2}$. To make comparisons for the same expected cost,

$$\pi_i = C_f \frac{z_i^{1/2} \hat{c}_i^{-1/2}}{\sum_{j \in U} z_j^{1/2} \hat{c}_j^{-1/2} c_j}.$$

The resulting anticipated variance is

$$AV_{ests} = \sigma^2 C_f^{-1} \left(\sum_{i \in U} \hat{c}_i^{1/2} z_i^{1/2} \right) \left(\sum_{j \in U} z_j^{1/2} \hat{c}_j^{-1/2} c_j \right) - \sigma^2 \sum_{i \in U} z_i. \tag{13}$$

If we assume that the values of b_i are unrelated to the values of c_i and z_i, then

$$\begin{aligned} AV_{ests} &= \sigma^2 C_f^{-1} \left(\sum_{i \in U} c_i^{1/2} z_i^{1/2} \right)^2 N^{-2} \left(\sum_{i \in U} b_i^{-1/2} \right) \left(\sum_{i \in U} b_i^{1/2} \right) \\ &\quad -\sigma^2 \sum_{i \in U} z_i \end{aligned} \tag{14}$$

(see the appendix for details.) If the coefficient of variation of b_i is small, then a Taylor Series approximation gives $N^{-2} \sum b_i^{-1/2} \sum b_i^{1/2} \approx 1 + \frac{1}{4} C_b^2$. Applying this, and the same approximations as in Subsection 3.1, (14) becomes

$$AV_{ests} = \frac{\sigma^2 C_f^{-1} N^2 \bar{c} \bar{z} \left(1 + \frac{1}{4} C_b^2\right)}{\left(1 + \frac{1}{4} C_c^2\right) \left(1 + \frac{1}{4} C_z^2\right)} \tag{15}$$

Comparing (15) and (12), the effect of using estimated cost parameters rather than no costs at all is to multiply the anticipated variance by $\left(1 + \frac{1}{4} C_b^2\right) / \left(1 + \frac{1}{4} C_c^2\right)$. Therefore cost information is worth using provided $C_b < C_c$. The coefficient of variation of the error factors has to be less than that of the true unit costs over the population.

5. EXAMPLES OF COST MODELS

The key quantities determining the usefulness of the unit cost data are C_b and C_c. Optimal designs using unequal cost information are not very common, so there is relatively little literature on the typical values of these measures. Unequal costs may be driven by a variety of factors, including mode effects, geography and willingness to respond, and literature on these issues is helpful to give a rough idea of cost models that may apply in practice.

Groves (1989, p.538) compares per-respondent costs of telephone interviewing ($38.00) and personal interviewing ($84.90) of the general population. If the preference of all units on a frame were known, and half preferred each mode, this would imply $C_c = 0.38$. Greenlaw and Brown-Welty (2009) compared paper and web surveys, and found per-respondent costs of $4.78 and $0.64, respectively, in a survey of members of a professional association. In a mixed mode option, two thirds of respondents opted for the web option. If preferences are known in advance, then $C_c = 0.76$.

Another reason for varying costs is that some respondents are more difficult to recruit than others, requiring more visits or reminders. Groves and Heeringa (2006, Section 2.2) trialled a survey where interviewers classified non-respondents from the first approach as either likely or unlikely to respond. In subsequent follow-up, the first group had a response rate of 73.7% compared to 38.5% for the second group. This suggests that the per-respondent cost for the second group would be at least 1.9 times higher than the first group. (In fact, the ratio would be higher, because more follow-up attempts would be made for the difficult group.) If 50% of respondents are in both groups, then $C_c = 0.31$.

Geography is another source of differential costs in interviewer surveys. In the Australian Labour Force Survey, costs have been modelled as having a per-block component and a per-dwelling component (Hicks 2001, Table 4.2.1 in Section 4.2) depending on the type of area (15 types were defined). Assuming a constant 10 dwellings sampled per block, the net per-dwelling costs range from \$4.98 in Inner City Sydney and Melbourne to \$6.71 in Sparse and Indigenous areas. While this is a significant difference in costs across area types, the great majority of the population are in three area types (settled area, outer growth and large town) where per-dwelling costs vary only between \$5.71 and \$6.07. As a result, C_c is estimated at a very small 0.054.

Table 1 shows the approximate percentage improvement in the anticipated variance from using estimated cost information for different values of C_c and C_b, some suggested by these examples. Negative values indicate that the design is less efficient than ignoring costs altogether. The table suggests that cost information is only worthwhile provided there is a fair variation in the unit costs, otherwise the benefit is very small, and can be erased when there is even small imprecision in the estimated costs. Mixed mode surveys are probably the best candidate for exploiting varying unit costs in sample design.

Table 1: Percentage improvement in anticipated variance from using estimated cost information compared to no cost information

Coefficient of Variation of Unit Costs (C_c) (%)	Possible Scenario	Coefficient of Variation of Error Factor (C_b) (%)			
		0	10	25	50
5		0.1	-0.2	-1.5	-6.2
10	interviewer travel due to remoteness	0.2	0.0	-1.3	-6.0
20		1.0	0.7	-0.6	-5.2
30	response propensity	2.2	2.0	0.7	-3.9
40	mixed mode (phone/personal int.)	3.8	3.6	2.3	-2.2
50		5.9	5.6	4.4	0.0
75	mixed mode (paper/web self-complete)	12.3	12.1	11.0	6.8

6. DISCUSSION

Incorporating unequal unit costs can improve the efficiency of sample designs. For the gains to be appreciable, the unit costs need to vary considerably. Even with no estimation error, a coefficient of variation of 50% may lead to a gain of only 6% in the anticipated variance. When this coefficient of variation is 75%, as can happen in a mixed mode survey, the reduction in the anticipated variance (or in the sample size for fixed precision) can be over 12%. Costs will be estimated with some error and this reduces the gain by a factor determined by the relative variation of the relative errors in estimating the costs at the individual level.

REFERENCES

ABS (2007), "Australian Bureau of Statistics Annual Report 2006-07," www.abs.gov.au/ausstats/abs@.nsf/mf/1001.0.

Brewer, K. and Gregoire, T. G. (2009), "Introduction to Survey Sampling," in *Handbook of Statistics 29A: Sample Surveys: Design, Methods and Applications*, eds. Pfeffermann, D. and Rao, C. R., Amsterdam: Elsevier/North-Holland, pp. 9–37.

Clark, R. G. (2009), "Sampling of subpopulations in two-stage surveys," *Statistics in Medicine*, 28, 3697–3717.

Clark, R. G. and Steel, D. G. (2007), "Sampling within households in household surveys," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170, 63–82.

Cochran, W. (1977), *Sampling Techniques*, New York: Wiley, 3rd ed.

Greenlaw, C. and Brown-Welty, S. (2009), "A comparison of web-based and paper-based survey methods testing assumptions of survey mode and response cost," *Evaluation Review*, 33, 464–480.

Groves, R. M. (1989), *Survey Errors and Survey Costs*, New York: Wiley.

Groves, R. M. and Heeringa, S. G. (2006), "Responsive design for household surveys: tools for actively controlling survey errors and costs," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169, 439–457.

Hansen, M., Hurwitz, W., and Madow, W. (1953), *Sample Survey Methods and Theory Volume 1: Methods and Applications*, New York: John Wiley and Sons.

Hicks, K. (2001), "Cost and Variance Modelling for the 2001 Redesign of the Monthly Population Survey," www.abs.gov.au/ausstats/abs@.nsf/mf/1352.0.55.037, Australian Bureau of Statistics Methodology Advisory Committee Paper.

Särndal, C., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

APPENDIX: DETAILED DERIVATIONS

Lemma 1: Let u_i be defined for $i \in U$. Let $u_i = \bar{u} + \theta e_i$, where $\sum_{i \in U} e_i = 0$ and θ is small. Then:

a. $\overline{\sqrt{u}} = \sqrt{\bar{u}} - \frac{1}{8}\theta^2 \bar{u}^{-3/2} S_e^2 + o\left(\theta^2\right)$.

b. $S_{\sqrt{u}}^2 = \frac{1}{4}\theta^2 \bar{u}^{-1} S_e^2 + o\left(\theta^2\right) = \frac{1}{4}\bar{u}^{-1} S_u^2 + o\left(\theta^2\right)$.

c. $N^{-2}\left(\sum_{i \in U} u_i^{1/2}\right)\left(\sum_{i \in U} u_i^{-1/2}\right) = 1 + \frac{1}{4}\theta^2 \bar{u}^{-2} S_e^2 + o\left(\theta^2\right) = 1 + \frac{1}{4} C_u^2 + o\left(\theta^2\right)$.

d. $C_{\sqrt{u}}^2 = \frac{1}{4}\theta^2 \bar{u}^{-2} S_e^2 + o\left(\theta^2\right) = \frac{1}{4} C_u^2 + o\left(\theta^2\right)$.

The notation $o\left(C_u^2\right)$ can be used in place of $o\left(\theta^2\right)$, since $C_u^2 = \theta^2 C_e^2$. This will be done in the remainder of the Appendix.

Proof:

We start by writing $\overline{\sqrt{u}}$ as a function of θ:

$$\overline{\sqrt{u}} \;=\; N^{-1} \sum_{i \in U} \sqrt{u_i} = N^{-1} \sum_{i \in U} \sqrt{\bar{u} + \theta e_i}$$

Call this $g\left(\theta\right)$, then differentiating about $\theta = 0$ gives $g(0) = \sqrt{\bar{u}}$, $g'(0) = 0$ and

$$g''(0) = -\frac{1}{4} N^{-1} \bar{u}^{-3/2} \sum_{i \in U} e_i^2 = -\frac{1}{4} \bar{u}^{-3/2} S_e^2.$$

Hence

$$\overline{\sqrt{u}} = g\left(\theta\right) = g(0) + g'(0)\theta + \frac{1}{2} g''(0)\theta^2 + o\left(\theta^2\right) = \sqrt{\bar{u}} - \frac{1}{8}\theta^2 \bar{u}^{-3/2} S_e^2 + o\left(\theta^2\right)$$

which is result a.

Result b is proven using result a:

$$\begin{aligned} S^2_{\sqrt{u}} &= N^{-1}\sum_{i\in U}\left(\sqrt{u_i}\right)^2 - \left(N^{-1}\sum_{i\in U}\sqrt{u_i}\right)^2 \\ &= \bar{u} - \left(\overline{\sqrt{u}}\right)^2 \\ &= \bar{u} - \left(\sqrt{\bar{u}} - \frac{1}{8}\theta^2\bar{u}^{-3/2}S_e^2 + o\left(\theta^2\right)\right)^2 \\ &= \bar{u} - \left(\bar{u} + \frac{1}{64}\theta^4\bar{u}^{-3}S_e^4 - \frac{1}{4}\theta^2\bar{u}^{-1}S_e^2 + o\left(\theta^2\right)\right) \\ &= \frac{1}{4}\theta^2\bar{u}^{-1}S_e^2 + o\left(\theta^2\right) = \frac{1}{4}\bar{u}^{-1}S_u^2 + o\left(\theta^2\right) \end{aligned}$$

To derive c, we firstly write $N^{-1}\sum_{i\in U}u_i^{-1/2}$ as a function $g()$ of θ and take a Taylor Series expansion:

$$\begin{aligned} N^{-1}\sum_{i\in U}u_i^{-1/2} &= N^{-1}\sum_{i\in U}\left(\bar{u}+\theta e_i\right)^{-1/2} \\ &= g\left(\theta\right) = g(0) + g'(0)\theta + \frac{1}{2}g''(0)\theta^2 + o\left(\theta^2\right) \\ &= \bar{u}^{-1/2} + 0\theta + \frac{1}{2}\frac{3}{4}\bar{u}^{-5/2}N^{-1}\sum_{i\in U}e_i^2\theta^2 + o\left(\theta^2\right) \\ &= \bar{u}^{-1/2} + \frac{3}{8}\bar{u}^{-5/2}S_e^2\theta^2 + o\left(\theta^2\right) \end{aligned} \tag{16}$$

Note that$N^{-1}\sum_{i\in U}u_i^{1/2} = \overline{\sqrt{u}}$. Multiplying the expression for $\overline{\sqrt{u}}$ in result a and (16) gives

$$\begin{aligned} & N^{-2}\left(\sum_{i\in U}u_i^{1/2}\right)\left(\sum_{i\in U}u_i^{-1/2}\right) \\ & = \left\{\sqrt{\bar{u}} - \frac{1}{8}\theta^2\bar{u}^{-3/2}S_e^2 + o\left(\theta^2\right)\right\}\left\{\bar{u}^{-1/2} + \frac{3}{8}\bar{u}^{-5/2}S_e^2\theta^2 + o\left(\theta^2\right)\right\} \\ = \; & 1 + \frac{1}{4}\bar{u}^{-2}S_e^2\theta^2 + o\left(\theta^2\right) \\ = \; & 1 + \frac{1}{4}C_u^2 + o\left(\theta^2\right) \end{aligned}$$

which is result c.

For result d, firstly note that $\overline{\sqrt{u}} = \sqrt{\bar{u}} + o\left(\theta\right)$ from result a, and so, from a first

order Taylor Series,

$$\left(\overline{\sqrt{u}}\right)^{-2} = \left(\sqrt{\bar{u}}\right)^{-2} + o\left(\theta\right) = \bar{u}^{-1} + o\left(\theta\right).$$

Combining this with result b, we obtain

$$\begin{aligned} C^2_{\sqrt{u}} &= S^2_{\sqrt{u}} \left(\overline{\sqrt{u}}\right)^{-2} \\ &= \left\{\frac{1}{4}\theta^2 \bar{u}^{-1} S^2_e + o\left(\theta^2\right)\right\} \left\{\bar{u}^{-1} + o\left(\theta\right)\right\} \\ &= \frac{1}{4}\theta^2 \bar{u}^{-2} S^2_e + o\left(\theta^2\right) \\ &= \frac{1}{4} C^2_u + o\left(\theta^2\right) \end{aligned}$$

giving result d.

Derivation of (8)

For the special case where $u_i = v_i$, (5) becomes

$$\sum_{i \in U} u_i^2 = N \bar{u}^2 \left(1 + C^2_u\right). \tag{17}$$

Applying (5),

$$\sum_{i \in U} c_i^{1/2} z_i^{1/2} = N \overline{\sqrt{c}}\ \overline{\sqrt{z}} \left(1 + C_{\sqrt{c},\sqrt{z}}\right) \tag{18}$$

where $\overline{\sqrt{c}} = N^{-1} \sum_{i \in U} \sqrt{c_i}$ and $\overline{\sqrt{z}} = N^{-1} \sum_{i \in U} \sqrt{z_i}$. Using (17), we can express $\overline{\sqrt{c}}$ in terms of $\bar{c}$:

$$\bar{c} = N^{-1} \sum_{i \in U} c_i = N^{-1} \sum_{i \in U} \left(\sqrt{c_i}\right)^2 = \left(\overline{\sqrt{c}}\right)^2 \left(1 + C^2_{\sqrt{c}}\right). \tag{19}$$

Similarly,

$$\bar{z} = \left(\overline{\sqrt{z}}\right)^2 \left(1 + C^2_{\sqrt{z}}\right). \tag{20}$$

Assuming the last term of (7) is negligible, applying (18), (19) and (20) gives (8).

Derivation of (9)

Lemma 1d implies that $C^2_{\sqrt{c}} = \frac{1}{4}C^2_c + o\left(C^2_c\right) \approx \frac{1}{4}C^2_c$ and $C^2_{\sqrt{z}} = \frac{1}{4}C^2_z + o\left(C^2_z\right) \approx \frac{1}{4}C^2_z$. Result (9) follows from (8) by using these approximations, as well as assuming that $C_{\sqrt{c},\sqrt{z}} = 0$.

Derivation of (12)

Firstly, $\sum_{i\in U} c_i z_i^{1/2} = N\bar{c}\overline{\sqrt{z}}\left(1 + C_{c,\sqrt{z}}\right)$, from (5), where $C_{c,\sqrt{z}}$ is the population relative covariance between the values of $z_i^{1/2}$ and c_i. It is assumed that the values of c_i and z_i are unrelated, so that $C_{c,\sqrt{z}} = 0$. It is also assumed that the second term of (11) is negligible, corresponding to small sampling fraction. Hence (11) becomes:

$$AV_{nocosts} = \sigma^2 N^2 C_f^{-1} \bar{c} \left(\overline{\sqrt{z}}\right)^2. \tag{21}$$

From (20), and Lemma 1d, we have

$$\left(\overline{\sqrt{z}}\right)^2 = \frac{\bar{z}}{1 + C^2_{\sqrt{z}}} \approx \frac{\bar{z}}{1 + \frac{1}{4}C^2_z}$$

Substituting into (21) gives (12).

Derivation of (14)

Two terms in (13) will be simplified using (5). Firstly,

$$\begin{aligned} \sum_{i\in U} \hat{c}_i^{1/2} z_i^{1/2} &= \sum_{i\in U} b_i^{1/2} c_i^{1/2} z_i^{1/2} \\ &= N\left(N^{-1}\sum_{i\in U} b_i^{1/2}\right)\left(N^{-1}\sum_{i\in U} c_i^{1/2} z_i^{1/2}\right) + C_{\sqrt{b},\sqrt{cz}} \end{aligned} \tag{22}$$

where $C_{\sqrt{b},\sqrt{cz}}$ is the covariance between the population values of $b_i^{1/2}$ and $c_i^{1/2} z_i^{1/2}$.

Secondly,

$$\begin{aligned}\sum_{i\in U} z_i^{1/2}\hat{c}_i^{-1/2}c_i &= \sum_{i\in U} b_i^{-1/2}c_i^{1/2}z_i^{1/2} \\ &= N\left(N^{-1}\sum_{i\in U} b_i^{-1/2}\right)\left(N^{-1}\sum_{i\in U} c_i^{1/2}z_i^{1/2}\right) + C_{1/\sqrt{b},\sqrt{cz}} \quad (23)\end{aligned}$$

where $C_{1/\sqrt{b},\sqrt{cz}}$ is the covariance between the population values of $b_i^{-1/2}$ and $c_i^{1/2}z_i^{1/2}$.

If we assume that the population values of b_i are unrelated to the values of c_i and z_i, so that $C_{\sqrt{b},\sqrt{cz}} = C_{1/\sqrt{b},\sqrt{cz}} = 0$, and subsitute (22) and (23) into (13), then we obtain (14).

Derivation of (15)

We can express (14) in terms of AV_{opt} which is defined in (7), assuming the last term of (7) is negligible, corresponding to small sampling fraction:

$$AV_{ests} \approx AV_{opt}N^{-2}\sum_{i\in U} b_i^{-1/2}\sum_{i\in U} b_i^{1/2} \quad (24)$$

Lemma 1c implies that

$$N^{-2}\sum_{i\in U} b_i^{-1/2}\sum b_i^{1/2} = 1 + \frac{1}{4}C_b^2 + o\left(C_b^2\right) \approx 1 + \frac{1}{4}C_b^2.$$

Substituting this, and (8), into (24) gives (15).