Variable Selection in Regression Via Repeated Data Splitting

Peter F. THALL, Kathy E. RUSSELL, and Richard M. SIMON

A new algorithm—backward elimination via repeated data splitting (BERDS)—is proposed for variable selection in regression. Initially, the data are partitioned into two sets $\{E, V\}$, and an exhaustive backward elimination (BE) is performed in E. For each p value cutoff α used in BE, the corresponding fitted model from E is validated in V by computing the sum of squared deviations of observed from predicted values. This is repeated m times, and the α minimizing the sum of the m sums of squares is used as the cutoff in a final BE on the entire data set. BERDS is a modification of the algorithm BECV proposed by Thall, Simon and Grier (1992). An extensive simulation study shows that, compared to BECV, BERDS has a smaller model error and higher probabilities of excluding noise variables, of selecting each of several uncorrelated true predictors, and of selecting exactly one of two or three highly correlated true predictors BERDS is also superior to standard BE with cutoffs 05 or .10, and this superiority increases with the number of noise variables in the data and the degree of correlation among true predictors An application is provided for illustration.

Key Words: Cross validation; Data splitting; Monte Carlo Simulation; Regression; Variable Selection.

1. INTRODUCTION

This article introduces and evaluates a new algorithm, *backward elimination via repeated data splitting* (BERDS), for selecting predictive covariates in regression. It is a modification of the algorithm backward elimination via cross-validation (BECV) proposed by Thall, Simon, and Grier (1992), and also may be regarded as a modification of the Monte Carlo cross-validation method of Shao (1993). We consider the usual linear regression setting where each case is of the form $(Y, \mathbf{X}) = (Y, X_1, \ldots, X_p)$ and it is appropriate to assume that $E(Y \mid \mathbf{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$. The problem is to select a subset of the p candidate covariates to obtain a final model that provides accurate and reliable predictions of future values of Y for given $\mathbf{X}$. A general review of subset

Peter F Thall is Associate Professor, Department of Biomathematics, M D Anderson Cancer Center, 1515 Holcombe Boulevard, Box 237, Houston, TX 77030 e-mail rex@odin mdacc tmc edu Kathy E Russell is Programmer, Department of User Services, Rice University, 6100 Main, Mail Stop 119, Houston, TX 77005, e-mail kathy@is rice.edu Richard M Simon is Chief Biometric Research Branch, Cancer Therapy Evaluation Program, National Cancer Institute, EPN--Room 739, 6130 Executive Boulevard, Rockville, MD 20892, e-mail rich@brb nci.nih gov

©1997 *American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of North America*
Journal of Computational and Graphical Statistics, Volume 6, Number 4, Pages 416–434

selection methods is given by Miller (1990).

If $\beta_j = 0$ and X_j is not highly correlated with a true predictor having nonzero β—that is, if X_j is a *noise variable*—then it is desirable to exclude X_j since its inclusion increases the predictive variance and also may be misleading substantively. If $\beta_j \neq 0$—that is, if X_j is a *true predictor*, and $|\ \beta_j\ |$ is sufficiently large relative to the standard deviation σ of $Y\ |\ \mathbf{X}$—and if moreover X_j is not highly correlated with other true predictors in the model, then it is desirable to include X_j. Because noise variables may masquerade as true predictors and vice versa due to random variation, the goals of screening out noise variables while retaining true predictors are inherently antagonistic. This problem has been studied by Freedman and Pee (1989) in the context of explaining why simulation results may show marked disagreement with analytical results in variable selection. Derksen and Keselman (1992) studied several automatic variable selection methods and found that, in typical cases, from 20% to 74% of the selected variables are noise variables, the number of noise variables selected varies with the number of candidate predictors, and the degree of collinearity among the true predictors has an effect on the probability each is selected. This is due to the well-known problem of variance inflation when predictors are correlated, which complicates the variable selection process since any mocels containing correlated predictors are unstable.

Here, we are especially interested in settings with a large number of noise variables, hence the problem of noise variables appearing to have predictive value is an important consideration. Our simulation results show that BERDS has high probabilities of both including true predictors and excluding noise variables, and also has a high probability of selecting exactly one out of two or three highly correlated true predictors.

Stepwise test-based variable selection procedures, such as forward selection (FS) or backward elimination (BE), rely on one or two predetermined p values, α_{STAY} and α_{ENTER} for use as test cutoffs at each stage of the selection algorithm. These cutoffs are arbitrary. Although a particular numerical p value p^* has a well-defined interpretation as the type I error probability in the context of a single test of hypothesis, when $\alpha_{\text{STAY}} = p^*$ is used as the cutoff in BE, denoted BE(p^*), it may have a broad range of different meanings depending upon the number of covariates, numerical values of the parameter estimates, sample size, and σ. The use of a standard value $\alpha_{\text{STAY}} = .10$ as a cutoff in BE thus is analogous to using the criterion $|\ \hat{\beta}/se(\hat{\beta})\ | > 1$ as a standard decision criterion for a single test of the hypothesis $\beta = 0$ in simple linear regression.

Both BECV and BERDS first choose a p value cutoff α^* that minimizes an objective function quantifying the fit of the model selected by BE(α), and then perform BE(α^*) on the entire data set. Thus, both algorithms determine α_{STAY} for BE empirically from the data, eliminating the need to choose an arbitrary value of α_{STAY}. The fundamental difference is that, whereas BECV relies on a single *K-fold cross-validation* (Brieman, Friedman, Olshen, and Stone 1984) to obtain an objection function, BERDS uses repeated data splitting. In BERDS, the data are partitioned into two sets $\{E, V\}$ and an exhaustive BE is performed in E. For each α, the fitted model obtained by BE(α) in E is validated in V by computing the sum of squared deviations of observed from predicted values. This is repeated m times, and the α minimizing the sum of the m sums of squares so obtained is used as the cutoff in a final BE on the entire data set.

This modification, combined with two additional refinements described later, reduces the variability of the objective function. Consequently, BERDS provides a substantial improvement over BECV in terms of both eliminating noise variables and selecting true predictors. Our simulations indicate that BERDS is greatly superior to standard BE with $\alpha_{\mathrm{STAY}} = .05$ or .10, and that this superiority increases with both the number of noise variables in the data and the degree of correlation among the true predictors.

Several model selection methods recently have been proposed. Sauerbrei and Schumacher (1992), extending a method of Chen and George (1989), proposed using the bootstrap to evaluate variables appearing in a set of models obtained from a preliminary subset selection procedure. Breiman (1992) proposed the *little bootstrap* to estimate model error as a criterion for model selection. Breiman (1995) proposed the *nonnegative (nn) garrote*, which estimates $\boldsymbol{\beta}$ by minimizing $\sum_i (Y_i - \sum_j c_j \beta_j X_{ij})^2$ subject to $c_j \geq 0$ and $\sum_j c_j \leq s$. Tibshirani (1995) proposed a similar method, the *lasso*, which minimizes $\sum_i (Y_i - \sum_j \beta_j X_{ij})^2$ in $\boldsymbol{\beta}$ subject to $\sum_j |\beta_j| \leq t$ The lasso may produce estimators with $\hat{\beta}_j = 0$, thus deleting X_j George and McCulloch (1993) proposed a Bayesian approach to subset selection in regression, relying on Gibbs sampling to approximate posterior probabilities.

Section 2 provides some background, and Section 3 gives a formal definition of BERDS. Section 4 presents the results of an extensive simulation study. We first compare a "standard" version of BERDS to both BECV and standard BE, then evaluate several versions of BERDS. We next evaluate the sensitivity of BERDS to sample size, number of noise variables in the data, strength of a single true predictor, and the number of highly correlated true predictors in the data. We then compare BERDS to repeated BECV and to a hybrid algorithm that uses the "little bootstrap" of Breiman (1992) in place of repeated cross validation. We also evaluate BERDS under a model having predictors that are highly nonlinearly associated. Section 5 uses BERDS to analyze the "automobile accident data" discussed in Weisberg (1985), and Section 6 contains a discussion.

2. NOTATION AND BACKGROUND

The data for the ith case consist of a response variable Y_i and p covariates $X_{i,1}, \ldots, X_{i,p}$, $i = 1, \ldots, n$. We denote $\mathbf{X_i} = (1, X_{i,1}, \ldots, X_{i,p})$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$, and $E(Y_i \mid \mathbf{X}_i) = \mathbf{X}_i\boldsymbol{\beta} = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}$.

We assume each $Y_i \mid \mathbf{X}_i$ is normally distributed with variance 1, denoted $Y_i \mid \mathbf{X}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, 1)$, with the n cases mutually independent For any model corresponding to a given subset of the p covariates, the parameter vector $\boldsymbol{\beta}$ still has $p + 1$ entries but $\beta_j = 0$ if X_j is not included in the model, $j = 1, \ldots, p$, hence the non-zero entries of $\boldsymbol{\beta}$ identify which predictors are in the model.

A useful criterion for evaluating the predictive value of a model $\boldsymbol{\beta}$ is its *prediction error* PE $= E(Y^{\mathrm{new}} - \mathbf{X}\hat{\boldsymbol{\beta}})^2$, where Y^{new} is a future observation with covariates $\mathbf{X}$. Model selection may be carried out by minimizing an estimate of PE as a function of a tuning parameter, such as the biasing constant in ridge regression. A common device for estimating PE is *cross-validation* which, in its simplest form, is carried out by partitioning the data into an *estimation set* E and a *validation set* V, obtaining a fitted model in E and computing an estimate of PE in V. The roles of E and V are then reversed and the process repeated. At the other extreme, the PRESS method of Allen (1971) successively removes each $(Y_i, \mathbf{X}_i)$ from the data set and obtains a fitted model $\hat{E}_{(i)}(Y \mid \mathbf{X})$, then estimates PE as

$$\frac{1}{n}\sum_{i=1}^{n}[Y_i - \hat{E}_{(i)}(Y_i \mid \mathbf{X}_i)]^2. \tag{2.1}$$

Cross-validation has been applied in many areas, including discriminant analysis (Lachenbruch and Mickey 1968) and smoothing splines (Golub, Heath, and Wahba 1979). Efron (1983) used bootstrap cross-validation in the context of logistic regression. Hocking (1976) and Roecker (1991) used cross-validation to define a stopping rule in FS. Picard and Cook (1984) discussed cross-validation in linear regression, and general discussions were given by Stone (1974), Geisser (1975), and Efron and Tibshirani (1993).

K-fold cross-validation (Breiman et al. 1985) is carried out by first partitioning the data into K subsets $\{V_1, \ldots, V_K\}$ of equal or nearly equal size. The estimation set $E_j = \cup_{k \neq j} V_k$ is the complement of the corresponding validation set V_j, $j = 1, \ldots, K$. For each j, the model is fit in E_j and an estimate SSV_j of PE is computed by evaluating this fitted model using the data in V_j. The objective function $\sum_j SSV_j$ is then used as a criterion to evaluate the model fit in the entire data set.

Thall, Simon, and Grier (1992) used K-fold cross-validation to select the α_{STAY} cutoff in BE, as follows. For each $j = 1, \ldots, K$, an exhaustive BE is performed in the set E_j by deleting each covariate in turn until all have been removed from the model, and the p values and models when each covariate is removed are recorded. For each $\alpha \epsilon (0, 1)$, the estimated parameter vector obtained from $BE(\alpha)$ performed in E_j is denoted by $\hat{\boldsymbol{\beta}}(\alpha, E_j)$. The criterion for assessing $BE(\alpha)$ as a function of α is the *cross-validation sum of squares*

$$SSV(\alpha) = \sum_{j=1}^{K} SSV(\alpha, E_j) = \sum_{j=1}^{K}\sum_{i \epsilon V_j}[Y_i - X_i\hat{\boldsymbol{\beta}}(\alpha, E_j)]^2. \tag{2.2}$$

The value α^* minimizing $SSV(\alpha)$ is determined and a final model is selected by performing $BE(\alpha^*)$ on the entire data set. This algorithm is *backward elimination via cross-validaion*, BECV. Because $SSV(\alpha)$ evaluates the predictive ability of the model selected by $BE(\alpha)$ rather than of a given model, it is inappropriate to regard $SSV(\alpha)$ as an estimate of PE.

Although BECV is effective at screening out noise variables, in many settings it has an undesirably high probability of also excluding true predictors. This problem arises from the fact that $SSV(\alpha)$ is very sensitive to the particular partition $\{V_1, \ldots, V_K\}$ chosen. That is, two different partitions are likely to produce very different $SSV(\alpha)$ functions, hence very different values of α^* and different final models. Moreover, in most cases $SSV(\alpha)$ is not a smooth function of α with a clearly defined minimum, but rather exhibits a high degree of local variation.

We hypothesized that repeating cross-validation, say m times, and averaging the m objective functions so obtained would produce a more reliable objective function and hence an improved algorithm. We found the use of K-fold cross-validation at each repetition to be redundant from the viewpoint of data reuse, however. Thus, we simply partitioned the data into two sets $\{E, V\}$, fit the model in E and validated the fitted model in V, and repeated this process m times to obtain an overall objective function. That is, we employed *repeated data splitting* in place of K-fold cross-validation. To guard further against extreme effects of particular partitions we used a trimmed mean of the m objective functions and also truncated the domain of α to guard against local variation near 0. Our simulation results indicate that the resulting algorithm, BERDS, is superior to BECV. This may be attributed to the fact that BERDS uses a Monte Carlo approximation to full-cross validation in place of K-fold cross-validation.

3. DEFINITION OF BERDS

3.1 BACKWARD ELIMINATION VIA REPEATED DATA SPLITTING

1. Randomly partition the data into two complementary subsets, the *estimation set* E and the *validation set* V.
2. Perform an exhaustive BE on the data in E, recording the p value α_j of the test of $\beta_j = 0$ versus $\beta_j \neq 0$ when X_j is deleted, $1 \leq j \leq p$, with $\alpha_0 = 1$ for the full model with all p predictors. Denote by $\overline{\alpha}(E)$ and $\underline{\alpha}(E)$ the maximum and minimum of $\{\alpha_1, \ldots, \alpha_p\}$, respectively. For each $\alpha \epsilon [0, 1]$ the validation sum of squares corresponding to the $\{E, V\}$ split is
$$SS_{E,V}(\alpha) = \sum_{i \epsilon V} [Y_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}(\alpha, E)]^2,$$
where $\hat{\boldsymbol{\beta}}(\alpha, E) = \hat{\boldsymbol{\beta}}(\alpha_j, E)$ for $\alpha_j = \min\{\alpha_i : \alpha_i \geq \alpha, i = 1, \ldots, p\}$.
3. Repeat steps (1) and (2) m times. For each $\alpha \epsilon [0, 1]$, denote the rth estimation set, validation set, and validation sum of squares by E_r, V_r, and $SS_r(\alpha) \equiv SS_{E_r, V_r}(\alpha)$, respectively, $r = 1, \ldots, m$, and define the overall validation sum of squares $SS(\alpha)$ to be the 20% trimmed mean of $\{SS_1(\alpha), \ldots, SS_m(\alpha)\}$.
4. For given percentage q, denote by L_q and U_{100-q} the qth and $(100-q)$th quantiles of $\{\underline{\alpha}(E_1), \ldots, \underline{\alpha}(E_m)\}$ and $\{\overline{\alpha}(E_1), \ldots, \overline{\alpha}(E_m)\}$, respectively, and define α^* to be the value that minimizes $SS(\alpha)$ over $[L_q, U_{100-q}]$.
5. Obtain the final model by performing standard BE(α^*) on the entire data set.

Steps (1) and (2) apply simple two-fold data splitting to obtain a validation sum of squares which estimates the predictive variability of the model chosen by $BE(\alpha)$ in E, for each empirical α. A general discussion of data splitting is given by Picard and Berk (1990). Step (3) creates a validation sum of squares that is an average of the values obtained by repeating (1) and (2). Steps (1)–(3) of BERDS thus replace the K-fold cross-validation steps of BECV. Step (4) symmetrically truncates the α-domain on which the objective function is minimized to avoid the high degree of local variation of $SSV(\alpha)$ near 0, which persists even if m is large. This may be considered an alternative to the ".25-s rule" of BECV, which defines α^* to be the smallest value of α such that $SSV(\alpha^*) \leq SSV(\alpha^0) + 25s$, where α^0 is the true minimizing value and s^2 is an empirical estimate of the variance of $SSV(\alpha^0)$. This local variation is due to the fact that, for a given partition {E,V} and $\alpha < \underline{\alpha}(E)$, the model chosen by BE is simply $(\beta_0, 0, \ldots, 0)$, hence $\hat{Y} = \overline{Y}$ and $SS_{E,V}(\alpha) = \sum_{i \epsilon V} [Y_i - \overline{Y}(E)]^2$. The variability in

$\{\underline{\alpha}(E_1), \ldots, \underline{\alpha}(E_m)\}$ thus produces high variability among $SS_1(\alpha), \ldots, SS_m(\alpha)$ and consequently high local variation in $SS(\alpha)$ for values of α near 0. This phenomenon is even more troublesome for BECV, where the analogous variability is among the summands of $SSV(\alpha)$ created by K-fold cross-validation.

The two-fold data splitting applied in each repetition in BERDS does not switch the roles of E and V, as is done in two-fold cross-validation. We use the term "data splitting" rather than "cross-validation" to underscore this distinction. In particular, BERDS is not the same as m repetitions of BECV with $K = 2$.

For a given α, the models $\beta(\alpha, E_1), \ldots, \beta(\alpha, E_m)$ obtained from the m repetitions in BERDS typically will not be the same. Hence, as is the case with $SSV(\alpha)$ in BECV, $SS(\alpha)$ really estimates the predictive ability of the model chosen by BE(α) rather than of a specific model. In particular, it is inappropriate to use $SS(\alpha^*)$ as the basis for estimating the PE of the final model selected by BERDS.

BERDS has four parameters. The first is the percentage of the data allotted to the estimation set. We denote this by E:V so that, for example, E:V = 90:10 means that E comprises 90% and V the remaining 10% of the data. The second parameter is the number of repetitions m. Two additional parameters are the amount of trimming in Step (3), or more generally the method of averaging the m terms, and the amount of α-domain truncation q in Step (4). Provided that there is a small to moderate degree of trimming of $\{SS_1(\alpha), \ldots, SS_m(\alpha)\}$, BERDS is relatively insensitive to the amount of trimming. Because the operating characteristics of BERDS begin to stabilize for $m \geq 10$, we settled on a 20% trim to ensure that at least the maximum and minimum of $\{SS_1(\alpha), \ldots, SS_m(\alpha)\}$ are trimmed in each of the applications considered, and we use this trim level throughout. A simulation study of the effects of varying each of E:V, m and q is summarized in Section 4.3

4. SIMULATIONS

4.1 Simulation Design

In the studies summarized in Sections 4.2–4.4, the covariates are generated as uniform random variables on the interval [−3, +3], denoted $U[-3, +3]$. Either $N(0, 1)$ or lognormal covariates are employed in the studies summarized in Section 4.5. In all cases, Y is generated as a $N(\mathbf{X}\beta, 1)$. Partial F tests are used throughout, although in general any reasonable test for $\beta_j = 0$ under a model containing X_j will do. We quantify the predictive strength of each X_j by the magnitude of β_j/σ for which a two-sided .05-level t test of $\beta_j = 0$ achieves a given power under the model $E(Y \mid \mathbf{X}) = X_j\beta_j$. Unless otherwise stated, all "true predictors" in the simulations have individual power .95. This implies that, for $U[-3, +3]$ covariates, $\beta = .306, .209$, and .102 for $n = 50, 100$, and 400, respectively. The lognormal covariates with power figures .25 and .99 for $n = 100$ used in the final study in Section 4.5 have $\beta = .062$ and .285, respectively. We determined these numerical β values in empirical calibration studies analogous to that of Thall, Simon, and Grier (1992). The method of L'Ecuyer and Cote (1991) was used to generate uniform random variables, and the method of Ahrens and Dieter (1973) to generate multivariate normal random vectors. Each k-vector of correlated uniform covariates was

Table 1 Operating Characteristics of BERDS[1] Compared to BECV and Standard Backward Elimination, Under Model 1

	10 noise variables				20 noise variables			
	BERDS	*BECV*	*BE(.05)*	*BE(10)*	*BERDS*	*BECV*	*BE(.05)*	*BE(.10)*
				$n = 50$				
Power	978	930	953	979	933	824	.924	.947
# Noise	824	1.21	590	1 20	1 38	922	1.60	2.99
ME	112	130	106	136	177	.160	.197	263
α^*	076	121	—	—	.047	035	—	—
				$n = 100$				
Power	979	928	939	968	941	852	.925	955
# Noise	928	1.17	.542	1 11	1 07	1 11	1.26	2.45
ME	060	068	.055	069	080	.083	.092	122
α^*	090	.126	—	—	.046	050	—	—
				$n = 400$				
Power	.977	925	944	969	948	.860	.934	961
# Noise	1 00	1 34	.540	1 04	1 01	1.09	1 03	2 06
ME	.015	.017	013	017	019	021	021	028
α^*	.099	141	—	—	052	058	—	—

[1] BERDS based on $m = 20$ repetitions of a 50 50 E V split.

obtained by first generating a k-variate normal $(Z_1, \ldots, Z_k)$ with mean $\mathbf{0}_k$, all variances 1 and all correlations .90, and then defining $X_j = 6\Phi(Z_j) - 3$, for $j = 1, \ldots, k$, where Φ denotes the standard normal cdf. The vector $(X_1, \ldots, X_k)$ has marginal $U[-3, +3]$ variables with correlations .89. Vectors of correlated lognormal covariates were obtained similarly by defining $X_j = \exp(Z_j)$, which yields multivariate standard lognormals with correlations .85.

We evaluate each selection algorithm in terms of (1) the mean probability of including each of one or more uncorrelated true predictors, or alternatively the probabilities of selecting exactly j out of k correlated true predictors, $j = 0, 1, \ldots, k$; (2) the number of of noise variables included in the final model; and (3) the model error

$$ME = \frac{1}{n}\sum_{i=1}^{n}[\hat{Y}_i - E(Y_i \mid \mathbf{X}_i)]^2. \tag{4.1}$$

Although ME cannot be computed in practice because $E(Y \mid \mathbf{X})$ is not known, it is a natural criterion for evaluating a fitted model in a simulation study.

All simulations involve 1,000 repetitions of each case, and each reported value is the mean from these repetitions. All computations reported here were carried out on a DEC AlphaServer 2100 5/250 running OSF/1.

4.2 Comparison of BERDS to BECV and BE

The first simulations compare a standard version of BERDS, with $m = 20$, E:V = 50:50, and α-domain $[L_{90}, U_{10}]$, to BECV, BE(.05), and BE(.10). We consider the 12 data structures defined by (1) $n = 50$, 100, or 400; (2) either one true predictor (Model 1) or two correlated true predictors (Model 2); and (3) 10 or 20 noise variables. We

Table 2. Operating Characteristics of BERDS[1] Compared to BECV and Standard Backward Elimination, Under Model 2

	10 noise variables				20 noise variables			
	BERDS	*BECV*	*BE(05)*	*BE(.10)*	*BERDS*	*BECV*	*BE(05)*	*BE(10)*
				n = 50				
Pr[Both]	007	199	037	107	000	046	032	109
Pr[One]	993	800	963	893	1 00	954	968	891
# Noise	149	1 62	723	1 31	288	1 11	1 73	3 13
ME	113	171	158	186	130	173	234	296
α^*	014	160	—	—	010	045	—	—
				n = 100				
Pr[Both]	000	217	011	082	000	.059	016	.081
Pr[One]	1 00	.783	989	918	1 00	941	984	919
# Noise	081	1.56	579	1 15	104	1 16	1 24	2 50
ME	052	.085	076	.091	054	089	106	140
α^*	012	.171	—	—	006	.060	—	—
				n = 400				
Pr[Both]	005	254	006	055	000	089	010	057
Pr[One]	995	746	994	945	1 00	.911	990	943
# Noise	095	1 67	527	1 02	096	1 35	1 05	2 07
ME	013	.022	.019	022	013	024	026	033
α^*	015	191	—	—	007	077	—	—

[1] BERDS based on $m = 20$ repetitions of a 50·50 E.V split.

include only one true predictor in Model 1 because additional simulations with multiple uncorrelated p'edictors produced the same substantive and qualitative results.

Tables 1 and 2 summarize the results under Models 1 and 2, respectively. When interpreting the simulation results, it is important to bear in mind that the performance of BE(.05) or BE(.10) under a given model is essentially a matter of luck, since α_{STAY} is arbitrary. For example, in the case $n = 100$ with 10 noise variables under Model 1, Table 1 shows that on average BERDS chooses empirical $\alpha^* = .090$ and has ME = .060, each of which is bracketed by the corresponding values $\alpha_{\mathrm{STAY}} = .05, .10$, and ME = .055, .069 for BE. Thus, in terms of ME, in this case by chance BE(.05) wins and BE(.10) loses compared to BERDS. If there are 20 rather than 10 noise variables with $n = 100$, however, then both versions of BE perform very poorly compared to BERDS, with $\mathrm{ME}_{\mathrm{BE}(.05)}/\mathrm{ME}_{\mathrm{BERDS}} = 1.15$ and $\mathrm{ME}_{\mathrm{BE}(.10)}/\mathrm{ME}_{\mathrm{BERDS}} = 1.52$. In all six cases considered under Model 1, BERDS has higher power than BECV, with substantially larger differences in power when more noise variables are present. Aside from the case n=50 with 20 noise variables, BERDS includes fewer noise variables and has smaller ME than BECV. The one anomalous case illustrates an important consideration in implementing BERDS, since here E:V = 50:50 with $n = 50$ requires fits of models with 22 parameters in estimation sets having 25 observations. If E V = 90:10 is used instead so that each E has 45 observations, then on average BERDS has power .946, selects 1.30 noise variables and has ME = .153, which is an improvement over BECV. It thus appears that BERDS is uniformly superior to BECV under Model 1, provided that some care is taken to ensure that the number of observations in the estimation set is not too close to p.

Model 2 is somewhat more realistic but also more problematic in that, if both true

Table 3 Operatıng Charactenstics of BERDS† for Varyıng *m* and E.V Proportıons

	m = 10 repetıtıons		*m = 20 repetıtıons*	
E:V	*50.50*	*90.10*	*50 50*	*90 10*
		Model 1		
Power	976	.978	976	.982
# Noıse	1 07	1 24	894	1.22
ME	.062	061	059	.061
α^*	.103	137	.089	.126
		Model 2		
Pr[Both]	.010	.170	.005	.151
Pr[One]	990	.830	995	849
# Noıse	197	1.17	147	981
ME	058	077	055	075
α^*	024	.136	015	114

† *n* = 100 wıth 10 noıse varıables

predictors are selected, then the fitted final model will be highly unstable. An attractive property of BERDS in this case is that ıt ıs nearly certain to select exactly one of the two highly correlated true predictors and exclude the other. In Table 2, we denote by Pr[Both] and Pr[One] the respective probabilities of selecting both true predictors and of selecting exactly one of the two. The values of Pr[One] for BERDS are larger than those for any of the other three algorithms in all cases considered. Consequently, the superıority of BERDS to standard BE is even more pronounced under Model 2, with $\mathrm{ME}_{\mathrm{BE}}/\mathrm{ME}_{\mathrm{BERDS}}$ ranging from 1.40 to 2.59 and this ratio increasing with the number of noise variables present. This illustrates the arbitrariness of the values α_{STAY} = .05 and .10 typıcally used in BE. A somewhat surprising result is that BERDS also includes a much smaller number of noise variables than does BECV when the true predictors are correlated. The combined result of these effects is that BERDS has a substantially smaller ME than BECV in all six cases considered.

We also considered the case of five true predictors, both uncorrelated and with common correlation .90, for n = 100 with 10 noise variables. BERDS and BECV have very similar operatıng characteristics when the five true predictors are uncorrelated, with ME = .138 for BERDS and .133 for BECV. In the case of five hıghly correlated true predictors, (1) exactly 1, 2, 3, 4, or 5 of the five are selected with respective probabılities .06, .57, .31, .04, and .02 by BERDS and with probabilities .01, .26, .35, .22, and .17 by BECV; (2) BERDS selects about 73% as many noıse variables as BECV; and (3) BERDS has ME = .142 compared to .148 for BECV.

4.3 EFFECTS OF THE BERDS PARAMETERS

To assess the sensitivity of BERDS to its parameters, we considered the case n = 100 with 10 noise variables for m = 10 or 20 repetitions, split sizes E:V = 50:50 and 90:10, and α-domain truncation parameter q = 0, 90, or 100. Recall that q = 0 corresponds to no truncation, q = 90 to truncating $[0, 1]$ below at the 90th percentile of $\{\underline{\alpha}(E_1),\ldots,\underline{\alpha}(E_m)\}$ and above at the 10th percentile of $\{\overline{\alpha}(E_1),\ldots,\overline{\alpha}(E_m)\}$, and $q = 100$ to truncating at the maximum and minimum of these sets.

Regardless of m or E:V, BERDS behaves very poorly under Model 1 without any α-domain truncation ($q = 0$), hence we do not discuss this case further. Table 3 summarizes simulation results for $q = 90$. Under Model 1 with E:V = 50:50 the use of $q = 90$ rather than $q = 100$ (not tabled) has little effect on power but reduces the number of noise variables selected by 20% to 36% and hence reduces the ME by 8% to 14%, with a smaller reduction in ME for E:V = 90:10. These results indicate that, under Model 1, $q = 90$ is superior to $q = 100$, and $m = 10$ or 20 with E:V = 50:50 or 90:10 give about the same ME and power. Under Model 2, however, the split proportion E:V = 50:50 yields a uniformly and substantially smaller ME than does E:V = 90:10, primarily because Pr[One] is much larger when E:V = 50:50.

To assess more thoroughly the effects of the number of repetitions, we carried out an extensive simulation study with $m = 2, 3, \ldots, 50$ under Model 1 for $n = 100$ with 10 noise variables. In Figure 1, ME is plotted as a function of m for $q = 90$ or 100 and E:V = 50:50 or 90:10. Corresponding plots of α^* on m (not included) appear very similar to those of ME on m. The use of $q = 100$ with E:V = 50:50 produces an algorithm with the very unattractive property that both α^* and ME increase monotonically with m, although this effect disappears for E:V = 90:10. In contrast, for $q = 90$ both α^* and ME stabilize by $m = 10$ to 20, and this is the case for either split size.

Based on all of these results, we conclude that a reasonable standard version of BERDS has α-domain $[L_{90}, U_{10}]$, E:V = 50:50, and $m = 20$, although a larger value of m can do no harm. For small sample sizes, however, it may be advisable to use E:V = 90:10 to ensure that the number of observations in E is not too close to p. We further discuss choice of split size in Section 6.

4.4 Sensitivity to the Data Structure

In this section we study the sensitivity of BERDS to different data structures. Specifically, we consider the case $n = 100$ with 10 noise variables and one true $U[-3, +3]$ predictor with $\beta = .209$—that is, individual power .95—and vary in turn each of the parameters (1) sample size n; (2) number of noise variables; (3) number of highly correlated true predictors; and (4) coefficient β of a single true predictor. The standard version of BERDS defined above is used throughout.

Figure 2 summarizes all four of these simulation studies in terms of the 16 plots of each of the outcome variables ME, number of noise variables selected, average power, and α^* on each of the four variables listed previously. The first column of plots indicates that the performance of BERDS improves with increasing sample size, and also that the empirically chosen p value α^* decreases with n and then appears to stabilize at an asymptote.

The second column summarizes what may called the "needle-in-a-haystack" scenario, since the number of noise variables grows large while there is only one true predictor. BERDS appears to be remarkably insensitive to the number of noise variables in the data, since α^* apparently adjusts to the number of noise variables automatically, decreasing

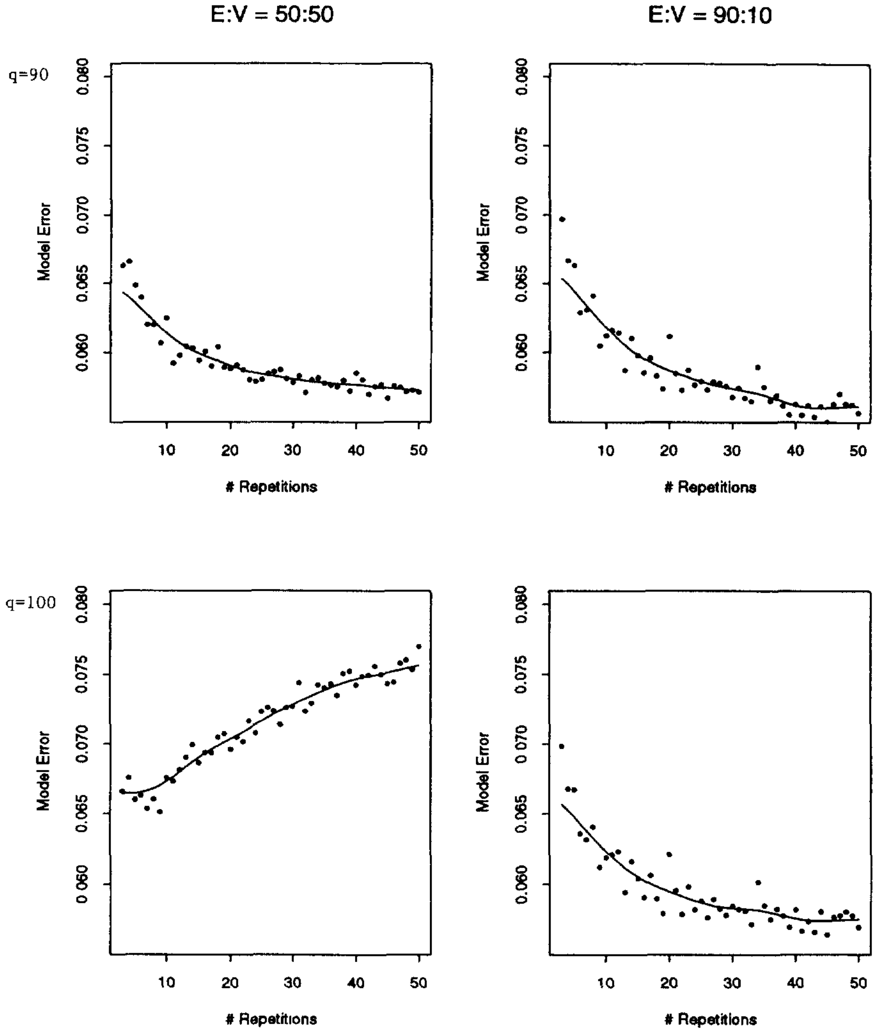

Figure 1 Model Error Versus Number of Repetitions in BERDS

with the amount of noise. Although there is a degradation of power with increasing noise, it is quite small since even with 30 noise variables the power is still .93.

The third column of plots summarizes the sensitivity of BERDS to the number p_{90} of true predictors with common correlation .90. The plots of the number of noise variables selected and α^* have similar patterns, with both jumping as p_{90} increases from 3 to 4. The actual number of noise variables selected is quite small, however. The "power" in the (3,3) plot of the matrix is [number of true predictors selected]/p_{90}, and this is remarkably flat at about .50. Recall that BERDS is almost certain to select exactly one of two true predictors with correlation .90, as noted in the lower half of Table 2, and even when $p_{90} = 3$ the mean number selected is 1.08. Thus, when $p_{90} = 2$ or 3, BERDS is nearly certain to include exactly one, with the number selected increasing thereafter to an asymptote of roughly $p_{90}/2$. That is, BERDS selects about half the number of

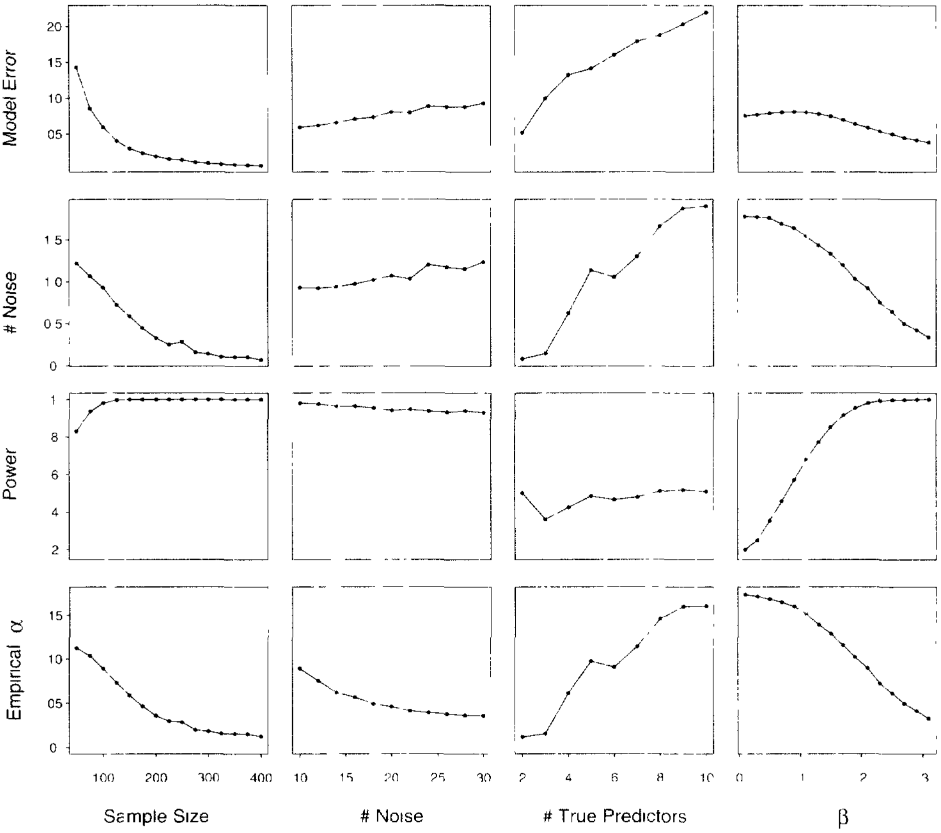

Figure 2 Model Error, Number of Noise Variables Selected, Power of Selecting One True Predictor, and Empirical α^*, Each as a Function of Sample Size, Number of Noise Variables Number of Correlated True Predictors, and the β of One True Predictor BERDS is implemented with $E{:}V = 50{:}50$ and $m = 20$ Except for the one parameter varied in each column of four plots, $n = 100$, there are 10 noise variables, there is one true predictor, and $\beta = 209$ ($\pi_\beta = .95$)

equicorrelated true predictors when $p_{90} > 3$.

The plots in column 4 illustrate the sensitivity of BERDS to the strength of one true predictor, in terms of the magnitude of its parameter β. Not surprisingly, the power increases with β, and this plot has the appearance of a conventional power curve. ME shows a gradual decrease with β, although the magnitude of this decrease is small compared to the drop in ME with increasing sample size. The decrease in α^* with β is quite pronounced over the domain studied. Together with the fact that fewer noise variables are selected as β increases, this illustrates the ability of BERDS to automatically eliminate noise variables while retaining true predictors.

It is also informative to consider the matrix of plots across its rows—that is, in terms of the output variables ME, number of noise variables selected, average power, and α^*. ME appears relatively insensitive to both the number of noise variables and β, but decreases with n and increases with p_{90}. It thus appears that, beyond $p_{90} = 2$ or 3, BERDS has trouble dealing with collinearity. This is further illustrated by the second row of plots, where the only thing that causes the number of noise variables selected to increase substantially is an increase in p_{90}. The third row indicates that BERDS has good power under a wide range of settings. The fourth row illustrates the manner in which the α^* chosen by BERDS automatically adjusts to the particular data structure.

4.5 Additional Studies

Various other ways to compute an objective function in the algorithm are possible. We consider the following two. Because each repetition in BERDS fits in E and validates in V but does not then reverse the roles of E and V, BERDS is not the same thing as repeating BECV with $K = 2$. A natural question is whether repeating the K-fold cross-validation in BECV can achieve the same improvement as BERDS. To address this we considered a modified version of BECV which uses repeated K-fold cross-validation, with either $K = 2$ or $K = 10$. A similar question is how well the algorithm performs using the little bootstrap proposed by Breiman (1992) in place of repeated data splitting to obtain an objective function. To implement this, we first carried out an exhaustive BE on the entire data set to obtain empirical p values $\{\alpha_1, \ldots, \alpha_p\}$ Following Breiman (1995, p. 375), we generated $\{\tilde\epsilon_1, \ldots, \tilde\epsilon_n\} \sim$ iid $N(0, t^2\sigma^2)$ using the full model MSE for σ^2 and t = .6, .7, or .8, then defined $\tilde Y_i = Y_i + \tilde\epsilon_i$, and performed BE($\alpha$) on $\{(\tilde Y_i, \mathbf{X}_i) \; 1 \leq i \leq n\}$ to obtain $\tilde\beta(\alpha)$ for each empirical α. The little bootstrap estimate of PE is

$$\sum_{i=1}^{n}(Y_i - \mathbf{X}_i\tilde\beta(\alpha))^2 + 2B_t(\alpha). \tag{4 2}$$

where

$$B_t(\alpha) = t^{-2}E\left[\sum_{i=1}^{n}\tilde\epsilon_i\mathbf{X}_i\tilde\beta(\alpha)\right] \tag{4.3}$$

and the expectation is over the simulated $\{\tilde\epsilon_i\}$ To obtain $B_t(\alpha)$ we repeated this 25 times and used the mean in place of the E operator

The results are summarized in Table 4 For comparability to standard BERDS, repeated BECV is based on a 20% trimmed mean of $\{SS_j(\alpha), j = 1, \ldots, m\}$ from $m = 10$ repetitions with α-domain $[L_{90}, U_{10}]$. Thus, BECV with repeated two-fold cross-validation and BERDS have the same amount of data reuse. BERDS is greatly superior to BECV under Model 2 according to all three criteria, while under Model 1 BECV with repeated two-fold cross-validation has slightly fewer noise variables, slightly less power, and slighly smaller ME. An unexpected result is that BECV with repeated 10-fold cross-validation performs relatively poorly under either model. It likewise appears that use of the little bootstrap in place of repeated data splitting to choose the cutoff for BE does not produce an algorithm with good properties. It is possible that modification of the algorithm, possibly by searching over a different set of α's or increasing the number of repetitions to estimate $B_t(\alpha)$, may produce an improved method

An associate editor has suggested evaluating BERDS under a model where the covariates are nonlinearly associated or have asymmetric marginals. If $\mathbf{X}$ is partitioned into $(\mathbf{X}_1, \mathbf{X}_2)$ with $\beta = (\beta_0, \beta_1, \beta_2)$, in general $E(Y \mid \mathbf{X}_1) = \beta_0 + \beta_1'\mathbf{X}_1 + \beta_2'E(\mathbf{X}_2 \mid \mathbf{X}_1)$

Table 4 Comparison of BERDS to Repeated BECV and BE with the Little Bootstrap[1]

	BERDS	Repeated BECV		Little bootstrap		
		2-fold	10-fold	$t = 6$	$t = 7$	$t = 8$
			Model 1			
Power	976	951	915	818	794	766
# Noise	.894	675	860	1 76	1 46	1 20
ME	059	054	061	086	084	083
α^*	089	070	094	135	105	082
			Model 2			
Pr[Both]	005	048	142	246	180	138
Pr[One]	.995	952	858	555	569	587
Pr[Neither]	000	000	000	199	251	275
# Noise	147	594	889	2 14	1 55	1 25
ME	056	074	074	176	192	199
α^*	015	067	106	193	131	099

[1] $m = 20$ for BERDS with 50.50 E V proportion, and $m = 10$ for repeated BECV, $n = 100$ and 10 noise variables

and $\mathrm{var}(Y \mid \mathbf{X}_1) = \sigma^2 + \boldsymbol{\beta}_2'\mathrm{var}(\mathbf{X}_2 \mid \mathbf{X}_1)\boldsymbol{\beta}_2$. hence if $\boldsymbol{\beta}_2 \neq \mathbf{0}$ a nonlinear association between the two subvectors of covariates may have a nontrivial effect on any variable selection procedure. To study this we define Model 3, which has six lognormal covariates obtained from standard normals with common correlation .50, yielding correlations .39 among the lognormals. The model has two strong predictors, two weak predictors, and two "noise" covariates. Specifically, $\beta_1 = \beta_2 = .285$, corresponding to individual power figures .99 for $n = 100$, $\beta_3 = \beta_4 = .062$, corresponding to individual power .25, and $\beta_5 = \beta_6 = 0$ Due to the associations between each of X_5, X_6 and the other covariates, $E(Y \mid X_5, X_6) \neq 0$ hence X_5 and X_6 are not really noise variables

A simulation study under Model 3 with $n = 100$ is summarized in Table 5. BERDS and BECV have very similar performance under this model, and the four algorithms have nearly identical ME. The average α^* values .27 selected by BERDS and .41 by BECV are quite large compared to the usual $\alpha_{\mathrm{STAY}} = .05$ or .10, essentially because there is very little "noise." Both BERDS and BECV have higher power figures than BE(.05) and BE(.10), with small differences for the strong predictors but larger differences for the weak predictors and covariates with $\beta = 0$ Thus, although BERDS is not markedly superior to standard BE with regard to ME in this case, it has a much higher proba-

Table 5 Operating Characteristics of BERDS[1] BEC V, and BE under Model 3

Selection probability	*BERDS*	*BECV*	*BE(.05)*	*BE(.10)*
Strong Covariates	988	992	972	978
Weak Covariates	458	585	245	344
Covariates With $\beta = 0$	290	433	077	128
ME	070	070	069	068
α^*	271	411	–	–

[1] BERDS based on $m = 20$ repetitions of a 50 50 E V split, $n = 100$

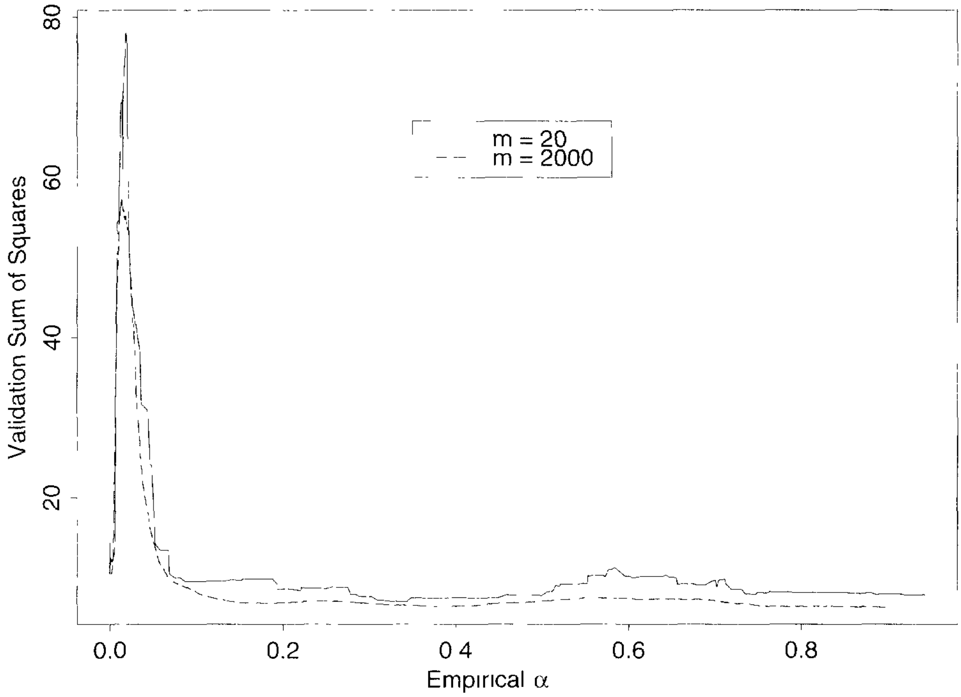

Figure 3 Cross-Validation Sum of Squares as a Function of α in BERDS Applied to the Highway Accident Data with m = 20 and 2,000

bility of selecting the correct model $\{X_1, X_2, X_3, X_4\}$. Moreover, the model $\{X_1, X_2\}$ selected most frequenlty by BE is incorrect and heteroscedastic. These results indicate that BERDS adapts quite well to situations where the predictors have nonlinear marginals and nonlinear associations.

5. THE HIGHWAY ACCIDENT DATA

In this section we apply BERDS to the highway accident data presented and analyzed in detail by Weisberg (1985) and also analyzed by Thall et al. (1992). The data consist of $n = 39$ cases (highway sections) with $p = 13$ covariates that are candidates for predicting the outcome Y = number of automobile accidents per million miles.

To obtain a more complete picture of the behavior of BERDS in this data set we applied it 100 times, since the results of BERDS are random, like those of any method involving sampling from the data. To examine the distribution of final models chosen, we did this for for $m = 10, 20, 50, 100, 500, 1{,}000$, and $2{,}000$. In a single application of BECV with $K = 19$ to these data, Thall et al. (1992) obtained the model $\{1, 4, 9\}$. Because increasing K in BECV is analogous to increasing the E,V proportion in BERDS, we repeated BECV 100 times for $K = 10$ and $K = 19$ The results for BECV and those for BERDS with $m = 10, 20$, and 50 are summarized in Table 6, and plots of $SS(\alpha)$ on α for $m = 20$ and 2,000 are given in Figure 3

Table 6. Applıcatıon to the Hıghway Accıdent Data

			Number of tımes selected out of 100				
			BERDS			BECV	
Model variables	C_p	R^2	$m = 10$	$m = 20$	$m = 50$	$K = 10$	$K = 19$
1,9	3 31	652	2	0	0	1	0
1,4,9	236	701	0	0	0	3	1
1,4,8,9,12	– 383	745	23	22	19	20	21
1,3.4,8,9.12	881	752	20	24	41	15	19
1 3,4,8,9,12,13	2 45	756	16	16	9	17	12
1,3,4,7,8,9,12,13	4 27	758	9	6	1	3	3
1.3,4.5,7,8,9,12,13	6 14	759	1	1	2	1	1
1,3,4,5,7,8,9,11,12,13	8 02	760	29	31	28	31	27
1,2,3,4,5,7,8,9,10,11 12,13	12 1	760	0	0	0	0	14
1,2,3,4,5,6,7,8,9,10,11,12,13	14 0	761	0	0	0	9	2

The figure shows that, in terms of ME, there is very little difference between the model $\{1,4,8\ 9,12\}$ and any model containing these five variables as a proper subset. As noted by Weisberg (1985), this is the model obtained by BE using 2.0 as the cutoff on the F-statistic domain, and also by FS using this as both the enter and stay cutoffs. An exhaustive BE shows that any α_{STAY} between .164 and .347 yields this model. While the distribution of final models chosen by BERDS must converge to a unit mass on one model as $m \rightarrow \infty$, this convergence is rather slow for this data set. For values of $m >$ 50, the two models $\{1,3,4,8,9,12\}$ and $\{1,3,4,5,7,8,9,11,12,13\}$ are chosen with successively higher frequencies, with the latter chosen 97% of the time when m =2,000. This application illustrates the usefulness of examining the plot of $SS(\alpha)$ on α when applying BERDS.

6. DISCUSSION

BE is widely used in linear regression. Although the typical values $\alpha_{\text{STAY}} = .05$ or .10 used as the p value cutoff in BE are arbitrary and do not correspond to the type I error of a single test of hypothesis, many reports of data analyses involving BE treat α_{STAY} as a standard statistical significance level. Although many statisticians have recognized that it is improper to interpret α_{STAY} in this way, there have been few methods proposed for selection of this parameter.

We have introduced and examined by simulation and application a new algorithm, BERDS, for variable selection in linear regression. It is a modification of the algorithm BECV proposed by Thall et al. (1992), and is a continuation of their attempt to select α_{STAY} empirically using the data. Both algorithms are based on the approach of choosing $\alpha_{\text{STAY}} = \alpha$ for use in BE by first minimizing an objective function that quantifies model fit as a function of α. In particular, BERDS does away with the problem of choosing an arbitrary cutoff when applying BE.

Thall et al. (1992) used 10-fold cross-validation to obtain a validation sum of squares $SSV(\alpha)$, analogously to the use of cross-validation by Breiman et al. (1985). The partition into 10 subsets was randomly selected and the resulting $SSV(\alpha)$ was quite noisy as a function of α and highly dependent on the particular partition chosen. We have determined that this noise can be reduced by using repeated data splitting and averaging the estimated functions $SS_1(\alpha), \ldots, SS_m(\alpha)$ over the replications to obtain an overall $SS(\alpha)$. We used a 20% trimmed mean of these values in order to further smooth the function. To reduce the noise at the endpoints of the domain of $SS(\alpha)$, we limited the range of α_{STAY} over which the minimization takes place, rather than using a version of the *ad hoc* 1-SE rule originally used by Breiman et al. (1985) and also used by Thall et al. (1992).

Breiman et al. (1985) used 10-fold cross-validation because it was less computationally demanding than leave-one-out cross-validation. K-fold cross-validation estimates PE for a model based on $(1 - K^{-1})n$ observations. This has been widely viewed as a reason for using large K, which approaches the leave-one-out procedure. The models determined in the K estimation subsets $E_1 \ldots \quad E_K$ are highly "correlated" however. Although using a large K provides an estimate of PE for a model developed on a sample size close to n, the estimate itself is of reduced precision because it is based on a smaller validation set V. The facts that K different E-V pairs are used and that the random selection of the partition itself may be replicated does not change the fundamental fact that V is small.

The optimal division between E and V in BERDS depends on the sample size, the number and distribution of the covariates, the residual variance, and other factors. Shao (1993) showed that leave-one-out cross-validation does not even provide a consistent estimate of PE. To obtain consistency, both the size n_E of E and the size n_V of V must increase without limit as $n \rightarrow \infty$. Ideally, one would like $n_E = n$ and $n_V = n$, but of course this is impossible when using internal cross-validation that requires data splitting. For the range of conditions considered in this article, we found that simple two-fold data splitting performed as well as, and in many cases substantially better than, K-fold cross-validation. This is due to the fact that BERDS uses a Monte Carlo approximation to full-cross validation in place of K-fold cross-validation, hence is more efficient. We also found that using $K = 2$ with $m/2$ replications was no better than using half the data for estimation and the other half for validation, without reversing the roles of the two sets, while replicating the selection of the random partition m times. Picard and Berk (1990) discussed the trade-off between making n_E large to obtain a reliable model fit and making n_V large to obtain a reliable validation, and they provided specific criteria for optimizing split size in linear regression. It seems reasonable that the behavior of BERDS can be improved by optimizing split size empirically, and this is an important issue for future investigation.

We found that using $m = 20$ replications enabled α_{STAY} to be determined in a manner that provided models with much better prediction characteristics than BECV. As noted in Section 5, however, this is not to say that the same model is always selected by BERDS with $m = 20$. When there are several models with similar $SS(\alpha)$'s, while the probability distribution of selected models must converge to a point mass on one model with increasing m, this convergence may be very slow. Although uniqueness of the model selected among models with similar $SS(\alpha)$'s may not be important, it may be worthwhile to identify models having very similar $SS(\alpha)$'s and decide which to use on other bases. While a single application of the procedure, with say $m = 20$, provides a single objective function, it may be worthwhile to replicate the analysis with other groups of m partitions in order to identify the models that provide similar $SS(\alpha)$'s. If it is important to uniquely identify one model, alternatives to random selection of partitions should be examined.

[Received April 1996. Revised November 1996.]

REFERENCES

Allen, D M (1971), "Mean Square Error Prediction as a Criterion for Selecting Variables" (with discussion), *Technometrics*, 13, 469–481

Ahrens, J H, and Dieter, U (1973), "Extensions of Forsythe's Method for Random Sampling from the Normal Distribution," *Mathematical Computation*, 27, 927–937

Breiman, L. (1992), "The Little Bootstrap and Other Methods for Dimensionality Selection in Regression X-Fixed Prediction Error," *Journal of the American Statistical Association*, 87, 738–754

——— (1995), 'Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, 37, 373–384

Breiman, L, Friedman, J H., Olshen, R A, and Stone, C J (1984), *Classification and Regression Trees*, Belmont, CA Wadsworth

Chen, C.-H., and George, S.L. (1989), "The Bootstrap and Identification of Prognostic Factors Via Cox's Proportional Hazards Regression Model," *Statistics in Medicine*, 4, 39–46

Cox, D R (1972), 'Regression Models and Life Tables,' (with discussion), *Journal of the Royal Statistical Society*, Ser B, 74, 187–220

Derksen, S, and Keselman, H.J. (1992), "Backward, Forward and Step wise Automated Subset Selection Algorithms· Frequency of Obtaining Authentic and Noise Variables," *British Journal of Mathematical and Statistical Psychology* 45, 265-282

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

Efron, B, and Tibshirani, R.J (1993), *An Introduction to the Bootstrap*, New York Chapman and Hall

Freedman, D A, and Pee, D (1989), "A Return to a Note on Screening Regression Equations" *The American Statistician*, 43 279–282

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

George, E I, and McCulloch, R E (1993), "Variable Selection Via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

Golub, G, Heath, M, and Wahba, G (1979), 'Generalized Cross Validation as a Method for Choosing a Good Ridge Parameter," *Technometrics*, 21, 215–224

Hocking, R R (1976), "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, 32, 1–49

Lachenbruch, P, and Mickey, M (1968), "Estimation of Error Rates in Discriminant Analysis,' *Technometrics*, 10, 1–11

L'Ecuyer, P, and Cote, S (1991), "Implementing a Random Number Package with Splitting Facilities," *ACM Transactions on Mathematical Software*, 17, 98–111

Miller, A J (1990), *Subset Selection in Regression*, New York Chapman and Hall

Picard, R R, and Berk, K.N (1990), "Data Splitting" *The American Statistician*, 44, 140–147

Picard, R R, and Cook, D (1984) "Cross-Validation of Regression Models," *Journal of the American Statistical Association*, 79, 575–583

Roecker, E B (1991), "Prediction Error and Its Estimation for Subset-Selected Models," *Technometrics*, 33, 459–468

Sauerbrei, W. and Schumacher, M. (1992), "A Bootstrap Resampling Procedure for Model Building: Application to the Cox Regression Model," *Statistics in Medicine*, 11, 2093–2109.

Shao, J. (1993), "Linear Model Selection by Cross-Validation," *Journal of the American Statistical Association*, 88, 486–494.

Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society*, Ser. B, 36, 111–147.

Thall, P.F., Simon, R., and Grier, D.A. (1992), "Test-Based Variable Selection via Cross-Validation," *Journal of Computational and Graphical Statistics*, 1, 41–61.

Tibshirani, R. (1995), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Ser. B, 58, 267–288.

Weisberg, S. (1985), *Applied Linear Regression*, New York: Wiley.