COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut

COWLES FOUNDATION DISCUSSION PAPER NO. 108

Estimation in the Linear Decision Model

Walter D. Fisher

January 1961

CONTENTS

Abstract

Using statistical decision theory with reference to a linear decision model with a quadratic welfare function in the endogenous variables, it is shown that (1) the loss function is different than the usual loss functions implied in prediction models; (2) under the Bayesian assumption that a prior distribution of the unknown parameters exists and under usual data assumptions, the minimum-risk decision implies a certain class of "optimal" estimates of the parameters, which are different from the usual estimates; (3) the optimal estimates require some knowledge on the part of the estimating statistician of the decision-maker's welfare function.

ESTIMATION IN THE LINEAR DECISION MODEL*

Walter D. Fisher

1. THE PROBLEM

General Character of the Problem -

A linear economic model may be used for two distinct purposes: (1) predicting future values of endogenous variables with given values of exogenous variables; (2) deciding on future values of certain controlled exogenous variables (with given values of the remaining uncontrolled exogenous variables) that will best accomplish some objective. In some cases the objective may be represented by a cardinal objective or welfare function of the endogenous variables.

While the distinction in purpose between prediction and decision models is well-known, it has been the custom to use for both models the same statistical procedures for estimating the unknown parameters. Implicit in this custom is the assumption that for the decision model, the estimator can and should estimate the unknown parameters without reference to the welfare function of the policy-maker. In the case of a model with no overidentifying restrictions, the almost universal practice is equivalent to fitting the data to the reduced form by classical least squares. It will be shown that the customary procedure is not the correct one.

---

*The writer is grateful to L. J. Savage and T. N. Srinivasan for stimulating discussion and helpful suggestions.

The type of problem to be considered is the following. A policy-maker has a welfare function involving certain endogenous variables. These variables also appear in a reduced form -- that is, are known to be linear functions of certain exogenous variables, some of which may be controlled by the policy-maker (the decision variables, or instruments). During a data-collection period, for which data are available on all types of variables, no deliberate control over the decision variables has been exercised, but it is desired in a subsequent decision period to select values of the decision variables so that welfare is maximized. The parameters of the reduced form are assumed to remain the same in both periods. Since random errors are present in both the data period and the decision period, the best that can be done is to maximize expected welfare. From the data, the reduced form can be estimated, and the decision-maker can then use the estimated reduced form as a "certainty equivalent" -- i.e., maximize his welfare function while assuming that his estimated reduced form is true.[1] What then, is the estimation procedure that

---

[1] It will be shown that a certainty equivalent exists for this problem. The term "certainty equivalent" is due to Theil. See [1], pp. 421-431.

---

should be used for estimating the reduced form when it is known in advance that this is the use to which it will be put?

This problem will be dealt with by means of statistical decision theory and in places the Bayesian assumption will be made that a probability distribution of the unknown parameters exists. The case to be considered is that of a quadratic welfare function; a linear reduced form in the endogenous variables that appear in the welfare function, with no overidentifying restrictions on the parameters; non-stochastic exogenous variables; and serially uncorrelated disturbances with a known covariance matrix.

The reduced form may, if desired, be postulated to arise from a set of structural behavior equations. If so, since it is assumed that no overidentifying restrictions are present, estimates of the reduced form imply estimates of the parameters of such structural equations as may be identifiable. It is convenient, however, in the decision problem here considered, to deal directly with the reduced form (more precisely, with that portion of the reduced form that involves endogenous variables appearing in the welfare function), and not to refer further to the structural equations.[2]

Historical note -- The term "decision model" appears to have been used first by Frisch [2]. The idea has been developed theoretically and practically to its greatest extent in the Scandinavian countries and the Netherlands, and extensive discussion by Tinbergen [3], [4], and Theil [1] is available. The specific model considered in this paper -- a quadratic welfare function subject to a linear reduced form -- and the estimation problems associated with it have also been considered by Theil.[2a]

---

[2] For a recent and interesting discussion of the relevance of reduced forms, see Klein [19].

[2a] [1], Ch. 8.5.

Theil's analysis includes a consideration of the approximate effects of estimation of the reduced form on welfare attained, but it does not include a specific answer to the question: "What, then, is the estimation procedure that should be used?" He suggests that classical least squares is hard to defeat.

A straightforward application of statistical decision theory to problems of prediction and decision in economic models has been recommended by Sverdrup [5], Hurwicz [6], and Malinvaud [7]. But most of the work that has been done on estimation of economic models has been based on a two-stage procedure with Stage 1 being estimation without reference to a utility or welfare function, and Stage 2 being application of the estimates for prediction or policy without reference to data or probability problems. This procedure is recommended specifically, for example, by Marschak [8] and Koopmans and Hood [9]. Koopmans and Hood recognize the possibility of using statistical decision theory directly, but defend the two-stage approach on the grounds that welfare functions are often not known in advance, and that estimates must frequently serve a multiplicity of future objectives.[3]

---

[3] [9], Sec. 2.

To the knowledge of the present writer, an explicit comparison of the prediction problem with the decision problem by means of the loss function of statistical decision theory has not yet been made, nor has a specific solution to the decision problem been derived by use of decision theory, nor has the

relationship of such a solution to conventional estimates been shown.[4]

---

[4] Sverdrup [10] has applied decision theory to the prediction problem by use of the minimax criterion. So has Radner [18].

Specific formulation -- Consider the cardinal welfare function

$$w = b'y - y'Cy \quad , \tag{1.1}$$

where: w is a scalar representing welfare,

y is a $G \times 1$ vector of endogenous variables,

b is a known $G \times 1$ vector, and

C is a known $G \times G$ symmetric positive-definite matrix.

Consider also the reduced form -

$$y = \Pi_D z_D + \Pi_E z_E + v \tag{1.2}$$

where: z_D is a $D \times 1$ vector of decision variables with $D \leq G$, sometimes called a "decision."

z_E is a $E \times 1$ vector of non-stochastic non-controllable exogenous variables,

Π_D is a $G \times D$ matrix of unknown parameters, of rank D.

Π_E is a $G \times E$ matrix of unknown parameters, and

v is a $G \times 1$ vector of random disturbances with $E(v) = 0$, and

$E(vv') = V$, a known $G \times G$ symmetric positive definite matrix.[5]

---

[5] The distribution of v is conditional on fixed (but unknown) parameters. The expectation symbol E without further symbols attached is taken in the conventional statistical sense to mean expected value conditional upon fixed values of unknown parameters. When the Bayes approach is used later in this paper, additional explanatory symbols will be attached to the E, showing with respect to which variables (including random parameters) the expectation is taken.

The equation (1.2) may be written in abbreviated form as

$$y = \Pi z + v , \tag{1.3}$$

where Π is the $G \times H$ matrix $(\Pi_D \; \Pi_E)$ and z is the $H \times 1$ column vector $\begin{pmatrix} z_D \\ z_E \end{pmatrix}$ with $H = D + E$.

The assumptions $D \leq G$ and rank $(\Pi_D) = D$ are made for simplicity, so that no more decision variables are in the problem than are needed for a unique solution. If, for some initial problem more than G decision variables were available, or if the rank of Π_D were less than D, some linear function of certain of the z_D could be defined so as to bring the number down with no loss of welfare and a new D defined consistent with the stated assumptions.

All of the definitions and results of this paper are conditional on given z_E, so this will not be repeated. The case where there are no variables in z_E $(E = 0)$ is included as a special case.

By substituting (1.3) into (1.1) we may write

$$(1.4) \qquad w(z_D,\Pi) = b'(\Pi z + v) = (\Pi z + v)'C(\Pi z + v) \quad ,$$

in which realized welfare is regarded as a function of the decision and the unknown parameters.

Now let P be some estimate or guess of the value of Π, however obtained, with or without data. As with Π, P is a $G \times H$ matrix, and may be partitioned as $P = (P_D \; P_E)$, where P_D is a $G \times D$ matrix and P_E a $G \times E$ matrix. In this paper it will be assumed that the rank of P_D is D.[6]

---

[6] This assumption that the rank of P_D is D is a natural one, since it makes P_D of the same rank as its counterpart Π_D, and this is the conventional approach in estimation problems. It may be pointed out, however, that there may be instances where the estimator prefers to simplify or reduce the system by only considering matrices P_D that are of rank less than Π_D, and aggregation of variables for convenience would be one such case. These instances are not studied in the present paper.

If a decision-maker modifies the reduced form (1.3) by substituting the estimate P in place of Π, and then constructs an assumed welfare function by substituting (1.3) as modified into (1.1) in place of the true (1.3), this assumed welfare function will take the form

$$(1.5) \qquad w(z_D,P) = b'(Pz + v) - (Pz + v)'C(Pz + v) \ ,$$

being a function of the decision and the estimated parameters.

DEFINITION 1. For any P, a <u>quasi-optimal decision</u> with respect to P, $\tilde{z}_D^P$, is a value of z_D that maximizes the expected value of assumed welfare $E[w(z_D,P)]$.

In other words, $\tilde{z}_D^P$ is the value of z_D obtained by using P as a certainty equivalent for Π.[7] It will be shown in the next section that $\tilde{z}_D^P$ exists and is unique.

---

[7]The notion of certainty equivalence is applied here in a slightly different way than in Theil's discussion in [1], pp. 421-431. In Theil's discussion the matrix Π_D is assumed known (what he calls the "multiplicative structure"), the uncertainty at issue being in the matrix Π_E and in the presence of the random v in the decision period (Theil's "additive structure"). It is only within this framework that Theil reaches his conclusion that "unbiased point predictions form then a certainty equivalent in decision-making" (p. 430). Here in the present paper, it is assumed that Π_D is also uncertain.

DEFINITION 2. An implied estimate, P , with respect to some decision z_D , is a P with respect to which z_D is a quasi-optimal decision, i.e., for which $z_D = \hat{z}_D^P$.

It will be shown in a later section that implied estimates exist for any decision, and, in general, are not unique.

DEFINITION 3. A certainty decision, z_D^o , is the value of z_D that maximizes the expected value of true welfare $E[w(z_D,\Pi)]$, if Π were known.

The certainty decision may also be regarded as a special case of a quasi-optimal decision in the event that the decision-maker just happened to hit on the true value of Π , and used it for his P .

DEFINITION 4. The loss is

$$(1.6) \qquad L(P,\Pi) \equiv M(\tilde{z}_D^P, \Pi) = E[w(z_D^o, \Pi)] - E[w(\tilde{z}_D^P, \Pi)] .$$

This is the loss of realized welfare entailed by making an incorrect estimate P of Π and therefore making a quasi-optimal decision that departs from the certainty decision. To maximize expected welfare under uncertainty is tantamount to minimizing loss.

Plan of the Paper -- In Section 2 the loss function is derived in terms of the unknown parameters and an arbitrary decision, or its implied estimates, and is compared with the loss function of a prediction problem from the same reduced form. In Section 3 a general Bayesian solution for the decision problem is derived, based on the notion of a probability distribution of the unknown parameters. In Section 4 a special Bayes solution is obtained from the assumptions of uniform prior distribution of ignorance and a random sample with normally distributed disturbances. In Section 5 is a summary of the results and some comments on the restrictive assumptions used in this paper.

2. THE LOSS FUNCTION

The general case. -- If the estimate P is used, the quasi-optimal decision, $\tilde{z}_D^P$, is obtained, from Definition 1, as that value of z_D that maximizes the expected value of assumed welfare, as given by (1.5). By setting

$P = (P_D\ P_E)$, by taking the expected value of (1.5), taking derivates of this expected value with respect to the elements of z_D , and setting the derivates equal to zero, the necessary condition is obtained

$$(2.1) \qquad P'_D C(P_D \hat{z}_D^P + P_E z_E - y^o) = 0 \ ,$$

where $$y^o = 1/2\ C^{-1} b \ .$$

Since P_D is of rank D and C is nonsingular, the inverse of the matrix $P_D' CP_D$ exists. By multiplying (2.1) through by this inverse, the quasi-optimal decision $\hat{z}_D^P$ is found to be

$$(2.2) \qquad \hat{z}_D^P = (P_D' C P_D)^{-1} P_D' C(y^o - P_E z_E) \ .$$

Moreover, because of the positive-definiteness of the matrix C , this decision provides a true maximum and is unique.

The certainty decision, z_D^o , by the same reasoning but using Π instead of P , is found to be

$$(2.3) \qquad z_D^o = (\Pi_D' C \Pi_D)^{-1} \Pi_D' C(y^o - \Pi_E z_E) \ .$$

The loss, as defined by (1.6), can be obtained from a Theorem of Theil.[8] It is

$$(2.4) \qquad L(\Pi,P) = (\tilde{z}_D^P - z_D^o)' \Pi_D' C \Pi_D (\tilde{z}_D^P - z_D^o) .$$

By substituting (2.2) and (2.3) into (2.4), there results

$$(2.5) \qquad L(P,\Pi) = (Q_P^{-1} r_P - Q_\Pi^{-1} r_\Pi)' Q_\Pi (Q_P^{-1} r_P - Q_\Pi^{-1} r_\Pi) ,$$

where $$Q_P = P_D' C P_D , \quad r_P = P_D' C(y^o - P_E z_E) ,$$

$$Q_\Pi = \Pi_D' C \Pi_D , \quad r_\Pi = \Pi_D' C(y^o - \Pi_E z_E) .$$

Special cases. -- More insight into the nature of this loss function (2.5) can be obtained by considering special cases. Consider first the case where $D = G$ and $E = 0$. Then the matrices Π_D and P_D are nonsingular and square, and there are no exogenous variables z_E in the system. Then the quasi-optimal decision and the certainty decision are found, from (2.2) and (2.3) to be, respectively,

$$(2.6) \qquad \tilde{z}_D^P = P_D^{-1} y^o ,$$

$$(2.7) \qquad \hat{z}_D = \Pi_D^{-1} y^o .$$

---

[8] [1], Theorem 3, p. 453.

The loss is, from (2.5)

$$(2.7a) \qquad L(P,\Pi) = y^{o\prime}(I - \Pi_D P_D^{-1})' C (I - \Pi_D P_D^{-1}) y^o ,$$

where I is the identity matrix.

Let $\tilde{y}^P$ denote the realized value of y resulting from the quasi-optimal decision $\tilde{z}_D^P$. Then, from (1.2) and (2.6) with z_E null,

$$(2.8) \qquad E(\tilde{y}^P) = \Pi_D P_D^{-1} y^o .$$

The loss may then be written, from (2.7a) and (2.8)

$$(2.9) \qquad L(P,\Pi) = E(\tilde{y}^P - y^o)' C (\tilde{y}^P - y^o) - tr(CV) .$$

Note that from its definition following (2.1) y^o may be given the following interpretation. It is the value of y that would maximize welfare in (1.1), and that could be realized under perfect estimation of Π and zero disturbance in the decision period, provided that G independent decision variables are available (which is so assumed in the special case now under consideration) [9]

---

[9] In the more general case where D may be smaller than G, y^o may not be attainable because of the constraints on y imposed by the reduced form. But in any case, y^o is a known constant, dependent on the constants of the welfare function.

Therefore, in (2.9) the loss function is represented by a certain positive-definite quadratic form in the discrepancies between the "best possible values" of the endogenous variables and their realized values under the erroneous estimate P, less the constant $tr(CV)$.

It can be shown that (2.9) still holds in the case where $D = G$, but E not necessarily zero. In this case, however, the analogue of (2.7a) is a more complicated expression that will not be presented here.

To take the very simplest case, that of one endogenous variable and one decision variable, consider the welfare function

$$w = by - cy^2 , \tag{2.10}$$

and the reduced form consisting of the single equation

$$y = \pi z + v , \tag{2.11}$$

where all variables are scalars. Then, for some estimate, p, of π

$$\tilde{z}^p = y^o/p , \tag{2.12}$$

where

$$y^o = b/2c . \tag{2.13}$$

The loss is then, from (2.7a),

$$(2.14) \qquad L(p,\pi) = cy^{02}(1 - \pi/p)^2 .$$

Prediction. -- Now consider the loss function for a prediction problem with the same reduced form that has been used for the decision problem.[10]

---

[10]Although forecasting economists usually abstain from using notions of loss or welfare functions, they frequently express the conviction that their forecasts are to be used somehow for "policy purposes."

Say that the objective of a forecasting economist is simply to forecast the values of the endogenous variables y with small error, using given values of the other variables z. The forecaster, unlike the decision-maker heretofore considered, is assumed to have no control over any of the variables z; he treats them all as the decision-maker treated the z_E -- as uncontrolled exogenous variables.

This objective can be formulated in a rather general way by postulating that the forecaster desires to minimize the loss function

$$(2.15) \qquad L_{pred}(P,\Pi) = E[\hat{y} - E(y)]' C[\hat{y} - E(y)] = E(\hat{y} - y)' C(\hat{y} - y) - tr(CV),$$

where y is given by the reduced form (1.3); $\hat{y}$ is a forecast of y, conditional on all of the z's, using an estimated reduced form

(2.16) $\hat{y} = Pz$,

and where C is the same C appearing in the welfare function (1.1).

This loss function, also suggested by Theil,[11] has the following properties.

(1) It treats positive and negative errors in any endogenous variable as having equal weight;

(2) It allows for the possibility of giving different weights to the errors in different endogenous variables and to the interactions between such errors; for convenience in making comparison with the decision model, these weights are taken as the elements of the matrix C appearing in the welfare function of the decision model; thus making (2.15) analagous to (2.9).

(3) It includes as a special case the minimizing of the sum of mean squared errors of the "estimate" $\hat{y}_i$ from its "parameter," $E(y_i)$.

Set $C = I$ in (2.15).

(4) If $\hat{y}$ is specified as unbiased, it can be shown that the minimizing of (2.15) is equivalent to minimizing the variance of each element of $\hat{y}$ simultaneously, whatever be C . See Appendix B .

(5) It includes as a special case the minimizing of the mean squared error of a linear function of the elements of $\hat{y}$. Make C a matrix of rank one.

The superficial resemblance between this loss function for the prediction problem, as given by (2.15) and the loss function for the decision problem in the special case as given by (2.9) may be deceiving. Their fundamental difference

---

[11] See [1], Appendix 8A.

is crucial, and results from the possibility of manipulating z_D in the decision problem. Consequently, different meanings attach to y^o , y^P , and $\hat{y}$. To see this difference, substitute y and $\hat{y}$, as given by (1.3) and (2.16) respectively, into (2.15). Then the loss for the prediction problem becomes

$$L_{pred}(P,\Pi) = z'(P - \Pi)'C(P - \Pi)\ z \ , \tag{2.17}$$

whereas the loss for the decision problem is given by (2.5) or by (2.7a) in the special case. The expressions are not the same.

The difference shows up clearly in the simple case of the single-equation model with one decision variable. In the decision problem the loss is given by (2.14), which may also be written

$$L(p,\pi) = c\tilde{z}^{p2}(p - \pi)^2 \ , \tag{2.18}$$

by using (2.12). In the prediction problem the corresponding loss function is, from (2.17)

$$(2.19) \qquad L_{pred}(p,\pi) = cz^2(p - \pi)^2 \quad .$$

In spite of the superficial resemblance of these last two equations, $\tilde{z}^p$ in (2.18) is a variable, a function of p from (2.12), while z in (2.19) is a known constant. When this difference is taken into account, the estimation problems are different.

Existence of implied estimates. -- The question is now raised whether, for any arbitrary decision, there exist implied estimates in the sense of Definition 2 -- that is, whether for any z_D a P can be found for which z_D is a quasi-optimal decision with respect to that P. The question may be posed in another way. The decision process may be imagined to be either a one-stage process, in which the decision-maker does not even make use of estimates, but makes his decisions directly, in some unexplained manner; or it may be a two-stage process, in which another man, an estimator, gets an estimate, P, in some unexplained manner, and then the decision-maker uses this P as a certainty-equivalent in making a decision that maximizes assumed welfare. We ask: for any decision whatever, obtained by the one-stage process, is there a corresponding two-stage process with an appropriate P which will yield this same decision?

That the answer to this question is in the affirmative for the model considered in this paper can be seen quickly. If the arbitrary decision z_D is to be a quasi-optimal decision with respect to some P, it is necessary and sufficient that there exist a P such that (2.1) holds when the arbitrary decision z_D is substituted for $\tilde{z}_D^P$; that is, such that

$$(2.20) \qquad P_D'C(P_D z_D + P_E z_E - y^o) = 0 .$$

Thinking of the elements of P_D and P_E as variables, this is a set of D equations in DH variables, quadratic in the elements of P_D and linear in the elements of P_E. For solutions to exist it is sufficient that the parenthesis be made null -- that is, that

$$(2.21) \qquad P_D z_D + P_E z_E = y^o ,$$

where z and y^o are arbitrary non-null vectors of order H and G respectively. P may always be selected to accomplish this, and, except in special cases, the alternatives are infinite in number. It may be concluded that implied estimates exist for any decision, and (2.20) states the conditions they must satisfy.

Two special cases may be noted. First, if there are no uncontrolled exogenous variables in the system, the term $P_E z_E$ in (2.20) may be ignored, so that (2.21) becomes

$$(2.22) \qquad P_D z_D = y^o .$$

Second, if $D = G$, P_D is a nonsingular square matrix. Then (2.20) may be multiplied through by $(P_D'C)^{-1}$, which gives (2.21) as a necessary, as well as sufficient condition on P. In this case the H elements in each row of P (the estimated coefficients of an equation of the reduced form) must lie on a

certain hyperplane in H-space. Other cases will be examined in later sections of the paper, where z_D is taken as a certain kind of "optimal" decision.

Implied estimates that are not unique also exist for the prediction problem, since a prediction problem may be regarded as a decision problem with a certain kind of loss function. Let the forecast $\hat{y}$ be the decision. Then an implied estimate with respect to $\hat{y}$ is any P that satisfies (2.16), where z is a given non-null vector.

Having shown that a unique quasi-optimal decision exists for any estimate P, and also that any decision z_D corresponds to some P, with respect to which it is quasi-optimal, we have shown in essence that a many-to-one correspondence exists between the entire set of possible P's and the entire set of possible z_D's, and that we may speak of a set of pairs (P, z_D) that exhaust these sets.

DEFINITION 5. An <u>estimate-decision pair</u> is a pair (P,z_D), where P is an implied estimate with respect to the decision z_D, and z_D is a quasi-optimal decision with respect to the estimate P.

It follows that for the elements of such a pair, it is immaterial whether the loss function be written as $L(P,\Pi)$ or as $M(z_D,\Pi)$ for any Π.

3. A BAYES SOLUTION

The Bayesian approach -- The case is now considered where the investigator wishes to indicate his uncertainty of the true Π by postulating a joint probability distribution representing his degree of belief in alternative possible values of the elements of Π . In this section no attempt is made to derive or explain the origin of this probability distribution. The investigator's beliefs may be more or less well founded, and may or may not be based on the observation of data. Hence the distribution may be called a personal probability distribution in the sense of Savage [11]. In a following section a special case is considered where the personal probability is a posterior probability, and is factored in classical fashion into a prior probability and a likelihood function based on sample data.

It is recognized that the notion of regarding unknown parameters as random variables is questioned by many statisticians, and that many have been taught to avoid this notion. The reasons for using the notion here in connection with decision theory, a use which may be called a Bayesian approach, are the following ones.

(1) The concept of personal probability is a useful and natural one in decision theory in general, and in the present problem in particular, when loss or utility functions are assumed. Minimizing expected loss with respect to a personal probability distribution seems often a more reasonable course of action than, for example, minimizing maximum conceivable loss, or other proposed statistical criteria. So it seems here to the present writer in the present problem.

(2) The concept is operational, as has been pointed out by Savage. That is, it is possible by experimentation to obtain information on a person's degree of belief in an event.

(3) The use of special personal probability distributions (such as uniform prior density of ignorance) to derive optimal decisions seems no more subjective or arbitrary than the use of special "nice" statistical properties, such as unbiasedness or minimum variance.

(4) The Bayesian approach has a distinguished history, deriving from Bayes and Laplace, and a renaissance is being urged by some contemporary statisticians with strong arguments.[12] Indications are that its current

---

[12] See Savage [11], [12] for the arguments and references.

---

eclipse in the majority of statistical thinking will be only temporary, and that it will have an increasing influence in the future.

The general case. -- For convenience, the personal probability distribution is assumed to be representable as a density. Consider the GH elements of Π as random variables.

DEFINITION 6. The personal probability density of Π, $f(\Pi)$, is the joint multi-variate density of the GH elements of Π, representing the investigator's degree of belief in alternative possible values of Π for which exist the $G \times H$ matrix of expected values

(3.1) $$\bar{\Pi} = (\bar{\Pi}_D \; \bar{\Pi}_E)$$

and the GH x GH covariance matrix V_Π, both of which are assumed known.

DEFINITION 7. The Bayes risk[12a] of an estimate-decision pair (P, z_D)

[12a] Wald defines the risk as the expected value of the Bayes risk over all possible samples. It can be shown, however, that his risk is minimized by some optimal decision if and only if the Bayes risk is minimized for each sample separately.

---

is the expected value of the loss with respect to the personal probability distribution $f(\Pi)$

(3.2) $$R(P) \equiv S(z_D) = E_f[L(P,\Pi)] \equiv E_f[M(z_D,\Pi)] = \int_{-\infty}^{\infty} M(z_D,\Pi) f(\Pi) d\Pi ,$$

where $\int d\Pi$ denotes the GH-fold multiple integral with respect to each element of Π.

DEFINITION 8. An optimal decision, z_D^*, is a decision that minimizes the Bayes risk $S(z_D)$.

By looking at the definition of loss as given by (1.6) and noting that $E_f[Ew(z_D^o,\Pi)]$ is a constant, it can be seen that an optimal decision also maximizes

expected realized welfare $E_f[Ew(\tilde{z}_D^P,\Pi)]$.[13]

---

[13] In these expressions the expectation sign E without a subscript denotes expected value with respect to the distribution of the disturbance, v, while E_f denotes expected value with respect to the personal probability distribution $f(\Pi)$.

DEFINITION 9. An optimal implied estimate, P^*, is an estimate, P of Π, that is an implied estimate with respect to an optimal decision, z_D^*.

It can be shown that, given the personal probability distribution, the set of optimal implied estimates just defined is precisely the same set of estimates P that minimize the Bayes risk when expressed as $R(P)$ in (3.2).[14]

---

[14] Proof: For any estimate-decision pair (P, z_D) and for any Π, $L(P,\Pi) = M(z_D,\Pi)$, by definition of (P, z_D). Hence for any (P, z_D) and for any $f(\Pi)$, $R(P) = S(z_D)$, by taking expected value of the last condition with respect to $f(\Pi)$. Hence, in particular, $R(P^*) = S(z_D^*)$. But since z_D^*, by Definition 8, is the z_D that gives minimum S, $R(P^*)$ must be the minimum R. Hence the set of P^*, the implied estimates of z_D^* by Definition 9, is precisely the set of P's that minimizes $R(P)$. I am indebted to W. C. Bainard of the Cowles Foundation for showing me this simple proof.

From what has gone before, it is also seen that P^* may be regarded as a certainty equivalent that gives the optimal decision z_D^*, and that we may speak of the pair (P^*, z^*_D) as a particular case of an estimate-decision pair as defined above in Definition 5.

DEFINITION 10. A Bayes solution to the decision problem is the optimal decision z_D^*, or an optimal implied estimate P^*, or the pair (P^*, z^*_D).

From (2 3) and (2.4)[16]

---

[16] $L(P,\Pi)$ is equivalent to $M(z_D,\Pi)$ from (1.6). It is permissible to replace $\hat{z}_D^P$ in (2.4) by z_D since it has been shown that any z_D is equal to some $\hat{z}_D^P$. Q_Π and r_Π are defined following (2.5).

$$(3.3) \qquad M(z_D, \Pi) = z_D' \; Q_\Pi \; z_D - 2z_D' \; r_\Pi + r_\Pi' \; Q_\Pi \; r_\Pi \quad ,$$

whence, from (3.2)

$$(3.4) \qquad S(z_D) = z_D' \; \bar{Q}_\Pi \; z_D - 2z_D' \; \bar{r}_\Pi + \overline{r_\Pi' \; Q_\Pi \; r_\Pi} \; ,$$

where a bar over a quantity denotes the expected value of the quantity with respect to f.

Since $\bar{Q}_{\Pi}$ is a convex combination of positive-definite matrices Q_{Π}, it is also positive definite. Hence a true minimum of $S(z_D)$ exists and is unique. By differentiating $S(z_D)$ with respect to z_D and setting partial derivatives equal to zero, the optimal decision is given by

$$(3.5) \qquad z_D^* = \bar{Q}_{\Pi}^{-1} \bar{r}_{\Pi} = (\overline{\Pi'_D C \Pi_D})^{-1} (1/2 \ \bar{\Pi}' b - \overline{\Pi'_D C \Pi_E} \ z_E) .$$

This expression involves the means, variances, and covariances of the personal probability distribution of Π. For example, the typical element of the matrix $\overline{\Pi'_D C \Pi_E}$ is

$$(3.6) \qquad E_f\left[\sum_{h=1}^{G} \sum_{i=1}^{G} c_{hi} \ \pi_{hj} \ \pi_{ik} \right] = \sum_{h=1}^{G} \sum_{i=1}^{G} c_{ni} \left[\bar{\pi}_{hj} \ \bar{\pi}_{ik} + \operatorname{cov}\left(\pi_{hj}, \ \pi_{ik}\right) \right],$$

where c_{hi} is an element of C, cov stands for covariance, and where j runs from 1 to D and k runs from 1 to E.

Then, from (2.1) and (3.5), optimal implied estimates must satisfy

$$(3.7) \quad (P_D^{*1} C P_D^*)^{-1} \left(\tfrac{1}{2} P_D^{*1} b - P_D^{*1} C P_E^* z_E\right) = (\overline{\Pi'_D C \Pi_D})^{-1} (1/2 \ \bar{\Pi}'_D b - \overline{\Pi'_D C \Pi_E} \ z_E) .$$

Involved in this expression are the constants of the welfare function.

Special cases. -- The unintelligibility of the last equation may be partially reduced by considering some special cases. In the case where Π_D is square, and hence P_D^* invertible, by equating the right side of (3.7) to z_D^* via (3.5), (3.7) reduces to

$$P_D^* z_D^* + P_E^* z_E = 1/2\, C^{-1} b \,. \tag{3.8}$$

The optimal implied estimates here lie on a set of hyperplanes. In the case of a single equation with one decision variable (D=1, G=1, E=0), the optimal decision and its implied estimate are the respective scalars:

$$z^* = \frac{b}{2c\pi} \,, \tag{3.9}$$

and

$$p^* = \bar{\pi} + \frac{\mathrm{var}(\pi)}{\bar{\pi}} \,, \tag{3.10}$$

where $\mathrm{var}(\pi)$ is the scalar variance of the personal probability distribution of π. In this particular case the implied estimate is unique and independent of the welfare function.

Another interesting simple case is that of a reduced form of two equations, each involving the same single decision variable (D=1, G=2, E=0), the reduced form being

$$(3.11) \qquad y_1 = \pi_1 z + v_1$$

$$y_2 = \pi_2 z + v_2 \quad ,$$

and the welfare function being

$$(3.12) \qquad w = b_1 y_1 + b_2 y_2 - c_{11} y_1^2 - 2c_{12} y_1 y_2 = c_{22} y_2^2 \quad .$$

Here the optimal decision is the scalar

$$(3.13) \qquad z^* = \frac{b_1 \bar{\pi}_1 + b_2 \bar{\pi}_2}{2[c_{11}\bar{\pi}_1^2 + 2c_{12}\bar{\pi}_1\bar{\pi}_2 + c_{22}\bar{\pi}_2^2 + c_{11}\mathrm{var}(\pi_1) + 2c_{12}\mathrm{cov}\,\pi_1\pi_2 + c_{22}\mathrm{var}(\pi_2)]}$$

and the optimal implied estimates must satisfy the condition

$$(3.14) \qquad \frac{b_1 p_1^* + b_2 p_2^*}{2(c_{11}p_1^{*2} + 2c_{12}p_1^* p_2^* + c_{22}p_2^{*2})} = z^* \quad ,$$

where z^* is given by (3.13). In this case the implied estimates are not unique, but lie on an ellipse, the position and shape of which are dependent on the constants of the welfare function, b and C, which here do not vanish. In Figure 3 this locus is graphed for the case $b' = (1\ 1)$, $C = I$, $\bar{\pi}_1 = 1$, $\bar{\pi}_2 = 2$, $\mathrm{var}(\pi_1) = 1$, and $\mathrm{var}(\pi_2) = 1$. The locus is the circle

$$(3.15) \qquad p_1^2 + p_2^2 - p_1/2z^* - p_2/2z^* = 0 \quad ,$$

which has its center at the point $(1/4z^*, 1/4z^*)$ and radius $\sqrt{2}/4z^*$. All points

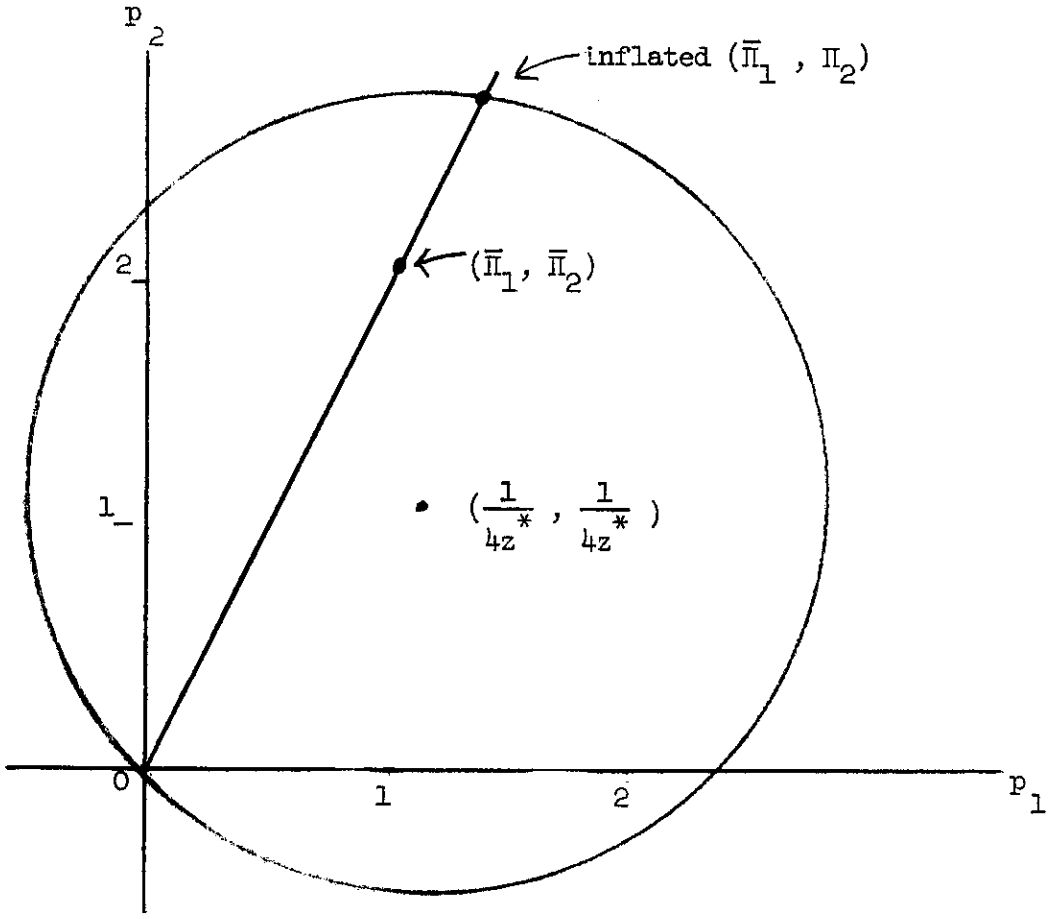


FIGURE 1

Locus of Optimal Implied Estimates (p_1^*, p_2^*) for a Two-Equation, One-Decision Model.

on the circle are valid implied estimates, except the origin because of the requirement that the vector $(p_1\ p_2)$ be of rank one.

Comparison with prediction problem -- The prediction problem may also be handled by the Bayesian approach, as is well known. When a point estimate (or guess), P, is made of the values of the reduced form parameters, a forecast, or decision, $\hat{y}$, results. The loss function (2.15) or its equivalent (2.17) may then be used to obtain the Bayes risk, either in the form $R(P)$ or $S(\hat{y})$. Putting (2.17) into (3.2) we have

$$(3.16) \quad R_{pred}(P) = E_f\,[L_{pred}(P, \Pi)] = E_f[z'(P - \Pi)'\ C\ (P - \Pi)\ z\ ] .$$

By expanding this expression and minimizing it with respect to P, the optimal implied estimates are found to be those that satisfy the relation[17]

---

[17] We are here taking advantage of the conclusion proved in Footnote 14 that the optimal estimates obtained by the minimization of $R(P)$ are the same as those that are obtained by finding the optimal decision and then finding the implied estimates with respect to that decision.

$$(3.17) \qquad P^*_{pred}z = \bar{\Pi}z .$$

It is noteworthy that this result is independent of the constants of the welfare function -- contrary to the result in the decision problem as given by (3.7), in which these constants appear.

In the special case D=1, G=1, E=0 , the risk of the estimate p becomes

$$(3.18) \qquad R_{pred}(p) = cz^2[(p - \pi)^2 + var(\pi)\ ] \ ,$$

and the optimal implied estimate attains the unique value

$$(3.19) \qquad p^*_{pred} = \bar{\pi} \ .$$

It is interesting to compare this case again with the corresponding case of the decision problem. Since the loss functions have already been shown to be different, the risk functions will be different also. For the decision problem it is -- from (2.14) --

$$(3.20) \qquad R(p) = cy^{o2} \ \frac{(p - \bar{\pi})^2}{p^2} + \frac{var(\pi)}{p^2} \ ,$$

giving the optimal implied estimate that was shown in (3.10)

$$p^* = \bar{\pi} + \frac{var(\pi)}{\bar{\pi}}$$

These two risk functions $R_{pred}(p)$ and $R(p)$ with their minimal points are graphed in Figure 2, setting $\bar{\pi} = 2$, $var(\pi) = 1$, and the multiplying constants equal to 1 . It can be seen from this figure that, crudely speaking, when the

loss is based on the "relative error" or "percentage error" between p and π, with the variable under control, p, as a base of the percentage -- as it is in the decision problem,--there is a premium on selecting a somewhat higher p (to get a large "base" and small relative error) than in the prediction problem, where the loss is based on "absolute error."

In the prediction problem with two equations and one independent variable ($G=2$, $D=1$, $E=0$), the optimal implied estimates of the vector $(\pi_1 \ \pi_2)$ is -- from (3.17) -- found to be the unique value

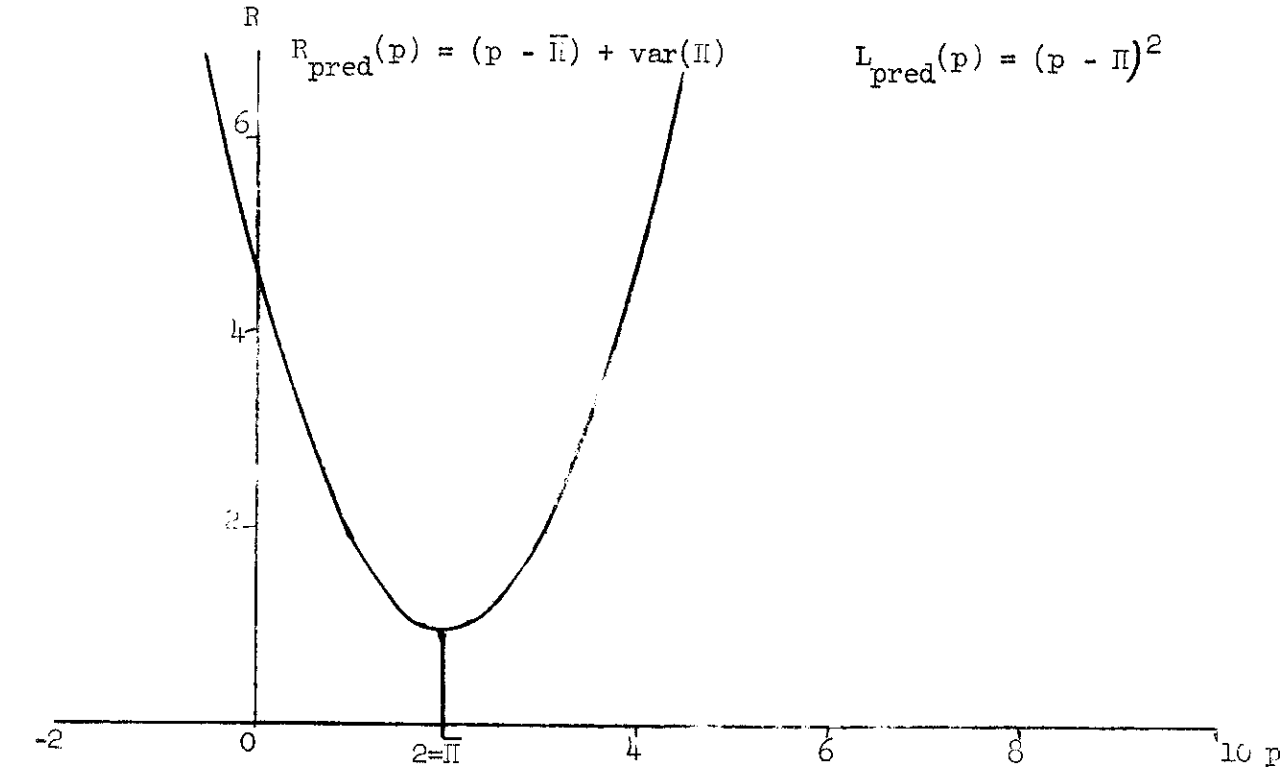
$$(3.21) \qquad (p_1 \ p_2)^*_{pred} = (\bar{\pi}_1 \ \bar{\pi}_2) \ .$$

which is independent of the welfare function. Looking back at the circle in Figure 1, this estimate is represented as a point within the circle -- the point (1,2). Thus, no member of the set of optimal implied estimates in the decision problem is the same as the (unique) estimate of the prediction problem.

4. A SPECIAL BAYES SOLUTION

Posterior distribution after observations. -- So far in this paper nothing has been said about observations, and next to nothing about conventional statistical estimates. This luxury has been made possible by the device of assuming a personal probability distribution of the unknown parameters.

But it is of interest to compare the Bayes solutions found in Section 3 above with the conventional estimates. To do this it will be necessary to make



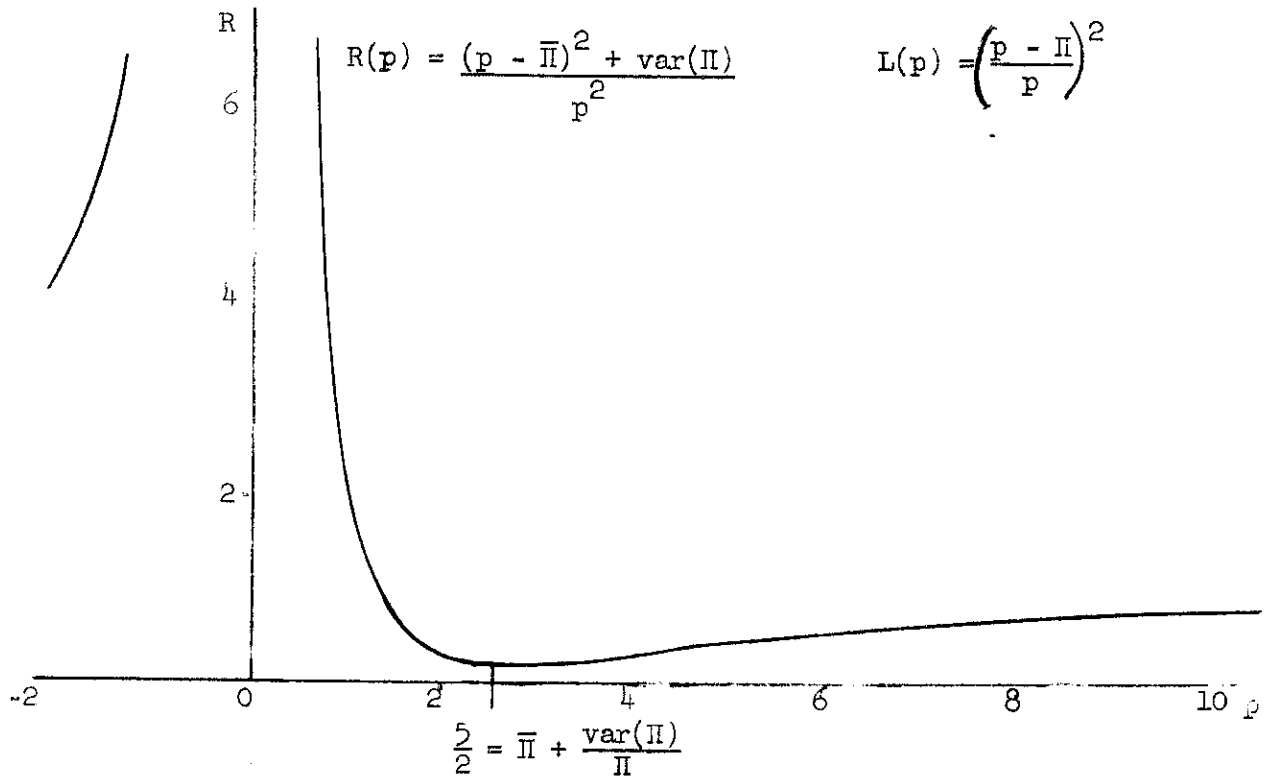


FIGURE 2
Two Bayes Risk Functions

some additional assumptions about the personal probability distribution and about the observed data.

Consider that a prior probability distribution exists, representing the investigator's degree of belief in alternative values of Π before observing any data. This will be regarded as a personal probability, as above. Then he observes data, assuming he knows something about the probability distribution of alternative data conditional on the value of Π. This conditional distribution is called the data distribution. Then the state of his beliefs after data observation may be called the posterior distribution. Since the posterior distribution is the later stage of his beliefs (probably revising the prior distribution), it is the personal probability distribution that is relevant for making the decision; so the posterior distribution of this section will be regarded as the personal probability distribution of the last section. More precise definitions of these terms are given in Appendix A, and more details may be found, for example, in Kendall.[18]

---

[18] [14], pp. 175-178.

If before the observation of data the investigator has no strong beliefs that certain values of Π are more probable than other values, if he feels he will "let the data decide for him" what Π is, the situation approaches that described in classical literature as "equal distribution of ignorance," or as "Bayes' postulate," where it is decided to assign equal prior probabilities to

all possible values of Π . Seldom can the claim be made that the probabilities are precisely equal. For this reason, and also because we are dealing with densities,[19] we shall take a limiting approach: assume the existence of a

---

[19] A uniform density over an infinite range does not exist, despite certain classical formulations that would so indicate (see [15], p.97). This technical flaw has been pointed out frequently in the literature. But the fundamental notion of prior ignorance is a valid one. As Savage has pointed out [12], the essential requirement is that it be possible to make an experiment with data that is sharp enough in relation to prior beliefs to modify them if need be. While here we shall assume, so to speak, an infinite regress of prior beliefs from some arbitrary distribution toward uniform ignorance, which gives us a limiting special case of the prior distribution, similar results to ours could be obtained by starting with some arbitrary but not too pathological prior distribution, and collecting increasing amounts of relevant data. See Savage's discussion of his "principle of precise measurement" in [12], pp. 13-14.

---

sequence of prior probability distributions that approach uniform density as a limit, and examine the limiting Bayes solutions of the sequence.

ASSUMPTION 1 . There exists a sequence of prior probability densities

$$\left\{ \frac{1}{h}\, g, \quad \Pi_o + \frac{\Pi - \Pi_o}{h} \right\}$$

where g_1 is any bounded density over the GH-dimensional parameter space, where Π_o is a value of Π where g_1 takes on its least upper bound, and where $h = 1, 2, \ldots$

This sequence might be referred to as "dilation around Π_o " .

Each succeeding density distribution, crudely speaking, maintains a shape similar to preceding distributions, but is compressed more along the g_1 axis and stretched more throughout the parameter space in all directions away from Π_o.[20]

---

[20] I am indebted to L. J. Savage for suggesting this particular sequence to me, which is more general than a sequence I had previously postulated.

The following assumptions, which imply a normal data distribution, are made for mathematical convenience and to facilitate the comparison between Bayes and conventional estimates.

ASSUMPTION 2. The distribution of the random disturbances v in the reduced form (1.3) has the normal multivariate density $n(0, V)$.

ASSUMPTION 3. A set of T observations

$$X = (Y\ Z) = \begin{pmatrix} y_1' & z_1' \\ \cdot & \cdot \\ \cdot & \cdot \\ y_T' & z_T' \end{pmatrix}$$

is available, where z_t is a nonstochastic vector and y_t is a random vector satisfying the reduced form (1.3). The random observations y_t are independent, and all made under the same Π, and at a time when no control was exercised

over the decision variables z_D . The matrix Z is of rank H.

LEMMA 1. Under Assumptions 1, 2, and 3, as h approaches infinity, the means $\overline{\Pi}$ of the posterior distribution of Π approach the least squares estimates[21] of Π , $\hat{P}$, and the covariance matrix, V_{Π} , of the posterior

---

[21] By "least squares estimates" are meant the classical or "simple" least squares estimates -- namely the elements of the matrix $\hat{P}$ that minimizes

$$\sum_{t=1}^{T} (y_{ti} - \hat{y}_{ti})^2$$

simultaneously for $i = 1 \ldots g$; where y_{ti} and $\hat{y}_{ti}$ are respectively the elements of Y and $\hat{Y}$; and where $\hat{Y}' = PZ'$.

---

distribution of Π approaches that of the least squares estimates, $V_{\hat{P}}$, thus:

$$(4.1) \qquad \lim_{h\to\infty} \overline{\Pi} = \hat{P} \quad ; \qquad \lim_{h\to\infty} V_{\Pi h} = V_{\hat{P}} \quad .$$

The proof of this lemma is given in Appendix A (Lemma A2) .

The solution and conventional estimates. -- With the notion of a sequence of prior distributions (and hence of posterior distributions) we may associate the notion of a limiting Bayes solution, which is the limit that the sequence of Bayes solutions approaches.

DEFINITION 11. Let $\{g_h(\Pi)\}$ denote a sequence of prior distributions. Let (P_h^* , z_{Dh}^*) denote a Bayes solution obtained when the personal probability distribution is considered to be the posterior distribution associated with $g_h(\Pi)$.

A limiting Bayes solution[22] is an estimate-decision pair (P^*, z_D^*) such that

$$\lim_{h\to\infty} z_{Dh}^* = z_D^* \quad . \tag{4.2}$$

---

[22] cf. Wald's "Bayes solution in the wide sense," [13], pp. 16-17.

DEFINITION 12. A prior-ignorance Bayes solution is a limiting Bayes solution under the particular sequence of prior distributions postulated in Assumption 1.

The following Lemma provides a means of relating prior-ignorance Bayes solutions to least squares estimates.

LEMMA 2. In any formula for a Bayes solution in Section 3 of this paper,[23]

---

[23] Recall that "Bayes solution" includes both optimal decisions and optimal implied estimates.

---

if the following substitutions are made:

(a) the means $\overline{\Pi}$ of the personal probability distribution[24] are replaced by the least squares estimates $\hat{P}$; and

---

[24] The formulas may contain these moments implicitly, rather than explicitly. For example, in (3.6) it is shown how the expression $\overline{\Pi_D' C \Pi_E}$ involves elements of $\overline{\Pi}$ and V_Π .

(b) the elements of the covariance matrix V_{Π} of the personal probability distribution[24] are replaced by corresponding elements of the covariance matrix $V_{\hat{P}}$ of the least squares estimates $\hat{P}$;

---

[24] Ibid.

---

then the formula becomes one for a prior-ignorance Bayes solution under Assumptions 1, 2, and 3 .

PROOF. In view of Definition 11, it is sufficient to consider optimal decisions. All of those presented in Section 3 are unique, and the formulas contain moments of the personal probability distribution no higher than the second order, i.e., $\overline{\Pi}$ and V_{Π} . Let these moments (which are now to be regarded as posterior moments) when associated with the prior distribution $g_h(\Pi)$, be denoted by $\overline{\Pi}_h$ and $V_{\Pi h}$. Then, from Lemma 1, as $h \to \infty$, these mements approach $\hat{P}$ and $V_{\hat{P}}$, respectively. Then it can be seen that in every formula for an optimal decision z_D^* , the limit $\lim_{h \to \infty} z_{Dh}^*$ exists, is unique, and is given by the substitute formula described.

Let us look at the prediction problem first, where the results are well-known. From formula (3.17) it can be seen that the least squares estimates are precisely prior-ignorance optimal implied estimates. This is evident also from the facts that the least squares estimates are also Maximum Likelihood

Estimates, that the posterior distribution tends to the likelihood function, which is symmetric, and that the means of the posterior distribution minimize the risk as given by (3.16). It also is true that the least squares estimates are also Best Linear Unbiased Estimates (see Appendix B).

In the decision problem, classical least squares estimates do not constitute optimal implied estimates under prior ignorance. This can be seen by making the substitutions of Lemma 2 into formula (3.7) and also its special cases. The discrepancies are the same as those between the optimal implied estimates of the decision problem and those of the prediction problem, which have been discussed above.

Consider the simple one-equation, one-decision case of formula (3.10). Under Assumptions 1, 2, and 3, $\bar{\pi}$ becomes the least squares estimate $\hat{p}$. Then (3.10) may be written

$$(4.3) \qquad p^* = \hat{p}\left(1 + \frac{\operatorname{var}\hat{p}}{\hat{p}^2}\right) = \hat{p}\left(1 + \frac{1}{t^2}\right),$$

where t is the conventional "t-ratio," the rate of a regression coefficient to its standard error.[25] The optimal implied estimate is here seen to be an

---

[25] It is still a t-ratio although we are assuming the standard error known -- a special case of the more usual situation where an unbiased estimate of the standard error is in the denominator.

"inflated" least squares estimate -- the least squares estimate multiplied by a positive factor involving the reciprocal of the t-ratio.

Consider the case of the two-equation, one-decision model of formula (3.11) and its special case graphed in Figure 1. The point $(\bar{\pi}_1, \bar{\pi}_2)$ becomes the pair of least squares estimates $(\hat{p}_1, \hat{p}_2)$, which is located at point (1, 2) on the graph, while the prior-ignorance optimal implied estimates are represented by any point on the circle (except the origin). Again, the optimal implied estimates may be obtained by "inflating" each least squares estimate by a positive factor. On the graph, the inflated estimates are found at the point where the circle is intersected by a ray from the origin passing through the point $(\bar{\pi}_1, \bar{\pi}_2)$. Algebraically, they are

$$(4.4) \qquad (p_1^* \; p_2^*) = \lambda \, (\hat{p}_1 \; \hat{p}_2)$$

where λ is the scalar inflator found by substituting (4.4) into (3.14); by setting

$$(4.5) \qquad (\bar{\pi}_1 \; \bar{\pi}_2) = (\hat{p}_1 \; \hat{p}_2)$$

and

$$(4.6) \qquad \begin{pmatrix} \text{var } \pi_1 & \text{cov}(\pi_1, \pi_2) \\ \text{cov}(\pi_1, \pi_2) & \text{var } \pi_2 \end{pmatrix} = V_{\hat{p}}$$

in (3.13), and equating (3.13) to (3.14). The result is that the scalar inflator is

$$(4.7) \qquad \lambda = \left| + \frac{\text{tr } C \, V_{\hat{p}}}{\hat{p}' C \, \hat{p}} \right| = \left| + \frac{1}{t^2} \right| \quad ,$$

where $\bar{t}^2$ is a suitably weighted quadratic mean of the individual t-ratios, the weights being functions of the variances, covariances, and elements of the C matrix. In Figure 1 the separate t-ratios are 1 and 2, C is the identity matrix, and the inflator is 1.4.

It is found that when only one decision variable appears in the model (D = 1), the prior-ignorance optimal implied estimates can be expressed by scalar inflators of the least squares estimates, even when some exogenous variables appear in z_E. But in the general case scalar inflators do not necessarily exist, although optimal implied estimates exist that are more complicated linear functions of the least squares estimates.

Practical differences. -- The differences between the two kinds of estimates have been stressed. The question may well be asked, however: how important are these differences? If an investigator wished to obtain prior-ignorance estimates that would lead to optimal decisions, would he go seriously astray by using conventional least squares estimates?

The preceding discussion of inflated least squares estimates throws some light on this question, but does not answer it decisively for large models. The answer depends on the accuracy of the least squares estimates. The more accurate they are (the higher the t-ratios), the less difference there is between least squares estimates and optimal implied estimates. The one-equation, one-decision model furnishes a crude example. The relationship between the t-ratio of the single regression coefficient and the inflator needed to convert the least squares estimate into the optimal implied estimate is shown in the following table.

t-ratio	inflator
1	2
2	1.25
3	1.11
4	1.06
5	1.04

This relationship is suggestive, even for larger models. With very low t-ratios, lacking statistical significance, the question of the magnitude of the difference between the two types of estimate is probably irrelevant, as the investigator will not have confidence in his least squares estimates anyway. For very high t-ratios, the magnitude of the difference will be very small, and least squares estimates could be used with confidence. It is when the t-ratios are on the borderline of significance -- say around 2 or so -- that the question of the discrepancy has relevance and the size of the discrepancy may be appreciable.

5. CONCLUSION

Summary of results. -- A decision model is postulated, which includes a linear reduced form with some of the non-stochastic variables under control of decision-maker, and a welfare function that is quadratic in the endogenous variables. For this model the loss function of statistical decision theory has been derived -- a function showing the cost in terms of lost welfare of making erroneous estimates P, and consequent erroneous decisions, when the true parameters are Π. This loss function is found to be different from another loss function suggested for the prediction problem with the same reduced form, a function representing the cost of making errors in forecasting the endogenous

variables when all of the non-stochastic variables are given and not under control.

The difference between the two loss functions derives from the fact that in the decision model some variables are under control, while in the prediction model all are uncontrolled. The difference in loss functions means that the optimal statistical estimates of the reduced form parameters in the two models would be expected to be different.

Optimal estimates and decisions are defined, using Bayesian concepts, as those that minimize expected loss. In the decision model, an optimal estimate is the implied estimate that, if used as a certainty equivalent by the decision-maker, would lead him to make an optimal decision. It is found that in the decision model the optimal estimates are different than in the prediction model. In the decision model, except in the simplest special case, the optimal estimates involve the constants of the welfare function; in the prediction model they do not. In the decision model, therefore, the statistician doing the estimating of the parameters needs to know at least some of the constants of the welfare function, in order to provide the decision-maker with optimal estimates.

When a limiting approach to Bayes' postulate of equal prior probabilities of the unknowns is made, along with usual data assumptions including normality of disturbances, the limiting optimal estimates in the prediction model are equivalent to the classical least squares estimates; in the decision model they are not; but they may be regarded as inflated, or adjusted least squares estimates, their closeness to the least squares estimates depending directly on the accuracy of the estimates.

Appraisal. -- The question naturally arises as to whether the results summarized above -- or similar results -- will prevail when the restrictive assumptions made are relaxed or generalized. These assumptions include those regarding personal probability, the structure of the reduced form, and the form of the welfare function. While this question cannot be answered definitely at present, indications are that the main point of this paper -- the showing of a fundamental difference in principle between the decision model and the prediction model -- will not be changed.

The use made here of the Bayesian concept of a personal probability distribution representing the investigator's degree of belief in alternative values of unknown parameters, and -- specifically -- the postulate of a sequence of prior distributions approaching uniform density -- has made it possible to derive specific optimal estimates in Sections 3 and 4. While these results are believed to be useful, and while the Bayesian apparatus has helped the author to gain insight into the main problem, the differences between decision and prediction models do not derive from this apparatus. Rather, they derive from the difference in loss functions -- that is to say, from the structure of the problems themselves -- the nature of the objectives of decision-maker and forecaster and their degree of control over variables in the economic model. A change in the prior distribution, for example, would somewhat alter the specific results, but there would still be a difference between optimal estimates in the decision model from those in the prediction model.

The same considerations indicate that generalizing the assumptions regarding the reduced form -- linearity, no over-identifying restrictions on

parameters, normality and serial independence of disturbances, known covariance matrix of disturbances -- would still leave differences between appropriate estimates in decision model and prediction model, although the estimates themselves would be expected to be different. It would be highly desirable to work out results for more general cases.

It should be noted that the specification of additional restrictions on the unknown parameters will affect the prior distribution. The presence of the usual over-identifying restrictions on them (from a model of structural equations) will place them in a domain of lower dimensionality. The placing of inequality constraints on them (usually quite realistic) will be inconsistent with the infinitely dilating sequence of prior distributions postulated in Section 4. If a bounded domain of the unknown parameters is in order, a rectangular prior distribution might be specified.

The relaxing of the assumption of known covariance matrix of disturbances will place this covariance matrix in the category of unknown parameters. While this change introduces some touchy questions with respect to the prior distribution of a variance, some preliminary work by the author on this aspect indicates that reasonable prior distributions exist, and that an estimated covariance matrix from the sample can be used.

The form of the welfare function will also affect the specific results, but it is hard to see how a different welfare function -- polynomial or logarithmic, for example -- would change the essential points that the optimal estimates for the decision and prediction models differ, and that some knowledge of the welfare function is needed by the estimating statistician. A broader

question would be whether it is realistic or useful to postulate a welfare function at all. This is an old question. While recognizing the worth of raising the question, and the difficulties of formulating welfare functions for large social groups, the author can only here state his belief that the concept has fundamental relevance.

As stated previously, the results of this paper do not necessarily point to large discrepancies from customary statistical procedures, nor from statistical criteria for "good" estimates. In many situations the conventional procedures will be fairly consistent with optimal decisions. Yet, conceptual differences exist, and occasionally practical differences will occur.

APPENDIX A. A Bayesian Interpretation of Multivariate Least Squares

The purpose of this Appendix is to show that the classical least squares estimates, $\hat{P}$, of Π in the reduced form of the text and their covariance matrix are also the limits approached by the means and covariance matrix of the posterior distribution of Bayesian theory when the prior distribution approaches a uniform density and when the disturbances are normal. The demonstration is made in two steps: (1) A general Bayesian set-up is described in which the moments of the posterior distribution approach those of the likelihood distribution; (2) the least squares estimates are shown to fit into the general set-up. The use made here of limiting processes in a multivariate situation is believed to constitute some generalization of previous results.

Bayesian set-up. -- The unknown parameter, Θ, is a random point or vector defined over Euclidean n-space, R_n. The prior distribution is a probability density, $\xi(\Theta)$ defined over R_n whose first and second moments exist. The data, x, is a random point or vector defined over Euclidean N-space, R_N. The data distribution is a conditional probability density, $p(x|\Theta)$, over R_N, whose first and second moments exist. The likelihood distribution (sometimes called just likelihood) is the probability density function, $p_x(\Theta)$, obtained from the data distribution $p(x|\Theta)$ by regarding Θ as a random variable, x as fixed, and multiplying by a suitable constant so that $\int_{-\infty}^{\infty} (\Theta)d\Theta = 1$. The posterior distribution is the conditional probability density, $g_x(\Theta)$, of the parameter Θ, given the observation of a particular set of data, x, this distribution having mean

$\bar{\Theta}$, a n x 1 vector, and covariance matrix V_Θ , a n x n matrix.

BAYES THEOREM. (cf, for example, Kendall [14], p. 176). The posterior distribution is proportional to the product of the prior distribution and the likelihood:

$$(A1) \qquad g_x(\Theta) = \frac{g(\Theta)\ p_x(\Theta)}{\int_{-\infty}^{\infty} g(\Theta)\ p_x(\Theta)\ d\Theta} \ .$$

ASSUMPTION A1. There exists a sequence of prior probability distributions

$$\{g_h(\Theta)\} = \{h^{-1} g_1 [\Theta_0 + h^{-1} (\Theta - \Theta_0)]\}$$

over R_n , where $h = 1, 2, \ldots,$ and where $g_1(\Theta)$ is any bounded density whose first and second moments exist, and where Θ_o is a value of Θ where g_1 takes on its least upper bound.

LEMMA A1. Under Assumption A1, as $h \to \infty$, all moments of the posterior distribution $g_x(\Theta)$ approach as a limit the corresponding moments of the likelihood $p_x(\Theta)$.

PROOF.[26] Let the positive integer m denote the order of the highest

---

[26] For the method of proving this lemma I am indebted to L. J. Savage, who is, however, not responsible for what I have done to his original suggestions.

moment of $g_x(\Theta)$ that it is desired to compute. Let $q(\Theta)$ denote the vector whose elements are those functions of Θ that, when integrated,[27] give all of

---

[27] All of the scalar integrals in this paper are multiple integrals.

---

the moments of $p_x(\Theta)$ around zero of orders 1 through m inclusive, i.e.,:

$$(A2)\quad q(\Theta) = \{(\Theta_i^{r1}\, p_x)\ (\Theta_i^{r2}\, \Theta_j^{s2}\, p_x)\ (\Theta_i^{r3}\, \Theta_j^{s3}\, \Theta_k^{t3}\, p_x)\ \ldots\}$$

where the quantities in parentheses are subvectors with i, j, k, ...= 1 ... n;

$$r_1 = 1,\quad r_2+s_2 = 2,\quad r_3 + s_3 + t_3 = 3,\ \ldots,\ r_m + s_m + t_m + \ldots = m\ .$$

Then the vector whose elements are the moments of $p_x(\Theta)$ around zero corresponding to all desired moments of $g_x(\Theta)$ is[28]

$$\int q = \int_{-\infty}^{\infty} q(\Theta)\ d\,\Theta\ .$$

---

[28] In order to avoid repetition of symbols the following abbreviations will be adopted from this point on in this Appendix: $q(\Theta) = q$; $p_x(\Theta) = p$; $g(\Theta) = g$; for any scalar or vector function of Θ, $f(\Theta), \int_{-\infty}^{\infty} f(\Theta)d\Theta = \int f$, where the integral sign over a matrix means the matrix whose elements are scalar integrals of each element of the matrix.

With respect to the prior distribution g_h, the moments of the posterior distribution g_{xh} that correspond to $\int q$ may, from Bayes Theorem, be represented by

$$\frac{\int q\, g_h}{\int p g_h}$$

For any vectors u and v, let $|u - v|$ denote the Euclidean distance between the two points. Then to prove the Lemma it is necessary and sufficient to prove that

$$\text{(A3)} \qquad \lim_{h\to\infty} \left| \frac{\int qgh}{\int pgh} - \int q \right| = 0 .$$

Let ϵ be an arbitrarily small number. Then, for any Θ there exists an h large enough so that

$$\text{(A4)} \qquad \frac{g_h(\Theta)}{g_h(\Theta_o)} = \frac{g_1[\Theta_o + h^{-1}(\Theta-\Theta_o)]}{g_1(\Theta_o)} \leq 1 + \epsilon .$$

This is so because of the continuity of the density. By multiplying (A4) by q and by p, in turn, and integrating, it is found that (A4) implies

$$\text{(A5)} \qquad \frac{\int q\, g_h}{g_h(\Theta_o)} \leq (1 + \epsilon) \int q ;$$

$$\text{(A6)} \qquad \frac{\int p\, g_h}{g_h(\Theta_o)} \leq 1 + \epsilon .$$

Subtracting $\int q$ from (A5) and 1 from (A6) and taking distances from 0, (A4) is found to imply

$$\text{(A7)} \qquad \left| \frac{\int q\, g_h}{\int g_h(\Theta_o)} - \int q \right| \leq |\epsilon| \int |q| \; ;$$

$$\text{(A8)} \qquad \left| \frac{\int p g_h}{g_h(\Theta_o)} - 1 \right| \leq |\epsilon| \; .$$

In general, a vector, v, and a scalar, s, may be chosen so that

$$\text{(A9)} \qquad \frac{\int q g_h}{g_h(\Theta_o)} - \int q = v \int |q| \; ; \quad \text{and}$$

$$\text{(A10)} \qquad \frac{\int p g_h}{g_h(\Theta_o)} - 1 = s \; .$$

By the properties of distance, moreover, it may be verified that (A9) and (A10) imply

$$\text{(A11)} \qquad \left| \frac{\int q\, g_h}{\int p g_h} - \int q \right| \leq \frac{(|v| + |s|) \int |q|}{|1 + s|} \; .$$

By comparing the distance from 0 of the left sides of (A9) and (A10) with (A7) and (A8) respectively, it is seen that there exists an h large enough so that $|v|$ and $|s|$ may both be made smaller than an arbitrarily small

positive number $|\epsilon|$. So the limit of the left side of (A11) as $h \to \infty$ is zero, which establishes (A3), which proves the lemma.

Least squares estimates. -- The reduced form (1.3) of the text

$$y = \Pi z + v$$

is now considered where the disturbance vector v has the normal multivariate distribution $n(0,V)$, and where T independent observations on y are available, in accordance with Assumptions 2 and 3 of the text. The joint likelihood function of the unknown parameters Π under these conditions is well known.[29]

---

[29] See, for example, Anderson [16], formula (22), p. 183.

Although it is not customary to regard the unknown parameters as random variables, nor the likelihood function as a probability distribution, by multiplying by a suitable normalizing constant, the likelihood function can in fact be put in the form of a normal multivariate density function of the elements of Π . The means of this density function are the classical least squares estimates $\hat{P}$, as computed from the observations at hand, and the covariance matrix of the elements of Π is the $GH \times GH$ matrix

$$\text{(A12)} \qquad \Sigma_{\hat{P}} = [\text{cov}(v_i,v_j)(Z'Z)^{-1}] = V \otimes (Z'Z)^{-1} ,$$

where V is the covariance matrix of the disturbance vector v , with typical element $\text{cov}(v_i,v_j)$, Z is the $T \times H$ data matrix on z , and $\otimes$ denotes the

Kronecker product, or direct product, operation.[30] Of course, as is well known,

---

[30] Ibid.

$\hat{P}$ is a maximum-likelihood estimate of Π .

A correspondence may now be made between this and the preceding section of this Appendix. Let $n = GH$, $N = GT$; let Θ denote the GH x 1 vector formed by placing all the elements of Π in a single column; let x denote the GT x 1 vector formed by placing the vectors y_t end to end in a single column. The likelihood density of the previous section $p_x(\Theta)$ is therefore the normal multivariate density $n(\hat{P}, V_{\hat{P}})$ just described in the preceding paragraph. The posterior means $\bar{\Theta}$ and covariance matrix V_Θ of the preceding section become $\bar{\Pi}_h$ and $V_{\Pi h}$. Assumption A1 becomes Assumption 1 of the text. Lemma A1 may therefore be applied to the moments around 0 of the posterior distribution $g_{xh}(\Pi)$. Hence we have the following result.

LEMMA A2. Under Assumptions 1, 2, and 3 of the text,

$$\lim_{h\to\infty} \bar{\Pi}_h = \hat{P} \quad , \tag{A13}$$

$$\lim_{h\to\infty} V_{\Pi h} = V_{\hat{P}} \tag{A14}$$

where $\hat{P}$ is the matrix of classical least squares estimates and $V_{\hat{P}}$ is given by (A12).

The proof of this lemma is immediate when it is remembered that the means and the elements of the covariance matrix are obtainable by simple subtraction from the first and second moments around zero. This lemma is Lemma 1 of the text.

APPENDIX B. A Note on Best Linear Unbiased Estimates in the Prediction Problem

The purpose of this note is to point out that under the data assumptions made in this paper, which are the usual ones, including serial independence of data, each of two different definitions of "best linear unbiased estimate" of the predictand, $E(y)$, in the prediction problem, lead to simple least-squares estimation of the individual equations one by one.

Generalized least squares. -- Let β be a fixed unknown $k \times 1$ vector, X a fixed and known $n \times k$ matrix of rank k, y a random $n \times 1$ vector with $E(y) = X\beta$ and with known covariance matrix Ω. Let α be the unknown parameter vector $\alpha = \psi\beta$ to be estimated, where ψ is a fixed and known $r \times k$ matrix, and let y^o be one observation that is available on y. Assume that it is desired to find an estimate, a, of α that is (1) linear in y^o; (2) unbiased; and that satisfies either condition (3a) or (3b).

(3a) For any symmetric positive-definite matrix Q the quadratic form $E(a - \alpha)' Q(a - \alpha)$ is minimized, subject to conditions (1) and (2).

(3b) The mean squared errors $E(a_i - \alpha_i)^2$ are minimized simultaneously for $i = 1 \ldots r$, subject to conditions (1) and (2).

By a slight generalization of the result given by Theil ([1], Appendix 8A), the desired estimate, using condition (3a) is found to be given by the generalized least squares formula

$$\text{(B1)} \qquad \hat{a} = \psi(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y^o \quad .$$

This formula can be derived by following Theil's derivation exactly, replacing his condition $AX = I$ by the condition $AX = \psi$, implied by the unbiased condition $E(a) = \alpha$, and making other minor notation changes.

Formula (B1) has been derived under condition (3b) by Chipman and Rao ([17] Theorem 2.1, p. 11). Since in both Theil's and Chipman and Rao's derivations the formula (B1) is shown to be both necessary and sufficient for the desired estimate $\hat{a}$, it follows that conditions (3a) and (3b) are logically equivalent to each other.

Prediction from reduced form. -- Consider now the prediction problem suggested in this paper of predicting $E(y)$, where y is given by the reduced form (1.3) and where z is a given vector of exogenous variables in the prediction period. Say that the data available, from a previous period, conform to Assumption 3 of the text (Section 4), and that it is desired that the prediction, $\hat{y}$, minimize the loss function (2.15) while at the same time being linear in the random data Y and unbiased.

This problem may be placed into correspondence with the generalized least squares problem by letting $G = r$, $GH = k$, $GT = n$, $E(y) = \alpha$, $\hat{y} = a$, and rewriting certain matrices as vectors. It happes that because of the special nature of the assumption of independent observations, the covariance matrix V cancels, and the following result is obtained:

$$\text{(B2)} \qquad \hat{y} = Y'Z(Z'Z)^{-1} z \ ,$$

where Y and Z are the data matrices reconstituted as defined in Assumption 3. This is the simple least squares result.

In accordance with the result cited from Chipman and Rao, the formula (B2) also holds if, instead of minimizing the loss function (2.15), it had been required to minimize the mean squared error (or variance) of each $\hat{y}_i$ separately, subject to linearity and unbiasedness. The two alternative definitions of "bestness" are logically equivalent.

References

[1] Theil, H. Economic Forecasts and Policy, Amsterdam, North Holland, 1958, Ch.8.

[2] Frisch, Ragnar, "L'Emploi des Modèles pour l'Élaboration d'une Politique Économique Rationelle," Revue d'Economie Politique, vol. 60 (1950), pp. 474-498, 601-634.

[3] Tinbergen, J. On the Theory of Economic Policy, Amsterdam, North Holland, 1952.

[4] Tinbergen, J. Economic Policy: Principles and Design, Amsterdam, North Holland, 1956.

[5] Sverdrup, E., "Prediction Problems and the Theory of Statistical Decision Functions" (abstract), Econometrica, vol. 19 (1951), p.61.

[6] Hurwicz, L., "Aggregation in Macroeconomic Models" (abstract), Econometrica, vol. 20 (1952), p. 489.

[7] Malinvaud, E., "l'Agrégation dans les Modèles Économiques" in R. Roy, ed., Cahiers du Sémmaire d'Économetrie No.4, Paris, 1956.

[8] Marschak, J. "Economic Measurements for Policy and Prediction," in W.C.Hood and T.C. Koopmans, ed., Studies in Econometric Method, New York, John Wiley, 1953, Ch. 1.

[9] Koopmans, T.C. and Hood, W.C., "The Estimation of Simultaneous Linear Economic Relationships," in W.C. Hood and T.C. Koopmans, ed., Studies in Econometric Method, New York, John Wiley, 1953, Ch. 6.

[10] Sverdrup, E., "Weight Functions and Minimax Procedures in the Theory of Statistical Inference," Archiv fur Mathematik og Naturvidenskab, vol. 51 (1952), pp. 1-76.

[11] Savage, L.J., The Foundations of Statistics, New York, John Wiley, 1954.

[12] Savage, L.J.,"The Foundations of Statistics Reconsidered," Fourth Berkeley Symposium, 1960.

[13] Wald, A., Statistical Decision Functions, New York, John Wiley, 1950.

[14] Kendall, M.G., The Advanced Theory of Statistics, vol.I, London, Griffin, 1943

References

[15] Jeffreys, H., Theory of Probability, Oxford, Clarendon, 1939.

[16] Anderson, T. W., Introduction to Multivariate Statistical Analysis, New York, Wiley, 1958.

[17] Chipman, J.S. and Rao, M.M., "On the Use of Idempotent Matrices in the Treatment of Linear Restrictions in Regression Analysis," (Technical Report No. 10, University of Minnesota, 1960 (Processed).

[18] Radner, R., "Minimax Estimation for Linear Regressions," Annals of Mathematical Statistics, Vol. 29 (December 1958), pp. 1244-1250.

[19] Klein, L. R., "The Efficiency of Estimation in Econometric Models," Essays in Economics and Econometrics, Chapel Hill, University of North Carolina Press, 1960.