# Decision Theoretic Approach to Some Multivariate Problems *

by

Herman Rubin

Department of Statistics

Division of Mathematical Sciences

Mimeograph Series No. 167

August 1968

* This research was partly supported by the Office of Naval Research Contract N00014-67-A-0226-0008. Reproduction in whole or in part is permitted for any purpose of the United States Government.

# Decision Theoretic Approach to Some Multivariate Problems *

by

Herman Rubin

## Introduction

There are many problems in multivariate analysis which can profitably be approached in a decision-theoretic manner. In fact, many of them must be so treated if they are to be other than results in theoretical mathematical statistics. This includes many which are not now so treated. The author makes no claim that the approaches in this paper are good ones; in fact, the precise formulation of any applied decision-theoretic problem must not be made by the statistician, but rather by the user. In some cases, no procedure is suggested. However, it is hoped that this work may suggest approaches which are more appropriate.

There is a deliberate omission of all references in this paper. These references would be to results in theoretical mathematical statistics which this author does not wish to urge on the reader as giving good practical procedures. Much further development is needed by mathematical statisticians with knowledge of the problems of application before this can be done, and the applicability of such procedures must be carefully circumscribed. In fact, there is a considerable misuse of statistical procedures by people who have no idea of the situations in which they are appropriate. Except for routine situations, such as occur in industrial quality control, adjustment of good observations to be compatible with known conditions, etc., it is necessary to apply statistical procedures with intelligence.

---

* This research was partly supported by the Office of Naval Research Contract N00014-67-A-0226-0008. Reproduction in whole or in part is permitted for any purpose of the United States Government.

It is at least somewhat annoying that researchers will frequently use procedures, the applicability of which depend heavily on assumptions which they do not understand, and which they would probably not be willing to make if the assumptions were understood, and object strenuously to being asked for some idea of the parameter values. As is also noted in section 5, they use the results of a statistical investigation even though an understanding of the problem would clearly indicate otherwise.

1. Inclusion of variables in a regression. Here the problem is complicated by possible structure in the space of variables, as well as the purpose of the regression. If the purpose of the regression is merely prediction, and the variables themselves have no structure, an appropriate loss function might be of the form

$$L(w_1 b) = E_w(y - b'x)^2 , \tag{1}$$

where the expectation is on the future distribution of the obserfations. Now if $M$ is the total moment matrix of the observed $x$'s, $m$ the moments vector of $y$ with $x$, the residuals are normal, and the regression coefficients are a priori normal $(\mu, \Sigma)$, and the variance of the residuals is known, the Bayes estimator is

$$\hat{b} = \mu + (\sigma^2 \Sigma^{-1} + M)^{-1} (m - M\mu). \tag{2}$$

All of this is well-known. Normality of the residuals is not too important. However, in general the a priori distribution is not normal and $\sigma^2$ is not known. Of course, we have the well-known asymptotic results, but with reasonable sample sizes (say $< 10^7$), there is reason to doubt the effectiveness of asymptotic approximations with the number of explanatory variables

being on the order of 1000 or even 100. That $\Sigma$ may be ill-conditioned is of no consequence, since

$$(3) \qquad (\sigma^2 \Sigma^{-1} + M)^{-1} = (\sigma^2 I + \Sigma M)^{-1} \Sigma$$

is a well-behaved function of $\Sigma$ ; in fact, (3) or

$$(4) \qquad (\sigma^2 \Sigma^{-1} + M)^{-1} = \Sigma^{\frac{1}{2}} (\sigma^2 I + \Sigma^{\frac{1}{2}} M \Sigma^{\frac{1}{2}})^{-1} \Sigma^{\frac{1}{2}}$$

enables (2) to be used even for certain infinite dimensional problems. Of course, none of this is relevant if there are, say 10 explanatory variables and 1,000,000 observations. However, in many educational and psychological situations, there are 50-500 explanatory variables, and in meteorological situations, 10,000 variables may not be too many!

Another problem which must be faced is that of computational cost. A frequently overlooked point is that the theory of rational behavior under uncertainty assumes zero cost of computation and zero computing time. If we consider a problem with $10^6$ observations and $10^4$ explanatory variables, and endow our computer with a multiply plus add time of two microseconds, it would take about 28,000 hours (about 3.2 years) to even compute the moments! The computation of the regression coefficients by (2) would take about 100 hours of computer time if $\Sigma^{-1}$ were given, and less than 400 hours if $\Sigma$ were given. However, if the number of variables were reduced to 1000, the moment time decreases to 280 hours and the computation time to .4 hours.

A practical problem with fitting large-scale regressions is the problem of multicollinearity. This is <u>not</u> due to chance, but is due to the fact that, in practice, the explanatory variables are highly related.

Because of these considerations, one may ask if standard procedures, such as stepwise choice of variables, can be modified to take into account decision-theoretic considerations. The prediction error is

$$\mu = (\beta_1^* - \hat{\beta}_1^*)' x_1 + \beta_2' (x_2 - M_{21} M_{11}^{-1} x_1) + v, \tag{5}$$

where $x_1$ is the included vector, $x_2$ the excluded vector, and $v$ the residual. If a Bayes procedure is used to obtain $\hat{\beta}_1^*$, the various terms are orthogonal, and

$$\rho = E[(\beta_1^* - \hat{\beta}_1^*)' H_{11}(\beta_1^* - \hat{\beta}_1^*) + \beta_2' H_{22}^* \beta_2 + \sigma^2]. \tag{6}$$

What can be done without using prior information? If we use standard regression techniques

$$E((\beta_1^* - \hat{\beta}_1^*)' H_{11}(\beta_1^* - \hat{\beta}_1^*)) = (\sigma^2 + \beta_2' \bar{M}_{22}^* \beta_2) \operatorname{tr} H_{11} M_{11}^{-1},$$

or we may rewrite (6) as

$$\rho = (\sigma^2 + \beta_2' \bar{M}_{22} + \beta_2)(1 + \operatorname{tr} H_{11} M_{11}^{-1}) + \beta_2' (H_{22}^* - \bar{M}_{22}^*) \beta_2 . \tag{7}$$

We even have an unbiased estimator of the "variance" term in (7).

It is clear that we are likely to believe $\beta_2' \bar{M}_{22}^* \beta_2$ is "small", but nothing can be said without using some prior information about the last term in (7). From this we can get the general hint, however, that if a predictor is going to be considerably more variable in the future, it is desirable to include it, even though it may not have been too important in the past. Of course, it is not the variability of the variable which counts, but the variability of its residual from the sample regression on the predictors included, that matters. Also several variables would have to be considered simultaneously.

Another thing which can be done is to use the methods of Stein and others to reduce the mean squared error of the estimates. It is hard to say how this would work in practice.

The next problem is which variables to include in the regression. If there are at most 20 predictors under consideration and a high-speed computer is available, all regressions can be computed cheaply, but the user will have to instruct the computer carefully so that storage of results and handling of output do not become unreasonable. It is very dangerous to eliminate variables; let the correlation matrix be

| $y$ | $x_1$ | $x_2$ |
|---|---|---|
| 1.0 | .7 | 0 |
| .7 | 1.0 | -.7 |
| 0 | -.7 | 1.0 |

;

now $x_2$ has 0 correlation with $y$, but $R^2 = .49$ if only $x_1$ is used, and is .9608 if both are used.

Nor should one wish to include variables which give large increases in $R^2$ immediately. Let the correlation matrix be

| $y$ | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| 1.0 | .7 | .7 | .84 |
| .7 | 1.0 | 0 | .6 |
| .7 | 0 | 1.0 | .6 |
| .84 | .6 | .6 | 1.0 |

The optimal regression is $y = .7x_1 + .7x_2$ with $R^2 = .98$ ; the best single variable, of course, is $x_3$ with $R^2 = .7056$ ; $x_1$ (or $x_2$) and $x_3$ gives $R^2 = .7656$. It might even be possible to reject as insignificant all variables individually which might improve the prediction, but the total set might give considerable improvement. This can be very easily seen if we change the last row of the correlation matrix to (.98 .7 .7 1.0); the best single variable yields $R^2 = .9604$, the best two, $R^2 = .98$ , two including the best one, $R^2 = .960784$.

The author would suggest, therefore, that if not more than n ($20 \leq n \leq 50$, depending on the value of the problem) predictors, it might be useful to compute all regressions, possibly using a Stein-type modification, and take that one for which the estimated prediction error, taking into account degrees of freedom, is smallest.

However, one would expect that the user should be prepared to give prior information which should enable better procedures to be devised. Also, it is rare that follow-up studies are made of the uses of regression; this should be done. As to what should be done with large numbers of predictors, I suspect that modifications of the preceding procedures could be used, but the computational costs mount rapidly.

It is even possible, say in polynomial regressions, that the number of variables available may be infinite, in which case it is necessary to finitize the number of terms in order to compute. We now have a risk function which can be approximately written as

$$\rho = W(\hat{\beta} - \beta) + R(\hat{\beta}) , \tag{8}$$

the first term being the prediction error and the second term due to the

inconvenience or cost of computation. R can be a very complicated function; for example, the sequence $\{nab^{n-1}\}$ in a polynomial regression is cheaper than its first 5 terms, and the sequence $\{ab^n/(n+1)\}$ is cheaper than its first 20 terms.

Furthermore, the multicollinearity problem is very bad for polynomial regression, which may require us to add a term of an extremely involved analytic nature to (8). Thus we need means of estimation which, at some stage, take into account both the complexity term and the computational cost term. With large amounts of data, this can be very difficult; the complexity term, is likely to lead to regressions non-linear in the parameters, and possibly a complete recalculation of moments at each stage. Multicollinearity may also require calculation of residuals and moment computations. Since except for error control problems the moment computation is likely to be considerably costlier than any other single phase, we may have a very expensive computational situation. There is the further problem of the assessment, particularly by a computer program, of the possibility of nearby values of $\hat{\beta}$ with $R(\hat{\beta})$ small. R will always be discontinuous everywhere it is finite if $\beta$ is infinite dimensional, but it is lower semi-continuous. As the examples cited show, these discontinuities may be very complicated.

2. "Scientific" purpose of a regression. Structural Inference. Here the loss function might be of the form (8), but the complexity term is much more important. The inclusion of many variables not functionally related, which before could not be seriously entertained, now can be of low cost. For example, if there is a "natural" enumeration of variables, $\{nab^{n-1}\}$ and $\{ab^{n-1}/n\}$ are likely to be of "complexity" between 2 and 10.

Furthermore, R is much harder to evaluate. Too much complexity may make it difficult for theoretical models to be formulated which will advance the field; too little may not provide a good enough fit, or may leave out important variables. This would indicate the possibility that the risk may depend on $\hat{\beta}$ and $\beta$ in a much more complicated manner. Note also that the complexity term is quite time dependent; if the current theoretical model takes into account all but one of the variables included in a regression, the complexity is low. We should remember that telescopic data would readily refute Kepler's laws at any reasonable significance level, yet these confirm Newtonian gravitation, which fits extremely well. There are also some relatively simple physical problems which have not yet been satisfactorily treated, although much good data is available.

Occasionally, the hypothesis that certain elements of $\beta$ are zero will have positive probability; in this case relatively simple Bayes or approximately Bayes procedures may suffice for large samples. Generally this will not be so. However, the prior distribution is likely to concentrate near the "simple" regressions. Again, follow-up of the results will be quite useful.

Similar comments apply to non-regression type problems of structural estimation, such as simultaneous equation models. There is the further effect here that certain incorrect assumptions may improve estimation substantially, but increase bias only slightly.

3. Discriminant Analysis. If we consider the case of linear discriminant analysis between several groups each assumed normal with constant known covariance matrix, we find ourselves almost exactly in the prediction situation for regression. However, if there are more than three groups,

uncorrelated with everything else, then the asymptotic theory is relatively simple. The above relaxation of assumptions also makes comparability of different samples easy.

5. Looking at the data. In most statistical problems, the tendency is to select a technique and "let the data speak for themselves". There are certainly dangers in "looking at the data"; the dangers in the other direction can be even greater. Of course, in complicated problems, it may be necessary to program the computer carefully so that the data can be intelligently looked at.

For example, suppose a regression of college grades on several criteria shows a negative coefficient for a mathematical aptitude score, there being no other criterion of a similar type in the list. Then for predicting college grades approximately this regression should be used (assuming large samples and few predictors), but anyone who, on the basis of such a study, would suggest using the regression for college entrance purposes or would conclude that the score does not reflect mathematical aptitude amply fulfills Disraeli's opinion of statisticians.

For another example, suppose that the following results are obtained on a common final examination of a many-section course, the instructors being evaluated as of two types, and the students being rated before the course in equal groups as good, average, and poor, the figures being average numberical grades.

| | A | B |
|---|---|---|
| good | 118 | 115 |
| average | 102 | 120 |
| poor | 100 | 70 |

Any policy decision on these numbers alone is dangerous, even assuming sampling errors are small. However, a typical investigator, noting that the mean in the A group is significantly greater than that in the B group, would jump to the conclusion that A-type teachers should be used. This type of "statistical" behavior is unfortunately becoming very common in many fields.

Unclassified
Security Classification

## DOCUMENT CONTROL DATA - R&D

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

| | |
|---|---|
| 1. ORIGINATING ACTIVITY (Corporate author)<br>Purdue University | 2a. REPORT SECURITY CLASSIFICATION<br>Unclassified<br>2b. GROUP |

3. REPORT TITLE

Decision Theoretic Approach to Some Multivariate Problems

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

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

Rubin, Herman

| 6. REPORT DATE | 7a. TOTAL NO. OF PAGES | 7b. NO. OF REFS |
|---|---|---|
| August 1968 | 11 | 0 |

| | |
|---|---|
| 8a. CONTRACT OR GRANT NO.<br>N00014-67-A-0226-0008<br>b. PROJECT NO.<br>c.<br>d. | 9a. ORIGINATOR'S REPORT NUMBER(S)<br>Mimeo Series #167<br>9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) |

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited.

| 11. SUPPLEMENTARY NOTES | 12. SPONSORING MILITARY ACTIVITY |
|---|---|
| | Office of Naval Research |

13. ABSTRACT

Some problems of multivariate analysis are considered from a decision-theoretic standpoint. Reasons are given for this to be done in practical situations, and some suggestions are made which, it is hoped, will lead to useful practical procedures.

DD FORM 1 JAN 64 1473

Unclassified
Security Classification