The Annals of Statistics
1973, Vol. 1, No. 3, 506-516

THE CHOICE OF VARIABLES FOR PREDICTION IN CURVILINEAR MULTIPLE REGRESSION

By R. J. Brooks

University College, London

A Bayesian formulation of the problem of analysing data from a curvilinear regression of y on $x_1, x_2, \cdots, x_r$ in order to predict a future value of y is considered. The problem is to obtain a criterion to decide which is the best subset of $x_1, x_2, \cdots, x_r$ to perform this prediction. Under very strict assumptions the criterion obtained is shown to use the same statistic as the orthodox (least squares) approach.

1. Introduction. The type of problem to be discussed in this paper can be illustrated by an example from Johnson and Leone ((1964) page 313). A rubber manufacturing company performed an experiment in which interest was focused on the relationship of the tear strength of a particular type of rubber with three variables (i) percentage of component A in rubber, (ii) percentage of component A in resin and (iii) percentage modifier. If these three variables are denoted by x_1, x_2 and x_3 respectively and the tear strength by y, then a quadratic regression model of y on x_1, x_2, x_3 was investigated. It was found that the values of the coefficients of x_3, x_3^2, x_1x_3 and x_2x_3 were all not significantly different from zero (using a 5% significance level). This could be taken to suggest that x_3 need not be included in the model and if we wanted to predict a future value of y we need only use (x_1, x_2). In this paper, this type of problem is investigated, but from a Bayesian viewpoint. If the model of interest is a curvilinear regression of y on $x_1, x_2, \cdots, x_r$ and the object in mind is to predict a future value of y, we have to decide which is the best subset of $x_1, x_2, \cdots, x_r$ to use for this prediction.

We use the decision-theoretic approach of Lindley (1968), who discussed the same problem for linear regression of y on $x_1, x_2, \cdots, x_r$. Consequently, the first four assumptions stated below are generalizations of those made by Lindley.

Let $\mathscr{E}$ denote the results of an experiment to investigate the regression plus $\mathscr{E}_0$, which denotes any knowledge prior to the experiment.

Assumption 1. $\boldsymbol{\theta}$ and $\mathbf{x}$ are two random variables in $\mathscr{R}^u$ and $\mathscr{R}^r$ respectively (Euclidean spaces of u and r dimensions) which are independent, given $\mathscr{E}$.

Assumption 2. y is a random variable in $\mathscr{R}^1$ with density $p(y \mid \boldsymbol{\theta}, \mathbf{x}, \mathscr{E}_0)$ whose form does not depend on the results of the experiment, such that

$$E(y \mid \boldsymbol{\theta}, \mathbf{x}, \mathscr{E}_0) = \boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) \tag{1}$$

and

$$\operatorname{Var}(y \mid \boldsymbol{\theta}, \mathbf{x}, \mathscr{E}_0) = \sigma^2,$$

a known constant. Here, if $\mathbf{x} = (x_1, x_2, \cdots, x_r)^T$, ξ is a known integer (> 0),

Received March 1971; revised May 1972.

$u = \binom{\xi+r}{r}$ and $A = \{$all sets $(\alpha_1, \alpha_2, \cdots, \alpha_r)$: α_i is integer $\geqq 0$, $0 \leqq \sum_{i=1}^{r} \alpha_i \leqq \xi\}$, $\boldsymbol{\phi}(\mathbf{x})$ is a $u \times 1$ vector with typical element $x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_r^{\alpha_r}$ where $(\alpha_1, \alpha_2, \cdots, \alpha_r) \in A$. The corresponding element of $\boldsymbol{\theta}$ is defined to be $\theta_{\alpha_1\alpha_2\cdots\alpha_r}$.

Equation (1) can then be written as

$$E(y \mid \boldsymbol{\theta}, \mathbf{x}, \mathscr{E}_0) = \sum_A \theta_{\alpha_1\alpha_2\cdots\alpha_r} x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_r^{\alpha_r} .$$

An example would be the two variable quadratic regression (i.e. $r = \xi = 2$),

$$E(y \mid \boldsymbol{\theta}, \mathbf{x}, \mathscr{E}_0) = \theta_{00} + \theta_{10}x_1 + \theta_{01}x_2 + \theta_{20}x_1^2 + \theta_{11}x_1x_2 + \theta_{02}x_2^2 \tag{2}$$

where $\boldsymbol{\theta} = (\theta_{00}, \theta_{10}, \theta_{01}, \theta_{20}, \theta_{11}, \theta_{02})^T$ and $\boldsymbol{\phi}(\mathbf{x}) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2)^T$.

In these two assumptions, $\mathbf{x}$ and y refer to the future values of the independent and dependent variables, respectively.

Lindley (1968) discussed the special case $\xi = 1$, which in his notation can be written for the u variable linear regression as $E(y \mid \boldsymbol{\theta}, \mathbf{x}, \mathscr{E}_0) = \theta_1x_1 + \theta_2x_2 + \cdots + \theta_ux_u$, where we put $x_1 = 1$ to allow for the constant term. If we replace the elements of the $u \times 1$ vector $(x_1, x_2, \cdots, x_u)^T$ by the elements of the $u \times 1$ vector $\boldsymbol{\phi}(\mathbf{x})$, then the initial part of Lindley's analysis, up to his expression (11), holds for the regression model we are considering. Lindley then assumes (Assumption 5) that $p(\mathbf{x} \mid \mathscr{E})$, which with our model will be $p(\boldsymbol{\phi}(\mathbf{x}) \mid \mathscr{E})$, has linear regressions. Clearly this does not hold for $\xi > 1$, so the remainder of his prediction analysis is not applicable to curvilinear regression.

Let I denote a subset of the integers $1, 2, \cdots, r$ containing s members $(0 \leqq s \leqq r)$, and J its complement. Define $\mathbf{x}_I$ as the vector with elements x_i, $i \in I$, and similarly define $\mathbf{x}_J$. Define $\boldsymbol{\phi}_I(\mathbf{x}_I)$ as the $v \times 1$ vector of all the elements of $\boldsymbol{\phi}(\mathbf{x})$ which do not contain any elements or function of elements belonging to $\mathbf{x}_J$. Let the remaining elements of $\boldsymbol{\phi}(\mathbf{x})$ be contained in $\boldsymbol{\phi}_J(\mathbf{x})$. Define vectors $\boldsymbol{\theta}_I$ and $\boldsymbol{\theta}_J$ corresponding to $\boldsymbol{\phi}_I(\mathbf{x}_I)$ and $\boldsymbol{\phi}_J(\mathbf{x})$ respectively.

In the above example when $r = \xi = 2$, these definitions would mean that when $I = (1)$, $\boldsymbol{\phi}_I(\mathbf{x}_I) = (1, x_1, x_1^2)^T$ and $\boldsymbol{\phi}_J(\mathbf{x}) = (x_2, x_1x_2, x_2^2)^T$ so that $\boldsymbol{\theta}_I = (\theta_{00}, \theta_{10}, \theta_{20})^T$ and $\boldsymbol{\theta}_J = (\theta_{01}, \theta_{11}, \theta_{02})^T$.

ASSUMPTION 3. The decision space consists of elements $(I, f(\cdot))$, where $f(\cdot)$ is a function from $\mathscr{R}^s$ to $\mathscr{R}^1$.

ASSUMPTION 4. The loss in predicting y using the subset of variables $\mathbf{x}_I$ giving the prediction $f(\mathbf{x}_I)$ is

$$[y - f(\mathbf{x}_I)]^2 + c_I ,$$

where $c_I(\geqq 0)$ is the cost of observing the variables in $\mathbf{x}_I$ per cost of unit error in prediction.

The loss function is assumed to consist of additive terms arising from the cost of the error $|y - f(\mathbf{x}_I)|$ in prediction and the cost of observing those variables in $\mathbf{x}_I$. The cost of error in prediction is assumed to be proportional to $[y - f(\mathbf{x}_I)]^2$.

Using results obtained by Lindley ((1968) (8) and (11)), we have:

LEMMA 1. *Under Assumptions* 1 *to* 4, (a) *the best set of independent variables to use to predict* y *are those* x_i *with* $i \in I$, *where* I *is chosen to be that subset of* $(1, 2, \cdots, r)$ *which minimizes*

$$E(\boldsymbol{\theta})^T \mathbf{V}[\boldsymbol{\phi}_J(\mathbf{x})] E(\boldsymbol{\theta}) + c_I \tag{3}$$

where

$$\mathbf{V}[\boldsymbol{\phi}_J(\mathbf{x})] = E([\boldsymbol{\phi}(\mathbf{x}) - E\{\boldsymbol{\phi}(\mathbf{x}) \mid \mathbf{x}_I\}][\boldsymbol{\phi}(\mathbf{x}) - E\{\boldsymbol{\phi}(\mathbf{x}) \mid \mathbf{x}_I\}]^T) ;$$

(b) *the optimum predictor of* y *is* $E(\boldsymbol{\theta})^T E[\boldsymbol{\phi}(\mathbf{x}) \mid \mathbf{x}_I]$.

In (a) *and* (b), *all expectations and dispersion matrices are conditional upon* $\mathscr{E}$.

2. The regression experiment. To choose the best set of variables, it will clearly be useful to simplify (3). To do this we need to know the density $p(\mathbf{x} \mid \mathscr{E})$. Consequently, we now discuss the nature of the experiment we conduct leading to the information denoted by $\mathscr{E}$.

The situations described by Assumptions 5a and 5b both lead to $p(\mathbf{x} \mid \mathscr{E})$ belonging to the same family of densities.

ASSUMPTION 5a. In the experiment, n independent observations $y_1, y_2, \cdots, y_n$ are obtained at values $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ of the independent variables, where y_k has density $p(y_k \mid \boldsymbol{\theta}, \mathbf{x}_k, \mathscr{E}_0)$ given in Assumption 2. Here $\mathbf{x}_k = (x_{k1}, x_{k2}, \cdots, x_{kr})^T$, $(k = 1, 2, \cdots, n)$. The $\mathbf{x}_k$ are independent random variables from a common multinormal distribution with unknown mean vector $\boldsymbol{\mu}$ and unknown dispersion matrix $\boldsymbol{\Sigma}$. Furthermore, the future value $\mathbf{x} = (x_1, x_2, \cdots, x_r)^T$ (see Assumption 1) is another independent value from the same distribution. Prior to the experiment, the parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1})$ are independent of $\boldsymbol{\theta}$ and have a Normal-Wishart distribution (as defined by Ando and Kaufman ((1965) Section 1.3)).

ASSUMPTION 5b. This is as Assumption 5a, but the $\mathbf{x}_k$ are selected. ($\mathbf{x}$ still has the distribution given in Assumption 5a.)

The experiment discussed in detail by Lindley ((1968) Section 4) is a special case of that described by Assumption 5a, but with a diffuse prior for $(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1})$, i.e. $p(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1} \mid \mathscr{E}_0) \propto |\boldsymbol{\Sigma}|$.

LEMMA 2. *Under Assumption* 5a *or* 5b, $p(\mathbf{x} \mid \mathscr{E})$ *is a Student density on* $\mathscr{R}^r$ *(the parameters in the two cases are different).*
(Note: throughout this paper, the definition of the Student density is that given by Raiffa and Schlaifer (1961) page 256.)

PROOF. Consider the situation of Assumption 5b. This is a designed experiment which does not give us any information about the distribution of $\mathbf{x}$; thus $p(\mathbf{x} \mid \mathscr{E}) = p(\mathbf{x} \mid \mathscr{E}_0)$. Now

$$p(\mathbf{x} \mid \mathscr{E}_0) = \int\int p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}, \mathscr{E}_0) p(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1} \mid \mathscr{E}_0)\, d\boldsymbol{\mu}\, d\boldsymbol{\Sigma}^{-1} , \tag{4}$$

where $p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}, \mathscr{E}_0)$ is the density of a multivariate normal distribution and $p(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1} \mid \mathscr{E}_0)$ is the density of a Normal-Wishart distribution. Ando and

Kaufman ((1965) Section 2.4) show that $p(\mathbf{x} \mid \mathscr{E}_0)$ is the density of a multivariate Student distribution.

In the situation of Assumption 5a (a random experiment), we have

$$p(\mathbf{x} \mid \mathscr{E}) = \int\int p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}, \mathscr{E}_0) p(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1} \mid \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n, \mathscr{E}_0)\, d\boldsymbol{\mu}\, d\boldsymbol{\Sigma}^{-1} \tag{5}$$

where $p(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1} \mid \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n, \mathscr{E}_0)$ is the density of the posterior distribution of $(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1})$ which is Normal-Wishart (Ando and Kaufman (1965) Section 1.3). By comparing (5) with (4), $p(\mathbf{x} \mid \mathscr{E})$ will be the density of a multivariate Student distribution.

The parameters of the densities (4) and (5) can be found by comparison with the results in Ando and Kaufman (1965) and should be such that the moments necessary for calculating $\mathbf{V}[\boldsymbol{\phi}_J(\mathbf{x}) \mid \mathscr{E}]$ exist.

3. Choice of variables for prediction. As a result of the discussion of Section 2, we assume $p(\mathbf{x} \mid \mathscr{E})$ is a Student density on $\mathscr{R}^r$ with ν degrees of freedom. We make use of the following result.

LEMMA 3. *If $p(\mathbf{x} \mid \mathscr{E})$ is a Student density on $\mathscr{R}^r$, then for fixed I,*

$$E[\boldsymbol{\phi}_J(\mathbf{x}) \mid \mathbf{x}_I, \mathscr{E}] = \mathbf{B}_{JI} \boldsymbol{\phi}_I(\mathbf{x}_I) \tag{6}$$

where $\mathbf{B}_{JI}$ is a $(u - v) \times v$ matrix with elements depending on $\mathscr{E}$.

The proof of Lemma 3 is found in the appendix.

By analogy with Lindley ((1968) Section 3), we find that

$$\mathbf{V}[\boldsymbol{\phi}_J(\mathbf{x}) \mid \mathscr{E}] = \mathbf{M}_{JJ} - \mathbf{M}_{JI}\mathbf{M}_{II}^{-1}\mathbf{M}_{IJ} \tag{7}$$

where $\mathbf{M}_{IJ} = E[\boldsymbol{\phi}_I(\mathbf{x}_I)\boldsymbol{\phi}_J(\mathbf{x})^T \mid \mathscr{E}]$ and $\mathbf{M}_{II}, \mathbf{M}_{JI}, \mathbf{M}_{JJ}$ are defined similarly. Here, in $\mathbf{V}[\boldsymbol{\phi}_J(\mathbf{x}) \mid \mathscr{E}]$, the zero rows and columns corresponding to the elements of $\boldsymbol{\phi}_I(\mathbf{x}_I)$ have been omitted. Thus, from (3), we have:

LEMMA 4. *Under Assumptions 1 to 5, to predict y the optimum set I is chosen to minimize*

$$E(\boldsymbol{\theta}_J \mid \mathscr{E})^T(\mathbf{M}_{JJ} - \mathbf{M}_{JI}\mathbf{M}_{II}^{-1}\mathbf{M}_{IJ})E(\boldsymbol{\theta}_J \mid \mathscr{E}) + c_I\,. \tag{8}$$

It is evident that we require moments of the form

$$E(x_1^{\eta_1} x_2^{\eta_2} \cdots x_r^{\eta_r} \mid \mathscr{E}) \tag{9}$$

for integers $\eta_i \geqq 0$ such that $0 \leqq \sum_{i=1}^r \eta_i \leqq 2\xi$. These moments can be expressed in terms of the elements of $E(\mathbf{x} \mid \mathscr{E})$ and $\mathbf{V}(\mathbf{x} \mid \mathscr{E})$. (As an example, see Lemma 5.) In the case of the random experiment these elements depend on $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n, \mathscr{E}_0$, but in the case of the designed experiment they only depend on $\mathscr{E}_0$.

4. An approximate method. In general, it is tedious to evaluate (9) in terms of the elements of $E(\mathbf{x} \mid \mathscr{E})$ and $\mathbf{V}(\mathbf{x} \mid \mathscr{E})$. However, in the case of the random experiment, we can find an approximation to (8) which is analogous to a result by Lindley ((1968) Section 5, (31)) for multiple linear regression. In addition to Assumptions 1 to 5a we need the following:

ASSUMPTION 6. The prior distribution of $\boldsymbol{\theta}$ is uniform over $\mathscr{R}^r$. Also, in Assumption 5a, the prior distribution of $(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1})$ is the limiting case in which $p(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1} \mid \mathscr{E}_0) \propto |\boldsymbol{\Sigma}|$.

In this situation it is well known that $E(\boldsymbol{\theta} \mid \mathscr{E})$ is the least squares estimator of $\boldsymbol{\theta}$.

Furthermore, Lindley ((1968) Section 4) gives $\nu = n - 1$,

$$E(\mathbf{x} \mid \mathscr{E}) = n^{-1} \textstyle\sum_{k=1}^{n} \mathbf{x}_k \tag{10}$$

and

$$\mathbf{V}(\mathbf{x} \mid \mathscr{E}) = n^{-1}\mathbf{W}\mathbf{W}^T \tag{11}$$

to order n^{-1}, where the (k, i)th element of $\mathbf{W}$ is $x_{ki} - \bar{x}_i$, and $\bar{x}_i = n^{-1} \sum_{k=1}^{n} x_{ki}$.

Let $\boldsymbol{\Phi}(\mathbf{X})$ be the matrix whose kth row is $\boldsymbol{\phi}(\mathbf{x}_k)$, where $\mathbf{X}$ denotes the matrix with elements $x_{ki} (k = 1, 2, \cdots, n;\ i = 1, 2, \cdots, r)$. Partition $\mathbf{X}$ into two matrices $\mathbf{X}_I$ and $\mathbf{X}_J$ corresponding to the variables in $\mathbf{x}_I$ and $\mathbf{x}_J$ respectively. Similarly, partition $\boldsymbol{\Phi}(\mathbf{X})$ into $\boldsymbol{\Phi}_I(\mathbf{X}_I)$ and $\boldsymbol{\Phi}_J(\mathbf{X})$ corresponding to $\boldsymbol{\phi}_I(\mathbf{x}_I)$ and $\boldsymbol{\phi}_J(\mathbf{x})$, respectively. For clarity, we write $\mathbf{Z} = \boldsymbol{\Phi}(\mathbf{X})$, $\mathbf{Z}_I = \boldsymbol{\Phi}_I(\mathbf{X}_I)$ and $\mathbf{Z}_J = \boldsymbol{\Phi}_J(\mathbf{X})$.

For n sufficiently large, from (10) and (11) we see that to order n^{-1}, $E(x_i \mid \mathscr{E})$ and $E(x_i x_j \mid \mathscr{E})$ equal

$$n^{-1} \textstyle\sum_{k=1}^{n} x_{ki} \quad \text{and} \quad n^{-1} \sum_{k=1}^{n} x_{ki} x_{kj}\,,$$

respectively. Also, $p(\mathbf{x} \mid \mathscr{E})$ is approximately a multinormal density, and if we approximate $E(x_1^{\eta_1} x_2^{\eta_2} \cdots x_r^{\eta_r} \mid \mathscr{E})$ (see (9)) by

$$n^{-1} \textstyle\sum_{k=1}^{n} x_{k1}^{\eta_1} x_{k2}^{\eta_2} \cdots x_{kr}^{\eta_r}\,,$$

then we replace $\mathbf{M}_{IJ}$ by $n^{-1}\mathbf{Z}_I{}^T\mathbf{Z}_J$, and replace $\mathbf{M}_{II}$, $\mathbf{M}_{JI}$, $\mathbf{M}_{JJ}$ similarly. Expression (7) is then approximately

$$n^{-1}[\mathbf{Z}_J{}^T\mathbf{Z}_J - (\mathbf{Z}_I{}^T\mathbf{Z}_J)^T(\mathbf{Z}_I{}^T\mathbf{Z}_I)^{-1}(\mathbf{Z}_I{}^T\mathbf{Z}_J)]\,. \tag{12}$$

In an orthodox (least squares) approach to this analysis we would calculate $R(I)$, the reduction in the residual sum of squares due to extending the model which includes only those independent variables in $\mathbf{x}_I$, (i.e. $E(\mathbf{y} \mid \boldsymbol{\theta}_I, \mathbf{X}_I, \mathscr{E}_0) = \mathbf{Z}_I\boldsymbol{\theta}_I$) to include all the independent variables (i.e. $E(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{X}, \mathscr{E}_0) = \mathbf{Z}\boldsymbol{\theta}$). Kendall and Stuart ((1966) (35.122)) give

$$R(I) = E(\boldsymbol{\theta}_J \mid \mathscr{E})^T[\mathbf{Z}_J{}^T\mathbf{Z}_J - (\mathbf{Z}_I{}^T\mathbf{Z}_J)^T(\mathbf{Z}_I{}^T\mathbf{Z}_I)^{-1}(\mathbf{Z}_I{}^T\mathbf{Z}_J)]E(\boldsymbol{\theta}_J \mid \mathscr{E}) \tag{13}$$

where $E(\boldsymbol{\theta}_J \mid \mathscr{E})$ is the least squares estimator of $\boldsymbol{\theta}_J$.

Using the approximation to (7) as given by (12), we have:

THEOREM 1. *Under Assumptions* 1 *to* 5a *and* 6, *for large* n, *to predict* y *the optimum set* I *is chosen to minimize*

$$n^{-1}R(I) + c_I\,. \tag{14}$$

This result is comparable to that obtained by Lindley ((1968) (31)) for multiple linear regression. To determine the best subset of independent variables to use to predict a future value of y, for the curvilinear regression defined in (1), both the orthodox and Bayesian approaches calculate the statistic $R(I)$ for sufficiently large sample size. However, this has been done with three restrictions:

(a) we have the full model as defined in (1),
(b) $\mathbf{x}$ has a multinormal distribution,
(c) $\mathbf{x}$ and $(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ have a common distribution.

If we have, for example, the model

$$E(y \mid \boldsymbol{\theta}, \mathbf{x}, \mathscr{E}_0) = \theta_{00} + \theta_{10}x_1 + \theta_{01}x_2 + \theta_{02}x_2^2$$

then we can still use (8) by considering the model given by (2) and assuming prior to the experiment that $\theta_{20} = \theta_{11} = 0$. However since the kth row of $\mathbf{Z}$ is then $(1, x_{k1}, x_{k2}, x_{k1}^2, x_{k1}x_{k2}, x_{k2}^2)$, this means that $E(\boldsymbol{\theta}_J \mid \mathscr{E})$ is not given as in (13) and the approximation (14) will no longer hold, so that the Bayesian method does not, in this case, use the same statistic $R(I)$ as the orthodox method. Hence, we impose the restriction (a).

An easier example to demonstrate the restriction (a) would be to consider simple linear regression through the origin, where we write $E(y \mid \boldsymbol{\theta}, x_1, \mathscr{E}_0) = \theta_0 + \theta_1 x_1$ assuming prior to the experiment $\theta_0 = 0$ and θ_1 is uniformly distributed over $\mathscr{R}^1$.

The restriction (b) can most easily be seen by considering simple linear regression, i.e. $\xi = r = 1$. Suppose that instead of (b), we assume x_1 has a uniform distribution over $(0, \lambda)$, where given $\mathscr{E}_0$, λ is independent of $\boldsymbol{\theta}$ and $p(\lambda \mid \mathscr{E}_0) \propto \lambda^{-(\alpha+1)} (\lambda > 0; \alpha \geqq 0$ is known). (This is a conjugate prior density for λ.) According to (3), we have to find the minimum of $c_{(1)}$ and

$$E(\theta_1 \mid \mathscr{E})^2 \operatorname{Var}(x_1 \mid \mathscr{E}) .$$

With (c), $\operatorname{Var}(x_1 \mid \mathscr{E})$ depends on the observations $x_{11}, x_{21}, \cdots, x_{n1}$ only through $M = \max_{k=1,2,\cdots,n} \{x_{k1}\}$. This can easily be seen, because

$$p(x_1 \mid \mathscr{E}) = \int p(x_1 \mid \lambda, \mathscr{E}_0) p(\lambda \mid x_{11}, x_{21}, \cdots, x_{n1}, \mathscr{E}_0)\, d\lambda$$

where $p(\lambda \mid x_{11}, x_{21}, \cdots, x_{n1}, \mathscr{E}_0)$ is the posterior density of λ, which depends on $x_{11}, x_{21}, \cdots, x_{n1}$ only through M (see Raiffa and Schlaifer (1961) page 54). Thus it is obvious that when $I = \phi$, $E(\theta_1 \mid \mathscr{E})^2 \operatorname{Var}(x_1 \mid \mathscr{E})$ is not approximated by $n^{-1}R(I)$, where $R(I)$ is obtained from (13), and so (14) does not hold for all I.

If instead of (c) we assume the experiment is designed (as in Assumption 5b), then $\mathbf{V}[\boldsymbol{\phi}_J(\mathbf{x}) \mid \mathscr{E}]$ does not depend on $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$. Consequently, (7) cannot be approximated by (12) and it is obvious that we do not use $R(I)$.

A further point of interest that should be mentioned while we are comparing the approach of this paper and orthodox methods, is that, for example, in the model given by (2), we are not investigating whether we should use a quadratic

model rather than a linear model. By our method, if we include x_1, we would include x_1^2, as we assume the extra cost of using x_1^2 is a computing cost, which may be negligible. This procedure may seem unreasonable if say ξ is greater than 2. However, many problems of this nature just require the use of a response surface for which $\xi = 2$, and it is towards this class of problems that the procedure is particularly directed.

5. An example. Consider the data from a regression experiment described in Williams ((1959) pages 42–45). The regression model is quadratic in the independent variable, thus $r = 1$, $\xi = 2$ and $E(y \mid \boldsymbol{\theta}, x_1, \mathscr{E}_0) = \theta_0 + \theta_1 x_1 + \theta_2 x_1^2$.

With the assumptions of Section 4, we examine the criteria for selecting the best subset of independent variables, both in the original form given by (3) or equivalently (8), and in the approximate form given by (14). The choice of variables is simply between observing x_1 or not observing x_1 in order to predict a future value of y, i.e. $I = (1)$ or ϕ respectively.

When $I = \phi$, (3), or (8), will equal S where

$$S = e_1^2 \operatorname{Var}(x_1 \mid \mathscr{E}) + 2e_1e_2 \operatorname{Cov}(x_1, x_1^2 \mid \mathscr{E}) + e_2^2 \operatorname{Var}(x_1^2 \mid \mathscr{E})$$

and $e_i = E(\theta_i \mid \mathscr{E})$, $i = 1, 2$. The quantities $\mu = E(x_1 \mid \mathscr{E})$ and $\tau^2 = \operatorname{Var}(x_1 \mid \mathscr{E})$ are obtained from (10) and (11) respectively, and in order to calculate S, we express $E(x_1^3 \mid \mathscr{E})$ and $E(x_1^4 \mid \mathscr{E})$ in terms of μ and τ^2 using the following lemma.

LEMMA 5. *If x has a Student distribution with ν degrees of freedom, mean μ and variance τ^2, then*

$$E(x^3) = 3\tau^2\mu + \mu^3,$$
$$E(x^4) = 3k\tau^4 + 6\tau^2\mu^2 + \mu^4$$

where $k = (\nu - 2)/(\nu - 4)$. Here we assume $\nu > 4$.

The proof of Lemma 5 is given in the appendix.

In this example, $\nu = n - 1$, and consequently

$$S = (3k - 1)e_2^2\tau^4 + (e_1 + 2\mu e_2)^2\tau^2$$

where $k = (n - 3)/(n - 5)$.

Alternatively, if we proceed as in Section 4, S is approximately $n^{-1}R(\phi)$.

When $I = (1)$, (3) and (14) both yield c, the cost of observing x per cost of unit error in prediction. Consequently, we choose $I = (1)$ if and only if $c < S$ or $c < n^{-1}R(\phi)$, depending on whether we approximate (3) or not.

The values of $\sum x_{k1}$, $\sum x_{k1}^2$, e_1, e_2 and $R(\phi)$ are given in Williams as 1646.4, 81747, 9.4743, 0.50863 and 21621500, respectively. The number of observations n is 36; hence $S = 580110$ and $n^{-1}R(\phi) = 600597$. The error in using $n^{-1}R(\phi)$ instead of S is just below $3\frac{1}{2}\%$ and therefore negligible.

Comparing the Bayesian and orthodox methods used on the above example, we see that a large value of S has corresponded to a highly significant value of the statistic used for testing the inclusion of (θ_1, θ_2) in the regression (see

Williams, Table 3.10). Intuitively, this is what we would expect, since a large S means that the cost of observing x would have to be very much larger than the cost of unit error in prediction before we would find it too expensive to observe x_1 in order to predict y.

Note that if we observe x_1, the optimal predictor of y is, by Lemma 2, $e_0 + e_1x_1 + e_2x_1^2$; otherwise it is $e_0 + e_1E(x_1 \mid \mathscr{E}) + e_2E(x_1^2 \mid \mathscr{E})$.

APPENDIX

PROOF OF LEMMA 3. That is, if $p(\mathbf{x} \mid \mathscr{E})$ is a non-degenerate Student density on $\mathscr{R}^r$ with ν degrees of freedom, then for fixed I,

$$E[\boldsymbol{\phi}_J(\mathbf{x}) \mid \mathbf{x}_I, \mathscr{E}] = \mathbf{B}_{JI}\boldsymbol{\phi}_I(\mathbf{x}_I) \tag{6}$$

where $\mathbf{B}_{JI}$ is a matrix of dimension $(u - v) \times v$, whose elements depend on $\mathscr{E}$ and are regarded here as constants. We omit the $\mathscr{E}$ for clarity.

Suppose without loss of generality, that $\mathbf{x}$ is partitioned so that $\mathbf{x}_I = (x_1, x_2, \cdots, x_s)^T$ and $\mathbf{x}_J = (x_{s+1}, x_{s+2}, \cdots, x_r)^T$. Then $\boldsymbol{\phi}_I(\mathbf{x}_I)$ contains all variables of the form $x_1^{\alpha_1}x_2^{\alpha_2} \cdots x_s^{\alpha_s}$, and $\boldsymbol{\phi}_J(\mathbf{x})$ all variables of the form $x_1^{\alpha_1}x_2^{\alpha_2} \cdots x_r^{\alpha_r}$ where at *least one* of $\alpha_{s+1}, \alpha_{s+2}, \cdots, \alpha_r$ is nonzero. An individual element of $E[\boldsymbol{\phi}_J(\mathbf{x}) \mid \mathbf{x}_I]$ can then be written as $x_1^{\alpha_1}x_2^{\alpha_2} \cdots x_s^{\alpha_s}E[\psi(\mathbf{x}_J) \mid \mathbf{x}_I]$ where $\psi(\mathbf{x}_J) = x_{s+1}^{\alpha_{s+1}}x_{s+2}^{\alpha_{s+2}} \cdots x_r^{\alpha_r}$ and at least one of $\alpha_{s+1}, \alpha_{s+2}, \cdots, \alpha_r$ is nonzero.

The definition of the non-degenerate Student density function on $\mathscr{R}^r$ is given by Raiffa and Schlaifer ((1961) page 256) as

$$p_s(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{H}, \nu) \equiv \int_0^\infty p_N(\mathbf{x} \mid \boldsymbol{\mu}, h\mathbf{H})p_\gamma(h \mid 1, \nu)\, dh\,, \tag{15}$$

where $\mathbf{H}$ is a positive semi-definite matrix, $p_N(\mathbf{x} \mid \boldsymbol{\mu}, h\mathbf{H})$ is the density of the multinormal distribution with mean $\boldsymbol{\mu}$ and dispersion matrix $h^{-1}\mathbf{H}^{-1}$, $p_\gamma(h \mid 1, \nu) = \exp(-\frac{1}{2}\nu h)(\frac{1}{2}\nu h)^{\frac{1}{2}\nu - 1}\frac{1}{2}\nu/\Gamma(\frac{1}{2}\nu)$ is the Gamma-2 density ($h \geqq 0$), as defined by Raiffa and Schlaifer ((1961) page 226).

Corresponding to the partition of $\mathbf{x}$ into $\mathbf{x}_I$ and $\mathbf{x}_J$, suppose that $\boldsymbol{\mu}$ and $\mathbf{H}$ are partitioned as

$$\begin{bmatrix} \boldsymbol{\mu}_I \\ \boldsymbol{\mu}_J \end{bmatrix}$$

and

$$\begin{bmatrix} \mathbf{H}_{II} & \mathbf{H}_{IJ} \\ \mathbf{H}_{JI} & \mathbf{H}_{JJ} \end{bmatrix}$$

respectively.

Using (15), it follows that

$$E_s[\psi(\mathbf{x}_J) \mid \mathbf{x}_I, \boldsymbol{\mu}, \mathbf{H}, \nu] = \frac{\int_0^\infty E_N[\psi(\mathbf{x}_J) \mid \mathbf{x}_I, \boldsymbol{\mu}, h\mathbf{H}]p_N(\mathbf{x}_I \mid \boldsymbol{\mu}_I, h\mathbf{H}_{II})p_\gamma(h \mid 1, \nu)\, dh}{p_s(\mathbf{x}_I \mid \boldsymbol{\mu}_I, \mathbf{H}_{II}, \nu)} \tag{16}$$

where

$$p_s(\mathbf{x}_I \mid \boldsymbol{\mu}_I, \mathbf{H}_{II}, \nu) = \int_0^\infty p_N(\mathbf{x}_I \mid \boldsymbol{\mu}_I, h\mathbf{H}_{II})p_\gamma(h \mid 1, \nu)\, dh\,.$$

Now

$$E_N[\psi(\mathbf{x}_J)\,|\,\mathbf{x}_I,\boldsymbol{\mu},h\mathbf{H}] = \int x_{s+1}^{\alpha_{s+1}}x_{s+2}^{\alpha_{s+2}}\cdots x_r^{\alpha_r}p_N(\mathbf{x}_J\,|\,\mathbf{x}_I,\boldsymbol{\mu},h\mathbf{H})\,d\mathbf{x}_J \tag{17}$$

where $p_N(\mathbf{x}_J\,|\,\mathbf{x}_I,\boldsymbol{\mu},h\mathbf{H}) = (2\pi)^{-\frac{1}{2}(r-s)}\,|h\mathbf{H}_{JJ}|^{\frac{1}{2}}\exp\left[-\tfrac{1}{2}h(\mathbf{x}_J-\mathbf{m}_J)^T\mathbf{H}_{JJ}(\mathbf{x}_J-\mathbf{m}_J)\right]$,

$$\begin{aligned}\mathbf{m}_J &= \boldsymbol{\mu}_J - (h\mathbf{H}_{JJ})^{-1}h\mathbf{H}_{JI}(\mathbf{x}_I-\boldsymbol{\mu}_I)\\ &= \boldsymbol{\mu}_J - \mathbf{H}_{JJ}^{-1}\mathbf{H}_{JI}(\mathbf{x}_I-\boldsymbol{\mu}_I)\,,\end{aligned} \tag{18}$$

(see Raiffa and Schlaifer (1961) page 250).

We make the transformation $h^{\frac{1}{2}}\mathbf{U}_{JJ}(\mathbf{x}_J-\mathbf{m}_J)=\boldsymbol{\omega}_J$ where $\mathbf{U}_{JJ}^T\mathbf{I}\mathbf{U}_{JJ}=\mathbf{H}_{JJ}$ and $\mathbf{I}$ is the unit matrix of order $r-s$; then

$$h(\mathbf{x}_J-\mathbf{m}_J)^T\mathbf{H}_{JJ}(\mathbf{x}_J-\mathbf{m}_J) = \boldsymbol{\omega}_J{}^T\mathbf{I}\boldsymbol{\omega}_J\,,$$

and the Jacobian of the transformation is $|h\mathbf{H}_{JJ}|^{-\frac{1}{2}}$. Each element of $\mathbf{x}_J$ will be of the form

$$x_j = h^{-\frac{1}{2}}\textstyle\sum_{l=s+1}^{r}u^{(jl)}\omega_l + m_j\,, \qquad j=s+1,\cdots,r\,,$$

where each $u^{(jl)}$ is a constant and the m_j are the corresponding elements of $\mathbf{m}_J$; the right-hand side of (17) then becomes

$$\begin{aligned}(2\pi)^{-\frac{1}{2}(r-s)}\int\textstyle\prod_{j=s+1}^{r}\left[h^{-\frac{1}{2}}\sum_{l=s+1}^{r}u^{(jl)}\omega_l+m_j\right]^{\alpha_j}\exp\left(-\tfrac{1}{2}\boldsymbol{\omega}_J{}^T\mathbf{I}\boldsymbol{\omega}_J\right)d\boldsymbol{\omega}_J\\ = \textstyle\sum_B K_{\beta_{s+1}\cdots\beta_r}m_{s+1}^{\beta_{s+1}}\cdots m_r^{\beta_r}h^{-\frac{1}{2}\sum_{j=s+1}^{r}(\alpha_j-\beta_j)}\end{aligned}$$

where $B=\{\text{all sets }(\beta_{s+1},\cdots,\beta_r)\colon \beta_j \text{ is integer} \geqq 0,\ 0\leqq\beta_j\leqq\alpha_j\}$ and each $K_{\beta_{s+1}\cdots\beta_r}$ is a constant. The numerator of the right-hand side of (16) is now

$$\textstyle\sum_B K_{\beta_{s+1}\cdots\beta_r}m_{s+1}^{\beta_{s+1}}\cdots m_r^{\beta_r}\int_0^\infty h^{-\frac{1}{2}\sum(\alpha_j-\beta_j)}p_N(\mathbf{x}_I\,|\,\boldsymbol{\mu}_I,h\mathbf{H}_{II})p_\gamma(h\,|\,1,\nu)\,dh$$

where $p_N(\mathbf{x}_I\,|\,\boldsymbol{\mu}_I,\ h\mathbf{H}_{II}) = (2\pi)^{-\frac{1}{2}s}\,|h\mathbf{H}_{II}|^{\frac{1}{2}}\exp\left[-(h/2)(\mathbf{x}_I-\boldsymbol{\mu}_I)^T\mathbf{H}_{II}(\mathbf{x}_I-\boldsymbol{\mu}_I)\right]$, and hence the integral in this expression is proportional to

$$\left[\nu+(\mathbf{x}_I-\boldsymbol{\mu}_I)^T\mathbf{H}_{II}(\mathbf{x}_I-\boldsymbol{\mu}_I)\right]^{-\frac{1}{2}(\nu+s)+\frac{1}{2}\sum(\alpha_j-\beta_j)}\,.$$

Now

$$p_s(\mathbf{x}_I\,|\,\boldsymbol{\mu}_I,\mathbf{H}_{II},\nu)\propto\left[\nu+(\mathbf{x}_I-\boldsymbol{\mu}_I)^T\mathbf{H}_{II}(\mathbf{x}_I-\boldsymbol{\mu}_I)\right]^{-\frac{1}{2}(\nu+s)}\,,$$

and for the right-hand side of (16) we then obtain

$$\textstyle\sum_B K'_{\beta_{s+1}\cdots\beta_r}m_{s+1}^{\beta_{s+1}}\cdots m_r^{\beta_r}\left[\nu+(\mathbf{x}_I-\boldsymbol{\mu}_I)^T\mathbf{H}_{II}(\mathbf{x}_I-\boldsymbol{\mu}_I)\right]^{\frac{1}{2}\sum(\alpha_j-\beta_j)}$$

where $K'_{\beta_{s+1}\cdots\beta_r}=L_{\beta_{s+1}\cdots\beta_r}K_{\beta_{s+1}\cdots\beta_r}$, each $L_{\beta_{s+1}\cdots\beta_r}$ being a constant.

We can write

$$(\mathbf{x}_I-\boldsymbol{\mu}_I)^T\mathbf{H}_{II}(\mathbf{x}_I-\boldsymbol{\mu}_I) = \textstyle\sum_\Gamma M_{\gamma_1\cdots\gamma_s}x_1^{\gamma_1}\cdots x_s^{\gamma_s}$$

where $\Gamma=\{\text{all sets }(\gamma_1,\cdots,\gamma_s)\colon \gamma_i \text{ is integer}\geqq 0,\ 0\leqq\sum_{i=1}^{s}\gamma_i\leqq 2\}$; furthermore, from (18), $m_j=\sum_{i=1}^{s}Q_{ji}x_i+R_j$ for constants R_j and Q_{ji} $(j=s+1,\cdots,r;\ i=1,2,\cdots,s)$, and thus

$$\textstyle\prod_{j=s+1}^{r}m_j^{\beta_j} = \sum_\Delta Q'_{\delta_1\cdots\delta_s}x_1^{\delta_1}\cdots x_s^{\delta_s}$$

where $\Delta=\{\text{all sets }(\delta_1,\cdots,\delta_s)\colon \delta_i \text{ integer}\geqq 0,\ 0\leqq\sum_{i=1}^{s}\delta_i\leqq\sum_{j=s+1}^{r}\beta_j\}$ and each $Q'_{\delta_1\cdots\delta_s}$ is a constant.

From (16) we finally obtain that an element of $E_s[\boldsymbol{\phi}_J(\mathbf{x}) \mid \mathbf{x}_I, \boldsymbol{\mu}, \mathbf{H}, \nu]$ has the form

$$(19) \qquad x_1^{\alpha_1} \cdots x_s^{\alpha_s} \sum_B K'_{\beta_{s+1} \cdots \beta_r} [\nu + \sum_\Gamma M_{\gamma_1 \cdots \gamma_s} x_1^{\gamma_1} \cdots x_s^{\gamma_s}]^{\frac{1}{2}\sum_{j=s+1}^{r}(\alpha_j - \beta_j)} \times \sum_\Delta Q'_{\delta_1 \cdots \delta_s} x_1^{\alpha_1} \cdots x_s^{\delta_s} .$$

The maximum power of any x_i, $i = 1, 2, \cdots, s$, is

$$\alpha_i + 2 \cdot \tfrac{1}{2} \cdot \sum_{j=s+1}^{r} (\alpha_j - \beta_j) + \sum_{j=s+1}^{r} \beta_j = \alpha_i + \sum_{j=s+1}^{r} \alpha_j \leqq \xi ,$$

and the maximum sum of the powers of x_i is

$$\sum_{i=1}^{s} \alpha_i + 2 \cdot \tfrac{1}{2} \cdot \sum_{j=s+1}^{r} (\alpha_j - \beta_j) + \sum_{j=s+1}^{r} \beta_j = \sum_{i=1}^{r} \alpha_i \leqq \xi .$$

Consequently (19) can be written as

$$\sum_Z b_{\zeta_1 \cdots \zeta_s} x_1^{\zeta_1} \cdots x_s^{\zeta_s} ,$$

where $Z = \{\text{all sets } (\zeta_1, \cdots, \zeta_s) : \zeta_i \text{ is integer} \geqq 0, 0 \leqq \sum_{i=1}^{s} \zeta_i \leqq \xi\}$ and each $b_{\zeta_1 \cdots \zeta_s}$ is a constant. Written in matrix notation, this result means that (6) is true, since all possible forms of $x_1^{\zeta_1} \cdots x_s^{\zeta_s}$ will be contained in the elements of $\boldsymbol{\phi}_I(\mathbf{x}_I)$.

PROOF OF LEMMA 5. Using the definition of the non-degenerate Student density function on $\mathscr{R}^1$ as given by (15), we can show

$$(20) \qquad E_s(x^\eta \mid \mu, H, \nu) = \int_0^\infty E_N(x^\eta \mid \mu, hH) p_\gamma(h \mid 1, \nu)\, dh .$$

Since, from Raiffa and Schlaifer ((1961), page 257, (8.29)), we have

$$E_s(x \mid \mu, H, \nu) = \mu \qquad \text{and} \qquad \mathrm{Var}_s(x \mid \mu, H, \nu) = \nu H^{-1}/(\nu - 2) ,$$

then μ is defined as in the statement of the lemma, and $H = \nu/(\nu - 2)\tau^2$.

It is easier to make the transformation $\omega = x - \mu$ and calculate $E_s(\omega^\eta \mid \mu, H, \nu)$ for $\eta = 3$ and 4 using (20), and then transform back to x to obtain $E_s(x^\eta \mid \mu, H, \nu)$. Since $E_N(\omega^3 \mid 0, hH) = 0$ and $E_N(\omega^4 \mid 0, hH) = 3(hH)^{-2} = 3(\nu - 2)^2\tau^4/h^2\nu^2$, then from (20) we obtain $E_s(\omega^3 \mid 0, H, \nu) = 0$ and $E_s(\omega^4 \mid 0, H, \nu) = 3k\tau^4$, where $k = (\nu - 2)/(\nu - 4)$.

Transforming back to x, we obtain

$$E_s(x^3 \mid \mu, H, \nu) = 3\mu\tau^2 + \mu^3$$

and

$$E_s(x^4 \mid \mu, H, \nu) = 3k\tau^4 + 6\tau^2\mu^2 + \mu^4 ,$$

these being the quantities $E(x^3)$ and $E(x^4)$ required in Lemma 5.

Acknowledgments. The author is grateful to Professor D.V. Lindley and a referee for their comments. He is also indebted to the Science Research Council under whose financial support the work was undertaken.

REFERENCES

ANDO, A. and KAUFMAN, G. M. (1965). Bayesian analysis of the independent multinormal process—neither mean nor precision known. *J. Amer. Statist. Assoc.* **60** 347–358.

JOHNSON, N. L. and LEONE, F. C. (1964). *Statistics and Experimental Design in Engineering and the Physical Sciences*, **2**. Wiley, New York.

KENDALL, M. G. and STUART, A. (1966). *The Advanced Theory of Statistics. Design and Analysis, and Time Series*. **3**. Griffin, London.

LINDLEY, D. V. (1968). The choice of variables in multiple regression. *J. Roy. Statist. Soc. Ser. B.* **30** 31–53.

RAIFFA, H. and SCHLAIFER, R. (1961). *Applied Statistical Decision Theory*. Harvard Graduate School of Business Administration, Boston.

WILLIAMS, E. J. (1959). *Regression Analysis*. Wiley, New York.

UNIVERSITY COLLEGE LONDON
GOWER STREET
LONDON WC 1, ENGLAND