The Relationship between Variable Selection and Data Agumentation and a Method for Prediction

Author(s): David M. Allen
Source: *Technometrics*, Vol. 16, No. 1 (Feb., 1974), pp. 125-127
Published by: Taylor & Francis, Ltd. on behalf of American Statistical Association and American Society for Quality
Stable URL: http://www.jstor.org/stable/1267500
Accessed: 01-07-2018 05:01 UTC

TECHNOMETRICS©, VOL. 16, NO. 1, FEBRUARY 1974

The Relationship Between Variable Selection and Data Agumentation and a Method for Prediction

David M. Allen

Department of Statistics
University of Kentucky
(Lexington, Kentucky)

We show that data augmentation provides a rather general formulation for the study of biased prediction techniques using multiple linear regression. Variable selection is a limiting case, and Ridge regression is a special case of data augmentation. We propose a way to obtain predictors given a credible criterion of good prediction.

KEY WORDS

Augmentation
Variable Selection
Prediction
Regression

1. INTRODUCTION

In exploratory experiments researchers often collect data on a response variable and several potential predictor variables. Additional predictor variables are frequently created by taking functions of the observed predictor variables. The traditional method of obtaining a prediction equation is multiple linear regression. The hazard of using too many predictor variables is widely known. The reason for the problem is that the addition of a variable to a least squares prediction equation almost always increases (never decreases) the variance of a predicted response [1, 9]. Addition of a variable may decrease the squared bias, but this decrease is often small relative to the increase in variance.

Nearly all of the proposed solutions to this problem can be classified into one of two categories:

(i) Variable selection techniques.
(ii) Augmentation of the variable matrix by dummy observations.

Our first objective is to show that (i) is a limiting case of (ii), and thus (ii) provides a more general framework for study. Secondly we propose a way to obtain predictors given a credible criterion of good prediction.

2. THE RELATIONSHIP BETWEEN VARIABLE SELECTION AND DATA AUGMENTATION

The classical multiple linear regression model is

$$\underset{(n\times 1)}{Y} = \underset{(n\times r)}{X}\ \underset{(r\times 1)}{\beta} + \underset{(n\times 1)}{\epsilon} \qquad (1)$$

Received Nov. 1972; revised June 1973

where Y is a vector of responses; X is a known, full rank matrix of nonstochastic variables; β is the unknown weight vector corresponding to X; and ϵ is a normally distributed vector of random variables having expected value 0 and dispersion matrix $I\sigma^2$.

We first review a technique for estimating β subject to linear restrictions. Let P be a $(p \times r)$ matrix of full rank. The value of β minimizing $(Y - X\beta)'(Y - X\beta)$ subject to $P\beta = 0$ is

$$b - (X'X)^{-1}P'(P(X'X)^{-1}P')^{-1}Pb \qquad (2)$$

where $b = (X'X)^{-1}X'Y$. This can be shown by a straight-forward application of Lagrange's multipliers.

We now consider the effect of augmenting X and Y with dummy observations. The generalized least squares estimate of β based on the augmented matrices

$$\left[\frac{X}{P}\right] \quad \text{and} \quad \left[\frac{Y}{O}\right]$$

is

$$(X'X + P'WP)^{-1}X'Y \qquad (3)$$

where $W = \text{Diag}\ (w_1 w_2 \cdots w_p)$ and $w_i > 0$, $i = 1, 2, \cdots, p$. In this formulation, $P\beta$ represents a set of linear combinations of the elements of β that we have prior belief is null. The matrix W represents the strength of this belief. Specifically, w_i represents the number of observations, having expected value equal to the i-th element of $P\beta$, that we believe necessary to provide information equal to our prior information. If Y is normally distributed, then (3) is the Bayesian estimator of β using a prior distribution of $P\beta$ given σ^2 of $N_r(0, W^{-1}\sigma^2)$. An expression algebraically identical to (3) is

$$b - (X'X)^{-1}P'(W^{-1} + P(X'X)^{-1}P')^{-1}Pb. \qquad (4)$$

Note that (4) approaches (2) as $\min\ (w_1 w_2 \cdots w_p)$ approaches infinity.

Throughout the remainder of this paper we will consider P to have a special structure. Let every row of P contain one "1" and all other elements be "0". Estimator (2) is equivalent to deleting variables corresponding to columns of P that contain a 1; calculating least squares estimators of the remaining coefficients; and then inserting zeros in the positions of the deleted variables. Since (2) represents a prediction equation with deleted variables and is a limiting form of (4), variable selection is said to be a limiting form of data augmentation.

2. Choosing A Predictor

For our special P the matrix $P'WP$ in (3) is diagonal and will be denoted by D. Thus we will consider estimators of the form

$$\hat{\beta} = (X'X + D)^{-1}X'Y \tag{5}$$

where $D = \text{diag}\ (d_1 d_2 \cdots d_r)$ and $d_j \geq 0$, $j = 1, 2, \cdots, r$. For any fixed $D \neq 0$, predictors using $\hat{\beta}$ have variances no larger than predictors using $\underline{b}$.

Hoerl and Kennard [5, 6] present a technique they call ridge regression. Ridge regression is a way to determine D in (5). They require that D be proportional to diag $(X'X)$ and use subjective rules for determining the constant of proportionality.

The procedure we propose depends upon having a credible criterion of good prediction. By credible criterion we mean a criterion that is made small when the predicted values are close to the observed values but also includes a penalty for increasing the variance of the predictors. It is not the purpose of this paper to compare criteria. However, for the purpose of illustration we will describe two credible criteria.

The total mean square error of prediction is

$$n\sigma^2 + \sum_{i=1}^{n} \text{Var}\ (\hat{Y}_i) + \sum_{i=1}^{n} (E(Y_i - \hat{Y}_i))^2. \tag{6}$$

If $\hat{Y}_i = x_i'\hat{\beta}$ where x_i' is the i-th row of X and $\hat{\beta}$ is as given by (5) with fixed D, then (6) is

$$n\sigma^2 + \text{tr}[(X'X + D)^{-1}X'X]^2\sigma^2 + \beta'X'(I - X(X'X + D)^{-1}X')^2X\beta. \tag{7}$$

It would be desirable to choose D such that (7) is minimum. However since β and σ^2 are unknown, an estimator of (7) can be used as a criterion. The quantity

$$M(D) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + 2\ \text{tr}\ (I + (X'X)^{-1}\ D)^{-1}S^2, \tag{8}$$

where S^2 is an estimator of σ^2, is an estimator of (7).

The criterion (8) can be considered an extension of Mallow's Cp Statistic [4, 7, 8] to a more general class of estimators.

Another credible criterion is the Prediction Sum of Squares (PRESS). This criterion is defined by

$$\text{PRESS}\ (D) = \sum_{i=1}^{n} (Y - \hat{Y}_{(i)})^2$$

where $\hat{Y}_{(i)}$ is the estimator of $E(Y_i)$ *excluding* the i-th observation. Explicitly, $\hat{Y}_{(i)} = (X'X - x_ix_i' + D)^{-1}(X'Y - x_iY_i)$. In other words, each observation is "predicted" using the other $n - 1$ observations. The resulting "errors of prediction" are squared and summed to form PRESS. PRESS is appealing because it simulates prediction. It does not use an observation to aid in the "prediction" of itself. PRESS as defined here is an extension of PRESS defined in [2] to a more general class of predictors.

The proposed procedure is to choose a credible criterion and then find the value of D such that the criterion is minimum. The coefficients of the Xs in the prediction equation are given by (5) evaluated at this value of D.

4. Discussion

A local minimum of the criteria presented here can be found by a standard nonlinear regression technique. We recommend that the user apply this technique several times using initial values of D corresponding to different subsets having technical appeal. Indeed, our experience has shown the existence of distinct prediction equations having nearly equal values of the criterion. A discussion of computation and a numerical example are presented in [3].

The usefulness of the proposed procedure depends upon the statistical properties of predictors associated with the criterion used. While little is known about these properties, the Cp statistic has apparently gained wide acceptance.

References

[1] Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics 13*, 469–75.

[2] Allen, D. M. (1971). The prediction sum of squares as a criterion for selecting predictor variables. Univ. of Ky. Dept. of Statistics, *Tech. Report 23*.

[3] Allen, D. M. (1972). Biased prediction using multiple linear regression. Univ. of Ky. Dept. of Statistics, *Tech. Report 36*.

[4] Gorman, J. W. and Toman, R. J. (1966). Selection of variables for fitting equations to data. *Technometrics, 8* 27–51.

[5] HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression. Biased estimation for nonorthogonal problems. *Technometrics 12*, 55–67.

[6] HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression. Applications to nonorthogonal problems. *Technometrics 12*, 69–82.

[7] MALLOWS, C. L. (1964). Choosing variables in a linear regression: A graphical aid. Presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas.

[8] MALLOWS, C. L. (1966). Choosing a subset regression. Presented at the Joint Statistical Meeting, Los Angeles, Calif.

[9] WALLS, R. C. and WEEKS, D. L. (1969). A note on the variance of a predicted response in regression. *Amer. Statistician* 23(3):24–26.