Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation

Author(s): Bradley Efron

Source: *Journal of the American Statistical Association*, Jun., 1983, Vol. 78, No. 382 (Jun., 1983), pp. 316-331

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

Stable URL: https://www.jstor.org/stable/2288636

Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation

BRADLEY EFRON*

We construct a prediction rule on the basis of some data, and then wish to estimate the error rate of this rule in classifying future observations. Cross-validation provides a nearly unbiased estimate, using only the original data. Cross-validation turns out to be related closely to the bootstrap estimate of the error rate. This article has two purposes: to understand better the theoretical basis of the prediction problem, and to investigate some related estimators, which seem to offer considerably improved estimation in small samples.

KEY WORDS: Bootstrap; Prediction problem; ANOVA decomposition; Logistic regression.

1. INTRODUCTION

In the prediction problem the statistician has available a set of cases $x_1, x_2, \ldots, x_n$ collectively called the *training set* $\mathbf{x}$. Each case consists of two parts $x_i = (t_i, y_i)$, where t_i is a vector of predictors and y_i is a response variable. For example, t_i might describe a medical patient's age, weight, sex, race, previous disease history, and so on, and y_i might indicate whether the patient survived a certain operation. On the basis of the training set, a prediction rule $\eta(t, \mathbf{x})$ is constructed. The intention is to use $\eta(t_0, \mathbf{x})$ to predict a future unobserved response y_0 on the basis of its predictor vector t_0.

We are mainly concerned with the situation where y_i is a dichotomy, such as "survived" or "didn't survive," and the prediction $\eta_i = \eta(t_i, \mathbf{x})$ is likewise dichotomous. Let $Q[y_i, \eta_i]$ indicate the correctness of the ith prediction,

$$Q[y_i, \eta_i] = 0 \quad \text{if } \eta_i = y_i$$
$$= 1 \quad \text{if } \eta_i \neq y_i. \tag{1.1}$$

The *true error rate* (Err) of the prediction rule $\eta(t, \mathbf{x})$ is its probability of incorrectly classifying a randomly selected future case $X_0 = (T_0, Y_0)$, in other words the expectation $E\, Q[Y_0, \eta(T_0, \mathbf{x})]$.

Our goal is to estimate Err on the basis of the training set $\mathbf{x}$. The most obvious estimate is the *apparent error rate* $\bar{err} = \sum_{i=1}^{n} Q[y_i, \eta(t_i, \mathbf{x})]/n$, which is the proportion of observed errors made by $\eta(t, \mathbf{x})$ on its own training set $\mathbf{x}$. Usually $\bar{err}$ tends to be smaller than Err, because the same data have been used both to construct and to evaluate $\eta(t, \mathbf{x})$. This is a familiar fact in ordinary linear regression, where $\bar{err}$ = (residual sum of squares from fitted model)/n underestimates the true residual variance, and so the denominator n is usually reduced.

Cross-validation circumvents this difficulty by removing each x_i from the data set used in its own prediction. Let $\mathbf{x}_{(i)}$ be the training set with x_i removed, and $\eta(t, \mathbf{x}_{(i)})$ be the corresponding prediction rule. The cross-validated error rate is

$$\hat{\text{Err}}^{(\text{CV})} = \frac{1}{n} \sum_{i=1}^{n} Q[y_i, \eta(t_i, \mathbf{x}_{(i)})]. \tag{1.2}$$

The well-known paper by Lachenbruch and Mickey (1968) is a good reference. Cross-validation is discussed in a wider context by Stone (1974) and Geisser (1975).

In the next section we introduce another estimate of Err, based on the *bootstrap*, Efron (1979): $\hat{\text{Err}}^{(\text{BOOT})}$ is essentially the nonparametric maximum likelihood estimate of Err, assuming only that the training cases x_i are a random sample from some unknown distribution F on the space of possible vectors $x = (t, y)$. We will see that in some ways $\hat{\text{Err}}^{(\text{BOOT})}$ outperforms $\hat{\text{Err}}^{(\text{CV})}$ as an estimator of Err, though the comparison is not totally one-sided. Other estimators introduced in later sections outperform both $\hat{\text{Err}}^{(\text{CV})}$ and $\hat{\text{Err}}^{(\text{BOOT})}$, in an admittedly small catalog of five sampling experiments.

This article has two main purposes: to understand the theoretical basis of the prediction problem, especially as it relates to cross-validation and the bootstrap; and to investigate some related estimators, which seem to offer considerably improved estimation of err when the training set is small. The discussion is actually in the opposite order. The related estimators, all of which are simple variants of the bootstrap, are introduced in Sections 3 through 6. Sections 7 and 8 concern a decomposition of the prediction problem, based on the ANOVA description of Efron and Stein (1981), clarifying the theoretical connections between the various methods. The article ends with some remarks and a summary of recommendations.

2. THE BOOTSTRAP AND CROSS-VALIDATION

This section describes the bootstrap estimate of the true error rate, and relates it to $\hat{\text{Err}}^{(\text{CV})}$, the cross-vali-

* Bradley Efron is Professor of Statistics and Biostatistics, Department of Statistics, Stanford University, Stanford, CA 94305. In addition to the referees and the editor, Stephen Stigler, the author is grateful to Professors Samprit Chatterjee, Jerome Friedman, Gail Gong, and Charles Stone for helpful discussions on the prediction problem. Financial support was provided by Public Health Service Grant 5 R01 GM21215, and National Science Foundation Grant MCS80-24649.

© Journal of the American Statistical Association
June 1983, Volume 78, Number 382
Theory and Methods Section

dation estimate. It is taken from a longer discussion in Chapter 7 of Efron (1982). We begin with a more careful description of the prediction problem.

Each case $x_i = (t_i, y_i)$ in the training set is the realization of a random quantity $X_i = (T_i, Y_i)$, where T_i is a p-dimensional row vector of predictors and Y_i is a real-valued response variable. We assume that there is some unknown distribution F on the $p+1$-dimensional sample space $\mathcal{X} = \mathcal{R}^{p+1}$ such that the training set $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is a random sample from F,

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F, \tag{2.1}$$

iid abbreviating "independent and identically distributed." For convenience let $X_0 = (T_0, Y_0)$ denote a future observation from F, independent of the training set.

We have at hand a specific recipe for constructing a prediction rule $\eta(t, \mathbf{x})$ on the basis of the training set. A familiar example is the ordinary least squares rule $\eta(t, \mathbf{x}) = t(\mathbf{t}'\mathbf{t})^{-1}\mathbf{t}'\mathbf{y}$, where $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ and $\mathbf{t}'$ is the $p \times n$ matrix $(t'_1, t'_2, \ldots, t'_n)$. Notice that this rule remains the same under any reordering of the cases $x_1, x_2, \ldots, x_n$ constituting $\mathbf{x}$. This is the usual case and will be assumed to hold in what follows. We use $\eta(t_0, \mathbf{x})$ to predict y_0 from t_0, and measure the prediction error according to some function $Q[y_0, \eta(t_0, \mathbf{x})]$. In the dichotomous case both y and η are either 0 or 1, and Q is described by (1.1). The true error rate $\text{Err}(\mathbf{x}, F)$ is the expected value

$$\text{Err} = E_F\, Q[Y_0, \eta(T_0, \mathbf{x})], \tag{2.2}$$

the expectation being taken over $X_0 = (T_0, Y_0) \sim F$, with $\mathbf{x}$ fixed at its observed value. We will sometimes write $Q(x_0, \mathbf{x})$ for $Q[y_0, \eta(t_0, \mathbf{x})]$.

The apparent error rate $\bar{err}(\mathbf{x})$ is the statistic

$$\bar{err} = \frac{1}{n}\sum_{i=1}^{n} Q[y_i, \eta(t_i, \mathbf{x})] \tag{2.3}$$

usually an underestimate of Err. Let $\text{op}(\mathbf{x}, F)$, "op" being short for optimism, indicate the random variable

$$\text{op} = \text{Err} - \bar{err} \tag{2.4}$$

with expectation ω,

$$\omega(F) = E_F\, \text{op}(\mathbf{X}, F) = E_F\{\text{Err}(\mathbf{X}, F) - \bar{err}(\mathbf{X})\}. \tag{2.5}$$

If ω were known, we could estimate Err with

$$\hat{\text{Err}}^{(\text{IC})} = \bar{err} + \omega, \tag{2.6}$$

IC standing for "ideal constant." In most cases ω is not known, and must itself be estimated from the training set $\mathbf{x}$. Cross-validation as described in Section 1 amounts to using the estimate

$$\hat{\omega}^{(\text{CV})} = \frac{1}{n}\sum_{i} Q[y_i, \eta(t_i, \mathbf{x}_{(i)})] - \bar{err}, \tag{2.7}$$

so that $\hat{\text{Err}}^{(\text{CV})} = \bar{err} + \hat{\omega}^{(\text{CV})}$ equals $\sum Q[y_i, \eta(t_i, \mathbf{x}_{(i)})]/n$.

The bootstrap estimate $\hat{\text{Err}}^{(\text{BOOT})}$ equals $\bar{err} + \hat{\omega}^{(\text{BOOT})}$, where $\hat{\omega}^{(\text{BOOT})}$ if the nonparametric maximum likelihood estimate $\omega(\hat{F})$, $\hat{F}$ being the empirical probability distribution putting mass $1/n$ on each observed case,

$$\hat{F}: \text{mass } \frac{1}{n} \text{ on } x_i, \quad i = 1, 2, \ldots, n. \tag{2.8}$$

We can describe $\omega(\hat{F})$ more explicitly, in a way that suggests how to actually evaluate it. Let $\mathbf{X}^*$ indicate a random sample of size n from $\hat{F}$,

$$X_1^*, X_2^*, \ldots, X_n^* \overset{\text{iid}}{\sim} \hat{F}, \tag{2.9}$$

and E_* indicate expectation with respect to the random mechanism (2.9), $\hat{F}$ fixed at its observed value. Then

$$\begin{aligned}\hat{\omega}^{(\text{BOOT})} &= \omega(\hat{F}) = E_*\, \text{op}(\mathbf{X}^*, \hat{F}) \\ &= E_* \sum_i \left(\frac{1}{n} - P_i^*\right) Q[y_i, \eta(t_i, \mathbf{X}^*)],\end{aligned} \tag{2.10}$$

with P_i^* indicating the proportion of the bootstrap sample on x_i,

$$P_i^* = \frac{\#\{X_j^* = x_i\}}{n}, \quad i = 1, 2, \ldots, n. \tag{2.11}$$

The last expression in (2.10) is obtained by following through definition (2.5) (see Efron 1982).

Usually $\hat{\omega}^{(\text{BOOT})}$ must be evaluated by Monte Carlo: independent bootstrap training sets $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$ are generated according to (2.9), and for each $\mathbf{x}^{*b}$ the prediction rule $\eta(t, \mathbf{x}^{*b})$ is calculated. This gives a bootstrap replication of op according to (2.10), $\text{op}^{*b} = \sum_{i=1}^{n} (1/n - P_i^{*b})\, Q[y_i, \eta(t_i, \mathbf{x}^{*b})]$, and we approximate $\hat{\omega}^{(\text{BOOT})}$ by the average $\sum_{b=1}^{B} \text{op}^{*b}/B$. As $B \to \infty$ this approaches definition (2.10). For practical purposes B in the range 25–200 seems quite adequate. *A better Monte Carlo method is given in Section 8.*

As an example, suppose $p = 2$, $n = 14$, and that each T_i is bivariate normal with mean vector either $\pm(\frac{1}{2}, 0)$,

$$\begin{aligned}Y_i &= \begin{matrix}0 & \frac{1}{2}\\ \text{prob} & \text{and}\\ 1 & \frac{1}{2}\end{matrix} \quad T_i \mid Y_i = y_i \\ &\sim N_2((y_i - \tfrac{1}{2}, 0), I).\end{aligned} \tag{2.12}$$

The prediction rule is Fisher's estimated linear discriminant,

$$\eta(t_0, \mathbf{x}) = \begin{matrix}0 & <0\\ \text{if } \hat{\alpha} + t_0\hat{\beta}' \text{ is} & .\\ 1 & \geq 0\end{matrix} \tag{2.13}$$

The coefficients $\hat{\alpha}$ and $\hat{\beta}$ are given in terms of $n_y = \#\{y_j = y\}$, $\bar{t}_y = \sum_{y_i = y} t_i/n_y$, and $S = [\sum_j t'_j t_j - n_1 \bar{t}'_1 \bar{t}_1 - n_2 \bar{t}'_2 \bar{t}_2]/n$: $\hat{\alpha} = [\bar{t}_1 S^{-1} \bar{t}'_1 - \bar{t}_2 S^{-1} \bar{t}'_2]/2$ and $\hat{\beta} = [\bar{t}_2 - \bar{t}_1]S^{-1}$.

In the sampling experiment subsequently called (2, 14), 100 independent trials of situation (2.12), (2.13) were generated. Results for the first 10 of these, and summary statistics for all 100, appear in Table 1. We see that op

Table 1. The First 10 Trials of Experiment (2,14), Described at (2.12), (2.13), and Summary Statistics for all 100 Trials; $\hat{\omega}^{(CV)}$ is Less Biased Than $\hat{\omega}^{(BOOT)}$, But Much More Variable. The Bootstrap Was Run With B = 200 Replications per Trial

Trial	Err	$\bar{err}$	op	$\hat{\omega}^{(CV)}$	$\hat{\omega}^{(JACK)}$	$\hat{\omega}^{(BOOT)}$
1	.458	.286	.172	.214	.214	.083
2	.312	.357	−.045	.000	.066	.098
3	.313	.357	−.044	.071	.066	.110
4	.351	.429	−.078	.071	.066	.107
5	.330	.357	−.027	.143	.148	.102
6	.318	.143	.175	.214	.194	.073
7	.310	.071	.239	.071	.066	.047
8	.380	.286	.094	.071	.056	.097
9	.360	.429	−.069	.071	.087	.127
10	.335	.143	.192	.000	.010	.048
100 Trials Exp	.360	.264	ω = .096	.091	.093	.080
100 Trials (Sd)	(.045)	(.123)	(.113)	(.073)	(.068)	(.028)

averaged .096, which is ω except for sampling error, and so $\bar{err}$ tends to seriously underestimate Err in this case, E Err/E $\bar{err}$ = 1.36. For each trial, $\hat{\omega}^{(CV)}$ was calculated according to (2.7), and $\hat{\omega}^{(BOOT)}$ calculated according to the Monte Carlo algorithm, using B = 200 bootstrap replications per trial. Err was calculated theoretically from (2.12), (2.13). The bootstrap estimate of ω was biased slightly downwards, averaging .080, but was far less variable than $\hat{\omega}^{(CV)}$.

The jackknife, or more precisely the jackknife estimate of bias, relates cross-validation to the bootstrap. It uses a quadratic expansion for op($\mathbf{X}^*$, $\hat{F}$) and properties of the multinomial distribution to show that E_* op($\mathbf{X}^*$, $\hat{F}$) can be approximated by

$$\hat{\omega}^{(JACK)} = \frac{1}{n}\sum_i Q[y_i, \eta(t_i, \mathbf{x}_{(i)})] - \frac{1}{n}\sum_i \{\sum_j Q[y_i, \eta(t_i, \mathbf{x}_{(j)})]/n\}. \quad (2.14)$$

Comparing (2.14) with (2.7), it is not surprising that $\hat{\omega}^{(JACK)}$ is usually close in value to $\hat{\omega}^{(CV)}$, as seen in Table 1. Gong (1982) shows that they have asymptotic correlation 1.00, under smoothness conditions on Q. The correlation over the 100 trials of experiment (2, 14) was .93. Derivation of (2.14) appears in Section 7.3 of Efron (1982). For an interesting application of almost the same idea to density estimation see Wong (1983), discussed here in Remark B of Section 9.

To summarize, $\hat{\omega}^{(BOOT)}$ is the obvious nonparametric MLE for ω; $\hat{\omega}^{(JACK)}$ is a quadratic approximation to $\hat{\omega}^{(BOOT)}$; and $\hat{\omega}^{(CV)}$ is similar in form and value to $\hat{\omega}^{(JACK)}$. All of this indicates that there is only one basic idea operating here: the substitution of $\hat{F}$ for F in whatever we are trying to estimate, that being $\omega(F)$ in the problem at hand.

3. FIVE SAMPLING EXPERIMENTS

Table 2 reports on five sampling experiments comparing cross-validation, the bootstrap, and several other methods of estimating Err, the true error rate. The first four experiments are (2, 14), described in Section 2, and three simple variations, (2, 20), (5, 14), and (5, 20). Experiment (2, 20) is exactly the same as (2, 14) except that the sample size n is increased from 14 to 20. Each trial of experiment (5, 14) involves n = 14 cases in p = 5 dimensions. The distribution of (T_i, Y_i) is as given at (2.12), except that $T_i \mid Y_i = y_i \sim N_5((2y_i - 1, 0, 0, 0, 0), I)$ a five-dimensional normal distribution. The prediction rule $\eta(t_0, \mathbf{x})$ is Fisher's estimated linear discriminant function, as described in Section 2. Experiment (5, 20) is the same as (5, 14) except that n is increased from 14 to 20. Each of these four experiments comprised 100 trials.

Experiment GG is taken from the Ph.D thesis of Gong (1982). The sample size is n = 20, the prediction dimension p = 4. Predictor vectors T_i have a four-dimensional normal distribution with mean vector 0, all standard deviations equal 1.0, and all correlations equal zero except for corr(T_{i2}, T_{i3}) = .8. Given $T_i = t_i$, the dichotomous response variable y_i equals 1 with probability $1/[1 + \exp -(\alpha + t_i\beta')]$, $\alpha = 0$, $\beta = (1, 2.25, 0, 0)$. The prediction rule $\eta(t, \mathbf{x})$ is based on a forward stepwise logistic regression, using a sequence of hypothesis tests to determine which of the components of t_i to include in making the prediction, and will not be further described here. Experiment GG comprised 171 trials.

We are estimating Err = $\bar{err}$ + op with statistics[1] of the form $\hat{\text{Err}} = \bar{err} + \hat{\omega}$. The mean squared error (MSE) is $E(\hat{\text{Err}} - \text{Err})^2 = E(\hat{\omega} - \text{op})^2$, or

$$\text{MSE} = (E\hat{\omega} - \omega)^2 + \text{var}(\hat{\omega}) - 2\,\text{cov}(\hat{\omega}, \text{op}) + \text{var}(\text{op}). \quad (3.1)$$

Notice that MSE measures how well, on the average, $\hat{\text{Err}}$ estimates Err($\mathbf{x}$, F) *for the given training set* $\mathbf{x}$. In this sense it is a measure of average conditional risk. An un-

[1] The notation $\hat{\omega}$ could be changed to $\hat{op}$, but isn't for reasons given in the last paragraph of Section 2.

conditional measure such as $E[\hat{\text{Err}} - E(\text{Err})]^2$ seems less appropriate.

Table 2 gives the MSE for each method, and also the information needed to compute the individual components on the right side of (3.1). For example, in experiment (2, 14), $\omega = E(\text{op}) = .096$ (100 trials), while the bootstrap estimate $\hat{\omega}^{(\text{BOOT})}$ has expectation .080. We see that the bias term in (3.1) contributes $(.080 - .096)^2 = .000256$, a negligible amount compared with the total MSE of .0179.

The large negative correlation of $\hat{\omega}^{(\text{BOOT})}$ with op, $-.64$ for experiment (2, 14), substantially increases MSE for the ordinary bootstrap. Cross-validation has correlations near zero, but suffers from high values of $\text{var}(\hat{\omega})$. The estimators performing well in Table 2 do so by reducing the negative correlation nearly to zero, without increasing $\text{var}(\hat{\omega})$ much above $\text{var}(\hat{\omega}^{(\text{BOOT})})$. It doesn't seem possible to make the correlation positive; see Remark G of Section 9.

Figure 1 graphically compares the performances of four of the estimators in Table 1. The MSE's are plotted on a relative inefficiency scale,

$$\text{REL} = \frac{\text{MSE} - \text{MSE}^{(\text{IC})}}{\text{MSE}^{(\text{ZERO})} - \text{MSE}^{(\text{IC})}}, \qquad (3.2)$$

where $\text{MSE}^{(\text{IC})}$ is the mean squared error for the ideal constant estimator (2.6), $\hat{\text{Err}} = \bar{err} + \omega$, and $\text{MSE}^{(\text{ZERO})}$ is the mean square error for the zero estimator $\hat{\text{Err}} = \bar{err}$, that is, the apparent error rate. Large numbers are bad here: REL > 100% as for cross-validation in Experiments (2, 20) and GG, indicate estimators worse than $\bar{err}$.

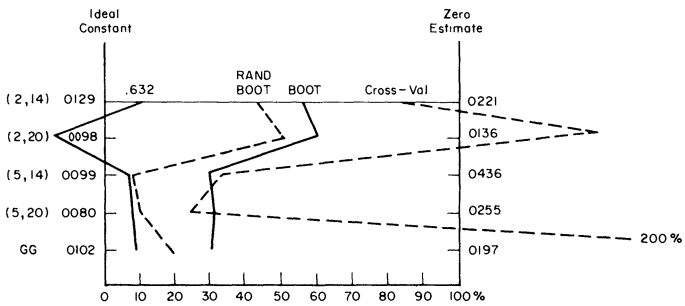

Figure 1. Relative inefficiencies (3.2) for four of the estimators in Table 1; 0% indicates performance equal to the ideal constant estimator, >100% indicates performance worse than the apparent error rate $\bar{err}$.

The small sample sizes, $n = 14$ or 20, are an important factor in the large behavioral differences evident in Table 1 and Figure 1. The equivalent of Figure 1 for $n = 100$ would show all the estimators, doing much better (REL in the range 0–40%), but with the ordinary bootstrap and cross-validation still performing noticeably worst. Efron gives numerical results for all five experiments scaled up to have $n = 100$, Stanford Technical Report #78.

Table 2. Five Sampling Experiments, Comparing Several Different Methods for Estimating Err

	(2,14)	(2,20)	(5,14)	(5,20)	GG
	Exp(Sd Corr) MSE	Exp(Sd Corr) MSE	Exp(Sd Corr) MSE	Exp(Sd Corr) MSE	Exp(Sd Corr) MSE
100 Trials (171 for *GG*)					
True op	096(113)	059(099)	184(099)	130(090)	098(094)
Ideal Constant	096(0 0) 0129	059(0 0) 0099	184(0 0) 0099	130(0 0) 0080	098(0 0) 0088
Zero Correction	0(0 0) 0221	0(0 0) 0134	0(0 0) 0432	0(0 0) 0249	0(0 0) 0184
Cross-Validation	091(073 −15) 0206	067(070 +00) 0148	170(094 −15) 0216	139(070 +03) 0126	113(120 −21) 0280
Bootstrap ($B = 200$)	080(028 −64) 0179	061(020 −47) 0122	103(031 −58) 0210	086(025 −69) 0136	083(022 −57) 0118
Randomized Bootstrap	087(026 −55) 0169	062(020 −38) 0118	147(020 −31) 0129	109(017 −46) 0101	082(023 −28) 0108
Simple Randomized Bootstrap	097(023 −62) 0166	072(019 −51) 0123	157(021 −54) 0133	121(020 −67) 0109	100(020 −49) 0110
Double Bootstrap	097(038 −59) 0195	070(029 −40) 0132	184(054 −57) 0190	114(034 −61) 0132	106(033 −48) 0129
Randomized Double Bootstrap	097(036 −54) 0186	068(029 −43) 0133	186(038 −52) 0152	120(032 −62) 0128	NA NA
$632(\hat{\epsilon}^{(0)} - \bar{err})$	076(035 −09) 0138	059(032 +22) 0095	152(038 −04) 0126	112(035 +02) 0094	080(042 +14) 0097
$\hat{w}^{(0)} = \hat{\epsilon}^{(0)} - \hat{\mu}$	101(034 −56) 0184	071(024 −44) 0128	176(044 −54) 0167	124(030 −53) 0119	107(029 −43) 0120
1000 Trials: *op Err*					
$\bar{err} \rightarrow$	.093 .356 .262	.060 .340 .280	.178 .250 .072	.120 .219 .099	

NOTE: Entry MSE is the mean squared error $E(\hat{Err} - Err)^2$; Exp = $E(\hat{\omega})$, Sd = Standard Dev $(\hat{\omega})$, Corr = correlation $(\hat{\omega},\text{op})$. All bootstrap methods used $B = 200$ bootstrap replications per trial except in experiment *GG*, where $B = 100$.

The next three sections discuss the new estimators appearing in Table 2, some of which clearly outperform cross-validation and the bootstrap.

4. RANDOMIZED BOOTSTRAP

The randomized bootstrap is a particularly simple variant of the ordinary bootstrap appropriate when y is dichotomous. The two versions of the randomized bootstrap appearing in Table 2 performed well.

The empirical probability distribution $\hat{F}$, (2.8), concentrates all of its mass on the n points (t_i, y_i), $i = 1, 2, \ldots, n$. The idea of the randomized bootstrap is to assign some probability mass to the n complementary points $(t_i, \bar{y}_i)$, $\bar{y}_i \equiv 1 - y_i$. Given the training set $\mathbf{x}$, we have in mind some way of assigning probabilities to all $2n$ points (t_i, y_i), $(t_i, \bar{y}_i)$, say

$$\text{Assigned probability on } (t_i, y) = \frac{1}{n}\pi_i(y, \mathbf{x}), \quad (4.1)$$

where $\pi_i(y, \mathbf{x}) + \pi_i(\bar{y}, \mathbf{x}) = 1$. This last condition means that (t_i, y_i) and $(t_i, \bar{y}_i)$ are assigned total probability $1/n$, as with the ordinary bootstrap. For the *simple randomized bootstrap*, line 6 of Table 2,

$$\pi_i(y_i, \mathbf{x}) = .9, \quad \pi_i(\bar{y}_i, \mathbf{x}) = .1. \quad (4.2)$$

(To put it another way, this rule shrinks the maximum likelihood estimates $\pi(y_i, \mathbf{x}) = 1$, $\pi(\bar{y}_i, \mathbf{x}) = 0$ toward the central value .5.)

Let $\hat{F}^{(\mathrm{RAND})}$ be the distribution on $2n$ points given by (4.1). Then the randomized bootstrap estimate of ω is

$$\hat{\omega}^{(\mathrm{RAND})} = E_*\, \mathrm{op}(\mathbf{X}^*, \hat{F}^{(\mathrm{RAND})}), \quad (4.3)$$

where now $\mathbf{X}^*$ is a random sample from $\hat{F}^{(\mathrm{RAND})}$,

$$X_1^*, X_2^*, \ldots, X_n^* \overset{\text{iid}}{\sim} \hat{F}^{(\mathrm{RAND})}, \quad (4.4)$$

and E_* indicates expectation with respect to this probability mechanism.

Defining $N_{i,y}^* = \#\{X_j^* = (t_i, y)\}$ and $N_i^* = N_{i0}^* + N_{i1}^*$, it is possible to express $\mathrm{op}(\mathbf{X}^*, \hat{F}^{(\mathrm{RAND})})$ as a sum of n terms,

$$\mathrm{op}^* = \sum_{i=1}^{n} \frac{1}{n}\{[(2\pi_i(y_i, \mathbf{x}) - 1) - (2N_{i,y_i}^* - N_i^*)] \times Q[y_i, \eta(t_i, \mathbf{X}^*)] + [\pi_i(\bar{y}_i, \mathbf{x}) - N_{i,\bar{y}_i}^*]\}. \quad (4.5)$$

Notice that this reduces to the expression for op* in (2.9) if $\pi_i(y_i, \mathbf{x}) = 1$, $\pi_i(\bar{y}_i, \mathbf{x}) = 0$. Since $E_*[\pi_i(\bar{y}_i, \mathbf{x}) - N_{i,\bar{y}_i}^*] = 0$, (4.5) gives the quite simple expression

$$\hat{\omega}^{(\mathrm{RAND})} = E_* \sum_{i=1}^{n} \frac{1}{n}\{(2\pi_i(y_i, \mathbf{x}) - 1) - (2N_{i,y_i}^* - N_i^*)\} \times Q[y_i, \eta(t_i, \mathbf{X}^*)]. \quad (4.6)$$

Both (4.5) and (4.6) remain valid if y_i is replaced by 1 and $\bar{y}_i$ is replaced by 0 everywhere.

Most often in dichotomous prediction problems, the prediction rule provides a probability assessment $\pi_i(y, \mathbf{x})$ as well as a specific prediction $\eta(t_i, \mathbf{x})$. For example Fisher's estimated linear discriminant is naturally associated with the probability assessment

$$\pi_i(1, \mathbf{x}) = 1/[1 + \exp - (\hat{\alpha} + t_i\hat{\beta}')], \quad (4.7)$$

(see Efron 1975, Sec. 1). Line 5 of Table 1, the randomized bootstrap, refers to the use of (4.7) in (4.1) through (4.4), except that the values $\pi_i(1, \mathbf{x})$ are restricted to lie in the range [.1, .9].

There is an obvious ad hoc component to the choice of the numbers .1, .9 for the two randomized bootstraps. In theory the statistician could make a subjective assessment of the uncertainty in each prediction $\eta(t_i, \mathbf{x})$, in order to assign $\pi_i(y_i, \mathbf{x})$ and $\pi_i(\bar{y}_i, \mathbf{x})$. In the sampling experiments the exact assignments seemed less important than keeping them away from 0 and 1. In particular the simple method (4.2) caused little bias (and as a matter of fact helped correct the bias in the ordinary bootstrap), and gave almost as much improvement as the more complicated method based on (4.7).

It is obvious that $\hat{F}$ can be a poor estimate of F, particularly if we know that F is smooth. Using $\hat{F}^{(\mathrm{RAND})}$ in place of $\hat{F}$ is a form of smoothing. The smoothing is carried out entirely in the y direction. This is handy since in real applications t may be very complicated, having high dimensionality, censored components, missing values, qualitative and quantitative components, and so on.

5. THE DOUBLE BOOTSTRAP

The bootstrap estimate $\hat{\mathrm{Err}}^{(\mathrm{BOOT})}$ was obtained in Section 2 by (a) writing Err as $\bar{err}$ + (Err − $\bar{err}$) where $\bar{err} \equiv S(\mathbf{X})$ is an observable statistic and (Err − $\bar{err}$) $\equiv R(\mathbf{X}, F)$ is a random variable, and (b) estimating Err by $S + E_* R^*$, where $E_* R^*$ is the bootstrap expectation $E_* R(\mathbf{X}^*, \hat{F})$. There is no obvious theoretical reason for the choice $S = \bar{err}$. For any statistic S we could write Err $= S +$ (Err − S) and estimate Err by $S + E_*(\mathrm{Err} - S)^*$. For example, choosing $S = 0$ gives the estimate $E_*\mathrm{Err}^*$. It will turn out, in Section 8, that this is a poor estimate of Err.

This section concerns bootstrapping the bootstrap. We take $S = \bar{err} + \hat{\omega}^{(\mathrm{BOOT})} = \hat{\mathrm{Err}}^{(\mathrm{BOOT})}$, what we have called the ordinary bootstrap estimate of Err, write

$$\mathrm{Err} = \hat{\mathrm{Err}}^{(\mathrm{BOOT})} + (\mathrm{Err} - \hat{\mathrm{Err}}^{(\mathrm{BOOT})}) \equiv S + R, \quad (5.1)$$

and estimate Err by the "double bootstrap" estimate

$$\hat{\mathrm{Err}}^{(\mathrm{DOUB})} = S + E_* R^*. \quad (5.2)$$

One motivation for doing so is the downward bias of the ordinary bootstrap evident in Table 2, in particular for experiment (5, 14). If $\hat{\mathrm{Err}}^{(\mathrm{BOOT})}$ is an underestimator of Err, then we can correct it by bootstrapping, as in (5.1), in the same spirit as we originally corrected $\bar{err}$. (Sec. 7 discusses the downward bias of the bootstrap.)

We can rewrite (5.1), (5.2) as

$$\begin{aligned}\hat{\mathrm{Err}}^{(\mathrm{DOUB})} &= (\bar{\mathrm{err}} + \hat{\omega}^{(\mathrm{BOOT})}) \\ &\quad + E_*(\mathrm{Err} - \bar{err} - \hat{\omega}^{(\mathrm{BOOT})})^* \\ &= (\bar{\mathrm{err}} + \hat{\omega}^{(\mathrm{BOOT})}) + E_*(\mathrm{Err} - \bar{err})^* \\ &\quad - E_*(\hat{\omega}^{(\mathrm{BOOT})})^* \\ &= \bar{err} + 2\hat{\omega}^{(\mathrm{BOOT})} - E_*(\hat{\omega}^{(\mathrm{BOOT})})^*, \end{aligned} \tag{5.3}$$

the last line following from $E_*(\mathrm{Err} - \bar{err})^* = E_*(\mathrm{op})^* = \hat{\omega}^{(\mathrm{BOOT})}$. Another way to say (5.3) is that $\hat{\omega}^{(\mathrm{DOUB})} = 2\hat{\omega}^{(\mathrm{BOOT})} - E_*(\hat{\omega}^{(\mathrm{BOOT})})^*$. Assuming that $\hat{\omega}^{(\mathrm{BOOT})}$ has already been computed, we need to calculate $E_*(\hat{\omega}^{(\mathrm{BOOT})})^*$ in order to find $\hat{\omega}^{(\mathrm{DOUB})}$ and $\hat{\mathrm{Err}}^{(\mathrm{DOUB})}$. This looks as if it involves two layers of bootstrapping, perhaps B^2 recomputations of the rule η, which would be $200^2 = 40{,}000$ recomputations in our case. It turns out that a total of $2B$ recomputations, 400 in our case, suffice, thanks to a Monte Carlo "swindle," so that the double bootstrap is computationally feasible, as well as properly named.

It is shown in the Appendix that

$$E_*(\hat{\omega}^{(\mathrm{BOOT})})^* = E_{**}\left\{\sum_{i=1}^{n} e(P_i^{**})\, Q[y_i, \eta(t_i, \mathbf{X}^{**})]\right\}, \tag{5.4}$$

where $\mathbf{X}^{**}$ is a second-level bootstrap sample,

$$X_1^{**}, X_2^{**}, \ldots, X_n^{**} \overset{\text{iid}}{\sim} \hat{F}^*, \tag{5.5}$$

$\hat{F}^*$ indicating the empirical distribution function of a first-level bootstrap sample $X_1^*, \ldots, X_n^*$. The quantities $\mathbf{P}_i^{**}$ are the proportions of $\mathbf{X}^{**}$ on the various original cases x_i,

$$P_i^{**} = \#\{X_j^{**} = x_i\}/n, \tag{5.6}$$

and the function $e(P_i^{**})$ is given to a good approximation as follows:

$nP_i^{**} =$	0	1	2	3	4	5	6
$ne(P_i^{**}) =$	.37	.37	−.36	−1.02	1.65	−2.30	−2.97

(5.7)

The expectation E_* in (5.4) is with respect to the marginal distribution of $\mathbf{X}^{**}$, first taking a bootstrap sample as at (2.9), then constructing its empirical distribution function $\hat{F}^*$, and finally drawing $\mathbf{X}^{**}$ as at (5.5). The training set $\mathbf{x}$ is considered fixed during this entire process.

Expression (5.4) can be approximated by Monte Carlo just as was (2.9) for the ordinary bootstrap: (a) A sequence of independent double bootstrap vectors $\mathbf{x}^{**b}$, $b = 1, 2, \ldots, B$, is obtained, each one generated by the process described in the previous paragraph; (b) For each $\mathbf{x}^{**b}$ the rule $\eta(t, \mathbf{x}^{**b})$ is constructed and; (c) $(\hat{\omega}^{(\mathrm{BOOT})})^{*b} = \sum_{i=1}^{n} e(P_i^{**b})\, Q[y_i, \eta(t_i, \mathbf{x}^{**b})]$ is evaluated. Then

$$E_*(\hat{\omega}^{(\mathrm{BOOT})})^* \doteq \frac{1}{B}\sum_{b=1}^{B} (\hat{\omega}^{(\mathrm{BOOT})})^{*b}, \tag{5.8}$$

with increasing accuracy as $B \to \infty$. For the entries in line 7 of Table 2, $B = 200$ in (5.8), taken in addition to the 200 replications used in calculating $\hat{\omega}^{(\mathrm{BOOT})}$.

The double bootstrap nicely corrects the bias in the ordinary bootstrap, as can be seen by comparing $E(\hat{\omega}^{(\mathrm{DOUB})})$ from line 7 of Table 2 with the actual values of ω. In terms of MSE its performance is about the same as the ordinary bootstrap.

Line 8 of Table 2 refers to a double bootstrap version of the simple randomized bootstrap defined at (4.2). This computation requires a formula like (5.4) referring to $(\hat{\omega}^{(\mathrm{RAND})})^*$, from (4.6), rather than to $(\hat{\omega}^{(\mathrm{BOOT})})^*$. No more will be said about the randomized double bootstrap here, except that it is not particularly difficult to compute and performs slightly better than the double bootstrap in Table 2.

6. THE .632 ESTIMATOR

The estimator of line 9, Table 2, $.632(\hat{\epsilon}^{(0)} - \bar{err})$, called "the .632 estimator" for short, was a clear winner in the sampling experiments. This section defines and motivates the .632 estimator. Unfortunately the motivation is weak, leaving open the possibility that the estimator's success here was a fluke. (It has continued to perform best in some additional, rather different, sampling experiments described in Gong 1982.)

Why does $\bar{err}$ tend to underestimate Err? Another answer, beside that it obviously does, can be given in terms of the distance of the point to be predicted from the training set: $\bar{err}$ is an error rate for points zero distance from the training set $\mathbf{x}$; Err is the expected error rate for a new point X_0 which may lie some distance away from $\mathbf{x}$. If the error rate of the prediction rule increases as the point being predicted moves away from $\mathbf{x}$, then $\bar{err}$ will underestimate Err.

To make this argument more concrete suppose that for each x and Δ the set $S(x, \Delta)$ is a neighborhood of x having probability content Δ under the true distribution F,

$$\mathrm{Prob}_F\{X_0 \in S(x, \Delta)\} = \Delta. \tag{6.1}$$

The neighborhoods $S(x, \Delta)$ are assumed to grow smaller as Δ decreases, going to the single point $\{x\}$ as $\Delta \to 0$. Define

$$\delta(x_0, \mathbf{x}) = \inf_{\Delta}\left\{x_0 \in \bigcup_{i=1}^{n} S(x_i, \Delta)\right\}, \tag{6.2}$$

δ is large or small as x_0 is far from or near to the nearest point in the training set $\mathbf{x}$. Denote $\bar{Q}(\Delta)$ by the following:

$$\bar{Q}(\Delta) = E\{Q(X_0, \mathbf{X}) \mid \delta(X_0, \mathbf{X}) = \Delta\}, \tag{6.3}$$

where $Q(X_0, \mathbf{X}) = Q[Y_0, \eta(T_0, \mathbf{X})]$, so $\bar{Q}(\Delta)$ is the ex-

pected error rate given that X_0 is distance Δ from the nearest point in the training set.

The curve marked Actual in Figure 2 shows $\bar{Q}(\Delta)$ as a function of Δ for experiment (2, 20). The neighborhoods $S(x, \Delta)$, $x = (t, y)$, were taken as circles in the t space,

$$S((t, y), \Delta) = \{x_0 = (t_0, y_0): y_0 = y \text{ and } \| t_0 - t \| \leq r_{x,\Delta}\}, \quad (6.4)$$

with $r_{x,\Delta}$ chosen to satisfy (6.1). The set $\cup_{i=1}^n S(x_i, \Delta)$ is a union of circles in the planes $y = 0$ and $y = 1$, each circle centered at x_i and having radius roughly inversely proportional to the density of model (2.12) at x_i.

As expected, $\bar{Q}(\Delta)$ is an increasing function of Δ. Notice that

$$\bar{Q}(0) = E\,\bar{err}, \quad (6.5)$$

$= .280$ for experiment (2, 20). Relation (6.5) is a consequence of the way we defined $\bar{Q}(\Delta)$ and $S(x_i, \Delta)$. As $\Delta \to 0$, the conditional distribution of a point X_0 in $\cup_{i=1}^n S(x_i, \Delta)$ approaches the empirical distribution $\hat{F}$, (2.8), so $\bar{Q}(\Delta) \to E\,\bar{err}$.

Let $D(\Delta)$ be the cumulative distribution function of $\delta(X_0, \mathbf{X})$, $D(\Delta) = \text{Prob}\{\delta(X_0, \mathbf{X}) \leq \Delta\}$ for $0 \leq \Delta \leq 1$. Then the expected true error rate is $E(\text{Err}) = \int_0^1 \bar{Q}(\Delta)\, dD(\Delta)$ and, since $E\,\bar{err} = \bar{Q}(0)$,

$$\omega = \int_0^1 [\bar{Q}(\epsilon) - \bar{Q}(0)]dD(\Delta) \quad (6.6)$$

is the expected optimism. Looking at Figure 2, and at the insert, which shows $D(\Delta)$, one can see that most of $\omega = .060$ for experiment (2, 20) comes from values of Δ in the range (0, .1).

The curve marked Bootstrap in Figure 2 is

$$\bar{Q}^{(B)}(\Delta) = E\{Q(X_0^*, \mathbf{X}^*) \mid \delta(X_0^*, \mathbf{X}^*) = \Delta\}, \quad (6.7)$$

the expectation of $Q(X_0^*, \mathbf{X}^*) = Q[Y_0^*, \eta(T_0^*, \mathbf{X}^*)]$ given that the independent point $X_0^* \sim \hat{F}$ is distance Δ away from the bootstrap training set $X_1^*, X_2^*, \ldots, X_n^*$ $\overset{\text{iid}}{\sim} \hat{F}$ used to construct $\eta(t, \mathbf{X}^*)$; E indicates marginal expectation over the choice of $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F$ and then $X_0^*, X_1^*, X_2^*, \ldots, X_n^* \overset{\text{iid}}{\sim} \hat{F}$. *Notice the agreement between* $\bar{Q}(\Delta)$ *and* $\bar{Q}^{(B)}(\Delta)$. A point say $\Delta = .05$ away from a bootstrap training set has the same expected probability of misclassification as a point .05 away from an actual training set.

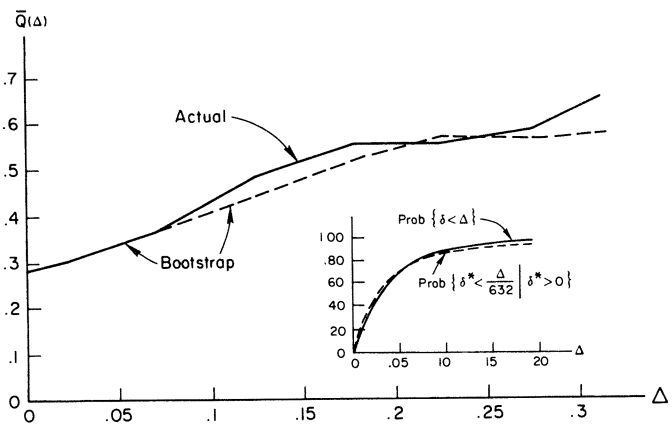

Figure 2. The actual and bootstrap expected error rates as a function of the distance from X_0 to the nearest point in the training set, experiment (2, 20). The insert shows the actual and bootstrap cumulative distribution function for the distance to the nearest point.

There is one big difference between the actual and the bootstrap situations: the distribution of the distance δ. The bootstrap distance $\delta(X_0^*, \mathbf{X}^*)$ has a high probability of equaling zero,

$$\text{Prob}\{\delta(X_0^*, \mathbf{X}^*) = 0\} = 1 - (1 - (1/n))^n \doteq 1 - e^{-1} = .632, \quad (6.8)$$

this being the probability that X_0^* falls on one of the support points of $\hat{F}$ occurring in the bootstrap sample $\mathbf{X}^*$ (i.e., that $X_0^* = x_i$ with $P_i^* > 0$). Given that $\delta(X_0^*, \mathbf{X}^*) > 0$ (i.e., $X_0^* = x_i$ with $P_i^* = 0$) we show in the Appendix that $\delta(X_0^*, \mathbf{X}^*)$ is roughly distributed as $\delta(X_0, \mathbf{X})/.632$,

$$\text{Prob}\left\{\delta(X_0^*, \mathbf{X}^*) > \frac{\Delta}{.632} \;\middle|\; \delta(X_0^*, \mathbf{X}^*) > 0\right\} \doteq \text{Prob}\{\delta(X_0, \mathbf{X}) > \Delta\}. \quad (6.9)$$

All the probabilities occurring in (6.8), (6.9) are marginal over the choice of $\mathbf{X}$ and $\mathbf{X}^*$, X_0^*.

For a given training set $\mathbf{x}$, let $\hat{\epsilon}^{(0)}$ be the bootstrap expected error rate at those points x_i not occurring in the bootstrap sample,

$$\hat{\epsilon}^{(0)} = E_*\{Q(X_0^*, \mathbf{X}^*) \mid X_0^* = x_i \text{ with } P_i^* = 0\}. \quad (6.10)$$

In any one trial of experiment (2, 20), $\hat{\epsilon}^{(0)}$ was computed by (a) examining all 4,000 entries (b, i), $b = 1, \ldots, 200$, $n = 1, \ldots, 20$; (b) looking at the approximately 36.8 percent of the entries having $P_i^{*b} = 0$; and (c) setting $\hat{\epsilon}^{(0)}$ equal to the observed error rate for these entries,

$$\hat{\epsilon}^{(0)} = \frac{\sum_{(b,i):P_i^{*b}=0} Q(x_i, \mathbf{x}^{*b})}{\#\{(b, i): P_i^{*b} = 0\}}. \quad (6.11)$$

Expression (6.11) approaches definition (6.10) as $B \to \infty$. Section 8 discusses $\hat{\epsilon}^{(0)}$ and also the conditional error rates for values of P_i^* besides 0.

The .632 estimator is $\hat{\omega}^{(.632)} = .632(\hat{\epsilon}^{(0)} - \bar{err})$. This gives the Err estimate

$$\hat{\text{Err}}^{(.632)} = \bar{err} + \hat{\omega}^{(.632)} = .368\,\bar{err} + .632\,\hat{\epsilon}^{(0)}. \quad (6.12)$$

The weights .368 and .632 are suggested by (6.9). The points X_0^* contributing to $\hat{\epsilon}^{(0)}$ are about 1/.632 too far out along the δ axis in Figure 2, whereas the points contributing to $\bar{err}$ are at $\delta = 0$. The weighted average (6.12) therefore involves points with about the right expected value of δ. If $\bar{Q}(\Delta) \doteq \bar{Q}^{(B)}(\Delta)$, and both are roughly linear in Δ, this makes $\hat{\text{Err}}^{(.632)}$ roughly unbiased for Err. In fact $\hat{\text{Err}}^{(.632)}$ displays a moderate downward bias in Table 2.

The reason for its remarkably low MSE is its lack of negative correlation with op, term three of (3.1). This is discussed further in Section 9, Remark G.

It will turn out in Section 8 that $\hat{\epsilon}^{(0)}$ is almost the same as $\hat{\text{Err}}^{(\text{HCV})}$, the estimated error rate based on a cross-validation that leaves out half of the sample at a time. This means that a good approximation $\hat{\text{Err}}^{(.632)}$ can be written as $.368\ \bar{err} + .632\ \hat{\text{Err}}^{(\text{HCV})}$.

7. ANOVA DECOMPOSITIONS

This section describes the ANOVA decomposition (Efron and Stein 1981) as it applies to $Q(X_0, \mathbf{X})$ and $Q(X_0^*, \mathbf{X}^*)$. It will give us a better theoretical basis for understanding the prediction problem, particularly the orders of magnitude involved, and the relationships between various estimates of err. The calculations are carried through formally and do not constitute valid asymptotic theorems, but nevertheless give quite accurate predictions in our numerical studies.

The ANOVA decomposition for $Q(x_0, \mathbf{x}) = Q[y_0, \eta(t_0, \mathbf{x})]$ is

$$\begin{aligned} Q(x_0, \mathbf{x}) &= \mu_{x_0} + \frac{1}{n}\sum_i \alpha_{x_0}(x_i) \\ &\quad + \frac{1}{n^2}\sum_{i<i'} \beta_{x_0}(x_i, x_{i'}) \\ &\quad + \frac{1}{n^3}\sum_{i<i'<i''} \gamma_{x_0}(x_i, x_{i'}, x_{i''}) + \ldots . \end{aligned} \tag{7.1}$$

The quantities μ_{x_0}, $\alpha_{x_0}(x_i)$, $\beta_{x_0}(x_i, x_{i'})$, and so on correspond to the grand mean, main effects, second-order interactions, and so on in the standard ANOVA decomposition of an n-way table:

$$\begin{aligned} \mu_{x_0} &= E\, Q(x_0, \mathbf{X}), \\ \alpha_{x_0}(x_i) &= n[E\{Q(x_0, \mathbf{X}) \mid X_i = x_i\} - \mu_{x_0}], \\ \beta_{x_0}(x_i, x_{i'}) &= n^2[E\{Q(x_0, \mathbf{X}) \mid X_i = x_i, X_{i'} = x_{i'}\} \\ &\quad - \alpha_{x_0}(x_i) - \alpha_{x_0}(x_{i'}) + \mu_{x_0}] \end{aligned} \tag{7.2}$$

and so on; x_0 is fixed in these expectations, with $\mathbf{X}$ random, subject to the indicated conditioning statements. The sums in (7.1) are over all integers $i, i', i'', \ldots$, in the range $1, 2, \ldots, n$, subject to ordering conditions as indicated. The right side of (7.1) terminates with a single nth order interaction term.

The factors of n in (7.2) give α, β, and so on, nondegenerate limiting distributions. As n grows large $\alpha_{x_0}(X_i)$ approaches the *influence function* for $Q(x_0, \mathbf{X})$, $\beta_{x_0}(X_i, X_{i'})$ approaches the second-order influence function, and so on. See Hampel (1974) for a good discussion of influence function ideas.

Expansion (7.1) is an orthogonal decomposition of $Q(x_0, \mathbf{x})$. The quantities $\alpha_{x_0}(X_i)$, $\beta_{x_0}(X_i, X_{i'})$, and so on have expectation zero and are mutually uncorrelated. In fact each of them has conditional expectation zero when conditioned upon all but one of its defining X_i,

$$E\,\alpha_{x_0}(X_i) = 0,\ E\,\beta_{x_0}(x_i, X_{i'}) = 0,$$

$$E\,\gamma_{x_0}(x_i, x_{i'}, X_{i''}) = 0 \tag{7.3}$$

and so on, $i < i' < i''$, only capitalized X's being random in these expectations.

Many quantities related to the prediction problem have simple expressions in terms of the ANOVA decomposition. As a first example we have

$$\omega = -E\,\alpha_{X_1}(X_1)/n. \tag{7.4}$$

This follows from

$$\begin{aligned} E\ \text{Err} &= E\, Q(X_0, \mathbf{X}) = E\,\mu_{X_0}, \\ E\,\bar{err} &= E\,\frac{1}{n}\sum Q(X_i, \mathbf{X}) = E\, Q(X_1, \mathbf{X}) \\ &= E\,\mu_{X_1} + \frac{1}{n}E\,\alpha_{X_1}(X_1), \end{aligned} \tag{7.5}$$

all other terms such as $E\,\beta_{X_1}(X_1, X_2)$ equaling zero by (7.3), so that $\omega = E(\text{Err} - \bar{err}) = -E\,\alpha_{X_1}(X_1)/n$.

As another example consider the following variant of cross-validation. Let $\mathbf{x}_{(i,j)}$ represent the modified training set with x_i removed and x_j included twice, and let $Q_{(i,j)} = Q(x_i, \mathbf{x}_{(i,j)})$. Then define

$$\hat{\omega}^{(\text{CV}+)} = \frac{1}{n(n-1)}\sum_{i\neq j} Q_{(i,j)} - \bar{err}. \tag{7.6}$$

This is a version of $\hat{\omega}^{(\text{CV})}$, (2.7), for which all the modified training sets $\mathbf{x}_{(i,j)}$ have sample size n. An easy calculation, very much like (7.4), shows that

$$E\,\hat{\omega}^{(\text{CV}+)} - \omega = (E\,\beta_{X_0}(X_1, X_1))/n^2. \tag{7.7}$$

(Because $\hat{\omega}^{(\text{CV})}$ involves samples of size $n-1$, (7.1) cannot be applied to it; see Efron and Stein 1981.) Letting $\hat{\text{Err}}^{(\text{CV}+)} = \bar{err} + \hat{\omega}^{(\text{CV}+)}$, (7.7) gives $E(\hat{\text{Err}}^{(\text{CV}+)} - err) = E\,\beta_{X_0}(X_1, X_1)/n^2$, compared with $E(\text{Err} - \bar{err}) = -E\,\alpha_{X_1}(X_1)/n$ from (7.4). *Cross-validation reduces bias of the error estimate from $O(1/n)$ to $O(1/n^2)$.*

There is an analog of decomposition (7.1) that applies to the bootstrap quantity $Q(x_0^*, \mathbf{x}^*) = Q[y_0^*, \eta(t_0^*, \mathbf{x}^*)]$:

$$\begin{aligned} Q(x_0^*, \mathbf{x}^*) &= \hat{\mu}_{x_0^*} + \frac{1}{n}\sum_i \hat{\alpha}_{x_0^*}(x_i^*) \\ &\quad + \frac{1}{n^2}\sum_{i<i'} \hat{\beta}_{x_0^*}(x_i^*, x_{i'}^*) + \ldots \end{aligned} \tag{7.8}$$

where

$$\begin{aligned} \hat{\mu}_{x_0^*} &= E_*\, Q(x_0^*, \mathbf{X}^*), \\ \alpha_{x_0^*}(x_i^*) &= E_*\{Q(x_0^*, \mathbf{X}^*) \mid X_i^* = x_i^*\} - \mu_{x_0^*} \end{aligned} \tag{7.9}$$

and so on, as in (7.2). The bootstrap analogy of (7.3) is

$$E_*\,\hat{\alpha}_{x_0^*}(X_i^*) = 0, \quad E_*\,\hat{\beta}_{x_0^*}(x_1^*, X_2^*) = 0 \tag{7.10}$$

and so on. The random variables $X_0^*, X_1^*, \ldots, X_n^*$ take

their values in the training set $\{x_1, x_2, \ldots, x_n\}$. We will use the shortened notation $\hat{\alpha}_{j_0}(j_1)$ for

$$\hat{\alpha}_{X_0^* = x_{j_0}}(X_1^* = x_{j_1}),$$

likewise $\hat{\beta}_{j_0}(j_1, j_2)$, and so on. Then (7.10) becomes

$$\frac{1}{n}\sum_{j_1=1}^{n} \hat{\alpha}_{j_0}(j_1) = 0, \quad \frac{1}{n}\sum_{j_2=1}^{n} \hat{\beta}_{j_0}(j_1, j_2) = 0, \quad (7.11)$$

and so on, the first relationship holding for all j_0, the second for all j_0, j_1, $1 \le j_0, j_1 \le n$.

We immediately get

$$\hat{\omega}^{(\mathrm{BOOT})} = -E_* \hat{\alpha}_{X_1^*}(X_1^*)/n = -\sum_{j=1}^{n} \hat{\alpha}_j(j)/n^2, \quad (7.12)$$

the proof being the same as for (7.4).

We can use (7.4), (7.12) to analyze the downward bias of $\hat{\omega}^{(\mathrm{BOOT})}$ as an estimator of ω noticed in Table 2 (details given in the Appendix):

$$E\,\hat{\omega}^{(\mathrm{BOOT})} - \omega = \frac{1}{n^2}[E\,\alpha_{X_1}(X_1) - E\,\beta_{X_1}(X_1, X_1) + E\,\beta_{X_0}(X_1, X_1) - \tfrac{1}{2}E\,\gamma_{X_0}(X_0, X_1, X_1)] + O\left(\frac{1}{n^3}\right). \quad (7.13)$$

Equation (7.13) shows that $\hat{\omega}^{(\mathrm{BOOT})}$, like $\hat{\omega}^{(\mathrm{CV}+)}$, estimates the $O(1/n)$ quantity ω with expected error $O(1/n^2)$. Comparing (7.13) and (7.7) shows that $\hat{\omega}^{(\mathrm{BOOT})}$ has three extra terms in the $O(1/n^2)$ expression, two of which turn out to be negative in our experiment.

Now we will evaluate the terms in (7.13). Let $\mathbf{X}_{(1,2)} = (X_2, X_2, X_3, \ldots, X_n)$. We already know that $E\,\alpha_{X_1}(X_1)/n^2 = -\omega/n$. Expressions (7.1), (7.3) give

$$E\,\beta_{X_0}(X_1, X_1)/n^2 = E\,Q(X_0, \mathbf{X}_{(1,2)}) - E\,Q(X_0, \mathbf{X}),$$

$$E\,\beta_{X_1}(X_1, X_1)/n^2 = E\,Q(X_2, \mathbf{X}_{(1,2)}) - E\,Q(X_1, \mathbf{X}) + E\,Q(X_0, \mathbf{X}),$$

$$E\,\gamma_{X_0}(X_0, X_1, X_1)/2n^2 = \left(\frac{n}{2}\right)[E\,Q(X_3, \mathbf{X}_{(1,2)}) - E\,Q(X_0, \mathbf{X}_{(1,2)}) - E\,Q(X_1, \mathbf{X}) + E\,Q(X_0, \mathbf{X})]. \quad (7.14)$$

Table 3 shows the components of $E\,\hat{\omega}^{(\mathrm{BOOT})} - \omega$ for experiments (5, 14), (5, 20), (2, 14) and, partially, (2, 20). These were obtained by Monte Carlo evaluation of the $E\,Q$ terms in (7.14). The sums (7.14) compare well with the actual biases $E\,\hat{\omega}^{(\mathrm{BOOT})} - \omega$, using 1,000 trials.

The first component $E\,\alpha_{X_1}(X_1)/n^2$ is negative, as we expect it to be since according to (7.4) it equals $-\omega/n$. (The modified estimator $n/(n-1)\,\hat{\omega}^{(\mathrm{BOOT})}$ has bias expression (7.13) except with the first component missing.) The third component $E\,\beta_{X_0}(X_1, X_1)/n^2$ is positive but small. This is also expected, from either (7.7) or the first line of (7.14).

The second component, $-E\,\beta_{X_1}(X_1, X_1)/n^2$, is negative and large. Its negativity amounts to a convexity relationship in the second line of (7.14): $E\,Q(X', \mathbf{X})$ is a convex decreasing function of the number of times, zero, once, or twice, that X' appears in the training set $\mathbf{X}$. We might say that Q is "deletion sensitive" in this case. All of our experiments were deletion sensitive, but artificial examples can be constructed going the other way. It is *not* a theorem that $E\,\hat{\omega}^{(\mathrm{BOOT})} - \omega < 0$, though that seems to be the usual case. In highly overfitted situations, where X' being in the training set even once makes $E\,Q(X', \mathbf{X})$ nearly zero, we expect $-E\,\beta_{X_1}(X_1, X_1)/n^2$ to be strongly negative because of (7.14). Experiment (5, 14) is a good example of this effect.

The fourth component, $-E\,\gamma_{X_0}(X_0, X_1, X_1)/2n^2$, is positive and large, though not as large as the second. The last line of (7.14) suggests that this component will always be positive in highly overfitted situations.

8. REPETITION ERROR RATES

The .632 estimator of Section 6 involves $\hat{\epsilon}^{(0)}$, the bootstrap error rate for cases having bootstrap weight zero. This section concerns the bootstrap error rates $\hat{\epsilon}^{(h)}$ for bootstrap weights $P_i^* = h/n$, $h = 0, 1, 2, \ldots, .$ The main result is a theorem relating $\hat{\epsilon}^{(h)}$ to the ANOVA expansion of Section 7. Among other things, this gives further information on how cross-validation relates to the bootstrap, and an improved Monte Carlo method for calculating the bootstrap. All proofs are deferred until the Appendix.

For a given training set $\mathbf{x}$, the hth repetition error rate $\hat{\epsilon}^{(h)}$ is defined to be the bootstrap error rate for values of

Table 3. Components of Bias for $\hat{\omega}^{(\mathrm{BOOT})}$, (7.13) (figures in parentheses are standard errors)

Exper.	$\frac{E\alpha_{X_1}(X_1)}{n^2}$	$\frac{-E\beta_{X_1}(X_1,X_1)}{n^2}$	$\frac{E\beta_{X_0}(X_1,X_1)}{n^2}$	$\frac{-E\gamma_{X_0}(X_0,X_1,X_1)}{2n^2}$	Sum (7.13)	$E\hat{\omega}^{(\mathrm{BOOT})} - \omega$ (1000 trials)
(5,14)	−.0127	−.1255	+.0083	+.0585	−.071	−.072
	(.0002)	(.0006)	(.0005)	(.0052)	(.006)	(.003)
(5,20)	−.0060	−.0639	+.0046	+.0355	−.030	−.033
	(.0001)	(.0009)	(.0002)	(.0092)	(.006)	(.003)
(2,14)	−.0066	−.0231	+.0047	+.0174	−.008	−.013
	(.0003)	(.0005)	(.0004)	(.0019)	(.002)	(.004)
(2,20)	−.0030	−.0106	+.0037	NA	NA	.000
	(.0002)	(.0009)	(.0004)			(.004)

the predicted point X_0^* equaling an x_i with bootstrap weight h/n,

$$\hat{\epsilon}^{(h)} = E_*\{Q(X_0^*, \mathbf{X}^*) \mid X_0^* = x_i \text{ with } P_i^* = h/n\}. \tag{8.1}$$

Usually $\hat{\epsilon}^{(h)}$ must be calculated by Monte Carlo as described following (6.10),

$$\hat{\epsilon}^{(h)} = \frac{\sum_{(b,i):P_i^{*b}=h/n} Q(x_i, \mathbf{x}^{*b})}{\#\{(b,i):P_i^{*b} = h/n\}}. \tag{8.2}$$

Figure 3 shows the average values of $\hat{\epsilon}^{(0)}$, $\hat{\epsilon}^{(1)}$, $\hat{\epsilon}^{(2)}$ in our five experiments. The decreasing convex nature of these plots demonstrates the deletion sensitivity mentioned in Section 7.

In the bootstrap ANOVA expression (7.8) let

$$\hat{\mu}_i = \hat{\mu}_{X^*_0 = x_i} \quad \text{and} \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_i. \tag{8.3}$$

Also define

$$\hat{A} = \frac{\sum_{i=1}^{n} \hat{\alpha}_i(i)}{n}\left[-\frac{1}{n-1}\right],$$

$$\hat{B} = \frac{\sum_{i=1}^{n} \hat{\beta}_i(i,i)}{n}\frac{\binom{n}{2}}{n^2}\left[-\frac{1}{n-1}\right],$$

$$\hat{C} = \frac{\sum_{i=1}^{n} \hat{\gamma}_i(i,i,i)}{n}\frac{\binom{n}{3}}{n^3}\left[-\frac{1}{n-1}\right]^3, \tag{8.4}$$

and so on.

Theorem. The repetition error rates $\hat{\epsilon}^{(h)}$ are linear combinations of $\hat{\mu}, \hat{A}, \hat{B}, \hat{C}, \ldots,$

$$\hat{\epsilon}^{(h)} = \hat{\mu} + \lambda_1^{(h)}\hat{A} + \lambda_2^{(h)}\hat{B} + \lambda_3^{(h)}\hat{C} + \ldots, \tag{8.5}$$

where the constants $\lambda_j^{(h)}$ are given by

$$\lambda_j^{(h)} = E[-(n-1)]^{H_j}, \tag{8.6}$$

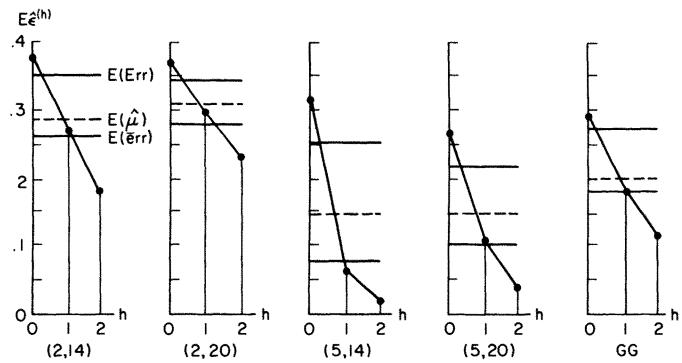

Figure 3. Repetition error rates for the five simulation experiments, h = 0, 1, 2, averaged over all trials in the experiments. Also shown are averages for Err, ērr, and μ̂.

Table 4. Some of the Coefficients $\lambda_j^{(h)}$ Appearing in (8.5)

		1 A	2 B	3 C	···	j
0	$\hat{\epsilon}^{(0)} - \hat{\mu}$	1	1	1	···	1
1	$\hat{\epsilon}^{(1)} - \hat{\mu}$	0	−1	−2	···	$(-j-1)$
2	$\hat{\epsilon}^{(2)} - \hat{\mu}$	−1	$-\frac{n-3}{n-1}$	$\frac{n+5}{n-1}$		
⋮	⋮	⋮				
h	$\hat{\epsilon}^{(h)} - \hat{\mu}$	$-(h-1)$				
	$\hat{\omega}^{(\text{BOOT})}$	$\frac{n-1}{n}$	0	0	···	0

H_j being the number of red balls in j draws without replacement from a population of h red balls and $(n-h)$ black balls.

Some of the coefficients $\lambda_j^{(h)}$ are given in Table 4. The coefficients for $\hat{\omega}^{(\text{BOOT})}$ are based on (7.12), which says that $\hat{\omega}^{(\text{BOOT})} = ((n-1)/n)\hat{A}$. The terms $\hat{\mu}, \hat{A}, \hat{B}, \hat{C}, \ldots,$ are in declining order of magnitude $O_p(1)$, $O_p(1/n)$, $O_p(1/n^2)$, $O_p(1/n^3)$, ... The right side of (8.5) has $n+1$ terms, and if all of these are included (8.5) is exactly true. Notice that the theorem applies to a single training set $\mathbf{x}$, and not just to expectations over random $\mathbf{X}$.

A wide class of interesting Err estimates can be obtained from the $\hat{\epsilon}^{(h)}$. As a first example consider the bootstrap estimate. Define

$$p_n^{(h)} \equiv \binom{n}{h}\frac{(n-1)^{n-h}}{n^n} = \text{Prob}_*\left\{P_i^* = \frac{h}{n}\right\},$$

$h = 0, 1, 2, \ldots, n$, the last equality following from (2.14). It will turn out that

$$\hat{\omega}^{(\text{BOOT})} = \sum_{h=0}^{n} p_n^{(h)}(1-h)\hat{\epsilon}^{(h)}. \tag{8.7}$$

If $\hat{\epsilon}^{(h)}$ is nearly linear in h, as in experiment (2, 20), say $\hat{\epsilon}^{(h)} \doteq \hat{c}_0 - h\hat{c}_1$, then (8.7) gives $\hat{\omega}^{(\text{BOOT})} \doteq ((n-1)/n)\hat{c}_1$, so $\hat{\omega}^{(\text{BOOT})}$ is the negative slope of the repetition plot, times $(n-1)/n$.

Formula (8.7) gives an improved way to calculate $\hat{\omega}^{(\text{BOOT})}$. First calculate the $\hat{\epsilon}^{(h)}$ as in (8.2), and then combine them as in (8.7). If the number of bootstrap replications B is small this method can be quite a bit more efficient than the obvious Monte Carlo algorithm described in Section 2. The improvement arises from not having to estimate by Monte Carlo the theoretical constants $p_n^{(h)}$.

The quantity $\hat{\mu}$ will be shown to equal

$$\hat{\mu} = \sum_{h=0}^{n} p_n^{(h)}\hat{\epsilon}^{(h)}. \tag{8.8}$$

Then the estimator $\hat{\omega}^{(0)}$ of line 10, Table 2 can be written as

$$\hat{\omega}^{(0)} \equiv \hat{\epsilon}^{(0)} - \hat{\mu} = \sum_{h=0}^{n}[I_{h=0} - p_n^{(h)}]\hat{\epsilon}^{(h)}. \tag{8.9}$$

(Notice that $\hat{\omega}^{(0)}$ is different than the quantity $\hat{\epsilon}^{(0)} - \bar{err}$ appearing in the .632 estimator.) It performed reasonably well in Table 2, having about the same MSE as the bootstrap and about the same bias as cross-validation. Comparing (7.14) with Table 4 shows why $\hat{\omega}^{(0)}$ has smaller bias than does $\hat{\omega}^{(\mathrm{BOOT})}$. The coefficient 1 rather than $(n - 1)/n$ on $\hat{A}$ removes the $E\,\alpha_{X_1}(X_1)/n^2$ term from (7.14). The coefficient 1 rather than 0 on $\hat{B}$ gives an added expectation of $E\,\beta_{X_1}(X_1, X_1)/2n^2 + O(1/n^3)$ thereby removing half the $-E\,\beta_{X_1}(X_1, X_1)/n^2$ term in (7.14). We could remove all of this term, for example with the estimate $(\hat{\epsilon}^{(0)} - \hat{\mu}) - (\hat{\epsilon}^{(1)} - \hat{\mu})$, but then the $E\,\gamma_{X_0}(X_0, X_1, X_1)$ term in (7.14) results in substantial upward biases.

There is an interesting connection between $\hat{\epsilon}^{(0)}$ and cross-validation. For n even we can define the *half sample cross-validation estimate*

$$\hat{\mathrm{Err}}^{(\mathrm{HCV})} = \frac{\sum_i \sum_S Q(x_i, \mathbf{X}^{*S})}{n\binom{n-1}{n/2}}, \tag{8.10}$$

the second sum being taken over all subsamples $\mathbf{X}^{*S}$ of $\{x_1, x_2, \ldots, x_n\}$ having $n/2$ elements and not containing the predicted case x_i. We will show that $\hat{\mathrm{Err}}^{(\mathrm{HCV})}$ has the expansion formula

$$\hat{\mathrm{Err}}^{(\mathrm{HCV})} = \hat{\mu} + \hat{A} + \hat{B} + \hat{C} - \frac{n-1}{6n^2(n-3)} \frac{\sum\sum_{i\neq j} \hat{\gamma}_i(j, j, j)}{n(n-1)} - \frac{1}{2n^2(n-3)} \frac{\sum\sum_{i\neq j} \hat{\gamma}_i(i, j, j)}{n(n-1)} + O_p\left(\frac{1}{n^4}\right). \tag{8.11}$$

Compared with Table 4, (8.11) shows that

$$(\hat{\mathrm{Err}}^{(\mathrm{HCV})} - \hat{\mu}) = (\hat{\epsilon}^{(0)} - \hat{\mu}) + c_n + O_p(1/n^{5/2}), \tag{8.12}$$

$c_n = -1/6n^2\, E\,\gamma_{X_0}(X_1, X_1, X_1)$ a constant of order $O(1/n^2)$. This suggests a high correlation between $\hat{\mathrm{Err}}^{(\mathrm{HCV})} - \hat{\mu}$ and $\hat{\epsilon}^{(0)} - \hat{\mu}$, and in fact the observed correlations were .86 experiment (2, 14), .98 experiment (2, 20), .94 experiment (5, 14), and .95 experiment (5, 20).

Expression (7.8), which gives (8.11), applies only when $\mathbf{x}^*$ has n component cases. A half-sample $\mathbf{X}^{*S} = \{x_{i_1}, x_{i_2}, \ldots, x_{i_{n/2}}\}$ can be regarded as a sample of size n by doubling each case, $\{x_{i_1}, x_{i_1}, x_{i_2}, x_{i_2}, \ldots, x_{i_{n/2}}, x_{i_{n/2}}\}$. With this understanding the quantity $Q(x_i, \mathbf{X}^{*S}) = Q[y_i, \eta(t_i, \mathbf{X}^{*S})]$ is well defined and can be evaluated by (7.8).

The cross-validation estimate $\hat{\omega}^{(\mathrm{CV}+)}$ introduced at (7.6), which deletes and adds single cases at a time, turns out to have expansion

$$\hat{\omega}^{(\mathrm{CV}+)} = \hat{A} + 2\hat{B} + 3\frac{n-1}{n-2}\hat{C} + \frac{1}{2n^2} \frac{\sum\sum_{i\neq j} \hat{\gamma}_i(i, j, j)}{n(n-1)} + O_p\left(\frac{1}{n^3}\right). \tag{8.13}$$

As shown at (7.8) this has little bias (cf. the remarks following (8.9)) but high variability, the same as $\hat{\omega}^{(\mathrm{CV})}$.

It seems reasonable for an estimator $\hat{\omega}$ of ω to begin $\hat{A} + O_p(1/n^2)$, as do $\hat{\omega}^{(\mathrm{BOOT})}$, $\hat{\omega}^{(0)}$, $err^{(\mathrm{HCV})} - \hat{\mu}$, and $\hat{\omega}^{(\mathrm{CV}+)}$. This makes $\hat{\omega}$ unbiased for ω order $O(1/n)$, the bias being $O(1/n^2)$. At a more primitive level, $\hat{A} = (-1/(n-1))(\sum \hat{\alpha}_i(i)/n)$ looks like $\omega = (-1/n)(E\,\alpha_{X_1}(X_1))$. One estimator that does not begin this way is connected with the Err estimate E_* (Err*) mentioned in the first paragraph of Section 5. Its use amounts to estimating ω by $\hat{\omega} = E_*(\mathrm{Err}^*) - \bar{err}$. Notice that $E_*\,\mathrm{Err}^* = \hat{\mu}$, by the first line of (7.9). The expansion of $\hat{\omega} = \hat{\mu} - \bar{err}$ turns out to be

$$\hat{\mu} - \bar{err} = \frac{\sum\sum_{i\neq j} \hat{\beta}_i(j, j)}{2n^3} + \left\{ \frac{\sum_i \hat{\beta}_i(i, i)}{2n^3} - \frac{\sum_i \sum_j \hat{\gamma}_i(j, j, j)}{3n^4} - \frac{\sum_i \sum_j \sum_{j'} \hat{\delta}_i(j, j, j', j')}{8n^5} \right\} + O_p\left(\frac{1}{n^3}\right). \tag{8.14}$$

The $O_p(1/n)$ term in (8.14) is $\sum \sum_{i\neq j} \hat{\beta}_i(j, j)/2n^3$, not $\hat{A}$, with expectation $E\,\beta_{X_0}(X_1, X_1)/2n + O(1/n^2)$. There is no theoretical reason for believing that this will be near ω, and the numerical results were terrible, for example $E\,\hat{\omega} = .024$ compared with $\omega = .093$ for experiment (2, 14).

The .632 estimator of Section 6 also has the "wrong" $O_p(1/n)$ term. Table 4 and (8.14) give

$$E\,\hat{\omega}^{(.632)} = \omega\left\{ .632 + \frac{.632}{2} \frac{E\,\beta_{X_0}(X_1, X_1)}{-E\,\alpha_{X_1}(X_1)} \right\} + O\left(\frac{1}{n^2}\right). \tag{8.15}$$

Table 4 gives these values for the bracketed factor,

Experiment:	(5, 14)	(5, 20)	(2, 14)	(2, 20)
{Factor}:	.84	.87	.86	1.02

(8.16)

The rationale for $\hat{\omega}^{(.632)}$ makes it unsurprising that these numbers are near one. On the other hand there is no guarantee that this will always happen, and arbitrarily bad counterexamples can be constructed.

9. REMARKS

Remark A. The sample sizes in our experiments, $n = 14$ or 20, are small. In practice small sample sizes can arise even when n is large, if we are interested in estimating the error rate for only a portion of the population. For example Efron and Gong (1983) consider a medical example with $n = 155$. Of particular interest are the 33 patients who died. Cross-validation, the bootstrap, and so on, can easily be modified to give the prediction rule's

estimated error rate for the population of those who die, but the effective sample size is then 33, not 155.

Remark B. The methods of this article can be applied to other problems besides estimating prediction error. Suppose we need to estimate a density function $f(x_0)$ on the basis of a random sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ from f, and wish to select among a family of possible density estimators $f_\lambda(x_0, \mathbf{x})$. Here λ might be the window width of a kernel estimator. Let $Q_\lambda(x_0, \mathbf{x}) = -\log f_\lambda(x_0; \mathbf{x})$ and $err_\lambda(\mathbf{x}, f) = E\, Q_\lambda(X_0, \mathbf{x}) = -\int[\log f_\lambda(x_0; \mathbf{x})]f(x_0)dx_0$. In an insightful paper Wong (1983) suggests selecting λ to minimize $E_*\, err_\lambda{}^*$. Using arguments much like those in Section 2, he shows that this is asymptotically equivalent to the older method of "modified likelihood." In light of the remarks in Section 8, we might prefer to define $\bar{err}_\lambda(\mathbf{x}) = -1/n \sum_{i=1}^n \log f_\lambda(x_i, \mathbf{x})$, and select λ minimizing $\bar{err}_\lambda + E_*(err_\lambda - \bar{err}_\lambda)^*$. Wong (1982) has shown that this last approach does in fact lead to a superior asymptotic theory.

Remark C. The statistician may want more than just an estimate of Err. Bootstrap methods are helpful in understanding the variability of all aspects of the prediction problem. As an example, in simulation 1 of experiment GG the forward stepwise logistic regressions selected variables 1 and 2 for inclusion in the fitted prediction rule $\eta(\cdot, \mathbf{x})$. In $B = 100$ bootstrap replications of simulation 1 the following sets of variables were selected:

set selected:	{12}	{123}	{13}	{3}	{2}	all others
# time selected:	53	15	11	8	5	8

(9.1)

Without attempting a quantitative assessment, we see that the "standard error" of the set of variables selected, about the central value {12}, is reasonably small in this case.

Remark D. Stone (1974) and Geisser (1975) emphasize the use of cross-validation to select among competing prediction rules. As a simple example suppose we observe a random sample $x_1, x_2, \ldots, x_n$ from a distribution F on the real line, and wish to choose between two estimators $\hat{\eta}_l = \eta_l(\mathbf{x})$, $l = 1, 2$, perhaps the sample median and the 10 percent trimmed mean. The goal is to minimize the expected squared error of prediction $E[X_0 - \eta(\mathbf{x})]^2$ for a future observation X_0 for F.

We wish to choose $l = 1$ or 2 minimizing $\mathrm{Err}_l(\mathbf{x}, F) = E_F[X_0 - \eta_l(\mathbf{x})]^2$. If F is actually normal, the difference $\mathrm{Err}_2 - \mathrm{Err}_1$ estimated by the bootstrap turns out to be

$$\hat{\mathrm{Err}}_2{}^{(\mathrm{BOOT})} - \hat{\mathrm{Err}}_1{}^{(\mathrm{BOOT})} = (\hat{\eta}_2 - \bar{x})^2 - (\hat{\eta}_1 - \bar{x})^2 + O_p\left(\frac{1}{n^{3/2}}\right), \quad (9.2)$$

so asymptotically the bootstrap selects the estimate nearest the sample mean $\bar{x}$. The same can be shown to hold for the cross-validation estimate $\hat{\mathrm{Err}}_2{}^{(\mathrm{CV})} - \hat{\mathrm{Err}}_1{}^{(\mathrm{CV})}$ *if the jackknife estimates of* $\mathrm{var}(\hat{\eta}_l)$ *and* $\mathrm{cov}(\hat{\eta}_l, \bar{x})$ *converge to their correct values.* If not, cross-validation can give strange answers. Stone (1977) shows that for $F \sim N(0, 1)$ the cross-validation method will select $\hat{\eta}_2$ the median as better than $\hat{\eta}_1$ the mean with asymptotic probability 0.5008.

Remark E. Cross-validation is often carried out removing large blocks of observations at a time. If $n = GH$ and $\mathbf{x}_{(g)} = (x_1, x_2, \ldots, x_{(g-1)H}, x_{gH+1}, \ldots, x_n)$, then $\hat{\mathrm{Err}}^{(\mathrm{CVG})} = \sum_{g=1}^G \sum_{h=1}^H Q(x_{(g-1)H+h}, \mathbf{x}_{(g)})/n$ requires only G recomputations of η. There are also theoretical reasons for preferring $\hat{\mathrm{Err}}^{(\mathrm{CVG})}$ to $\hat{\mathrm{Err}}^{(\mathrm{CV})}$. As explained in Section 6.2 of Efron (1982), quadratic approximation formulas like (2.16) tend to be more trustworthy for H large. In other words $\hat{\mathrm{Err}}^{(\mathrm{CVG})}$ is likely to be closer in value to $\hat{\mathrm{Err}}^{(\mathrm{BOOT})}$ than is $\hat{\mathrm{Err}}^{(\mathrm{CV})}$. As an example, grouped cross-validation like the bootstrap, selects the mean in preference to the median with asymptotic probability one in Stone's problem, Remark D, if H is suitably large.

If H is large than $\hat{\mathrm{Err}}^{(\mathrm{CVG})} - \bar{err}$ will have substantial upward bias as an estimate of ω. For example $\hat{\mathrm{Err}}^{(\mathrm{HCV})} - \bar{err}$, (8.10), corresponding to $G = 2$, $H = n/2$, is upwardly biased $O(1/n)$. This will be true for any choice $H = cn$, c fixed as $n \to \infty$. The bias can be removed by estimating Err with $\bar{err} + (\hat{\mathrm{Err}}^{(\mathrm{CVG})} - \hat{\mu})$ instead of $\hat{\mathrm{Err}}^{(\mathrm{CVG})}$, but that involves calculating the bootstrap quantity $\hat{\mu}$. At this point it becomes simpler to estimate ω by $\hat{\omega}^{(0)} = \hat{\epsilon}^{(0)} - \hat{\mu} \doteq \hat{\mathrm{Err}}^{(\mathrm{HCV})} - \hat{\mu}$.

Remark F. Cross-validation behaves more like the bootstrap if $Q[y, \eta]$ is a smooth function, like $(y - \eta)^2$ rather than (1.1), and if $\eta(\cdot, \mathbf{x})$ is also a moderately smooth function of $\mathbf{x}$. Then (2.16) gives more accurate approximations. In this case it is reasonable to estimate Err by $\hat{\mathrm{Err}}^{(\mathrm{CV})}$, though bootstrap calculations may still be helpful for other purposes, as in Remark C.

Remark G. Consider the ordinary least squares (OLS) situation $y_i = t_i\,\beta + e_i$, $e_i \overset{\mathrm{iid}}{\sim} N(0, \sigma^2)$, $i = 1, \ldots, n$, and $\eta(t_0, \mathbf{x}) = t_0(\mathbf{t}'\mathbf{t})^{-1}\mathbf{t}'\mathbf{y}$. If the predictors t_i are p dimensional then $\omega = (2p/n)\,\sigma^2$. The UMVU estimate of ω is $\hat{\omega} = (2p/n)\,\hat{\sigma}^2$, $\hat{\sigma}^2$ the usual unbiased estimate of σ^2, and it is easily shown that $\mathrm{corr}(\mathrm{op}, \hat{\omega}) = -\sqrt{1 - p/n}$. For $p = 2$, $n = 14$, the correlation is $-.93$.

The .632 estimator had $\mathrm{corr}(\mathrm{op}, \hat{\omega})$ nearly zero, and this largely accounted for its good performance in the sampling experiments. The OLS example suggests that we cannot always get $\mathrm{corr}(\mathrm{op}, \hat{\omega}) \doteq 0$ for a good estimator of ω.

We can change OLS to be more like the dichotomous models by assuming σ^2 a known function of β, say $\sigma_\beta{}^2 = a_0 + b_0(\beta - \beta_0)$ for β near some fixed vector β_0, a_0 and b_0 given. Then if β is estimated by least squares, the obvious parametric estimate $\hat{\omega} = (2p/n)\,[a_0 + b_0(\hat{\beta} - \beta_0)]$ has $\mathrm{corr}(\mathrm{op}, \hat{\omega}) = 0$, a similar result holding if β is estimated by maximum likelihood. In this case, as in the sampling experiments, we can expect good nonparametric estimators to have $\mathrm{corr}(\mathrm{op}, \hat{\omega})$ nearly zero.

10. SUMMARY

1. There are a variety of nonparametric methods available for estimating err in the dichotomous prediction problem (1.1), (1.2), all of which are closely related to nonparametric maximum likelihood estimation, that is, to the bootstrap.

2. In practical situations the different methods can give considerably different answers.

3. Cross-validation (1.4) gives a nearly unbiased estimate of err, but often with unacceptably high variability, particularly if n is small.

4. The ordinary bootstrap (2.10) gives an estimate of err with low variability, but with a possibly large downward bias, particularly in highly overfitted situations.

5. The double bootstrap of Section 5 and the $\hat{\omega}^{(0)}$ estimator (8.9) automatically correct the bias of the ordinary bootstrap without increasing the MSE of estimation.

6. The randomized bootstrap, Section 4, requires a modest amount of additional input from the statistician, but results in substantially lower MSE. Overall it performed second best in the sampling experiments.

7. The .632 estimator of Section 6 performed best in the sampling experiments, but has the weakest theoretical justification. It is recommended with caution, pending further numerical and theoretical study.

APPENDIX

Derivation of (5.4), (5.7)

The second-level bootstrap vector $\mathbf{P}^{**}$ has, given $\mathbf{P}^*$, a conditional multinomial distribution, divided by n,

$$\mathbf{P}^{**} \mid \mathbf{P}^* \sim \text{Mult}_n(n, \mathbf{P}^*)/n. \tag{A.1}$$

For instance if $P_i^* = h/n$, then $P_i^{**}|\mathbf{P}^* \sim bi(n, h/n)/n$, the proportion of heads observed in n flips of a coin having probability of heads h/n.

Denote by $E_{**}^{(*)}$ the expectation with respect to probability mechanism (A.1), with $\mathbf{P}^*$ held fixed. Also, let E_{**} indicate expectation with respect to the marginal distribution of $\mathbf{P}^{**}$, obtained from (A.1) and the distribution of $\mathbf{P}^*$,

$$\mathbf{P}^* \sim \text{Mult}_n(n, \mathbf{P}^\circ)/n \quad (\mathbf{P}^\circ = (1, 1, \ldots, 1)/n, \tag{A.2}$$

which agrees with its use in (5.4). Finally, let $E_*^{(**)}$ indicate expectation with respect to the conditional distribution of $\mathbf{P}^*$ given $\mathbf{P}^{**}$. In all these expectations the data $\mathbf{x}$ are fixed.

We need to evaluate $E_*(\hat{\omega}^{(\text{BOOT})})^*$, where $\hat{\omega}^{(\text{BOOT})} = E_*\ \text{op}(\mathbf{X}^*, \hat{F})$. Suppose that $r(\mathbf{x}) \equiv E_*\ R(\mathbf{X}^*, \hat{F})$ is the bootstrap expectation of a random variable $R(\mathbf{X}, F)$, which is invariant under all permutations of the coordinates of $\mathbf{X}$ (as is $\text{op}(\mathbf{X}, F)$). A bootstrap replication of r is of the form

$$r(\mathbf{X}^*) = E_{**}^{(*)}\, R(\mathbf{X}^{**}, \hat{F}^*) = E_{**}^{(*)}\, R(\mathbf{P}^{**}, \mathbf{P}^*). \tag{A.3}$$

The last expression makes sense because, with data $\mathbf{x}$ fixed $\mathbf{P}^*$ determines F^*, and $\mathbf{P}^{**}$ determines $\mathbf{X}^{**}$ up to permutations of its components.

By carefully following through the various definitions we can apply (A.3) to $r(\mathbf{x}) = \hat{\omega}^{(\text{BOOT})} = E_*\ \text{op}(\mathbf{X}^*, \hat{F})$ and obtain

$$(\hat{\omega}^{(\text{BOOT})})^* = E_{**}^{(*)} \left\{ \sum_{i=1}^{n} (P^* - P^{**})\, Q[y_i, \eta(t_i, \mathbf{X}^{**})] \right\}. \tag{A.4}$$

Since $E_* E_{**}^{(*)}\, R(\mathbf{P}^{**}, \mathbf{P}^*) = E_{**} E_*^{(**)}\, R(\mathbf{P}^{**}, \mathbf{P}^*)$ for any function $R(\mathbf{P}^{**}, \mathbf{P}^*)$, (A.4) gives

$$E_*(\hat{\omega}^{(\text{BOOT})})^* = E_{**} \left\{ \sum_{i=1}^{n} e(P_i^{**})\, Q[y_i, \eta(t_i, \mathbf{X}^{**})] \right\}, \tag{A.5}$$

where

$$e(P_i^{**}) = E_*^{(**)}\, (P_i^* - P_i^{**}). \tag{A.6}$$

This shows that (5.4) holds with $e(P_i^{**})$ given by (A.6).

The conditional expectation $E_*^{(**)}\, (P_i^* - P_i^{**})$ is actually a function of the entire vector $\mathbf{P}^{**}$ and not just of the ith component P_i^{**}. However, the effect of the other components turns out to be quite small, and will be ignored in what follows. Let $nP_i^{**} \equiv N_i^{**}$ and $nP_i^* \equiv N_i^*$. Then a standard Bayesian calculation gives

$$e(P_i^{**}) = \frac{\sum_{n_i^*=0}^{n} \frac{(n_i^* - N_i^{**})}{n}\, bi\left(n, \frac{1}{n}; n_i^*\right)\, bi(n, n_i^*/n; N_i^{**})}{\sum_{n_i^*=0}^{n} bi\left(n, \frac{1}{n}; n_i^*\right)\, bi(n, n_i^*/n; N_i^{**})} \tag{A.7}$$

where $bi(n, p; h)$ is the binomial probability $\binom{n}{h} p^h (1 - p)^{n-h}$.

As $n \to \infty$ the distribution of $N_i^* \to Po(1)$, a Poisson with parameter one, and $N_i^{**} \mid N_i^* \to Po(N_i^{**})$. In this case (A.7) simplifies considerably, and can be rewritten as

$$e(P^{**}) = \frac{1}{n} \left\{ \frac{EZ^{N_i^{**}+1}}{EZ^{N_i^{**}}} - N_i^{**} \right\} \tag{A.8}$$

where $Z \sim Po(e^{-1})$, a Poisson with parameter $\lambda = e^{-1} = .3679$. Formula (A.8) was used to calculate (5.7). These values are within a few percent of (A.7), even for small n. For example with $n = 10$, $P_i^{**} = 0$, (A.7) gives $e(P_i^{**}) = .0359$ compared with $e(P_i^{**}) = .0368$ for (A.8).

Verification of (6.9)

Define the set $T(x_0, \Delta) = \{x : x_0 \in S(x, \Delta)\}$. Then

$$\text{Prob}\{\delta(X_0, \mathbf{X}) > \Delta\} = \text{Prob}\left\{ X_0 \notin \bigcup_{i=1}^{n} S(X_i, \Delta) \right\}$$

$$= \text{Prob}\{X_i \notin T(X_0, \Delta),\ i = 1, 2, \ldots, n\}$$

$$= E[1 - \text{Prob}\{T(X_0, \Delta)\}]^n. \tag{A.9}$$

Suppose that Prob$\{T(X_0, \Delta)\}$ approximately equals Δ. This will be true, at least for small values Δ, if the original neighborhoods $S(x, \Delta)$ are based on a distance function symmetric in x and x_0, as in (6.4). Then (A.9) gives the approximation

$$\text{Prob}\{\delta(X_0, \mathbf{X}) > \Delta\} \doteq (1 - \Delta)^n. \qquad \text{(A.10)}$$

Let n^* be the number of cases x_i having $P_i^* > 0$ in a given bootstrap sample $\mathbf{X}^*$. The same reasoning as in (A.9) gives

$$\text{Prob}\{\delta(X_0^*, \mathbf{X}^*) > \Delta \mid \delta^* > 0\}$$
$$= E[1 - \text{Prob}\{T(X_0^*, \Delta)\}]^{n^*}. \qquad \text{(A.11)}$$

Both the probability and the expectation in (A.11) are marginal over $\mathbf{X}$ and $\mathbf{X}^*$, X_0^*. The approximation Prob$\{T(X_0^*, \Delta)\} \doteq \Delta$ gives

$$\text{Prob}\{\delta(X_0^*, \mathbf{X}^*) > \Delta \mid \delta^* > 0\} \doteq (1 - \Delta)^{.632n}, \qquad \text{(A.12)}$$

where we have substituted $E\, n^* \doteq .632n$ for n^*.

If (A.11) and (A.12) can be trusted then Prob$\{\delta(X_0, \mathbf{X}) > \Delta\} \doteq e^{-n\Delta}$ and Prob$\{\delta(X_0^*, \mathbf{X}^*) > \Delta/.632\} \doteq e^{-n\Delta}$, verifying (6.9). The insert in Figure 2 compares Prob$\{\delta(X_0, \mathbf{X}) < \Delta\}$ with

$$\text{Prob}\left\{\delta(X_0^*, \mathbf{X}^*) < \frac{\Delta}{.632} \;\middle|\; \delta^* > 0\right\},$$

showing excellent agreement.

Derivation of (7.13)

There are simple algebraic relationships between the terms in (7.1) and those in (7.8),

$$\hat{\mu}_j = \mu_{x_j} + \alpha_{x_j}(\cdot) + \frac{\binom{n}{2}}{n^2}\beta_{x_j}(\cdot, \cdot) + \frac{\binom{n}{3}}{n^3}\gamma_{x_j}(\cdot, \cdot, \cdot) + \dots,$$

$$\hat{\alpha}_j(j_1) = [\alpha_{x_j}(x_{j_1}) - \alpha_{x_j}(\cdot)] + \frac{\binom{n-1}{1}}{n}[\beta_{x_j}(x_{j_1}, \cdot) - \beta_{x_j}(\cdot, \cdot)] + \frac{\binom{n-1}{2}}{n^2}[\gamma_{x_j}(x_{j_1}, \cdot, \cdot) - \gamma_{x_j}(\cdot, \cdot, \cdot)] + \dots,$$

$$\hat{\beta}_j(j_1, j_2) = [\beta_{x_j}(x_{j_1}, x_{j_2}) - \beta_{x_j}(x_{j_1}, \cdot) - \beta_{x_j}(x_{j_2}, \cdot) + \beta_{x_j}(\cdot, \cdot)] + \frac{\binom{n-2}{1}}{n}[\gamma_{x_j}(x_{j_1}, x_{j_2}, \cdot) - \gamma_{x_j}(x_{j_1}, \cdot, \cdot) + \gamma_{x_j}(x_{j_2}, \cdot, \cdot) + \gamma_{x_j}(\cdot, \cdot, \cdot)] + \dots, \qquad \text{(A.13)}$$

the dot notation indicating averages, $\alpha_{x_j}(\cdot) = \sum_{j_1=1}^n \alpha_{x_j}(x_{j_1})/n$, $\beta_{x_j}(x_{j_1}, \cdot) = \sum_{j_2=1}^n \beta_{x_j}(x_{j_1}, x_{j_2})/n$, and so on. (This last average involves terms like $\beta_{x_j}(x_2, x_1)$, which do not seem to exist in (7.1). However, the corresponding terms there are really random variables $\beta_{x_j}(X_i, X_{i'})$, which have well-defined values for $X_i = x_2$, $X_{1'} = x_1$.) Relationships (A.13) are familiar in the comparison of random effects with fixed effects models in ANOVA.

The second line of (A.13) shows that $E\,\hat{\alpha}_j(j) = E\,\alpha_{X_j}(X_j) + O(1/n)$. The $O(1/n)$ term can be computed explicitly giving, from (7.4), (7.12), formula (7.13).

Proofs for Results of Section 8

For any bootstrap random variable $R^* = R(\mathbf{X}^*, \hat{F})$, and for any choice of $i = 1, 2, \dots, n$,

$$E_* R^* = \sum_{h=0}^{n} p_n^{(h)}\, E_*\{R^* \mid P_i^* = h/n\} \qquad \text{(A.14)}$$

since $p_n^{(h)} = \text{Prob}_*\{P_i^* = h/n\}$. Applying (A.14) to $R^* = Q(x_i, \mathbf{X}^*)$ gives $\hat{\mu}_i = E_*\, Q(x_i, \mathbf{X}^*) = \sum_{h=0}^n p_n^{(h)}\, \hat{\epsilon}_i^{(h)}$, where

$$\hat{\epsilon}_i^{(h)} = E_*\{Q(x_i, \mathbf{X}^*) \mid P_i^* = h/n\}. \qquad \text{(A.15)}$$

It is easy to see that $\sum_{i=1}^n \hat{\epsilon}_i^{(h)}/n = \hat{\epsilon}^{(h)}$, (8.1), so $\hat{\mu} = \sum_{i=1}^n \hat{\mu}_i/n = \sum_{h=0}^n p_n^{(h)}\, \hat{\epsilon}^{(h)}$, as claimed at (8.8). Likewise letting R^* in (A.14) equal $R_i^* \equiv (1/n - P_i^*)\, Q(x_i, \mathbf{X}^*)$ gives $E_*\, R_i^* = \sum_{h=0}^n p_n^{(h)}\, ((1 - h)/n)\, \hat{\epsilon}_i^{(h)}$, and then $\hat{\omega}^{(\text{BOOT})} = E_* \sum_{i=1}^n R_i^* = \sum_{h=0}^n p_n^{(h)}\, (1 - h)\hat{\epsilon}^{(h)}$, verifying (8.7).

Rather than prove the theorem we will prove the stronger result

$$\hat{\epsilon}_i^{(h)} = \hat{\mu}_i + \lambda_1^{(h)} \frac{\hat{\alpha}_i(i)}{[-(n-1)]} + \lambda_2^{(h)} \frac{\hat{\beta}_i(i, i)}{[-(n-1)]^2} \frac{\binom{n}{2}}{n^2} + \dots. \qquad \text{(A.16)}$$

Averaged over $n = 1, 2, \dots, n$, (A.16) gives (8.5). Because we have assumed that $Q(x_i, \mathbf{X}^*) \equiv Q[y_i, \eta(t_i, \mathbf{X}^*)]$ is unchanged under permutations of $\mathbf{X}^*$'s components, (A.15) can be written as

$$\hat{\epsilon}_i^{(h)} = E_*\{Q(x_i, \mathbf{X}^*) \mid X_1^*, \dots, X_h^* = x_i \text{ and } X_{h+1}^*, \dots, X_n^* \neq x_i\}. \qquad \text{(A.17)}$$

Now we use (7.8) and (7.11) to evaluate this conditional expectation. For example,

$$E_*\{\hat{\alpha}_i(X_{j_1}^*) \mid X_{j_1}^* \neq x_i\} = \frac{1}{n-1}\sum_{i_1 \neq 1} \hat{\alpha}_i(i_1) = \frac{-\hat{\alpha}_i(i)}{n-1} \qquad \text{(A.18)}$$

and so

$$E_*\left\{\frac{1}{n}\sum_{j=1}^{n}\hat{\alpha}_i(X_j^*) \mid X_1^*, \ldots, X_h^* = x_i \text{ and } X_{h+1}^*, \ldots, X_n^* \neq x_i\right\} = \left[\frac{h}{n} + \frac{n-h}{n}\left(-\frac{1}{n-1}\right)\right]\hat{\alpha}_i(i). \quad \text{(A.19)}$$

Similar calculations for the higher-order terms in (7.8) evaluate (A.17) as

$$\begin{aligned}\hat{\epsilon}_i^{(h)} = \hat{\mu}_i &+ \frac{1}{n}\left[\binom{h}{1} + \binom{n-h}{1}\left(-\frac{1}{n-1}\right)\right]\hat{\alpha}_i(i) \\ &+ \frac{1}{n^2}\left[\binom{h}{2} + \binom{h}{1}\binom{n-h}{1}\left(-\frac{1}{n-1}\right) + \binom{n-h}{2}\left(-\frac{1}{n-1}\right)^2\right]\hat{\beta}_i(i, i) \\ &+ \frac{1}{n^3}\left[\binom{h}{3} + \binom{h}{2}\binom{n-h}{1}\left(-\frac{1}{n-1}\right) + \binom{h}{1}\binom{n-h}{2}\left(-\frac{1}{n-1}\right)^2 + \binom{n-h}{3}\left(-\frac{1}{n-1}\right)^3\right]\hat{\gamma}_i(i, i, i) + \ldots . \end{aligned} \quad \text{(A.18)}$$

The bracketed terms are obviously related to hypergeometric expectations of the form (8.6), and results (A.16), (8.5) follow easily.

The proof of (8.11) also relies on (7.8) and (7.11). In the notation of (8.10),

$$\begin{aligned}Q(x_i, \mathbf{X}^{*S}) = \hat{\mu}_i &+ \frac{2}{n}\sum_{j\in S}\hat{\alpha}_i(j) + \frac{4}{n^2}\sum_{j<j'\in S}\sum\hat{\beta}_i(j, j') \\ &+ \frac{1}{n^2}\sum_{j\in S}\hat{\beta}_i(j, j) + \frac{8}{n^3}\sum_{j<j'<j''\in S}\sum\sum\hat{\gamma}_i(j, j', j'') \\ &+ \frac{2}{n^3}\sum_{j<j'\in S}\sum[\hat{\gamma}_i(j, j', j') + \hat{\gamma}_i(j, j, j')] + \ldots , \end{aligned} \quad \text{(A.19)}$$

all sums being only over cases x_j in $\mathbf{X}^{*S}$. Then

$$\hat{\mathrm{Err}}_i^{(\mathrm{HCV})} \equiv \frac{\sum_S Q(x_i, \mathbf{X}^{*S})}{\binom{n-1}{n/2}} = \hat{\mu}_i + \frac{\sum_{j\neq i}\hat{\alpha}_i(j)}{n-1} + \frac{\sum\sum_{j<j'\neq i}\hat{\beta}_i(j, j')}{n(n-1)} + \frac{\sum_{j\neq i}\hat{\beta}_i(j, j)}{2n(n-1)} + \ldots , \quad \text{(A.20)}$$

or

$$\begin{aligned}\hat{\mathrm{Err}}_i^{(\mathrm{HCV})} &= \hat{\mu}_i + \hat{\alpha}_i(i)\left[-\frac{1}{n-1}\right] + \hat{\beta}_i(i, i)\frac{\binom{n}{2}}{n^2}\left[-\frac{1}{n-1}\right]^2 \\ &+ \hat{\gamma}_i(i, i, i)\frac{\binom{n}{3}}{n^3}\left[-\frac{1}{n-1}\right]^3 - \frac{\sum_{j\neq 1}\hat{\gamma}_i(j, j, j)}{n-1}\frac{n-1}{6n^2(n-3)} \\ &- \frac{1}{2n^2(n-3)}\frac{\sum_{j\neq 1}\hat{\gamma}_i(i, j, j)}{n-1} + O_p\left(\frac{1}{n^4}\right), \end{aligned} \quad \text{(A.21)}$$

(A.21) following from (A.20) by (7.11). But $\hat{\mathrm{Err}}^{(\mathrm{HCV})} = \sum_{i=1}^{n}\hat{\mathrm{Err}}_i^{(\mathrm{HCV})}/n$ so (A.20) is an improved version of (8.11). Formulas (8.13) and (8.4) follow from similar algebraic manipulations of (7.8) and (7.11).

The first line in (A.21) is the beginning of expansion (A.16) for $\hat{\epsilon}^{(0)}$. The quantity $\sum_{j\neq i}\hat{\gamma}_i(j, j, j)/(n-1)$ appearing in the next term of (A.21) can be written as

$$\begin{aligned}\frac{\sum_{j\neq i}\hat{\gamma}_i(j, j, j)}{n-1} &= \sum_j\sum_{j'}\sum_{j''}c_i(j, j', j'')\,\gamma_{x_i}(x_j, x_{j'}, x_{j''}) \\ &+ \sum_j\sum_{j'}\sum_{j''}\sum_{j'''}d_i(j, j', j'', j''')\,\delta_{x_i}(x_j, x_{j'}, x_{j''}, x_{j'''}) + \ldots . \end{aligned} \quad \text{(A.22)}$$

The constants $c_i(j, j', j'')$, $d_i(j, j', j'', j''')$, . . . , are calculated from (A.13). For example, $c_i(j, j, j) = 1/(n-1) - 3/n(n-1) + 3/n^2(n-1) - 1/n^3$ if $i \neq j$. The variance of $\sum_j\sum_{j'}\sum_{j''}c_i(j, j', j'')\gamma_{X_i}(X_j, X_{j'}, X_{j''})$ can be calculated exactly and has the limiting form, as $n \to \infty$, $[\mathrm{var}\,\gamma_{X_0}(X_1, X_1, X_1)]/n$. Likewise its expectation approaches $E\,\gamma_{X_0}(X_1, X_1, X_1)$. Ignoring further terms in (A.22) this gives $\sum_{j\neq i}\hat{\gamma}_i(j, j, j)/(n-1) = E\,\gamma_{X_0}(X_1, X_1, X_1) + O_p(1/n^{1/2})$. Returning to (A.21)

$$(\hat{\mathrm{Err}}_i^{(\mathrm{HCV})} - \hat{\mu}_i) = (\hat{\epsilon}_i^{(0)} - \hat{\mu}_i) + \frac{E\,\gamma_{X_0}(X_i, X_1, X_1)}{6n^2} + O_p(1/n^{5/2}), \quad \text{(A.23)}$$

which is an improved version of (8.12).

[*Received May 1982. Revised October 1982.*]

REFERENCES

EFRON, B. (1975), "The Efficiency of Logistic Regression Compared to Normal Discrimanant Analysis," *Journal of the American Statistical Association*, 70, 892–898.

——— (1979), "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics*, 7, 1–26.

——— (1982), "The Jackknife, The Bootstrap, and Other Resampling Plans," SIAM NSF-CBMS, Monograph #38.

EFRON, B., and GONG, G. (1983), "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," *The American Statistician*, 37, 36–48.

EFRON, B., and STEIN, C. (1981), "The Jackknife Estimate of Variance," *Annals of Statistics*, 9, 586–596.

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

GONG, G. (1982), "Cross-Validation, the Jackknife, and the Bootstrap: Excess Error Estimation in Forward Logistic Regression," Ph.D. dissertation, Stanford University Technical Report No. 80, Dept. of Statistics.

HAMPEL, F. (1974), "The Influence Curve and its Role in Robust Estimation," *Journal of the American Statistical Association*, 64, 1303–1317.

LACHENBRUCH, P., and MICKEY, M. (1968), "Estimation of Error Rates in Discrimant Analysis," *Technometrics*, 10, 1–11.

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

——— (1977), "Asymptotics For and Against Cross-Validation," *Biometrika*, 64, 29–38.

WONG, W. (1983), "A Note on the Modified Likelihood for Density Estimation," *Journal of the American Statistical Association*, 78, 461–463.