A FINITE SELECTION MODEL FOR EXPERIMENTAL DESIGN OF THE HEALTH INSURANCE STUDY

Carl Morris, The Rand Corporation

1. INTRODUCTION

The nation currently is debating what form, if any, of national health insurance should be adopted. Evaluating alternative proposals requires the best knowledge concerning the consequences of alternative ways of financing medical care services. To increase current knowledge, the Department of Health, Education, and Welfare is sponsoring the Health Insurance Study (HIS) which is being designed and analyzed by The Rand Corporation and conducted with the help of several subcontractors. A summary description of the study design is provided in (Newhouse 1974).

The main part of this study involves collecting from three to five years of longitudinal data on each of a large number of families after assigning to them one of several kinds of health insurance plans (treatments). The finite selection model (FSM) has been developed primarily to determine which families should be assigned to which treatments, although it also is used for other purposes.

The FSM is both a set of concepts and a computer package for defining and then carrying out the required operations. The concepts are discussed below. They include formulas for assessing the worth of a particular assignment, and algorithms for finding optimal or near-optimal assignments. Good assignments must satisfy requirements of precision, balance, and robustness. Statistics needed to assess precision and balance are provided in Sections 2, 4, 5 when a linear model is specified. The main idea behind the optimality algorithm is given in Section 2, but it is modified and adapted to provide balanced and robust samples in more general situations in Sections 4, 6, 7.

The work here is an outgrowth of that of Conlisk and Watts (1969). Their "allocation model" has been used to find optimal sample sizes for several large-scale public policy experiments before the HIS. While their model dictates classical designs in situations where the inputs and needs (costs, precision, etc.) have sufficient symmetry, it has the desirable effect of dictating nonclassical designs in the asymmetrical situations usually encountered in complex experiments.

The Conlisk-Watts model, however, is not entirely satisfactory for assigning families to treatment groups in the presence of a substantial battery of preexperimental data collected for each family. Their model generates optimal *stratified* samples for these situations. Stratified samples have the disadvantage that (i) the number of preexperimental variables used must be severely limited, (ii) families with dissimilar values on a continuous variable, such as income, must be grouped into the same stratum, (iii) more families than available may be required from a given stratum (because their model samples from an infinite population).

The finite selection model is so named because it circumvents these difficulties by making *selections* for treatments from the *finite* population of subjects known to be available for assignment. Objection (iii) is thereby eliminated. Preexperimental measurements are permitted to be continuous, if desired, thereby eliminating objection (ii) and dealing with objection (i) by permitting a substantial increase in the number of preexperimental variables utilized.

We proceed in the next section to discuss the model, building up the concepts of optimality, balance and robustness stepwise in Sections 2-6. Discussion of the specific application to HIS is reserved for Section 7.

2. THE FINITE SELECTION MODEL AS AN OPTIMIZATION MODEL

As stated before, the main application of the FSM to the Health Insurance Study is to assign experimental subjects to treatment groups according to the general design goals of: optimality (i.e., precision); balance; and robustness. Rather than consider all these objectives at once, they will be considered separately in different sections. All these concepts derive from the definition and algorithm for optimality considered in this section.

Assume that k measurements on each of N experimental subjects have been obtained and are available in the form of an N x k matrix X_N. Associated with the i-th subject is a cost $c_i > 0$, i = 1(1)N. A budget constraint C is specified, and any subset $S \subset \{1, \ldots, N\}$ of subjects may be selected provided

$$\sum_{i \in S} c_i \leq C. \qquad (2.1)$$

Let $\mathcal{S}_C$ be the set of all subsets satisfying (2.1). The optimization problem is to choose the element in $\mathcal{S}_C$ which minimizes a given objective. In principle, since $\mathcal{S}_C$ is finite, this may be done by enumeration. That is, in a sense, the problem is solved as soon as an objective function is specified. However, for a problem of the magnitude of the HIS, a complete enumeration would take about 10^{500} years of computer time, so an algorithm also must be specified.

The objective function for this work borrows heavily from Conlisk and Watts, who used similar criteria in designing the negative income tax experiments, and also from others before them, especially (Kiefer 1959), (Elfving 1952) who studied the mathematical consequences of these criteria under the name "A-optimality." The reader is referred especially to (Conlisk-Watts 1969) for background material.

A response variable y will be measured for each subject during the experiment. It cannot be used to design the experiment, except: to suggest an appropriate scaling of the variables in X_N; and to choose the budget or number of subjects needed for specified precision. This is assumed to have been done.

If n subjects are available in the analysis, a regression equation of the form

$$y = X_n\beta + e, \; E(e) = 0, \; V(e) = \sigma^2 I \qquad (2.2)$$

is specified, X_n being an n x k matrix. Define

$$S_n = (X_n'X_n)^{-1}. \qquad (2.3)$$

Then the Gauss-Markov estimator b is the most efficient unbiased estimator of the unknown k-vector of regression coefficients β, where

$$b = S_n X_n' y \qquad V(b) = \sigma^2 S_n. \qquad (2.4)$$

Certain linear combinations of the regression coefficients are to be estimated. If there are p of them, then a known matrix P of dimensions p × k is specified and the vector $\alpha = P\beta$ is to be estimated. The dimension p may be either larger or smaller than k. The best unbiased estimate of α and its variance is

$$a = Pb, \qquad V(a) = \sigma^2 P S_n P'. \qquad (2.5)$$

Precise estimation of all the components of α would follow if the diagonal elements of V(a), the variances of the a_i, could be simultaneously minimized by appropriate choice of X_n. This being impossible, a real-valued functional on V(a) must be specified. Several choices have been considered in the literature (Kiefer 1959), one of the most popular being to minimize the generalized variance, or determinant, $|V(a)|$. The determinant criterion, being independent of P when $p \geq k$, is an inefficient choice for the HIS since specified linear combinations of β are of special interest. Instead we follow (Conlisk-Watts 1969) and minimize the weighted sum of variances

$$\varphi(X_n) \equiv \frac{1}{\sigma^2} \Sigma\, W_j\, \mathrm{Var}(a_j), \qquad (2.6)$$

with weights $(W_1, \ldots, W_p)$, $W_j \geq 0$ specifying the relative importance of estimating the α_j. The scalar σ^2, which is not assumed known, is immaterial to this criterion. Including its value in (2.6) makes the expression for φ independent of σ^2.

Denoting W as the p × p diagonal matrix having diagonal elements $(W_1, \ldots, W_p)$, and then defining $T = P'WP$, a known k × k matrix, (2.6) may be rewritten

$$\varphi(X_n) = \mathrm{tr}(TS_n) \qquad (2.7)$$

with tr, the trace function of a matrix, being the sum of diagonal elements of the matrix.

The first and simplest problem considered in this paper is the following:

Problem 1

Find the n rows of the N rows of X_N which minimize (2.7) among choices which have total cost (2.1) not exceeding the cost constraint C.

Problem 1 requires solution to a nonlinear integer programming problem. An optimization method analogous to the method of steepest descent is described below. This method can be proved to coverge in the continuous case, and always has given satisfactory results in discrete situations for which it is mainly intended. This convergence issue will be discussed more fully in Section 3. The algorithm is based on the matrix identity (2.9). Suppose X_n already has been specified such that the inverse S_n in (2.3) exists. Now consider a row of x' of X_N which is not a row of X_n as the $n+1^{st}$ candidate, so that

$$X_{n+1} = \begin{pmatrix} X_n \\ x' \end{pmatrix}, \quad X_{n+1}'X_{n+1} = X_n'X_n + xx'. \qquad (2.8)$$

Then defining $S_{n+1} = (X_{n+1}'X_{n+1})^{-1}$,

$$S_{n+1} = S_n - \frac{S_n xx' S_n}{1 + x'S_n x}. \qquad (2.9)$$

From (2.9) and (2.7), it follows that

$$\varphi(X_{n+1}) = \varphi(X_n) - \frac{x'S_n T S_n x}{1 + x'S_n x}. \qquad (2.10)$$

The value of x that maximizes the right-most term in (2.10) therefore will produce the optimal X_{n+1} from X_n. Since cost must be considered, let c_i be the cost of the experimental subject located in the i-th row of X_N, having vector of characteristics x_i. Define

$$CE_i(X_n) = \frac{1}{c_i} \frac{x_i'S_n T S_n x_i}{1 + x_i'S_n x_i} \qquad (2.11)$$

as the cost-effectiveness of experimental unit i, conditional on X_n already having been selected.

The implied algorithm now is obvious. At each step choose i from the remaining available units to maximize CE_i. After a choice is made, update S_n using (2.9). Repeat this process until the budget constraint is met. Each of the computations (2.10) is relatively inexpensive since only two inner products are involved, and the inversion (2.9) is much cheaper than inverting a full matrix.

The preceding algorithm does not cover the initial cases when $X_n'X_n$ is not invertible. For these cases an initial invertible matrix Q_0 is specified. The choice Q_0 proportional to $X_N'X_N$ or some guess at the final form of the optimal $X_n'X_n$ often is convenient. Then

$$S_{n,\epsilon} = (X_n'X_n + \epsilon Q_0)^{-1} \qquad (2.12)$$

does exist for any $\epsilon > 0$, and is used in place of S_n in (2.11), and (2.9). The value of ϵ usually would be quite small. In tests, the final efficiency and the selection order of experimental units has depended little on the choice of ϵ. As $\epsilon \to 0$, if Q_0 is the identity matrix, the limit in (2.12) is the "pseudo-inverse" of $X_n'X_n$. The pseudo-inverse could be used to provide another method of treating the noninvertability problem.

The routine just described, which moves from no assignments to a complete assignment is termed "build-up." Further improvement may be sought by proceeding then to a "substitution" phase. Obviously substitution may be used to try to improve any full assignment, not only one obtained through build-up.

Define

$$CE_i^*(X_n) = \frac{1}{c_i} \frac{x_i'S_n T S_n x_i}{1 - x_i'S_n x_i}. \qquad (2.13)$$

Given the matrix X_n, it may be seen that removal from X_n of the subject i which minimizes CE_i^* will

cause the minimum increase in variance per unit cost. This follows because

$$S_{n-1} = S_n + \frac{S_n x x' S_n}{1 - x' S_n x} \tag{2.14}$$

and so

$$\varphi(X_{n-1}) = \varphi(X_n) + \frac{x_i' S_n T S_n x_i}{1 - x_i' S_n x_i} \tag{2.15}$$

if X_{n-1} is X_n with the row corresponding to x_i removed.

The substitution procedure simply removes the already selected subject which minimizes CE* (2.13) and puts this subject back into the pool of unselected subjects. Then a search is made over the pool of unselected subjects and the subject that maximizes CE (2.11) is determined. This subject then is *substituted* for that one just removed. Iteration stops when a subject is substituted for himself.

In our experience, only small gains have been obtained from substitution after build-up. This provides an empirical demonstration that the build-up algorithm is effective in obtaining a near optimum. Substitution also may be used as a convergence check in any particular case.

Algorithms developed from (2.13) can be effective from other viewpoints. For example, if n is nearly as large as N, then considerable computational expense is saved by assigning all N subjects initially and then "building down" to n by weeding out those that are ineffective. And, in general, if both build-up and substitution algorithms are available, then there are trade-offs between the number of computations that should be attempted in either phase in order to attain a nearly optimal set overall. For example, during build-up it is cheaper to search over only a small set of available subjects, not all subjects, before making each selection. (This concept is discussed more fully in Section 7 under the name "search length.") The result is a fairly good assignment that can be further improved without too many substitutions. Another application of the substitution algorithm, to the allocation model, will be discussed in Section 3.

3. THE FINITE SELECTION MODEL AS AN ALLOCATION MODEL

It was noted in Section 1 that the finite selection model just described differs from the allocation model considered in (Conlisk-Watts 1969) primarily because their model has a limited number m of distinct rows, but each row may be selected an unlimited number of times. The X of FSM may have an unlimited number of distinct rows, but no row may be selected more than once.

Still, the substitution algorithm just described may be used to solve for the optimum in the allocation model and in that case (the "continuous case") it can be shown that the optimum is attained. Because these considerations are important from both a mathematical standpoint and for application to the allocation model, we shall digress briefly in this section to discuss them.

Let X be a specified m × k matrix. This is as before, except each row of X may be chosen not only once, but any nonnegative number of times, including a fractional number. The solution to the allocation model problem then is a vector $(\pi_1, \ldots, \pi_m)$ of proportions, π_i specifying the proportion of times the row x_i' of X is to be allocated. The cost constraint does not affect the proportions, only the relative values of the costs c_i do. Initially, let N be a large integer, and choose any integers $n_1, \ldots, n_m$ such that $\Sigma n_i = N$. This is the initial allocation and its inner product matrix is

$$X_N' X_N = \Sigma n_i x_i x_i' . \tag{3.1}$$

The FSM then is employed in the substitution mode, using (2.13) and (2.11) to find the least cost-effective row i^* of X, reducing n_{i^*} by one. It then finds the most cost-effective row i of X and augments n_i by one. Each substitution moves a step closer to the optimum, which cannot be achieved in general if N is finite. If ever $i^* = i$, then N is doubled and so is each value n_i. The procedure is repeated for the new value of N. This procedure ends when N is quite large and the optimum is declared to be

$$\pi_i = n_i/N, \qquad i = 1(1)m = 1, 2, \ldots, m. \tag{3.2}$$

When N is large, the contribution $x_i' S_N x_i$ in (2.11), (2.13) is small compared to unity (a typical value is 1/N). Consequently, for large N, it is advantageous to ignore the difference between CE and CE*, replacing both by

$$CE_i(n_1, \ldots, n_k) = \frac{x_i' S_N T S_N x_i}{c_i} , \tag{3.3}$$

over all points i = 1, ..., m where

$$S_N = (\Sigma n_i x_i x_i')^{-1}. \tag{3.4}$$

The substitution then is based on the largest and smallest values of (3.3).

This algorithm is the continuous extension of the substitution algorithm used in the discrete situation. It is also the method of "steepest descent" applied to minimization of the criterion for the allocation model, and is known to converge to a unique minimum since φ is a convex function of $(\pi_1, \ldots, \pi_m)$. The allocation model always has been solved very inexpensively using this algorithm.

By analogy, the algorithms for the FSM also must produce nearly optimum solutions in the discrete case (where each x_i may be selected only once) provided N is fairly large. This has been supported by our experience using these algorithms, even including ones with small N.

In very large problems, such as actual selection and assignment of HIS families, the goal of optimality is too expensive, even if it is achievable. In these situations, we settle for the less expensive uses of the preceding algorithms described in Section 7 and for substantial precision gains without insisting upon optimal assignment. Optimality of the algorithm is an issue of greater theoretical than practical importance provided that there is evidence that most of the gains realized by optimal assignments has been achieved. Of course the computer could be programmed to continue only so long as the value of the marginal increase

in experimental precision exceeds the marginal cost of computer time.

4. USING THE FINITE SELECTION MODEL TO CONSTRUCT BALANCED GROUPS

In practice the FSM must select subjects for each of several treatment groups from the same subject pool. There then are two objectives: to achieve high efficiency within each group; and to balance the design matrices between the groups. We show here how the version of FSM which is designed to produce efficient samples may be extended to produce both efficient and balanced groups.

Let n_i be the number of subjects to be assigned to the i-th treatment group, i = 1(1)g. Let N be the total number of subjects available for experimentation. Of course $n_1 + \ldots + n_g \leq N$. (While costs actually are to be allocated, for ease of exposition here they are taken to be equal. This justifies formulation in terms of prespecified sample sizes.)

Problem 2

The characteristics of the N potential experimental subjects are given in the N rows of X_N. They are to be assigned to g groups of prechosen sizes $\{n_i\}$ such that the groups are (i) "balanced" and (ii) each group has good precision.

The concept of "balance" is difficult to define and measure when it is imperfect. If $n_1 = n_2 = \ldots = n_g$, then perfect balance corresponds to

$$X_1 = \ldots = X_g \tag{4.1}$$

with X_i the design matrix of the i-th treatment group. This is possible only when n blocks of g perfectly matched experimental subjects are available to be assigned as balanced blocks. Although this can be caused to happen artifically, by stratifying and ignoring certain variables (for example in the HIS, income could be split into three groups only and all other preexperimental variables ignored), perfect matching almost always is impossible when assigning experimental subjects to treatments. When the n_i are unequal, the analogy to (4.1) is even more restrictive.

In general, measures which compare imperfectly balanced assignments are needed so that the best of several imperfect alternatives may be selected. As a first cut, when the linear relationship between the covariates and the dependent variables is certain, (4.1) is no better for estimating treatment contrasts than requiring that the vector of column means for each matrix X_i be the same for all i = 1(1)g (Haggstrom 1975).

In the case $n_1 = \ldots = n_g$, a measure of balance which is congenial to the development in Sections 2, 3 requires that the set of numbers $\varphi_i \equiv \varphi(X_i)$ be close to one another. Then the measure of balance

$$B \equiv \Sigma(\varphi_i - \bar{\varphi})^2, \quad \bar{\varphi} = \Sigma\varphi_i/g \tag{4.2}$$

may be used. When the n_i are not all equal, the φ_i cannot be expected to be equal, and a modification must be made. We shall return to this point in Section 6.

The algorithm used by the FSM to produce both balance and efficiency simply requires the experimental treatments to take turns selecting the cost-effective experimental subject from the remaining list of unselected subjects during the build-up phase. Each treatment group in turn chooses one subject, making the cost-effective choice for its needs according to the criterion (2.11). The next choice is made by a different treatment. At the end of the build-up phase, when the full quota $n_1, \ldots, n_g$ has been reached, the $n_0 \equiv N - \Sigma n_i$ unselected subjects which are least useful remain unassigned, while the Σn_i subjects which provide maximum information are assigned and are fairly well balanced over the treatment groups.

Fixed, random, or sequential methods may be used to determine selection order. The simplest and most powerful "fixed" procedure is the percentage method. At each step, the treatment group which has had the smallest percentage of choices at that point makes the next selection. A fully random procedure would determine the selection order at random, subject to the constraints that at the end of Σn_i choices there must have been n_i selections for plan i. There are many possible subrandomizations which lie between these two strategies, for example blocked and stratified randomization. The percentage method for selection order would be expected to provide the optimum balance, random selection order the least (it's possible that one group would make all the first choices in this case). However, randomization still may be an advantage insofar as visible randomization often is viewed as a necessary input to experimentation. Although this form of randomization is least efficient in the FSM, even then the balance produced is a substantial improvement over that achieved by ignoring the FSM and assigning subjects at random to treatments.

Why do these methods produce balanced samples? Heuristically, because each treatment selects according to the same algorithm (it doesn't even matter if the objective function is misspecified badly), and with selection times distributed similarly to the other treatments. The results would be satisfactory even if the least valuable families were evaluated as the most valuable by the selection criterion. This produces balanced samples for the same reason that taking turns when choosing teams for competition leads to balanced teams when the choosers are equally capable. In one example, the case when n blocks of g identical subjects are available, it is clear that the fixed selection order will lead to complete balance, producing the result (4.1). The rest of the proof is based on experience: selections that have been made following these procedures have been much better balanced than those that occur from simple random sampling. Finally, after any particular allocation, the balance of the resulting sample may be calculated to make sure it is satisfactory.

The percentage method and random selection order are "nonsequential" in that they may be specified in advance before any selections actually are made. Truly "sequential" selection methods would depend on the current value of φ_i for each treatment group, the group with the largest (worst) value being permitted to make the next selection. A

substitution phase (cf., Section 3), which *is necessarily* sequential, can be carried out if the φ_i can be compared. Then at each step the least efficient treatment would be permitted to draw from the most efficient. When the $\{n_i\}$ are unequal, this can be done only if an entirely satisfactory scaling of the $\{\varphi_i\}$ has been achieved, one which accounts for differences due purely to unequal sample sizes.

5. SCALING THE OBJECTIVE FUNCTION φ FOR REPRESENTATIVE SAMPLING

If the assigned weights $\{W_i\}$ are to have a meaningful effect, it is necessary that they be defined in terms of an expected assignment. If for example we expect $\mathrm{Var}(a_1) = 1{,}000{,}000\ \mathrm{Var}(a_2)$ with random sampling, then an algorithm based on (2.6) will ignore a_2 and concentrate almost exclusively on $\mathrm{Var}(a_1)$ for any moderate value of the ratio W_1/W_2.

Relative precisions like 1,000,000 are not as rare as one might think. Suppose a_1 is the regression coefficient of the variable x_1. The precision for estimating a_1 is drastically reduced by the inclusion of another independent variable x_2 in the model which is highly collinear with x_1. The researcher who specifies the weights $\{W_i\}$ needs protection against the effect that the inclusion or exclusion of an extraneous variable has on the design by making some weights meaningless.

It also is desirable to specify appropriate centering and scaling of the design matrix. For example, if a target population is specified then it is appropriate to center the independent variables for the regression at their expected values for that population. In a significant number of examples, the N subjects being assigned provide the best estimate of the target population, because they represent the largest random sample taken from that population.

The methods about to be developed in this section account for the difficulties just described by using the expected outcome of random sampling as a bench mark. By good fortune, these methods also will contribute to measurement of balance for situations where the group sizes $n_1, \ldots, n_g$ are unequal (cf., Section 4).

Suppose X_N is constructed in the following manner. The N subjects are sampled at random from a population, the i-th subject having k-1 vector of measurable characteristics $z_i' = (z_{i2}, \ldots, z_{ik})$. Now for i = 1(1)N, let X_N be the matrix with elements

$$x_{i1} = 1, \quad x_{ij} = z_{ij} - \bar{z}_j, \quad j = 2(1)k \qquad (5.1)$$

with $\bar{z}_j \equiv \Sigma z_{ij}/N$. The first column of X_N is for fitting a constant term, the remaining columns are centered at the best estimates of the population means available.

Suppose n of the N rows are sampled from X_N at random. Define X_n as the n × k matrix comprised of these n rows, and again let $S_n = (X_n'X_n)^{-1}$. The expected value of S_n given the matrix of values S_N is needed. The case of the multivariate normal distribution will be considered first.

Under the assumption that the N independent variables z_i which define the rows of X_N are sampled at random from a multivariate normal distribution with mean vector μ and covariance matrix Σ,

$$z_i \sim N_{k-1}(\mu, \Sigma), \text{ independently } i = 1(1)N, \qquad (5.2)$$

then this expectation is given by

$$(n-k-1)ES_n|\bar{z}_N, S_N = \begin{pmatrix} 1 - \frac{2}{n} - \frac{k-1}{N} & 0' \\ 0 & \hat{\Sigma}^{-1} \end{pmatrix} \qquad (5.3)$$

with $\bar{z}_N \equiv \Sigma z_i/N$ and $\hat{\Sigma}$ the (k-1) × (k-1) matrix with (j, ℓ) element, $2 \le j, \ell \le k$, defined by

$$\sum_{i=1}^{N} x_{ij}x_{i\ell}/(N-k-1). \qquad (5.4)$$

The result in (5.3) assumes $n > k + 1$. Its proof, which is somewhat technical, will not be given here.

Denote by Λ_n the k × k matrix on the right-hand side of (5.3), and let p_j' be the j-th row of the matrix P in (2.5) so $a_j = p_j'b$. Formulas (2.6) and (2.7) now are modified, provided $n > k + 1$, by redefining the weights:

$$\varphi_n^*(X_n) = (n-k-1)\Sigma W_j \frac{\mathrm{Var}(a_j)}{\sigma^2 p_j'\Lambda_n p_j} \qquad (5.5)$$

$$= (n-k-1)\Sigma W_j \frac{p_j'S_n p_j}{p_j'\Lambda_n p_j}. \qquad (5.6)$$

If the redefined weights are normalized to add to unity

$$\Sigma W_j = 1 \qquad (5.7)$$

then for random sampling, with the assumption (5.2),

$$E\varphi_n^*(X_n) = 1. \qquad (5.8)$$

Therefore a specific choice of X_n improves on the expected outcome of random sampling if and only if (5.5) is less than unity. Of course when n = N, no improvement on random sampling is possible, and it may be verified in this case that $\varphi_N^*(X_N) = 1$.

The weights defined by (5.5)-(5.7) acquire a relative meaning, with respect to random sampling, rather than the absolute meaning of earlier sections. Regardless of the relative precisions of $\{a_j\}$, each of the weights do have an effect on the sample.

The notion of scaling the weights may be used even when the validity of the assumption (5.2) is suspect. In general, any estimate $v_j(n)$ of $\mathrm{Var}(a_j)$ may be specified and then

$$\varphi_n^{**}(X_n) = \Sigma W_j p_j'S_n p_j/v_j(n) \qquad (5.9)$$

used in place of (5.6). Values of $v_j(n)$ may be estimated using simulations from random sampling, but alternative choices for $v_j(n)$ are $p_j'S_n^*p_j$ with S_n^* the design matrix for a particular assignment, or a value proportional to $p_j'(X_N'X_N)^{-1}p_j$. This

last choice is equivalent to (5.6) if $1/(n-k-1)$ is the constant of proportionality, and provides support for using (5.6) even when (5.2) is not true.

In applications thus far we have found (5.6) to be a useful approximation even when several of the independent variables are discrete or categorical, and therefore are not approximately multivariate normal. Actually, (5.3) is infinite for random samples if categorical or discrete variables are present since there is the possibility that random sampling will fail to represent all categories. Use of (5.5) and (5.6) as a bench mark therefore favors random sampling, since it makes the expected outcome of random sampling appear to be better than it actually would be.

The definition in (5.6) may be used to generalize the measurement of balance provided in (4.2). Suppose an assignment of subjects to the g treatment groups has been made with unequal group sizes $n_1, \ldots, n_g$. Let $\varphi_1^*, \ldots, \varphi_g^*$ be the g values of (5.6). Each of these values has unit expectation for random sampling. Formula (4.2) may be used with φ_i replaced by φ_i^* to define balance. More generally, (4.2) may be replaced by a weighted variance with weights increasing in n_i chosen to reflect the variability of φ_i^* for random sampling. The best choice of weights is still under investigation.

These measures of balance allude to the balance of final efficiencies only, although the FSM balance algorithm has led to substantially better balance than that expected from random sampling for the variances of all the $\{a_i\}$. Balance with respect to each individual coefficient may be measured in this same manner. The weight vector in (5.6) simply is reset to indicate the particular value j for which estimation is being considered. That is, one sets $W_j = 1$ and all other values of the weights at zero before computing the statistic (4.2). This then is done for each $j = 1(1)p$, resulting in p separate measures.

The expected balance, assuming random sampling, may be computed either by applying the same theory used to derive (5.3) or by using simulation methods. The theoretical results, which are based on assumption (5.2), will be published elsewhere.

After each assignment of subjects to treatments the FSM prints out a battery of statistics which are used to assess the design, including the g x p matrix H of normalized variances of regression coefficients

$$h_{ij} = (n_i - k - 1)\,\frac{p_j' S_{n_i} p_j}{p_j' \Lambda_{n_i} p_j} \qquad (5.10)$$

and the measures of precision and balance just described, each of which is a function of the elements (5.10). Therefore, gains in precision and balance relative to random sampling can be observed and the design inputs or selection method reconsidered if the results are unsatisfactory.

For convenience of exposition, only the situation of equal cost c_i per experimental subject has been discussed in Sections 4, 5. The general case with costs of subjects, not numbers of subjects, being allocated to treatments raises no markedly new problems. There is insufficient space here to treat this generalization.

6. ROBUSTNESS

The act of selecting the most efficient subjects for experimentation and rejecting the least efficient can lead to biases if: (a) the model and its inputs are improperly specified; or (b) if there are measurement errors associated with the independent variables, these errors being correlated with the dependent variable. To illustrate (b) with a somewhat unrealistic example, suppose family income were included as a single independent variable specified to have a linear effect. Then families with extremely high or low incomes are preferred by the FSM. If actual incomes of the target population are fairly homogeneous, then the extreme incomes reported might belong mainly to families who either are careless in reporting such data or who lie. The treatment groups thus derived would not be optimal with respect to income, but instead would be overrepresented by careless and untruthful people. If such people differ with respect to the dependent variable, as they would be likely to do, then "optimal" selection from the initial population leads to biased estimates for the general population, and without compensating gains in precision. This same example also may be used in case (a) to illustrate that biased results may be obtained from improper specification of the income term, even if income always is reported accurately.

An "heroic assumption" may be made to circumvent these concerns, namely that all variables which could have an effect on the dependent variable are already included in a properly specified model. Of course the heroic assumption is too strong because improper specification still could yield samples which result in correct estimates of treatment contrasts, if treatment groups are balanced. While the heroic assumption probably is valid for many experiments in the physicial sciences, there tend to be too many latent variables (unmeasured variables correlated with the dependent variables) for its validity in social experimentation.

The effect of latent variables must then be accounted for by

(i) reducing their variance by balancing variables likely to be correlated with them,

(ii) making the residual effect of the latent variables random with respect to the treatments.

The best way to meet these objectives with the FSM, although at a substantial sacrifice of optimality if n_0 (6.1) is large, is to make all groups be representative of X_N. Equivalently, the group of subjects being discarded must be balanced with the treatment groups. Letting

$$n_0 = N - \sum_{i=1}^{g} n_i \qquad (6.1)$$

be the number of experimental subjects which must be discarded, we define one extra group, the discard group of n_0 subjects, and require its members to be selected in the same fashion as the other treatment groups.

Constructing representative groups in this manner makes the design very robust. If the

independent variables of X_N are pure random noise, the result will be a simple random assignment. If the independent variables of X_N are measured with error but are correlated with the "true values" then a better than random assignment is achieved. This argument applies equally to problems of model misspecification, to errors in variables, and to the use of proxy variables. To the extent that latent variables are correlated with the independent variables their effect is reduced. In the limiting case, any latent variable having a multiple correlation coefficient of unity with the set of independent variables will be as well balanced as the independent variables are. Finally, the values of the independent variables act as a randomizing agent so that the residuals of the latent variables (net of their predicted values from the independent variables) would be related to the treatments in random fashion (Rubin 1974).

Some of the arguments just presented apply also to the situation of Section 4 with balanced treatments, even if the discard group is not balanced.

7. APPLICATION OF THE FSM TO THE HEALTH INSURANCE STUDY; FURTHER EXTENSIONS

The finite selection model has been used in two modes for the Health Insurance Study. The Conlisk-Watts allocation model, implemented by FSM (Section 3) was used to determine which and how many health insurance plans (treatments) should be used and how many $(n_1, n_2, \ldots, n_g)$ families (subjects) should be assigned to each plan. Then in the mode of Sections 4-6 the model has been used over a three-year period to assign families to treatments and controls in the sites chosen for the experiment.

Nothing will be said here about the allocation application, except that certain extensions of the Conlisk-Watts model were developed to cover the case where the rows of X are randomly related to the variable of choice (in that case, the plan and distribution of family income). The modified algorithm is not unlike the one to be described in (7.1) which solves the "multiple family unit" problem.

In each of several sites, roughly N = 600 families are assigned to g = 13 groups of treatments, controls, and discards. (These numbers vary from site to site.) This is done on the basis of 24 variables constructed from 14 demographic factors provided from baseline interviews of families. These factors are wage and nonwage income, family size, education, insurance, welfare status, number of family heads, race, age, sex, health status, physician and hospital visits, and location of residence. Location is included because it could be a proxy variable for certain unmeasured variables. The other factors are chosen because each is known to predict future utilization of health services and future health status. If the 24 variables just described, after preliminary scaling, are denoted $z_2, \ldots, z_{25}$, then the matrix X_N with k = 25 columns is defined according to (5.1). The population from which X_N is sampled at random is the target population.

With this target population and centering using (5.1), the p x k policy matrix P has p = 29, k = 25 with the upper 25 x 25 portion being the identity matrix. The last four rows account for interest in the missing categories of the family size, education, health status, and age factors, because these were divided into at least three categories. Thus, the regression coefficients themselves are essentially the parameters of interest.

The 29 weights $\{W_j\}$ were chosen to be fairly equal. Formula (5.6) is used to define the objective function, although S_n is modified to (2.12) with $Q_0 \equiv X_N'X_N$. The assignment strategy combines the methods of Sections 4-6 to produce a combination of optimality, balance and robustness. Certain additional modifications of the procedures described before have been necessary, and will be discussed in the ensuing paragraphs. The results from data in the first site show that, used in this conservative way with numbers similar to those mentioned above, a 25 percent improvement over random sampling in φ (formula (5.6)) may be obtained. The measure of imbalance (formula (4.2)) is only 4 percent of that expected for random samples. The computational cost of FSM in making these selections is about twenty-five dollars.

The entire assignment is made in several *stages*, each spaced several weeks apart. This is required because selections must be made for the first enrollments before all baseline data are received. The sequential nature of the FSM in its build-up mode fits in naturally with this requirement; selections at each stage optimally augment those already made.

Because the final sample size on some plans is small, and is reduced further by about one-half at the first stage, the condition $n > k + 1$ is not met on several plans during the first assignment. Therefore, the 13 groups are created in a series of "splits" to improve this situation. The first split includes fewer groups, each meeting the condition $n > k + 1$ so that precision and balance may be adequately evaluated by the methods of Section 5. Further splitting of these groups into smaller groups is carried out using fewer than all 24 independent variables so that again $n > k + 1$ for each group. By the time the final stage is reached this problem is no longer serious.

"Multiple Family Units" (MFU) often occur. These are groups of families who live in the same household unit and must be assigned to the same treatment group. The algorithm (2.11) for an MFU is modified to

$$CE_{MFU}(X_n) = \frac{\sum_{i \in MFU} c_i CE_i(X_n)}{\sum_{i \in MFU} c_i} . \qquad (7.1)$$

This is a linear approximation to the total improvement to X_n divided by the total cost for all families in the MFU. The linear approximation is good if n is large, i.e., if the impact of any one family on the objective function is small.

The computational cost of making the assignments with the numbers just described would be about 250 dollars for each site. While this is not excessive, it can be reduced by a substantial factor by limiting the "search-length" for each selection, and we do this. Rather than search over *all* unassigned families at each step to find the best, the search is limited to a few families. We have found that choosing the best of 20 families at each

step cuts computational costs ten-fold with only a modest loss of efficiency. Each successive selection is made by cycling through the list of unselected families in intervals of 20, and returning to the beginning of the list as often as is necessary.

Introducing methods with limited search length makes the resultant selection dependent on the initial listing order of the families. For this reason, unselected families initially are listed in random order. A search length of unity therefore would result in a random selection, and longer search lengths will outperform random selection. Besides cutting computational costs, limited search length procedures have the advantage of adding randomness to the assignment.

Stratified sampling may be combined with the optimality of the FSM. If the unselected subjects are partitioned into strata indexed 1, 2, ..., s then the model may be instructed to assign a prespecified number of subjects from each stratum to each treatment group. Selections meeting these constraints are most easily specified by inputting a "selection order matrix." The selection order matrix is an N x 2 matrix, the first row indicating the treatment t_1 which makes the first selection and the second row the specified stratum s_1 from which the selection is to be made. The second row indicates that t_2 may select from s_2, and so on. The selection order thus specified constitutes a generalization of fixed selection order discussed in Section 4 to the case of more than one stratum. As noted there, it offers a second opportunity (besides reduced search length) to add randomness to the assignment by randomizing the order in which treatments make choices.

The strata just described are used in HIS assignments to get prespecified counts for income groups, family sizes, multiple family units, and to separate families by selection stages.

8. SUMMARY

Theory and algorithms have been described which provide the basis for the development of a computer model, the "finite selection model," which is being used in several ways in the statistical design of the Health Insurance Study. The main purpose of the model is to assign experimental subjects to treatments, meeting experimental objectives of optimality, efficiency, and balance. The software already is available and has been used, although it continues to be modified and improved.

Most of the general ideas for deriving optimal, balanced, and robust samples can be applied to other objective functions, such as the generalized variance, and to other design situations. Related work on D-optimality in various situations, besides references already mentioned, is reported (Harville 1975), (Dykstra 1971), and (Wynn 1975), and also in additional references cited by them.

9. ACKNOWLEDGMENTS

The research reported herein was performed pursuant to a grant from the Department of Health, Education, and Welfare, Washington, D.C. The opinions and conclusions expressed herein are solely those of the author and should not be construed as representing the opinions or policy of any agency of the United States Government.

Keith H. Davis, a Rand computer scientist, has programmed the finite selection model. He also has made numerous suggestions during its development. Adapting the FSM algorithm to find the optimum in the allocation model as described in Section 3, was his idea. Bradley Efron, statistical consultant to the HIS, proposed the idea of randomizing the order of treatment selection.

REFERENCES

1. Conlisk, John, and Harold Watts, "A Model for Optimizing Experimental Designs for Estimating Response Surfaces," *Proceedings of the Social Statistics Section*, American Statistical Association, 1969, pp. 150-156.

2. Dykstra, O., "The Augumentation of Experimental Data to Maximize $|\underset{\sim}{X}'\underset{\sim}{X}|$," *Technometrics*, 13, 1971, pp. 682-688.

3. Elfving, Gustav, "Optimum Allocation in Linear Regression Theory," *Annals of Mathematical Statistics*, 23, 1952, pp. 255-262.

4. Haggstrom, Gus, *The Pitfalls of Manpower Experimentation*, The Rand Corporation, P-5449, April 1975.

5. Harville, David A., "Computing Optimum Designs for Covariance Models," in J. N. Srivastava (ed.), *A Survey of Statistical Design and Linear Models*, North-Holland Publishing Company, 1975, pp. 209-228.

6. Kiefer, J., "Optimum Experimental Designs," *Journal of the Royal Statistical Society, Series B*, 21, 1959, pp. 272-319.

7. Newhouse, Joseph P., *The Health Insurance Study --A Summary*, The Rand Corporation, R-965-1-OEO, March 1974.

8. Rubin, Donald B., "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, Vol. 66, No. 5, 1974, pp. 688-701.

9. Wynn, H. P., "Simple Conditions for Optimum Design Algorithms," in J. N. Srivastava (ed.), *A Survey of Statistical Design and Linear Models*, North-Holland Publishing Company, 1975, pp. 571-579.