Biometrika (1989), **76**, 3, *pp.* 503–14
Printed in Great Britain

A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods

BY PRABIR BURMAN
Division of Statistics, University of California, Davis, California 95616, U.S.A.

SUMMARY

Concepts of v-fold cross-validation and repeated learning-testing methods have been introduced here. In many problems, these methods are computationally much less expensive than ordinary cross-validation and can be used in its place. A comparative study of these three methods has been carried out in detail.

Some key words: Ordinary cross-validation; Repeated learning-testing method; v-fold cross-validation.

1. INTRODUCTION

Model selection has attracted the attention of many researchers. One of the most well-known methods is ordinary cross-validation. Much work has been done on this, for example Stone (1974, 1977), Bowman (1984) and Härdle & Marron (1985). However, the difficulty with ordinary cross-validation is that it can be computationally very expensive in a number of practical problems, for example L^1 regression, tree structured methods for classification and regression (Breiman et al., 1984), estimation of the optimal transformations of variables for correlation and regression (Breiman & Friedman, 1985). We have therefore recently introduced two techniques (Burman, 1989) in a study of the optimal transformations of variables. We call them corrected v-fold cross-validation and repeated learning-testing methods. These methods can be used in the place of ordinary cross-validation whenever the latter is computationally very expensive. In this present paper we study these methods in detail.

The organization of the paper is as follows. In § 2 we introduce the concepts of v-fold cross-validation and the repeated learning-testing methods. In § 3 we discuss the stability of our estimates. In § 4 we present a few simulation results on various aspects of our proposed estimates. We then summarize our findings in § 5. Section 6 presents a few technical results to support our simulation study, and finally we give proofs in § 7.

2. v-FOLD CROSS-VALIDATION AND REPEATED LEARNING-TESTING METHODS

2·1. *General description*

In this section we describe v-fold cross-validation and repeated learning-testing methods, demonstrate the need for bias correction and introduce the proper correction terms. The concepts of v-fold cross-validation and repeated learning-testing methods are not new. They are discussed by Breiman et al. (1984, Ch. 3, 8). Based on their simulation studies they concluded that these methods do not always work. To rectify that we have introduced certain correction terms. Burman (1989) showed that with these correction terms v-fold cross-validation and repeated learning-testing methods work very well for

model selection purposes. Here we study various aspects of these methods in detail. Is v-fold cross-validation a better method than repeated learning-testing? For what values of v does v-fold cross-validation work best?

Let $Z_1, \ldots, Z_n$ be independent and identically distributed observations with distribution function F and let F_n be the empirical estimate of F. Our goal is to estimate a quantity of the form

$$s_n = \int T(z, F_n)\, dF(z). \tag{2·1}$$

We may also wish to estimate $E(s_n)$, which could be estimated by ordinary cross-validation. We introduce three other methods for estimating s_n. We first give a few examples of T and s_n.

Example 1. Let $Z_i = (Y_i, X_i)$ $(i = 1, \ldots, n)$ be independent and identically distributed random variables and let $\mu(x) = E(Y_1 | X_1 = x)$ be the regression function. If we approximate the regression function by a polynomial of degree k, estimate the parameters by least-squares and denote the estimated regression function by $\hat{\mu}_k(x)$, then $z = (y, x)$,

$$T(z, F_n) = \{y - \hat{\mu}_k(x)\}^2, \quad s_n = \int \{y - \hat{\mu}_k(x)\}^2\, dF(x, y).$$

Here s_n is the predictive error. In model selection problems we estimate s_n for various values of k and choose that value of k for which the estimated value of s_n is the smallest.

Example 2. Let $Z = (Y, X_1, \ldots, X_d)$ be a vector of univariate random variables. Breiman & Friedman (1985) considered transformations $h(Y), \phi_1(X_1), \ldots, \phi_d(X_d)$ of the variables $Y, X_1, \ldots, X_d$ such that

$$E\{h(Y)\} = E\{\phi_1(X_1)\} = \ldots = E\{\phi_d(X_d)\} = 0, \quad E\{h^2(Y)\} = 1, \quad E\{\phi_j^2(X_j)\} < \infty.$$

Transformations $h^*, \phi_1^*, \ldots, \phi_d^*$ are called optimal if they minimize

$$e^2(h, \phi_1, \ldots, \phi_d) = E\{h(Y) - \phi_1(X_1) - \ldots - \phi_d(X_d)\}^2.$$

Assume that we have a data set of size n and estimate the optimal transformations using splines with $k_0, k_1, \ldots, k_d$ knots respectively. If $\hat{h}, \hat{\phi}_1, \ldots, \hat{\phi}_d$ are the sample estimates of the optimal transformations, then the predictive error is

$$e^2(\hat{h}, \hat{\phi}_1, \ldots, \hat{\phi}_d) = E[\{\hat{h}(Y) - \hat{\phi}_1(X_1) - \ldots - \hat{\phi}_1(X_d)\}^2 | Y_i, X_i, i = 1, \ldots, n].$$

In this example, s_n is the quantity above and we are interested in estimating it for various values $k_0, k_1, \ldots, k_d$.

Example 3: L^1 *regression.* Consider the regression problem in Example 1. Let $\hat{\theta}_0, \ldots, \hat{\theta}_d$ be the estimates obtained by minimizing

$$\sum_{i=1}^{n} |Y_i - \theta_0 - \theta_1 X_i - \theta_2 X_i^2 - \ldots - \theta_k X_i^k|.$$

Let (X, Y) be independent of (X_i, Y_i) $(i = 1, \ldots, n)$. Then the prediction error is

$$E(|Y - \hat{\theta}_0 - \hat{\theta}_1 X - \ldots - \hat{\theta}_k X^k| \,|\, Y_i, X_i, i = 1, \ldots, n).$$

Here s_n is the quantity given above. The L^1 method is typically used to get a robust estimate of the regression function. It is known that a good deal of computing is needed to calculate the estimates $\hat{\theta}_0, \ldots, \hat{\theta}_d$.

In the last two examples ordinary cross-validation could be computationally very expensive and there thus arises the need for alternative methods. Let us now formally introduce the concepts of v-fold cross-validation, the repeated learning testing and the repeated v-fold cross-validation methods.

2·2. *v-fold cross-validation*

We first divide the data randomly in v groups so that their sizes are as nearly equal as possible. Let the size of the αth group be m_α and assume that $[n/v] \leqslant m_\alpha \leqslant [n/v]+1$ for all α. Let $F_{n\alpha}$ be the empirical estimate of F based on all the $n_\alpha = n - m_\alpha$ observations outside group α and let $\bar{F}_{n\alpha}$ be the estimate of F based on the m_α observations in group α. Then a v-fold cross-validated estimate of s_n is given by

$$\text{CV}_{nv} = \sum_{\alpha=1}^{v} p_\alpha \int T(z, F_{n\alpha})\, d\bar{F}_{n\alpha}(z),$$

where $p_\alpha = m_\alpha/n$. Note that $p_\alpha \doteq v^{-1}$ for all α.

Let $s = \int T(z, F)\, dF(z)$. If we assume that $T(., F_n)$ can be expanded about $T(., F)$ and note that $n_\alpha \doteq (v-1)v^{-1}n$ for all α, then

$$E(s_n - s) = E \int \{T(z, F_n) - T(z, F)\}\, dF(z) \doteq c_0 n^{-1},$$

$$E(\text{CV}_{nv} - s) = \sum p_\alpha E \int \{T(z, F_{n\alpha}) - T(z, F)\}\, dF(z) \doteq v(v-1)^{-1} c_0 n^{-1},$$

where c_0 is a constant depending on T and F. Consequently,

$$E(\text{CV}_{nv} - s_n) \doteq (v-1)^{-1} c_0 n^{-1}. \tag{2·2}$$

For ordinary cross-validation, that is $v = n$, the right-hand side of the last expression is $O(n^{-2})$, but when v is small, say $v = 3$ or 4, this term is not necessarily very small. Simulation results presented in § 4 clearly bring this out. Also the term c_0 is of the order of k, the number of parameters being estimated. Indeed, for the regression case in Example 1, if we fit a polynomial of degree k, then c_0 is a constant times $k+1$. So CV_{nv} may turn out to be a poor estimate of s_n if the number of parameters is not small. This observation is consistent with Stone's response to Geisser's comment (Stone, 1974).

Thus a correction term for CV_{nv} is needed. We show later that the following works well. Let

$$\text{CV}^*_{nv} = \sum p_\alpha \int T(z, F_{n\alpha})\, d\bar{F}_{n\alpha}(z) + \int T(z, F_n)\, dF_n(z) - \sum p_\alpha \int T(z, F_{n\alpha})\, dF_n(z). \tag{2·3}$$

As shown in § 6,

$$E(\text{CV}^*_{nv} - s_n) = E \sum p_\alpha \int \{T(z, F_{n\alpha}) - T(z, F_n)\} d(F_n - F)(z) \doteq (v-1)^{-1} c_1 n^{-2} \tag{2·4}$$

for some constant c_1 depending on T and F. Expressions (2·2) and (2·4) show that the corrected v-fold cross-validated estimate CV^*_{nv} works much better than CV_{nv} asymptotically as n increases; the correction term is negligible when $v \doteq n$.

In the regression case, Example 1, let $\hat{\mu}_k$ be as in Example 1. Let $\hat{\mu}_{k\alpha}$ be the regression estimate based on the $n_\alpha = n - m_\alpha$ observations not in group α. Then the uncorrected v-fold cross-validated estimate of s_n is

$$\text{CV}_{nv} = \sum_{\alpha=1}^{v} (m_\alpha/n)[\sum \{Y_i - \hat{\mu}_{k\alpha}(X_i)\}^2/m_\alpha],$$

where the sum inside the bracket is over all the m_α observations in group α. We have already discussed the need for a correction term as given in (2·3). The corrected v-fold cross-validated estimate of s_n is

$$\text{CV}^*_{nv} = \text{CV}_{nv} + \sum_{i=1}^{n} \{Y_i - \hat{\mu}_k(X_i)\}^2/n - \sum_{\alpha=1}^{v} (m_\alpha/n)\left[\sum_{i=1}^{n} \{Y_i - \hat{\mu}_{k\alpha}(X_i)\}^2/n\right].$$

2·3. *Repeated learning-testing method*

In this method, we repeatedly split the data randomly in two parts, a learning set of size n_0 and a test set of size m_0 $(m_0 + n_0 = n)$. Typically $n_0 \geqslant m_0$. For each split, estimates are developed based on the data in the learning set and then these are tested on the data in the test set. On the αth split of the data let $F_{n\alpha}$ and $\bar{F}_{n\alpha}$ be the estimates of F based on the observations in the learning set and the test set respectively. If the data is split v times, then a repeated learning-testing estimate of s_n is given by

$$\text{LT}_{nv} = v^{-1} \sum_{\alpha=1}^{v} \int T(z, F_{n\alpha})\, d\bar{F}_{n\alpha}(z).$$

As for v-fold cross-validation, this does not provide an adequate estimate of s_n and a bias correction term is needed here too. The following gives a corrected repeated learning-testing estimate of s_n.

$$\text{LT}^*_{nv} = \text{LT}_{nv} + \int T(z, F_n)\, dF_n(z) - v^{-1} \sum \int T(z, F_{n\alpha})\, dF_n(z).$$

The corrected v-fold cross-validated and repeated learning-testing estimates have exactly the same form. It can be shown that

$$E(\text{LT}_{nv} - s_n) \doteq (m_0/n_0) c_0 n^{-1}, \quad E(\text{LT}^*_{nv} - s_n) \doteq (m_0/n_0) c_1 n^{-2}, \tag{2·5}$$

where c_1 is the same as in (2·4). The bias does not depend on v, the number of repeats, and is reduced if the ratio of the size of the test set to the size of the learning set is small.

The predictive sample reuse method of Geisser (1975) could be regarded as a version of the repeated learning-testing method. However, he did not introduce a correction term. Since he considered all possible partitions of the data into learning and test sets of sizes n_0 and m_0, the number, v, of repeats is $n!/\{m_0!(n-m_0)!\}$, and consequently his method involves a lot more computing than ordinary cross-validation.

2·4. *Repeated v-fold cross-validation*

This method is a combination of the v-fold cross-validation and the repeated learning-testing methods described in this section. This method is quite simple. We repeat the method of v-fold cross-validation t times. On the βth repeat we randomly split the data in v groups as described in § 2·2, get a bias corrected estimate $\text{CV}^*_{nv\beta}$ of s_n, and finally we take a simple average of these t estimates. Let us call this estimate RCV^*_{nvt}, so that

$$\text{RCV}^*_{nvt} = t^{-1} \sum \text{CV}^*_{nv\beta};$$

RCV_{nvt} has about the same bias as CV^*_{nv}:

$$E(\text{RCV}^*_{nvt} - s_n) \doteq (v-1)^{-1} c_1 n^{-2}.$$

However, a variance calculation shows that it may be better to use a vt-fold corrected cross-validated estimate rather than RCV^*_{nvt}; see § 3.

3. Stability of the estimates

We now compare the stability of the ordinary cross-validated estimate of s_n with those defined in § 2.

Retaining only the terms of order n^{-2} and higher, we get, see § 6,

$$\begin{aligned} \text{var}\,(\text{CV}_{nn} - s_n) &\doteq \gamma_0 n^{-1} + \gamma_1 n^{-2} + 2\gamma_2 n^{-2}, \\ \text{var}\,(\text{CV}^*_{nv} - s_n) &\doteq \gamma_0 n^{-1} + v(v-1)^{-1}\gamma_1 n^{-2} + 2\gamma_2 n^{-2}. \end{aligned} \tag{3·1}$$

Here $\gamma_0 > 0$, $\gamma_1 > 0$ and γ_2 are constants depending on T and F. Thus

$$\text{var}\,(\text{CV}^*_{nv} - s_n) > \text{var}\,(\text{CV}_{nn} - s_n).$$

However, the difference is $(v-1)^{-1}\gamma_1 n^{-2}$ and this decreases as v increases. So we can expect the difference to be small for $v > 2$ and this is supported by simulation. The expression for $\text{var}\,(\text{CV}_{nv} - s_n)$ is not pretty, but it can be shown that

$$\text{var}\,(\text{CV}_{nv} - s_n) = \text{var}\,(\text{CV}^*_{nv} - s_n) + O((v-1)^{-1} n^{-2});$$

see § 6. In the regression case it can be shown that it is larger than $\text{var}\,(\text{CV}^*_{nv} - s_n)$ and $\text{var}\,(\text{CV}_{nn} - s_n)$. We believe this to be true always, but do not have a proof. However, as expected, the difference between $\text{var}\,(\text{CV}^*_{nv} - s_n)$ and $\text{var}\,(\text{CV}_{nv} - s_n)$ is negligible as v becomes large.

Now consider the repeated learning-testing estimate. Even though LT^*_{nv} has a small bias as an estimate of s_n, its variance is not as small as we want it to be. Retaining terms of order n^{-1} we get

$$\begin{aligned} \text{var}\,(\text{LT}^*_{nv} - s_n) &\doteq \{1 + n_0/(m_0 v)\}\gamma_0 n^{-1}, \\ \text{var}\,(\text{LT}^*_{nv} - s_n) - \text{var}\,(\text{CV}_{nn} - s_n) &\doteq n_0/(m_0 v)\gamma_0 n^{-1}. \end{aligned}$$

Recall from § 2 that

$$E(\text{LT}^*_{nv} - s_n) \doteq (m_0/n_0) c_1 n^{-2}.$$

As the ratio m_0/n_0 decreases the bias becomes smaller but the variance increases. The only way we can reduce variance in the learning-testing case is by increasing v, the number of repeats. From a computational point of view, it seems that v-fold cross-validation is better than the repeated learning-testing method. As in the case of v-fold cross-validation, the expression for $\text{var}\,(\text{LT}_{nv} - s_n)$ is not a nice one; see § 6.

Finally, consider the repeated v-fold cross-validated estimate as given in § 2·4. It turns out that

$$\text{var}\,(\text{RCV}^*_{nvt} - s_n) \doteq \gamma_0 n^{-1} + t(t-1)^{-1}(v-1)^{-1}\gamma_1 n^{-2} + 2\gamma_2 n^{-2}, \tag{3·2}$$

where γ_0, γ_1 and γ_2 are as in (3·1). An easy calculation shows that the expression in (3·2) is larger than the variance of the corrected vt-fold cross-validated estimate, though the difference $t^{-1}(v-1)^{-1}\gamma_1 n^{-2}$ is rather small. In repeated v-fold cross-validation, the estimate $T(.,.)$ has to be calculated vt times, as it has to be in the case of vt-fold cross-validation, so that we may be better off using vt-fold cross-validation than repeating v-fold cross-validation t times.

4. SOME SIMULATION RESULTS

In this section we present some simulation results for a regression model. The correct regression is a quadratic model but we fit a linear model to it. Let

$$Y = \mu(X) + \varepsilon,$$

where

$$X \sim N(0, 1), \quad \varepsilon \sim N(0, 2), \quad \mu(x) = x|x| = \text{sgn}\,(x)x^2,$$

X and ε being independent. Based on a data set of size n we estimate the linear regression $\hat{a} + \hat{b}x$ by least-squares. We want to estimate

$$s_n = E\{(Y - \hat{a} - \hat{b}X)^2 | (X_i, Y_i), i = 1, \ldots, n\}.$$

We consider two sample sizes, $n = 12$ and $n = 24$. For v-fold cross-validation we randomly split the data in v groups and calculate the linear regression $\hat{a}_\alpha + \hat{b}_\alpha x$ by least-squares based on all the observations not in group α. The uncorrected v-fold cross-validated estimate of s_n is

$$\text{CV}_{nv} = \sum_{\alpha=1}^{v} (m_\alpha/n)\{\sum (Y_i - \hat{a}_\alpha - \hat{b}_\alpha X_i)^2/m_\alpha\},$$

where the sum inside the second bracket is over all the m_α observations in group α. The corrected v-fold cross-validated estimate of s_n is

$$\text{CV}^*_{nv} = \text{CV}_{nv} + \sum_{i=1}^{n} (Y_i - \hat{a} - \hat{b}X_i)^2/n - \sum_{\alpha=1}^{v} (m_\alpha/n)\left\{\sum_{i=1}^{n} (Y_i - \hat{a}_\alpha - \hat{b}_\alpha X_i)^2/n\right\}.$$

For the repeated learning-testing method, n_0 and m_0 are the sizes of the learning and the test sets and v is the number of times the data has been split. The uncorrected repeated learning-testing estimate is

$$\text{LT}_{nv} = v^{-1} \sum_{\alpha=1}^{v} \{\sum (Y_i - \hat{a}_\alpha - \hat{b}_\alpha X_i)^2/m_0\},$$

where the second sum inside the second bracket is over all the m_0 observations in the test set on the αth split of the data.

The corrected repeated learning-testing estimate is

$$\text{LT}^*_{nv} = \text{LT}_{nv} + \sum_{i=1}^{n} (Y_i - \hat{a} - \hat{b}X_i)^2/n - v^{-1} \sum_{\alpha=1}^{v} \left\{\sum_{i=1}^{n} (Y_i - \hat{a}_\alpha - \hat{b}_\alpha X_i)^2/n\right\}.$$

The simulations in Tables 1, 2 and 3 are based on 40 000 repeats. In Table 1 we have calculated $E(\text{CV}^*_{nv} - s_n)$, $E(\text{CV}_{nv} - s_n)$, and the standard deviations of $\text{CV}^*_{nv} - s_n$ and $\text{CV}_{nv} - s_n$ for various values of v. The case $v = n$ is evaluated only for the uncorrected case; CV_{nn} is the ordinary cross-validated estimate of s_n.

Table 1 shows that the bias of CV^*_{nv} is always smaller than that of CV_{nv} and CV^*_{nv} has a smaller variance than CV_{nv} even though the difference is quite small when $n = 24$ and $v > 3$. The simulations support the observations in §§ 2 and 3 that the performance of CV^*_{nv} is worst when $v = 2$. Though it seems that the bias for CV^*_{nv} initially decreases and then increases, there is no reason to support that simply because the standard error is about 0·01 when $n = 12$ and is about 0·004 when $n = 24$.

Table 2 gives results for the repeated learning-testing method for various values of $p_0 = m_0/n$, the proportion of observations in the test set. For each p_0, we consider two values of v, $v = [1/p_0]$ and $v = [2/p_0]$, where [.] is the integer part of a number. The reason for these two choices is as follows. If $p_0 = \frac{1}{3}$ and $v = [1/p_0] = 3$ then we are calculating

Table 1. *Biases and variances of uncorrected and corrected v-fold cross-validated estimates*

n	$E(s_n)$		$v=2$	$v=3$	$v=4$	$v=5$	$v=6$	$v=n$†
12	2·566	$E(\mathrm{CV}_{nv}-s_n)$	0·78	0·31	0·20	0·14	0·12	0·06
		$E(\mathrm{CV}^*_{nv}-s_n)$	0·16	0·05	0·04	0·01	0·03	
		st. dev. $(\mathrm{CV}_{nv}-s_n)$	2·66	1·79	1·62	1·54	1·49	1·41
		st. dev. $(\mathrm{CV}^*_{nv}-s_n)$	1·79	1·54	1·48	1·44	1·42	
24	2·321	$E(\mathrm{CV}_{nv}-s_n)$	0·25	0·11	0·07	0·05	0·04	0·02
		$E(\mathrm{CV}^*_{nv}-s_n)$	0·02	0·002	0·0001	0·0004	0·002	
		st. dev. $(\mathrm{CV}_{nv}-s_n)$	1·05	0·88	0·84	0·82	0·80	0·78
		st. dev. $(\mathrm{CV}^*_{nv}-s_n)$	0·86	0·81	0·80	0·79	0·78	

† The case for $v=n$ corresponds to ordinary cross-validation.

Table 2. *Biases and variances of uncorrected and corrected repeated learning-testing estimates*; *v, number of repeats*

			$p_0=\frac{1}{2}$		$p_0=\frac{1}{3}$		$p_0=\frac{1}{4}$		$p_0=\frac{1}{5}$		$p_0=\frac{1}{6}$	
n	$E(s_n)$		$v=2$	$v=4$	$v=3$	$v=6$	$v=4$	$v=8$	$v=5$	$v=10$	$v=6$	$v=12$
12	2·566	$E(\mathrm{LT}_{nv}-s_n)$	0·74	0·75	0·31	0·32	0·19	0·20	0·11	0·12	0·13	0·12
		$E(\mathrm{LT}^*_{nv}-s_n)$	0·14	0·14	0·05	0·06	0·03	0·04	0·02	0·03	0·04	0·02
		st. dev. $(\mathrm{LT}_{nv}-s_n)$	2·59	2·19	1·93	1·74	1·87	1·65	1·85	1·65	1·82	1·61
		st. dev. $(\mathrm{LT}^*_{nv}-s_n)$	1·78	1·59	1·68	1·53	1·71	1·53	1·76	1·58	1·73	1·54
24	2·321	$E(\mathrm{LT}_{nv}-s_n)$	0·26	0·24	0·12	0·12	0·08	0·08	0·05	0·04	0·04	0·04
		$E(\mathrm{LT}^*_{nv}-s_n)$	0·02	0·01	0·01	0·01	0·01	0·01	0·01	−0·003	−0·001	−0·001
		st. dev. $(\mathrm{LT}_{nv}-s_n)$	1·09	0·98	1·01	0·92	1·02	0·92	1·06	0·94	1·02	0·90
		st. dev. $(\mathrm{LT}^*_{nv}-s_n)$	0·93	0·85	0·95	0·87	0·98	0·88	1·03	0·91	0·99	0·88

the regression line three times for three learning sets. Three-fold cross-validation corresponds to the case $p_0=\frac{1}{3}$ and $v=3$ in the sense that test set sizes are about one-third of the data and the regression line has been calculated three times. When $p_0=\frac{1}{3}$ and $v=6$ we have 6 learning sets and consequently we have to estimate the regression line 6 times.

As we have discussed earlier, the variances of the repeated learning-testing estimates, corrected and uncorrected, decrease as v, the number of repeats, increases. However, as shown in (2·5), the biases do not depend on the number of repeats. The simulation results support these facts and also show that the bias of LT_{nv} is always larger than that of LT^*_{nv}. As for v-fold cross-validation, it appears that LT^*_{nv} has a smaller variance than LT_{nv}, but the difference seems to be quite small when $n=24$ and $v>3$. Comparison of Tables 1 and 2 shows that CV^*_{nv} and LT^*_{nv} have about the same biases, but $\mathrm{var}\,(\mathrm{CV}^*_{nv}) < \mathrm{var}\,(\mathrm{LT}^*_{nv})$.

For a fixed data set the value of the corrected and the uncorrected v-fold cross-validated estimates vary from partition to partition. Table 3 is based on a simulation study of this variability. For the quantities $E\{\mathrm{var}\,(\mathrm{CV}_{nv}|\mathscr{Z}_n)\}$, $E\{\mathrm{var}\,(\mathrm{CV}^*_{nv}|\mathscr{Z}_n)\}$ given in Table 3, the variance is first taken over random partitions for a given data set $\mathscr{Z}_n=\{Z_i=(Y_i, X_i)\colon i=1,\ldots,n\}$ and then the expectation is taken over $\mathscr{Z}_n$. Tables 1 and 3 clearly show that $E\{\mathrm{var}\,(\mathrm{CV}_{nv}|\mathscr{Z}_n)\}$ and $E\{\mathrm{var}\,(\mathrm{CV}^*_{nv}|\mathscr{Z}_n)\}$ are quite small compared to the variances of $\mathrm{CV}_{nv}-s_n$ and $\mathrm{CV}^*_{nv}-s_n$.

Table 3. *Variability of the estimates over the partitions*

	$n=12$					$n=24$				
	$v=2$	$v=3$	$v=4$	$v=5$	$v=6$	$v=2$	$v=3$	$v=4$	$v=5$	$v=6$
$E\{\mathrm{var}(\mathrm{CV}_{nv}\|\mathscr{Z}_n)\}$	0·046	0·037	0·029	0·025	0·032	0·014	0·011	0·008	0·008	0·006
$E\{\mathrm{var}\,(\mathrm{CV}^*_{nv}\|\mathscr{Z}_n)\}$	0·023	0·023	0·021	0·019	0·025	0·008	0·008	0·006	0·007	0·005

5. CONCLUSIONS

In summary whenever ordinary cross-validation is computationally very expensive, we can use v-fold cross-validation or the repeated learning-testing methods. If v is not large the uncorrected v-fold cross-validated estimate or the uncorrected repeated learning-testing estimate may have large biases. Thus we recommend that the correction terms given in § 2 should be routinely used.

The corrected v-fold cross-validated estimate is preferable to the corrected repeated learning-testing estimate because the former has a smaller variance.

The bias and the variance of the v-fold cross-validated estimate decrease as v increases. From a practical point of view, 2-fold corrected cross-validation should be avoided if possible.

For the corrected repeated learning-testing estimate, the bias depends on the ratio m_0/n_0, the ratio of the size of the test set to the learning set, and not on the number of repeats v. The smaller the ratio m_0/n_0, the smaller the bias. However, the smaller the ratio $n_0/(m_0 v)$, the smaller the variance.

6. SOME TECHNICAL RESULTS

This section gives a few technical results on the biases and the variances. More details appear in an unpublished technical report. We assume that for any z, $T(z, .)$ is von Mises differentiable with derivatives T_1 and T_2 (Serfling, 1980, Ch. 6). This condition can be relaxed a little. For all the results in this section, we keep only the dominant terms. All proofs we present here can be made rigorous with careful analysis (Beran, 1984; Serfling, 1984). Let $D_n = F_n - F$, $D_{n\alpha} = F_{n\alpha} - F$, $\bar{D}_{n\alpha} = \bar{F}_{n\alpha} - F$,

$$nE \int \{T(z, F_n) - T(z, F)\}\, dF(z) \doteq c_0,$$

$$-E \int \{T_1(z_1, z, F_n) - T_1(z_1, z, F)\}\, dD_n(z_1)\, dD_n(z) = c_1 n^{-2} + O(n^{-3}).$$

Note that the constants c_0 and c_1 depend on T and F.

THEOREM 6·1. *We have that*

(a) $E(\mathrm{CV}_{nv} - s_n) \doteq (v-1)^{-1} c_0/n$,
(b) $E(\mathrm{LT}_{nv} - s_n) \doteq (m_0/n_0) c_0/n$,
(c) $E(\mathrm{CV}^*_{nv} - s_n) \doteq (v-1)^{-1} c_1/n^2$,
(d) $E(\mathrm{LT}^*_{nv} - s_n) \doteq (m_0/n_0) c_1/n^2$,
(e) $E(\mathrm{RCV}^*_{nvt} - s_n) \doteq (v-1)^{-1} c_1/n^2$.

To find c_0 and c_1 in the regression case of Example 1, let us assume that the regression function μ is linear,

$$Y = a + bX + \varepsilon,$$

where ε and X are independent normal random variables, $E(\varepsilon) = 0$ and $\mathrm{var}(\varepsilon) = \sigma^2$. Then $c_0 = 2\sigma^2$ and $c_1 = 4\sigma^2$. Now

$$E(\mathrm{CV}_{nv} - s_n) \doteq 2\sigma^2 (v-1)^{-1} n^{-1}, \quad E(\mathrm{CV}_{nn} - s_n) \doteq 2\sigma^2 n^{-2},$$
$$E(\mathrm{CV}^*_{nv} - s_n) \doteq 4\sigma^2 (v-1)^{-1} n^{-2}.$$

If the regression function μ is not linear, but we fit a linear regression, then the expressions for c_0 and c_1 are not so simple because they include terms involving the bias

function $b(x)=\mu(x)-\mu_1(x)$, where $\mu_1(x)=a+bx$ is the least-squares linear model closest to μ.

Next, for any α, let

$$nE\left\{\int T(z)dD_n(z)\right\}^2=\gamma_0.$$

$$\tfrac{1}{2}(v-1)v^{-2}n^2E\left\{\int T_1(z_1,z)\,dD_{n\alpha}(z_1)\,d\bar{D}_{n\alpha}(z)+\int T_1(z_1,z)\,d\bar{D}_{n\alpha}(z_1)\,dD_{n\alpha}(z)\right\}^2=\gamma_1+O(n^{-1}).$$

$$(v-1)v^{-2}n^2E\left\{\int T(z)\,d\bar{D}_{n\alpha}(z)\right\}\times\left\{\int T_2(z_2,z_1,z)\,dD_{n\alpha}(z_2)\,dD_{n\alpha}(z_1)\,d\bar{D}_{n\alpha}(z)\right\}=\gamma_2+O(n^{-1}).$$

Then γ_0, γ_1 and γ_2 are constants depending on T and F and $\gamma_1>0$. Since $m_\alpha\doteq n/v$ for all α, two quantities described above are meaningful.

THEOREM 6·2. *We have*

(a) $\operatorname{var}(\mathrm{CV}^*_{nv}-s_n)\doteq\gamma_0/n+v(v-1)^{-1}\gamma_1/n^2+2\gamma_2/n^2$,
(b) $\operatorname{var}(\mathrm{LT}^*_{nv}-s_n)\doteq\{1+n_0/(m_0v)\}\gamma_0/n$,
(c) $\operatorname{var}(\mathrm{CV}_{nn}-s_n)\doteq\gamma_0/n+\gamma_1/n^2+2\gamma_2/n^2$,
(d) $\operatorname{var}(\mathrm{RCV}^*_{nvt}-s_n)\doteq\gamma_0/n+t(t-1)^{-1}(v-1)^{-1}\gamma_1/n^2+2\gamma_2/n^2$.

As a consequence of Theorem 6·2 we get the following result.

COROLLARY 6·3. *It follows that*

(a) $\operatorname{var}(\mathrm{CV}^*_{nv}-s_n)\doteq\operatorname{var}(\mathrm{CV}_{nn}-s_n)+(v-1)^{-1}\gamma_1/n^2$,
(b) $\operatorname{var}(\mathrm{LT}^*_{nv}-s_n)\doteq\operatorname{var}(\mathrm{CV}_{nn}-s_n)+\{n_0/(m_0v)\}\gamma_0/n$,
(c) $\operatorname{var}(\mathrm{LT}^*_{nv}-s_n)\doteq\operatorname{var}(\mathrm{CV}_{nv}-s_n)+\{n_0/(m_0v)\}\gamma_0/n$.

Theorem 6·2 and Corollary 6·3 show that LT^*_{nv} is a poorer estimate of s_n than CV^*_{nv}. Also $\operatorname{var}(\mathrm{CV}^*_{nv}-s_n)>\operatorname{var}(\mathrm{CV}_{nn}-s_n)$, but the difference becomes small as v increases.

The expressions for the variances of $\mathrm{CV}_{nv}-s_n$ and $\mathrm{LT}_{nv}-s_n$ are quite complicated. Here we give only the expression for $\operatorname{var}(\mathrm{LT}_{nv}-s_n)$. Let

$$\bar{T}_1(z_1)=\int T(z_1,z)/dF(z),$$

$$\gamma_3=nE\left\{\int\bar{T}_1(z_1)\,dD_n(z_1)\right\}^2,\quad\gamma_4=nE\left\{\int T(z)\,dD_n(z)\right\}\left\{\int\bar{T}_1(z_1)\,dD_n(z_1)\right\}.$$

THEOREM 6·4. *We have*

(a) $\operatorname{var}(\mathrm{CV}_{nv}-s_n)\doteq\operatorname{var}(\mathrm{CV}^*_{nv}-s_n)+O\{(v-1)^{-1}n^{-2}\}$,
(b) $\operatorname{var}(\mathrm{LT}_{nv}-s_n)\doteq\{1+n_0/(m_0v)\}\gamma_0/n+\{m_0/(n_0v)\}\gamma_3/n-2\gamma_4/(nv)$.

Corollary 6·3 shows that CV^*_{nv}, the corrected v-fold cross-validated estimate, has a larger variance than CV_{nn}, the ordinary cross-validated estimate of s_n. Even though we have been unable to show in our general setting that CV^*_{nv} has a smaller variance than

CV_{nv}, we believe this to be so. In the regression case when the regression function μ is linear, $\gamma_0 = 2\sigma^4$, $\gamma_1 = 16\sigma^4$ and $\gamma_2 = \gamma_3 = \gamma_4 = 0$. Then the following can be shown:

$$\text{var}\,(\text{CV}_{nn} - s_n) \doteq 2\sigma^4\{n^{-1} + 8n^{-2}\},$$

$$\text{var}\,(\text{CV}^*_{nv} - s_n) \doteq 2\sigma^4\{n^{-1} + 8n^{-2} + 8(v-1)^{-1}n^{-2}\},$$

$$\text{var}\,(\text{CV}_{nv} - s_n) \doteq 2\sigma^4\{n^{-1} + 8n^{-2} + 8(v-1)^{-1}n^{-2} + 4(v-1)^{-2}n^{-2} + 2(v-1)^{-3}n^{-2}\}.$$

As we have discussed earlier, if the regression function μ is not linear and we are fitting a linear model, then the expressions for $\gamma_0, \ldots, \gamma_4$ involve the model bias.

ACKNOWLEDGEMENT

It is a pleasure to thank Professors Leo Breiman and Charles Stone for many helpful discussions. Thanks are also due to two referees whose suggestions have led to the improvement in the presentation of this paper.

APPENDIX

Outline of proofs of theorems

For Theorem 6·1, note for (a) that $\text{CV}_{nv} - s_n$ can be written as

$$\int T(z)\, dD_n(z) + \sum p_\alpha \int \{T(z, F_{n\alpha}) - T(z, F)\}\, d\bar{D}_{n\alpha}(z)$$

$$+ \sum p_\alpha \int \{T(z, F_{n\alpha}) - T(z, F_n)\}\, dF(z) = I_1 + I_2 + I_3, \quad \text{(A·1)}$$

say. Now $E(I_1) = E(I_2) = 0$, whereas

$$I_3 = \sum p_\alpha \int \{\bar{T}_1(z_1)\, dD_{n\alpha}(z_1) - \bar{T}_1(z_1)\, dD_n(z_1)\}$$

$$+ \sum p_\alpha \int \{\bar{T}_2(z_2, z_1)\, dD_{n\alpha}(z_2)\, dD_{n\alpha}(z_1) - \bar{T}_2(z_2, z_1)\, dD_n(z_2)\, dD_n(z_1)\}.$$

The first term in the last expression is zero since $\sum p_\alpha D_{n\alpha} = D_n$. Since $p_\alpha \doteq v^{-1}$ and $n_\alpha \doteq (v-1)v^{-1}n$ for all α, the expectation of the second term approximately is

$$\sum p_\alpha (1/n_\alpha - 1/n)c_0 \doteq (v-1)^{-1}c_0/n.$$

The proof of (b) is the same as in (a).

For (c) note that

$$\text{CV}^*_{nv} - s_n = \int T(z)\, dD_n(z) + \sum p_\alpha \int \{T(z, F_{n\alpha}) - T(z, F)\}\, d\bar{D}_{n\alpha}(z)$$

$$+ \sum p_\alpha \int \{T(z, F_n) - T(z, F_{n\alpha})\}\, dD_n(z)$$

$$= I_4 + I_5 + I_6, \quad \text{(A·2)}$$

say. Now $E(I_4) = E(I_5) = 0$. The result holds since it can be shown that

$$E(I_6) \doteq (v-1)^{-1}n^{-2}c_1 + O(n^{-3}).$$

The proofs of (d) and (e) follow as in (c).

For (a) of Theorem 6·2, as in (A·2), $\text{CV}^*_{nv} - s_n = I_4 + I_5 + I_6$. Since $E(\text{CV}^*_{nv} - s_n) = O(n^{-2})$ as given in Theorem 6·1 and since we are keeping terms of order n^{-2} and higher,

$$\text{var}\,(\text{CV}^*_{nv} - s_n) \doteq E\{(\text{CV}^*_{nv} - s_n)^2\}.$$

Now,

$$E\{(\text{CV}^*_{nv} - s_n)^2\} = E(I_4^2) + E(I_5^2) + E(I_6^2) + 2E(I_4 I_5) + 2E(I_4 I_6) + 2E(I_5 I_6).$$

The result is proved via

$$E(I_4^2) = \gamma_0/n, \quad E(I_5^2) \doteq v(v-1)^{-1}\gamma_1/n^2,$$

$$E(I_6^2) \doteq O(n^{-3}), \quad E(I_4 I_5) \doteq v(v-1)^{-1}\gamma_2/n^2,$$

$$E(I_4 I_6) \doteq -(v-1)^{-1}\gamma_2/n^2, \quad E(I_5 I_6) \doteq O(n^{-3}).$$

For (b), the proof follows from part (a) of Theorem 6·4 by putting $v = n$.

For (c) note that $\text{LT}^*_{nv} - s_n$ can be written as

$$v^{-1}\sum \int T(z)\, d\bar{D}_{n\alpha}(z) + v^{-1}\sum \int \{T(z, F_{n\alpha}) - T(z, F)\}\, d\bar{D}_{n\alpha}(z)$$

$$+ v^{-1}\sum \int \{T(z, F_n) - T(z, F_{n\alpha})\}\, dD_n(z) = I_7 + I_8 + I_9,$$

say. The result holds since it can be shown that

$$E\{(\text{LT}_{nv} - s_n)^2\} = E(I_7^2) + O(n^{-2}) = \{1 + n_0/(m_0 v)\}\gamma_0 n^{-1} + O(n^{-2}).$$

The proof of (d) is the same as in parts (a) and (c).

For Theorem 6·4 the proof of (a) follows using (A·1) and calculations similar to the ones in Theorem 6·2.

For (b) note that

$$\text{LT}_{nv} - s_n = v^{-1}\sum \int T(z)\, d\bar{D}_{n\alpha}(z) + v^{-1}\sum \int \{T(z, F_{n\alpha}) - T(z, F)\}\, d\bar{D}_{n\alpha}(z)$$

$$+ v^{-1}\sum \int \{T(z, F_{n\alpha}) - T(z, F_n)\}\, dF(z)$$

$$= I_{10} + I_{11} + I_{12},$$

say. Since we are interested only in the terms of order n^{-1},

$$E(\text{LT}_{nv} - s_n) \doteq E(I_{10}^2) + E(I_{12}^2) + 2E(I_{10} I_{12}).$$

The proof follows from

$$E(I_{10}^2) = \{1 + n_0/(m_0 v)\}\gamma_0 n^{-1}, \quad E(I_{10} I_{12}) \doteq -(nv)^{-1}\gamma_4, \quad E(I_{12}^2) \doteq \{m_0/(n_0 v)\}\gamma_3/n.$$

References

BERAN, R. J. (1984). Jackknife approximation to bootstrap estimates. *Ann. Statist.* **12**, 101–18.

BOWMAN, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71**, 353–60.

BREIMAN, L. & FRIEDMAN, J. (1985). Estimating optimal transformations for regression and correlation. *J. Am. Statist. Assoc.* **80**, 580–619.

BREIMAN, L., FRIEDMAN, J., OLSHEN, R. & STONE, C. (1984). *Classification and Regression Trees.* Belmont, California: Wadsworth.

BURMAN, P. (1989). Estimation of optimal transformations using v-fold cross validation and repeated learning-testing methods. *Sankhyā* A **51**. To appear.
GEISSER, S. (1975). The predictive sample reuse method with applications. *J. Am. Statist. Assoc.* **70**, 320–28.
HÄRDLE, W. & MARRON, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13**, 1465–81.
SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: Wiley.
SERFLING, R. J. (1984). Generalized L-, M-, and R-statistics. *Ann. Statist* **12**, 76–86.
STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. R. Statist. Soc.* B **36**, 111–47.
STONE, M. (1977). Cross validation: a review. *Math. Oper. Statist., ser. Statist.* **9**, 127–39.

[*Received April* 1988. *Revised January* 1989]