USING EXTERNAL INFORMATION IN LINEAR REGRESSION: WITH A COMMENT--ETC(U)

JAN 82 T J MITCHELL, N R DRAPER

DAA629-80-C-0041

UNCLASSIFIED MRC-TSR-2327
MRC Technical Summary Report #2327

USING EXTERNAL INFORMATION IN LINEAR REGRESSION, WITH A COMMENTARY ON RIDGE REGRESSION. I. MIXED ESTIMATION AND BAYESIAN METHODS

Toby J. Mitchell and Norman R. Draper

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

January 1982

Received July 18, 1980

Approved for public release
Distribution unlimited

Sponsored by
U. S. Army Research Office
P.O. Box 12211
Research Triangle Park
North Carolina 27709
ABSTRACT

In this article, we discuss ways of using "dummy data" and mixed estimation (Theil and Goldberger, 1961) to bring external information formally into linear regression problems when the experimental data/model are inadequate. This is a useful way of attacking the same practical problems that motivated the development of ridge regression (Hoerl and Kennard, 1970).

The main practical issues considered are (i) what form should the "dummy data" take?, and (ii) how much weight should it be given relative to the experimental data? When specific prior information is unavailable, it is suggested that the dummy data should reflect a preference for "stable" response functions and it is shown how this can be accomplished. Guidelines for the choice of the weighting parameter $k$ (equivalent to the choice of the ridge parameter in ridge regression) are given. Upper limits for $k$ are based on various tests of compatibility between the external (dummy) data and the experimental data. Lower limits for $k$ are determined by the inadequacy of the data/model for the purpose(s) of the analysis. Finally, the parallel Bayesian approach is discussed, with emphasis on Box's (1980) framework of model estimation and criticism.

AMS(MOS) Subject Classification: 62J99
Key Words: Collinearity; Dummy data; External information in regression; Ridge regression.

Work Unit Number 4 - Statistics and Probability

*Oak Ridge National Laboratories, P. O. Box Y, Oak Ridge, TN 37830.

Sponsored by the United States Army under Contract Nos. DAAG29-80-C-0041 and DAAG29-80-C-0113.
SIGNIFICANCE AND EXPLANATION

We discuss ways of using "dummy data" to bring external information formally into linear regression problems when the experimental data/model are inadequate. The main practical issues are (i) what form should the "dummy data" take?, and (ii) how much weight should it be given relative to the experimental data? When specific prior information is unavailable, it is suggested that the dummy data should reflect a preference for so-called "stable" response functions, and it is shown how this can be accomplished. Guidelines for the choice of weighting are also given. Upper limits for a weighting parameter are based on various tests of compatibility between the external (dummy) data and the experimental data. Lower limits are determined by the limits of inadequacy of the data/model for the purpose(s) of the analysis. The parallel Bayesian approach is also discussed.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the authors of this report.
1. INTRODUCTION.

In this report we focus on the problems that arise in linear regression when there is insufficient information in the data/model to obtain useful estimates of the regression coefficients and/or linear functions of them. Methods for augmenting the experimental data with external "information" lead to a form of ridge regression which differs in several major respects from that which is currently practiced.

We shall consider models for the observed response variable \( y_d \), given associated "predictor data" \( d \), to be of the form:

\[
Y_d = \eta_d + \epsilon_d
\]

where

\[
\eta_d = \beta' x(d) = x'(d)\beta,
\]

the \( p \) predictor variables \( x_i(d) \) are specified functions of \( d \), the regression coefficients \( \beta_i \) are unknown constants, and \( \epsilon_d \) is a random variable with mean 0 and variance \( \sigma^2 \). For \( d \neq d' \), \( \epsilon_d \) is independent of \( \epsilon_{d'} \). The error variance \( \sigma^2 \) assumed to be constant for all \( d \), and is generally unknown.

We assume that (1.1) and (1.2) hold over a model region \( R \) in the space of the predictor data. We shall refer to (1.2) as the model function, although it is really a family of model functions, indexed by the parameter vector \( \beta \). We shall view the problem of "estimating" \( \beta \) to be the problem of choosing from this family a single model function which is to be used to make inferences about \( \eta_d \) given \( d \in R \), or more generally, to make inferences about specified linear functions of the \( \beta_i \).

---

*Oak Ridge National Laboratories, P. O. Box Y, Oak Ridge, TN 37830.*

Sponsored by the United States Army under Contracts Nos. DAAG29-80-C-0041 and DAAG29-80-C-0113.
Given a sample of \( n \) independent observations \( \mathbf{x}' = (y_1, y_2, \ldots, y_n) \), where \( y_u \) is associated with known \( d_u, \ u = 1, 2, \ldots, n \), (1.1) and (1.2) imply the statistical model for the experimental data:

\[
E(y) = \mathbf{x} \hat{\beta}
\]

\[
V(y) = \sigma^2 \mathbf{I},
\]

where the matrix \( \mathbf{X} \) is \( n \times p \) with \((u, i)\)th element \( X_{ui} \) equal to \( x_i(d_u) \) and \( \mathbf{I} \) is the \( n \times n \) identity matrix.

The most common estimation procedure follows from the least squares criterion: choose \( \hat{\beta} \) to minimize the sum of squares: \( SS(\hat{\beta}) = \mathbf{y} - \mathbf{X}\hat{\beta} \). This yields the least squares estimator

\[
\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},
\]

where we assume that \( \mathbf{X}'\mathbf{X} \) is non-singular so that \( \hat{\beta} \) is unique. (If \( \mathbf{X}'\mathbf{X} \) were singular, (1.5) would contain a generalized inverse, and \( \hat{\beta} \) would not be unique.)

The least-squares estimator may be justified on several statistical grounds, the most prominent of which are:

(i) The Gauss-Markov Theorem: Among all unbiased estimators for \( \mathbf{B} = \mathbf{X}\hat{\beta} \) that are linear in the \( y \)'s, \( \mathbf{X}\hat{\beta} \) has minimum variance, no matter what \( \mathbf{X} \) is. (An alternative version that does not impose unbiasedness is: Among all linear estimators for \( \mathbf{B} \), \( \mathbf{X}\hat{\beta} \) minimizes the maximum mean squared error of estimation (Barnard, 1977).) Since the purpose of the model is often to estimate linear functions of the \( \beta \)'s, this is generally considered a strong argument in favor of least squares, although the restriction to linear estimators seems arbitrary.

(ii) If \( \mathbf{X} \) has a multinormal distribution with mean and variance-covariance matrix given by (1.3) and (1.4), then \( \hat{\beta} \) is the maximum likelihood estimator of \( \beta \). Intuitively, we would expect the model function that is favored most by the available data, i.e., the maximum likelihood model, to serve at least as well as any other in making inferences about general linear functions of the \( \beta \)'s.
These arguments in favor of least squares lose force if the model as given by (1.1) and (1.2) is inadequate, or if, in the case of (ii), \( y \) is not normally distributed. Minimum bias estimation (Karson, Manson, and Hader, 1969; Kupper and Meydrech, 1974) and robust regression (Huber, 1981) are alternatives to least squares in such situations. We shall not consider these methods here, since we shall assume that the model assumptions (including normality, when needed) are satisfied. Even then, least squares estimation may seem unsatisfactory because the variances of the \( \beta \)'s may be extremely high for certain configurations of the \( d \)'s. Various alternative estimation methods have been proposed for these situations also; predominant among them are variable selection and ridge regression. (See Hocking (1976) for an excellent overview of these methods.)

The motivation for our work here is our dissatisfaction with the way ridge regression is perceived and used. There is an enormous literature on this subject, stemming from the basic papers of Hoerl and Kennard (1970a, 1970b). Unfortunately, much of this work presents ridge regression primarily as a biased estimation technique, based on the experimental data alone, that is "superior" in some sense to least squares estimation. This approach has misled users and authors alike. The main problem in situations where ridge regression is often invoked is one of lack of information in the data (with respect to the choice of a single model from the given family), and not with least squares estimation per se. External information or additional assumptions are needed; what form should these take and how should they be implemented? These are the topics of Part I of this report. We have deliberately written Part I almost as if ridge regression did not exist, in order to set out first the fundamentals that underlie our point of view. However, we do consider only the kind of external information that leads to "ridge-type" estimators. In Part II, we shall offer a commentary on some specific aspects of ridge regression, adding further support to recent criticisms of various ridge regression myths (Thisted 1976; Draper and Von Vorstrand, 1979; Smith and Campbell, 1980; Smith, 1980; Egerton and Laycock, 1981).
Certainly the idea of considering ridge regression as a means of bringing external information into the estimation procedure is not new. Hoerl and Kennard (1970a) mentioned a Bayesian derivation of ridge estimation, and Lindley and Smith (1972) have provided an extensive Bayesian treatment. We shall rely more on the "mixed estimation" approach of Theil and Goldberger (1961), in that the external information is introduced by means of dummy data rather than prior distributions. We offer (in Section 3) a rationale for specifying such data in the absence of specific external information about the regression coefficients. In Section 4, we suggest some guidelines for selecting the amount of influence, or weight, to give the dummy data; this is equivalent to the choice of the ridge parameter \( k \) in ridge regression. Bayesian methods are discussed in Section 5 with emphasis on Box's (1980) illuminating approach to model estimation and criticism.
2. BRINGING IN EXTERNAL INFORMATION

2.1. The need for external information.

Suppose we wish to estimate one or more linear functions of the regression coefficients. We shall denote these by \( \mu_1, \mu_2, \ldots, \mu_m \), where \( \mu_1 = \xi_1 \hat{\beta} \) and the t's are known. If the response vector \( y \) is normally distributed:

\[
x \sim N(\xi \hat{\beta}, \sigma^2 I),
\]

(2.1)

the set of "acceptable" values of \( \hat{\beta} \) lies inside the ellipsoid:

\[
(\hat{\beta} - \hat{\beta})'X'X(\hat{\beta} - \hat{\beta}) < a
\]

(2.2)

for some suitably chosen positive constant \( a \). This follows from a standard likelihood argument which requires only that every "acceptable" \( \hat{\beta} \) has a likelihood greater than that of every "unacceptable" \( \hat{\beta} \). By maximizing and minimizing \( \xi_1^2 \hat{\beta} \) subject to (2.2), we find that the "acceptable" values for \( \mu_1 \) must be in the interval with end points:

\[
\xi_1^2 \hat{\beta} \pm (a \xi_1^2 (X'X)^{-1} \xi_1)^{1/2},
\]

(2.3)

The choice \( a = \sigma^2 \chi_1^2 \) in (2.3) where \( \chi_2^2 \) is the upper \( 1-a \) point of the chi-squared distribution with one degree of freedom, results in the usual 100(1-\( a \))% confidence interval for \( \mu_1 \) when \( \sigma^2 \) is known. The other factor that determines the interval width, namely:

\[
\omega_1 = \xi_1^2 (X'X)^{-1} \xi_1
\]

(2.4)

is affected jointly by the model and "design", (i.e., the configuration of the d's), which express themselves through \( Y \).

If \( \omega_1 \) in (2.4) is such that the interval (2.3) is too wide to be of value, we conclude that the data and model assumptions alone are inadequate for the purpose of estimating \( \mu_1 = \xi_1^2 \hat{\beta} \). Ideally, we would like to have more experimental data, but we shall assume here that this is not possible and that our only recourse is to bring in external, often vague, information about \( \hat{\beta} \). For another purpose (i.e., another \( \xi_1 \)), of course, there might be need at all for additional information or assumptions.

We have deliberately avoided presenting \( \omega_1 \) as the variance of \( \xi_1^2 \hat{\beta} \) (divided by \( \sigma^2 \)) to emphasize that its role as an indicator of the information provided by the data/model with respect to \( \mu_1 \) is independent of whatever point estimation procedure for \( \hat{\beta} \) is
used. If \( w_1 \) is "too wide", the problem cannot be alleviated by blaming least squares estimation and seeking a "better" estimator in some class of biased estimators. External information must be used.

2.2. Mixed estimation.

Our discussion will center on the use of "mixed estimation", proposed by Theil and Goldberger (1961) and developed further by Theil (1963), as a vehicle for bringing in external information. (We shall briefly summarize the Bayesian parallel in Section 5.)

For mixed estimation, one augments the observed data with an \( n_0 \)-vector of "dummy data" \( X_0 \), i.e., guessed values of linear functions of the \( \beta \)'s. These guesses are modeled as:

\[
X_0 = X_0 \hat{\beta} + e_0
\]  

(2.5)

where \( X_0 \) is a specified \( n_0 \times p \) matrix and

\[
e_0 \sim N(0, \sigma^2 e_0).
\]  

(2.6)

We shall consider \( X_0 \) to be known throughout; its choice will be considered in Section 3. The guesses are consistent in that there exists some \( \hat{\beta} \) (not necessarily the "true" \( \beta \)) such that

\[
X_0 \hat{\beta} = X_0.
\]  

(2.7)

For known \( \sigma^2 \) and \( \sigma_0^2 \), the mixed estimator is the weighted least squares estimator of \( \hat{\beta} \) based on all the data:

\[
\hat{\beta} = (W + W_0)^{-1}(W \hat{\beta} + W_0 \hat{\beta}^0)
\]  

(2.8)

where

\[
W = X'X/\sigma^2
\]  

(2.9)

and

\[
W_0 = X_0'X_0/\sigma_0^2.
\]  

(2.10)

If we let

\[
k = \sigma^2 / \sigma_0^2
\]  

(2.11)

and

\[
T = X_0'X_0/k
\]  

(2.12)
Then (2.8) can be written

$$\bar{\bar{\beta}} = (X'X + kT)^{-1}(X'Y + kT\bar{\bar{r}}).$$

The variance-covariance matrix of $\bar{\bar{\beta}}$ is

$$\text{V}(\bar{\bar{\beta}}) = (Y' + W_0)^{-1} = \sigma^2(X'X + kT)^{-1}.$$ (2.14)

When $\sigma^2$ is not known, Theil and Goldberger (1961) and Theil (1963) proposed to replace $\sigma^2$ by $s^2$, the residual mean square from the least squares analysis of the experimental data. They also mentioned an iterative approach in which a new estimate of $\sigma^2$ and hence a new weight matrix $W$ is derived from the residual mean square after each iteration. Under this procedure, successive estimates of $\bar{\bar{\beta}}$ converge to the maximum likelihood estimate of $\bar{\bar{\beta}}$ based on all the data. Unfortunately, this estimate of $\bar{\bar{\beta}}$ is not linear in the observations, so further statistical results are difficult to obtain.

We shall adopt here a more tractable approach, and treat $\sigma^2$ as an unknown multiple of $\sigma^2$:

$$\sigma^2 = k^{-1}n^2.$$ (2.15)

For fixed $k$, this will allow us to make inferences using standard weighted least squares procedures, where the experimental data are given weight 1 and the dummy data are given weight $k$. Thus $\bar{\bar{\beta}}$ is given by (2.13) with variance-covariance matrix given by (2.14).

The estimate of $\sigma^2$ is based on the weighted residual mean square

$$\hat{\sigma}^2 = (s^2 + q)/(v + n_0)$$ (2.16)

where

$$q = (\bar{Y} - \bar{x})'X'(X'X + kT)^{-1}kT(\bar{Y} - \bar{x}),$$ (2.17)

and $v = n-p$ is the number of degrees of freedom for $s^2$. Confidence intervals for linear functions of the $\beta$'s can be constructed in the usual way, using Student's $t$ distribution with $v+n_0$ degrees of freedom. Thus, a 100(1-$\alpha$)% confidence interval for $\bar{Y} = \bar{x}^T\bar{\bar{\beta}}$ is given by

$$\bar{Y} \pm t_{v+n_0,\alpha/2}[k'(X'X + kT)^{-1}k]'^{1/2}s.$$ (2.18)

We note that if $X$ has been "centered and scaled" so that $X'X$ is in the form of a correlation matrix, and if $Y = \bar{x}$ and $\bar{\bar{r}} = \bar{\bar{r}}$, then $\bar{\bar{\beta}}$ reduces to the standard Hoerl-Kennard (1970a) ridge estimator. Proponents of classical ridge regression, while admitting
that ridge regression can be viewed as the result of using external information, would deny
the formal use of that information in determining confidence intervals. They would then
arrive at the standard "least squares" confidence intervals rather than (2.18). (See
Obenchain, 1977.)

The difficulties in finding a "best" value for $k$ are well known to followers of
ridge regression. Our approach here will be to regard $k$ as fixed by assumption, then
check that assumption for compatibility with the experimental data, much in the spirit of
"model criticism" put forward by Box (1980). We shall discuss these procedures in Section
4, after we deal with the problem of specifying $\mathcal{T}$ and $\mathcal{G}$. 
3. USE OF PRIOR PREFERENCES TO DETERMINE T and $\hat{\beta}^*$. 

3.1. Prior preferences.

We shall assume that we can establish (by the approach to be suggested in Section 3.2) a convention for ranking proposed values of $\hat{\beta}$ a priori, where the ranking is such that $\hat{\beta}^{(1)}$ is preferred to $\hat{\beta}^{(2)}$ if and only if

$$
(\hat{\beta}^{(1)} - \beta^0)^T T^0 (\hat{\beta}^{(1)} - \beta^0) < (\hat{\beta}^{(2)} - \beta^0)^T T^0 (\hat{\beta}^{(2)} - \beta^0)
$$

(3.1)

for specified $\beta^0$ and non-negative definite $T^0$. To ensure that the likelihood for the dummy data reflects these preferences, we would choose $T^0 = T$ and $\beta^0 = \beta$ in our mixed estimation procedure.

A special case of this kind of rule is Hoerl and Kennard's (1970) preference for "short" parameter vectors, expressed in (3.1) as $\beta^0 = 0$, $T^0 = I$, where $X'X$ is in correlation form. This particular preference has been criticized (rightly, we believe) because the length of $\hat{\beta}$ depends on the parametrization of the model. (See Smith and Campbell (1980).)

Here we present a rationale for choosing $T^0$ and $\beta^0$ which is in the same spirit as the Hoerl-Kennard preference, and at the same time produces a family of estimators that is invariant under reparametrizations of the model. This rationale, which is based on the concept of "stability" of the expected response $\eta_d$, also extends naturally to quadratic and higher order models. In some circumstances, of course, one's prior information may permit the choice of $T^0$ and $\beta^0$ to be made directly.

3.2. "Stable response" preferences.

Suppose $R$ is a selected subregion of the model region $R$. We define the instability $\Gamma$ of the response function $\eta_d$ over $R$ to be the variance of $\eta_d$ induced by a uniform probability distribution $\phi(d)$ over $R$. (This is just the squared deviation of $\eta_d$ from its average, integrated over $R$. We assume, of course, that $R$ and the functions $\eta_d$ are sufficiently well-behaved for $\Gamma$ to exist.) We emphasize that $\Gamma$ is an inherent property of the model function $\eta_d$; it has nothing to do with the data.

From (1.2), $\Gamma$ can be written as a quadratic form:

$$
\Gamma = \hat{\beta}' T \hat{\beta}
$$

(3.2)
where
\[ u = V_f(x(d)) = E_f(x(d) - \xi)(x(d) - \xi)' = \xi - \xi' \tag{3.3} \]

where
\[ \xi = E_f(x(d)x'(d)) \tag{3.4} \]

and
\[ \xi = E_f(x(d)). \tag{3.5} \]

In the absence of any other external information about \( \xi \), we might simply express a preference for those \( \xi \)'s that give a more stable response \( \eta_d \) over \( R_s \), i.e., we choose \( \xi_0 = \eta \) and \( \xi_0 = 0 \) in (3.1). This leads, via mixed estimation, to a "stable response" (SR) estimator of form (2.13) with \( \xi = \eta \) and \( \xi_0 = 0 \), i.e.,
\[ \xi_{SR} = (\xi' + k\xi)' \xi. \tag{3.6} \]

Now suppose the model function (1.2) is reparametrized as follows:
\[ \eta_d = \sum_{i=1}^{n} \omega f_i(d) = f'(d)\eta \tag{3.7} \]

where
\[ f'(d) = f'(d)\eta, \quad \xi = \xi', \tag{3.8} \]

\( \eta \) being a nonsingular \( p \times p \) matrix. Then the appropriate \( u \)-matrix is:
\[ u_x = V_f(x(d)) = A^t \tag{3.9} \]

the predictor data matrix is \( X^t \), and the SR estimator for \( \xi \) is:
\[ \xi_{SR} = (A' \xi + kA' \xi)' \xi \tag{3.10} \]

Thus the SR estimator is invariant under reparameterizations of the model. This is what we would expect, since the instability of the model function depends only on \( \eta_d \) and \( R_s \), neither of which is affected by reparameterization.

It is interesting to examine the influence of various choices of the "region of stability" \( R_s \) on \( \xi \), and hence on \( \xi_{SR} \). In the three cases that follow we suppose that there is a constant term in the model, so it will be convenient to change notation slightly by writing the model function (1.2) as
\[ \eta_d = \beta_0 + \frac{1}{p} \sum_{i=1}^{p} x_{i}(d) = \beta_0 + x'(d) \bar{\beta} \quad (3.11) \]

and the expected response vector (1.3) as

\[ E(\chi) = \beta_0 + \bar{\eta} \bar{\beta} \quad (3.12) \]

where \( \bar{\eta} \) is a vector of 1's. Note that \( p \) is now 1 less than the total number of coefficients in the model.

**Case 1.** \( R_s \) consists of the data points \( d_1, d_2, \ldots, d_n \) themselves.

Since \( \phi(d) = 1/n \) at every \( d_u \) in the data set, (3.5) yields

\[ \bar{\xi}' = (1, \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p), \]

where \( \bar{x}_i = \frac{1}{n} \sum_{u=1}^{n} x_{ui} \). It then follows from (3.3) that

\[ \bar{U} = \begin{bmatrix} 0 & 0' \\ 0 & \frac{1}{n} \bar{X} \bar{X}' \end{bmatrix} \quad (3.13) \]

where \( \bar{X} \) is obtained from \( X \) by "centering" the columns, i.e.,

\[ \bar{X} = (1, \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p) \quad (3.14) \]

It can be shown that the corresponding SR estimator (3.6) is given by:

\[ \bar{\beta}_{SR} = \frac{n}{np} \bar{\xi} \quad (3.15) \]

The effect of the external data here is to "shrink" the least squares estimator.

**Case 2.** "Factorial" \( R_s \) defined by the levels of the observed predictor variables - first order model.

We shall assume here that \( \eta_d \) is a first-order model function, i.e., the predictors \( x_i(d) \) in (3.11) are not functionally related. We now define the region of stability \( R_s \) to correspond to the factorial "design" generated by the observed levels of the individual predictors. Let \( \Omega_1 = \{ x_{11}, x_{12}, \ldots, x_{1n} \} \), i.e., \( \Omega_1 \) is the set of observed values of \( x_1 \), and let \( \Omega \) be the matrix whose rows are elements of the lattice \( \Omega_1 \times \Omega_2 \times \ldots \times \Omega_p \). This lattice has \( n^p \) elements, some of which may be identical. For example,
if $\bar{x} = \begin{bmatrix} 1 & 2 \\ 1 & 3.5 \\ 2 & -3 \end{bmatrix}$ then $\bar{y} = \begin{bmatrix} 1 & 2 \\ 1 & 3.5 \\ 1 & -3 \\ 1 & 2 \\ 1 & 3.5 \\ 1 & -3 \\ 2 & 2 \\ 2 & 3.5 \\ 2 & -3 \end{bmatrix}$.

We take $R_n$ to be the set of points $(w_j)$ corresponding to the rows of $\bar{y}$, with $\phi(w_j) = 1/n$. (Note, however, that under this definition $R_n$ depends on the parametrization of the model, so the invariance result (3.10) will not hold.)

Under the above conditions, $y$ as defined by (3.3) - (3.5) is:

$$y = \frac{1}{n} \begin{bmatrix} 0 & - & 0' \\ - & 0 & - \\ 0 & - & \frac{S}{n} \end{bmatrix}$$

(3.16)

where $S$ is a diagonal matrix with ith diagonal element

$$S_i = \sum_u (X_u - \bar{x}_i)^2, \quad i = 1, 2, \ldots, P.$$  

(3.17)

To see this, note that $U_{i+1,i+1}$ is, by (3.3), the variance of $X_i$ with respect to the distribution $\phi$. Under $\phi$, the variance of $X_{u_1}$ is just $1/n$, so the mean of $X_i$ is $\bar{x}_i$ and the variance of $X_i$ is $S_i/n$. Moreover, because of the independence of $X_i$ and $X_j$ for all pairs $i \neq j$, the off-diagonal terms of $S$ are 0. Finally, the constancy of $x_0 = 1$ under $\phi$ results in the zero values in the first row and first column of $y$.

With $y$ given by (3.16), the SR estimator (3.6) is given by:

$$\tilde{\beta}_{SR} = (\bar{z}'\bar{z} + \lambda_k)^{-1}\bar{z}'\chi$$

(3.18)

$$\tilde{\beta}_{0,SR} = \tilde{\gamma} - \frac{1}{\hat{\beta}_{1,SR}} \bar{x}_1$$

where $\bar{z}$ is the "correlation form" of $\bar{x}$, i.e.,

$$\bar{z} = \frac{\bar{x} - \bar{x}}{\bar{x}'}$$

(3.19)
Note that (3.18) is the same as the Hoerl-Kennard family of ridge estimators. We consider this derivation to be a curiosity rather than a rationale for using the Hoerl-Kennard estimators. Generally, we would prefer to define $R_s$ so that it depends neither on the observed data nor on the parametrization of the model.

**Case 3.** Rectangular $R_s$ - first order model.

In most regression problems, the predictor data $d$ take the form of a vector of $r$ "basic" predictor variables, and the region of stability $R_s$ can be taken to be some geometrically convenient region in $R$. In this example, we again assume a first-order model, so $p = r$, and we suppose that the predictor variables have been coded so that

$$-1 < x_j(d) < 1 \iff d \in R_s.$$  \hspace{1cm} (3.20)

Taking $\phi(d)$ to be uniform over $R_s$, we find

$$y = \begin{bmatrix} 0 & 2 \varepsilon \\ 0 & \frac{1}{3} I \end{bmatrix}.$$ \hspace{1cm} (3.21)

With $y$ given by (3.21), $\hat{E}_{SR}$ is given by

$$\hat{E}_{SR} = (\hat{X}'\hat{X} + \frac{1}{3} I)^{-1}\hat{X}'\hat{X}.$$ \hspace{1cm} (3.22)

$$\hat{\beta}_{0,SR} = \hat{y} - \sum_{i=1}^{p} \hat{E}_{i,SR} \hat{x}_i.$$  

We note that if one were to use the Hoerl-Kennard formula, after centering, but not scaling $\hat{X}$, one would arrive at this same family of estimators. ($\hat{X}$ has already been effectively scaled by the coding (3.20).) As in Case 2 above, however, the application of the stable response preference to higher order models results in a more complicated form of $y$, and the analogy with the Hoerl-Kennard approach no longer holds.

In the above Cases 1-3, we have given some examples of the derivation of the stable response family of estimators for different choices of the region of stability $R_s$. In Cases 1 and 2, which lead to the more familiar "shrunken" estimator and Hoerl-Kennard estimator, the data define $\phi$. We prefer the approach taken in Case 3, however, since we feel that the notion of response surface stability over some region should not depend on the distribution of the current data over that region.
We offer the "stable response" preference primarily as a route to follow in the absence of specific prior information about individual $\delta$'s or about the expected response $n_d$ at certain points $d \in R$. The notion of instability of the true response surface over a specified region seems to us to lend itself to physical interpretation much more readily than the notion of length of the vector of regression coefficients, even though the choice of region of stability is still rather arbitrary. The operational advantage of the SR method is that one can apply it without having to worry about parametrization or about the specific meaning of individual coefficients.

3.3. Setting up the dummy data.

External preferences of the form (3.1) can be transformed into dummy data for mixed estimation by choosing $X_0$, $Y_0$, and $X_0$ to satisfy (2.12) and (2.7) with $X = x_0$ and $y^* = y^0$. The $n_0$ rows of $X_0$ could, for example, be the eigenvectors corresponding to the positive eigenvalues $\lambda_i$, $i = 1, 2, \ldots, n_0$, of $T$. $Y_0$ would then be a diagonal matrix with those eigenvalues on the diagonal. Since the analysis depends only on $T$ and $y^*$, the only point in choosing $X_0$, $y_0$, and $X_0$ in this way is to take advantage of standard computer routines for doing weighted least squares. The weight for the $i$th dummy observation would be $w_0$, each experimental observation would have weight 1.
4. SOME GUIDELINES FOR CHOOSING k.

4.1. Omnibus compatibility test.

Throughout this section, it will be convenient to express \( \tilde{\beta}, \sigma^2, \) and \( q \), defined respectively by (2.13), (2.16), and (2.17), as functions of \( k \).

Although we doubt that a "best" procedure for choosing \( k \) exists, we can present some guidelines. For example, we shall certainly not want to choose \( k \) so large that the external (dummy) data is incompatible with the experimental data. Theil's (1963) test statistic for detecting incompatibility is, after adapting for fixed \( k \) rather than fixed \( \sigma_0^2 \) and expressing in terms of \( \tilde{\beta} \) and \( \Sigma \):

\[
\Psi(k) = \frac{q(k)}{n_0\sigma^2},
\]

(4.1)

where \( q(k) \) is given by (2.17). Under the assumed model for the experimental and dummy data, \( \Psi(k) \) is distributed as \( \mathcal{F}_{n_0,v} \) where \( n_0 \) is the rank of \( \Sigma \). Large values of \( \Psi(k) \) indicate incompatibility.

This test for compatibility is equivalent to the usual test of the hypothesis:

\( \mathbb{E}(x_0) = x_0 \hat{\beta} \) against the alternative: \( \mathbb{E}(x_0) \neq x_0 \hat{\beta} \), a test frequently used in regression to evaluate a subset of the observations (in this case \( x_0 \)) jointly for the presence of outliers. (Gentleman and Wilk, 1975.)

We propose that \( \Psi(k) \) be used to set an upper limit \( k_a \) on \( k \), chosen so that

\[
\Psi(k_a) = \mathcal{F}_{n_0,v,a}
\]

(4.2)

for some suitably chosen percentage point (e.g. \( a = .10 \)) of the \( \mathcal{F}_{n_0,v} \) distribution.

Remark 1. If \( k \leq k_a \) (i.e., if the external information "passes" the compatibility test) then \( \tilde{\beta}(k) \) is within the usual \((100(1-a))\% \) confidence region for \( \hat{\beta} \) based on the experimental data alone, i.e.,

\[
(\tilde{\beta}(k) - \hat{\beta})'g'g(\tilde{\beta}(k) - \hat{\beta}) < p s^2 \mathcal{F}_{p,v,a}
\]

(4.3)

That is, \( \tilde{\beta}(k) \) is "a-acceptable" in the sense of McCabe (1978). This follows from the fact that, for \( k \leq k_a \),

\[
(\tilde{\beta}(k) - \hat{\beta})'g'g(\tilde{\beta}(k) - \hat{\beta}) < q(k) < n_0\sigma^2 \mathcal{F}_{n_0,v,a} < ps^2 \mathcal{F}_{p,v,a}
\]

(4.4)

since it can be shown that \( n_0 \mathcal{F}_{n_0,v,a} < \mathcal{F}_{p,v,a} \).
Remark 2. The choice of \( k \) associated with the estimator RIDGM, which performed well in the large simulation study of ridge estimators conducted by Dempster, Schatzoff and Wermuth (1977) is equivalent to the solution of \( \Psi(k) = 1 \), where \( \hat{\xi}^* = 0 \) and \( \Sigma = I \). For any reasonable \( \alpha \), therefore, the \( k \) for RIDGM would be less than \( k_{\alpha} \).

Remark 3. As a first approximation to \( k_{\alpha} \), one could approximate \( \Psi(k) \) by a linear function near \( k = 0 \); in this neighborhood

\[
q(k) = (\hat{\xi} - \hat{\xi}^*)'k\hat{\xi}(\hat{\xi} - \hat{\xi}^*),
\]

so

\[
k_{\alpha} = n_\alpha^2P_n^0\n_\alpha^0/v_\alpha^0/(\hat{\xi} - \hat{\xi}^*)'I(\hat{\xi} - \hat{\xi}^*).
\]

In the case of Hoerl-Kennard ridge regression, where \( \hat{\xi}^* = 0 \) and \( \Sigma = I \), (4.6) becomes

\[
n_\alpha^2P_n^0\n_\alpha^0/v_\alpha^0(\hat{\xi} - \hat{\xi}^*). \quad \text{If additionally we replace } P_n^0\n_\alpha^0/v_\alpha^0 \text{ by 1, this becomes the well-known prescription for } k \text{ proposed by Hoerl, Kennard, and Baldwin (1975).}
\]

4.2. Directional compatibility tests.

The compatibility of the external data with the experimental data can also be tested with respect to specific linear combinations \( \mu_t = \xi_t'\hat{\xi} \) of interest, where \( \xi_t \) is in the row space of \( \hat{X} \) (or, equivalently, the row space of \( \hat{\Sigma} \)). Since

\[
\hat{\psi}_t = \hat{\xi}_t' - \xi_t'\hat{\xi}^* \sim N(0,\sigma^2_t I'((\xi_t'\xi_t)^{-1} + k^{-1}\xi_t'\xi_t)t),
\]

a test statistic for compatibility associated with the direction \( \xi_t \) is:

\[
\psi_t(k) = \frac{(\hat{\xi}_t' - \xi_t'\hat{\xi}^*)^2}{\xi_t'((\xi_t'\xi_t)^{-1} + k^{-1}\xi_t'\xi_t)\xi_t'\xi_t'},
\]

which is distributed as \( F_{1,v} \) (i.e., \( \psi_t(k) \) is distributed as \( |t| \)). Here we have used the notation \( \Sigma^- \) to refer to the Moore-Penrose generalized inverse of \( \Sigma \), to cover cases in which \( \Sigma \) may be singular (\( n_0 < p \)).

4.3. Maximum "safe" \( k \).

The direction in which the incompatibility between experimental and external data is most statistically significant can be found by maximizing the numerator of \( \psi_t(k) \) in (4.8) while holding the denominator fixed. This is simply a matter of maximizing a non-negative definite quadratic form over a given contour of a positive definite quadratic form. The
result is
\[ \psi^*(k) = \max_{\xi} \psi_{\xi}(k) = s^{-2}(\text{max. eigenvalue of } \zeta(k)) \]
\[ = s^{-2}q(k) \]
where
\[ \zeta(k) = X'X + k\mathbb{I}^{-1}k\mathbb{I}(\hat{e} - \bar{e})'(\hat{e} - \bar{e})', \]
and is independent of the particular contour over which the maximization is done. Since the rank of \( \zeta \) is one, the maximum eigenvalue of \( \zeta \) is the same as the trace, which is \( q(k) \). The maximizing direction \( \xi \) is proportional to \( \hat{e} - \bar{e} \).

We may now define the maximum "safe" \( k, k_s' \), by
\[ q(k_s') = s^2F_{1, \nu; \alpha} \]
for some suitably chosen percentage point of the \( F_{1, \nu} \) distribution. For \( k < k_s' \), no directional compatibility test statistic will be significant at level \( \alpha \).

4.4. Lower bounds for \( k \)

Compatibility tests are useful for setting upper limits on \( k \), but not for setting lower limits, since the closer \( k \) is to zero, the more compatible is the external data with the experimental data.

Recall that the reason we resort to the use of external data in the first place is that the data/model for the experiment are inadequate, i.e., the confidence intervals for some estimates of interest are too wide. This point of view immediately establishes as a lower bound for \( k \) the smallest value that ensures that the confidence interval (2.18) for all \( \xi'\hat{e} \) of interest will be "sufficiently narrow". If this is achieved at \( k = 0 \), we would be happy with the standard least squares analysis, since we would not need to use the external information, with all its attendant difficulties.

There will be some difficulty in practice if one is unwilling or unable to specify the maximum interval width "necessary" for the interval to be "useful". We would recommend that the end points of the interval (2.18) be plotted as a function of \( k \), or as a function of the more meaningful parameter \( \rho \) to be defined in equation (4.13). It should then be
apparent how much one gains (in the sense of narrowing the interval) and at what cost (in
the sense of increasing the reliance on the external information). This is unavoidably a
subjective judgment, but one which we feel is more relevant to the choice of a reasonable
value of $k$ than is the "stability" of the Hoerl-Kennard (1970) "ridge trace", a criterion
frequently used by practitioners of ridge regression.

Our approach will clearly lead to different choices of $k$ for different end uses,
where each end use is the estimation of a linear combination of regression coefficients.
For some end uses it may not be necessary to bring in any external information at all,
while for others, the external information may need to be weighted heavily. If this is too
much trouble, however, one may choose an overall measure of precision to plot against $k$
instead. A reasonable choice would be $\bar{V}_{\text{av}}(k)$, the average estimated variance of a
fitted value over the region of interest in $R$. Using (2.14) we obtain

$$\bar{V}_{\text{av}}(k) = \sigma^2(k) \text{tr}(\Sigma^\prime X + k\Sigma)^{-1}$$  \hspace{1cm} (4.12)

where $\Sigma$ is the region moment matrix $(E(g(d)g'(d)))$ for a uniform probability
distribution $\phi$ over the region of interest) and $\sigma^2(k)$ is given by (2.16).
Alternatively, if one were interested only in changes in response from one set of
conditions to another, one would replace $\Sigma$ in (4.12) by $2(\Sigma - \Sigma')$ where
$\Sigma' = E(g'(d))$. Then (4.12) would be the average estimated variance of the difference in
predicted response between two randomly chosen points in the region.

4.5. Percent information attributed to external data.

In communicating the results of a regression analysis of the sort we are considering
here, it is helpful to provide a quantitative measure of the shares of information
attributable to the experimental and external data. Again we rely on the work of Theil
(1963), who showed that the function

$$\beta = 100\text{tr}[k\Sigma^\prime X + k\Sigma]^{-1}$$  \hspace{1cm} (4.13)

is the only function that meets a certain set of reasonable criteria for expressing the per-
cent information in $\bar{V}(k)$ attributable to the external data. Interchanging $X^\prime X$ with
$k\Sigma$ in (4.13) yields the percent information attributable to the experimental data. This
is easily seen to equal $100 - \rho$, as it should. An attractive feature of this definition is that $\rho$ is invariant under nonsingular linear reparametrizations of the model.

In plotting the end points of intervals of type (2.18) as recommended above, it may be useful to use $\rho$ as the horizontal axis rather than $k$ itself. We feel that $\rho$ should be computed in any case, since it is more easily interpretable than $k$, and since its use places proper emphasis on the fact that $\tilde{\beta}$ is the direct result of incorporating external information.
5. BAYESIAN METHODS

As one might expect, there is a Bayesian parallel to the foregoing regression analysis for fixed \( k \) using mixed estimation. The evaluation of the choice of \( k \) using tests of compatibility is not a strictly Bayesian concept, however, but is a diagnostic check in the sense of Box (1980).

We shall review here what we would consider to be a standard Bayesian approach, based on "noninformative" priors, and shall indicate that the modifications needed to make it tractable and useful lead to the analysis we have described above. We shall utilize primarily the results of Box and Tiao (1973) and Box (1980). The reader is also referred to the paper of Lindley and Smith (1972) for additional discussion.

We assume the following subjective probability model

\[
P(\beta, \sigma^2, \beta^*, \sigma^2, \sigma_0^2) = \frac{(\sigma_0^2)^{(1+n/2)}}{(\sigma^2)^{(1+n/2)}(\sigma^2)^{(v/2-1)}}
\]

\[
\times \exp\left[-\frac{1}{2}(\beta^* - \beta_0)^T \Sigma (\beta^* - \beta_0) - \frac{1}{2} (\beta^* - \beta_0)^T \sum (\beta^* - \beta_0) - \frac{\nu \sigma^2}{2}\right].
\]

This may be derived by assuming that \( \ln \sigma_0^2, \ln \sigma^2, \) and \( \beta^* \) have independent locally uniform prior distributions and that

\[
P(\beta^*, \sigma^2, \sigma_0^2) = (\sigma_0^2)^{-n/2} \exp\left(-\frac{1}{20 \sigma_0^2} (\beta^* - \beta_0)^T \sum (\beta^* - \beta_0)\right),
\]

\[
P(\beta, \sigma^2, \sigma_0^2) = (\sigma^2)^{-n/2} \exp\left(-\frac{1}{20 \sigma^2} (\beta - \beta^*)^T \sum (\beta - \beta^*)\right),
\]

\[
P(\sigma^2 | \beta^*, \beta_0, \sigma_0^2) = (\sigma^2)^{v/2-1} \exp\left(-\frac{\nu \sigma^2}{2}\right).
\]

This model is equivalent to that considered by Box and Tiao (1973, Ch. 9), for the estimation of common regression coefficients using data from two independent sources. Direct adaptation of their results yields the following posterior density for \( \beta^* \):

\[
P(\hat{\beta}^* | \hat{\beta}, \hat{\sigma}^2, \hat{\sigma}_0^2) = ((\hat{\beta} - \beta_0)^T \sum (\hat{\beta} - \beta_0) + \nu \hat{\sigma}^2)^{-n/2} \exp\left(-\frac{\nu \hat{\sigma}^2}{2}\right).
\]

Unfortunately, this is an improper density since it is not integrable. (The portion of the integral in the neighborhood of \( \hat{\beta}^* \) can be made arbitrarily large.)
An alternative approach would be to fix $\sigma_0^2$ by assumption, and then use $P(\beta, s^2, \beta^2 | \sigma_0^2)$ as the basis for a diagnostic check on $\sigma_0^2$, in the spirit of Box (1980).

The posterior distribution of $\beta$ is then:

$$P(\beta | \hat{\beta}, s^2, \hat{\beta}^2, \sigma_0^2) = \frac{1}{2\sigma_0^2} \exp\left(-\frac{1}{2\sigma_0^2} (\beta - \hat{\beta})' \Sigma^{-1} (\beta - \hat{\beta}) \right)$$

(5.6)

Although this is a proper density of relatively simple form, we do not know of an easy way to obtain the posterior mean and variance of $\beta$. (One could do it by integrating the moments $E(\hat{\beta})$ and $E(\hat{\beta}^2)$, conditional on $\hat{\beta}, s^2, \hat{\beta}^2, \sigma_0^2$, numerically with respect to the $\sigma_0^2$.) Alternatively, one could seek the mode of (5.6) for use as an estimate of $\beta$.

This would require an iterative procedure similar to that suggested by Lindley and Smith (1972), who pointed out, however, that this provides only a point estimate of $\beta$. A quadratic Taylor's series expansion of the logarithm of (5.6) about the mode would lead to an approximate variance-covariance matrix. Even if all this were done, however, there are further difficulties in obtaining an appropriate test, in the sense of Box (1980), for use as a diagnostic check on the assumed $\sigma_0^2$.

All things considered, therefore, it is not surprising that we should want to fix $k$ by assumption, instead of $\sigma_0^2$, as Box (1980) did in an illustration of his estimation/criticism analysis. Box recommended that estimation be based on the (multivariate t) posterior distribution for $\beta$:

$$P(\beta | \hat{\beta}, s^2, \hat{\beta}^2, k) = \frac{(\hat{\beta} - \beta)' \Sigma (\hat{\beta} - \beta) - (n_o + n)/2}{(n_o + n)\sigma^2}$$

(5.7)

where $\hat{\beta}$ is given by (2.13) and $\widehat{\sigma^2}$ is given by (2.16). In particular, posterior confidence intervals for any linear combination of the regression coefficients are given by (2.18).

For diagnostic checking, we note that:

$$P(\beta, s^2, \hat{\beta}^2 | k) = \frac{(s^2)^{v/2-1}(q(k) + v\sigma^2)^{-(n_o + v)/2}}{2\pi \Gamma(v/2)}$$

(5.8)
If we regard this as a likelihood function for $k$, then the likelihood ratio test of any hypothesized $k$ is based on the statistic $\psi(k)$ in (4.1), i.e., it is the same as Theil's test for compatibility between experimental and external information. Box (1980) also derived $\psi(k)$ as one of several diagnostic checks based on the distribution of the experimental data conditional on all the underlying assumptions, and noted its equivalence to Theil's statistic.
REFERENCES


<table>
<thead>
<tr>
<th>REPORT DOCUMENTATION PAGE</th>
<th>READ INSTRUCTIONS BEFORE COMPLETING FORM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. REPORT NUMBER</td>
<td>2. GOVT ACCESSION NO.</td>
</tr>
<tr>
<td>2327</td>
<td></td>
</tr>
<tr>
<td>4. TITLE (and Subtitle)</td>
<td>5. TYPE OF REPORT &amp; PERIOD COVERED</td>
</tr>
<tr>
<td>USING EXTERNAL INFORMATION IN LINEAR REGRESSION, WITH A COMMENTARY ON RIDGE REGRESSION</td>
<td>Summary Report - no specific reporting period</td>
</tr>
<tr>
<td>I. MIXED ESTIMATION AND BAYESIAN METHODS</td>
<td></td>
</tr>
<tr>
<td>6. PERFORMING ORG. REPORT NUMBER</td>
<td></td>
</tr>
<tr>
<td>7. AUTHOR(a)</td>
<td></td>
</tr>
<tr>
<td>Toby J. Mitchell and Norman R. Draper</td>
<td></td>
</tr>
<tr>
<td>8. CONTRACT OR GRANT NUMBER(s)</td>
<td></td>
</tr>
<tr>
<td>DAAG29-80-C-0041</td>
<td></td>
</tr>
<tr>
<td>DAAG29-80-C-0113</td>
<td></td>
</tr>
<tr>
<td>9. PERFORMING ORGANIZATION NAME AND ADDRESS</td>
<td>10. PROGRAM ELEMENT, PROJECT, TASK AREA &amp; WORK UNIT NUMBERS</td>
</tr>
<tr>
<td>Mathematics Research Center, University of Wisconsin</td>
<td>4-Statistics &amp; Probability</td>
</tr>
<tr>
<td>610 Walnut Street</td>
<td></td>
</tr>
<tr>
<td>Madison, Wisconsin 53706</td>
<td></td>
</tr>
<tr>
<td>11. CONTROLLING OFFICE NAME AND ADDRESS</td>
<td>12. REPORT DATE</td>
</tr>
<tr>
<td>U. S. Army Research Office</td>
<td>January 1982</td>
</tr>
<tr>
<td>P.O. Box 12211</td>
<td></td>
</tr>
<tr>
<td>Research Triangle Park, North Carolina 27709</td>
<td></td>
</tr>
<tr>
<td>13. NUMBER OF PAGES</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
</tr>
<tr>
<td>14. MONITORING AGENCY NAME &amp; ADDRESS(if different from Controlling Office)</td>
<td>15. SECURITY CLASS. (of this report)</td>
</tr>
<tr>
<td>15a. DECLASSIFICATION/DOWNGRADING SCHEDULE</td>
<td>UNCLASSIFIED</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>16. DISTRIBUTION STATEMENT (of this Report)</td>
<td></td>
</tr>
<tr>
<td>Approved for public release; distribution unlimited.</td>
<td></td>
</tr>
<tr>
<td>17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>18. SUPPLEMENTARY NOTES</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>19. KEY WORDS (Continue on reverse side if necessary and identify by block number)</td>
<td></td>
</tr>
<tr>
<td>Collinearity; Dummy data; External information in regression; Ridge regression</td>
<td></td>
</tr>
<tr>
<td>20. ABSTRACT (Continue on reverse side if necessary and identify by block number)</td>
<td></td>
</tr>
<tr>
<td>In this article, we discuss ways of using &quot;dummy data&quot; and mixed estimation (Theil and Goldberger, 1961) to bring external information formally into linear regression problems when the experimental data/model are inadequate. This is a useful way of attacking the same practical problems that motivated the development of ridge regression (Hoerl and Kennard, 1970). The main practical issues considered are (i) what form should the &quot;dummy data&quot; take?, and (ii) how much weight should it be given relative to the experimental data? When specific prior information is unavailable, it is—</td>
<td></td>
</tr>
</tbody>
</table>
suggested that the dummy data should reflect a preference for "stable" response functions and it is shown how this can be accomplished. Guidelines for the choice of the weighting parameter $k$ (equivalent to the choice of the ridge parameter in ridge regression) are given. Upper limits for $k$ are based on various tests of compatibility between the external (dummy) data and the experimental data. Lower limits for $k$ are determined by the inadequacy of the data/model for the purpose(s) of the analysis. Finally, the parallel Bayesian approach is discussed, with emphasis on Box's (1980) framework of model estimation and criticism.