AD A059561

DIFFERENTIAL WEIGHTING FOR PREDICTION AND DECISION-MAKING STUDIES

SOCIAL SCIENCE RESEARCH INSTITUTE
UNIVERSITY OF SOUTHERN CALIFORNIA

J. Robert Newman

(12) LEVEL II

DDC FILE COPY

D D C
RECEIVED
OCT 6 1978
RECEIVED
B

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

ADVANCED
DECISION TECHNOLOGY
PROGRAM

CYBERNETICS TECHNOLOGY OFFICE
DEFENSE ADVANCED RESEARCH PROJECTS AGENCY
Office of Naval Research • Engineering Psychology Programs

78 08 15 006

TECHNICAL REPORT SSRI 77-1

DIFFERENTIAL WEIGHTING FOR PREDICTION AND DECISION-MAKING STUDIES:

A STUDY OF RIDGE REGRESSION

by

J. Robert Newman

Sponsored by

Defense Advanced Research Projects Agency
ARPA Order No. 3052
Under Subcontract from
Decisions and Designs, Incorporated

August 1977

D D C
RECEIVED
OCT 6 1978
B

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

SOCIAL SCIENCE RESEARCH INSTITUTE
University of Southern California
Los Angeles, California 90007

78 08 15 006

Summary

This paper is another in a series exploring the conditions under which either differential or simple unit weighting of predictor variables in prediction and/or decision studies will be appropriate. Some of the difficulties of applying the ordinary least squares (OLS) regression analysis to practical problems are described and an alternative regression model called ridge analysis (RIDGE) is offered as a substitute to OLS. The trouble with OLS is that when the predictor variables are intercorrelated then the regression coefficients estimated by OLS are often quite deviant from the "true" coefficients. They are often too large in absolute value and the sign of the coefficient can be wrong. The RIDGE solution to this is very simple: just add small positive values to the main diagonal of the correlation matrix depicting the intercorrelations between the predictor variables, and re-estimate the coefficients in the usual manner. The resulting estimates are called ridge estimates and in theory they will be superior to OLS estimates in the sense of producing smaller error in cross validation samples. That is, when OLS and RIDGE estimates are estimated in one sample of data, and then tested on a new sample of data the RIDGE estimates will result in fewer errors of prediction than the OLS estimates.

Several empirical studies were conducted using computer simulated data for various prediction situations. The OLS and RIDGE models were compared as to their efficacy in prediction and both models were compared against the simplest model possible, that of unit weighting (UNIT), in which no weighting is performed; the variables are simply added up and

the sum used for prediction. The results of these studies indicate that OLS and RIDGE, with one exception, always outperformed UNIT with respect to producing smaller errors of prediction and, what is more important, RIDGE always did better than OLS. The one exception in which UNIT did better than OLS and RIDGE is for the case in which all the "true" coefficients are positive, not too far apart, and the sample size is relatively small ($\leq$ 50). This is a very restricted class of conditions. The general conclusion is that UNIT weighting will be appropriate only in unusual situations. Regression models are to be preferred as a way of generating differential weights. Also, the ridge method of estimation (RIDGE) always should be the preferred model over OLS. One practical implication of this is that if an investigator does not have the luxury to do cross validation then RIDGE estimation can be used as a substitute for cross validation.

ACCESSION for
NTIS White Section ☑
DDC Buff Section □
UNANNOUNCED □
JUSTIFICATION ________
BY ________
DISTRIBUTION/AVAILABILITY CODES
Dist. AVAIL. and/or SPECIAL
A

Contents

Figures

Tables

Acknowledgment

This research was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Office of Naval Research under Contract N00014-76-C-0074 under subcontract from Decisions and Designs, Inc.

Disclaimer

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or of the United States Government.

Differential Weighting for Prediction and Decision Making Studies: A Study of Ridge Regression

Introduction

A major problem in prediction and decision studies is how to differentially weight relevant information and form a composite model based on those weights which can then be used to make a decision and/or prediction. The most widely used model for doing this is the multiple regression model. However, this model is often overly complex or leads to the wrong weighting scheme. Many investigators have suggested replacing this model with a simpler one and the simplest model of all is the so called unit weighting model. With unit weighting no attempt is made to estimate what the optimal differential weights might be, instead they are all assigned the same value, namely 1. This paper is another in a series (Newman, Seaver, and Edwards, 1976; Newman, 1976) designed to investigate under what conditions differential weighting is appropriate.

The paper focuses on some of the properties of the regression model that lead to difficulties in its use and how those difficulties can be remedied thus allowing for easier and more productive use of the regression model. In particular I will discuss a modification of the regression model called ridge regression. Before defining what ridge regression is and how it works, however, I will first review briefly some of the basic features of the regression model and its difficulties.

The regression model and its difficulties.

The regression model assumes that some criterion variable Y can be preaicted from a set of predictor variables by forming the linear combination of the N predictor variables, i.e., in vector-matrix notation we write

$$Y' = XB + e \qquad (1)$$

where Y' is a column vector containing N predicted values of another column vector containing the actual values of the criterion, B is a vector of the regression coefficients, and X is a Nxp matrix of N observations on each of p predictor variables, e is the residual vector containing the deviations of the actual values of Y from the predictor values Y'. The vector B, of course, is unknown but assumed to have "true" values in the population from which the observations in Y and X were taken. In ordinary least squares theory (OLST) the vector B is estimated using the least squares principle, i.e., by minimizing the sum of squares of residuals $\sum_{i}^{N} e_i^2 = \sum_{i=1}^{N} (Y_i - Y_i')^2$. If all the variables are transformed into Z scores such that all have 0 mean and unit standard deviation then a well known matrix solution to finding B, the vector of estimated standardized coefficients, is:

$$\hat{B}_j = R_{ij}^{-1} \, r_{yj} \qquad (2)$$

where R_{ij}^{-1} is the inverse of the simple correlation matrix containing the intercorrelation coefficients between each of the predictor variables and r_{yj} is the vector of the correlations between each predictor and the criterion variable (validity coefficients). The use of standardized values and the correlation matrix and validity vector does not restrict generality of results since after the solution is obtained the reverse transformation can also be made to obtain the "raw score" regression coefficients. The transformation should always be made since that places all variables on the same scale no matter what units they were originally expressed in raw score form.

Once the regression coefficients are estimated, the regression model can be used to predict the criterion variable on data that was not used as the estimation data. As a matter of fact if the model is to be used for making practical predictions or decisions such as selecting students into professional schools, this procedure, called cross validation, should be done often to see how well the model works. It is clear that the model will not do as well on new data as it did on the data in which the coefficients were estimated. This is so since the least squares criterion minimizes the error in the estimating sample, much of the predictability thus obtained could be due to chance fluctuations in the sample data. There is no guarantee that the estimated coefficients will stand up well when applied to new data.

If certain conditions pertain in the estimating sample, then the estimated coefficients will not stand up well on cross validation. Some of the conditions that will cause difficulty in initial estimation are:

(a) Poor sampling procedures leading to non-representative sampling.

(b) Small sample size relative to the number of predictors and thus the number of coefficients that need to be estimated.

(c) The presence of measurement error in the variables, in particular measurement error in the criterion variable.

(d) The presence of "outliers" in the data, i.e., data points that lie outside the normal range of the numerical values for the variables.

(e) Intercorrelations between the predictor variables, a condition sometimes referred to as multi-collinearity.

When some or all of these conditions exist in an estimating data sample, the estimated coefficients are often poorly estimated, i.e., are far removed from the true values of the coefficients in the population from which the sample came. Of course any competent investigator will do everything possible to adhere to sound sampling procedures, careful study design, and so on, to control the above conditions as much as possible. However, even in carefully controlled studies they are never completely eliminated and the last condition, multi-collinearity, is often a fact of life and it is often difficult to reduce it. I will show shortly that item (e), the problem of multi-collinearity, is very serious for the regression model leading to very poor estimates of the regression coefficients. Fortunately, its effect can be reduced considerably as I will also show.

Before discussing that, it is of interest to review briefly how investigators, at least in the behavioral science, have reacted to the use of regression models that are known not to yield good predictive results. Many investigators have argued that any differential weighting model such as the regression model should be replaced with the simplest model possible, i.e., a unit weighting model in which no attempt is made to do differential weighting. The predictor variables are just added up and this sum is used to predict the criterion variable. Although this sounds counter-intuitive, there is a long and accumulating body of evidence that such unit weighting may be as good and in some cases better than differential weighting. This evidence has a theoretical and analytic underpinning as provided by the work of Wilks (1938), Gulliksen (1950, Ch. 20), and more recently Einhorn and Hogarth (1975), Wainer (1976), Wainer and Thissen (1976) and Green (1974). There have also been several empirical studies as represented by the works of Lawshe and Shucker (1959), Wesman and Bennett (1959), and Fischer (1972). There have been at least three computer simulation studies (Schmidt, 1971, 1972; Claudy, 1972), and the approach we take is similar to such simulations. In an important review and analysis, Dawes and Corrigan (1974) argue cogently that simple additive (unit weighting) models are quite appropriate and indeed desirable in many decision making situations.

Recently, Newman, Seaver, and Edwards (1976) and Newman (1976) investigated unit versus differential weighting and some of the conditions in which one model might be superior to the other such as sample size and measurement error. Their results strongly indicated that the differential weighting via the regression model was always superior to unit weighting except for small sample sizes. More recently Keren and Newman (1977), arguing that unit weighting will be appropriate only in very restricted conditions, demonstrate that there is a wide class of conditions in which the regression model will always be superior to unit weighting even for small sample sizes.

I am now of the opinion in light of the evidence that unit weighting is rarely appropriate in practical prediction or decision studies. I also believe, however, that it should always be considered as possibly appropriate because of its simplicity. Among other things, it gets rid of the problem of estimating what the appropriate weights should be.

It also relieves the investigator of the need for cross validation. Since nothing is estimated from the data, there is no need for cross validation. My sympathy for unit weighting, when it is appropriate, is echoed in a remark by Ward Edwards who, in the context of applied decision analysis, stated: "...if such an approximation (unit weighting) isn't too bad, what an enormous simplification of elicitation methods it offers us!" (Edwards, 1977, p. 339).

There is another reason for always considering unit weighting. It represents the simplest model possible and therefore, constitutes a base comparison against which all other, more sophisticated models, may be compared. In the studies reported below this is always done.

These studies also use a simulation method, developed by me, and described in detail in Newman, Seaver, and Edwards (1976). I digress briefly to describe this method and the technique used for model comparison.

Data Simulation

The simulation is a Monte Carlo simulation of a multivariate process. The simulation generates a random variable vector $X = (x_1, x_2, ...x_m)$ from a multivariate normal distribution. The program uses as an input a standardized variance-covariance matrix such as that given in Table 1 which depicts the intercorrelations between four variables. In Table 1, variable 4 is the criterion and variables 1 - 3 are the predictors. The program then generates a N x M data matrix with N rows depicting observations, for example persons, and M columns depicting measurements such as psychological tests. The elements of the data matrix represent a

Table 1

Example of a Correlation Matrix used as an input to the Simulation

	Variable			
	1	2	3	4
1	1			
2	.50	1		
3	.43	.45	1	
4	.47	.81	.74	1

random sample from a multivariate normal population having the correlation structure given in Table 1 and thus simulate a "score" for each observation on each of the column variables. Using Table 1 as an input, for example, each row of the simulated data matrix can be considered as a person being considered for admission to medical school. The first three columns of the data matrix could represent the score a person received on three tests being used for selection purposes and the fourth column could represent some criterion for selection such as "success in medical school."

Basis for model comparison.

Mean square error. With data matrices generated by the simulation, prediction models can be formed from the data and their efficacy in predicting can be evaluated. For example, we can compare the regression model against the simple unit weighting model to see which is best. One way to compare two such models is to calculate the mean square error (MSE) for each model and the model with the smallest MSE can be considered best. Since the simulation method enables the generation of as many samples of data as one chooses, this model comparison can be repeated any number of times.

Consider any two models 1 and 2. Then in comparing model 1 against model 2 we can form

$$MSE1 = \sum_{i=1}^{N} e_1^2/DF1 \qquad (3)$$

and

$$MSE2 = \sum_{i=1}^{N} e_2^2/DF2 \qquad (4)$$

where MSEi i=1,2 is the mean square error for model i, DFi is the degrees of freedom for model i, and $\sum_{i=1}^{N} e_i^2$ is the sum of the squared residuals of the ith model.

The relative predictive efficiency of the two models may be assessed by comparing the ratios of their mean squared errors:

$$\frac{MSE1}{MSE2} = \frac{(DF2)\sum_{i=1}^{N} e_1^2}{(DF1)\sum_{i=1}^{N} e_2^2} \qquad (5)$$

which, since the criterion variable is the same in both models and $1-R^2$, is equal to the mean squre error of residuals, 5 is re-expressed in the convenient form

$$\frac{MSE1}{MSE2} = \left(\frac{DF2}{DF1}\right)\frac{(1-R_1^2)}{(1-R_2^2)} \qquad (6)$$

R^2 is the squared multiple correlation. When the ratio in (6) is less than one, model 1 performs more accurately than model 2. When the ratio is greater than one, model 2 will perform better than model 1. If model 2 is the unit weighting model, then there is no loss indegrees of freedom, and DF2 = N the number of observations. Thus, for this case, the ratio $\frac{DF2=N}{DF1} > 1$, therefore favoring unit weighting. This is so since DF1 = N-n-1, where n is the number of regression coefficients estimated for model 1. This is a real advantage for the unit weighting model and should be retained. One of the nice features of unit weighting is that it does not "chew up" degrees of freedom. If model 1 is the regression model, then $(1-R_1^2)/(1-R_2^2) < 1$ on initial fit, thus favoring the regression model. This is not a real advantage and should not be retained. It is not fair to evaluate the effectiveness of a model on the sample data used to estimate that model. It is clear that the model will do best on the data which was used to fit the model. For this reason (6) should only be calculated on cross validation, i.e., on sample data not used for estimation.

The loss function. Another way of comparing models is the expected quadratic loss function $E(L^2)$, defined as:

$$E(L^2) = E\left[\sum_{i=1}^{p} (\hat{b}_i - b_i)^2\right] \qquad (7)$$

where $\hat{b}_i$ is an estimated coefficient and b_i is the "true" coefficient. Of course, in practice the "true" b_i's are not known. With the simulation method described above, however, we can always state what the "true" b_i's are since the input to the simulation can be considered as the true correlation matrix in the population. Thus, equation 2 allows us to calculate the "true" b_i's.

In my opinion the best method of comparing models is the MSE calculated in cross validation since it is a direct measure of how well you will predict on sample data not used for estimation. Forming the ratio of the MSE's as suggested by (6) for two competing models is also direct but can be misleading. With the accuracy of computers,it is possible for (6) to yield a very high percentage favoring one model over another yet the two MSE's averaged over many replications could not be very much different. Also, the loss function of (7), since it is calculated on the sample being used for estimating the b_i's, can favor one model over another but the MSE calculated in the cross validation sample may favor the other model. I will show examples of this condition later.

An example of results: *The effects of measurement error*. As an example of the use of the simulation and the model comparison we present some results,not previously reported, about the effects of different types of measurement error and sample size in comparing the efficacy of unit versus regression models. With Table 1 representing the input matrix to the simulation four sample sizes 25, 50, 75 and 100 were investigated. Also two types of measurement error were added to the criterion variable: completely random or uniform error and Normal or Gaussian error. For the former a value of the random variable defined over the unit interval (0,1) was selected and added to the criterion variable. For the Normal (Gaussian) error a value of the random variable was selected from a normal distribution with 0 mean and standard deviations ranging from .2 to 1.0 and added to the criterion variable. For completeness we included the case of no error being added. Thus we had a four (sample size) by seven (error condition) experimental design. For each of the 28 conditions 100 replications were made. At each replication a regression model was formed and a unit weighting model was formed by simply adding up the three predictor variables. The mean square errors (MSE) were calculated and the ratio of (6) calculated for the purposes of tabulating the number of times the regression model outperformed the unit weighting model (or vice versa). Since the MSE for the regression model will always be less than the MSE for unit weighting on <u>initial fit</u>, (6) was used to compare the two models only on <u>cross validated</u> regression models. This cross validation is accomplished sequentially in the sampling process; that is, the coefficients estimated in sample 1 are used to predict the actual values in sample 2; those estimated in sample 2 are used to predict the values in sample 3, and so on.

The results are presented in Tables 2 and 3. Table 2 presents the percentage of times the regression model outperformed the unit weighting model using the ratio given in (6). Note that the case of no error or

Table 2

Percentage of Times Regression Model Outperforms

Unit Weighting Model for Various Sample

Sizes and Number of Variables with Uniform or Gaussian Measurement Error

Type of Measurement Error

Sample Size	None	Uniform[a]	Gaussian (μ = 0, σ)[b] .2	.4	.6	.8	1.0
25	89	82	86	66	46	34	16
50	98	94	94	92	90	66	56
75	99	98	98	98	92	78	68
100	99	99	99	99	92	88	84

Note: Each percentage is based on 100 replications of each sample size

[a] selected over unit interval (0, 1)

[b] selected from Gaussian distribution with mean 0 and standard deviations ranging from .2 to 1.0

Table 3

Average Squared Error ($\overline{\text{MSE}}$) for the Two Models for Various Sample Sizes (N) and Type of Measurement Error

	N							
	25		50		75		100	
Type of Error	($\overline{\text{MSEU}}$)	($\overline{\text{MSER}}$)	($\overline{\text{MSEU}}$)	($\overline{\text{MSER}}$)	($\overline{\text{MSEU}}$)	($\overline{\text{MSER}}$)	($\overline{\text{MSEU}}$)	($\overline{\text{MSER}}$)
None	.30	.20	.30	.19	.32	.20	.29	.18
Uniform[a]	.35	.26	.36	.26	.36	.26	.35	.24
Gaussian[b]								
.2	.34	.25	.32	.23	.32	.22	.32	.23
.4	.40	.36	.39	.31	.40	.30	.40	.31
.6	.50	.45	.49	.42	.48	.42	.48	.41
.8	.57	.58	.59	.57	.57	.53	.58	.54
1.0	.65	.67	.64	.63	.66	.63	.64	.60

[a] selected over unit interval (0, 1)

[b] selected from Gaussian distribution with mean 0 and standard deviations ranging from .2 to 1.0

uniform or moderate Gaussian error (S.D.=.2) that except for the smallest sample size (N=25), the regression model almost always outperforms the unit weighting model. However, as the Gaussian error increases in severity with increasing values of the standard deviation, the regression model gets progressively worse when compared with unit weighting. Note in particular the line for the smallest sample size (N=25) that for the S.D.=.6, the unit weighting model is actually outperforming the regression model on a percentage basis. When the Gaussian error is most severe (S.D.=1.0), the regression model is not doing well even for fairly large sample size (N=75). Actually error this severe is probably quite unrealistic. However, it does show the vulnerability of the regression model to error. Note that except for the smallest sample size, adding uniform error does not effect the regression model, i.e., it still outperforms the unit weighting model. This is because this is "gentle" error. It has the tendency to move the distribution of values to the right and also tends to flatten the distribution. But it does not create "outliers." Gaussian error with large S.D., on the other hand, tends to push the tails of the distribution out or create "outliers" and this has serious deleterious effect on the efficacy of the regression model since outlying values result in estimated regression coefficients that are far removed from the true coefficients.

Table 3 presents the results using the mean squared error (MSE) as the means for comparing the two models. Since each experimental condition was repeated 100 times, the MSE's were themselves averaged and are referred to as the average mean squared error ($\overline{MSE}$) with ($\overline{MSEU}$) for the unit weighting model and ($\overline{MSER}$) for the regression model. The larger these errors, the poorer the models are performing. The results in Table 3 confirm what has already been stated. Severe error results in poor performance of the regression model when compared to the unit weighting model and this is especially true for smaller sample sizes.

I now turn to the problem of improving the regression model. One improvement on ordinary least squares regression (OLS) is called Ridge regression described in the next section.

Ridge Regression

The problem of multi-collinearity.

Ridge regression was first introduced by Hoerl (1962) who recommended it as a considerable improvement over conventional regression. Hoerl offered ridge regression as a possible solution to a vexing problem in multiple regression. The problem: If there are intercorrelations between the predictor variables (multi-collinearity) in a regression problem, then the conventional least squares estimates of the regression coefficients will often be far removed from the "true" regression coefficients. They can be wrong in absolute value, typically being larger in absolute value than they should be, and the signs of the coefficients can even be wrong. What happens when there are intercorrelations between the predictor variables is that the correlation matrix is ill-conditioned, which in matrix algebra terms means at least one of the eigenvalues of the correlation matrix is close to 0. To illustrate what effect this has, consider the correlation matrix given in Table 4. Table 4 was calculated from the data given in Hoerl's (1962) paper. Hoerl actually postulated a regression model which had the following form

$$Y = 2X_1 + 3X_2 + 5X_3 + 10$$

Note that all the regression coefficients are positive. However, Table 4 indiciates a horribly ill-conditioned matrix with the intercorrelations between the predictor variables being as high or higher than their respective validity coefficients with the criterion (variable 4). What this means in terms of eigenvalues is explained next.

Table 4

Correlation Matrix Used to

Illustrate the Problem of Multi-Collinearity

	Variable			
	1	2	3	4
1	1			
2	.98	1		
3	.94	.92	1	
4	.94	.91	.97	1

Eigenvalues

Given a correlation matrix such as that presented in Table 4, there exists a set of eigenvalues λ such that,

$$|\lambda I - R| = 0$$

where R depicts the correlation matrix, I is the unity matrix, and the vertical lines | | depict the determinant of the matrix. For the simple correlation matrix, depicting the correlations between the predictor variables of Table 4, the above expression would be written as follows:

$$\begin{vmatrix} \lambda_1 - r_{11} & -r_{12} & -r_{33} \\ -r_{21} & \lambda_2 - r_{22} & -r_{23} \\ -r_{31} & -r_{32} & \lambda_3 - r_{33} \end{vmatrix}$$

where r_{ij} is the correlation between variables i and j.

The expansion of this determinant results in a polynomial function $f(\lambda)$ of degree p, the number of predictors, in λ. The equation $f(\lambda)$ is called the characteristic equation of R, the correlation matrix, and its roots λ_1 , λ_2, $\lambda_{p=3}$ in this case,are called the characteristic roots or eigenvalues of R. For an orthogonal matrix each eigenvalue would equal one, and the sum of the eigenvalues would equal p the number of predictor variables. For a nonorthogonal or ill-conditioned matrix of predictor variables, the eigenvalues will not equal one. Instead, some will be greater than one and others very much smaller than one. The worse the the ill-conditioning, the greater the range of values. For example, the eigenvalues for Table 4 are 2.9, .09, and .01 respectively. These sum to p = 3, the number of predictors, but the first eigenvalue 2.9 represents 97% of the total variation. Other important features of eigenvalues are:

$$\text{(a)} \quad \prod_{i=1}^{p} \lambda_i = D$$

The product of the eigenvalues equals the determinant (D) of the matrix which for an orthogonal matrix will be equal to p. The higher the product, the more orthogonal are the predictor variables. The smaller this product and thus the smaller the determinant, the more ill-conditioned the matrix is. And,

$$\text{(b)} \quad \sum_{i=1}^{N} 1/\lambda_i = IC$$

the sum of the reciprocals of the eigenvalues is also an index of matrix ill-conditioning (IC). The higher the degree of intercorrelations between the predictor variables, the smaller some of the eigenvalues will become and therefore the larger the sum of their reciprocals. This has a direct interpretation of what to expect by the degree of non-orthogonality or ill-conditioning since Hoerl and Kennard (1970a) deomonstrated that

$$E(L^2) = E\left[\sum_{i=1}^{p} (\hat{b}_i - b_i)^2\right] = \sigma_e^2 \sum_{i=1}^{p} 1/\lambda_i$$

where as before $E(L^2)$ is the expected quadratic loss using $\hat{b}_i$ as the least squares estimate of b_i, the true coefficient, and σ_e^2 is the error variance. As an illustration of these ideas the determinant (D) of the matrix of intercorrelations of the predictor variables in Table 4 is .003, a value much smaller than p = 3. The sum of the reciprocals of the eigenvalues is 37.15, more than ten times what it would be for an orthogonal matrix. This is a most important point for it indicates the extreme variability of conventional least squares estimates of the regression coefficients when the simple correlation matrix is ill-conditioned.

The ridge method of estimation.

The ridge solution is to reduce ill-conditioning in the simple correlation matrix by the procedure of adding a small positive value (k), typically between 0 and .4, to the main diagonal of the simple correlation matrix. For example, to apply ridge just add a small positive value to the first three diagonal elements of Table 4 and re-estimate the regression coefficients from the correlation matrix using traditional methods. The resulting coefficients are called ridge estimates.[1] The question immediately arises as to what that value of k should be?

The choice of k.

There is now a growing list of methods for "optimally" selecting k. Some of these will be mentioned below. In my opinion, however, it is still desirable that an investigator using any correlation matrix for developing a regression prediction model follow the suggestion of Hoerl and Kennard (1970a) and Hoerl and Kennard (1970b) and display graphically what is called a RIDGE TRACE. In constructing a RIDGE TRACE it is recommended that you start with 0 and increase the value of k, the positive constant, in small increments and plot each set of estimated coefficients as a function of k. To

[1] The term "ridge" was chosen by Professor Hoerl because of its similarity to a type of mathematics called "ridge analysis." In a personal communication, Professor Hoerl had the following to say:

> Ridge analysis originally was developed as a method of interpreting quadratic response functions in p-variables over a bounded domain. The term relates to the technique of tracing paths (ridges and inverted ridges) of ascent and descent as one moves out from the center point. In the special case of regression (an unbounded domain) the only important one is the path of steepest descent from the center point $\underline{b}' = (0, 0, ---0)$, defined by $k = \infty$ and the least squares point $\hat{\beta}$, by $k = 0$. Since the concept of ridge analysis was applied to regression the latter terminology was chosen.

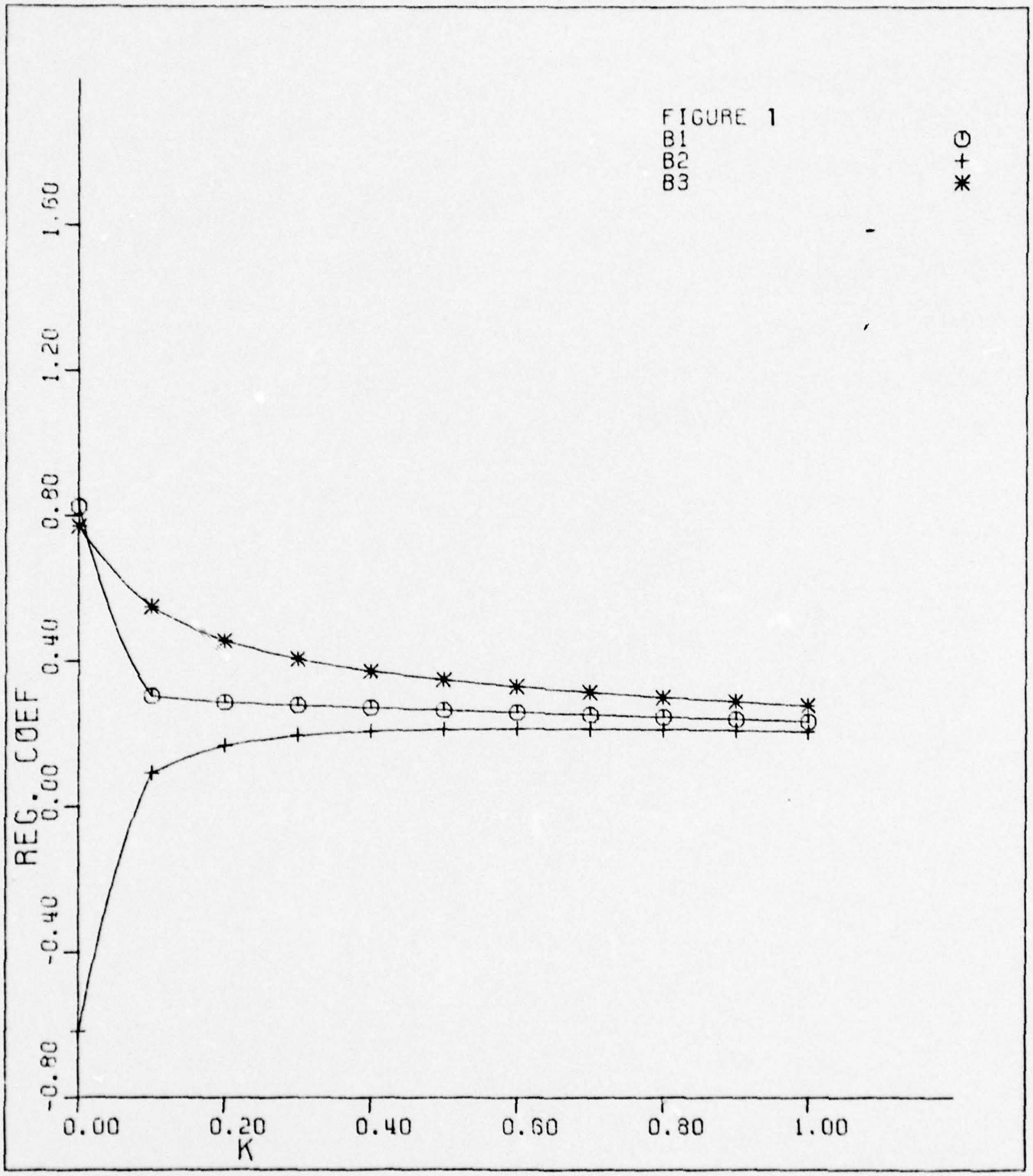


Figure 1. An example of a RIDGE TRACE used to select a value of K.

illustrate, consider Table 4 which is based on Hoerl's (1962) paper. With $k = 0$, the least squares regression coefficients are .824, -.618, and .771, respectively. Note that the second coefficient has a negative sign even though in the true model that Hoerl had formed all coefficients were positive. With any $k>0$ added to the first three diagonal elements of Table 4, we can obtain ridge estimates of these coefficients. For example if $k = .1$, the ridge estimates of the regression coefficients are .303, .091 and .549, respectively. Continuing in this way, adding k in increments of .1 to the main diagonal of the simple correlation matrix and estimating the coefficients each time, we generate a RIDGE TRACE as depicted in Figure 1. Note in Figure 1 the three regression coefficients are plotted as a function of k, the positive constant. When $k = 0$, we have the conventional least squares estimates. For any $k>0$ we have ridge estimates. It is recommended that a value for k should be chosen at that point where the curves for the three coefficients "flatten out." For example, in Figure 1 at $k = .20$ the curves are no long changing much and therefore $k = .20$ should be chosen as the final value. This may seem arbitrary, but there are now algorithms to obtain k from data. Lindley and Smith (1972), arguing within the framework of Bayesian statistics, recommend the value of k be calculated from the data as

$$k = \frac{\sigma_e^2}{\sigma_{\hat{b}}^2} \qquad (8)$$

where σ_e^2 is the error variance (MSE) of the regression model and $\sigma_{\hat{b}}^2$ is the average variance of the regression coefficients. Hoerl Kennard, and Baldwin (1975) recommend:

$$k = p\,\sigma_e^2 / \hat{B}'\,\hat{B} \qquad (9)$$

where p is the number of predictors, $\hat{B}$ is the vector of estimated least squares coefficients, and $\hat{B}'$ is the transpose of $\hat{B}$. Lawless and Wang (1976) recommend:

$$k = 1/F = \frac{\sigma_e^2}{\sigma_{reg}^2} \quad (10)$$

where F is the familiar F ratio, σ_{reg}^2 is the variance due to regression in the least squares solution to regression.

Discussions aboutthe proper choice of k are given by Price (1977) and Dempster, Schatzoff, and Wermuth (1975). I have not yet decided what is the best way to estimate k, if indeed there is a best way. I am investigating this problem and will report the results at a later date.

Properties of ridge estimates.

Ridge estimates of the regression coefficients are biased in the sense that their expected values do not equal the true regression coefficients in the population, i.e.,

$$E(b^*) \neq b$$

where b* is the ridge estimate, b is the true value, and E denotes the expected value. At first blush this seems horrible. Who needs biased estimates? However, it is easy to demonstrate that while ordinary least squares (OLS) estimates are unbiased, they also have much higher standard errors than do ridge estimates. We will show shortly that ridge estimates hold up much better, that is, will result in lower mean-squared-error (MSE) than least squares estimates on cross validation. Since every prediction equation should be cross validated, this has the implication that ridge estimates are to be preferred over least squares estimates. In ridge regression as k increases, the mean-square-error (MSE) for the regression equation in the estimated sample increases. For this reason k is called the bias constant. This is illustrated in Figure 2, which plots MSE calculated from Table 4 as a function of k. However, as k increases, the Variance Inflation Factor (VIF) for each regression coefficient decreases.

An appreciation for what the VIF's can do to increase the variability of estimated coefficients is to consider the well known formula for the standard error of a regression coefficient given as:

$$SE_{b_j} = \sqrt{MSE \cdot C_{jj}} \qquad (11)$$

where SE_{b_j} is the standard error of the regression coefficient b_j, and C_{jj} equals the values in the diagonal of the inverse of the simple correlation matrix. Thus the VIF's are the diagonal elements of the inverse of the simple correlation matrix between the predictor variables. The VIF's for each coefficient in the regression model measures the collective impact of these simple correlations on the variance of the coefficient inthe model. This is depicted in Figure 3 which plots VIF for three regression coefficients calculated from Table 4 as a function of k. Note that the VIF curves drop rapidly and seem to asymptote out for a k = .20.

I should emphasize that things are not always so poorly ill-conditioned as Table 4. For example, the correlation matrix given in Table 1 which is actually more realistic is not too poorly ill-conditioned. The VIF's for Table 1 for OLS (k = 0) are 1.434, 1.466, and 1.349, respectively. Not too bad! Using ridge estimation with the bias constant k being greater than 0, will reduce these VIF's, but there is not too much room for improvement. I will present evidence for this shortly.

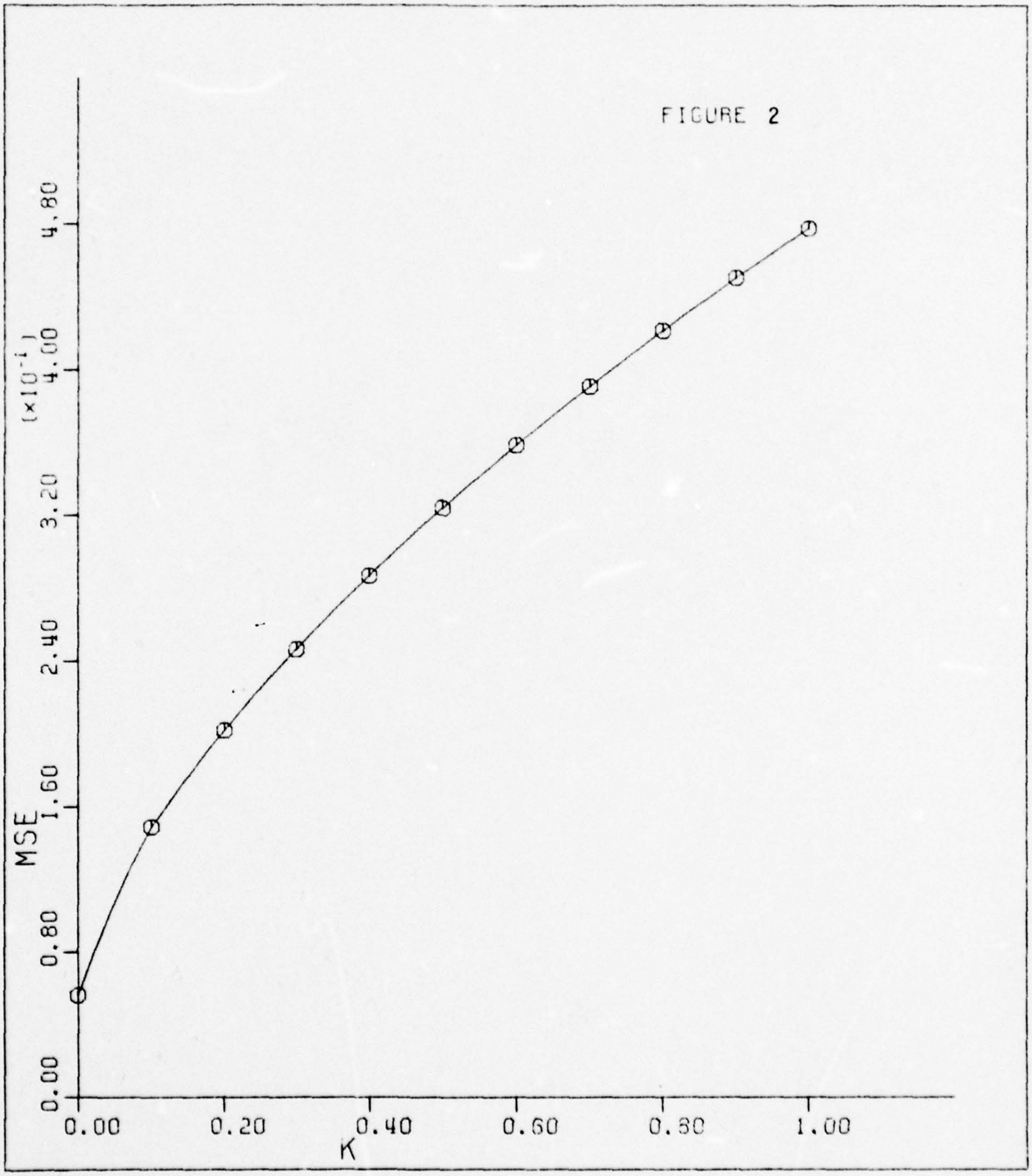


Figure 2. Mean Squared Error (MSE) as Function of K.

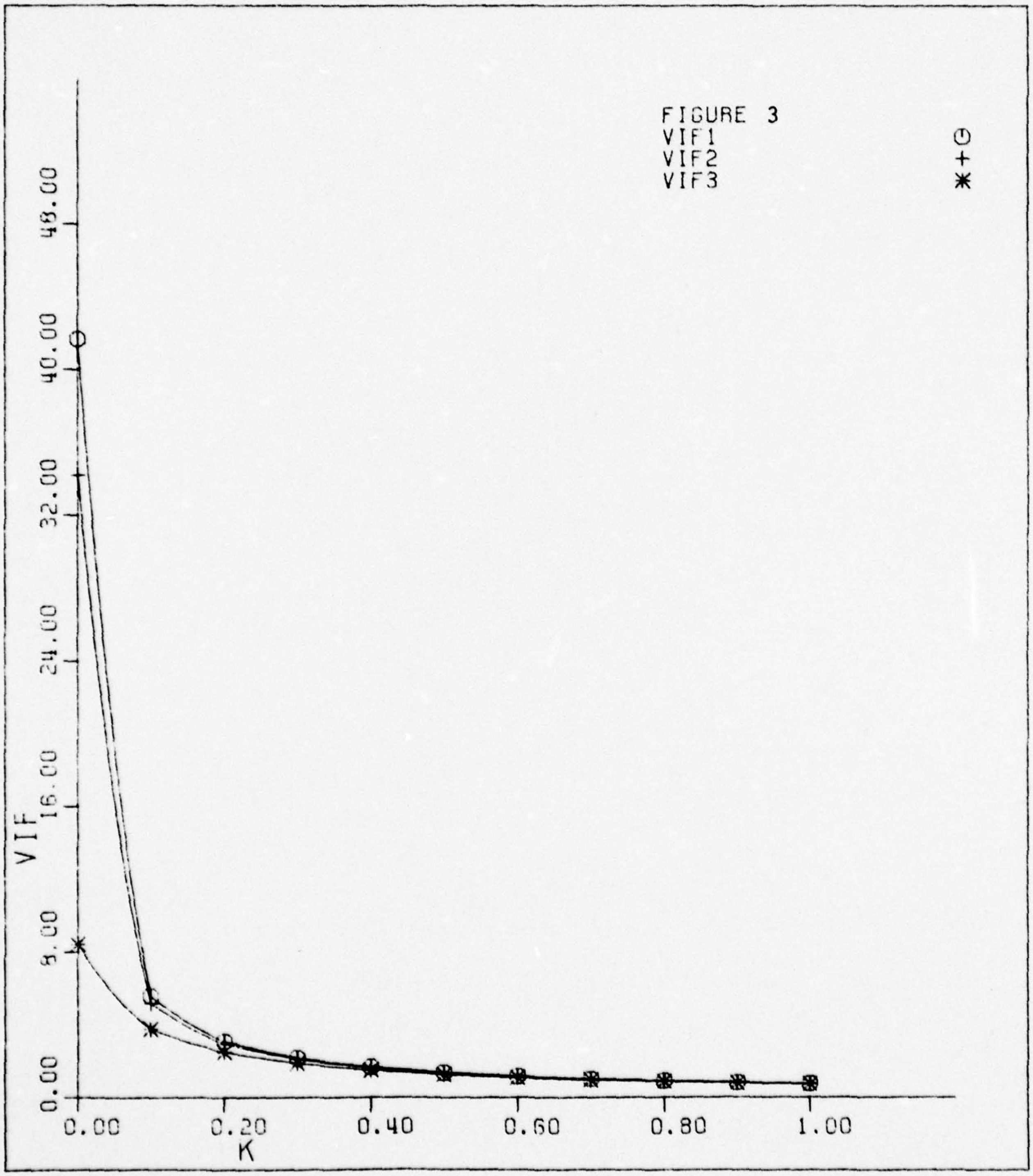


Figure 3. Variance Inflation Factors (VIF) as a Function of K.

Bayesian interpretation of ridge regression.

Several investigators, Hoerl and Kennard (1972), Lindley and Smith (1972), Marquardt and Snee (1975), Dempster, Schatzoff, and Wermuth (1975) have noted that ridge regression fits nicely into Bayesian statistical theory. Concentrating on reducing the VIF's is equivalent to introducing a "tight" prior distribution around the regression coefficients. Ordinary least squares (OLS) regression assumes a relatively flat prior distribution. OLS concentrates on reducing the MSE in the estimating sample. Ridge estimation allow this MSE to be higher than OLS, but in concentrating on reducing the VIF's, ridge estimation is paying attention to how well the prediction equation will do on future samples of data not used for estimation. This is equivalent to having a predictive posterior distribution over the estimated coefficients that is much more precise, i.e., has smaller variance for Ridge estimation than OLS estimation. This is nicely illustrated in the Marquardt and Snee (1975) paper (page 5).

Simulation Studies

In this section I present several studies comparing ridge regression (RIDGE) with ordinary least squares regression (OLS). I will compare both methods of estimation with that of unit weighting (UNIT) or no estimation at all.

A study of sign reversal.

One of the most disturbing aspects of conventional least squares regression is that the sign of the estimated coefficient is the opposite of what it should be. Of course, there is no guarantee that the ridge estimates of the coefficients will yield the correct sign. There is no analytic way, that I know of, that enables one to demonstrate, which estimation procedure will be more apt to yield the correct sign. However, the effects of adding k, the biasing constant to the main diagonal, is to reduce the absolute magnitude of the estimated coefficients. As k gets very large, the overall effect is to drive all the coefficients to zero. This would seem to suggest that the ridge estimates would be less likely to be wrong in sign. The following study was designed to investigate this.

The input matrix.

For this particular study I chose an input matrix the intercorrelational structure of which yields all positive regression coefficients. The matrix is given in Table 5 which is taken from Guilford (1965, p. 395).[2]

[1] I am indebted to McGraw Hill Publishing Company for permission to reproduce Table 5.

Table 5

Intercorrelations among Five Variables

	X_1	X_2	X_3	X_4	Y
X_1					
X_2	.562				
X_3	.401	.396			
X_4	.197	.215	.345		
Y	.465	.583	.546	.365	

N = 174

X_1 = Arithmetic test in Ohio State Psychological Examination

X_2 = Analogy test in the same examination

X_3 = An average grade in high school work

X_4 = Student interest inquiry (measuring breadth of interest)

Y = An average grade for first semester in the university

The correlations in Table 5 are between four predictors of College grades and the criterion was the average grade point average for the first semester of 174 students. The "true" regression coefficients for the four predictors were calculated from Table 5 and are .1039, .3703, .3022 and .1607, respectively.

Design and method of analysis.

Using the simulation method described previously sets of data were generated as if those data came from a population having the correlational structure given in Table 5. Four sample sizes of 25, 50, 100 and 200 were chosen. For each sample size either no error, low error, or high error was added to the dependent variable. The added error was Gaussian and low error was defined as a Gaussian with mean 0, standard deviation of .2 and high error was defined as a Gaussian with standard deviation of .4. Thus we had 4 (sample sizes) X 3 (error levels) = 12 simulated experimental conditions. For each of these conditions the ordinary least squares model (OLS) was fitted to the data, a ridge model(RIDGE) was fitted and a unit weighting model was formed (UNIT). The biasing constant k for fitting RIDGE was estimated from each sample size generated using the Lawless method, i.e., k = 1/F, where F is the F ratio calculated after the conventional least squres regression model was fitted to the data.

For measures of performance three indices were calculated: (a) the mean square error (MSE) calculated on cross validation for each model, (b) the number of sign reversals; i.e., since all the "true" coefficients were positive, every time an estimated coefficient received a negative sign in the estimating sample, a tally was made (this was done for OLS and

RIDGE); and (c) the quadratic loss function, $L^2 = \sum_{i=1}^{p} (\hat{b}_i - b_i)^2$, with $\hat{b}_i$ the estimated coefficient and b_i being the "true" coefficient, was calculated for each estimation procedure OLS and RIDGE.

For each of the simulated conditions, 100 replications were made. Sequential cross-validation was accomplished, i.e., the estimated coefficients on sample 1 were applied to sample 2, those estimated in sample 2 were applied to sample 3, and so on.

Results.

Table 6 presents the average MSE ($\overline{\text{MSE}}$), the average LOSS ($\overline{L}^2$), and total number of sign reversals for the three models OLS, RIDGE, and UNIT where appropriate. There are no sign reversals for the UNIT model. Also for the UNIT model the average LOSS can be calculated once and for all since there is no estimation and $L^2 = \sum_{i=1}^{4} (1 - b_i)^2 = 2.391$. Consider the smallest sample size (N = 25) first. Note that the $\overline{\text{MSE}}$ for UNIT is always smaller than either OLS and RIDGE and therefore UNIT is outperforming the differential weighting models. Note also that RIDGE $\overline{\text{MSE}}$ is always less than the OLS $\overline{\text{MSE}}$ and thus is doing better than OLS. The advantage of RIDGE over OLS is even more dramatic if we use average LOSS ($\overline{L}^2$) as our criterion for comparing the models since $\overline{L}^2$ is considerably smaller for RIDGE than OLS and both RIDGE and OLS are much better than UNIT since it has a huge $\overline{L}^2 = 2.391$ relative to the other two models. This disadvantage is of little consequence for actual prediction however, and we can see why UNIT does so well when we look at the number of sign reversals for OLS and RIDGE. With OLS, for N = 25, these are 47, 50, and 66, respectively, for measurement error ranging from none to high. The total number of sign

Table 6

Average Mean Square Error ($\overline{MSE}$), Average Loss ($\overline{L}^2$), and Number of Sign Reversals for Ordinary Least Squares Regression (OLS), Ridge Regression (RIDGE), and Unit Weighting (UNIT) for Different Sample Sizes and Degree of Error

	Sample Size											
	25			50			100			200		
	ERROR											
	None	Low	High	None	Low	High	None	Low	High	None	Low	High
MSE												
OLS	.560	.593	.651	.576	.598	.593	.538	.559	.607	.525	.551	.596
RIDGE	.546	.564	.604	.538	.592	.588	.533	.556	.587	.525	.551	.596
UNIT	.532	.536	.577	.533	.570	.583	.534	.563	.603	.528	.566	.613
Loss ($\overline{L}^2$)												
OLS	.170	.178	.117	.076	.080	.082	.028	.028	.035	.014	.017	.019
RIDGE	.101	.105	.070	.050	.050	.063	.023	.024	.030	.013	.016	.017
Total Number of[a] Sign Reversals												
OLS	47	50	66	37	43	34	10	20	25	3	11	17
RIDGE	36	32	30	29	33	24	6	16	19	2	10	14

Note: Number of replications = 100

[a] Total number over all four estimated coefficients

reversals for RIDGE are always less than OLS being 36, 32, and 30 for the case of the smallest sample size. The number of sign reversals for UNIT, as already mentioned, is zero. This is the primary reason why UNIT does so well for this case since it nevers gets the sign of the estimated coefficient wrong!

Now consider the larger sample sizes. For N = 50, the UNIT model is still outperforming both OLS and RIDGE in terms of having smaller MSE. RIDGE still outperforms the other two models with respect to $\bar{L}^2$, and outperforms OLS with respect to the number of sign reversals. (UNIT will never lose its advantage in this respect.) Now when we consider the larger sample sizes of 100 and 200, the UNIT model no longer has the distinct advantage over OLS and RIDGE with respect to the prediction criterion of MSE. With N = 100, RIDGE is now superior to both OLS and UNIT. With N = 200, OLS and RIDGE are equivalent with respect to MSE and both are doing better than UNIT. Note also that with respect to $\bar{L}^2$ and number of sign reversals OLS and RIDGE are almost equivalent with RIDGE doing slightly better than OLS. Overall, the results presented in Table 6 clearly indicate that OLS is the worst estimating procedure. It is not surprising that UNIT weighting does so well for the small sample sizes. As mentioned previously, the model is impervious to the vagaries of sampling error that are so prevalent in small samples. It is surprising, at least to me, that UNIT does as well as it does for the larger sample sizes. There is just not much difference between the MSE's for the three models at sample sizes of 100 and 200. For the strong advocates of unit weighting mentioned earlier in this report, these results should be encouraging at least for the case investigated here.

The reader may be curious about how the sign reversals distributed themselves over the four coefficients being estimated. This is presented in Table 7 which displays the number and percent of sign reversals for each estimated coefficient for the different sample sizes and degree of error. Table 7 indicates that sign reversal is restricted for the most part to one or two coefficients in the smaller sample sizes. It is virtually eliminated for the case of sample sizes 100 and 200 with no error and is only present in one coefficient for the cases of high error and the largest sample size (N = 200).

We turn now to a study in which unit weighting does not fare well at all.

Study 2. An example of low variance inflation factor (VIF).

The input matrix. In this study we used as the input to the simulation program the correlation matrix given in Table 1. This is the matrix in which the multi-collinearity is not too bad as indicated by the fact that the variance inflation factors (VIF) are not large. We would not expect under these conditions for RIDGE to be significantly better than OLS. However, the results to be presented shortly will demonstrate that RIDGE and OLS are much superior to UNIT.

Design and method of analysis.

Two sample sizes of 25 and 50 were chosen. Preliminary study indicated nothing to be learned by choosing larger sample sizes. Also, as in the previous study, the degree of error measurement added to the dependent variable increases from none to high in two steps as defined in the previous study. Thus we have 2 X 3 = 6 experimental conditions.

Table 7

Number (%) of Sign Reversals for Each Estimated Regression Coefficient for Different Sample Sizes and Degree of Error for OLS and RIDGE

	Sample Size											
	25			50			100			200		
	ERROR											
	None	Low	High	None	Low	High	None	Low	High	None	Low	High
OLS Coefficients												
$\hat{b}_1$	21	20	36	25	30	22	6	14	21	3	11	17
$\hat{b}_2$	4	3	5	0	0	0	0	0	0	0	0	0
$\hat{b}_3$	6	6	8	0	0	2	0	0	0	0	0	0
$\hat{b}_4$	16	21	17	12	13	10	4	6	2	0	0	0
RIDGE Coefficients												
b_1*	15	14	14	19	21	15	2	10	18	2	10	14
b_2*	1	2	2	0	0	0	0	0	0	0	0	0
b_3*	5	4	2	0	0	0	0	0	0	0	0	0
b_4*	15	12	12	10	12	9	4	6	1	0	0	0

Note: The number of replications = 100. Thus the numbers on the body of the table can be interpreted as frequencies or percentages.

Each experimental condition was replicated 100 times. For each replication, as in the previous study, the OLS, RIDGE were fitted to the data and a UNIT model was formed. For k, the bias constant, a constant value of .15 was chosen and all ridge estimates of the regression ceofficients were made using this value added to the main diagonal of the simple correlation matrix. Preliminary study indicated that k = .15 was close to optimal for this situation. For each replication the MSE was calculated for each model. For OLS and RIDGE this MSE was only calculated on cross-validated samples of data.

Results.

The results are presented in Table 8 using as the criterion for model comparison the average mean square error ($\overline{\text{MSE}}$). As can be seen from Table 8, although the RIDGE model is doing slightly better than OLS for all practical purposes these two models are doing about the same with respect to their respective MSE's. However, both OLS and RIDGE are doing very much better than UNIT.

The reason that UNIT does not do well under this particular condition can perhaps be explained by looking at the "true" regression coefficients which can be calculated from Table 1. These are -.0452, .6155, and .4825, respectively. The coefficient for variable 1 has a negative sign but for all practical purposes is 0. Assigning a value of 1 to this, which UNIT does, is clearly "way off".

The fact that one of the "true" coefficients had a negative value, as in this case, brings up an interesting case in regression analysis. This is the case in which all the correlations between the variables are positive in the population but the structure of these correlations are such that one or more of the regression coefficients can have a negative

Table 8

Average Mean Square Error $(\overline{MSE})$ for the Three Models for Two Sample Sizes and Degree of Measurement Error

	Sample Size					
		25			50	
Error	None	Low	High	None	Low	High
Model						
OLS	.230	.255	.336	.200	.230	.310
RIDGE	.227	.252	.333	.196	.227	.309
UNIT	.303	.328	.392	.306	.325	.368

sign. This is a type of suppressor variable condition called net or negative suppression (Cohen and Cohen, 1975). This type of situation is investigated more extenseively in the next study.

Study 3. A study of suppressor variables.

The classical definition of a suppressor variable is a variable that has zero correlation with the criterion variable but, due to its correlation with other predictors, its inclusion in the regression equation has the effect of raising the multiple correlation coefficient, thus increasing predictability. It accomplishes this, presumably, by "suppressing out" error variance in the predictor variables thus enhancing their ability to predict the criterion. Recently, Conger (1974) and Cohen and Cohen (1975) have noted the possible presence of two other types of suppressor variables. A second type called net or negative suppression occurs when all the correlations between the variables are positive, but one or more of the calculated regression coefficients turn out to be negative. Or to put it another way, if any variable is positively correlated with all predictors and also has a positive validity coefficient, but its regression coefficient turns out to be negative, that variable is serving as a net suppressor. A third type of suppression is called cooperative or reciprocal suppression. This will occur when predictors correlate positively with each other but negatively with the criterion (or, equivalently, the reverse). Another way of discovering cooperative suppression is to note whenever a variable in cooperation with other predictors has a standardized regression coefficient which exceeds in absolute value its validity coefficient but retains the same sign.

It is generally noted, correctly so, that classical and cooperative suppressors are very rare, at least in the behavioral sciences. Net or negative suppressors are not, however, necessarily rare and indeed might be fairly common (Darlington, 1968). It has been suggested by Keren and Newman (1977) that investigators might be able to improve prediction using regression models if they found such variables which could be included in the analysis. In any event the following study was designed to investigate the effects of suppressor variables with respect to comparing OLS, RIDGE, and UNIT models.

Input matrices.

In this study we used three predictor variables and one criterion variable. The simple correlation matrix contained as elements a constant correlation of .50. The correlations of each predictor with the criterion, however, changed in a way to define different types of suppressor variables. As an example, consider the two matrices labeled a and b below:

$$\begin{array}{cccc}
1.00 & & & \\
.50 & 1.00 & & \\
.50 & .50 & 1.00 & \\
.50 & .50 & 0 & 1.00 \\
 & (a) & &
\end{array}
\qquad\qquad
\begin{array}{cccc}
1.00 & & & \\
.50 & 1.00 & & \\
.50 & .50 & 1.00 & \\
.70 & .70 & .30 & 1.00 \\
 & (b) & &
\end{array}$$

With matrix a we have the classical suppressor (variable 3) which correlates zero with the criterion (variable 4). With matrix b, on the other hand, we have net or negative suppressor with variable 3 now correlating .30 with the criterion. Since, in simulation, such matrices can be considered the "true" correlations in the population the "true" regression coefficients can be calculated. These are for matrix a: .5, .5, and -.5, respectively for the three predictors. For matrix b the three regression weights are .35, .35, and -.06 respectively.

Design and method of analysis.

Only one sample size was investigated (N=25). Preliminary investigation indicated nothing to be learned by choosing larger sample sizes for investigation. As in the previous studies we had three error conditions with the degree of error of measurement added to the criterion variable increasing from none to high. Nine different patterns of validity coefficients for the two predictors and the suppressor (variable 3) defined the other experimental condition being manipulated. Examples of two of these patterns are given in matrices a and b above. The remaining seven are given in the results. Each experimental condition was replicated 50 times. For each replication, the OLS, RIDGE, and UNIT models were formed. The choice of k, the bias constant for ridge estimation, was chosen by the Lawless method as in study 1, i.e., k = 1/F, where F is the F ratio calculated after the original least squares model is fitted to the data.

The MSE was calculated for all three models for each replication. As in the previous studies the MSE for OLS and RIDGE was calculated only on cross validated samples and the cross validation was accomplished in the same manner as in the previous studies.

Results.

Table 9 presents the average mean square error ($\overline{MSE}$)for UNIT, OLS, and RIDGE as a function of the validities of both the suppressor (variable 3) and the two predictors. The results in Table 9 are for the high error condition only. This was the condition for which the simplest model UNIT did the best, i.e., the $\overline{MSE}$ for UNIT was the smallest for this condition. The major result and conclusion evident in Table 7 thus is only strengthened by considering the low and no measurement error conditions.

Note that with the exception of one case (Ex. No. 3) OLS is doing better than UNIT but, what is more important, RIDGE is outperforming both OLS and UNIT by a considerable margin. It seems safe to conclude at least for the conditions investigated here that unit weighting will not be appropriate when a suppressor variable is present. Also while OLS does better than UNIT we can also conclude that RIDGE should be used instead of OLS since it does best of all.

Discussion

On the issue of equal or unit versus differential weighting, the results reported in this paper lend strong evidence on the side of differential weighting. If unit weighting is to be compared to some model designed to produce optimal differential weights such as multiple regression then it is the rare case when unit weighting will do as well or better than the regression model. This assumes, however, that there is a well defined criterion variable available for prediction and that

Table 9

Average Mean Square Error ($\overline{MSE}$) for UNIT, OLS and RIDGE as a Function of the Validities of Both the Suppressor (Variable 3) and the Other Predictors

	Validities			$\overline{MSE}$		
Ex.No.	r_{14}	r_{24}	r_{34}	UNIT	OLS	RIDGE
1.	.5	.5	0	.858	.642	.596
2.	.5	.5	.1	.835	.714	.665
3.	.5	.5	.3	.775	.790	.730
4.	.6	.6	0	.798	.455	.418
5.	.6	.6	.1	.771	.546	.501
6.	.6	.6	.3	.701	.654	.593
7.	.7	.7	0	.719	.222	.206
8.	.7	.7	.1	.691	.332	.304
9.	.7	.7	.3	.691	.475	.433

Note: Each $\overline{MSE}$ is based on 50 replications. Sample size was N=25, and high error was added to the criterion variable.

the assumptions of the regression model are not grossly violated. Poor sampling procedures, large amounts of measurement error, two few data points relative to the number of predictors, multi-collinearity are some of the conditions that can degrade the ordinary least squares model and thereby make unit weighting look good by comparison. Also, if all the "true" regression coefficients should have a positive sign then unit weighting can outperform regression models, since the latter have a tendency to assign the wrong, i.e., negative sign, in estimation. This is particularly true for small sample sizes.

However, it seems that the strong proponents of unit weighting such as Dawes and Corrigan (1974) and Wainer (1976) have overstated their case. In particular, Wainer, in giving a proof of a so-called equal weights theorem, states that in many circumstances "almost no loss in accuracy" will obtain when least squares coefficients are replaced by equal weights. Laughlin (1977) refutes this strong statement and also demonstrated that the Wainer paper had a serious error. Laughlin showed that the loss in explained variance by substituting equal weights for optimal regression weights is twice as great as Wainer concluded. Still, there will be occasions when equal or unit weights might be appropriate. In a thoughtful paper Einhorn and Hogarth (1975) have provided guidelines to follow to determine when ordinary least squares weights may be replaced by equal weights.

This paper has also demonstrated that the ordinary least squares (OLS) model can be improved upon considerably. Ridge regression, herein called the RIDGE model, always outperformed OLS in the studies reported here. These results are in complete agreement with those reported by Lawless and Wang (1976), Hoerl, Kennard, and Baldwin (1976), Price (1977), and Dempster, Schatzoff, and Wermuth (1976). The Dempster et al.

paper is particularly interesting in the sense that it reports on an investigation of 56 alternatives to OLS and the general conclusion was that RIDGE was the best regression procedure. It is not true that RIDGE will always outperform OLS. However, I am willing to make the following statement: with proper choice of k, the bias constant, RIDGE will always be better than OLS when there are intercorrelations between the predictor variables and will do as well as OLS when the intercorrelations between the predictors are zero or near zero. On the basis of what has been discovered so far, it is now clear that OLS should not, in general, be used by behavioral scientists.

My strong recommendation of RIDGE over OLS has a practical as well as theoretical and empirical basis. In practical prediction and decision making studies we do not always have the luxury of cross validating the model we are using for prediction and/or decision. The results presented here and elsewhere, however, have demonstrated that RIDGE is very robust under cross validation and certainly much more robust than OLS. Thus if you are using a regression model and you do not have the time, energy or data to do cross validation then you will be safer, more conservative, and are apt to be closer to being "correct" in applying the regression model if you use ridge estimation of the coefficients. Thus if you are in a bind and cannot cross validate use RIDGE as a substitute for cross validation.

References

Claudy, J.G. A comparison of five variable weighting procedures. Educational and Psychological Measurement, 1973, 32, 311-322.

Cohen, J. and Cohen, P. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. New York, Wiley, 1975.

Conger, A.J. A revised definition for suppressor variables: A guide to their identification and interpretation. Educational and Psychological Measurement, 1974, 34, 35-46.

Darlington, R.B. Multiple regression in psychological research and practice. Psychological Bulletin, 1968, 69, 161-182.

Dawes, R.M. and Corrigan, B. Linear models in decision making. Psychological Bulletin, 1974, 81, 95-106.

Dempster, A.P., Schatzoff, M., and Wemuth, N. A simulation study of alternatives to ordinary least squares (Report No. S-35). Cambridge, Mass.: Harvard University, Department of Statistics, June 1975.

Edwards, W. Social Utilities. In Decision and Risk Analysis: Powerful New Tools for Management. Proceedings of the Sixth Triennial Symposium, June 1971, Hoboken: The Engineering Economist, 1972, 119-129.

Einhorn, H.J. and Hogarth, R.M. Unit weighting schemes for decision making. Organizational Behavior and Human Performance, 1975, 13, 171-192.

Fisher, G.W. Four methods for assessing multiattribute utilities: An experimental validation. Engineering Psychology Laboratory, University of Michigan, 1972.

Green, B.J. Parameter sensitivity in multivariate methods. Mimeographed manuscript. Baltimore: Department of Psychology, Johns Hopkins University, 1974.

Guilford, J.P. Fundamental Statistics in Psychology and Education. New York, McGraw Hill, 1965.

Gulliksen, H. Theory of Mental Tests. New York: Wiley, 1960.

Hoerl, A.E. Application of ridge analysis to regression problems. Chemical Engineering Progress, 1962, 58, 54-59.

Hoerl, A.E., Kennard, R.W. Ridge regression: Biased estimation for non-orthogonal problems. Technometrics, 1970a, 72, 55-67.

Hoerl, A.E. and Kennard, R.W. Ridge regression: Application to nonorthogonal problems. Technometrics, 1970b, 12, 69-82.

Hoerl, A.E., Kennard, R.W., and Baldwin, K. Ridge regression: Some simulations. Communications in Statistics, 1975, 4, 105-123.

Keren, C. and Newman, J.R. Additional considerations with respect to multiple regression and equal weighting. Unpublished paper available from the authors, Social Science Research Institute, University of Southern California.

Laughlin, J.E. Comments on estimating coefficients in linear models: It don't make no nevermind. Psychological Bulletin, in press.

Lawless, J.J. and Wang, P. A simulation study of some ridge and other regression estimators. Communications in Statistics, 1976, 4, 307-323.

Lawshe, C.H., and Schucker, R.E. The relative efficiency of four test weighting methods in multiple regression. Educational and Psychological Measurement, 1959, 19, 103-144.

Lindley, P.V. and Smith, A.F.M. Bayes estimates for the linear model. Journal of the Royal Statistical Society, Series B, 1972, 34, 1-41.

Marquardt, D.W., and Snee, R.D. Ridge regression in practice. American Statistician, 1975, 29, 3-20.

Newman, J.R. Differential weighting in multiattribute utility measurement: When it should not and when it does make a difference. Social Science Research Institute Technical Report, University of Southern California,SSRI 76-6, 1976.

Newman, J.R., Seaver D., and Edwards, W. Unit versus differential weighting schemes for decision making: A method of study and some preliminary results. University of Southern California, Social Science Research Institute Research Report, SSRI 76-5, 1976.

Price, M. Ridge regression: Application to nonexperimental data. Psychological Bulletin, 1977, 84, 759-766.

Schmidt, R.L. The relative efficiency of regression and simple unit predictor weights in applied differential psychology. Educational and Psychological Measurement, 1971, 31, 699-714.

Wainer, H. Estimating coefficients in linear models: It don't make no nevermind. Psychological Bulletin, 1976, 83, 213-217.

Wainer, H., and Thissen, D. Three steps towards robust regression. Psychometrika, 1976, 41, 9-33.

Wesman, A.G. and Bennett, G.K. Multiple regression vs. simple addition of scores in prediction of college grades. Educational and Psychological Measurement. 1959, 19, 243-246.

Wilks, S.S. Weighting systems for linear functions of correlated variables when there is no dependent variable. Psychometrika, 1938, 3, 23-40.

CONTRACT DISTRIBUTION LIST
(Unclassified Technical Reports)

Director — 2 copies
Advanced Research Projects Agency
Attention: Program Management Office
1400 Wilson Boulevard
Arlington, Virginia 22209

Office of Naval Research — 3 copies
Attention: Code 455
800 North Quincy Street
Arlington, Virginia 22217

Defense Documentation Center — 12 copies
Attention: DDC-TC
Cameron Station
Alexandria, Virginia 22314

DCASMA Baltimore Office — 1 copy
Attention: Mr. K. Gerasim
300 East Joppa Road
Towson, Maryland 21204

Director — 6 copies
Naval Research Laboratory
Attention: Code 2627
Washington, D.C. 20375

Office of Naval Research — 6 copies
Attention: Code 102IP
800 North Quincy Street
Arlington, Virginia 22217

Decisions and Designs, Incorporated — 20 copies
8400 Westpark Drive, Suite 600
McLean, Virginia 22101

Revised December 1977

SUPPLEMENTAL DISTRIBUTION LIST
(Unclassified Technical Reports)

Department of Defense

Director of Net Assessment
Office of the Secretary of Defense
Attention: MAJ Robert G. Gough, USAF
The Pentagon, Room 3A930
Washington, DC 20301

Assistant Director (Net Technical Assessment)
Office of the Deputy Director of Defense
Research and Engineering (Test and
Evaluation)
The Pentagon, Room 3C125
Washington, DC 20301

Assistant Director (Environmental and Life
Sciences)
Office of the Deputy Director of Defense
Research and Engineering (Research and
Advanced Technology)
Attention: COL Henry L. Taylor
The Pentagon, Room 3D129
Washington, DC 20301

Director, Defense Advanced Research
Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209

Director, Cybernetics Technology Office
Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209

Director, ARPA Regional Office (Europe)
Headquarters, U.S. European Command
APO New York 09128

Director, ARPA Regional Office (Pacific)
Staff CINCPAC, Box 13
APO San Francisco 96610

Dr. Don Hirta
Defense Systems Management School
Building 202
Ft. Belvoir, VA 22060

Chairman, Department of Curriculum
Development
National War College
Ft. McNair, 4th and P Streets, SW
Washington, DC 20319

Defense Intelligence School
Attention: Professor Douglas E. Hunter
Washington, DC 20374

Vice Director for Production
Management Office (Special Actions)
Defense Intelligence Agency
Room 1E863, The Pentagon
Washington, DC 20301

Command and Control Technical Center
Defense Communications Agency
Attention: Mr. John D. Hwang
Washington, DC 20301

Department of the Navy

Office of the Chief of Naval Operations
(OP-951)
Washington, DC 20450

Office of Naval Research
Assistant Chief for Technology (Code 200)
800 N. Quincy Street
Arlington, VA 22217

Office of Naval Research (Code 230)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Naval Analysis Programs (Code 431)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Operations Research Programs (Code 434)
800 North Quincy Street
Arlington, VA 22217

Office of Naval Research
Information Systems Program (Code 437)
800 North Quincy Street
Arlington, VA 22217

Director, ONR Branch Office
Attention: Dr. Charles Davis
536 South Clark Street
Chicago, IL 60605

Director, ONR Branch Office
Attention: Dr. J. Lester
495 Summer Street
Boston, MA 02210

Director, ONR Branch Office
Attention: Dr. E. Gloye
1030 East Green Street
Pasadena, CA 91106

Director, ONR Branch Office
Attention: Mr. R. Lawson
1030 East Green Street
Pasadena, CA 91106

Office of Naval Research
Scientific Liaison Group
Attention: Dr. M. Bertin
American Embassy - Room A-407
APO San Francisco 96503

Dr. A. L. Slafkosky
Scientific Advisor
Commandant of the Marine Corps (Code RD-1)
Washington, DC 20380

Headquarters, Naval Material Command
(Code 0331)
Attention: Dr. Heber G. Moore
Washington, DC 20360

Head, Human Factors Division
Naval Electronics Laboratory Center
Attention: Mr. Richard Coburn
San Diego, CA 92152

Dean of Research Administration
Naval Postgraduate School
Attention: Patrick C. Parker
Monterey, CA 93940

Naval Personnel Research and Development
Center (Code 305)
Attention: LCDR O'Bar
San Diego, CA 92152

Navy Personnel Research and Development
Center
Manned Systems Design (Code 311)
Attention: Dr. Fred Muckler
San Diego, CA 92152

Naval Training Equipment Center
Human Factors Department (Code N215)
Orlando, FL 32813

Naval Training Equipment Center
Training Analysis and Evaluation Group
(Code N-OOT)
Attention: Dr. Alfred F. Smode
Orlando, FL 32813

Director, Center for Advanced Research
Naval War College
Attention: Professor C. Lewis
Newport, RI 02840

Naval Research Laboratory
Communications Sciences Division (Code 54(
Attention: Dr. John Shore
Washington, DC 20375

Dean of the Academic Departments
U.S. Naval Academy
Annapolis, MD 21402

Chief, Intelligence Division
Marine Corps Development Center
Quantico, VA 22134

Department of the Army

Alan H. Curry
Operations and Management Science Divisio
U.S. Army Institute for Research in Manag
ment Information and Computer Science
730 Peachtree St., N.E. (Suite 900)
Atlanta, Georgia 30308

THIS PAGE IS BEST QUALITY PRACTICABLE
FROM COPY FURNISHED TO DDC

Deputy Under Secretary of the Army
(Operations Research)
The Pentagon, Room 2E621
Washington, DC 20310

Director, Army Library
Army Studies (ASDIRS)
The Pentagon, Room 1A534
Washington, DC 20310

U.S. Army Research Institute
Organizations and Systems Research Laboratory
Attention: Dr. Edgar M. Johnson
5001 Eisenhower Avenue
Alexandria, VA 22333

Director, Organizations and Systems
Research Laboratory
U.S. Army Institute for the Behavioral
and Social Sciences
1300 Wilson Boulevard
Arlington, VA 22209

Technical Director, U.S. Army Concepts
Analysis Agency
8120 Woodmont Avenue
Bethesda, MD 20014

Director, Strategic Studies Institute
U.S. Army Combat Developments Command
Carlisle Barracks, PA 17013

Commandant, Army Logistics Management Center
Attention: DRXMC-LS-SCAD (ORSA)
Ft. Lee, VA 23801

Department of Engineering
United States Military Academy
Attention: COL A. F. Grum
West Point, NY 10996

Commanding General
Headquarters, DARCOM
Attention: DRCED - Richard Murray
5001 Eisenhower Avenue
Alexandria, VA 22333

Marine Corps Representative
U.S. Army War College
Carlisle Barracks, PA 17013

Chief, Studies and Analysis Office
Headquarters, Army Training and Doctrine
Command
Ft. Monroe, VA 23351

Commander, U.S. Army Research Office
(Durham)
Box CM, Duke Station
Durham, NC 27706

Department of the Air Force

Assistant for Requirements Development
and Acquisition Programs
Office of the Deputy Chief of Staff for
Research and Development
The Pentagon, Room 4C331
Washington, DC 20330

Air Force Office of Scientific Research
Life Sciences Directorate
Building 410, Bolling AFB
Washington, DC 20332

Commandant, Air University
Maxwell AFB, AL 36112

Chief, Systems Effectiveness Branch
Human Engineering Division
Attention: Dr. Donald A. Topmiller
Wright-Patterson AFB, OH 45433

Director, Advanced Systems Division
(AFHRL/AS)
Attention: Dr. Gordon Eckstrand
Wright-Patterson AFB, OH 45433

Commander, Rome Air Development Center
Attention: Mr. John Atkinson
Griffis AFB
Rome, NY 13440

IRD, Rome Air Development Center
Attention: Mr. Frederic A. Dion
Griffis AFB
Rome, NY 13440

HQS Tactical Air Command
Attention: LTCOL David Dianich
Langley AFB, VA 23665

Other Government Agencies

Chief, Strategic Evaluation Center
Central Intelligence Agency
Headquarters, Room 2G24
Washington, DC 20505

Director, Center for the Study of
Intelligence
Central Intelligence Agency
Attention: Mr. Dean Moor
Washington, DC 20505

Mr. Richard Heuer
Methods & Forecasting Division
Office of Regional and Political Analysis
Central Intelligence Agency
Washington, DC 20505

Office of Life Sciences
Headquarters, National Aeronautics and
Space Administration
Attention: Dr. Stanley Deutsch
600 Independence Avenue
Washington, DC 20546

Other Institutions

Department of Psychology
The Johns Hopkins University
Attention: Dr. Alphonse Chapanis
Charles and 34th Streets
Baltimore, MD 21218

Institute for Defense Analyses
Attention: Dr. Jesse Orlansky
400 Army Navy Drive
Arlington, VA 22202

Director, Social Science Research Institute
University of Southern California
Attention: Dr. Ward Edwards
Los Angeles, CA 90007

Perceptronics, Incorporated
Attention: Dr. Amos Freedy
6271 Variel Avenue
Woodland Hills, CA 91364

Stanford University
Attention: Dr. R. A. Howard
Stanford, CA 94305

Director, Applied Psychology Unit
Medical Research Council
Attention: Dr. A. D. Baddeley
15 Chaucer Road
Cambridge, CB 2EF
England

Department of Psychology
Brunel University
Attention: Dr. Lawrence D. Phillips
Uxbridge, Middlesex UB8 3PH
England

Decision Analysis Group
Stanford Research Institute
Attention: Dr. Miley W. Merkhofer
Menlo Park, CA 94025

Decision Research
1201 Oak Street
Eugene, OR 97401

Department of Psychology
University of Washington
Attention: Dr. Lee Roy Beach
Seattle, WA 98195

Department of Electrical and Computer
Engineering
University of Michigan
Attention: Professor Kan Chen
Ann Arbor, MI 94135

Department of Government and Politics
University of Maryland
Attention: Dr. Davis B. Bobrow
College Park, MD 20747

Department of Psychology
Hebrew University
Attention: Dr. Amos Tversky
Jerusalem, Israel

Dr. Andrew P. Sage
School of Engineering and Applied Science
University of Virginia
Charlottesville, VA 22901

Professor Raymond Tanter
Political Science Department
The University of Michigan
Ann Arbor, MI 48109

Professor Howard Raiffa
Morgan 302
Harvard Business School
Harvard University
Cambridge, MA 02163

Reprinted from:

Judgment and Decision in Public Policy Formation

Edited by Kenneth R. Hammond

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher.

Copyright ©1978 by the American Association for the Advancement of Science

Published in 1978 in the United States of America by Westview Press, Inc.

AAAS Selected Symposium 1

Summary

In summary this paper indicates some of the basic conditions of public policy research and some of the misconceptions or institutional constraints on doing such public service. The paper offers optimistic advice in that the understanding of these conditions and constraints will facilitate useful good work. There are enormous opportunities to benefit the public through the successful address of the policy issues. While you may not know how to do it now, with the proper mind set there is a good chance you can learn and can be effective in working the vague, shifting, uncertain policy terrain. It may, therefore, be appropriate to end this advisory essay to those who would give advice with A Garland of Precepts by Phyllis McGinley:

Though a seeker since my birth.
Here is all I've learned on earth,
This the gist of what I know:
Give advice and buy a foe.
Random truths are all I find
Stuck like burs about my mind.
Salve a blister. Burn a letter.
Do not wash a cashmere sweater.
Tell a tale but seldom twice.
Give a stone before advice.

Pressed for rules and verities,
All I recollect are these:
Feed a cold to starve a fever.
Argue with no true believer.
Think-too-long is never-act.
Scratch a myth and find a fact.
Stitch in time saves twenty stitches.
Give the rich, to please them, riches.
Give to love your hearth and hall.
But do not give advice at all.

NOTE

Several people have usefully commented on earlier drafts of this paper, their successes in my education can be best attributed to their insightful and salient criticism. Their failures are due to my opacity and intransigence.

These benefactors include John Gilmore, Jack Nilles, Dennis Little, Don Kash, Jack White, Dennis Miller, Dorothy Nelkin, Anton Schmalz and Vary Coates.

4

Technology for Director Dubious: Evaluation and Decision in Public Contexts

Ward Edwards

In preparing this paper, I had the enormous advantage of having read the companion paper prepared by Mr. Joseph F. Coates, of the Office of Technology Assessment, U.S. Congress. Mr. Coates's incisive and provocative analysis of the nature of public policy decision making and the difficulties that experts have in providing useful inputs to that process merits extravagant admiration. It is a frank, penetrating review of virtually all of the issues that bemused academics like myself who have fluttered around the fringes of the Federal policy community for many years have vaguely sensed as being characteristic of policy making.

I would like to underline a few points made by Mr. Coates, as a preliminary to some suggestions about what might be done to address them. Perhaps his most important single point is that policy is not made in a problem-oriented vacuum. Instead, it is made in an embattled arena, usually by a man or an organization upon whom are focused the efforts of a wide variety of conflicting stake holders, each having his own perception of both problems and issues--often with his own collection of "facts" to back up that perception. As Mr. Coates says, "The key issue or issues are not obvious, since they usually have not been presented in a clear, cogent, or neutral way by any of the parties concerned. It is not in their interest to do so." In such an embattled context, "The resolution of an issue in almost all cases must be a compromise rather than a clear victory for any party to the conflict." This gladiatorial atmosphere presents problems to the would be policy-influencer, because "In general experts cannot deal with tradeoffs which are the essence of public policy. Experts cannot deal with compromise situations and conflict, as experts."

If one looks for the underlying issues of any conflict, they seem to fall into two categories: probabilities (measures of uncertainty) and utilities (measures of values). Concerning probabilities, Mr. Coates says "The future course of every public policy issue of necessity is involved in uncertainty. Much uncertainty is not accidental but intrinsic, and cannot be eliminated for several reasons. First, the future is not fully anticipatable; second, we do not have adequate models of social change; and third, many of the consequences of actions associated with policy cannot be understood until the actions themselves are taken." I would add that often those consequences cannot be understood even after the actions have been taken. As a result, Mr. Coates says that "Another primary task for government is to manage uncertainty, i.e., to take those measures that in one way or another eliminate, hedge, reduce, or compensate for uncertainty so as to permit the institutions of society to move ahead in an organized fashion." From my own point of view, such measures for uncertainty management have a necessary preliminary: first one must measure uncertainty.

The other issue that Mr. Coates identifies as crucial is the one that he calls value, but I would prefer for history-of-science reasons to call utility. He says, "The subject of values has engendered an alarming amount of intellectual trash, useless discussion, uninformed deliberation, and pointless hand wringing.... Values are difficult to discern. Individuals often cannot see their own; when they can see them, they cannot give weights to them. Values are often ill formed. They are latent, they are dark, they cannot necessarily be related to public decisions without a great deal of intermediate work."

On the question of measuring values, Mr. Coates seems to me to be somewhat ambivalent. At one point he says, "Since values are heterogeneous and overlapping among the parties of interest, it is difficult to identify and sort them into tidy bundles. An effective way to reveal the values of the parties to the conflict is important. That revelation is not likely to result from simple direct inquiry." At another point, he derides"... the false conclusion that making those values explicit is a worthwhile activity in all public policy processes.... Many private motives are in conflict, are latent, are dark, uncongenial, and even unspeakable. Consequently the universal call for making them explicit in public is really an invitation to hypocrisy."

From reading Mr. Coates's paper, one can formulate a picture of two different Federal Government policy-makers, whom I shall call Director Devious and Director Dubious. Mr. Coates describes Director Devious quite well. "The crucial question facing public policy in any given time is striking a fresh balance among conflicting forces. . . The search for information is often a delaying tactic. It can be a mechanism for apparently taking action while taking no action. . . Even those most intimately associated with the issues. . . often find it to their advantage not to confront (them), not to define them, not state them clearly, and not use them as a basis for discourse, analysis, evaluation, and decision making. . . There is a tendency to misunderstand the role of the elected official and the senior decision maker in wanting him to make the values explicit. For him to make his values explicit would be a travesty. The decision maker's role is to adjudicate and to keep his values internal so he can affectively adjudicate the value-laden material put forward to him by others."

I have much more difficulty in finding in Mr. Coates's paper a description of Director Dubious. Mr. Coates says "Government is not a religion and bureaucrats are not moral athletes." But I believe that, in this as in other areas of performance, a desire for athletic excellence is built into many of us, whatever the level of our capabilities for fulfilling that desire. My image of Director Dubious is that he is perplexed by the multiplicity of the uncertainties and the value orientations with which he must cope. While he recognizes the necessity of functioning as a middle-man mediating among conflicting stake holders with conflicting values, in the face of technological and political realities that are often rather vaguely and uncertainly defined, he genuinely would like to perform this function as best he can, and would welcome tools that might help him to do so. Nor, I think, would he endorse Mr. Coates's advice that he should keep his own values deeply hidden from others, and perhaps even from himself. If some of his values are, as Mr. Coates says, dark, uncongenial, and even unspeakable, he wishes they weren't. He would like to have some way of inspecting values, both his own and those of others, and attempting to make some kind of moral sense out of them in their relation to the facts of the problem.

If I may lapse for a moment into psychoanalytic jargon, perhaps Director Devious might be taken as a representation of the ego of one kind of elected official or senior decision maker. If so, perhaps Director Dubious is a representation of the same person's superego.

I feel reasonably confident that Mr. Coates would regard the tools that I am going to propose for use as idealistic and naive, and therefore unlikely to be of much use to a public policy maker. Contexts exist in which I would agree with him. Nevertheless, each of the two major tools I plan to discuss is in fact in current use in significant public decision making contexts. Unfortunately, I will not present examples of the actual application of those tools to public decisions. For one thing, many of the details of those applications as they now are in progress are classified or otherwise confidential. For another thing, even if they were not, the character of each detailed application is typically so complicated that any attempt to present the basic ideas at appropriate length would inevitably fail. Consequently, I will talk about two relatively simple tools, both currently in use, in contexts in which they obviously bear on public policy, and could be used by public policy makers, but so far have not been.

Evaluating Radiologic Efficacy by Bayesian Methods

My first tool is addressed to the first of the two key problems that Mr. Coates identified: the problem of uncertainty. The work that I will be reporting comes from the Efficacy Study of the American College of Radiology, and is a collaborative effort involving Lee Lusted, Russell Bell, Harry Roberts, David Wallace, and myself, among a good many others. The funds supporting it came from the National Center for Health Services Research of the U.S. Public Health Service. (For a report on the results so far, see Lusted, Bell, Edwards, Roberts, and Wallace, in press.)

The essential purpose of the Efficacy Study is to explore the usefulness of the very large number of X-rays and other radiologic diagnostic procedures being carried out in the United States. This particular report is based on 7,976 case studies in various emergency room settings. The study is ongoing; ultimately, it hopes to explore something on the order of 60,000 cases in a very wide variety of settings for radiological practice.

Back in 1971 the American College of Radiology set up a Committee on Efficacy. Among its motives were a finding by Bell and Loop (1971) that an X-ray examination of the skull following a trauma was quite unlikely to show skull fracture unless certain signs and symptoms were present, and that the probability was even lower that the radiographic findings would affect patient management or the final outcome. Bell and Loop estimated that society was paying $7,650.00 per skull fracture found in patients X-rayed under those conditions, and they questioned whether the benefits were worth the cost. More generally, the ACR's Board of Chancellors had been concerned because the demand for radiologic services was, and is, growing faster than the supply, even though costs were also increasing. No rational basis existed at that time, or now, for setting priorities for available radiologic services. Customarily the radiologist performs the radiographic examination that the attending physician requests whether or not the request is appropriate. Although some data do exist suggesting what X-ray examinations are appropriate under what conditions, most radiologists know that on occasion a physician will request a radiologic examination that appears unnecessary and the radiologist receiving the request is likely to meet it.

At its first meeting in 1971, the ACR committee on Efficacy, chaired by Professor Lee Lusted of the University of Chicago, attempted to formulate the problem of what efficacy was and how it might be measured. Three different conceptions of efficacy were proposed, varying both in relevance to the long range problem and in measurability. The most relevant, but also hardest to measure, has come to be called Efficacy-3. *Efficacy-3 is long run efficacy* from the patient's point of view; that is, a diagnostic procedure is Efficacious-3 if the patient is, in the long run, better off as a result of that procedure and its consequences than he would have been had it not been performed. Obviously, *knowledge of long run outcomes is difficult to obtain*, and knowledge of hypothetical long run outcomes for sequences of diagnostic and therapeutic procedures other than the one actually carried out is even more difficult to obtain. Consequently, we next considered Efficacy-2. A diagnostic procedure is *Efficacious-2* if and only if the course of subsequent therapeutic action taken by the attending physician is different as a result of performance of the procedure than it would have been otherwise.

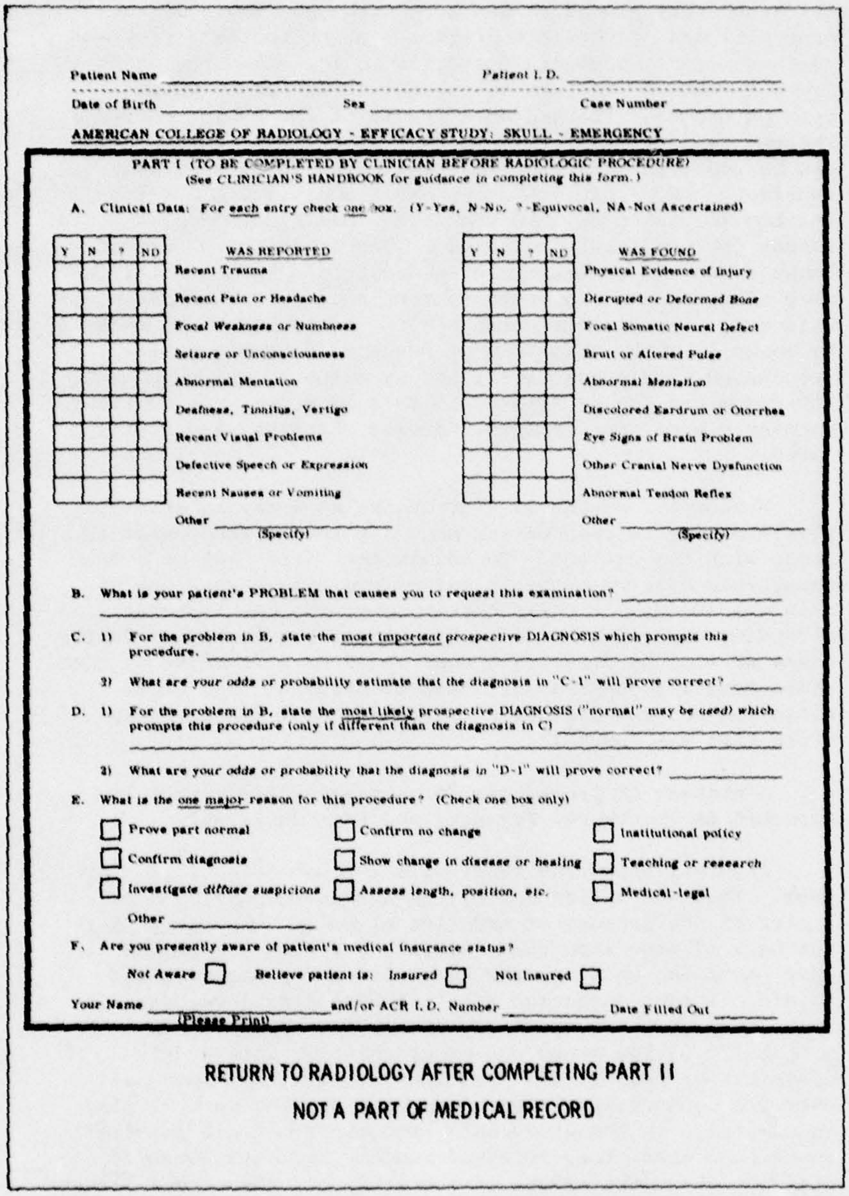
Obviously Efficacy-2 is easier to measrue than Efficacy -3, since it refers only to events in the immediate future. However, one must still discover what would have been done had constraints existed that did not in fact exist, and that too presents measurement difficulties. So, as a final fall-back position, we proposed Efficacy-1. A procedure is Efficacious-1 if and only if the procedure influences the diagnostic thinking of the attending physician. This definition turns out to lead to relatively straightforward measurements. All one must do is to discover what the attending physician was thinking at the time he ordered the X-ray, what he thinks at the time he receives the result, and compare the two; if they are different, the procedure is Efficacious-1, and the size of the difference measures the amount of efficacy.

How does one measure what the attending physician is thinking? Our procedure was to collect judgments of the probabilities of possible diagnoses prior to the X-ray, and another set of judgments posterior to it. Then, by using Bayes's theorem, one can calculate the extent to which opinion has been changed as a result of the X-ray. Bayes's theorem is a trivially simple fact about probability, and can be represented for our current purposes by the following equation: LFO = LIO + LLR. In this equation, LIO stands for Log Initial Odds, LFO stands for Log Final Odds, and LLR stands for Log Likelihood Ratio. The logarithmic form of Bayes's theorem is used here in order to make the relationship additive, and in order to make the measure of diagnostic efficacy, LLR, symmetric around 0. The mathematical details by means of which this form of Bayes's theorem can be translated into other forms, and by means of which probability judgments can be related to this equation, can be found in many places, for example, Edwards, Lindman, and Phillips (1965).

Obviously, at the time he orders an X-ray an attending physician may be considering many hypotheses about what is wrong with the patient. To reduce this large set to a more manageable set, we chose to define two diagnoses. One of them was the most important diagnosis, the one that the attending physician would be most eager not to miss. In the cases we will be discussing that would be a fracture or some other medically unpleasant state of affairs. The other diagnosis was the diagnosis considered most likely; very often that was "normal".

A pretest of procedures for measuring Efficacy-1 is reported in Thornburg, Fryback, and Edwards (1975).

Figure 1 shows the front of a typical data collection form. This was filled out by the attending physician as a part of the process of ordering an X-ray. Figure 2 shows the back of that same form, which was filled out by the same physician when the result of the X-ray was returned to him. I must emphasize that the attending physicians in this study were not specially chosen for expertise in probability. The study was geographically very widely distributed; radiological settings in emergency rooms all over the country were used. Radiologists who were willing to cooperate in the study were brought from those settings to Chicago where they received roughly two days worth of training about the nature of the study and about some rather elementary rules for assessing probabilities. When they returned to their native heaths, they recruited attending physicians from among those who frequently requested them to perform radiological services. They trained the



Patient Name ____________ Patient I. D. ____________

Date of Birth ____________ Sex ____________ Case Number ____________

AMERICAN COLLEGE OF RADIOLOGY - EFFICACY STUDY: SKULL - EMERGENCY

PART I (TO BE COMPLETED BY CLINICIAN BEFORE RADIOLOGIC PROCEDURE)
(See CLINICIAN'S HANDBOOK for guidance in completing this form.)

A. Clinical Data: For each entry check one box. (Y-Yes, N-No, ?-Equivocal, NA-Not Ascertained)

Y	N	?	ND	WAS REPORTED	Y	N	?	ND	WAS FOUND
				Recent Trauma					Physical Evidence of Injury
				Recent Pain or Headache					Disrupted or Deformed Bone
				Focal Weakness or Numbness					Focal Somatic Neural Defect
				Seizure or Unconsciousness					Bruit or Altered Pulse
				Abnormal Mentation					Abnormal Mentation
				Deafness, Tinnitus, Vertigo					Discolored Eardrum or Otorrhea
				Recent Visual Problems					Eye Signs of Brain Problem
				Defective Speech or Expression					Other Cranial Nerve Dysfunction
				Recent Nausea or Vomiting					Abnormal Tendon Reflex

Other ____________ (Specify)   Other ____________ (Specify)

B. What is your patient's PROBLEM that causes you to request this examination? ____________

C. 1) For the problem in B, state the most important prospective DIAGNOSIS which prompts this procedure. ____________

2) What are your odds or probability estimate that the diagnosis in "C-1" will prove correct? ______

D. 1) For the problem in B, state the most likely prospective DIAGNOSIS ("normal" may be used) which prompts this procedure (only if different than the diagnosis in C) ____________

2) What are your odds or probability that the diagnosis in "D-1" will prove correct? ______

E. What is the one major reason for this procedure? (Check one box only)

□ Prove part normal   □ Confirm no change   □ Institutional policy
□ Confirm diagnosis   □ Show change in disease or healing   □ Teaching or research
□ Investigate diffuse suspicions   □ Assess length, position, etc.   □ Medical-legal

Other ____________

F. Are you presently aware of patient's medical insurance status?

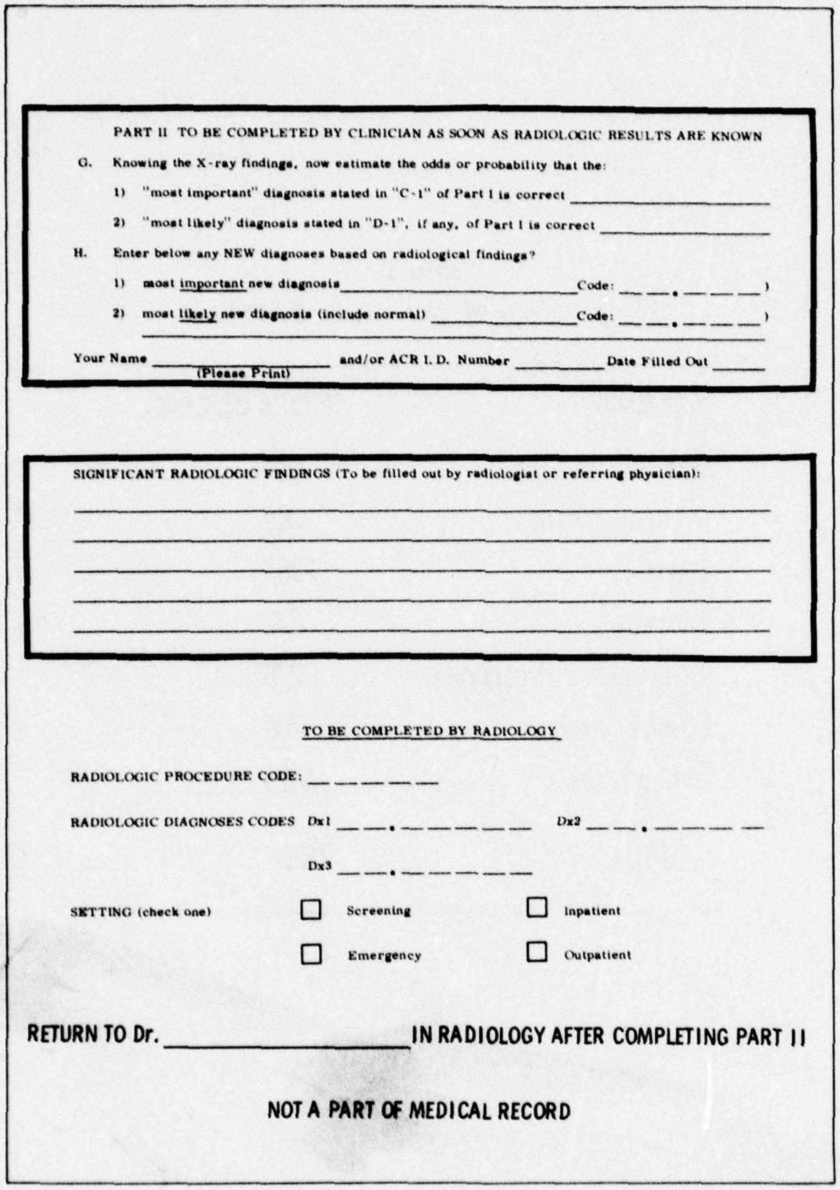
Not Aware □   Believe patient is: Insured □   Not Insured □

Your Name ____________ (Please Print) and/or ACR I. D. Number ____________ Date Filled Out ______

RETURN TO RADIOLOGY AFTER COMPLETING PART II

NOT A PART OF MEDICAL RECORD

Figure 1. Collection Form: Front Side



PART II TO BE COMPLETED BY CLINICIAN AS SOON AS RADIOLOGIC RESULTS ARE KNOWN

G. Knowing the X-ray findings, now estimate the odds or probability that the:

1) "most important" diagnosis stated in "C-1" of Part I is correct ____________

2) "most likely" diagnosis stated in "D-1", if any, of Part I is correct ____________

H. Enter below any NEW diagnoses based on radiological findings?

1) most important new diagnosis ____________ Code: __ __ . __ __ __)

2) most likely new diagnosis (include normal) ____________ Code: __ __ . __ __ __)

Your Name ____________ (Please Print) and/or ACR I. D. Number ________ Date Filled Out ______

SIGNIFICANT RADIOLOGIC FINDINGS (To be filled out by radiologist or referring physician):

TO BE COMPLETED BY RADIOLOGY

RADIOLOGIC PROCEDURE CODE: __ __ __ __ __

RADIOLOGIC DIAGNOSES CODES Dx1 __ __ . __ __ __ __ __ Dx2 __ __ . __ __ __ __

Dx3 __ __ . __ __ __ __ __

SETTING (check one) □ Screening □ Inpatient

□ Emergency □ Outpatient

RETURN TO Dr. ____________ IN RADIOLOGY AFTER COMPLETING PART II

NOT A PART OF MEDICAL RECORD

Figure 2. Collection Form: Back Side

Table 1

Distribution of Cases Over Procedures

PROCEDURE	NUMBER OF CASES
SKULL	958
CERVICAL SPINE	862
CHEST	2353
ABDOMEN	839
INTRAVENOUS PYELOGRAM	278
LUMBAR SPINE	708
EXTREMITIES	1878
TOTAL	7876

attending physicians in how to estimate probabilities. Under the circumstances we have been delighted with the relatively high quality of the probability estimates that we have obtained.

The sampling procedure used in this study, like that used in many other studies of medical practice, has one overriding principle: those who participated were those who were willing to participate. We make no apologies for this, since we know of no very satisfactory way of proceeding otherwise. Nevertheless, such sampling does present possibilities of bias in generalization to a national population either of radiologists or of attending physicians. Consequently, pending the outcome of further detailed analyses we are performing to explore the possibility of sample bias, generalizations from our results to such national populations should be done with extreme caution and nontrivial amounts of skepticism.

Various procedures explained in detail in Lusted et al. (in press) were used to spread cases widely over 47 different emergency rooms and about the same number of radiologists, between large and small hospitals, between teaching and non-teaching hospitals, and over a wide variety and number of attending physicians.

As of July, 1976, the data base was distributed over X-ray procedures as is shown in Table 1.

As usual in any kind of statistical study, there are technical problems, and I must discuss one: the truncation effect. Some respondents responded in probabilities and some responded in odds, but either way most of them worked with relatively small numbers of discrete levels of the quantities they were estimating. In the middle range of uncertainty, this hardly matters, but the extreme ends of the scale required particular attention. The problem is more severe for clinicians who reported in probabilities. Many of these, in spite of emphatic attempts to train them otherwise, made estimates of 0 or 1; both of those numbers are uninterpretable in Bayesian arithmetic. We adopted an editing convention of calling 0, .0001 and calling 1, .9999. These rounding conventions, combined with the fact that most attending physicians responded in probabilities and used only discrete sets of numbers, produced rather peculiar structures in the analyzed data. Figure 3 presents a scatter plot of log likelihood ratio against log initial odds over all procedures. You can see several parallelogram patterns that correspond to different common truncation limits used by groups of attending physicians, or imposed

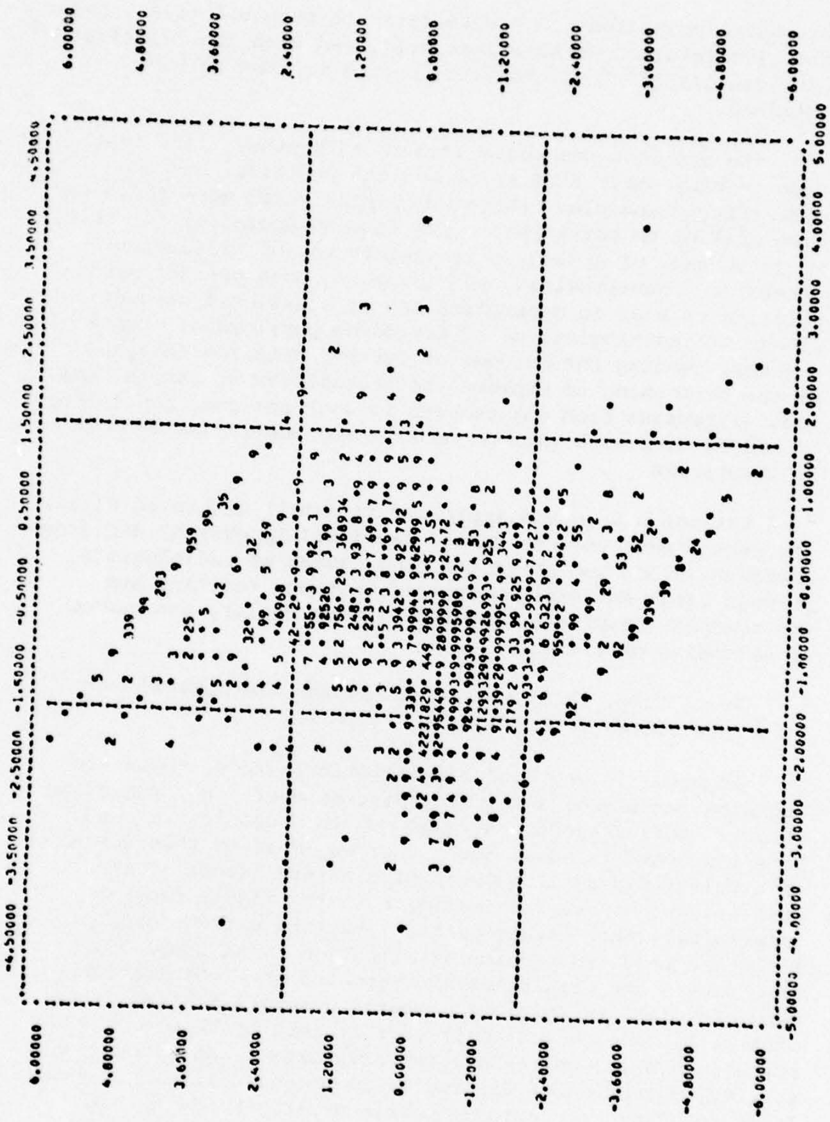


Figure 3. Log likelihood ratio as a function of log initial odds. Physician responses in 4005 cases for all seven radiologic diagnostic procedures.

by us since we could not work with estimates of 0 or 1. We have, of course, devised methods of analysis that are insensitive to what happens at the extremes of the probability scale. For a more detailed discussion of this technical topic, see Lusted et al.(in press).

Although the study is far from complete, it is possible to base some reasonably convincing conclusions on the data so far. First, the procedure is feasible; that is, such probabilistic assessments can be made in an orderly way and do provide information about the diagnostic thinking of attending physicians. We base this conclusion less on data analysis than on informal contact with the physicians who in fact made the assessments.

Our second conclusion is that the impact of X-ray examinations on diagnostic thinking was evident in the vast majority of cases and was substantial in most. Overall, not more than 10% of examinations seemingly had no influence on diagnostic thinking (that is, produced a 0 log likelihood ratio). A more detailed and refined analysis of the data suggest that the actual percentage of 0-information X-rays may be less than 5%.

Our third conclusion is that at the time X-rays were requested, the requesting physician was normally uncertain about the correctness of his tentative diagnosis. About 4 times in 5, however, the probability of the tentative most important diagnosis was assessed at less than 1/2; over half the time, it was assessed at less than about .15. In other words, the most important diagnosis often had the character of a not-very-likely medical disaster.

Our fourth conclusion is that about 3/4 of the examinations produced a lowering of the clinician's initial probabilities for the tentataive most important diagnosis. In other words, on the whole, the effect of radiology in the emergency room setting tends to be one of reassurance rather than one of confirming alarm. This conclusion has implications for the relationship between Efficacy-1, diagnostic efficacy, and Efficacy-2, treatment efficacy. Reassurance is clearly just as appropriate, from the point of view of Efficacy-1, as would be confirmation of one's worst fears. On the other hand, it seems quite likely that this finding might imply that X-ray procedures that are highly Efficacious-1 may not be especially Efficacious-2. We propose to attack that question in later studies, if we succeed in establishing that our current rather tentative ideas about how to measure Efficacy-2 are in fact workable.

Table 2
Percentage of Cases with Log Odds
Less than -1.75 or Greater than +1.75

Procedure	Before Radiography	After Radiography	Net Increase
Skull	15.9	69.9	54.0
Cervical Spine	20.8	77.4	56.6
Intravenous Pyelogram	6.1	54.7	48.6
Lumbar Spine	15.8	74.2	58.4
Chest	8.4	55.0	46.6
Abdomen	5.9	45.7	39.8
Extremities	8.4	75.8	67.4
All Procedures (7876 cases)	11.0	65.0	54.0

Our fifth conclusion is that the major effect of X-rays is to reduce uncertainty. This was no surprise. Even after examination, however, nearly 40% of clinicians assess probabilities for the most important tentative diagnosis at more than .02 but less than .98. This suggests that a substantial fraction of diagnostic decisions in the emergency room setting are based on weight of evidence rather than proof beyond reasonable doubt. Table 2 shows for various X-ray procedures the percentage of cases with log odds that are either less than -1.75 or greater than +1.75. Those numbers correspond to probabilities of .02 and .98 respectively.

An interesting sixth conclusion, at least from the study so far, is that the influence of X-ray examinations on diagnostic thinking was broadly similar for interns, resident physicians in training, and practicing physicians. Also other characteristics, such as the distribution of initial probabilities for diagnoses and the use of odds or probabilities in the expression of uncertainty, were similar for the three groups.

Some other conclusions can be reached from the data, particularly having to do with the question of how well attending physicians used the probabilities they estimated to express their uncertainty. Since these are highly technical in character, I will not review them. I will only add that in general, attending physicians tend to overassess the probability of the relatively unlikely medical disasters that were usually taken as most important diagnoses. Exactly the same kind of finding, of overassessment of the probability of highly undesirable events, has occurred in a number of other contexts in which probability estimators have the opportunity to confuse their judgment of probability with their assessments of the value of the consequence of the event whose probability was being judged. (See Kelly and Peterson, 1971)

A final implication of the study may surprise some. One of the questions asked on the initial form was whether or not the X-ray study was being performed for medical-legal reasons. This box was sometimes checked and sometimes not. Though minor differences between the results when it was checked and when it was not did occur, we were quite surprised at how small they were. In general, X-rays taken for medical-legal reasons are fully as Efficacious-1 as X-rays for which the attending physician does not indicate that he has such reasons in mind.

How does this study bear on public policy? At the moment, it has no direct bearing. It does suggest that the methodology used is in fact useable, and yields significant information about the behavior of the individuals performing socially important and policy-relevant functions. It is conceivable that refinements of the same methods, combined with methods for measuring Efficacy-2 and perhaps even Efficacy-3, might lead to policy-relevant recommendations about the conditions under which it is or is not most advisable to recommend that X-rays be taken. If such a happy result were to occur, the potential for improving the distribution of health care services might be significant.

Beyond that, however, there is a much more general implication of the study. It shows that decision makers, in this case attending physicians, can and will, with a little training and encouragement, make probability assessments concerning the issues with respect to which they are making decisions. Since uncertainty enters into every decision and probability is the appropriate metric by means of which to quantify uncertainties, this means that the hope of assessing the probabilities that enter into decisions affecting public policy may not be a vain one.

I need not rest this assertion solely on this particular study. Many other decision makers besides physicians must deal with uncertainty, and are in process of finding the explicit use of probabilities a helpful tool for doing so. We are all acquainted with the fact that probabilistic weather forecasting is coming to be more and more widely performed. (See for example Murphy and Winkler, 1974.) Even more interesting, at least to me, is the growth in use of explicit probabilities among public officials responsible for providing informational input to decision makers concerned with vast issues of global public policy. For public discussions of relevant technology, see Edwards, Phillips, Hays, and Goodman (1968), Kelly and Peterson (1971), Barclay and Randall (1975).

In sum, then, Director Dubious, eager to come to terms not only with his own uncertainties but with the uncertainties of those who advise or attempt to influence him, has available to him a quite elaborate technology, based on explicit assessment of probabilities. That technology is already in use, and its generality and simplicity invites optimists like me to suppose that that use may extend and spread into other contexts. Perhaps Director Dubious can be helped to become at least somewhat less dubious about uncertainties.

Multiattribute Utility Measurement as a Tool for the Explication and Aggregation of Social Values

As I read Mr. Coates's discussion of the latent, dark uncongenial, and even unspeakable nature of private motives, I was quite unclear whether he considered this to be desirable, deplorable, or simply a fact of life. But since I don't believe Mr. Coates's premise about the unattractive character of private motives, whether that premise is desirable or deplorable seems to be beside the point. Most motives, public or private, are mundane, ordinary, and reasonably well organized toward the problem at hand. My own motives in deciding what to include in this paper, for example are to present two intellectual tools that I think may be useful to public decision makers in as effective a light as I can manage, and in the process to be entertaining and perhaps to get a gentle argument going with Mr. Coates. Behind those surface motives, I may well have better-concealed motives to the effect that if the technologies that I am advocating are in fact perceived as useful, I may gain in prestige, in research funding, in opportunities for consultancies, and the like. None of these motives seem too latent, dark, or uncongenial; and I can guarantee that they are not unspeakable, since I just spoke (or at any rate wrote) about them. Many, perhaps most, of the motives that affect ordinary executives in their working lives have essentially this character.

Mr. Coates made eloquent reference in his paper to the two real problems about motives. One is that different people, and especially different pressure groups, have different motives, whereas the decision maker must make a decision that is responsive both to wishes of those whom he serves and to the technological facts of his problem. The other is that any single person's motives, whether private or public and whether latent or explicit, are virtually always in conflict. And, of course, every public policy decision requires value tradeoffs. In order to do better with respect to some dimensions of value, we must do worse with respect to others. But what are the appropriate exchange rates?

A new technology of value tradeoffs has been developing very rapidly over the course of the last nine years. It is called multiattribute utility measurement, and it is particularly prominent in the writings of Howard Raiffa, Ralph Keeney, R.A. Howard, and myself. Relevant references include Raiffa (1969), Keeney and Raiffa (1976), Howard (1973), and Edwards (1977, in press).

The essential idea of multiattribute utility measurement is that every significant value can in effect be partitioned into a set of sub-values on each of a number of dimensions. Technological devices exist for ascertaining what those dimensions are, for locating each one of the actions, objects, or whatever is being evaluated on each of these dimensions for judging how important each dimension is to the aggregate value of the thing being evaluated, and then for performing the aggregation. Details of this technology vary substantially from one of its advocates to another, but the description as I have just given it would probably be agreed to by all.

As in the case of probabilities, I intend to review an application that has potential public policy relevance rather than an application in being. There are in fact several applications already in being, and they have been described in open literature. However they are quite complicated. Two examples are: Chinnis, Kelly, Minckler, and O'Connor (1976); and O'Connor, Reese, and Allen (1976). See also Edwards, Guttentag, and Snapper (1975), and Keeney and Raiffa (1976). The particular application that I intend to discuss is to the selection of nuclear waste disposal sites. The work was performed in collaboration with Dr. Harry J. Otway, who is Director of the Research Project on Technological Risk Assessment, sponsored by the International Atomic Energy Authority and the International Institute for Applied Systems Analysis. For a more complete report of this study, see Otway and Edwards (in press).

Otway's project has two main goals. One is to measure the attitudes of various publics toward the risks associated with various modern technologies in general, and with nuclear power production technology in particular. The other is to explore methods by means of which the technological decision makers who must manage nuclear power activities can be aided in taking public attitudes into account in their decisions. This particular study was addressed to the latter question. The study was conducted during the course of an international meeting of high level technologists concerned with the problem of nuclear waste disposal. The ten participants included representatives from eight countries with advanced nuclear energy programs. Since the conference was in part about problems of risk assessment and risk management in nuclear waste disposal, they were very much concerned with the problem and very cooperative. Otway planned the study. enlisted the cooperation of the respondents, and collected the data. I did not attend the meeting.

The first task, of course, was to find what dimensions of value were relevant to the problem of selecting waste disposal sites. Since Otway's goal was to demonstrate how to take social attitudes toward those sites into account in the decision process, obviously social attitudes had to be one such value dimension, and indeed it was the first one listed.

Elicitation of value dimensions was done by simply asking all the respondents, together in a room, to identify what issues seemed to them important in making such decisions. Table 3 shows value dimensions and measures for six sites. After Otway had suggested social attitudes as the first such dimension, there was some question about how such attitudes should be scaled, and it was agreed that for the purpose of this demonstration a simple 0 to 100 scale would be appropriate with 100 as a highly favorable attitude and 0 as a highly unfavorable one.

The next dimension, proposed by one of the participants, was remoteness of the waste disposal site from a population center, measured in km. 160 km. was considered as having a value of 100 and 0 km. was considered as having a value of 0. The third dimension was the geospheric path length in km. Roughly, that is the distance a radioactive particle must travel, typically through the ground, to reach the nearest point used by people. Again 160 km. scores 100 and 0 km. scores 0. The fourth dimension was proximity of the waste disposal site to natural resources such as mines. 160 km. scores 100, 0 km. scores 0. The fifth dimension was geological disturbance probability--the probability of one or more significant-sized earthquakes in a year. 10^{-6} (one chance in a million) scores 100 and 1 scores 0. The sixth dimension was the relative migration rate of the critical nuclide, in the geological formation, allowing for adsorption and desorption, compared with the rate of movement of ground water (assumed constant at 0.3 m/day). Since this dimension is a ratio, it has no units; 10^{-5} was scored as 100 and 1 was scored as 0. The seventh dimension, elicited from the respondents only after a great deal of struggle and effort, was transportation distance between the nuclear plant and the waste disposal site. Zero km. scores 100 and 1.600 km. scores 0.

Note that all dimensions are transformed onto the 0 to 100 scale in such a fashion that higher scores are preferable to lower ones. The scaling of the dimensions was chosen in such a way that the respondents seemed likely to be

Table 3

Descriptions of Six Hypothetical Nuclear Waste Disposal Sites

	Value Dimension, Range, and Scaling	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6
D1.	Public attitude. 0 = extremely negative; 100= extremely positive	40	20	10	40	60	70
D2.	Remoteness from population center, km (90 km = 0; 160 km = 100)	40	12	12	120	40	120
D3.	Geospheric path length, km (0 km = 0; 160 km = 100)	40	12	12	4	4	40
D4.	Proximity to natural resources, km (0 km = 0; 160 km = 100)	50	150	150	50	15	15
D5.	Geologic disturbance probability per year (1 = 0; 10^{-6} = 100; linear in exponent)	10^{-4}	10^{-5}	10^{-4}	10^{-6}	10^{-5}	10^{-6}
D6.	Relative migration rate of critical nuclide (1=0; 10^{-5} = 100; linear in exponent)	10^{-3}	10^{-3}	10^{-2}	10^{-1}	10^{-2}	10^{-1}
D7.	Transportation distance, km (1600 km = 0; 0 km = 100)	1500	500	500	1500	150	150

willing to treat the single dimension utilities as linear with the physical measures involved--and indeed they were. In the case of dimension 5 and dimension 6 this linearity is, of course, with the exponent rather than with the number itself.

In retrospect, several features of the scaling of the dimensions were questionable. The most obvious is the use of 1 as the highest probability of an earthquake in a year. No one would seriously propose a nuclear waste disposal site with so high a probability of an earthquake; a lower probability should have been used as the upper bound.

It is important to emphasize that all sites were assumed to have the same biological characteristics, and that use of any of them was assumed to fall within appropriate budget constraints.

The value model to be used in this particular exercise was a simple weighted average model. Such value models are quite common, and have been exposed to a great deal of criticism by decision analysts (e.g. Keeney and Raiffa, 1976) who complain, quite correctly, that they do not capture subleties in the value structure that people may bring to a problem. Those, like myself, who like to use simple structures, and who feel that the simplicity of eliciting numbers built around those structures is more important than getting the model structure just right at the cost of enormously enhanced complexity of elicitation technique, are happy that a number of approximation theorems show that value structures elicited in this way will, under conditions such as prevailed in this experiment, often be very close approximations to much more elaborate and sophisticated value structures that would have required very much more difficult, complicated and socially unacceptable judgments. (See Yntema and Torgerson, 1961; Dawes and Corrigan, 1974, Wainer, 1976; and von Winterfeldt and Edwards, 1973(a), 1973 (b).)

In order to perform a simple evaluation of this kind, the next necessary step is to obtain the weights that are to be associated with the various dimensions. My preferred procedure for doing this is to ask each respondent, working separately, first to rank the dimensions in order of importance, from most to least important. Then he arbitrarily assigns an importance weight of 10 to the least important dimension, and then moves up through the dimensions making ratio judgments about the relative importances of each of the more important dimensions compared with the least important dimension. Since he can also make ratio judgments of the various dimensions to one another, he can obtain a great many internal consistency checks to make sure that he is in fact not unduly succumbing to whole number tendencies or any of the other vices to which this kind of judgmental procedure is subject. This was done for each respondent.

Finally, in order to see whether the apparatus that thus had been developed for assessing the attractiveness of waste disposal sites was appealing to the respondents, it was necessary actually to consider some waste disposal sites. So far, the entire process had been carried out without reference to any specific site. However, a number of sites that have been proposed as possible ones for nuclear waste disposal were used as the basis for judgment on the seven relevant dimensions, and the result is shown in Table 3. The ranges of the various dimensions that were actually encountered in the sites were much smaller than the ranges that had been anticipated as possible; this fact has important methodological consequences which I will discuss in a moment.

So far as the respondents were concerned, the final procedure was to ask them to make holistic evaluations, which means ratings on a 0 to 100 scale, of each site, for comparison with the multiattribute utility evaluations.

Otway asked each respondent to judge the importance weights of the seven value dimensions twice and consequently we could calculate test-retest reliabilities of these judgments. Correlations between first and second judgments were very high; the mean was .93. For convenience, all subsequent calculations used the second set of weights. The interrespondent agreement about importance weights was, as you would expect, much lower. Correlations among second judgment weights between pairs of respondents range from +.97 to -.27, with a mean of +.39. Actually, this is a somewhat higher level of inter-judge agreement than has been found in some other applications of this particular technique (e.g. the OCD example in Edwards, Guttentag, and Snapper, 1975). I have argued elsewhere (Edwards, 1971, in press; Edwards, Guttentag and Snapper, 1975) that individual differences in values should show up primarily in assessments of the importance of value dimensions. Single-dimension utilities are often technical judgments rather than value judgments.

Obviously, the question that would be of primary interest to Mr. Coates, and also to me, is: How do we go about reducing, removing or otherwise dealing with these individual differences in values?

At this point, unfortunately, time pressure problems arose. The best way to do it would be to normalize the importance weights for each individual separately, to average them, to calculate the ratios of importance weights specified by the averages, and then to feed those ratios back to the judges, sitting as a group, and ask them to debate them until they reach some form of agreement about a final set of such judgments that they were willing to allow to be used in a decision process. We did indeed normalize and average, but Otway could not feed back and reconcile differences. In a different context, I have tried this process of feeding back and reconciling differences, with quite good results. (See Edwards, in press.) And I would anticipate that some procedure of that sort would be the essential ingredient in any large-scale application of this technology to decisions over which there are major social conflicts. In the contexts in which the technology has so far been applied, however, the issues involved have been so profoundly technological that such a procedure has not generally been used. Instead, the experts on each of the kinds of numbers were asked to reach consensus about the numbers within the field of their expertise, and were usually able to do so quite well. Perhaps this technology is more easily applicable to fields in which this kind of technological resolution of conflict is appropriate than it is to contexts involving broader kinds of social conflicts.

Now we must turn our attention to the range problem that I mentioned earlier. Consider, for example, dimension 3, geospheric path length. Its actual range covers only 22.5% of the range that originally had been assigned to it. This can easily happen in situations, such as this one, in which the evaluation scheme is developed before the entities to be evaluated are known. Yet exactly that must often be done.

The reason why this presents a problem is that the range of utility values of a value dimension is in a sense a kind of importance weight. A dimension whose utility values range from 0 to 50 is effectively only half as important in controlling evaluation as one having the same weight whose utility values range from 0 to 100.

This problem can be solved only by judgmental methods. However, some mathematical techniques exist that help to put it into perspective. It is possible to transform both of the single-dimension utility values and the importance weights in such a fashion as to preserve unchanged the preference ordering over the options and the utility spacing

Dimensions	Sites					
	S1	S2	S3	S4	S5	S6
Public attitude	50	16.7	0	50	83.3	100
Remoteness from population center	25.9	0	0	100	25.9	100
Geospheric path length	100	22.2	22.2	0	0	100
Proximity to natural resources	25.9	100	100	25.9	0	0
Geologic disturbance probability per year	0	50	0	100	50	100
Relative migration rate of critical nuclide	100	100	50	0	50	0
Transportation distance	0	74.1	74.1	0	100	100
Aggregate utility $(\sum_i W_i u_{ij})$	45.6	57.3	40.4	38.2	41.0	57.9

Table 4

Rescaled single-dimension utilities and aggregate utilities at six nuclear waste disposal sites between options, while putting all of the single-dimension utility functions on a scale whose minimum in fact falls at 0 and whose maximum in fact falls at 100. Table 4 shows the result of doing so. Inspection of that table will show that no one could possibly pick site 3. In technical jargon, site 2 dominates site 3; that is, site 2 is at least as good as site 3 on every dimension, and definitely better on at least one. No other site is dominated. Also note that site 6, although evaluated as best by the weighted utility criterion, does not dominate site 3; site 3 is better than site 6 on the dimensions of proximity to natural resources and transportation distance.

The transformations which I have discussed permit exploration of the extent to which the scaling of the single dimension utility functions influences the ultimate outcome. I won't go into the details, but I can say that in this particular instance, which is rather extreme in deviations of the actual from the anticipated ranges, the effect on preference orderings was extremely modest. In other words, this procedure is rather robust to errors of anticipation of that sort.

Finally, consider the relation between the holistic ratings for the other sites by the respondents and the multiattribute utility ratings. The mean correlation in holistic ratings between pairs of correspondents is +.20, and the range is from +.97 to -.55. Note that the respondents are even less in agreement ab out holistic ratings than they were about importance weights. That too is a common finding in applications of this method. The correlation between mean holistic ratings and multiattribute utility ratings is +.58. Both procedures consider site 6 to be best and site 3 to be worst. This correlation between multiattribute utilities and holistic ratings is somewhat high compared with most other such correlations in the multiattribute utility literature, although it still shows that the two procedures do lead to different results. That on the whole is gratifying. After all, there would be no point in procedures like multiattribute utility measurement if direct numerical assessments produced exactly the same results.

Except for various technical details having to do with intercorrelations among dimensions, both in value and in physical characteristics, and with the effect of these on scaling procedures, that's the end of the story of this particular study, except for one important addition. Harry Otway informs me that the respondents thoroughly enjoyed the study, found the importance weights that they had judged extremely enlightening, and requested him to be prepared to repeat the study at their next meeting, with a considerably more realistic setting and paying considerably more attention to the details of how the study is done.

As I have said before, much more sophisticated and complicated versions of exactly the same technology have been used and are now being used to make major socially important decisions. Several have been published in unclassified sources. For example, one (Chinnis et al, 1976) has to do with the selection of the winning bidder from among a number of bids in a very large-scale procurement of an important and expensive item of military hardware. The additional complexities of the method were concerned primarily with the much larger number of dimensions that were taken into account, the use of a hierarchical value model rather than the simple value model I have presented here, and the introduction of scenarios and scenario probabilities as a tool for the assessment of values. While these technological details are all of fundamental importance to real applications, nothing in them changes the basic idea I have presented in this rather simple-minded exposition.

Nor are all the examples military. In one published application, (Edwards, Guttentag, and Snapper, 1975) a technique of essentially this character was used to help a major agency within the Department of Health, Education, and Welfare to make decisions about the allocation of its research budget for a year. In another application, now in progress, the same kind of technology is being used in planning the rate at which a government agency should encourage a boom town to boom. Still another application now in progress is to the National Program for Decriminalization of Status Offenders. A great deal of data has been collected by Professor Solomon Kobrin and his collaborators at the Social Science Research Institute of USC on the impact of this program both on the juveniles with whom it deals and on the criminal justice and related agencies who must deal with these juveniles. We are now collecting multiattribute utility measurements from a number of experts on juvenile delinquency, crime, the juvenile justice system, and the like, and expect to use these judgments in the process of assessing what the overall effects of this major national program in fact have been, and whether those effects are good or bad, and how good or how bad.

Conclusion

This paper, after some initial questioning of the assertion that major issues of public policy are inaccessible to technological tools, has attempted to illustrate the nature of two technological tools, and to suggest how they can be and are being used in the course of making major social policy decisions. Obviously, I would not want to claim that these tools are optimal, that they are fully developed, or that they should be used for all such decisions. Their applicability is quite limited, as I have attempted to suggest in the course of sketching their nature. Within that area of applicability, however, I believe that they can help those charged with responsibility for social policy in dealing with the two key problems that Mr. Coates identified: uncertainty, and difficulties in assessing and reconciling values.

As Mr. Coates correctly pointed out, no technological tool is likely to be of very great use to Director Devious. His conception of his function, and his goal structure, makes him essentially uninfluenceable by the technology of decision making. Indeed, only the part of that technology that has to do with budgeting and the assessment of costs is likely to get very much of his attention.

On the other hand, as I suggested at the beginning of this paper, Director Dubious is less impervious, mostly because he is less convinced that social policy making must continue to be done in the way in which it always has been done. I conceive of Director Dubious as a skeptical but open-minded man, interested in technological innovation and willing to explore the possibility that a particular technological innovation may have something useful to offer him. I have suggested two possible candidate technologies for his attention.

References

Barclay, S. & Randall, L. S. Interactive decision analysis aids for intelligence analysts. Technical Report DT/TR 75-4. McLean, Va.: Decisions and Designs, Inc., December, 1975.

Bell, R. S., & Loop, J. W. The utility and futility of radiographic skull examination for trauma. New England Journal of Medicine, 1971, 284, 236-239.

Chinnis, J. O., Kelly, C. W., III, Minckler, R. D., & O'Connor, M. F. Single channel ground and airborne radio system (SINCGARS) evaluation model. Technical Report DT/TR 75-2. McLean Va.: Decisions and Designs, Inc., August, 1976.

Coates, J. F. What is a public policy issue? Unpublished manuscript.

Edwards, W., Lindman, H., & Phillips, L. D. Emerging technologies for making decisions. In New directions in psychology II. New York: Holt, Rinehart, and Winston, 1965.

Edwards, W., Phillips, L. D., Hays, W. L., & Goodman, B. C. Probabilistic information processing systems: Design and evaluation. IEEE Transactions on Systems Science and Cybernetics, 1968, SSC-4, 248-265.

Howard, R. A. Decision analysis in systems engineering. In Miles, R. F., Jr., (Ed.), Systems concepts. New York: Wiley, 1973.

Keeney, R. L., & Raiffa, H. Decisions with multiple objectives: Preferences and value tradeoffs. New York: Wiley, 1976.

Kelly, C. W. III, & Peterson, C. R. Probability estimates and probabilistic procedures in current-intelligence analysis. IBM Rep. 71-5047. Gaithersburg, Md.: International Business Machines, 1971.

Lusted, L. D., Bell, R. S., Edwards, W., Roberts, H. V., & Wallace, D. L. Evaluating the efficacy of radiologic procedures by Bayesian Methods: A progress report. In Snapper, K. (Ed.), Models and metrics for decision makers. Washington, D. C.: Information Resources Press, in press.

Murphy, A. H., & Winkler, R. L. Probability forecasts: A Survey of national weather service forecasters. Bulletin of the American Meteorological Society, 1974, 55, 1449-1453.

O'Connor M. F., Reese, T. R., & Allen, J. J. A multiattribute utility approach for evaluating alternative Naval aviation plans. Technical Report DT/TR 76-16. McLean, Va.: Decisions and Designs, Inc., September, 1976.

Otway, H. J., & Edwards, W. Application of a simple multiattribute rating technique to evaluation of nuclear waste disposal sites: A demonstration. Vienna, Austria: International Atomic Energy Authority, in Press.

Raiffa, H. Preferences for multiattributed alternatives. RM-5868-DOT/RC. Santa Monica, CA: The Rand Corporation, 1969.

Thornbury, J. R., Fryback, D. G., & Edwards, W. Likelihood ratios as a measure of the diagnostic usefulness of the excretory urogram information. Radiology, 1975, 114, 561-565.

von Winterfeldt, D., & Edwards, W. Costs and payoffs in perceptual research. University of Michigan, Engineering Psychology Laboratory Report 011313-1-T, October, 1973. (a)

von Winterfeldt, D., & Edwards, W. Flat maxima in linear optimization models. University of Michigan, Engineering Psychology Laboratory Report 011313-4-T, November, 1973. (b)

Wainer, H. Estimating coefficients in linear models: It don't make no nevermind. Psychological Bulletin, 1976, 83, 213-217.

Yntema , D. B., & Torgerson, W. S. Man-computer cooperation in decisions requiring common sense. IRE Transactions on Human Factors in Electronics, 1961, HFE-2, 20-26.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
01855-1-T

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
(6) Differential Weighting for Prediction and Decision Making Studies: A Study of Ridge Regression,

5. TYPE OF REPORT & PERIOD COVERED
Technical 10/76-9/77

6. PERFORMING ORG. REPORT NUMBER
77-1

7. AUTHOR(s)
(10) J. Robert Newman

8. CONTRACT OR GRANT NUMBER(s)
Prime Contract
N00014-76-C-0074 new
Subcontract 76-0308-0715

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Social Science Research Institute
University of Southern California

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Advanced Research Projects Agency
1400 Wilson Blvd.
Arlington, Virginia 22209

12. REPORT DATE
(11) August 1977

13. NUMBER OF PAGES
47

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
Decisions and Designs, Inc.
Suite 100, 7900 Westpark Drive
McLean, Virginia 22101
(under contract from Office of Naval Research)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

(9) Technical rept. Oct 76-Sep 77

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

(12) 95p.

18. SUPPLEMENTARY NOTES

(15) N00014-76-C-0074, ARPA Order-3052

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Prediction
Decision making
Regression
Multi-collinearity
Differential weighting
Cross validation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper is another in a series exploring the conditions under which either differential or simple unit weighting of predictor variables in prediction and/or decision studies will be appropriate. Some of the difficulties of applying the ordinary least squares (OLS) regression analysis to practical problems are described and an alternative regression model called ridge analysis (RIDGE) is offered as a substitute to OLS. The trouble with OLS is that when the predictor variables are intercorrelated then the regression coefficients estimated by OLS are often

→ next page

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

390 664

quite deviant from the "true" coefficients. They are often too large in absolute value and the sign of the coefficient can be wrong. The RIDGE solution to this is very simple: just add small positive values to the main diagonal of the correlation matrix depicting the intercorrelations between the predictor variables, and re-estimate the coefficients in the usual manner. The resulting estimates are called ridge estimates and in theory they will be superior to OLS estimates in the sense of producing smaller error in cross validation samples. That is, when OLS and RIDGE estimates are estimated in one sample of data, and then tested on a new sample of data the RIDGE estimates will result in fewer errors of prediction than the OLS estimates.

Several empirical studies were conducted using computer simulated data for various prediction situations. The OLS and RIDGE models were compared as to their efficacy in prediction and both models were compared against the simplest model possible, that of unit weighting (UNIT), in which no weighting is performed; the variables are simply added up and the sum used for prediction. The results of these studies indicate that OLS and RIDGE, with one exception, always outperformed UNIT with respect to producing smaller errors of prediction and, what is more important, RIDGE always did better than OLS. The one exception in which UNIT did better than OLS and RIDGE is for the case in which all the "true" coefficients are positive, not too far apart, and the sample size is relatively small (≤ 50). This is a very restricted class of conditions. The general conclusion is that UNIT weighting will be appropriate only in unusual situations. Regression models are to be preferred as a way of generating differential weights. Also, the ridge method of estimation (RIDGE) always should be the preferred model over OLS. One practical implication of this is that if an investigator does not have the luxury to do cross validation then RIDGE estimation can be used as a substitute for cross validation.