Recent Developments in Sample Design using Mathematical Programming

Sun-Woong Kim[1]

Abstract

We discuss why sample design by mathematical programming can be beneficial to practical surveys. We illustrate some developments of software for sample design using mathematical programming in several statistical organizations. Also, we present certain restrictions on the use of mathematical programming.

Key Words: Sample Allocation, Linear Programming, Nonlinear Programming

1. Introduction

Survey sampling statisticians have been principally concerned with efficient sample designs. Survey sample design can profit by a mathematical programming formulation. A variety of sample designs using mathematical programming have recently developed, and some statistical organizations such as the U.S. Bureau of Labor Statistics (BLS), the Westat, and the Institute for Social Research (ISR) at the University of Michigan etc. have been successful in developing software to easily explore mathematical programming solutions for practical surveys.

In this paper, we introduce how mathematical programming can offer solutions to typical or atypical sample designs and discuss some benefits of an application of mathematical programming. Also, we suggest certain restrictions in using mathematical programming and future applications.

2. Sample Design and Mathematical Programming

2.1 Sample Allocation

Concerning survey sample design, we may focus on sample allocation problems. An explicit solution is available especially within the simple context of a single variable of interest. However, given the multi-variate nature of sample surveys or related to multi-purpose surveys sponsored by governments, some different techniques to solve sample allocation problems are required, and they would be far more complicated due to two basic criteria: (a) the minimization of a cost function subject to restrictions with regard to the sampling variances of the estimates (b) the minimization of a variance function subject to a cost model.

[1]Full-time lecturer, Department of Statistics, Dongguk University, e-mail:sunwk@dongguk.edu

The optimum allocation by mathematical programming has been used to solve complex sample optimization problems. Traditionally, mathematical programming problems can be divided into linear programming (LP) and nonlinear programming (NLP).

A general linear programming problem may be expressed as follows:

$$\begin{aligned} &\text{minimize} \quad \phi = c^T x \\ &\text{subject to} \quad Ax = b \ , \\ &\qquad\qquad\quad x \geq 0 \ , \end{aligned}$$

where x is an n vector $x=(x_1, \cdots, x_n)^T$, c is an n vector, b is an m vector and A is an $m \times n$ matrix.

On the other hand, we can specify a nonlinear programming problem below:

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad g_i(x) \leq 0 \ \text{ for } \ i = 1, \cdots, m \ , \\ &\qquad\qquad\quad h_i(x) \leq 0 \ \text{ for } \ i = 1, \cdots, l \ , \end{aligned}$$

where $f(x)$ is a nonlinear(or linear) function of an n vector $x=(x_1, \cdots, x_n)^T$ and $g_1, \cdots, g_m, h_1, \cdots, h_l$ are linear(or nonlinear) constraints on n decision variables.

We need to take a look at some allocation problems that have actually used in the U.S., as in the followings.

2.1.1 Multi-Dimensional Stratification Problem

Green (2000), who has been working for the Westat, considered a national probability sample of children for the National Center for Educational Statistics'(NECS) Early Childhood Longitudinal Study in the U.S. Detailed data was available on mother's and father's race/ethnicity (domain H), the child's birth weight (domain I) and plurality such as single birth, twins, triplets etc. (domain J). Using each domain as an independent stratification variable, the sample allocation problem was treated as a three-dimensional stratification problem. He suggested the following mathematical programming problem for deciding the actual sample size in each cell (hij):

$$\begin{aligned} &\text{Minimize} \quad \phi = \sum_{h=1}^{H} \sum_{i=1}^{I} \sum_{j=1}^{J} n_{hij} \\ &\text{Subject to} \quad 0 < n_{hij} \leq N_{hij} \ , \\ &\qquad\qquad\quad \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} n_{hij}}{d_h} \geq t_h \ \text{ etc.,} \end{aligned}$$

where n_{hij} is the actual sample size in cell,

N_{hij} is the population size in cell hij ,

t_h, t_i, t_j are the target effective sample sizes of levels h, i, j in domains H, I, J ; and

$d_h = \frac{n_h}{N_h^2} \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{N_{hij}^2}{n_{hij}}$ etc. are the design effects for levels h, i, j in domains H, I, J .

Although it looks like a linear programming problem, it is a nonlinear programming problem because the constraints that depend on d_h is a function of n_{hij}^{-1} . Note that Green (2000) does not give the details on optimization algorithms he used for obtaining the solution that consists of sample sizes in cells.

2.1.2 Multivariate Sample Allocation

The sample allocation problem in multipurpose surveys is complicated due to the fact that an efficient allocation for some estimates may not be efficient for others. Also, certain constraints on costs and minimum sample sizes for strata should be added to permit variance estimation with regard to precision goals.

Valiant (1994) and Valiant and Gentle (1997), who were working for the U.S. Bureau of Labor Statistics, have formulated more or less complicated optimization problem, in order to decide the sample sizes in multipurpose surveys such as the Employment Cost Index (ECI) and the Employee Benefits Surveys (EBS). Two surveys use the same two-stage sample of establishments and occupations to estimate personnel costs and the percentages of employees receiving benefits. Since the population as a whole and domains are considered, estimates from the surveys are made of an index of change in costs between time periods, of the average cost per employee per hour worked, and of the percentage of employees receiving various benefits within domains, including industry group (e.g., construction, manufacturing, wholesale, services), establishment size, class of worker, and geographic region. An objective function for sample allocation problems is a weighted combination of the relvariances (the variance of estimator divided by the square of its expected value) of different estimators and the total cost. Each weight reflects on the importance of each relvariance and the cost. Formally, the objective function under certain constraints is as follows:

Minimize $\phi = \sum_{i=l}^{L}{}_{l} w_l \delta_l + w_{L+1} c$

subject to

(1) $n_{h,\min} \le n_h \le N_h$ for establishment sample sizes n_h ,

(2) $n = \sum_h n_h \le n_0$, a bound on the total number of sample establishment,

(3) $m_{h,\min} \leq \overline{m}_h \leq m_{h,\max}$, i.e. the number of occupations sampled per establishment in stratum h is bounded above and below,

(4) $\dfrac{\sum_{h \in S} n_h \overline{m}_h}{\sum_{h \in S} n_h} \leq \overline{m}_{S,\max}$, i.e. the average number of occupations sampled per establishment is bounded above in a subset S of strata,

(5) $\delta_l^{1/2} \leq \delta_{l0}^{1/2}$ for $l \in S_E$, i.e. the coefficient of variation of an estimator l is bounded for all estimators in some set S_E ,

where $h=(1,\cdots,H)$ is a stratum, w_l is a weight with estimator $l(l=1,\cdots,L)$, δ_l is an anticipated relvariance of the estimator and a nonlinear function of n_h and $\overline{m}_h$ (See Valiant and Gentle (1997), pp. 342-344), and c is the total survey cost that is a function of sample size and has its weight, which is either o or 1.

This optimization problem is a nonlinear programming on both objective function and constraints. It should be noted that the weights $\{w_l\}_{l=1}^{L}$ in the objective function depend on subjective judgments as to the relative importance of each estimator. They developed a software called ALLOCATE with a graphical user interface (GUI) to assign the weights of the optimization problem and then solve the problem. This software calls a C program know as GRG2 (Windward Technologies Inc. 1994) to solve the nonlinear programming problem. It seems that development of software is essential because analysts may have different opinions and modifying the weights should be flexible.

Valiant and Gentle (1997) emphasizes that nonlinear optimization can be a powerful technique in sample allocation in multipurpose surveys and commercial versions of some optimization algorithms are available. However, they do not provide the details on the optimization algorithms for nonlinear programming they adopted to solve their problems.

2.2. Other Applications

In this section we deal with other sample designs using mathematical programming, including controlled selection, raking, overlap control, area sampling, etc.

First, controlled selection problems have been recognized as a mathematical programming since Causey, Cox, and Ernst (1985). Subsequently, Rao and Nigam (1990, 1992), Sitter and Skinner (1994), and Kim, Heeringa, and Solenberger (2002) regarded controlled selection as a linear programming problem, while Causey, Cox, and Ernst (1985) treated as a transportation problem. Kim, Heeringa, and Solenberger (2002) at the ISR showed that their suggested sample design is more efficient than others concerning the minimization of the overall distortion of controlled selection problem and they developed software called SOCSLP for solving the problem.

Second, the problem of iterative proportional fitting (or raking) was regarded as a linear programming problem (transportation problem) by Arthanari and Dodge (1993) and Causey, Cox, and Ernst (1985).

Third, optimal method for maximizing the overlap between surveys has been treated as a linear programming since Causey, Cox, and Ernst (1985). Kim, Corteville, and Flanagan (2002) used a linear programming to maximize the PSU overlap between the 1990 and 2000 redesigns for the Current Population Survey (CPS), National Crime Victimization Survey (NCVS) etc. They purchased SUNSET software for solving their linear programming problem.

Forth, Westat in the U.S. developed standard software for area sampling (Krenzke and Green (2002)) and automatic PSU formation software (WesPSU). WesPSU especailly used mathematical programming approaches that are conceived as binary programming problems, given an appropriate objective function (i.e., cost or variance) and a number of constraints. Green, Chowdhury, and Krenzke (2002) showed the result that WesPSU was considerably faster than a typical manual effort.

2.3 Suggested Research and Restrictions

Since the 1980's, sample designs using mathematical programming in practical surveys has developed by a few of statistical organizations, especially in the U.S. Optimization algorithms for mathematical programming may not be familiar to most survey practitioners. As mentioned above, there are commercial versions of specialized software for mathematical programming, although we should develop a graphical user interface or command-line interface to be linked to commercial software.

Optimization algorithms may be probably inefficient if an objective function has a number of factors or the number of constraints is extremely large. So there should be an appropriate objective function or limited constraints, even though we could purchase an excellent software at a price.

Despite there being very few studies on nonlinear programming for sample design including sample allocation, compared to linear programming, an application of nonlinear programming to sample design may be useful for organizations that conduct a variety of surveys and must periodically redesign to update.

Fortunately, SAS/OR software includes some procedures to solve both linear programming and nonlinear programming. In our situation where SAS is broadly used, it may be suitable. However, some restrictions fall under the followings:

(a) There is no efficient method of solving the nonlinear programming problem in full generality.

(b) Though there are some exceptions, a global minimum point is not found consistently in most nonlinear programming algorithms. All the algorithms in SAS/OR NLP Procedure find a local minimum.

3. Concluding Remarks

In this paper, we discuss why and how survey sample design can benefit from a mathematical programming formulation.

There have been a lot of fascinated studies on applications of mathematical programming for sample design, including sample allocation, controlled selection, overlap control, raking, area sampling etc. For example, sample design for multi-dimensional or multivariate sample allocation is far more interesting than Neyman allocation based on a single variable. But development of some software should be essential in most situations. In the face of a number of restrictions mentioned above, an application of LP or NLP for sample surveys would be desirable with regard to both precision goals and costs.

References

[1] Arthanari, T. S. and Dodge, Y. (1993). *Mathematical Programming in Statistics*. Wiley, New York.

[2] Causey, B. D., Cox, L. H., and Ernst, L. R. (1985). Applications of Transportation Theory to Statistical Problems, *The Journal of the American Statistical Association*, Vol. 80, 903-909.

[3] Green, J., Chowdhury, S., and Krenzke, T. (2002). WesPSU-Development and Demonstration of Primary Sampling Unit (PSU) Formation Software, *Proceedings of the Section of Survey Research Methods of the American Statistical Assocation.*

[4] Green, J. (2000). Mathematical Programming for Sample Design and Allocation Problems, *Proceedings of the Section of Survey Research Methods of the American Statistical Assocation.*

[5] Kim, Jay J., Corteville, D. R. (2002). Maximizing Retention of Primary Sampling Units (PSUs) in a Two-PSU per Stratum Design, *Proceedings of the Section of Survey Research Methods of the American Statistical Assocation.*

[6] Kim, Sun W., Heeringa, S., and Solenberger, P. (2002). Optimizing Solutions in Two-way Controlled Selection Problems, *Proceedings of the Section of Survey Research Methods of the American Statistical Assocation.*

[7] Krenzke, T. and Green, J. (2002). When, Why and How to Develop Widely Used Standard Software for Area Sampling, *Proceedings of the Section of Survey Research Methods of the American Statistical Assocation.*

[8] Rao, J. N. K. and Nigam, A. K. (1990). Optimal Controlled Sampling Design, *Biometrika*, Vol. 77, 807-814.

[9] Rao, J. N. K. and Nigam, A. K. (1992). Optimal Controlled Sampling : A Unified Approach, *International Statistical Review*, Vol. 60, 89-98.

[10] SAS/OR (2001). *User's Guide : Mathematical Programming*, Version 8, SAS Institute Inc.

[11] Valiant, R. (1994). An Application of Mathematical Programming to Sample Allocation, *Proceedings of the Section of Survey Research Methods of the American Statistical Assocation.*

[12] Valiant, R. and Gentle, J. (1997). An Application of Mathematical Programming to Sample Allocation, *Computational Statistics & Data Analysis*, Vol. 25, 337-360.