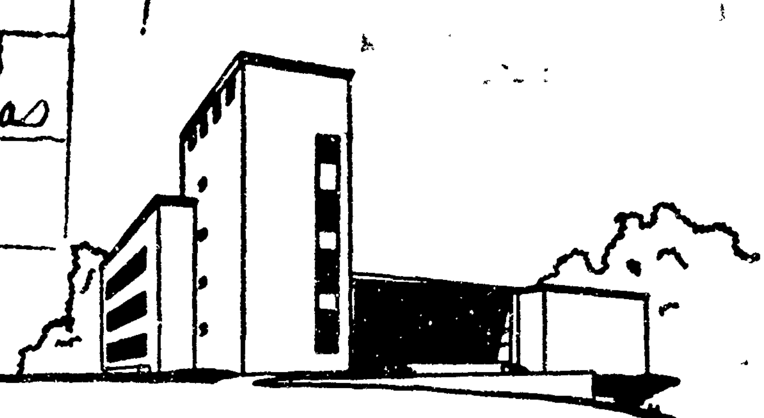
AD623394

Carnegie Institute of Technology

Pittsburgh 13, Pennsylvania



GRADUATE SCHOOL of INDUSTRIAL ADMINISTRATION

William Larimer Mellon, Founder

Management Sciences Research Report No. 50

ON STATISTICAL COST ANALYSIS
AND OPERATIONS RESEARCH

by

Charles H. Kriebel

August, 1965

MANAGEMENT SCIENCES RESEARCH GROUP
GRADUATE SCHOOL OF INDUSTRIAL ADMINISTRATION
CARNEGIE INSTITUTE OF TECHNOLOGY
PITTSBURGH, PENNSYLVANIA 15213

This report was prepared as part of the activities of the Management Sciences Research Group, Carnegie Institute of Technology, (under Contract Nonr 760(24) NR 047-048 with the U. S. Office of Naval Research).

1. Introduction

Optimization analyses within firms usually presuppose the specification and estimation of a functional which describes terminal actions and state variables in units of some criterion, such as operating costs. In most operations research studies optimization problems have the character of maximization or minimization. For example, in linear programming resource allocations, one seeks to minimize operating cost expressed as a linear function of the levels of different activities and a given set of cost coefficients, subject to a defined system of constraints. Within the programming framework one can explore the sensitivity of proposed solutions to errors in the estimation of model coefficients with the aid of a post-optimality analysis using parametric techniques. However, the task of obtaining initial estimates of all coefficients remains with the analyst's best judgment. Moreover, in many operations research investigations, such as those employing non-linear programming models, the form of the functional to be optimized (or relations within the constraint set) is not known *a priori* and, consequently, a statistical analysis of environmental and historical data must precede the normative development of explicit decision-making procedures. In this context the applied scientist often must develop the theoretical model *and* complete the empirical analysis to successfully implement his recommendations.

A recent paper by Professor Theil [9] discusses some of the general interactions between the fields of econometrics and management science. Clearly, at one time or another the operations research or management scientist is both a theoretician and, like the econometrician, an empiricist. In contrast to discussions on theory, the literature on empirical problems in operations research is notably sparse. The purpose of this report is to illustrate the overlap between data analysis problems in operations research and the theoretical development of a model. The case in point to be considered is statistical cost estimation in quadratic programming models.

In the next section we review the basic characteristics of quadratic programming analysis and then introduce an example model based on the study of linear decision rules for production planning by Holt, Modigliani, Muth, and Simon [2]. The discussion then proceeds to consideration of alternative approaches to estimating the coefficients in a specified objective function based on operating costs. Within this framework several general approaches are discussed, including simple multiple regression, the application seriatim of single equation techniques to relations in the model, and the simultaneous estimation of equation systems, as with k-class estimates in econometrics. In conclusion, some practical considerations for error analysis in estimation and sensitivity analysis in optimization are reviewed.

2. Background: Linear Decision Rules and Quadratic Programming

Quadratic programming problems concern the optimization of a quadratic objective function subject to a linear system of constraint equations (or inequations). Optimization problems with this mathematical structure have arisen in various applied areas, such as capital budgeting in investment portfolio analysis, aggregate production and employment scheduling, and so on (e.g., see Boot [1]).

The basic problem of minimizing (or maximizing) a quadratic function with respect to a column vector of actions can be stated as

$$\min_{\{\underline{a}\}} c(\underline{a}, \underline{x}) = \lambda_o(x) + 2\underline{a}'\, \underline{\lambda}(x) - \underline{a}'\underline{\underline{Q}}\, \underline{a} \,, \qquad [1]$$

where the vectors $\underline{a}$ and $\underline{x}$ are of dimensions (n×1), $\lambda_o(x)$ is a scalar function of $\underline{x}$, $\underline{\lambda}(x)$ is a vector function of $\underline{x}$, such as $\underline{\lambda}(x) = \underline{\underline{D}}\, \underline{x}$ for $\underline{\underline{D}}$ an (m×n) dimensional matrix of rank $m \leq n$, and $\underline{\underline{Q}}$ is a non-singular matrix of dimensions (n×n). Let $\partial/\partial\underline{a}$ denote taking partial derivatives with respect to the column vector $\underline{a}$. A necessary condition for $\underline{a} = \underline{a}^*$ to be a local minimum of [1] is that

$$\frac{\partial}{\partial\underline{a}} c(\underline{a},\underline{x})\Big|_{\underline{a}=\underline{a}^*} = \underline{\lambda}(x) - \underline{\underline{Q}}\, \underline{a}^* = \underline{0} \,, \qquad [2]$$

or equivalently that

$$\underline{a}^* = \underline{\underline{Q}}^{-1}\, \underline{\lambda}(x) = \underline{\underline{Q}}^{-1}\underline{\underline{D}}\, \underline{x} \,. \qquad [3]$$

A sufficient condition that [3] be the global minimum is that $c(\underline{a},\underline{x})$ be convex or, specifically, that $\underline{\underline{Q}}$ is positive definite.

The addition of linear equality constraints to the problem in [1] can be incorporated within this framework with little difficulty. For example, suppose the original problem is

$$\min_{\{\underline{a}\}} c(\underline{a},\underline{z}) = k + 2\,\underline{\gamma}'\underline{a} + 2\,\underline{\beta}'\underline{z} - (\underline{a}'\,\underline{\underline{A}}\,\underline{a} + \underline{z}'\,\underline{\underline{B}}\,\underline{z} + a'\underline{\underline{C}}\,z + z'\,\underline{\underline{C}}'\underline{a}) \qquad [4]$$

$$\text{subject to:} \qquad \underline{z} = \underline{\underline{R}}\,\underline{a} + \underline{x}\,, \qquad [5]$$

where $\underline{\underline{R}}$ is a (txn) dimensioned matrix of full row rank and $t < n$. The problem stated in [4] subject to [5] can be reformulated as the problem posed in [1] by substitution of [5] into [4] for the vector $\underline{z}$. That is, the relations in [1] become

$$\left.\begin{aligned} \lambda_o(x) &= k + 2\underline{\beta}'\underline{x} + \underline{x}'\underline{\underline{B}}\,\underline{x} \\ \underline{\lambda}\,(x) &= \underline{\gamma} + \underline{\underline{R}}'\underline{\beta} + (\underline{\underline{C}} + \underline{\underline{R}}'\underline{\underline{B}})\,\underline{x} \\ \underline{\underline{Q}} &= \underline{\underline{A}} + \underline{\underline{R}}'\underline{\underline{B}}\underline{\underline{R}} + \underline{\underline{C}}\underline{\underline{R}} + \underline{\underline{R}}'\underline{\underline{C}} \end{aligned}\right\} \qquad [6]$$

The solution to the problem as now stated is identical to that given by [3].

If the system of constraints in [5] is one of inequality relations, such as

$$\underline{\underline{E}}\,\underline{a} + \underline{x} \leq \underline{z}\,, \qquad [7]$$

where $\underline{\underline{E}}$ is a matrix of full row rank with (t×n) dimensions, and $t<n$, the above procedure can still be employed through the introduction of a (t×1) dimensioned slack vector $\underline{w}$. That is, write

$$\begin{aligned} &\underline{\underline{E}}\,\underline{a} + \underline{\underline{I}}\,\underline{w} + \underline{x} = \underline{z} \quad \text{, or} \\ &\underline{\underline{R}}^o\underline{a}^o + \underline{x} = \underline{z}\,, \end{aligned} \qquad [8]$$

where $\underline{\underline{R}}^o = [\underline{\underline{E}}\ \underline{\underline{I}}]$ and $\underline{a}^o = [\underline{a}\ \underline{w}]'$. Then appropriately augmenting the statements for $\underline{\lambda}(x)$ and $\underline{\underline{Q}}$ in [1] to conform to the (n+t) dimensions of $\underline{a}^o$, the solution sequence can proceed as before to obtain $(\underline{a}^o*) \Rightarrow \underline{a}*$, that is, we now have $\underline{\lambda}^o(x) = (\lambda(x)', \underline{0}')'$ and $\underline{\underline{Q}}^o = \begin{bmatrix} \underline{\underline{Q}} & \underline{\underline{o}} \\ \underline{\underline{o}} & \underline{\underline{o}} \end{bmatrix}$ for $\underline{\underline{Q}}^o$ a square (n+t) matrix.

Similar procedures can be introduced to handle non-negativity restrictions on $\underline{a}$, that is, $a_i \geq 0$ for i=1, 2, ..., n. For example, the initial problem specification in terms of a_i might be redefined in units of an arbitrary norm, say $a_i(N)$, so that $\hat{a}_i = a_i(N) - a_i$ and the new variables $\hat{a}_i$ for i=1, ..., n are unrestricted in sign. In general, quadratic programming problems are a special case of non-linear convex programming and can be solved by reference to the Kuhn-Tucker theorem[1/] which gives necessary and sufficient conditions for an optimal solution vector $\underline{a}*$.

Referring to the general solution for $\underline{a}*$ in equation [3] above we note that this expression can be written simply as

$$\underline{a}* = \underline{\underline{K}}\ \underline{x}\ ,$$

$$\text{or} \qquad a_i^* = \Sigma_{j=1}^n\ k_{ij}\ x_j \quad \text{for } i=1, 2, \ldots, n\ . \qquad [9]$$

That is, the procedures which determine the optimal actions under the quadratic programming problem are simple linear decision rules whose

---

1/ See Kuhn and Tucker [6]. In the discussion by Boot [1] a number of computational algorithms for solving quadratic programming problems are detailed. Several data processing equipment manufacturers have quadratic programming computer codes available based on these algorithms.

arguments are the state variables $x_i \in \underline{x}$, typically not controlled by the decision-maker.

3. Specification of Quadratic Cost Functionals

From the above we see that for decision problems which can be expressed in the form of [1], the general solution is that given by [3], or equivalently [9]. To illustrate this class of problems, and the corresponding issues in estimation, consider the aggregate planning problem first studied by Holt, Modigliani, Muth, and Simon in [2].1/ The decision problem in the HMMS model was to find for an individual firm production and work force levels, P_t and W_t for $t=1, 2, \ldots, T$ periods, that minimize expected total costs $E[C(T)]$, where

$$C(T) = \Sigma_{t=1}^{T} C_t \qquad \text{(the sum of operating costs in each period),} \qquad [10]$$

and

$$C_t = [C_{11} + C_{12} W_t \qquad \text{(regular payroll costs ...} \qquad [10\text{-}1])$$

$$+ C_{21}(W_t - W_{t-1} - C_{22})^2 \qquad \text{(hiring and layoff costs ...} \qquad [10\text{-}2])$$

$$+ C_{31} (P_t - C_{32} W_t)^2 + C_{33}P_t - C_{34}W_t + C_{35}P_tW_t \qquad \text{(overtime costs....} \qquad [10\text{-}3])$$

$$+ C_{61}(I_t - C_{62} - C_{63} S_t)^2] , \qquad \text{(inventory connected costs...} \qquad [10\text{-}4]$$

subject to the restrictions that

$$\left.\begin{array}{l} I_{t-1} + P_t - S_t = I_t \\ P_t \text{ and } W_t \geq 0 \end{array}\right\} \quad t = 1, 2, \ldots, T ; \qquad [11]$$

1/ In subsequent discussion the work by these authors will be referred to as the "HMMS model."

where sales S_t is a stochastic variable with known probability distribution for all t, I_t represents the ending inventory balance for all t, and the cost coefficients c_{ij} are known or can be estimated. By expanding the relations in [10] and regrouping terms we see that [10] is a special case of [4]; similarly, writing the inventory constraint as

$$I_t = I_o + \Sigma_{\tau=1}^{t} P_\tau - \Sigma_{\tau=1}^{t} S_\tau, \quad t=1,2,\ldots,T$$

we see that [11] is a special case of [5]. Hence, given cost coefficient values, the mathematical problem in [10] and [11] can be solved using the previous analysis. The operational problem is then: How can best estimates of the cost coefficients be obtained using available company data? More basically, one might ask: How can the best specification of the cost relations in [10-1] through [10-4] be determined?

For example, referring to the specification for hiring and layoff costs in [10-2], several alternative specifications could have been considered, such as

$$(\text{Hiring and layoff cost})_1 = C_{23} H_t + C_{24} F_t \qquad t=1,2,\ldots,T \quad [12\text{-}1]$$

or

$$(\text{Hiring and layoff cost})_2 = C_{25}(H_t - F_t)^2 + C_{26}(H_t + F_t)^2, t=1,2,\ldots,T \quad [12\text{-}2]$$

where H_t corresponds to workers hired in period t and F_t similarly for workers laid off and

$$W_t = W_o + \Sigma_{\tau=1}^{t} H_\tau - \Sigma_{\tau=1}^{t} F_\tau, \quad t=1, 2,\ldots,T.$$

Alternative specifications might be considered for cost components [10-3] and [10-4] as well.[1/]

Arguments for the specifications of the HMMS model chosen in [10] are detailed in Chapters 2, 3, and 9 of reference [2] and are analyzed further in Van de Panne and Bosje [10]. In the interests of brevity we will not review this discussion here. Suffice it to say that for the company environment analyzed the HMMS model specification is as reasonable as any alternative, and perhaps more preferred. However, the arguments and rationale for this specification may lose appeal when considering a different environment. In this regard, the applied scientist must exploit the statistical properties of his empirical investigation for guidelines.

To illustrate this last point in some detail, we introduce an alternative model based on the HMMS study which was first discussed in Kriebel [4]. Referring to this case as the HLA model, the initial HLA specification is as follows: find non-negative production and work force levels, P_{it} and W_t for $i=1,2,3$ locations and $t=1,2,\ldots,T$ periods which minimize expected total costs $E[C(T)]$ where

$$C(T) = \Sigma_{t=1}^{T} C_t \,, \quad \text{(the sum of operating costs in each period)} \qquad [13]$$

and

$$C_t = [C_{11} + C_{12} L_t \qquad \text{(regular payroll costs ...} \qquad [13\text{-}1]\ )$$

$$+ C_{21}(W_t - W_{t-1} + C_{22})^2 \quad \text{(hiring and layoff costs ...} \qquad [13\text{-}2]\ )$$

$$+ C_{31}(P_t - C_{32}L_t + C_{33})^2 + C_{34} \quad \text{(overtime costs ...} \qquad [13\text{-}3]\ )$$

$$+ C_{41}(W_t - L_t)^2 + C_{51}(P_t - P_{t-1} + C_{52})^2 \quad \text{(other variable production costs...} \qquad [13\text{-}4]\ )$$

1/ The specification for regular payroll cost given in [10-1] is perhaps the most difficult to improve upon, given the ordinary accounting procedures of most firms.

$$+ \Sigma_{i=1}^{3} C_{61i}(I_{it} - C_{62i} - C_{63i} S_{it})^2] \text{ (inventory connected costs at all locations ... [13-5])}$$

subject to the restrictions that

$$\left.\begin{aligned} I_{it} &= I_{it-1} + P_{it} - S_{it} \quad i=1,2,3 \\ L_t &= W_t - r_t \\ P_t &= \Sigma_{i=1}^{3} P_{it} \end{aligned}\right\} \text{ for } t=1,2, \ldots, T \ . \qquad [14]$$

The variable L_t represents the number of direct labor employees actually reporting for work within a particular time period t, and is stochastically determined for each period by W_t and the value of r_t, corresponding to the number of absentees. The subscript i on I_{it}, P_{it} and S_{it} serves to identify three separate locations where inventory is stored and sales transactions occur. With the exception of the overtime cost specification in equation [13-3] and the inclusion of the relation in [13-4], the specification of the HLA model in [13] is directly compatible with the HMMS model in [10]. Equation [13-4], labeled "other variable production costs" consists of two expressions, one corresponding to an absenteeism cost component and the other corresponding to a cost component associated with changing production levels. It is apparent from the preceding discussion that the mathematical problem in [13] and [14] is one of quadratic programming and that the HLA specification can be formulated to provide a solution as given by [9]. For example, if we partition the action vector by time periods, such that $\underline{a}' = (\underline{a}_1, \ldots, \underline{a}_t, \ldots, \underline{a}_T)$ where $\underline{a}_t = (P_{1t}, P_{2t}, P_{3t}, W_t)$ for t=1, 2, ..., T.

We now proceed to an empirical analysis of the HLA model and the determination of estimates for the coefficients of the specification in [13], given historical data on costs and the decision and state variables. To simplify this discussion, however, we will omit consideration of the inventory connected costs component [13-5] in [13], since the added complication introduces no new issues for the empirical analysis. Thus, in subsequent discussion we refer to the HLA model cost specification simply as

$$C(T) = \Sigma_{t=1}^{T} [C_{11} + C_{12}L_t + C_{21} (W_t - W_{t-1} + C_{22})^2 + C_{31} (P_t - C_{32}L_t + C_{33})^2 + C_{34} + C_{41}(W_t - L_t)^2 + C_{51} (P_t - P_{t-1} + C_{52})^2] , \qquad [13b]$$

where all of the previous definitions apply.

4. Single Equation Estimation of Cost Coefficients 1/

In conjunction with the preliminary analysis of the HLA company environment which . to the cost specification above, data was obtained on all variables and costs covering a history of approximately fifty consecutive time periods. As a first approach to obtaining coefficient estimates, a simple linear regression model was hypothesized of the form

$$Y_t = \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdot \cdot \cdot + \beta_7 x_{7t} + \epsilon_t \qquad [15a]$$

for

$$C_t = A + b_1L_t + b_2L_t^2 + b_3 (\Delta W_{t-1}) + b_4(\Delta W_{t-1})^2 + b_5P_t + +b_6 P_t^2 + b_7 (P_tL_t) , \qquad [15b]$$

1/ Computations for the statistical analyses discussed in the following two sections were performed on the CIT G-20 computer with time financed by the Graduate School of Industrial Administration. In this regard, the author acknowledges the programming assistance of Henry Townsend, a graduate student.

where $\Delta W_{t-1} = W_t - W_{t-1}$, and ordinary least squares estimates of the eight coefficients obtained from fifty-two observations. The results of this analysis are summarized in Exhibit 1. The coefficient of multiple

EXHIBIT 1

ORDINARY LEAST SQUARES ANALYSIS OF REGRESSION MODEL 1, EQUATION [15]

Regression Variable	Mean Value*	Regression Coefficient	Least Squares Estimate	Standard Error of Coefficient	t - Statistic
C_t	10,729	--	--	1,128	--
constant	--	A	-5074.5	--	--
L_t	86.3	b_1	-26.63	178.9	0.14
L_t^2	7498.7	b_2	0.544	1.38	0.39
ΔW_{t-1}	0.32	b_3	-24.98	17.79	1.40
$(\Delta W_{t-1})^2$	8.2	b_4	6.65	3.41	1.95
P_t	252.4	b_5	60.52	90.2	0.67
P_t^2	63882.9	b_6	-0.099	0.17	0.58
P_tL_t	21818.4	b_7	0.232	0.59	0.39

*For P_t expressed in 1000's and P_t^2 expressed in millions and L_t, ΔW_{t-1} expressed in units.

determination adjusted for degrees of freedom in this regression was 0.91 with corresponding F-statistic of 74. On the basis of the multiple correlation criterion this model provides good estimates of costs, $\hat{C}_t$, however, inspection of the last column in Table 1 indicates that only the estimate of coefficient b_4 has a significance level greater than 0.9. This result is partially explained by reference to the simple correlation matrix for the data. That is, referring to the simple correlations between the independent variables x_1 to x_7 corresponding to L_t to P_tL_t in the regression model of [15], the following correlations exceed a value of .50:

$$\rho_{12} = \rho(L_t, L_t^2) = .99$$

$$\rho_{17} = \rho_{27} = .88$$

$$\rho_{56} = \rho(P_t, P_t^2) = .99$$

$$\rho_{57} = \rho_{67} = .73$$

We conclude therefore that multicollinearity exists between the first and second degree terms for L_t and P_t in the regression, even though the actual relation between these variables is known a priori to be nonlinear. Further inspection of the data reveals that the source of this difficulty lies in the narrow range of the observations recorded for these variables, viz., the variance to mean ratios for the observations on L_t and P_t are .53 and 1.0, respectively.

If multicollinearity were the only problem in the regression results, we could circumvent the difficulty in this case by applying a linear transformation to the variables effected. That is, consider the regression

$$Z_t = \alpha + \beta_1 Y_{1t} + \beta_2 Y_{2t} + u_t \qquad [16]$$

and assume $Y_{1t} = g_t$ and $Y_{2t} = g_t^2$. Let $y_{1t} = (Y_{1t} - \bar{Y}_1)$ and $y_{2t} = y_{1t}^2$, and consider the alternate regression

$$Z_t = \gamma + \delta_1 y_{1t} + \delta_2 y_{2t} + u_t \ . \qquad [17]$$

The relation between the coefficients in [16] and those in [17] is simply

$$\alpha = \gamma - \delta_1 \bar{Y}_1 + \delta_2 \bar{Y}_1^{\,2} = \gamma - \delta_1 \bar{g} + \delta_2 (\bar{g})^2 \ ,$$

$$\beta_1 = \delta_1 - 2 \bar{Y}_1 \delta_2 = \delta_1 - 2 \bar{g} \delta_2 \ ,$$

$$\beta_2 = \delta_2 \ .$$

Subsequent analyses employing least squares estimates of the regression coefficients can now be implemented based on the results obtained from [17], discarding the initial regression in [16]. Revising the regression in [15] by this procedure gives the following changes to Table 1:

new constant $\equiv \gamma = 10{,}585.4$,

new coefficient for $L_t \equiv \delta_1 = 123.56$, with t statistic 10.3,

new coefficient for $P_t \equiv \delta_5 = 30.53$, with t statistic 5.95,

the remaining coefficient estimates essentially unaltered.

However, multicollinearity is not the only difficulty with the regression model in [15]. A more basic problem concerns the regression specification and subsequent identification of the cost parameters in the original model of [13b] based on the estimated regression coefficients. For example, even if for convenience we assume a priori that cost coefficients c_{41} and c_{51} are identically zero, so that the number of remaining cost coefficients in [13b] equals the number of regression coefficients in [15], the initial cost model is over-identified. That is, referring to the regression model we see that the value of cost coefficient c_{33} can be either 1/2 (b_5/b_6) or $\sqrt{b_2/b_6}$. It can be shown in this case that neither of these estimates based on the least squares regression will be a maximum likelihood "best" estimate of the coefficient value.[1/]

One approach to resolving the identification problem is to introduce restrictions _a priori_ on relations between the admissable values of the regression estimates. Such restrictions could be incorporated into the regression analysis in a variety of ways. For example, the restrictions could be included as equality constraints on the regression parameters, obtaining constrained least squares estimates through a quadratic programming analysis. Alternatively, additional relations between variables could be introduced into either the cost specification or the regression model until an exact correspondence between the coefficients was realized. As in the above example, however, exact identification by this procedure is not always possible.

An allied, though separate, problem with the results obtained in our initial regression analysis concerns the question of admissable values of the coefficient estimates. That is, referring to the cost

1/ An interesting modification of the standard regression procedure which yields maximum likelihood estimates when the parameters in a normal regression model are overidentified has been suggested by Lovell [7]. Basically, Lovell's approach seeks values of the coefficients which minimize the standard error of estimate while applying a search procedure, such as the Fibonacci routine, over the range of values for the overidentified coefficient--in this case, for c_{33}.

specification in [13b] it is clear whatever procedure is used to obtain coefficient estimates, $\hat{c}_{ij}$, we require that the estimated cost equation be non-negative for all positive values of L_t, W_t and P_t. For example, we might require in particular that $\hat{c}_{12} \geq 0$, $\hat{c}_{21} \geq 0$, $\hat{c}_{31}$ and $\hat{c}_{32} \geq 0$, and so on. Similarly, we may possess a priori qualitative information on the range of admissable values for certain coefficient estimates which would be appropriate to include within our analysis in addition to the observed information on the variables. One approach to this problem could be to introduce constraints and proceed as suggested above under a quadratic programming analysis. Another approach has been described by Theil [8] as "mixed estimation." In Theil's framework the initial regression model comparable to [15], say

$$\underline{y} = \underline{\underline{X}}\ \underline{\beta} + \underline{u}\ , \qquad [18]$$

where $\underline{\underline{X}}$ represents the matrix of observational information, is augmented by

$$\begin{bmatrix} \underline{y} \\ \underline{z} \end{bmatrix} = \begin{bmatrix} \underline{\underline{X}} \\ \underline{\underline{R}} \end{bmatrix} \underline{\beta} + \begin{bmatrix} \underline{u} \\ \underline{v} \end{bmatrix} , \qquad [19]$$

where $\underline{\underline{R}}$ represents the matrix of non-observational (qualitative) information. The generalized least-squares estimator of the elements in $\underline{\beta}$ is then

$$\hat{\underline{\beta}} = (\underline{\underline{X}}'\underline{\underline{\Sigma}}^{-1}\ \underline{\underline{X}} + \underline{\underline{R}}'\ \underline{\underline{H}}^{-1}\ \underline{\underline{R}})^{-1}\ (\underline{\underline{X}}'\ \underline{\underline{\Sigma}}^{-1}\ \underline{y} + \underline{\underline{R}}'\ \underline{\underline{H}}^{-1}\ \underline{z})\ , \qquad [20]$$

where $E\ [\underline{u}\ \underline{u}'] = \underline{\underline{\Sigma}}$, $E[\underline{v}\ \underline{v}'] = \underline{\underline{H}}$, and $E\ [\underline{u}\ \underline{v}'] = \underline{\underline{0}}$.

Rather than pursuing these considerations in detail to resolve the difficulties in the initial regression model, we turn our attention now to a different approach for obtaining cost coefficient estimates.

5. Estimation of Simultaneous Equation Systems

Consider again the question of specifying an estimating relationship for operating cost in the HLA model. Ignoring our original specification of period operating costs given by equation [13b], we have simply that total costs

$$C(T) = \Sigma_{t=1}^{T} C_t \qquad \text{(the sum of operating costs in each period)}, \qquad [21]$$

and C_t = [(hiring and layoff costs)$_t$ + (regular payroll costs)$_t$ + (overtime costs)$_t$ + (other variable production costs)$_t$]

$$= C_{1t} + C_{2t} + C_{3t} + C_{4t} \,. \qquad [22]$$

Retaining our earlier definitions on the variables, we might refine this statement of operating costs by stipulating the components as

$$C_{1t} = \alpha_1 + f_1(W_t - W_{t-1}) + u_{1t} \quad \text{(hiring and layoff costs)}, \qquad [22\text{-}1]$$

$$C_{2t} = \alpha_2 + \beta_{21} L_t + u_{2t} \quad \text{(regular payroll costs)}, \qquad [22\text{-}2]$$

$$C_{3t} = \alpha_3 + f_3(L_t, P_t) + u_{3t} \quad \text{(overtime costs)}, \qquad [22\text{-}3]$$

$$C_{4t} = \alpha_4 + f_4(W_t - L_t, P_t - P_{t-1}) + u_{4t} \quad \text{(other variable production costs)}, \qquad [22\text{-}4]$$

where the functions $f_i(\cdot)$ are quadratic in the arguments shown, and the variables u_{it} for i=1, ..., 4 correspond to disturbance terms for which we assume

$$E[u_{it}] = 0 \quad \text{for all } t \text{ and each } i$$

$$E[u_{it}u_{i,t+j}] = \begin{cases} \sigma_i^2 & \text{for } j=0, \text{ all } t \text{ and each } i \\ 0 & \text{for } j\neq 0, \text{ all } t \text{ and each } i \end{cases}$$

and that each u_{it} is independent of the predetermined variables (that is, P_t, W_t, and L_t).1/

Clearly, the three cost components in [22-2], [22-3], and [22-4] will be correlated, since they contain common arguments in the right-hand-side relations. For example, given the linear expression for regular payroll costs in equation [22-2], the expression for overtime costs can be rewritten as

$$c_{3t} = \alpha_5 + f_5\,(c_{2t}, P_t) + u_{3t}$$

and it is apparent that c_{3t} will be influenced by the disturbance term u_{2t}. More generally, if strong association (i.e., high positive correlation) exists between the components in [22], this information becomes lost when coefficient estimates are obtained from a regression employing the aggregated model. That is, the disturbance terms are additive between components and a corresponding increase occurs in the standard errors of the coefficient estimates. Better results can be obtained by refining the estimation procedure to take into account the component relationships, either individually or as a system of equations.

1/ We will also assume zero autocorrelations for the disturbance terms for each i.

From the detailed accounting data available in the firm, an initial estimate of the correlation between cost components--and hence, potential cost specifications--was obtained by the simple regression

$$c_t = \gamma_1 c_{1t} + \gamma_2 c_{2t} + \gamma_3 c_{3t} + \gamma_4 c_{4t} + u_t , \qquad [23]$$

where each of the γ_i assume unit values. On the basis of this analysis, it was decided to group overtime and other variable costs as one component and to consider regular payroll cost as a separate component, since the former showed negligible correlation and the latter indicated high correlation with the remaining costs. For other reasons, principally because a different accounting basis had been employed, it was decided also to treat hiring and layoff costs as a separate cost equation.

From these and earlier considerations a variety of model specifications were considered seriatim for each of the cost component relationships. In the interests of brevity only the final model is presented:[1/]

$$\hat{c}_t = d_1 \hat{c}_{1t} + d_2 c_{2t} + d_3 \hat{c}_{5t} \quad \text{(estimated operating costs for period } t) \qquad [24]$$

where

$$\hat{c}_{1t} = b_1 (H_t + F_t)^2 + b_2 (H_t - F_t)^2 \quad \text{(estimated hiring and layoff costs)} \qquad [24\text{-}1]$$

$$\hat{c}_{2t} = b_3 L_t \quad \text{(estimated regular payroll costs)} \qquad [24\text{-}2]$$

1/ The constant 2.76 appearing in the first term of equation [24-3] corresponds to an independent estimate of the labor productivity coefficient for direct production work force obtained from available data within the firm. An analysis of variance for different levels of the workforce and production accepted the constant variance hypothesis for this figure. Had this estimate not been available an equation could have been added to the model expressing production as a function of the work force level, and the analysis proceed as below.

$$\hat{c}_{5t} = A + b_4 (P_t - 2.76\ L_t)^2 \quad \text{(estimated other variable production costs)} \qquad [24-$$

$$+ b_5 P_t + b_6 P_t L_t + b_7 L_t + b_8 (W_t - L_t)^2 + b_9 (P_t - P_{t-1})^2.$$

Estimates for the coefficients in this model were obtained using ordinary two-stage least squares. That is, first the coefficients for the cost relations expressed by the system of equations in [24-1] to [24-3] which contain only predetermined variables were estimated by taking the least-squares regression of these actual costs on the right-hand-side relationships. The corresponding cost components in the original equation for operating costs were replaced then by their estimated values from the first-stage regression and least squares analysis was again applied to this reformulated relation. 1/ The estimating model obtained by this procedure was:

$$\hat{c}_t = \{17{,}607 + 1.55\ (H_t + F_t)^2 + 1.41\ (H_t - F_t)^2 - 137.7\ L_t \qquad [25]$$

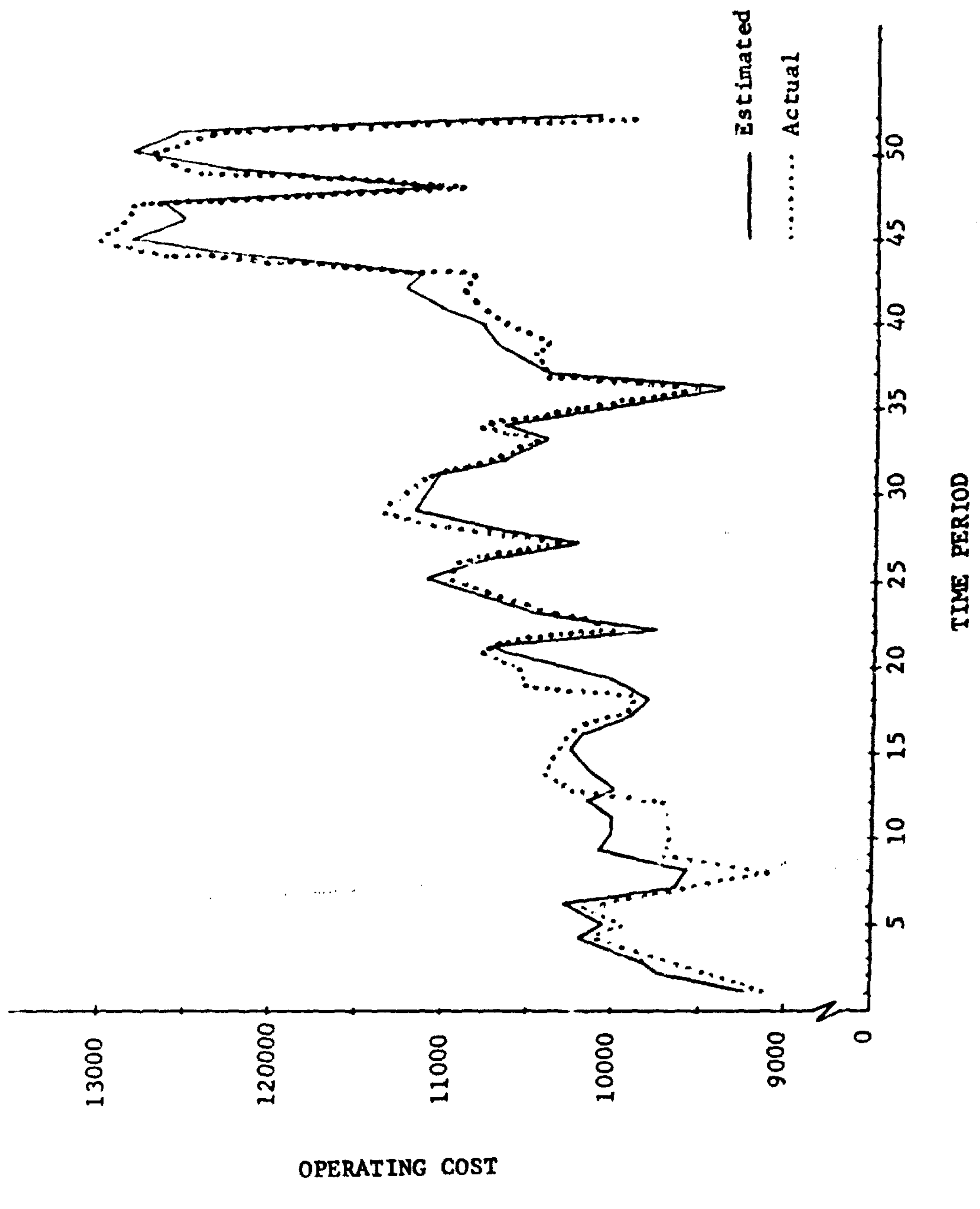
$$+ 0.15\ (P_t - 2.76\ L_t)^2 - 72.22\ P_t + 1.05\ P_t L_t + 0.97\ (W_t - L_t)^2 -$$

$$- 0.26\ (P_t - P_{t-1})^2\}$$

which has an adjusted coefficient of multiple determination of .93 with corresponding F-statistic of 219. The relative goodness-of-fit for the estimating model in equation [25] is illustrated by the graphs of Exhibit 2 which trace actual and estimated operating costs at HLA over fifty-two time periods.

1/ For an excellent discussion of two-stage least squares and other simultaneous equation techniques in econometrics see Theil [8]. An interesting discussion of statistical cost analysis within the framework of economic theory which reports on empirical studies is available in Johnston [3].

EXHIBIT 2: PERIOD COMPARISON OF ACTUAL AND ESTIMATED OPERATING COSTS, HLA MODEL



Within the context of our original problem some additional observations are worth noting. First, our earlier concern with identification of cost coefficients is not an issue in the above model since exact correspondence exists between the regression parameters and model coefficients. Second, the quadratic form which results from the coefficient estimates given in equation [25] is a convex function, and hence the quadratic programming solution presented earlier can be employed directly. Finally, tests for the reasonableness of the estimated parameters in the component cost relationships in general are satisfied. For example, basic economics suggests that the marginal cost of overtime should be positive for increasing production and constant work force levels and conversely negative for increasing work force and constant production. Isolating the estimated overtime cost relationship at the first stage gives

$$\hat{c}_{3t} = 18437 + 0.16\,(P_t - 2.76L_t)^2 - 75.64\,P_t + 1.097\,P_tL_t - 205.16\,L_t \qquad [26]$$

This function is convex for all positive values of P_t and L_t, $(\partial c_{3t}/\partial P_t) > 0$ for L_t constant, and $(\partial c_{3t}/\partial L_t) < 0$ for P_t constant. It is important to include such tests for the reasonableness of estimates obtained by any mechanical procedure, such as least squares, since typically there is no a priori guarantee that the procedure will not indicate nonsense results when the estimated model is literally translated.

6. Conclusion: Sensitivity Analysis and Cost Estimation

In conclusion, we turn our attention to the general question of errors in the specification and estimation of the objective cost function and their consequences. The considerations outlined below are expanded more fully in Kriebel [4] and van de Panne and Bosje [10].

Recall from the discussion of quadratic programming the optimal decisions $\underline{a}^*$ which minimize the objective cost function $c(\underline{a}, \underline{x})$ are given as

$$\underline{a}^* = \underline{\underline{K}}\ \underline{x} = -\underline{\underline{Q}}^{-1}\ \underline{\lambda}\ ,$$

where $c(\underline{a}, \underline{x}) = \lambda_o(a) + 2\ \underline{a}'\ \underline{\lambda}\ (x) + \underline{a}'\underline{\underline{Q}}\ \underline{a}$ and, to simplify notation, $\underline{\lambda} \equiv \underline{\lambda}\ (x)$. The minimum cost associated with the implementation of $\underline{a}^*$ is proportional therefore to

$$c(\underline{a}^*) = -\ \underline{\lambda}\ \underline{\underline{Q}}^{-1}\ \underline{\lambda}\ . \qquad [27]$$

Assuming the coefficients are equal to their estimated values, the decision-maker acts in accordance with $\underline{a}^*$. Since this assumption is not valid generally, the decision-maker commits a decision error. We can consider this decision error resulting from errors in the specification or estimation of the cost coefficients as a perturbation, say $\delta(\underline{a}^*)$, about the optimal actions $\underline{a}^*$. That is, the actual non-optimal decisions, $\underline{a}$, based on coefficient errors can be expressed as

$$\underline{a} = \underline{a}^* + \delta(\underline{a}^*)\ . \qquad [28]$$

This decision function can be evaluated implicitly as the Taylor series expansion

$$\underline{a} = \underline{a}^* + d(\underline{a}^*) + 1/2\ d^2(\underline{a}^*) + \cdots + \frac{1}{N!}\ d^N(\underline{a}^*) + \cdots\ , \qquad [29]$$

provided this series converges. Letting $\underline{R}$ represent the remainder terms for higher order differentials in the series, a second order approximation[1/] of $\delta(\underline{a}^*)$ in terms of the original model is

$$\delta(\underline{a}^*) = - \underline{\underline{Q}}^{-1}(d\underline{\lambda} - d\underline{\underline{Q}}\underline{\underline{Q}}^{-1}\underline{\lambda}) + \underline{\underline{Q}}^{-1} d\underline{\underline{Q}}\underline{\underline{Q}}^{-1} (d\underline{\lambda} - d\, \underline{\underline{Q}}\underline{\underline{Q}}^{-1}\underline{\lambda}) - 1/2\, \underline{\underline{Q}}^{-1} (d^2 \underline{\lambda} - d^2\underline{\underline{Q}}\underline{\underline{Q}}^{-1} \underline{\lambda}) + \underline{R} \,. \qquad [30]$$

The increase in operating cost which results from employing $\underline{a}$ is thus

$$\Delta c(\underline{a}) = \underline{\lambda}'\underline{\underline{Q}}^{-1}\underline{\lambda} + 2\, \underline{b}'\underline{\lambda} + \underline{b}'\underline{\underline{Q}}\, \underline{b} \,, \qquad [31]$$

where the second order approximation of $\delta(\underline{a}^*)$ gives

$$\underline{b} = -\underline{\underline{Q}}^{-1}[d\underline{\lambda} + \underline{\lambda} + 1/2\, (d^2\underline{\lambda} - d^2\, \underline{\underline{Q}}\underline{\underline{Q}}^{-1} \underline{\lambda}) - d\underline{\underline{Q}}\underline{\underline{Q}}^{-1} (\underline{\lambda} + d\underline{\lambda} - d\underline{\underline{Q}}\underline{\underline{Q}}^{-1} \underline{\lambda})] \,.$$

The practical consequence of this analysis is that it provides the empiricist with guidelines to consider when obtaining estimates of the individual cost elements. For example, in the HLA model if we consider the cost specification for $c(\underline{a}, \underline{x})$ as given by equation [13], then for $\underline{\lambda} = \underline{\underline{D}}\, x$, the primary coefficients in $\underline{\underline{D}}$ are $\{c_{12}, c_{31}, c_{32}, c_{33}, c_{51}, c_{61i}, c_{62i}, c_{63i}\}$ and those in $\underline{\underline{Q}}$ are $\{c_{21}, c_{31}, c_{32}, c_{51}, c_{61i}\}$. Clearly, any empirical analysis should focus attention on information pertaining to this second set of coefficients and the estimation of the corresponding cost elements. On the other hand, little or no attention should be devoted to obtaining estimates of the coefficients not included in either sub-set,

---

1/ Reference to the cost specifications considered in the HLA model indicates that the equations are of degree 3 in the cost coefficients so that, in fact, only fourth and higher differentials of $\underline{\underline{Q}}$ and $\underline{\lambda}$ vanish completely.

such as c_{41} in [13], since they will have no bearing on the analytical results. Furthermore, the general analysis of [30] can be greatly simplified if we bypass the simultaneous occurrence of coefficient errors, and consider the consequences of errors in each coefficient individually. For example, again referring to the specification in [13] and restricting consideration to only those coefficients which appear linearly in $\underline{\lambda}$ or $\underline{Q}$, such as c_{12} and c_{21} respectively, the evaluation of [30] simplifies to the first order differentials in $\underline{\lambda}$ and $\underline{Q}$, and the Taylor series expansion is now exact for these coefficients upon substituting their differences, Δc_{ij}, for differentials.

This report has reviewed a number of considerations in statistical cost estimation as these problems relate to empirical studies in operations research. To aid the discussion, the empirical issues were illustrated within the specific context of quadratic programming and a case history was presented. In this regard, the quadratic programming model was selected because its mathematical structure and solution can be stated readily, the estimation of its parameters is a nontrivial problem, and research on applications (such as the HMMS analysis) is available and documented. Although many of the empirical questions have only been outlined, the discussion has helped to point out several conclusions.

First, the implementation of management science models clearly requires proportionate attention to empirical, as well as, formal problems of analysis even when, before the fact, these problem areas may appear to be relatively decoupled. The empirical and formal analyses interact throughout the course of an investigation and serve to reinforce recommendations. For example, in the case analysis the initial model specification of equation [13] expressed decisions in terms of production and work force levels P_t and W_t for $t=1,2,\ldots, T$; however, the empirical results leading to the final specification in equation [25] necessitated reformulating the model in terms of the decisions P_t, H_t, and F_t and adding the definitional constraint: $W_t = W_{t-1} + H_t - F_t$, for $t=1, 2, \ldots, T$.

Second, extension of the formal analysis at the outset can substantially assist in the conduct of the empirical investigation. The preceding discussion on sensitivity analysis serves as a good example of this point. That is, such an analysis beforehand can help to identify the priorities that should be considered in planning the effort to obtain estimates of model parameters and relationships.

Third, within the empirical study, an analysis of sampling errors (such as the covariance matrix for the random disturbances in a regression) provides a natural basis for refining the procedures by which model estimates are obtained, e.g., the rationale leading to the two-stage least squares analysis.

Finally, qualitative information can and should be included within the empirical analysis in addition to available observational data. In this regard, recall the inclusion of an independent estimate for the labor productivity parameter in the final overtime cost specification and the tests on the reasonableness of the derived estimates at the conclusion of the least squares analysis.

As more and more decision making procedures are programmed for electronic computers, and these programs are extended within the firm, the empirical problems of data analysis and estimation will become the increasing concern of the management scientist.

References

[1] J. C. G. Boot, Quadratic Programming (North-Holland Publishing Co. and Rand McNally and Co., 1964).

[2] C. C. Holt, F. Modigliani, J. F. Muth, and H. A. Simon, Planning Production, Inventories, and Work Force (Prentice-Hall, Inc., 1960).

[3] J. Johnston, Statistical Cost Analysis (McGraw-Hill Book Co., Inc., 1963).

[4] C. H. Kriebel, "Optimum Design of Production Decision and Information Systems," Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, Mass. (June, 1964).

[5] C. H. Kriebel, "Information Processing in Programmed Decision Systems," Management Sciences Research Report, Carnegie Institute of Technology (forthcoming, 1965).

[6] H. W. Kuhn and A. W. Tucker, "Nonlinear Programming", pp. 481-492 in J. Neyman (ed.), Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability (University of California, Berkeley and Los Angeles, 1951).

[7] M. C. Lovell, "Sales Anticipations, Planned Inventory Investment, and Realizations," Paper presented to the National Bureau of Economic Research Conference on Investment Behavior in Madison, Wisconsin, June 10-12, 1965.

[8] H. Theil, Economic Forecasts and Policy (Second Revised Edition), (North-Holland Publishing Co., 1961), pp. 232 ff.

[9] H. Theil, "Econometrics and Management Science: Their Overlap and Interaction," Management Science (June, 1965), pp. B-200 to B-212.

[10] C. Van de Panne and P. Bosje, "Sensitivity Analysis of Cost Coefficient Estimates: The Case of Linear Decision Rules for Employment and Production," Management Science (Oct. 1962), pp. 82-107.

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Graduate School of Industrial Administration Carnegie Institute of Technology	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP Not applicable

3. REPORT TITLE

ON STATISTICAL COST ANALYSIS AND OPERATIONS RESEARCH

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report, August, 1965

5. AUTHOR(S) (Last name, first name, initial)

Kriebel, Charles H.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
August, 1965	26	10

8a. CONTRACT OR GRANT NO. NONR, 760(24) b. PROJECT NO. NR 047-048 c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) Management Sciences Research Report No. 50 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) None

10. AVAILABILITY/LIMITATION NOTICES

Releasable without limitation on dissemination

11. SUPPLEMENTARY NOTES None	12. SPONSORING MILITARY ACTIVITY Logistics and Mathematical Statistics Branch Office of Naval Research Washington, D. C. 20360

13. ABSTRACT

Although the operations research or management scientist is both a theoretician and an empiricist, the literature on data analysis problems in operations research is relatively sparse. This report discusses a variety of empirical problems in obtaining statistical estimates of operating costs for a quadratic programming model and presents a case history.

14. KEY WORDS	LINK A ROLE	LINK A WT	LINK B ROLE	LINK B WT	LINK C ROLE	LINK C WT
statistical cost estimation						
quadratic programming						
linear decision rules						
empirical research						
econometrics						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization (*corporate author*) issuing the report.

2*a*. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2*b*. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7*a*. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7*b*. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8*a*. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8*b*, 8*c*, & 8*d*. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9*a*. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9*b*. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers (*either by the originator or by the sponsor*), also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through ________________."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through ________________."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through ________________."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring (*paying for*) the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, roles, and weights is optional.