Cross-Validation of Regression Models

Author(s): Richard R. Picard and R. Dennis Cook

Source: *Journal of the American Statistical Association*, Sep., 1984, Vol. 79, No. 387 (Sep., 1984), pp. 575-583

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/2288403

Cross-Validation of Regression Models

RICHARD R. PICARD and R. DENNIS COOK*

A methodolgy for assessment of the predictive ability of regression models is presented. Attention is given to models obtained via subset selection procedures, which are extremely difficult to evaluate by standard techniques. Cross-validatory assessments of predictive ability are obtained and their use illustrated in examples.

KEY WORDS: Data splitting; Model selection; Prediction; Optimism principle.

1. INTRODUCTION

Issues of model selection arise in many statistical problems that deal with a large number of observed variables. Some of the more commonly encountered problems involve the use of multiple-regression techniques, where it is often desired to find a relatively simple function that models some underlying phenomenon. For example, many statistical-package programs (BMDP, SAS, SPSS, etc.) contain a number of subset selection algorithms, such as forward and backward stepwise methods, as one means of achieving this goal. There is a good deal of literature on the subject of developing a "best" model (Weisberg 1980 and Draper and Smith 1981 provide useful discussions), but much less has been done concerning proper interpretation of the selected fitted equation.

This article examines the assessment of the predictive ability of a fitted multiple-regression model. For illustrative purposes, fitted models obtained through the use of subset selection methods are emphasized, though it is important to recognize that the general concepts apply to many other equally important problems. The motivation for cross-validation is discussed in Section 2. Objectives of predictive assessment and the shortcomings of standard techniques in achieving them are also summarized in Section 2. Theoretical results that lead to specific cross-validatory estimators are developed in Section 3, Section 4 addresses the topic of data splitting, and examples are given in Section 5. Concluding remarks are found in Section 6.

2. MOTIVATION AND BACKGROUND

2.1 The Optimism Principle

A major tenet of conventional statistical folklore is that a model chosen via some selection process provides a much more optimistic explanation of the data used in its derivation than it does of other data that will arise in a similar fashion. One of the more eloquent statements of this principle was given by Mosteller and Tukey (1977):

> Testing the procedure on the data that gave it birth is almost certain to overestimate performance, for the optimizing process that chose it from among many possible procedures will have made the greatest use possible of any and all idiosyncrasies of those particular data. . . . As a result, the procedure will likely work better for these data than for almost any other data that will arise in practice. (p. 37)

This doctrine appears to be based on a long history of unfortunate experiences encountered by statisticians.

The most important words in the above quotation are "from among many possible procedures," since the selection process plays a key role. For subset selection procedures in multiple regression, as an example, providing a demonstration of this phenomenon is not difficult. Consider the standard linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\mathbf{X}$ is an $n \times p$ full-rank matrix of known constants, $\boldsymbol{\beta}$ is a p vector of unknown parameters, and the errors $\boldsymbol{\epsilon} = (\epsilon_i)$ are independent and identically distributed with mean 0 and variance σ^2. The model (2.1) is called the full model, and the possibility that some of the components of $\boldsymbol{\beta}$ are zero is often entertained. The purpose of subset selection algorithms is to extract a parsimonious fitted equation.

It is well known that the usual residual mean square $\hat{\sigma}_{\text{full}}^2$ from the least squares fit to the full model is an unbiased estimator of σ^2. The same does not hold for residual mean squares from fitted models chosen by subset selection procedures. Suppose, for example, that all $2^p - 1$ least-squares subset fits are considered and the subset corresponding to the smallest residual mean square $\hat{\sigma}_{\text{min}}^2$ is selected. Because $\hat{\sigma}_{\text{min}}^2 \leq \hat{\sigma}_{\text{full}}^2$ for all values of $\mathbf{Y}$ and $\hat{\sigma}_{\text{full}}^2$ has expectation σ^2, it is obvious that $\hat{\sigma}_{\text{min}}^2$ does not. Clearly, $\hat{\sigma}_{\text{min}}^2$ is optimistic and in some cases the bias is substantial. This means that a naive user of the selected model could easily be misled into believing that predictions are of higher quality than is actually the case.

Exact distributional results are virtually impossible to obtain, even for the simplest of common subset selection algorithms. Using simulation, Berk (1978a) found realistic examples in which bias in residual mean squares from

* Richard Picard is a Staff Member in the Statistics Group at the Los Alamos National Laboratory, MS-F600, P.O. Box 1663, Los Alamos, NM 87545. Dennis Cook is Professor and Chairman, Dept. of Applied Statistics, University of Minnesota, St. Paul, MN 55108. The authors thank Dick Beckman and Dave Whiteman for their helpful comments during the preparation of this manuscript.

© Journal of the American Statistical Association
September 1984, Volume 79, Number 387
Theory and Methods Section

fitted models chosen via stepwise algorithms exceeded 20% of the actual value of σ^2. Many other measures of fit are similarly biased over the subset selection; observed values of C_p and the prediction sum of squares (PRESS) tend to be much lower than might otherwise be expected (Berk 1978b), whereas values of R^2 are greatly inflated (Diehr and Hoflin 1974, Rencher and Pun 1980, Freedman 1983). Imitation of standard statistical techniques that rely on a known—as opposed to a selected—model can lead to conclusions in error. For example, substitution of $\hat{\sigma}_{\min}^2$ in the usual formulas for confidence intervals is not justified, and ensuing optimism promotes erroneous beliefs.

While the preceding discussion is posed in the context of subset selection, similar comments apply whenever the fitted model is obtained through application of techniques whose statistical behavior cannot be adequately described theoretically. Such techniques arise because totally automated selection procedures are, of course, not usually recommended. Advocates of subset selection carefully point out that it does not guarantee a best model but instead provides a useful ordering of various candidate models. A common recommendation is to use a primary criterion (e.g., C_p, Mallows 1973, or PRESS, Allen 1971) to acquire a relatively small collection of models and then employ secondary criteria in the final selection. Furthermore, subset selection is often only one part of a more general process of model development that is intended to deal with outliers, heteroscedasticity, and so on.

2.2 Characterizing Predictive Ability

Consider model (2.1) and any selection or fitting method that yields a p-vector $\hat{\boldsymbol{\beta}}$ as an estimator of $\boldsymbol{\beta}$. If the full model is not selected, some of the components of the realized value of $\hat{\boldsymbol{\beta}}$ are zero. The distribution of $\mathbf{Y}$ together with the selection algorithm thus induce a distribution on R^p for the ensuing estimator $\hat{\boldsymbol{\beta}}$. Assuming that moments of order 2 exist, denote the mean vector and covariance matrix of $\hat{\boldsymbol{\beta}}$ by $\boldsymbol{\mu}_{\hat{\beta}}$ and $\boldsymbol{\Sigma}_{\hat{\beta}}$, respectively.

The predictive ability of the associated fitted model can be characterized using the moments of $\hat{\boldsymbol{\beta}}$. Suppose that $\mathbf{x}_f$ is a known p vector and $Y_f = \mathbf{x}_f^T\boldsymbol{\beta} + \epsilon_f$ is an observation conforming to the structure of (2.1) and independent of $\hat{\boldsymbol{\beta}}$. That is, $(Y_f, \mathbf{x}_f^T)$ can be thought of as a future observation. The predicted value of Y_f based on the selected model is $\hat{Y}_f = \mathbf{x}_f^T\hat{\boldsymbol{\beta}}$ and the error of prediction is $Y_f - \hat{Y}_f$.

The predictive ability of the model is reflected by statistical properties of $Y_f - \hat{Y}_f$ for different choices of $\mathbf{x}_f$. Ideally, we might hope to obtain the distribution function of $Y_f - \hat{Y}_f$, but this is unrealistic in cases of model selection. Instead, the second moment about 0, or mean squared error (MSE), is used as a summary. Other summaries could be adopted, of course.

Two distinct MSE's are important measures of predictive ability. The first is the conditional (on $\hat{\boldsymbol{\beta}}$) MSE at $\mathbf{x}_f$,

$$\begin{aligned}\mathrm{MSE}(\mathbf{x}_f \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0) &= E_{Y_f}[(Y_f - \hat{Y}_f)^2 \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0] \\ &= \sigma^2 + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\mathbf{x}_f\mathbf{x}_f^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0), \qquad (2.2)\end{aligned}$$

and the second is the unconditional MSE,

$$\begin{aligned}\mathrm{MSE}(\mathbf{x}_f) &= E_{\hat{\beta}}\,\mathrm{MSE}(\mathbf{x}_f \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0) \\ &= \sigma^2 + (\boldsymbol{\beta} - \boldsymbol{\mu}_{\hat{\beta}})^T\mathbf{x}_f\mathbf{x}_f^T(\boldsymbol{\beta} - \boldsymbol{\mu}_{\hat{\beta}}) \\ &\quad + \mathrm{tr}\boldsymbol{\Sigma}_{\hat{\beta}}\mathbf{x}_f\mathbf{x}_f^T. \qquad (2.3)\end{aligned}$$

Overall measures of predictive ability can be obtained by integrating the MSE's with respect to a distribution F_X for $\mathbf{x}_f$. This amounts to replacing $\mathbf{x}_f\mathbf{x}_f^T$ in (2.2) and (2.3) with $\mathbf{C} = \int \mathbf{x}_f\mathbf{x}_f^T\, dF_X$ and thus obtaining the integrated mean squared errors IMSE $(\mathbf{C} \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0)$ and IMSE $(\mathbf{C})$, respectively.

The conditional MSE summarizes the predictive ability of the actual fitted equation, whereas the unconditional MSE averages this with respect to the distribution of $\hat{\boldsymbol{\beta}}$. The unconditional version is more appropriate for comparison of selection methods; in practice, however, the conditional version is more relevant to researchers, who are generally interested in the performance of the particular predictor to be used and are not as concerned about a hypothetical average. Over the distribution of $\hat{\boldsymbol{\beta}}$, large differences between the conditional MSE and its unconditional counterpart may occur. In subsequent sections, methods for estimating the conditional MSE are proposed.

2.3 Validation

In the analysis of large data sets, an appropriate way to proceed is often not immediately apparent. Consequently, some aspects of exploratory data analysis naturally arise. Competing models may be temporarily entertained and the final choice of a predictive equation is influenced by many factors, including the personal experiences and prejudices of the investigator.

It is important to develop the underpinnings of proper assessment of models produced by such selection procedures. A central point to keep in mind is that when a model is chosen because of qualities exhibited by a particular set of data, predictions of future observations that arise in a similar fashion will almost certainly not be as good as might naively be expected. Obtaining an adequate estimator of MSE requires future data and, in the extreme, model evaluation is a long-term, iterative endeavor. To expedite this process, the future can be constructed by reserving part of the present, available data.

The concept of splitting the data into two parts (not necessarily of equal size), with one part used to divine the fitted model and the other part reserved for validation, is by no means new. Historical background is provided by Stone (1974,1978) and Geisser (1975), who also present their own viewpoints on assessment of predictive ability.

Other useful discussions are given by Snee (1977) and Mosteller and Tukey (1977), who consider aspects of data splitting from a researcher's perspective.

Despite its early origins, the subject of validation has not yet been thoroughly examined. Much of the past work is illustrative in nature, emphasizing various techniques and their range of application rather than laying a theoretical foundation for a methodology that can be used in conjunction with standard methods of analysis that are strongly data analytic in nature.

A possible objection to the use of splitting is the loss of information incurred in model development. To be sure, deriving $\hat{\boldsymbol{\beta}}$ using only a portion of the data is a clear violation of the sufficiency principle (e.g., see Basu 1980, p. 576). We view this as an inevitable cost of using data-analytic methods and simultaneously requiring sound estimates of predictive ability. Furthermore, in moderate and large data sets (where splitting is most practical), this cost is typically quite small. For illustration, consider the partitioned form of model (2.1), $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$, in combination with the estimator

$$\hat{\boldsymbol{\beta}}_u = \begin{pmatrix} (\mathbf{X}_1{}^T\mathbf{X}_1)^{-1}\mathbf{X}_1{}^T\mathbf{Y} \\ \mathbf{0} \end{pmatrix},$$

obtained by underfitting the full model using ordinary least squares. The moments of $\hat{\boldsymbol{\beta}}_u$ are well known and the corresponding unconditional IMSE of prediction is

$$\text{IMSE}(\mathbf{C}) = \sigma^2 + \int [\mathbf{x}^T(\boldsymbol{\beta} - E\hat{\boldsymbol{\beta}}_u)]^2 dF_X + \int \text{var}(\mathbf{x}^T\hat{\boldsymbol{\beta}}_u) dF_X. \quad (2.4)$$

Next, consider a second model and estimator with the same structure except that each row of $\mathbf{X}_1$ and $\mathbf{X}_2$ is replicated k times so that there are kn observations. It is easily verified that the unconditional IMSE in the second case is identical to that in the first except for the final term in (2.4), which is multiplied by the factor $1/k$. The first term in (2.4), σ^2, can be thought of as error due to prediction of the future and cannot be reduced by taking more observations. The second term is the average squared bias and, when $\boldsymbol{\beta}_2 \neq 0$, represents a penalty for fitting a model of the wrong form. As illustrated, this term also does not decrease intrinsically with sample size. The third term, which reflects variability in parameter estimation, declines as $1/k$ and in many cases is dominated by the sum of the first two terms. Therefore IMSE(C) is not particularly sensitive to the magnitude of n in this case. Section 5 contains similar examples of insensitivity involving model selection. Formal demonstrations of this diminishing-returns phenomenon in more realistic settings are not easily established, but there is little reason to believe that predictive ability in such settings behaves in a fundamentally different fashion.

A second possible objection to the use of splitting concerns the stability of the validation estimate. Suppose that all of the data were used for model development and then reused for assessment. The variability of the assessment would depend on, among other things, the total number of observations. If the data were split prior to model development, however, the number of observations reserved for validation would largely determine variability in the resulting assessment. In practice, at most half of the data (and usually less) are so reserved, and estimates based on splitting have comparatively high variability. This can cause splitting to be ineffective in cases in which the optimism induced by model development is small or in which the development itself can be well characterized.

3. FORMAL DEVELOPMENTS

Consider model (2.1), again, with the additional structure $E\,\epsilon_i{}^3 = 0$ and $E\,\epsilon_i{}^4/\sigma^4 - 3 = \gamma$; and partition $\mathbf{Y}$ and $\mathbf{X}$ as $\mathbf{Y}^T = (\mathbf{Y}_E{}^T, \mathbf{Y}_V{}^T)$ and $\mathbf{X}^T = (\mathbf{E}^T, \mathbf{V}^T)$, where $\mathbf{Y}_V$ is an n_v vector and $\mathbf{V}$ is a full-rank $n_v \times p$ matrix. The derivation of an estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is based solely on the estimation data $(\mathbf{Y}_E, \mathbf{E})$, whereas the validation data $(\mathbf{Y}_V, \mathbf{V})$ are reserved for assessing the predictive ability of the fitted model.

Thus the validation data satisfy

$$\mathbf{Y}_V = \mathbf{V}\boldsymbol{\beta} + \boldsymbol{\epsilon}_v, \quad (3.1)$$

where $\boldsymbol{\epsilon}_v$ is independent of $\hat{\boldsymbol{\beta}}$. The validation residual vector, or vector of errors of prediction of $\mathbf{Y}_V$ based on $\hat{\boldsymbol{\beta}}$, is

$$\mathbf{r}_v = \mathbf{Y}_V - \mathbf{V}\hat{\boldsymbol{\beta}} = \mathbf{V}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \boldsymbol{\epsilon}_v. \quad (3.2)$$

In the squared-error framework, it is natural to develop estimators of $\text{IMSE}(\mathbf{C} \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0)$ by using quadratic forms in $\mathbf{r}_v$. Many statistics commonly used in related problems are quadratic forms in $\mathbf{Y}$ (PRESS) or functions of such forms (C_p).

For arbitrary real symmetric matrices $\mathbf{Q}$, it is easily established that

$$E(\mathbf{r}_v{}^T\mathbf{Q}\mathbf{r}_v \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0) = \sigma^2\text{tr}(\mathbf{Q}) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\mathbf{V}^T\mathbf{Q}\mathbf{V}(\boldsymbol{\beta} - \boldsymbol{\beta}_0).$$

Recall from Section 2.2 that

$$\text{IMSE}(\mathbf{C} \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0) = \sigma^2 + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\mathbf{C}(\boldsymbol{\beta} - \boldsymbol{\beta}_0).$$

Not surprisingly, expected values of quadratic forms in $\mathbf{r}_v$ strongly resemble the IMSE of prediction. It is clear that $\mathbf{r}_v{}^T\mathbf{Q}\mathbf{r}_v$ is conditionally unbiased for $\text{IMSE}(\mathbf{C} \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0)$ for all $\boldsymbol{\beta}$ and σ^2 when

$$\text{(a)}\quad \text{tr}(\mathbf{Q}) = 1 \qquad \text{and} \qquad \text{(b)}\quad \mathbf{V}^T\mathbf{Q}\mathbf{V} = \mathbf{C}. \quad (3.3)$$

To further understand the implications of (3.3), consider the situation in which $\mathbf{Q} = \text{diag}(q_1, \ldots, q_{n_v})$ and the ith row of $\mathbf{V}$ is $\mathbf{v}_i{}^T$. Constraint (3.3a) ensures that $\mathbf{r}_v{}^T\mathbf{Q}\mathbf{r}_v$ is simply a linear combination of squared validation residuals, and constraint (3.3b) ensures that the weights $\{q_i\}$ generally reflect the relative dispersion of the $\{\mathbf{v}_i\}$ within the region of predictive interest as characterized by $\mathbf{C}$:

$$\mathbf{V}^T\mathbf{Q}\mathbf{V} = \mathbf{C} \leftrightarrow \sum q_i\mathbf{v}_i\mathbf{v}_i{}^T = \int \mathbf{x}_f\mathbf{x}_f{}^T\, dF_X.$$

When $\mathbf{Q}$ is not diagonal, the same comments apply to the rotated validation residual vector $\mathbf{\Gamma r}_v$, where the $n_v \times n_v$ orthogonal matrix $\mathbf{\Gamma}$ is obtained from the spectral decomposition of $\mathbf{Q}$, $\mathbf{\Gamma}^T\mathbf{Q}\mathbf{\Gamma} = \text{diag}(d_1, \ldots, d_{n_v})$. Because $\text{IMSE}(\mathbf{C} \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0)$ depends on $\mathbf{C}$, it is important that the assessment of prediction be similarly dependent. Unbiased quadratic forms in $\mathbf{r}_v$ allow for this through the weighting of validation residuals.

Let U denote the class of $n_v \times n_v$ matrices that satisfy (3.3). The matrix $\mathbf{Q}^*$ that minimizes $\text{var}(\mathbf{r}_v^T\mathbf{Q}\mathbf{r}_v \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0)$ was derived by Picard (1981). The form of $\mathbf{Q}^*$ is not relevant to this discussion because its dependence on the unknown parameters $\boldsymbol{\beta}$, σ^2, and γ diminishes its immediate usefulness in practice. Two important cases (expressed here as theorems), however, merit special attention:

Theorem 1. If $\mathbf{C} = \mathbf{V}^T\mathbf{V}/n_v$, then $\text{var}(\mathbf{r}_v^T\mathbf{Q}\mathbf{r}_v \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0)$ is minimized over $\mathbf{Q} \in U$ by the choice $\mathbf{Q} = n_v^{-1}\mathbf{I}$.

Theorem 2. If $\gamma = 0$, then $\text{var}(\mathbf{r}_v^T\mathbf{Q}\mathbf{r}_v \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0)$ is minimized over $\mathbf{Q} \in U$ by the choice

$$\mathbf{Q}_0(\mathbf{C}) = \mathbf{V}(\mathbf{V}^T\mathbf{V})^{-1}\,\mathbf{C}(\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T + k(\mathbf{I} - \mathbf{V}(\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T), \quad (3.4)$$

where

$$k = (1 - \text{tr}(\mathbf{V}^T\mathbf{V})^{-1}\mathbf{C})/(n_v - p).$$

The proofs of these theorems are given in the Appendix.

Theorem 1 reflects the case in which $F_X = \hat{F}_V$, the empirical distribution function of $\{\mathbf{v}_i\}$. When $\hat{F}_V$ accurately represents the region of predictive interest, all validation residuals are weighted equally. Theorem 2 provides a means of assessing the predictive ability of the fitted model when the errors are approximately normal ($\gamma = 0$).

A few additional properties of the estimators associated with Theorems 1 and 2 are noteworthy. First, if F_X is a mixture of distribution functions $F_X = \sum \alpha_i F_i$, the corresponding estimator $\mathbf{r}_v^T\mathbf{Q}_0(\mathbf{C})\mathbf{r}_v$ can be written as a convex combination of estimators based on the $\{F_i\}$. Letting $\mathbf{C}_i = \int \mathbf{x}_f\mathbf{x}_f^T\, dF_i$, it follows that

$$\mathbf{C} = \sum \alpha_i \mathbf{C}_i,$$

$$\text{IMSE}(\mathbf{C} \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0) = \sum \alpha_i\, \text{IMSE}(\mathbf{C}_i \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0),$$

and

$$\mathbf{r}_v^T\mathbf{Q}_0(\mathbf{C})\mathbf{r}_v = \sum \alpha_i \mathbf{r}_v^T\mathbf{Q}_0(\mathbf{C}_i)\mathbf{r}_v.$$

It is also easily verified that $\mathbf{r}_v^T\mathbf{Q}_0(\mathbf{C})\mathbf{r}_v$ is invariant with respect to nonsingular, linear reparameterizations.

4. SPLITTING

Until now, little has been said concerning data splitting, which is the method of acquiring observations for validation. An ostensibly well-motivated procedure is to split the data at random. If the original $(\mathbf{Y}, \mathbf{X})$ can be viewed as a sample randomly drawn from a population of interest, a randomly selected validation set should mimic a sample of future observations. Such a split is easy to implement; in fact, many statistical-package routines provide users with the means and motivation (e.g., Dixon 1981, p. 707) for pursuing this approach. Although analyses based on random splits are, on the whole, superior to those based on no splitting at all, there is room for improvement.

In general, there are two important considerations regarding splitting. The first is that the integrity of the validation data be ensured to allow those data to imitate future observations. It is imperative that independence of the validation data from $\hat{\boldsymbol{\beta}}$ be achieved; for practical purposes, this consideration mandates that the response vector $\mathbf{Y}$ be ignored by the splitting mechanism.

The second major consideration involves localized assessments—that is, assessments that weight residuals whose $\mathbf{v}_i$ are local to the region of predictive interest more heavily than they weight other residuals. Were validation data unavailable from a particular region, a localized evaluation of the fitted model in that region would be impossible to obtain. Thus it is important to examine the $\{\mathbf{x}_i\}$ that constitute $\mathbf{X}$ and avoid such an occurrence. Random splits often fail to comprehensively cover the factor space, as gaps can occur with high probability when the number of carriers is large. Even under more agreeable circumstances, a poor split may still arise, albeit with lower probability.

A formal algorithm for data splitting, DUPLEX, was proposed by Snee (1977). Roughly, this contructs a distance metric on R^p that is used sequentially to select validation data. The basic concept is sound, and the splits provided generally allocate observations whose $\{\mathbf{v}_i\}$ are spread over the factor space.

An alternative to DUPLEX stems from a goal that the empirical distribution of the $\{\mathbf{v}_i\}$ be similar to that of the $\{\mathbf{x}_i\}$. One implementation of this idea is to extract $\mathbf{V}$ by matching moments of order 2 and to require the approximate equality

$$\mathbf{V}^T\mathbf{V}/n_v \simeq \mathbf{X}^T\mathbf{X}/n$$

to hold. Such matched splits are quite useful in cross-validatory data analyses.

Splits that are (nearly) matched allow for the development of predictors with good bias properties more easily than other splits do. In any regression problem of practical consequence, $\Pr\{\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}\} = 0$. Therefore, conditional on the realized value of $\hat{\boldsymbol{\beta}}$, prediction is invariably biased.

Conditional squared bias represents a substantial portion of the conditional MSE of prediction. In choosing a split of the data prior to analysis, any researcher would be interested in knowing which choice would lead to selection of the model that minimized this bias. Such a determination, however, depends on the realized values of the response variable used to derive the fitted model. For reasons described earlier, these values must be ignored by the splitting mechanism, and thus the relative merits of different splits can be fairly compared only in an unconditional sense.

It can be shown, following principles that foster the construction of minimum-bias experimental designs (Box and Draper 1959), that in the least-squares framework matched splits have excellent unconditional bias properties. Partition the model for the estimation data as

$$\mathbf{Y}_E = \mathbf{E}\boldsymbol{\beta} + \boldsymbol{\epsilon}_E = \mathbf{E}_1\boldsymbol{\beta}_1 + \mathbf{E}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_E,$$

and consider the least-squares subset fit corresponding to the estimator

$$\hat{\boldsymbol{\beta}}_E = \begin{pmatrix} (\mathbf{E}_1{}^T\mathbf{E}_1)^{-1}\mathbf{E}_1{}^T\mathbf{Y}_E \\ \mathbf{0} \end{pmatrix}.$$

Summing the conditional squared bias of prediction of observations in a future data set whose $\{\mathbf{x}_i\}$ are the same as those at hand and taking the expectation gives

$$E\left(\sum_{i=1}^{n} (\mathbf{x}_i{}^T\boldsymbol{\beta} - \mathbf{x}_i{}^T\hat{\boldsymbol{\beta}}_E)^2\right)$$
$$= \boldsymbol{\beta}_2{}^T(\mathbf{A}_E - \mathbf{A}_X)^T\mathbf{X}_1{}^T\mathbf{X}_1(\mathbf{A}_E - \mathbf{A}_X)\boldsymbol{\beta}_2$$
$$+ \sigma^2\mathrm{tr}(\mathbf{X}_1{}^T\mathbf{X}_1)(\mathbf{E}_1{}^T\mathbf{E}_1)^{-1}$$
$$+ \boldsymbol{\beta}_2{}^T\mathbf{X}_2{}^T(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1{}^T\mathbf{X}_1)^{-1}\mathbf{X}_1{}^T)\mathbf{X}_2\boldsymbol{\beta}_2, \quad (4.1)$$

where

$$\mathbf{A}_E = (\mathbf{E}_1{}^T\mathbf{E}_1)^{-1}\mathbf{E}_1{}^T\mathbf{E}_2$$

and

$$\mathbf{A}_X = (\mathbf{X}_1{}^T\mathbf{X}_1)^{-1}\mathbf{X}_1{}^T\mathbf{X}_2.$$

The dependence of (4.1) on the unknown parameters $\boldsymbol{\beta}_2$ and σ^2 makes it difficult to determine which split of the data leads to the fitted model with the best overall unconditional bias properties. Since the first term in (4.1), however, typically dominates the second (whereas the third depends only on $\mathbf{X}$ and not $\mathbf{E}$), splits for which $\mathbf{A}_E \simeq \mathbf{A}_X$ are excellent choices. Matched splits satisfy this condition.

Beyond the good overall prediction afforded by matched splits, good overall assessment is achieved as well. The validation mean square $\mathbf{r}_v{}^T\mathbf{r}_v/n_v$ is optimal (in the sense of Theorem 1) for evaluation of IMSE $(\mathbf{X}^T\mathbf{X}/n \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0)$. For splits that are poorly matched, the validation mean square does not properly weight residuals in this regard. Hence assessment of IMSE$(\mathbf{X}^T\mathbf{X}/n \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0)$ can be pursued only with fewer effective degrees of freedom. In this way, matched splits strike a balance between prediction and validation.

5. EXAMPLES

5.1 Localized Assessment

The data used in this example were collected by the Canadian Wildlife Service (Cook and Jacobson 1978) for purposes of estimating the snow-goose population in a region near West Hudson Bay. An aerial survey was conducted in which the numbers of adult geese in selected flocks were estimated by a member of the wildlife service. In addition, the flocks were photographed and the pictures later inspected to determine the true numbers of adults. Results for 20 such flocks are listed in Table 1 and displayed in Figure 1.

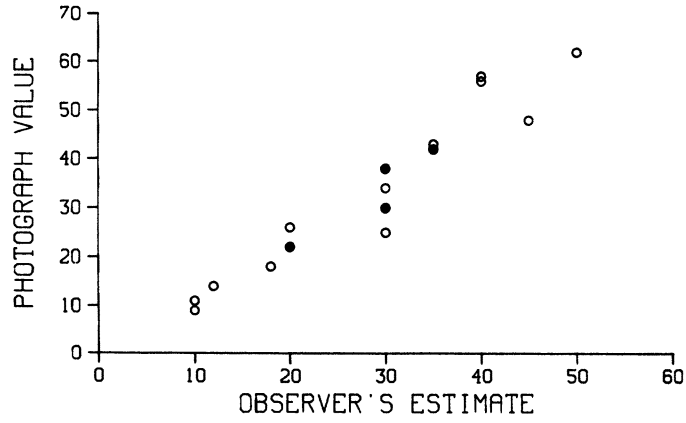

Figure 1. Scatterplot of the Data for the Example of Section 5.1. The four solid circles denote replicated observations.

Of interest is the development of a predictor of the true flock sizes $\{y_i\}$, based on the estimates $\{x_i\}$ given by the observer from the wildlife service. In the analysis of small data sets such as this one, the researcher considering use of cross-validation has something of a dilemma. If the data are split, both the predictor and its validation may be based on relatively small amounts of information. The drawbacks of such an approach are obvious. Alternatively, if the data are not split, then the assessment is subject to the optimism principle. As described in Section 2, the optimistic consequences tend to be severe when model selection is pursued and n is small.

To illustrate some of the issues involved, consider splitting the data in Table 1 based on an assumed simple linear-regression model. The design matrix $\mathbf{X}$ is 20×2. Arbitrarily choosing $n_v = 6$, it is possible to extract $\mathbf{V}$ such that

$$\mathbf{X}^T\mathbf{X}/20 = \begin{pmatrix} 1.0 & 28.5 \\ 28.5 & 933.4 \end{pmatrix}$$

and

$$\mathbf{V}^T\mathbf{V}/6 = \begin{pmatrix} 1.0 & 28.7 \\ 28.7 & 932.3 \end{pmatrix}$$

The class of such matched splits of the data has several members because there is more than one observation at

Table 1. Data Used in the Example of Section 5.1

Observer's Estimates $\{x_i\}$	Photograph Values $\{y_i\}$
10	9, 11
12	14
18	18
20	22, 22, 26
30	25, 30, 30, 34, 38, 38
35	42, 42, 43
40	56, 57
45	48
50	62

many values of x_i. Within this class, no split is preferred a priori over any other, and a random selection determines the validation data to be

$$\{(x_i, y_i) \mid (x_i, y_i) = (12, 14), (20, 22), (30, 30), (30, 38), (35, 43), (45, 48)\}.$$

Upon analysis of the estimation data, the presence of heteroscedasticity becomes apparent (larger flocks tend to be less accurately predicted than smaller ones) and a weighted least-squares approach is pursued. The decision of how to weight, a type of model selection problem, is often subjective and is usually made after examination of various diagnostic plots.

Suppose that a weighted least-squares predictor is acquired after optimally weighting as if $\text{var}(y_i) = x_i\sigma^2$. Evaluation of the fitted model's predictive ability in different regions of the factor space should be a part of any analysis, especially when the researcher is interested in how well the "next" observation, $Y_f = \mathbf{x}_f^T\boldsymbol{\beta} + \epsilon_f$, will be predicted but does not know what value $\mathbf{x}_f$ will take. Consider the cases in which the observer from the wildlife service estimates that two flocks contain 16 and 40 adult geese, respectively. Predicted values of the true flock sizes are easily obtained from the fitted model, but how good are these predictions?

Following the methodology presented earlier, the expected cross-product matrices

$$\mathbf{C}_{16} = \begin{pmatrix} 1 & 16 \\ 16 & 256 \end{pmatrix} \quad \text{and} \quad \mathbf{C}_{40} = \begin{pmatrix} 1 & 40 \\ 40 & 1600 \end{pmatrix}$$

are formed and substituted into the general expression (3.4) for $\mathbf{Q}_0(\mathbf{C})$. The resulting matrices are given in Table 2.

Inspection of $\mathbf{Q}_0(\mathbf{C}_{16})$ and $\mathbf{Q}_0(\mathbf{C}_{40})$ reveals the localized nature of assessment described in Section 3. For example, the diagonal elements of $\mathbf{Q}_0(\mathbf{C}_{16})$ corresponding to the validation residuals at $x_i = 12$ and 20 are much larger than those corresponding to validation residuals at $x_i = 35$ and 45, whereas the reverse is true for $\mathbf{Q}_0(\mathbf{C}_{40})$.

Table 2. Matrices of the Quadratic Forms in the Example of Section 5.1

	x_i					
	12	*20*	*30*	*30*	*35*	*45*
$Q_0(C_{16}) =$	.30	.10	.05	.05	.02	−.03
		.22	.02	.02	.00	−.04
			.14	.00	−.02	−.05
				.14	−.02	−.05
					.11	−.05
						.09
$Q_0(C_{40}) =$	.08	−.06	−.04	−.04	−.03	−.01
		.11	−.02	−.02	−.01	.02
			.17	.01	.02	.05
				.17	.02	.05
					.20	.07
						.27

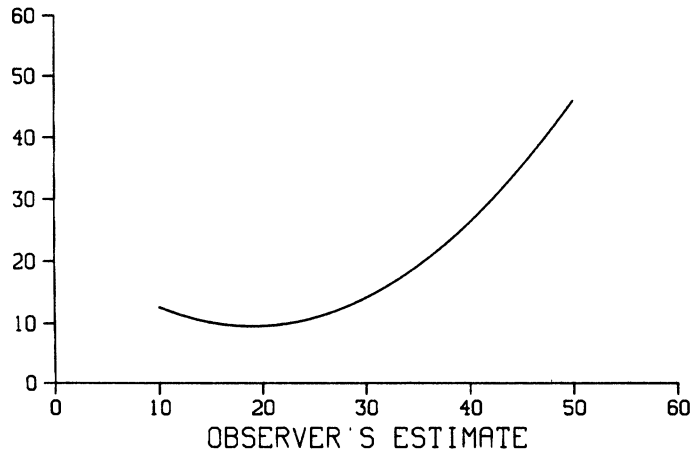

Figure 2. Estimated Mean Squared Error as a Function of the Observer's Estimate.

The resulting quadratic forms are

$$\mathbf{r}_v^T\mathbf{Q}_0(\mathbf{C}_{16})\mathbf{r}_v = 9.9$$

and

$$\mathbf{r}_v^T\mathbf{Q}_0(\mathbf{C}_{40})\mathbf{r}_v = 26.3.$$

Thus the estimated MSE of prediction for the case in which the observer claims 40 geese are present is more than twice as large as when he claims the flock contains 16 geese. More generally, Figure 2 illustrates the change in estimated MSE as a function of the observer's value.

The development of the methodology of the previous sections was pursued under the standard assumption of homoscedasticity. As indicated in the example, however, good performance can also be expected under heteroscedastic conditions. Consider the general case in which independent observations satisfy $Y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_i$, where $\epsilon_i \sim [0, \sigma^2 w(\mathbf{x}_i)]$ and $w(\cdot)$ is a function that describes the nature of the heteroscedasticity present. After splitting the data, analysis produces a fitted model $\hat{Y} = \mathbf{x}^T\boldsymbol{\beta}_0$. For a given expected cross-product matrix $\mathbf{C} = \int \mathbf{x}\mathbf{x}^T\, dF_X$, the conditional MSE of prediction of future observations is

$$\text{IMSE}(\mathbf{C} \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0) = \sigma^2 E_{F_X}[w(\mathbf{x})] + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\mathbf{C}(\boldsymbol{\beta} - \boldsymbol{\beta}_0). \quad (5.1)$$

This generalizes (2.2), where $w(\mathbf{x}) \equiv 1$ was assumed.

The validation data satisfy $\mathbf{Y}_V = \mathbf{V}\boldsymbol{\beta} + \boldsymbol{\epsilon}_v$, where $\boldsymbol{\epsilon}_v$ has the indicated heteroscedastic structure. Letting $\mathbf{v}_i^T$ denote the ith row of $\mathbf{V}$, it can be shown that

$$E(\mathbf{r}_v^T\mathbf{Q}_0(\mathbf{C})\mathbf{r}_v \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0) = \sigma^2 \sum_i [\mathbf{Q}_0(\mathbf{C})]_{ii} w(\mathbf{v}_i) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\mathbf{C}(\boldsymbol{\beta} - \boldsymbol{\beta}_0). \quad (5.2)$$

Recall that $\sum \mathbf{Q}_0(\mathbf{C})_{ii} = 1$. Under comparison of (5.2) to (5.1), it is clear that good assessment requires the averages

$$E_{F_X}(w(\mathbf{x})) \quad \text{and} \quad \sum [\mathbf{Q}_0(\mathbf{C})]_{ii} w(\mathbf{v}_i)$$

to be close. This phenomenon was illustrated in the analysis of the geese data.

When the number of observations reserved for validation, n_v, is large, the theoretical optimality of Theorem 2 (Section 3) comes at some cost in localized assessment. If failures in the usual distributional assumptions occur, as in the preceding example, a localized approach to validation is important. Striking the proper balance is not a simple matter, since assessments that are too local lack stability because they are essentially based on too few degrees of freedom, whereas those too global are insensitive to local changes in predictive ability. When n_v is large, a useful procedure is to select a subset of validation residuals whose $\{\mathbf{v}_i\}$ are sufficiently near the region of predictive interest and assess the fitted model via unbiased quadratic forms based on those residuals alone.

5.2 Selection Bias

In the previous example, relatively little model selection was involved (only the subjective choice of weights for the weighted least squares). Consider now an example in which the selection process is well defined but entails more serious selection bias. Let the observed data satisfy $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is 64×16 and is obtained by concatenating four replicates of the design matrix for a 2^4 factorial—that is, each element of $\mathbf{X}$ is either $+1$ or -1—and the components of the error vector $\boldsymbol{\epsilon}$ are iid standard normal. Two values for the parameter vector $\boldsymbol{\beta}$ are considered: $\boldsymbol{\beta}_1{}^T = (0, 0, \ldots, 0)$ and $\boldsymbol{\beta}_2{}^T = (1.766, 1.285, \ldots, -1.766)$. Here the values of the elements of $\boldsymbol{\beta}_2$ are the expected values of order statistics from a random sample of size 16 from a $N(0, 1)$.

Data simulated from the preceding are input to the BMDP9R program (Dixon 1981), and a fitted model is selected using the C_p option. For $\boldsymbol{\beta} = \boldsymbol{\beta}_1 = \mathbf{0}$ the selected model overfits the data, whereas for $\boldsymbol{\beta} = \boldsymbol{\beta}_2$ it typically underfits. In each instance, two passes are made. On the first pass, all 64 observations are part of a full analysis and are used to develop a fitted model. The value of PRESS (Allen 1971) is recorded from BMDP9R output, and the overall predictive ability $\text{IMSE}(\mathbf{X}^T\mathbf{X}/64 \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0)$ is also computed. On the second pass, 32 observations (2 of the 4 replicates of the 2^4 factorial) are used to develop the fitted model, and the remaining 32 observations from this split analysis are used for validation. The values of the validation mean square $\mathbf{r}_v{}^T\mathbf{r}_v/32$ and of the overall conditional predictive ability $\text{IMSE}(\mathbf{X}^T\mathbf{X}/64 \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0{}^*)$ of the fitted model are obtained. A summary of results from 200 simulated data sets is given in Table 3.

The PRESS statistic is often advocated for use in assessing overall predictive ability of the fitted model. Although the value of PRESS output by BMDP9R may not be identical to the true PRESS, which would be obtained by repeating the entire model selection process 64 times, the output value and true value are likely very close in this example. Moreover, Efron (1983, p. 327, Remark F) has stated that the true PRESS is a reasonable estimate of overall predictive ability and behaves similarly to the bootstrap.

Table 3. Simulation Results From 200 Trials

$\boldsymbol{\beta}$	Mean	Mean	Variance	MSE
Full Analysis—$n = 64$				
	$\text{IMSE}(\mathbf{X}^T\mathbf{X}/n \mid \boldsymbol{\beta} = \boldsymbol{\beta}_0)$	$\text{PRESS-IMSE}(\mathbf{X}^T\mathbf{X}/n \mid \boldsymbol{\beta} = \boldsymbol{\beta}_0)$		
zero	1.16	−.217	.042	.089
rankits	1.27	−.031	.067	.067
Split Analysis—$n_E = n_V = 32$				
	$\text{IMSE}(\mathbf{X}^T\mathbf{X}/n \mid \boldsymbol{\beta} = \boldsymbol{\beta}_0{}^*)$	$\mathbf{r}_v{}^T\mathbf{r}_v/n_v - \text{IMSE}(\mathbf{X}^T\mathbf{X}/n \mid \boldsymbol{\beta} = \boldsymbol{\beta}_0{}^*)$		
zero	1.32	−.003	.105	.105
rankits	1.56	.001	.139	.139

As can be seen from Table 3, the optimistic bias in PRESS output from BMDP9R is substantial for the overfit case, amounting to 19% of the actual predictive ability despite the seemingly large sample size. For the splitting case, there is no bias in $\mathbf{r}_v{}^T\mathbf{r}_v/32$, although variability is large relative to PRESS because fewer data are involved. The trade-off of bias versus variability slightly favors PRESS here, and the mean squared error of assessment is .089 as opposed to .105 for splitting.

For the second case, $\boldsymbol{\beta} = \boldsymbol{\beta}_2$, the parameter vector is large and there is little selection taking place. For both the full and split analyses, the fitted models contain 13 of the 16 variables, on the average. Here PRESS has very small bias and is much superior to splitting.

The setting of the simulation involved a design matrix with orthogonal columns and an explicit, formal recipe for the selected model. In this setting, either PRESS or some form of bootstrapping may be preferred to data splitting, although the advantage can be slight when the optimism or "downward bias" (Efron 1983, p. 320) is serious. In settings in which an explicit recipe for $\hat{\boldsymbol{\beta}}$ does not exist, splitting may be the only viable option.

6. CONCLUDING REMARKS

The preceding sections have described the usefulness of cross-validation in a data-analytic context. In the classical linear models framework, data splitting is admittedly of less value. Perhaps this diminished value is the main reason behind the deficient motivation in much of the previous cross-validatory literature. When distributional properties of an estimator $\hat{\boldsymbol{\beta}}$ can be well characterized and it is of interest to assess the unconditional predictive ability of the associated fitted model, the application of standard statistical principles will, of course, lead to methods superior to those discussed here. When the theory is intractable but bootstrapping or similar methodology can be pursued as a secondary measure, results may again be superior. This has apparently trapped many proponents of data splitting in a catch-22 situation: without formal comparisons to standard techniques, there is little reason to expect splitting to be widely used in prac-

tice; on the other hand, the tight theoretical structure needed for rigorous comparison must define a methodology superior to splitting in that situation (i.e., if the theory holds, use the theory to its fullest and accomplish the objective). In either case, little motivation is afforded.

It is important to recognize that the classical linear model framework is often not appropriate in practice. Formal recipes for $\hat{\boldsymbol{\beta}}$ cannot be written down in many cases, precluding proper application of alternative methods of assessment. Once the data analytic roles of model selection, diagnostic methods, and so forth are considered, splitting becomes useful.

In isolated instances, it may not be possible to obtain a good assessment of the model. One difficulty arises when all sources of variability are not present in the observed data. For example, in an experiment in which substantial day-to-day variation exists but observations are collected from only one day, the actual variability would likely be underestimated from the information at hand. It has been noted that cross-validation is ineffective in such cases and "is often weaker, by an unknown amount, than it appears to be" (Mosteller and Tukey 1977, p. 39). In fairness, however, no other analytic tool is available to estimate variation that is neither exhibited in the data nor otherwise apparent to the experimenter. The potential weakness in assessment should be recognized, but this should not be construed as a flaw in the methodology.

It has often been suggested that cross-validatory concepts can be used in the model-selection process. Most of these suggestions apply the "leave out one observation at a time" approach that is at the heart of the PRESS statistic. The PRESS criterion was a precursor of Stone's (1974, p. 121) "cross-validatory choice," and other authors (e.g., Geisser 1975; Wahba 1977; Golub, Heath, and Wahba 1979; Eastment and Krzanowski 1982; and Chow, Geman, and Wu 1983) have pursued parallel philosophies regarding other problems. In a more data-analytic context, the "leave out one" approach has proved valuable in the field of regression diagnostics. As a case in point, Cook's (1977) distance measure employed this concept.

Data splitting can be similarly employed. Snee (1977) argued that examination of competing regression models with respect to how well they predict validation data can aid in model selection. Brown (1982) argued similarly concerning calibration problems. In the regression framework, the quadratic form $\mathbf{r}_v^T\mathbf{Q}_0(\mathbf{X}^T\mathbf{X}/n)\mathbf{r}_v$ acts as the cross-validation version of the C_p statistic when $\hat{\boldsymbol{\beta}}$ is the ordinary least-squares estimator corresponding to the fit of a subset model (Picard 1981).

Implementation of cross-validation in the derivation of $\hat{\boldsymbol{\beta}}$ does not, however, alter the fundamentals of predictive assessment. If a proper evaluation of a selected fitted model is to be realized, the optimism principle cannot be ignored. When using data splitting, this implies that all aspects of model selection (even those that are cross-validatory) should be confined to analysis of the estimation data and that the validation data be reserved solely for assessment.

APPENDIX

The development of the theoretical results of Section 3 is discussed here. To establish Theorem 1, first note that $n_v^{-1}\mathbf{I}$ satisfies the unbiasedness constraints (3.3) when $\mathbf{C} = \mathbf{V}^T\mathbf{V}/n_v$. Consider any alternative solution in U, say

$$\tilde{\mathbf{Q}} = n_v^{-1}\mathbf{I} + \mathbf{R} \quad \text{where} \quad \mathbf{R} \neq \mathbf{0}.$$

Since $\tilde{\mathbf{Q}} \in U$,

$$\mathbf{R} = \mathbf{R}^T, \ \operatorname{tr}\mathbf{R} = 0, \quad \text{and} \quad \mathbf{V}^T\mathbf{R}\mathbf{V} = \mathbf{0}. \tag{A.1}$$

Under the assumptions of the model, it follows from standard variance formulas that

$$\operatorname{var}(\mathbf{r}_v^T\tilde{\mathbf{Q}}\mathbf{r}_v \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0) = 2\sigma^4\left\{\operatorname{tr}\tilde{\mathbf{Q}}^2 + \frac{\gamma}{2}\sum_i (\tilde{\mathbf{Q}}_{ii})^2\right\} + 4\sigma^2(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\mathbf{V}\tilde{\mathbf{Q}}^2\mathbf{V}(\boldsymbol{\beta} - \boldsymbol{\beta}_0).$$

Letting $\mathbf{B}_0 = \mathbf{V}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\mathbf{V}^T/\sigma^2$ and expanding $\tilde{\mathbf{Q}}^2 = (n_v^{-1}\mathbf{I} + \mathbf{R})^2$ gives

$$\begin{aligned}
&\operatorname{var}(\mathbf{r}_v^T\tilde{\mathbf{Q}}\mathbf{r}_v \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0) \\
&= 2\sigma^4\left\{\operatorname{tr}(n_v^{-1}\mathbf{I})^2 + \frac{\gamma}{2}\sum_i n_v^{-2} + 2\operatorname{tr}(n_v^{-1}\mathbf{I})^2\mathbf{B}_0\right\} \\
&\quad + 2\sigma^4\left\{\operatorname{tr}\mathbf{R}^2 + \frac{\gamma}{2}\sum_i (\mathbf{R}_{ii})^2 + 2\operatorname{tr}\mathbf{R}^2\mathbf{B}_0\right\} \\
&\quad + 2\sigma^4\{2n_v^{-1}\operatorname{tr}\mathbf{R} + \gamma n_v^{-1}\sum_i \mathbf{R}_{ii} + 4n_v^{-1}\operatorname{tr}\mathbf{R}\mathbf{B}_0\} \\
&= \operatorname{var}(\mathbf{r}_v^T\mathbf{r}_v/n_v \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0) \\
&\quad + 2\sigma^4\left\{\operatorname{tr}\mathbf{R}^T\mathbf{R} + \frac{\gamma}{2}\sum_i (\mathbf{R}_{ii})^2 + 2[\mathbf{R}\mathbf{V}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)/\sigma]^T[\mathbf{R}\mathbf{V}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)/\sigma]\right\} \\
&\quad + 2\sigma^4\{n_v^{-1}(2 + \gamma)\operatorname{tr}\mathbf{R} + 4n_v^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\mathbf{V}^T\mathbf{R}\mathbf{V}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)/\sigma^2\}.
\end{aligned} \tag{A.2}$$

For all error distributions such that $\operatorname{var}(\epsilon_i^2) > 0$, the kurtosis satisfies $\gamma > -2$. Thus the second term of (A.2) is strictly positive for all nonzero $\mathbf{R}$. The third term vanishes as a consequence of (A.1), and the conditional variance is, therefore, minimized by $n_v^{-1}\mathbf{I}$.

Theorem 2 is proved similarly. Note $\mathbf{Q}_0(\mathbf{C}) \in U$ and again consider an alternative solution $\tilde{\mathbf{Q}} = \mathbf{Q}_0(\mathbf{C}) + \mathbf{R}$, where $\mathbf{R}$ is nonzero and conforms to (A.1). For the case $\gamma = 0$, following the preceding gives

$$\begin{aligned}
&\operatorname{var}(\mathbf{r}_v^T\tilde{\mathbf{Q}}\mathbf{r}_v \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0) \\
&= 2\sigma^4\{\operatorname{tr}(\mathbf{Q}_0(\mathbf{C}) + \mathbf{R})^2 + 2\operatorname{tr}(\mathbf{Q}_0(\mathbf{C}) + \mathbf{R})^2\mathbf{B}_0\} \\
&= \operatorname{var}(\mathbf{r}_v^T\mathbf{Q}_0(\mathbf{C})\mathbf{r}_v \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0) \\
&\quad + 2\sigma^4\{\operatorname{tr}\mathbf{R}^T\mathbf{R} + 2[\mathbf{R}\mathbf{V}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)/\sigma]^T[\mathbf{R}\mathbf{V}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)/\sigma]\} \\
&\quad + 2\sigma^4\{2\operatorname{tr}\mathbf{R}\mathbf{Q}_0(\mathbf{C}) + 2\operatorname{tr}\mathbf{R}\mathbf{Q}_0(\mathbf{C})\mathbf{B}_0 + 2\operatorname{tr}\mathbf{Q}_0(\mathbf{C})\mathbf{R}\mathbf{B}_0\}.
\end{aligned} \tag{A.3}$$

The second term of (A.3) is strictly positive for nonzero $\mathbf{R}$, the third term vanishes as $\mathbf{R}$ satisfies (A.1), and the proof is concluded.

The above theorems are special cases of more general results that may be found in Picard (1981). For arbitrary γ and $\mathbf{C}$ the matrix $\mathbf{Q}^* \in U$ minimizing $\text{var}(\mathbf{r}_v{}^T\mathbf{Q}\mathbf{r}_v \mid \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0)$ is difficult to derive (knowledge of vec operators (Henderson and Searle 1979) and Kronecker products (Rao and Mitra 1971) is quite useful in this regard) and is of limited practical value bacause of its dependence on unknown parameters.

[*Received March 1983. Revised April 1984.*]

REFERENCES

ALLEN, D.M. (1971), "The Prediction Sum of Squares as a Criterion for Selecting Predictor Variables," Technical Report 23, University of Kentucky, Dept. of Statistics.

BASU, D. (1980), "Randomization Analysis of Experimental Data: The Fisher Randomization Test," *Journal of the American Statistical Association*, 75, 575–582.

BERK, K.N. (1978a), "Comparing Subset Selection Procedures," *Technometrics*, 20, 1–6.

——— (1978b), "Sequential PRESS, Forward Selection, and the Full Regression Model," *Proceedings of the Statistical Computing Section, American Statistical Association*, 309–313.

BOX, G.E.P., and DRAPER, N.R. (1959), "A Basis for the Selection of a Response Surface Design," *Journal of the American Statistical Association*, 54, 622–654.

BROWN, P.J. (1982), "Multivariate Calibration," *Journal of the Royal Statistical Society*, Ser. B, 44, 287–308.

CHOW, Y.S., GEMAN, S., and WU, L.D. (1983), "Consistent Cross-Validated Density Estimation," *Annals of Statistics*, 11, 25–38.

COOK, R.D. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics*, 19, 15–18.

COOK, R.D., and JACOBSON, J.O. (1978), "Analysis of 1977 West Hudson Bay Snow Goose Surveys," unpublished report, Canadian Wildlife Service.

DIEHR, G., and HOFLIN, D.R. (1974), "Approximating the Distribution of the Sample R^2 in Best Subset Regressions," *Technometrics*, 16, 317–320.

DIXON, W.J., ed. (1981), *BMDP Statistical Software*, Berkeley, Calif.: University of California Press.

DRAPER, N.R., and SMITH, H. (1981), *Applied Regression Analysis*, New York: John Wiley.

EASTMENT, H.T., and KRZANOWSKI, W.J. (1982), "Cross-Validatory Choice of the Number of Components From a Principal Component Analysis," *Technometrics*, 24, 73–77.

EFRON, B. (1983), "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation," *Journal of the American Statistical Association*, 78, 316–331.

FREEDMAN, D.A. (1983), "A Note on Screening Regression Equations," *The American Statistician*, 37, 152–155.

GEISSER, S. (1975), "The Predictive Sample Reuse Method With Applications," *Journal of the American Statistical Association*, 70, 320–328.

GOLUB, G.H., HEATH, M., and WAHBA, G. (1979), "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter," *Technometrics*, 21, 215–223.

HENDERSON, H.V., and SEARLE, S.R. (1979), "Vec and Vech Operators for Matrices, With Some Uses in Jacobians and Multivariate Statistics," *Canadian Journal of Statistics*, 7, 65–81.

MALLOWS, C.L. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–675.

MOSTELLER, F., and TUKEY, J.W. (1977), *Data Analysis and Regression*, Reading, Mass.: Addison-Wesley.

PICARD, R.R. (1981), "On Assessment of the Predictive Ability of Linear Regression Models," Ph.D. dissertation, University of Minnesota, Dept. of Applied Statistics.

RAO, C.R., and MITRA, S.K. (1971), *Generalized Inverse of Matrices and Its Applications*, New York: John Wiley.

RENCHER, A.C., and PUN, F.C. (1980), "Inflation of R^2 in Best Subset Regression," *Technometrics*, 22, 49–53.

SNEE, R.D. (1977), "Validation of Regression Models: Methods and Examples," *Technometrics*, 19, 415–428.

STONE, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society*, Ser. B, 36, 111–133.

——— (1978), "Cross-Validation: A Review," *Mathematische Operationsforschung und Statistik*, 9, 127–139.

WAHBA, G. (1977), "A Survey of Some Smoothing Problems and the Method of Generalized Cross-Validation for Solving Them," *Applications of Statistics*, ed. P.R. Krishnaiah, New York: North-Holland.

WEISBERG, S. (1980), *Applied Linear Regression*, New York: John Wiley.