Division of Research
Graduate School of Business Administration
The University of Michigan

May, 1981

MODEL-BASED STATISTICAL SAMPLING FOR COST ALLOCATION

Working Paper No. 267

Roger L. Wright

The University of Michigan

FOR DISCUSSION PURPOSES ONLY

Model-Based Statistical Sampling for Cost Allocation

Roger L. Wright*

1. Introduction

Statistical sampling and multiple regression analysis can be identified with the two stages of many managerial accounting and auditing projects, namely data collection and data analysis. These two stages can be integrated through a new methodology called model-based statistical sampling. This paper outlines the methodology and illustrates its use in allocating the cost of services. A specific class of applications in public utility load research is discussed.

COST EFFECTIVE MANAGEMENT INFORMATION SYSTEMS

Modern managerial accounting emphasizes the relationship between the cost of information and its contribution to better management decisions. Horngren [1977, p. 7] says, "the optimal accounting measure or system is the one that produces the greatest benefit net of the costs of obtaining the information." Demski and Feltham [1976] have provided a rigorous formulation of this approach to managerial accounting along the lines of Bayesian decision theory.

Bayesian decision theory originated in the efforts of mathematical statisticians to strengthen the foundations of statistical inference. Much of

*The author is Associate Professor of Statistics, Graduate School of Business Administration, The University of Michigan, Ann Arbor, MI 48109. This paper was prepared with the support of the U.S. Department of Energy, Grant No. DE-FG02-80ER10125. However, any opinions, findings, conclusions, or recommendations expressed herein are those of the author and do not necessarily reflect the views of the D.O.E.

The author wishes to thank C. F. Belknap, Jr. and other members of the Rate Research Department of Consumers Power Company for their cooperation, and C. E. Sarndal for helpful comments on the methodology.

the pioneering work was done by L. J. Savage [1954]. Arnold Zellner [1971] is leading a group of current workers who are applying the Bayesian approach to regression analysis and econometrics. Other statisticians, notably Carl Sarndal, [e.g., Cassel, et al., 1977] are examining the foundations of statistical inference in survey sampling.

Statistical sampling and regression analysis are at the heart of cost-effective procedures for collecting and analyzing managerial information. Statistical sampling is effectively employed in the valuation of inventories and receivables. Regression analysis is widely used in cost estimation and demand analysis. But until recently, there has been only a loose and often contradictory theoretical connection between these two methodologies--statistical sampling for data collection, and regression for data analysis.

Now however, the work of Sarndal and others is conceptually unifying the foundations of statistical sampling and regression analysis, and is providing the basis for an integrated methodology called model-based statistical sampling. The model-based statistical sampling methodology described in this paper relies on certain simplifying approximations and is not intended for audit populations with low error rates or for applications involving very small sample sizes. But in typical managerial applications, this methodology can produce more objective and reliable management information in a more systematic fashion and at a lower cost than conventional statistical and accounting methods. The approach generalizes Newman [1976].

MANAGEMENT INFORMATION FOR ALLOCATING THE COST OF SERVICES

Model-based statistical sampling can be described in the context of the cost-of-service allocation problem. Statistical projects are often undertaken to produce managerial information for allocating the cost of a central service department to a number of consuming departments, to be called cost centers in this paper. The cost-of-service allocation problem is discussed in many managerial accounting texts [e.g., Dopuch, et al., 1974, pp. 579-590; Horngren, 1977, pp. 524-529]. Thomas [1974] gives a comprehensive analysis of allocation from a financial accounting viewpoint.

The goal of effective cost-of-service allocation is to distribute the relevant costs that are incurred by the service center (the cost pool) to the individual cost centers in proportion to the actual benefit received by each cost center. In the situations of interest, the major difficulty is that there are a large number of cost centers receiving benefits from the service center and, moreover, the actual benefit received by each cost center can be accurately assessed only at a considerable expense.

In these circumstances, the cost-benefit principle of managerial accounting is often invoked to justify an allocation procedure that uses a readily available base as a proxy for the accurate assessment of benefit. This is justified if the base is highly correlated with the benefit. But the validity of this assumption typically involves a highly subjective judgment concerning the homogeneity of cost centers, or more accurately, the homogeneity of the relationship between base and benefit throughout the set of cost centers.

The preceding discussion suggests two approaches to cost allocation. Approach A follows the course of directly assessing the benefit received by each cost center. This approach produces allocations that are highly equitable and informative, but the expense of assessing benefits is likely to be prohibitive. Approach B eliminates the assessment expense by substituting a readily available base, but yields allocations that may be regarded as subjective, biased, and uninformative. Model-based statistical sampling provides a third approach that combines the objectivity of Approach A with the low cost of Approach B.

WHAT IS MODEL-BASED STATISTICAL SAMPLING?

The model-based statistical sampling approach builds on statistical sampling for data collection and multiple regression for data analysis. In the data collection stage, benefits are directly and accurately assessed for a relatively small number of cost centers that are selected following a carefully designed sampling plan. In the data analysis stage, the directly assessed benefits are related through regression analysis to auxiliary information describing the cost centers. Then the estimated regression relationship is used to objectively estimate the benefits received by the cost centers remaining outside of the sample. These estimated benefits are regarded as being attached to each cost center, and they can be accumulated by any function, costing objective, or classification of interest. A statistical error limit can be provided for any of these estimates of aggregate benefit. The sampling plan can be tailored to yield an acceptably small expected error limit for any specific set of functions.

A common impression is that statistical sampling is not appropriate unless the cost centers are homogeneous in some sense. While this may be true for ordinary statistical sampling, model-based statistical sampling turns lack of homogeneity to an advantage. By formulating a sampling plan that is based on the heterogeneity that is believed to exist among cost centers, the benefit assessments are directed toward the cost centers that are most relevant to the purpose of the study. This can yield a substantial savings in assessment expenditure, often exceeding 50%.

Even greater savings can usually be realized by taking advantage of auxiliary information in the analysis stage. In most applications it is not difficult to identify readily available auxiliary information describing each cost center that would be useful in predicting the benefit received by the cost center. This auxiliary information can be brought into the analysis through multiple regression. The relevance of this auxiliary information is measured by the coefficient of determination, R^2, between the benefit received by the function of interest at each cost center and the auxiliary information. The savings due to using the auxiliary information are directly related to R^2. For example, if the auxiliary information explains 80% of the variation in the benefit, then the required sample size will be reduced by 80%.

To summarize, model-based statistical sampling offers an objective and practical way of estimating benefits for cost-of-service allocation studies. Project costs are minimized by effectively combining auxiliary information with the direct assessment of benefit for a few suitably selected cost centers. Objectivity is guaranteed by selecting these cost centers on a statistical sampling basis, and by using estimation procedures based on multiple regression analysis.

The next two sections of this paper outline the model-based sampling methodology, first for data collection and then for data analysis. The following section discusses a class of cost-of-service applications underway in public utility load research. This section also provides a numerical example.

2. Planning the Data Collection Stage

The success of the data collection stage of a cost-of-service study depends on three factors:

*** Appropriate selection of cost centers to be included in the sample,

*** A suitable technique for measuring the benefit received by individual cost centers, and

*** Effective control of quality throughout the data collection activity.

Each of these three is crucial to the success of the project and must be given close attention by the project's management. However, this paper will focus on the selection of the sample. In a very literal sense, the bottom line of our discussion will be the number of cost centers to be included in the sample.

To make progress, the cost-of-service allocation problems of interest must be described rather precisely. Service is provided by a central service department or organization to each of a large number N of cost centers. Any particular cost center i receives a benefit that is quantified as y_i, which is not routinely recorded but can be accurately measured on a sampling basis. Interest is in the estimation of the aggregate benefit received by a function or cost objective; this aggregate benefit is assumed to be additive across cost centers so that it can be written as $\sum_{i=1}^{N} a_i y_i$. Here a_i is determined by the function of interest, and is assumed to be known for all cost centers. If the function is composed of a subset of the cost centers, then a_i is the indicator variable of the subset, i.e., $a_i = 1$ if i is included in the function, and $a_i = 0$ otherwise. More generally, a_i may represent the fraction of the benefit of cost center i that is received by the function of interest.

A project is to be undertaken in which the service benefit will be directly measured for each of n cost centers included in a statistical sample, denoted by s. The data collection part of the project involves selecting the sample s, and assessing the benefit y_i for each cost center i included in the sample, i.e., for each $i \varepsilon s$.

The output of the data collection stage is the sample database. Each of the n records of the sample database describes a particular cost center in the sample. Each record stores a number of variables or pieces of information describing the cost center, namely:

a) Information identifying the cost center, denoted by i,

b) The assessed benefit received by the cost center, denoted by the variable y_i, and

c) Any additional auxiliary information about the cost center that is readily available and believed to be relevant to the benefit received. This auxiliary information is represented by the k variables $x_{1i}\ x_{2i} \dots x_{ki}$.

THE MODEL

The basis for planning the data collection stage is past experience concerning the nature of the relationship between the benefit and the auxiliary information. This relationship can be conveniently and effectively formulated using a regression model which is denoted by the symbol M and is comprised of the following assumptions:

A) The expected benefit received by each cost center i, denoted by $E_M(y_i)$, is a linear function of the auxiliary information:

$$E_M(y_i) = \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}. \tag{1}$$

B) The actual benefit received is equal to the expected benefit plus a randomly distributed residual component u_i:

$$y_i = E_M(y_i) + u_i. \tag{2}$$

C) The standard deviation of the residual component of the benefit of each cost center i is known from past experience and is denoted σ_i.

D) The residual components of the N cost centers are mutually independent.

A few observations about each assumption are noteworthy:

a) The model assumes a sort of homogeneity of expected benefit among all cost centers. Equation (1) implies that a unit increase in the quantity x_{1i} will increase the expected benefit by β_1 for any cost center. In other words, the expected benefit is directly proportional to x_{1i} if $x_{2i}, \ldots, x_{ki}$ are held fixed. The accuracy of this assumption can often be increased by suitably transforming the original auxiliary information. For example, suppose that the original auxiliary information is comprised of a conventional allocation base x_i and a classification of cost centers into two groups, Group 1 and Group 2. Suppose also that the expected benefit received by each cost center is thought to be proportional to the base x_i, but with different constants of proportionality β_1 and β_2 within the two groups:

$$E_M(y_i) = \beta_1 x_i \text{ if i is in Group 1, and}$$

$$E_M(y_i) = \beta_2 x_i \text{ if i is in Group 2.}$$

In this example, it may not seem possible to combine the two groups of cost centers without violating Assumption A. However, define the two variables x_{1i} and x_{2i} as follows:

$$\begin{aligned} x_{1i} &= x_i \text{ if i is Group 1, } x_{1i} = 0 \text{ if i is Group 2, and} \\ x_{2i} &= 0 \text{ if i is Group 1, } x_{2i} = x_i \text{ if i is Group 2.} \end{aligned} \qquad (3)$$

Then $E_M(y_i) = \beta_1 x_{1i} + \beta_2 x_{2i}$ as required by (1). This device can be used very generally to combine alternative bases and classifications. Other transformations can be introduced to adapt to other features that arise in particular applications. In particular, a constant or intercept can be included in (1) simply by defining $x_{1i} = 1$.

b) The residual component u_i represents the composite effect of the variety of additional factors influencing the benefit received by a particular cost center. The residual component is sometimes called an error component but this terminology is somewhat misleading in these applications. The expected value of each residual component is necessarily zero as a result of (B).

c) The standard deviation of the residual component, σ_i, is allowed to vary between cost centers. This flexibility is necessary to deal with any residual heterogeneity between cost centers. In practice, past experience is used to estimate a fairly simple model relating σ_i to a suitable base or classification. This aspect of the planning can become rather technical [e.g., Chattergee and Price, 1977, pp. 101-114; Harvey, 1976]; it will not be emphasized in this paper. To the extent that the σ_i do vary between cost centers, the model M violates the homoscedasticity assumption of ordinary regression, i.e., M is heteroscedastic. Model-based statistical sampling turns this heteroscedasticity into an advantage both in data collection and in the data analysis.

d) If common factors significantly influence the residual components of the benefits received by several cost centers, assumption D will be violated. While assumption D can be relaxed, the price is a substantial increase in the complexity of the methodology. If reasonable efforts are made to include common factors in the auxiliary information factored into $E_M(y_i)$, then assumption D may be sufficiently accurate for our purpose.

THE SAMPLING PLAN

The sampling plan specifies how the cost centers are to be selected for direct assessment of benefit. The sampling plan specifies both the number n of cost centers to be included in the sample, and also the procedure for selecting them. Conventional statistical methodology emphasizes simple random sampling procedures that give each of the N cost centers the same probability of being included in the sample. Simple random sampling is the best plan only if units are homogeneous; otherwise model-based statistical sampling provides a better sampling plan.

Two sources of heterogeneity affect the efficiency of a sampling plan. One factor is the heteroscedasticity of the model M, that is, the variation of the residual standard deviations. The second factor is variation in the fraction a_i of the benefit of each cost center received by the function of interest. We define the relevance of cost center i to be the quantity $a_i\sigma_i$. Simple random sampling is best only if all N cost centers are homogeneous in the sense of being equally relevant.

In this context, a sampling plan is said to be best for a particular function if the sampling plan yields the most accurate estimate of the aggregate benefit of the function with the minimal number of direct assessments. This definition implicitly assumes that the expense of assessing benefit is equal for all cost centers. This assumption will be maintained throughout the paper.

When cost centers are heterogeneous in terms of their relevance, a varying probability sampling plan will be better than simple random sampling. Audit sampling often employs a particular varying probability sampling plan called dollar unit sampling. Under dollar unit sampling, accounts are selected for auditing with probability proportional to their dollar balance. We consider a generalization of dollar unit sampling in which cost centers are selected for benefit-assessment with probability proportional to a prechosen quantity P_i. For our purpose, the choice of the sampling plan can be identified with the choice of P_i for all N cost centers.

The principal basis for choosing P_i is the statistical reliability of the resulting estimates. Assume that the model M holds, and that the sample data will be used to estimate the aggregate benefit received by a function characterized by a_i, following the generalized regression procedure described in the next section. Then the expected standard error of this estimate is

$$se = \frac{N}{\sqrt{n}} \sqrt{\text{mean}(a^2\sigma^2/w) - (n/N)\text{mean}(a^2\sigma^2)}. \tag{4}$$

Some new notation has been introduced in (4). We use "mean" to denote an average calculated over all N cost centers. For example,

$$\text{mean}(a^2\sigma^2) = N^{-1} \sum_{i=1}^{N} a_i^2 \sigma_i^2 .$$

Moreover, the sampling plan is described in terms of w_i where

$$w_i = P_i/\text{mean}(P).$$

The quantity $\pm$ 2 se can be used in planning as an expected error limit for the estimate of the aggregate benefit. This assumes a 95% level of confidence. The derivation of (4) involves simplifying asymptotic approximations, so that (4) should not be used in applications with very small sample sizes or audit populations with low error rates [e.g., Beck, 1980; Garstka and Ohlson, 1979]. In specific applications, the accuracy of (4) can be checked through computer simulation. Work in this direction is underway. The mathematical details of (4) and related results are available in Sarndal [1980] and Wright [1981].

Equation (4) provides qualitative insights that are useful for planning. As is usual in sampling, (4) shows that the standard error increases in proportion to the total number N of cost centers, and decreases in proportion to the square root of the sample size. The term $(n/N)\text{mean}(a^2\sigma^2)$ generalizes the conventional finite population correction factor and is often negligible.

The remaining term in (4), mean($a^2\sigma^2/w$), is often the key factor in the standard error. This term reflects the interaction of the function of interest, the residual standard deviations assumed in the model M, and the inclusion probabilities of the sampling plan.

Equation (4) shows that the standard error can be decreased in three ways:

1) Increasing the sample size,

2) Bringing in additional auxiliary information to reduce the residual standard deviations, and

3) Making a more suitable choice of the P_i (or w_i) used in the sampling plan.

The first option, increasing the sample size, directly increases the expense of data collection. For example, to reduce the standard error by 50%, the number of cost centers to be assessed must be increased by 300%. So relying on an increased sample size for reliable estimation can have a disasterous impact on the budget of the project.

Fortunately, the remaining two options offer improved statistical reliability at virtually no added expense. Alternatively, these options can be used to reduce the sample size that is required for any given error limit.

Consider the second option--bringing in auxiliary information. The analysis is easiest if simple random sampling is assumed and the finite population correction factor is neglected. Define the coefficient of determination of the auxiliary information for the function of interest to be:

$R^2 = 1 - \text{mean}(a^2\sigma^2)/\text{var}(ay)$, with

$\text{var}(ay) = \text{mean}(a^2y^2) - \text{mean}(ay)^2$.

Under the simplifying assumptions, R^2 is equal to the reduction in the required sample size due to the use of the auxiliary information. For example, if the auxiliary information explains 80% of the variation in the benefit received by the function of interest, then the use of the auxiliary information reduces the required sample size by 80%.

Like all good things, the use of auxiliary information can be overdone. Equation (4) assumes that the sample size is substantially larger than the number (k) of auxiliary variables in the model M. The approximations behind (4) will break down if the model is too complex for the sample, or more generally, if there is strong multicollinearity in the sample database. These problems can be avoided if care is taken in specifying the auxiliary information included in the model.

Now consider the third option for decreasing the standard error or the required sample size--choosing the P_i of the sampling plan. An important principal of model-based sampling is that the best sampling plan for a particular function is to select cost centers with probability proportional to their relevance for the function. So the best sampling plan uses

$$P_i = a_i\sigma_i. \qquad (5)$$

There are often advantages to choosing P_i that are not best in that they violate (5). The efficiency of any such suboptimal sampling plan is defined to be the ratio n_2/n_1, where n_1 is the sample size required to achieve a certain standard error using the suboptimal sampling plan, and n_2 is the sample size required to achieve the same standard error with the best sampling plan. The efficiency of any plan using P_i or w_i can be calculated as

$$\text{eff} = \text{mean}(a\sigma)^2/\text{mean}(a^2\sigma^2/w). \qquad (6)$$

Stratification can be regarded as a technique for obtaining reasonably efficient approximations to the best sampling plan. Suppose relevance is constant within all strata. In this case, (5) gives the optimal stratified sampling plan following Neyman allocation. More commonly, relevance will vary almost continuously. Then the model-based approach can be used to design stratified sampling plans that are highly efficient and may be easier to implement than a sampling plan following (5).

These methods make sampling straightforward as long as there is a single principal function of interest. If several functions are important, then additional analysis may be required to identify a sampling plan that efficiently achieves the multiple objectives of the study.

CHOOSING THE SAMPLE SIZE

For any given function a_i and model M, equations (4)-(6) can be adapted to calculate the sample size that is required to achieve a prescribed expected error limit, say c. The most convenient approach is to follow three steps:

1) Determine the sample size n_0 required if simple random sampling is followed and the finite population correction factor is neglected:

$$n_0 = (2N/c)^2 \text{mean}(a^2\sigma^2). \qquad (7)$$

2) Determine the sample size n_1 required if simple random sampling is followed and the finite population correction factor is included:

$$n_1 = n_0/(1 + n_0/N). \qquad (8)$$

3) Determine the sample size n_2 required if the best sampling plan is followed and the finite population correction factor is included:

$$n_2 = n_1 \text{ mean}(a\sigma)^2/\text{mean}(a^2\sigma^2). \qquad (9)$$

These equations have been formulated for an expected error limit c equal to ± 2 se, i.e., an error limit at the 95% level of confidence. If another confidence level is desired, equation (7) should be modified by replacing 2N by zN where z is determined from a table of the standard normal distribution, e.g., z = 1.645 for 90% confidence or z = 2.576 for 99% confidence.

The role of heterogeneity in model-based sampling can be easily seen by rewriting (9) as $n_2 = n_1(1 + cv^2)^{-1}$. Here cv is the coefficient of variation of the relevance of the cost centers. This means that the best model-based sampling plan will be advantageous to the extent that the cost centers are heterogeneous in terms of their relevance for the function of interest. If, as is often the case, the coefficient of variation of relevance exceeds one, then the best model-based sampling plan will reduce the sample size by more than 50% compared to simple random sampling.

3. Generalized Regression Procedures For Data Analysis

When the sample database is completely assembled as described in the previous section, then the cost-of-service project enters the data analysis stage. Under model-based statistical sampling, the data analysis procedures are organized around a generalization of multiple linear regression introduced by Cassel, et al., [1977].

Multiple linear regression analysis is commonly used in managerial accounting to estimate cost behavior patterns from past experience. Good introductions are provided by Dopuch, et al., [1974, pp. 62-88], Horngren [1977, pp. 777-799], Johnston [1960], and Neter and Wasserman [1974]. The use of regression analysis in cost allocation applications is closely related to its use in cost estimation, but certain generalizations are needed to take account of the following features of our setup:

*** The heteroscedasticity of the model M,

*** The varying probability sampling plan used to collect the data, and

*** The interest in estimating the aggregate benefit received by one or more functions throughout the N cost centers.

The approach followed in the analysis will be natural to anyone familiar with multiple regression:

Step 1: Use the sample data to estimate the regression coefficients of the model M, i.e., to estimate the parameters $\beta_1, \beta_2, \ldots, \beta_k$ in (1).

Step 2: Use the estimated regression coefficients to estimate the benefit received by each of the N cost centers.

Step 3: Calculate the aggregate estimated benefit for any function or classification of interest.

ESTIMATING THE REGRESSION COEFFICIENTS

Since the model M has heterogeneous residual standard deviations, i.e., M is heteroscedastic, the sample data should be analyzed following an adaptation of ordinary regression analysis called model-based weighted-least-squares (WLS), [Neter and Wasserman, 1974, pp. 131-136; Maddala, 1977, pp. 259-268]. The model-based WLS procedure can be implemented by transforming the sample database and then using ordinary regression analysis to estimate the regression coefficients in the usual fashion.

To describe the procedure, it is convenient to rewrite the residual standard deviation σ_i as $\sigma_0 z_i$ where z_i is a known variable describing each cost center. Assumption C of the model M can be modified slightly at this stage by assuming that the parameter σ_0 is unknown and is to be estimated from the sample data. The variable z_i can be regarded as a base representing the collective magnitude of the various factors affecting the residual component of benefit; often z_i is taken to be a measure of the size of the cost center. As previously suggested, specification analysis procedures are available to evaluate alternative choices of z_i, but these will be discussed elsewhere.

The model-based WLS procedure is implemented by applying a simple division transformation to each variable in the analysis database:

$$\begin{aligned} y_i{*} &= y_i/z_i \\ x_{1i}{*} &= x_{1i}/z_i \\ x_{2i}{*} &= x_{2i}/z_i \qquad (10) \\ &\vdots \\ x_{ki}{*} &= x_{ki}/z_i. \end{aligned}$$

Then an ordinary regression procedure is followed to calculate the estimated multiple regression coefficients relating $y_i{*}$ to $x_{1i}{*}$, $x_{2i}{*}$, ..., $x_{ki}{*}$ using a zero-intercept option. These estimated regression coefficients can be denoted by b_1, b_2, ..., b_k.

Assuming that the model M is realistic, the ordinary regression output from the model-based WLS procedure can be used in the usual fashion to calculate confidence intervals and to test hypotheses for the regression coefficients of the regression equation (1). These and other specification analysis techniques can be used to refine the model M based on the sample data. Moreover, the standard error of the regression can be used to estimate σ_0. However, the multiple correlation coefficient and sample coefficient of determination are generally misleading under WLS procedures and should not be reported.

While model-based WLS is generally accepted by statisticians as the most appropriate data analysis procedure as long as the model M is accurate, survey sampling statisticians have tended to prefer alternative procedures that might be called design-based WLS. A design-based WLS procedure is obtained by substituting a different weight, say q_i, for z_i in (10). The weight q_i, which must be nonnegative, is determined from the sampling plan; often $q_i = \sqrt{P_i}$ or $q_i = \sqrt{w_i}$ [Sarndal, 1980].

Design-based WLS has both disadvantages and advantages. The principal disadvantage is that the usual confidence intervals and measures of significance obtained from the regression output are usually biased and should not be used for inference. The principal advantage is that with a suitable choice of q_i, design-based WLS will often yield comparatively simple and intuitively reasonable data analysis procedures in line with conventional sampling practice [e.g., Cochran, 1977].

The principles of data colection of Section 2 are valid for both model-based and design-based WLS data analysis procedures. In some circumstance, it may be advantageous to follow a hybrid strategy that combines model-based data collection with design-based WLS data analysis.

ESTIMATING BENEFITS

Once the estimated regression coefficients have been calculated through a suitable WLS procedure, the benefit received by each cost center and by any function of interest can be estimated from the auxiliary information describing each cost center. This step depends on the availability of this information for all cost centers, especially those not included in the sample.

Recall that y_i denotes the actual benefit (known or unknown) received by cost center i, and $\sum_{i=1}^{N} a_i y_i$ denotes the aggregate benefit received by the function of interest. An estimate of the benefit of cost center i will be denoted by $\hat{y}_i$; the corresponding estimate of the aggregate benefit can be calculated as $\sum_{i=1}^{N} a_i \hat{y}_i$ since the a_i are assumed to be known.

Three different procedures for estimating the benefit received by each cost center must be distinguished; the resulting estimates will be denoted by $\hat{y}_{1i}$, $\hat{y}_{2i}$, and $\hat{y}_{3i}$.

1) The conventional procedure uses the estimated regression equation to calculate estimates for all cost centers:

$$\hat{y}_{1i} = b_1x_{1i} + b_2x_{2i} + \dots + b_kx_{ki}. \qquad (11)$$

2) The second procedure adjusts the conventional estimates for the actual benefit directly assessed for the cost centers included in the sample s:

$$\hat{y}_{2i} = y_i \text{ if } i\varepsilon s\text{, and}$$
$$\hat{y}_{2i} = \hat{y}_{1i} \text{ if } i\not\varepsilon s. \qquad (12)$$

3) The generalized regression procedure adjusts the conventional estimates for the estimated residual components observed in the sample:

$$\hat{y}_{3i} = \hat{y}_{1i} + (N\hat{u}_i)/(nw_i) \text{ if } i\varepsilon s\text{, and}$$
$$\hat{y}_{3i} = \hat{y}_{1i} \text{ if } i\not\varepsilon s. \qquad (13)$$

Here the estimated residual component is

$$\hat{u}_i = y_i - \hat{y}_{1i} \text{ for } i\varepsilon s.$$

Which of these alternative procedures should be chosen? The answer depends upon a rather subtle interplay between the purpose at hand and the credibility of the model M. If the principal purpose is to estimate the benefit received by the individual cost centers, then either the first or second procedure should be followed.

The issue is more complicated if the principal concern is with the aggregate benefit received by a function involving a number of cost centers. In this case, the choice depends on the credibility of the model M. If M is accurate, then the second procedure generally provides the most reliable estimate of aggregate benefit. However, the estimates provided by either procedure one or two may be seriously biased if M is even slightly wrong. At the cost of somewhat poorer reliability if M is accurate, the generalized regression procedure gives protection against this bias. As long as the sample size is reasonably large, the generalized regression estimate is approximately unbiased from a survey sampling viewpoint regardless of the validity of the model M. So (13) will usually provide a conservative choice, and one that is in line with more conventional survey sampling practice.

The accuracy of the model M is important in another way. The sampling procedures discussed in the previous section assume that M is reasonably accurate. If M is erroneous, the expected error limits may be biased but the extent of this bias is not well understood at this time. These and related issues need further investigation.

Despite these limitations, model-based statistical sampling seems to provide the best approach to cost-of-service studies. The methodology is organized around a central model M. This model may be rather simple if little past experience is available. M may be systematically refined as experience is accumulated. The model M determines data collection and estimation procedures that are highly efficient if M is accurate, but remain free of significant bias even if M is somewhat wrong.

4. Public Utility Load Research

Many electric utility companies engage in load research studies which investigate the consumption of electricity by time of day within various classes of their customers. Load research serves a variety of purposes involving rate design, system operation, and planning. Under the Public Utility Regulatory Policies Act of 1978 (PURPA), all large public utilities are required to begin collecting load research data on a statistical sampling basis. Some references are Aigner [1978], Brandenburg and Higgins [1974], and Taylor [1977].

Model-based statistical sampling is an ideal methodology for load research. In fact, vital support for the development of the methodology has come from the Rate Research Department of Consumers Power Company. This group routinely uses model-based sampling procedures to plan its load research studies [e.g., Load Research Committee Report, 1980, pp. 66-97].

One important purpose of load research is to provide data for cost-of-service allocations. In the load research application, the service department is taken to be the entire utility company, and the cost centers are its customers.

Most utility cost accounting systems are organized around three primary cost pools: (1) fixed costs, (2) costs that are thought to vary by the total amount of energy produced and distributed throughout the year (called energy costs), and (3) costs that vary by the amount of system capacity that is maintained to meet peak usage during the year (demand costs). Costs are also classified as generation, transmission and distribution, and the distribution costs are subclassified according to the voltage level.

The two variable cost pools are allocated among customers in proportion to the estimated energy-related or demand-related benefit received by each customer during the year. The energy benefit received by each customer i is considered to be proportional to the customer's total consumption of electricity during the year, say x_i (called usage). Since usage is usually directly metered, no significant estimation problem is involved in allocating energy costs, although the voltage level provided to the customer may introduce some rate differentials.

The allocation of demand costs is another matter. For an individual customer, the demand-related benefit can be directly assessed through the use of a time-of-day meter which records the customer's consumption of electricity almost continuously--often for each fifteen minute interval throughout the year. The customer's demand-related benefit, or simply demand, y_i, is usually taken to be his consumption of electricity during one or more hours of peak system-wide consumption. This peak period itself is usually regarded as fixed. The goal is to allocate the demand-related cost pool in proportion to the demand y_i of each customer.

The problem is that time-of-day metering is far too expensive to be used for all customers--averaging \$400-\$500 per customer per year. So load research projects are undertaken to estimate demand on a sampling basis.

The effectiveness of the model-based sampling approach depends on the availability of auxiliary information that is highly correlated with demand y_i. In load research cost allocation projects, two sorts of auxiliary information are usually used: (1) usage, x_i, and (2) a classification of customers into k rate groups. A model M can be formed that combines these two sources of information by extending (3) to k groups. With this model, the regression coefficients β_1, β_2, ..., β_k represent demand/usage ratios within each of the k rate groups. These ratios are closely related to the parameters that utility engineers call load factors.

A central feature of the model M is its residual standard deviations. In our load research work we have estimated the residual standard deviations from past load research data using the assumed relationship $\sigma_i = \sigma_0 x_i^{\gamma}$. The parameters σ_0 and γ are allowed to vary between different rate groups but are assumed to be constant for all customers within each rate group. The parameter γ is introduced to integrate sampling theory, in which γ is often assumed to equal .5, and empirical experience which suggests γ closer to one. Both parameters can be estimated for each rate group from available data using an adaptation of Harvey [1976].

In load research, we are interested in k different functions, one function identifying the customers included in each rate group of the study. Let G denote a particular rate group. For the corresponding function, $a_i = 1$ if $i \varepsilon G$ and $a_i = 0$ otherwise. The aggregate demand-related benefit of this function is the total demand within rate group G.

The sampling plan of a load research project is usually designed to yield reliable estimates of the total demand within each of the k rate groups, or equivalently, of the aggregate benefit of each of the k functions of interest. Although this sounds like a multiple-objective planning problem, the theory of model-based sampling leads to a natural separation of the project by rate group. The only customers relevant to the function associated with rate group G are the customers in G itself, so the best sampling plan for this function would select a subsample exclusively from G. Because of the construction of the model M, the WLS estimation procedure also separates into independent subprocedures for each rate group. This has many practical advantages for planning and implementation.

In particular, the relevant model for rate group G is simply

$$E_M(y_i) = \beta x_i, \text{ and}$$
$$\sigma_i = \sigma_0 x_i^{\gamma} \text{ for } i \varepsilon G. \qquad (14)$$

This model, denoted by M_G, is the ratio model that arises in many other applications as well. Here β, σ_0, and γ are parameters identified with G, and $a_i = 1$ throughout G.

Before the composite model M is completely forgotten, we should note its relevance to questions regarding the definition of rate groups. If the regression coefficients of two existing groups are not materially different, it may be desirable to simplify the rate structure by merging the two groups. Conversely, if two subclasses have substantially different coefficients, they may be recognized as groups for the sake of equity. Statistical significance tests, developed within the context of M, can help to evaluate differences between coefficients. More generally, the definition of rate groups can be usefully regarded as a special case of the variable selection problem in regression analysis. This approach extends a suggestion of Demski and Feltham [1976, p. 131] for dealing with aggregation in cost determination.

THE BRADENBURG-HIGGINS EXAMPLE

A numerical illustration of the model-based approach can be developed from a dataset that Brandenburg and Higgins [1974] have used to demonstrate conventional sample design in load research. The dataset provides demand y_i (in kw) and monthly usage (in mwh) for each of $n = 210$ commercial or industrial customers. We assume that this is a sample database collected under a simple random sampling plan from a single rate group G of $N = 840$ customers. The model M_G is assumed to be given by (14).

The illustration will be presented in two parts:

1) Data analysis of this sample using WLS to estimate the parameters of M_G, and using generalized regression to estimate total demand within the rate group, and

2) Developing a sampling plan for a future load research study of this rate group, using the model M_G estimated in part one.

DATA ANALYSIS

The first step in model-based data analysis is to use the sample database to estimate the parameters σ_0 and γ of (14). The estimated relationship is found to be

$$\sigma_i = .9223\ x_i^{.9832}. \tag{15}$$

Then, using $z_i = x_i^{.9832}$ in (10), the model-based WLS procedure gives the estimated regression equation

$$\hat{y}_{1i} = 2.737\ x_i. \tag{16}$$

This result can be used to calculate the estimated residual component $\hat{u}_i$ for each sample customer i, and also the sample mean of the estimated residual components:

$$\begin{aligned} \text{mean}_s(\hat{u}) &= n^{-1} \sum_{i\varepsilon s} \hat{u}_i \\ &= -592.6 \text{ kw.} \end{aligned}$$

The next step is to utilize the distribution of x_i throughout the entire rate group of N customers. This distribution would ordinarily be readily available, but unfortunately it was not published for this example. So reasonable assumptions will be made for required summary statistics based on the sample database. In particular it will be assumed that

$$\begin{aligned} \text{mean}(x) &= N^{-1} \sum_{i=1}^{N} x_i \\ &= 1690 \text{ mwh.} \end{aligned}$$

The total demand within G might be estimated following any one of the three procedures discussed in Section 3:

$$1)\quad \sum_{i=1}^{N} \hat{y}_{1i} = 2.737 \sum_{i=1}^{N} x_i$$

$$= 2.737\ N\ \mathrm{mean}(x)$$

$$= (2.737)(840)(1690)$$

$$= 3{,}885{,}445 \text{ kw.}$$

$$2)\quad \sum_{i=1}^{N} \hat{y}_{2i} = \sum_{i=1}^{N} \hat{y}_{1i} + \sum_{i\varepsilon s} (y_i - \hat{y}_{1i})$$

$$= \sum_{i=1}^{N} \hat{y}_{1i} + n\ \mathrm{mean}_s(\hat{u})$$

$$= 3{,}885{,}445 - (210)(592.6)$$

$$= 3{,}760{,}999 \text{ kw.}$$

$$3)\quad \sum_{i=1}^{N} \hat{y}_{3i} = \sum_{i=1}^{N} \hat{y}_{1i} + (N/n) \sum_{i\varepsilon s} \hat{u}_i$$

$$= \sum_{i=1}^{N} \hat{y}_{1i} + N\ \mathrm{mean}_s(\hat{u})$$

$$= 3{,}885{,}445 - (840)(592.6)$$

$$= 3{,}387{,}661 \text{ kw.} \qquad (17)$$

Which of these three estimates is to be preferred? If the model M_G, (14), is exact, the second estimate is probably most accurate. Experience suggests that (14) is a good approximation for many purposes, but it may be slightly erroneous. For example, demand may be related to usage in a slightly curvilinear fashion. Even small errors in M_G may lead to substantial bias in the first two estimates so they are risky.

The third, generalized regression estimate (17) includes a residual correction for this potential bias so that it is likely to be preferred over the first and second estimates.

An alternative approach is to follow a design-based WLS procedure. For the ratio model, the usual choice of q_i is $q_i = (x_i w_i)^{1/2}$, giving

$$b = \sum_{i \varepsilon s} w_i^{-1} y_i \Big/ \sum_{i \varepsilon s} w_i^{-1} x_i .$$

With simple random sampling as in this case, $w_i = 1$ and b becomes the simple ratio estimator:

$$\begin{aligned} b &= \text{mean}_s(y)/\text{mean}_s(x) \\ &= 3757/1589 \\ &= 2.364 \text{ kw/mwh.} \end{aligned}$$

With this value of b, $\text{mean}_s(\hat{u}) = 0$ so that all three procedures give the same estimate of the total demand within G:

$$\begin{aligned} \sum_{i=1}^{N} \hat{y}_i &= b \sum_{i=1}^{N} x_i \\ &= (2.364)(840)(1690) \\ &= 3,355,934 \text{ kw.} \end{aligned}$$

This estimate may be favored because of its simplicity.

DEVELOPING A SAMPLING PLAN

In planning a new study, it is often worthwhile to pool data from several past studies to determine long run averages and perhaps trends or other changes in model parameters. However, to simplify the discussion, the one sample database will be used as the sole basis for planning. The key inputs to the planning process are the residual standard deviations determined by (15) together with the distribution of x_i throughout the rate group. These are used to calculate the statistics:

$$\text{mean}(\sigma) = N^{-1} \sum_{i=1}^{N} \sigma_i$$

$$= 1362 \text{ kw},$$

$$\text{mean}(\sigma^2)^{1/2} = (N^{-1} \sum_{i=1}^{N} \sigma_i^2)^{1/2}$$

$$= 2475 \text{ kw}.$$

These statistics can be used with the desired error limit to determine the sample sizes required under alternative sampling procedures. Following PURPA, the error limit c is taken to be ten percent of the estimated total demand, or c = 338,766 kw using (17). PURPA also specifies 90% level of confidence. Then, following (7)-(9):

$$n_0 = [(1.645)(840)(2475)/338766]^2$$

$$= 102 \text{ customers},$$

$$n_1 = 102/(1 + 102/840)$$

$$= 91 \text{ customers},$$

$$n_2 = 91(1362/2475)^2$$

$$= 28 \text{ customers}.$$

So if a simple random sampling plan is followed, about 91 customers will be required in the new load study, but if the best model-based sampling plan is followed, this is reduced to about 28 customers.

Under the best sampling plan, customers are selected with probability proportional to σ_i, given by (15). Since γ is so close to one, a reasonable simplification is a sampling plan which select customers with probability proportional to their size (PPS) as measured by x_i. The efficiency of the PPS sampling plan can be calculated using (6) together with the statistic:

$$\text{mean}(\sigma^2/w)^{1/2} = 1363 \text{ kw.}$$

In fact,

$$\begin{aligned} \text{eff} &= \text{mean}(\sigma)^2/\text{mean}(\sigma^2/w) \\ &= (1362/1363)^2 \\ &= 99.9\%. \end{aligned}$$

So the PPS design is virtually equivalent to the best plan in this case.

Another convenient approach is to use a stratified sampling plan to approximate the best sampling plan. In this example, a stratified sampling plan can be developed which achieves 95% efficiency with only six strata comprised of customers having approximately equal relevance.

CONCLUSION

Model-based statistical sampling has proven to be highly effective in rate research. Specific practical sampling plans can be easily developed based on the circumstances of each project. In the simplest applications, these sampling plans are closely related to conventional stratified regression and ratio procedures. Even in these circumstances, model-based sampling offers important advantages over present methods, especially in handling various forms of heteroscedasticity. In other applications, multivariate auxiliary information is available, including multiple bases for allocation and multiple classifications. Model-based statistical sampling can take advantage of this multivariate auxiliary information much more effectively than conventional methods. Moreover, the model-based approach provides a useful link between survey sampling and conventional regression analysis.

Although this paper has emphasized allocation applications, model-based statistical sampling is equally effective in most management information projects in which data can be efficiently collected on a sampling basis. Some other important areas of application are in determining physical inventory, in estimating the replacement cost of property and equipment, and in valuing loans and receivables.

REFERENCES

Aigner, D. J. "Bayesian Analysis of Optimal Sample Size and a Best Decision Rule for Experiments in Direct Load Control." Journal of Econometrics 9 (1979): 209-222.

Beck, P. A. "A Critical Analysis of the Regression Estimator in Audit Sampling." Journal of Accounting Research 18 (Spring 1980): 16-37.

Brandenburg, L. and C. E. Higgins, Jr. "Stratified Random Sampling Methods for Class Load Surveys for Electric Utilities," in Applied Statistics in Load Research, Vol. III. New York: Association of Edison Illuminating Companies, 1974. 234-284.

Cassel, C. M., C.E. Sarndal, and J. H. Wretman. Foundations of Inference in Survey Sampling. New York: Wiley, 1977.

Chatterjee, S. and B. Price. Regression Analysis by Example. New York: Wiley, 1977.

Cochran, W. G. Sampling Techniques Third Edition. New York: Wiley, 1977.

Demski, J. and G. Feltham. Cost Determination: A Conceptual Approach. Ames, Iowa: Iowa State University Press, 1976.

Dopuch, N., J. G. Birnberg, and J. Demski. Cost Accounting Data for Management's Decisions Second Edition. New York: Harcourt, Brace, Jovanovich, 1974.

Garstka, S. J. and P. A. Ohlson. "Ratio Estimation in Accounting Populations with Probabilities of Sample Selection Proportional to Size of Book Values." Journal of Accounting Research 17 (Spring 1979): 23-59.

Harvey, A. C. "Estimating Regression Models with Multiplicative Heteroscedasticity." Econometrica 44 (1976): 461-464.

Horngren, C. E. Cost Accounting: A Managerial Emphasis Fourth Edition. Englewood Cliffs, New Jersey: Prentice-Hall, 1977.

Johnston, O. Statistical Cost Analysis. McGraw-Hill Book Company, 1960.

Load Research Committee Report on Development of General Service Class Load Curves. New York: Association of Edison Illuminating Companies, 1980.

Maddala, G. S. Econometrics. New York: McGraw-Hill, 1977.

Neter, J. and W. Wasserman. Applied Linear Statistical Models. Homewood, Illinois: Irwin, 1974.

Newman, M. S. Financial Accounting Estimates Through Statistical Sampling by Computer. New York: Wiley, 1976.

Sarndal, C. E. "On π-Inverse Weighting Versus Best Linear Unbiased Weighting in Probability Sampling." Biometrika 67 (December 1980): 639-650.

Savage, L. J. The Foundations of Statistics. New York: Wiley, 1955.

Taylor, L. D. "On Modeling the Residential Demand for Electricity by Time-of-Day." In Forecasting and Modeling Time-of-Day and Seasonal Electricity Demand. Palo Alto: Electric Power Research Institute, 1977.

Thomas, A. L. The Allocation Problem: Part Two. Sarasota, Florida: American Accounting Association, 1974.

Wright, R. L. "Sample Design with Multivariate Auxiliary Information." Working Paper, School of Business Administration, The University of Michigan, 1981.

Zellner, A. An Introduction to Bayesian Inference in Econometrics. New York: Wiley, 1971.