THE PREDICTOR'S AVERAGE ESTIMATED VARIANCE CRITERION FOR THE SELECTION-OF-VARIABLES PROBLEM IN GENERAL LINEAR MODELS

By

Ronald W. Helms

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 777

October 1971

Errata: The Predictor's Average Estimated Variance Criterion for the Selection-of-Variables Problem in General Linear Models.

1. The matrix X'X should be replaced by $\frac{1}{n}$ X'X

   (a) in lines 3, 4 and 12 on p. 17.

   (b) in line 7, p. 22.

   (c) in line 2, p. 25.

2. The expression of $AEV(\hat{y}) = s^2p$ should read $AEV(\hat{y}) = s^2p/n$ on line 13, p. 17, and $AEV(\hat{y}_i) = s_i^2p/n$ on line 2, p. 25.

3. In Table 2 (pp. 23-24) the calculated AEV statistics should be divided by 68.

THE PREDICTOR'S AVERAGE ESTIMATED VARIANCE CRITERION FOR THE SELECTION-OF-VARIABLES PROBLEM IN GENERAL LINEAR MODELS

Ronald W. Helms

Department of Biostatistics
University of North Carolina at Chapel Hill

1. Introduction

1.1 Background. The linear models selection-of-variables problem has troubled statisticians and researchers for many years and the substantial number of recent papers on this topic in statistical literature indicates considerable dissatisfaction with the available procedures. Draper and Smith (1966) surveyed a number of procedures, including forward selection, backward elimination, and stepwise regression (Efroymson, 1960), perhaps the most widely used least squares algorithm. Recent papers, such as Garside (1965), Hocking and Leslie (1967), LaMotte and Hocking (1970), and Schatzoff, Feinberg, and Tsao (1968) have tended to concentrate upon finding computationally efficient algorithms for selecting a "best" subset of independent variables in regression problems. These authors define "best" in terms of minimizing a residual sum of squares subject to various restrictions. Other authors have taken the alternative approach of defining new criteria for choosing the "best" subset of independent variables. Box and Draper (1959), recognizing the importance of bias, particularly in polynomial models, chose to work with the unweighted integrated mean square error (IMSE). Karson, Manson, and Hader (1969) developed a "minimum bias" estimator, the properties of which are based on the integrated mean square error. Michaels (1969) and Helms (1969) extended this work, and

Allen (1971) developed a selection procedure and estimator for minimizing the mean square error of prediction (MSEP) for a given vector $\underline{x}$ of values of the independent variables.

This paper builds on the work of this latter type: a new (?) criterion is proposed for the selection of variables or, equivalently, for selection among competing models. The criterion provides a basis for model selection procedures which are not subject to some of the weaknesses of IMSE- and MSEP-based procedures. However, as with all known stepwise procedures the exact statistical properties of the ones proposed here seem difficult to discover.

1.2 Notation. We adopt, essentially, the notation used in Graybill (1969). Let a linear model be given by

$$\underset{n\times 1}{\underline{y}} = \underset{n\times p}{X}\ \underset{p\times 1}{\underline{\beta}} + \underset{n\times 1}{\underline{\varepsilon}} \tag{1.1}$$

where $\underline{y}$ is observed, $\underline{\varepsilon}$ is random and unobservable with $E(\underline{\varepsilon})=\underline{0}$, $D(\underline{\varepsilon})=E(\underline{\varepsilon}\underline{\varepsilon}')=\sigma^2 I$, $\underline{\beta}$ is the unknown vector of "regression coefficients", and X is a matrix of known constants, x_{ij} denoting the i-th value of the j-th "independent" variable. We do not require that X be full rank; i.e., X'X may be singular. A least squares estimator of $\underline{\beta}$ (also a Best Linear Unbiased Estimator if $(X'X)^{-1}$ exists), is any vector $\underline{b}$ satisfying the "normal equations" $(X'X)\underline{b}=X'\underline{y}$. We shall be interested in solutions of the form $\underline{b}=(X'X)^{-}X'\underline{y}$, where $(X'X)^{-}$ is any symmetric generalized inverse of (X'X) with non-negative eigen values.

Under the assumptions of the model an unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = (\frac{1}{n-r})\underline{y}'[I-X(X'X)^{-}X']\underline{y},$$

where r=rank (X). This estimator is the same for all choices of the generalized inverse.

2. The Average $\mathrm{Var}(\hat{y})$

Let $\underline{x}$ be a $1\times p$ row vector of variables. (Each row of X is a particular value of $\underline{x}$.) Assume that the observed value of y at $\underline{x}$ really is a function of $\underline{x}$, say $\eta(\underline{x})$, plus a random error, ε, which satisfies the assumptions on the elements of $\underline{\varepsilon}$ in the model above. That is,

$$y = \eta(\underline{x}) + \varepsilon.$$

Given an estimator $\underline{b}$ and a value of $\underline{x}$ we estimate $\eta(\underline{x})$ by

$$\hat{\eta}(\underline{x}) = \hat{y}(\underline{x}) = \underline{x}\underline{b}.$$

Clearly, under the assumptions, $E(\hat{y}(\underline{x})) = \eta(\underline{x})$, and

$$\mathrm{Var}(\hat{y}(\underline{x})) = \sigma^2\, \underline{x}(X'X)^{-}\underline{x}'$$

which has the unbiased estimator

$$\hat{\mathrm{Var}}(\hat{y}(\underline{x})) = \hat{\sigma}^2\, \underline{x}(X'X)^{-}\underline{x}'.$$

Notice that $\hat{y}$ is a function defined for all $\underline{x}$ in some set $\Omega \subset E^p$, which we shall call the "Region of Interest." One would hope that the function $\hat{y}$ is a good approximation of the function η. The value of $\mathrm{Var}(\hat{y}(\underline{x}))$ is a relative measure of how well $\hat{y}$ estimates η at the single point $\underline{x}$, but it does not indicate how well $\hat{y}$ estimates η over the region of interest, Ω. One measure of the quality of the whole function $\hat{y}$ as an estimator of η on Ω is an average of $\mathrm{Var}(\hat{y}(\underline{x}))$ over Ω. The averaging procedure must be a weighted average, the weight at a point $\underline{x}$ being a quantitive expression of the importance that $\mathrm{Var}(\hat{y}(\underline{x}))$ be

small at $\underline{x}$. The following paragraph describes a quite general weighted average technique.

Let F denote a probability distribution function on E^p satisfying the following

(1) $\int_{E^p-\Omega} dF(\underline{x}) = 0$

(2) $\int_{\Omega} dF(\underline{x}) = 1$

(3) $\int_{\Omega} \underline{x}\, dF(\underline{x}) = \underline{\mu}$ (finite)

(4) $\int_{\Omega} \underline{x}'\underline{x}\, dF(\underline{x}) = \underset{(p\times p)}{M}$ (finite)

(5) $\int_{\Omega} (\underline{x}-\underline{\mu})'(\underline{x}-\underline{\mu}) dF(\underline{x}) = V$ (finite).

Integration of a function over Ω with respect to F is a general way of taking a weighted average of the function. Condition (1) simply means that zero weight is given to all subsets outside the region of interest, Ω. Condition (2) implies that F is "normalized" and to get the weighted average it is not necessary to divide by $\int dF(X)$. Conditions (3), (4), and (5) define $\underline{\mu}$ ($1\times p$), M ($p\times p$), and V ($p\times p$), the mean vector, the matrix of moments about the origin, and the covariance matrix of the distribution function, F, and require that μ, M, and V be finite.

Theorem: Weighted Average of $\mathrm{Var}(\hat{y}(\underline{X}))$. If F is a distribution function satisfying conditions (1) - (5) above, then

$$\int_{\Omega} \mathrm{Var}(\hat{y}(\underline{x}))dF(\underline{x}), \text{ denoted } E_F[\mathrm{Var}(\hat{y})],$$

has value

$$E_F[\mathrm{Var}(\hat{y})] = \sigma^2 \ \mathrm{tr}[(X'X)^{-}M]$$

$$= \sigma^2\{\mathrm{tr}[(X'X)^{-}V] + \underline{\mu}(X'X)^{-}\underline{\mu}'\}.$$

V and M are not required to be nonsingular. Note that the value depends upon the particular generalized inverse chosen, $(X'X)^{-}$, if X'X is singular.

Proof. Even though $\underline{x}$ is not a random vector, expected value notation and results are used in order to simplify and motivate the proof. Using well-known properties of the trace and expected value operators, we derive as follows:

$$E_F(\mathrm{Var}(\hat{y})) = E_F[\sigma^2 \ \underline{x}(X'X)^{-}\underline{x}']$$

$$= \sigma^2 \ E_F[\mathrm{tr} \ \underline{x}(X'X)^{-}\underline{x}'] = \sigma^2 \ E_F[\mathrm{tr}(X'X)^{-}\underline{x}'\underline{x}]$$

$$= \sigma^2 \ \mathrm{tr}[(X'X)^{-}E(\underline{x}'\underline{x})] = \sigma^2 \ \mathrm{tr}[(X'X)^{-}M]$$

$$= \sigma^2 \ \mathrm{tr}[(X'X)^{-}(V + \underline{\mu}'\underline{\mu})]$$

$$= \sigma^2\{\mathrm{tr}[(X'X)^{-}V] + \mathrm{tr}[(X'X)^{-}\underline{\mu}'\underline{\mu}]\}$$

$$= \sigma^2\{\mathrm{tr}[(XX)^{-}V] + \mathrm{tr}[\underline{\mu}(X'X)^{-}\underline{\mu}']\}$$

$$= \sigma^2\{\mathrm{tr}[(X'X)^{-}V] + \underline{\mu}(X'X)^{-}\underline{\mu}'\}. \quad \text{Q.E.D.}$$

Restriction: If the variables in $\underline{x}$ are not functionally independent (as in polynomial regression, for example) the region of interest, Ω, must contain only points which are possible values of $\underline{x}$. By condition (1) any subset of E^p containing only impossible values of $\underline{x}$ will have zero weight.

Note that the final result, $\sigma^2\ \mathrm{tr}[(X'X)^{-}V + \underline{\mu}(X'X)^{-}\underline{\mu}']$, depends on Ω and F only through the parameters $\underline{\mu}$ and V. In practice it will not be necessary to specify Ω and F explicitly; only $\underline{\mu}$ and V are needed.

Note also that the lemma applies to weight functions which are either continuous (as a probability density function) or discrete (as a discrete probability function). In the first case the weighted average is of the form: $\int_{\Omega} \mathrm{Var}(\hat{y}(\underline{x}))f(\underline{x})d\underline{x}$, $f(\underline{x})$ being the weight function, and in the second case the weighted average is of the form: $\sum_{\underline{x}\in\Omega} \mathrm{Var}(\hat{y}(\underline{x}))f(\underline{x})$, $f(\underline{x})$ being the weight at the point $\underline{x}$. The proof is sufficiently general to permit mixed weight functions, continuous in some variables and discrete in others. Even for such complicated weighting schemes the final result depends only upon $\underline{\mu}$ and V.

Estimation of Avg $\mathrm{Var}(\hat{y})$. Given the estimator $\hat{\sigma}^2$ for σ^2, the corresponding estimator of $E_F[\mathrm{Var}(\hat{y})]$ is obviously:

$$\hat{\sigma}^2\ \mathrm{tr}[(X'X)^{-}V + \underline{\mu}(X'X)^{-}\underline{\mu}'].$$

We call this value the Average Estimated Variance of $\hat{y}$, denoted $\mathrm{AEV}(\hat{y})$.

Choice of a Generalized Inverse. Assume X'X is singular and consider its spectral decomposition

$$X'X = PDP',\ P'P = PP' = I$$

$$D = \mathrm{diag}(d_1, d_2, \ldots, d_r, 0, \ldots, 0)$$

where r = rank (X'X). A symmetric generalized inverse of X'X can be taken to have the form

$$(X'X)^{-} = PD^{-}P',$$

where D^- is a diagonal generalized inverse of D. The condition $D = DD^-D$ is equivalent to

$$d_i^- = \begin{cases} 1/d_i \text{ if} & 1 \le i \le r \\ \text{arbitrary if} & r < i \le p. \end{cases}$$

where $D^- = \text{diag}(d_1^-, d_2^-, \ldots, d_p^-)$.

Consider the value of

$$\text{tr}[(X'X)^-M] = \text{tr}[PD^-P'M] = \text{tr}[D^-(P'MP)]$$
$$= \sum_{i=1}^{r} \frac{1}{d_i} \text{diag}_i(P'MP) + \sum_{i=r+1}^{p} d_i^- \text{diag}_i(P'MP)$$

where $\text{diag}_i(P'MP)$ denotes the i-th diagonal element of P'MP. The first r terms of the sum are completely specified. If d_i^- is completely arbitrary for $r<i\le p$ then the sum can be made to take any positive or negative value, hence the restriction that $d_i^- \ge 0$. Since the matrix P'MP must be positive semidefinite and has nonnegative diagonal elements, the sum is minimized for $d_i^- = 0$, $i=r+1,\ldots,p$. Thus the particular generalized inverse $(X'X)^- = PD^-P'$, where

$$D^- = \text{diag}(1/d_1, \ldots, 1/d_r, 0, \ldots, 0),$$

causes $\text{tr}[(X'X)^-M]$ to be minimized, subject to the restrictions.

3. Average Var($\hat{y}$) and Estimated Var($\hat{y}$) for Incomplete Models

Let us partition the model (1.1) as

$$\underline{y} = X_1\underline{\beta}_1 + X_2\underline{\beta}_2 + \underline{\varepsilon} \tag{3.1}$$

where X_i is $n \times p_i$, $p_1+p_2 = p$, and $\underline{\beta}_i$ is $p_i \times 1$. In this section we shall assume that rank (X) = rank $[X_1, X_2]$ = p, so that X'X, $X_1'X_1$ and $X_2'X_2$ are all nonsingular. If we fit only the $X_1\underline{\beta}_1$ part of the model to the data we obtain the following results:

$$\underline{b}_1 = (X_1'X_1)^{-1}X_1'\underline{y}$$

$$E(\underline{b}_1) = \underline{\beta}_1 + (X_1'X_1)^{-1}X_1'X_2\underline{\beta}_2$$

$$D(\underline{b}_1) = \sigma^2(X_1'X_1)^{-1}$$

$$\hat{y}_1(\underline{x}) = \underline{x}_1\underline{b}_1$$

$$E[\hat{y}_1(\underline{x})] = \underline{x}_1\underline{\beta}_1 + \underline{x}_1(X_1'X_1)^{-1}X_1'X_2\underline{\beta}_2$$

Let

$$\underline{r} = [\underline{x}_2 - \underline{x}_1(X_1'X_1)^{-1}X_1'X_2]$$

Then

$$\text{Bias}[\hat{y}_1(\underline{x})] = \underline{r}\underline{\beta}_2$$

$$\text{Bias}^2[\hat{y}_1(\underline{x})] = \underline{\beta}_2'\,\underline{r}'\underline{r}\underline{\beta}_2$$

$$\text{Var}[\hat{y}_1(\underline{x})] = \sigma^2\,\underline{x}_1(X_1'X_1)^{-1}\underline{x}_1'$$

Let

$$A_1 = X_1(X_1'X_1)^{-1}X_1'.$$

$$s_1^2 = (\frac{1}{n-p_1})\underline{y}'[I-X_1(X_1'X_1)^{-1}X_1']\underline{y} = (\frac{1}{n-p_1})\underline{y}'(I-A_1)\underline{y}.$$

$$E(s_1^2) = \sigma^2 + (\frac{1}{n-p_1})\underline{\beta}_2'X_2'(I-A_1)X_2\underline{\beta}_2.$$

Consider the (obviously biased) estimator of $\mathrm{Var}(\hat{y}_1(\underline{x}))$:

$$\hat{\mathrm{Var}}[\hat{y}_1(\underline{x})] = s_1^2\ \underline{x}_1(X_1'X_1)^{-1}\underline{x}_1'.$$

From $E(s_1^2)$ given above,

$$E_{\underline{\varepsilon}}\{\hat{\mathrm{Var}}[\hat{y}_1(\underline{X})]\} = \underline{x}_1(X_1'X_1)^{-1}\ \underline{x}_1'[\sigma^2 + \underline{\beta}_2'X_2'(I-A_1)X_2\underline{\beta}_2/(n-p_1)]$$

and the average estimated variance of $\hat{y}_1$ over Ω with respect to the weighting distribution function F can be found as follows. We partition $\underline{\mu}$ and V to match the partition of $\underline{x}$:

$$E_F(\underline{x}) = E_F[(\underline{x}_1,\underline{x}_2)] = (\underline{\mu}_1,\underline{\mu}_2)$$

$$E_F(\underline{x}'\underline{x}) = M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} = \begin{bmatrix} E_F(\underline{x}_1'\underline{x}_1) & E_F(\underline{x}_1'\underline{x}_2) \\ E_F(\underline{x}_2'\underline{x}_1) & E_F(\underline{x}_2'\underline{x}_2) \end{bmatrix}.$$

and $V_{ij} = M_{ij} - \underline{\mu}_i'\underline{\mu}_j$.

Then

$$E_F[\hat{\mathrm{Var}}(\hat{y}_1)] = \int_{\Omega} \hat{\mathrm{Var}}(\hat{y}_1(\underline{X}))dF(\underline{X})$$

$$= s_1^2\ \mathrm{tr}[(X'X_1)^{-1}M_{11}]$$

$$= \text{estimated average } \mathrm{Var}(\hat{y}_1).$$

The expected value (with respect to the distribution of $\underline{\varepsilon}$) of this estimator is

$$E_{\underline{\varepsilon}}[E_F\ \hat{Var}(\hat{y}_1)] = E_{\underline{\varepsilon}}(s_1^2)\ tr[(X_1'X_1)^{-1}(V_{11}+\underline{\mu}_1'\underline{\mu}_1)]$$

$$= [\sigma^2 + \underline{\beta}_2'X_2'(I-A_1)X_2\underline{\beta}_2/(n-p_1)]tr[(X_1'X_1)^{-1}(V_{11}+\underline{\mu}_1'\underline{\mu}_1)].$$

$$= E_F\ Var(\hat{y}_1) + (\frac{1}{n-p_1})\underline{\beta}_2'X_2'(I-A_1)X_2\underline{\beta}_2\ tr[(X_1'X_1)^{-1}(V_{11}+\underline{\mu}_1'\underline{\mu}_1)]$$

It can be seen that the averaged estimated $Var(\hat{y}_1)$, where the estimate is based on s_1^2, contains contributions from both variance and bias. It can be shown that the integrated (with respect to F) mean square error of $\hat{y}$, is:

$$IMSE(\hat{y}_1) = \int_{\Omega} E[\hat{y}_1(\underline{x}) - (\underline{x}_1\underline{\beta}_1+\underline{x}_2\underline{\beta}_2)]^2 dF(\underline{x})$$

$$= \sigma^2\ tr(X_1'X_1)^{-1}M_{11} + \begin{bmatrix}(X_1'X_1)^{-1}X_1'X_2\underline{\beta}_2\\ \underline{\beta}_2\end{bmatrix}'\begin{bmatrix}M_{11} & M_{12}\\ M_{21} & M_{22}\end{bmatrix}\begin{bmatrix}(X_1'X_1)^{-1}X_1'X_2\underline{\beta}_2\\ \underline{\beta}_2\end{bmatrix}$$

where

$$M = V + \underline{\mu}'\underline{\mu} = \begin{bmatrix}V_{11} + \underline{\mu}_1'\underline{\mu}_1 & V_{12} + \underline{\mu}_1'\underline{\mu}_2\\ V_{21} + \underline{\mu}_2'\underline{\mu}_1 & V_{22} + \underline{\mu}_2'\underline{\mu}_2\end{bmatrix}.$$

One can see that the expected averaged estimated $Var(\hat{y}_1)$, namely $E_\varepsilon\ E_F\ \hat{Var}(\hat{y}_1)$, is not generally equal to the $IMSE(\hat{y}_1)$ unless $\underline{\beta}_2=\underline{0}$; however, if $\underline{\beta}_2\tilde{=}\underline{0}$ the two should be approximately equal, both being continuous functions. It should also be noted that $X_1'X_2 = 0$ is not sufficient to make the two equal, for if $X_1'X_2 = 0$,

$$E_\varepsilon\ E_F\ \hat{Var}(\hat{y}_1) = tr[(X_1'X_1)^{-1}M_{11}][\sigma^2 + \underline{\beta}_2'X_2'X_2\underline{\beta}_2/(n-p_1)]$$

$$IMSE(\hat{y}_1) = \sigma^2\ tr[(X_1'X_1)^{-1}M_{11}] + \beta_2'M_{22}\underline{\beta}_2.$$

Although these similarities are interesting they are not very useful. It is useful to note that when the fitted model is incomplete the averaged estimated $\mathrm{Var}(\hat{y}_1)$ can be made large by a large value of any of:

(a) $\mathrm{tr}[(X_1'X_1)^{-1}M_{11}]$, the ill-condition of $(X_1'X_1)$ with respect to the weight function.

(b) σ^2, the variance part of $E(s_1^2)$

(c) $\underline{\beta}_2'X_2'(I-A_1)X_2\underline{\beta}_2$, the bias part of $E(s_1^2)$.

Thus the averaged estimated variance (AEV) of $\hat{y}_1$ can be used as a relative measure of how well $\hat{y}_1$ approximates and estimates the true response function η. If $\hat{y}_1$ fails due to a bad experimental design, X_1, due to a large underlying variance, σ^2, or due to large bias, or any combination of these, the failure will be reflected by an increase in the $\mathrm{AEV}(\hat{y}_1)$.

4. Using Averaged Estimated Variance (AEV) as a Criterion for Comparing Competing Models.

The utility of the AEV statistic stems from its use as a criterion for choosing among several models under consideration to fit one set of data. Let us assume that the model (1.1) is valid and construct h "sub-models".

Let:

Z_i $(n \times p_i)$ be a sub-matrix of X,

$\underset{1\times p_i}{z_i}$ be a sub-vector of $\underline{x}$, and

$\underset{p_i \times 1}{\alpha_i}$ be a sub-vector of $\underline{\beta}$, i=1,2,...,h.

In each case we fit the model $E(\underline{y}) \cong Z_i\underline{\alpha}_i$ by taking $\underline{a}_i$, the least squares estimate of $\underline{\alpha}_i$, as a solution to the normal equations:

$$(Z_i'Z_i)\underline{a}_i = Z_i'\underline{y},$$

or

$$\underline{a}_i = (Z_i'Z_i)^-Z_i'\underline{y}$$

where $(Z_i'Z_i)^-$ is the generalized inverse indicated in the subsection "Choosing a Generalized Inverse", unless there is a compelling reason to choose another. We compute an estimate of the residual variance about the i-th model:

$$s_i^2 = \frac{1}{(n-p_i)}\,[\underline{y}'\underline{y}-\underline{a}_i'Z_i'\underline{y}] = \left(\frac{1}{n-p_i}\right)\,\underline{y}'[I-Z_i(Z_i'Z_i)^-Z_i']\underline{y},$$

which is invariant to choices of $(Z_i'Z_i)^-$. Now, $\hat{y}_i(\underline{x}) = \underline{z}_i\underline{a}_i$, which we also denote by $\hat{y}_i(\underline{z}_i)$. We assume a weighting distribution function, F, and region of interest of $\underline{x}$ have been chosen and have finite moments given by:

$$\underset{1\times p}{\underline{\mu}} = E_F(\underline{x})$$

$$\underset{p\times p}{M} = E_F(\underline{x}'\underline{x})$$

$$\underset{p\times p}{V} = E_F(\underline{x}-\underline{\mu})'(\underline{x}-\underline{\mu}) = M-\underline{\mu}'\underline{\mu}$$

For convenience we shall use the notation

$\underset{1\times p_i}{\underline{\mu}_i} = E_F(\underline{z})$, a sub-vector of $\underline{\mu}$,

$M_i = E_F(\underline{z}_i'\underline{z}_i)$, a $p_i\times p_i$ sub-matrix of M, and

$V_i = E_F(\underline{z}_i-\underline{\mu}_i)'(\underline{z}_i-\underline{\mu}_i)$

$= M_i-\underline{\mu}_i'\underline{\mu}_i$, a $p_i\times p_i$ sub-matrix of V.

Then we compute

$$AEV(\hat{y}_i) = s_i^2\ tr[(Z_i'Z_i)^-M_{ii}]$$

$$= s_i^2\{tr[(Z_i'Z_i)^-V_{ii}] + \underline{\mu}_i(Z_i'Z_i)^-\underline{\mu}_i'\}.$$

We then choose to use the sub-model which has the smallest $AEV(\hat{y}_i)$ value.

4.1 A Stepwise Regression Procedure. The AEV criterion can be used as the guiding criterion for selecting and discarding variables in a modification of Efroymson's stepwise regression procedure in the following manner. First choose Z_1 and $\underline{z}_1$ to include variables which are to be "forced" into the model. For example, $\underline{z}_1$ may contain a constant term so that one of the coefficients in $\underline{\alpha}_1$ is an intercept. $AEV(\hat{y}_1)$ is computed. At the i-th stage, the correlations are computed between the residuals, $\underline{y} - Z_i\underline{a}_i$, and each of the variables not in $\underline{z}_i$. The vector $\underline{z}_{i+1}$ is taken as the $p_{i+1} = p_i+1$ vector formed by adjoining the variable with the highest correlation to $\underline{z}_i$. $AEV(\hat{y}_{i+1})$ is computed; if $AEV(\hat{y}_{i+1}) < AEV(\hat{y}_i)$, another variable is added in the same manner. If $AEV(\hat{y}_{i+1}) > AEV(\hat{y}_i)$, the procedure is terminated and $\hat{y}_i$ is taken as the "best" sub-model. This is actually a forward selection procedure in that once a variable enters it is never deleted. Also, variables are selected for entering only on the basis of reduction of s_i^2, not on the basis of reduction of $AEV(\hat{y}_i)$. However, this procedure allows efficient calculation procedures and can be added with little difficulty to existing stepwise general linear models (regression through the origin) programs.

4.2 An Alternative Stepwise Regression Procedure. This procedure is based directly on the AEP criterion but efficient computational algorithms have not been developed for it. It is also defined recursively.

Step 1. Let $\underline{z}_1$ be the vector of variables to be "forced" into the model. Compute $AEV(\hat{y}_1)$.

Step i(i>1). Assume model i, i.e., $\underline{z}_i$, $\underline{a}_i$, $\hat{y}_i(\underline{x})$, etc., are given.

(a) For each variable in $\underline{z}_i$, compute the value AEV would take if that variable were deleted, and denote the result c_j, where

the variable under consideration is x_j.

(b) For each variable not in $\underline{z}_i$, compute the value AEV would take if that variable were added, and denote the result c_j, where the variable under consideration is x_j.

(c) Let $c_{min} = \min_{1 \le j \le p} c_j$. If $c_{min} > AEV(\hat{y}_i)$, stop, for adding or deleting any variable would increase the average estimated variance. If $c_{min} < AEV(\hat{y}_i)$, act upon x_j where $c_j = c_{min}$. That is, if x_j is in $\underline{z}_i$, delete it, producing $\underline{z}_{i+1}$, and go to next step. If x_j is not in $\underline{z}_i$, add it, producing $\underline{z}_{i+1}$, and go to the next step. It is not now known whether this procedure can be implemented as an efficient computational algorithm; this will be a subject of further research.

4.3 Advantages of the AEV Criterion. The AEV criterion for selection among sub-models has two clear advantages over criteria now in common use.

(1) The AEV statistic is a measure of how well the function $\hat{y}$ approximates and estimates the function η over the whole region of interest. Both the quality of the approximation (manifested via the squared-bias terms in s^2) and the quality of the estimation (manifested in the variance term in s^2 and in the term $tr[(X'X)^{-}M]$) are reflected in the statistic. The statistic is based on the performance of $\hat{y}$ over the whole region of interest, Ω, not on just one point.

(2) The use of the AEV statistic is stepwise procedures results in a clear-cut stopping rule which is readily understandable and non-arbitrary: if the next step would increase the average estimated variance of the prediction function, it is not taken.

5. Some Weighting Functions and Moments

The AEV criterion depends explicitly upon the moments of the weighting function-region of interest combination and might appear, therefore, to be somewhat arbitrary or subjective. However, note that the estimate of the coefficient vector in a linear model is an explicit function of the X matrix, which is just as "arbitrary" as the moments of the weighting function.

5.1 The Case $M = X'X$. Neither the X matrix nor the moments of the weight function are arbitrary, of course. If the experiment is designed, i.e., if X is determined by the experimenter before $\underline{y}$ is observed, then X clearly reflects the fact that the experimenter has in mind a region of interest and a weighting function; X'X is probably close to the matrix $M = E_F(\underline{x}'\underline{x})$. If X "occurs", i.e., if X is random and it is desired to make conditional inferences about the distribution of y given $\underline{x}$, then the X matrix again reflects the region of interest and the weighting function because the rows of X are observations from the region of interest and natural weighting function and X'X is again probably close to M. We see that in two very different situations the X'X matrix is a good M matrix if there are no reasons to choose otherwise.

It is at this point that we recall that it is not necessary to specify the weighting function and region of interest if the M matrix (or V and $\underline{\mu}$) is known. However it is also worthwhile to illustrate a weight function-region of interest combination which leads to $M = X'X$. Let the region of interest be $\Omega = E^p$ and define the weight function as:

$$f(\underline{x}) = \frac{1}{n} \cdot (\text{number of rows in X that equal } \underline{x}), \text{ where}$$

$$n = \text{number of rows in X.}$$

Clearly f is a discrete weight function which takes non-zero values only at points $\underline{x}$ which are equal to a row of X, i.e., only at the data points. For discrete weighting functions the integral is a sum and $E_F(\underline{x}'\underline{x}) = X'X = M$. Also $\underline{\mu} = (\frac{1}{n})\ \underline{1}'X$ and $V = X'X - \underline{\mu}'\underline{\mu}$.

If the fact that f is discrete, i.e., that $\mathrm{Var}(\hat{y})$ is being averaged over only n specific points, is bothersome, notice that there is also a p-variate normal distribution over E^p with mean $\underline{\mu}$ and covariance V and, in effect, $\mathrm{Var}(\hat{y})$ is being averaged over E^p with this weighting function. One can also construct a number of other distributions (weighting functions) over various subsets of E^p with the moments $\underline{\mu}$ and V.

It is interesting to note that when M = X'X the AEV statistic has the form:

$$\mathrm{AEV}(\hat{y}) = s^2\ \mathrm{tr}[(X'X)^-(X'X)] = s^2\ \mathrm{rank}\ (X'X).$$
$$= s^2\ p, \text{ if } X'X \text{ is nonsingular.}$$

This result depends upon the choice of $(X'X)^-$ specified in the section "Choosing a Generalized Inverse"; any other allowable choice would yield a larger $\mathrm{AEV}(\hat{y})$.

In this special case if one is comparing two sub-models, say model i and model j with p and p+1 terms respectively, then model j is judged "better than" model i only if

$$s_i^2 p > s_j^2(p+1), \text{ i.e., if}$$

$$s_j^2 < s_i^2(\frac{p}{p+1}), \text{ or}$$

$$s_i^2 - s_j^2 > s_j^2/p.$$

This result is the basis for a stopping rule which could easily be added to present stepwise regression computer programs.

5.2 The Case: V is Diagonal. Consider the situation in which the experimenter specifies the region of interest by specifying the "range of interest" for each variable. That is, he specifies ℓ_i, and u_i such that $\ell_i \leq x_i \leq u_i$, i=1,2,...,p. In such a case if may be reasonable to construct the weight function as the product of p individual weight functions, one per variable. For the i-th variable, two extremes in choices of weight function are:

(a) The uniform distribution on $[\ell_i, u_i]$. Here $\mu_i = (\ell_i + u_i)/2$ and $v_{ii} = (u_i - \ell_i)/12$.

(b) The normal distribution arranged so that $\mu_i - 2.5\sqrt{v_{ii}} = \ell_i$; $\mu_i + 2.5\sqrt{v_{ii}} = u_i$. Then $\mu_i = (\ell_i + u_i)/2$; $v_{ii} = (u_i - \ell_i)^2/25$.

These distributions represent extremes, for although U-shaped and J-shaped distributions can be used they will probably not be used often in real problems.

When the weight function is constructed in this fashion as the product of individual ("marginal") weight functions the joint weight function behaves like the joint frequency function of independent random variables, which implies that V is diagonal. When V is diagonal the AEV statistic takes the form

$$\begin{aligned} AEV(\hat{y}) &= s^2\{tr[(X'X)^- V] + \underline{\mu}(X'X)^- \underline{\mu}'\} \\ &= s^2\, \underline{\mu}(X'X)^- \underline{\mu}' + s^2 \sum_{i=1}^{p} v_{ii}\, diag_i[(X'X)^-] \end{aligned}$$

It is often the case that $x_1 \equiv 1$ (to provide the "intercept term") and all the other variables are "centered" ($\mu_i = 0, i>1$). In this very special case

$$AEV(\hat{y}) = s^2\, diag_1(X'X)^- + s^2 \sum_{i=2}^{p} v_{ii}\, diag_i(X'X)^-,$$

and only the diagonal elements of $(X'X)^-$ enter the calculations.

Just as there are many possible designs (X-matrices) for an experiment there are many possible ($\underline{\mu}$,V)-moments for the AEV statistic used in the analysis. The final results will depend upon both the design and the moments and careful consideration should go into the selection of both. Fortunately the design and the moments are both based on an experimenter's region of interest and weighting of interest and are therefore closely related.

6. An Example

This example comes from the area of algal assays. An environmental scientist can assay the growth potential of a water sample by filtering or killing indigenous algae, innoculating the sample with a particular number of algae from a specified species, incubate the algae culture under standard conditions, and determine the "accumulated biomass" after, say, 21 days. The "accumulated biomass" is, by definition, the dry weight of the algae present at the end of the period. Unfortunately, dry weight determinations are difficult and expensive to make and are prone to accidents. Also, the dry weight cannot be determined until the end of the incubation period, while it is very desirable to be able to make measurements at intermediate times. Several non-destructive, less expensive measurements are closely related to dry weight, including (1) the optical density, (2) a determination of the net algal carbon in a small sample, (3) the amount of chlorophyll fluorescence under a certain wavelength of light, and (4) the number of cells in a small sample, counted electronically or by hemacytometer.

An experiment was performed by Dr. C. M. Weiss of the Department of Environmental Sciences and Engineering at the University of North Carolina to determine functions for estimating dry weight from the other measurements. The experiment was designed to provide a representative range of final biomasses. Summary statistics for the data (excluding several definite outliers) are given in Table 1.

From the high simple correlations, from previous experience, and from examination of data plots it was clear that a simple linear regression of dry weight on any one of the other variables would produce a nice fit of the data. However, there were several cases in which one or two points appeared to be

outliers for one regression (e.g., dry weight vs. cell count) but lay almost directly on another regression line (e.g., dry weight vs. optical density). This suggests that, in spite of the high intercorrelations among the "independent" variables, the different variables might carry enough distinct information to justify a multiple regression.

Table 1. Descriptive Statistics for the Weiss Algae Data.

A. Means and Standard Deviations (N=68 observations)

	Variable	Mean	Standard Derivation
1	Net Carbon	27.13	33.60
2	Chlorophyll	171.35	202.17
3	Optical Density	0.09425	0.10842
4	Cell Count	1920.51	2143.79
5	Dry Weight	40.17	48.76

B. Covariance/Correlation Matrix. (Covariances above diagonal, correlations below)

Variable	1	2	3	4	5
1	1128.71	6147.77	3.581	66996.6	1623.44
2	0.905	40872.18	20.814	419881.3	9074.97
3	0.983	0.950	0.012	225.82	5.22
4	0.930	0.969	0.972	4595845.	98082.0
5	0.991	0.921	0.988	0.938	2377.76

The BMD02R stepwise regression program was run on the data with results summarized in Table 2. The program's default options for "F to enter" and "F to delete" were used. All variables entered the estimation function.

In this case the variables are all random; it is desired to make inferences about the conditional distribution of dry weight given the values of the other variables. It is important for the estimating function to fit the true function where the data are, which suggests that $M = X'X$, $\underline{\mu} - (\frac{1}{n})\underline{1}'X$ and V = sample variance covariance matrix are appropriate moment matrices for the weight function-region of interest. These matrices are shown in Table 1.

Table 2. Summary of BMD02R Stepwise Regression Fitting of Weiss Algae Data.

SUB-PROBLM 1
DEPENDENT VARIABLE 5
MAXIMUM NUMBER OF STEPS 10
F-LEVEL FOR INCLUSION 0.010000
F-LEVEL FOR DELETION 0.005000
TOLERANCE LEVEL 0.001000

VARIABLE NO.	IDENT. NAME
1	NET CARBON
2	CHLOROPHYLL
3	OPTICAL DENS.
4	CELL COUNT
5	DRY WEIGHT

STEP NUMBER 1
VARIABLE ENTERED 1

MULTIPLE R 0.9910
STD. ERROR OF EST. 6.5863

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE	F RATIO
REGRESSION	1	156446.750	156446.750	3606.479
RESIDUAL	66	2863.039	43.379	

$AEV(\hat{y}) = 86.76$

VARIABLES IN EQUATION				VARIABLES NOT IN EQUATION			
VARIABLE	COEFFICIENT	STD. ERROR	F TO REMOVE	VARIABLE	PARTIAL CORR.	TOLERANCE	F TO ENTER
(CONSTANT	1.13678)						
1	1.43832	0.02395	3606.4773 (2)	2	0.41383	0.1807	13.4323 (2)
				3	0.55404	0.0333	28.7903 (2)
				4	0.33429	0.1347	8.1776 (2)

STEP NUMBER 2
VARIABLE ENTERED 3

MULTIPLE R 0.9938
STD. ERROR OF EST. 5.5250

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE	F RATIO
REGRESSION	2	157325.625	78662.813	2576.914
RESIDUAL	65	1984.189	30.526	

$AEV(\hat{y}) = 91.58$

VARIABLES IN EQUATION				VARIABLES NOT IN EQUATION			
VARIABLE	COEFFICIENT	STD. ERROR	F TO REMOVE	VARIABLE	PARTIAL CORR.	TOLERANCE	F TO ENTER
(CONSTANT	-0.35593)						
1	0.85765	0.11007	60.7148 (2)	2	-0.02195	0.0741	0.0309 (2)
3	182.99854	34.10553	28.7903 (2)	4	-0.31171	0.0374	6.8877 (2)

Table 2. (cont'd). Summary of BMD02R Stepwise Regression Fitting of Weiss Algae Data.

STEP NUMBER 3
VARIABLE ENTERED 4

MULTIPLE R 0.9944
STD. ERROR OF EST. 5.2906

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE	F RATIO
REGRESSION	3	157518.375	52506.125	1875.849
RESIDUAL	64	1791.398	27.991	

$AEV(\hat{y}) = 111.96$

VARIABLES IN EQUATION				VARIABLES NOT IN EQUATION			
VARIABLE	COEFFICIENT	STD. ERROR	F TO REMOVE	VARIABLE	PARTIAL CORR.	TOLERANCE	F TO ENTER
(CONSTANT	-0.21529)						
1	0.66167	0.12917	26.2398 (2)	2	0.15043	0.0575	1.4587 (2)
3	321.32886	62.00612	26.8553 (2)				
4	-0.00409	0.00156	6.8877 (2)				

STEP NUMBER 4
VARIABLE ENTERED 2

MULTIPLE R 0.9945
STD. ERROR OF EST. 5.2718

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE	F RATIO
REGRESSION	4	157558.938	39389.734	1417.335
RESIDUAL	63	1750.859	27.791	

$AEV(\hat{y}) = 138.96$

VARIABLES IN EQUATION				VARIABLES NOT IN EQUATION			
VARIABLE	COEFFICIENT	STD. ERROR	F TO REMOVE	VARIABLE	PARTIAL CORR.	TOLERANCE	F TO ENTER
(CONSTANT	-0.12213)						
1	0.69573	0.13176	27.8802 (2)				
2	0.01605	0.01329	1.4587 (2)				
3	301.94946	63.83473	22.3745 (2)				
4	-0.00510	0.00177	8.3613 (2)				

F-LEVEL OR TOLERANCE INSUFFICIENT FOR FURTHER COMPUTATION

The AEV statistic was computed for each step taken by the program, with results given in Table 2. (Because M = X'X, $AEV(\hat{y}_i) = s_i^2 p$, where p = number of variables in the model, including constant term.) It is interesting to note that the BMD program, with default values for entering and deleting variables, entered all of the variables, while if the AEV criterion had been used only the intercept and net carbon would have been entered.

The values of the Averaged Estimated Variances for the multivariate models are interesting, too, in that they form an increasing sequence as more variables are added, each model being worse than the one before.

7. Summary

The Averaged Estimated Variance is introduced as a measure of how well an estimation function, $\hat{y}(\underline{x})$, estimates (stochastic property) and approximates (bias, non-stochastic property) a response function, $\eta(\underline{x})$ over an entire region of interest, Ω, of values of $\underline{x}$. The statistic is the weighted average of an estimate of $\mathrm{Var}(\hat{y}(\underline{x}))$, averaged over Ω with respect to a normalized weight function. It is shown that the $\mathrm{AEV}(\hat{y})$ depends on the weight function and region of interest only through the first two moments of the weight function over Ω, and computational formulas are derived.

The expected value of the $\mathrm{AEV}(\hat{y})$ statistic is derived for complete and incomplete models and is compared with the Integrated Mean Square Error of $\hat{y}$.

Several types of weight functions are discussed and special forms of $\mathrm{AEV}(\hat{y})$ are shown for each. A stepwise regression procedure based on the AEV statistic is outlined and discussed. An example is given in which the AEV criterion is compared with the usual "F to enter" and "F to remove" criteria of the BMD stepwise regression procedure.

8. Acknowledgements

This work was supported in part by National Institute of Health, Institute of General Medical Science Grant No. GM-17868-04 and GM-13625 and by the Environmental Protection Agency - Water Quality Office Project - 16010DQT through a contract with the Department of Environmental Sciences and Engineering, School of Public Health, University of North Carolina.

Data processing for the example was performed by Mr. Robert Middour and Mrs. Diane Gabriel.

9. References

Allen, D. M. (1971). Mean Square Error of Prediction as a Criterion for Selecting Variables. Technometrics 13, pp. 469-476.

Box, G. E. P. and Draper, N. R. (1959). A Basis for the Selection of a Response Surface Design. J. Amer. Statistical Assoc. 54, pp. 622-654.

Draper, N. R. and Smith, H. (1966). Applied Regression Analysis. John Wiley and Sons, Inc. New York.

Efroymson, M. A. (1960). Multiple Regression Analysis, in Mathematical Methods for Digital Computers, A. Ralston and H. S. Wilf, editors. John Wiley and Sons, Inc. New York.

Garside, M. J. (1965). The Best Sub-set in Multiple Regression Analysis. Applied Statistics 14, pp. 196-200.

Helms, R. W. (1969). A Procedure for the Selection of Terms and Estimation of Coefficients in a Response - Surface Model with Integration-Orthogonal Terms. Institute of Statistics Mimeo Series No. 646, Department of Biostatistics, University of North Carolina, Chapel Hill, N. C.

Hocking, R. R. and Leslie, R. B. (1967). Selection of the Best Subset in Regression. Technometrics 9, pp. 531-540.

Karson, M. J., Manson, A. R., and Hader, R. J. (1969). Minimum Bias Estimation and Experimental Design for Response Surfaces. Technometrics 11, pp. 461-476.

Lamotte, L. R. and Hocking, R. R. (1970). Computational Efficiency in the Selection of Regression Variables. Technometrics 12, pp. 83-93.

Michaels, S. E. (1969). Optimum Design and Test/Estimation Procedures for Regression Models. Unpublished Ph.D. thesis, Department of Experimental Statistics, North Carolina State University.

Schatzoff, M., Feinberg, S., and Tsao, R. (1968). Efficient Calculations of All Possible Regressions. Technometrics 10, pp. 769-779.