Research Report

Cross-Validation and *U*-Statistics

ETS RR–19-27

Shelby J. Haberman

December 2019

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Cross-Validation and *U*-Statistics

Shelby J. Haberman

Consultant, Jerusalem, Israel

Cross-validation is a common statistical procedure applied to problems that are otherwise computationally intractable. It is often employed to assess the effectiveness of prediction procedures. In this report, cross-validation is discussed in terms of U-statistics. This approach permits consideration of the statistical properties of cross-validation as an approach to assessment of prediction accuracy.

Keywords Resampling; variance estimation; jackknifing

doi:10.1002/ets2.12263

Cross-validation methods have a long history in statistics (Geisser, 1975; Golub, Heath, & Wahba, 1979; Picard & Cook, 1984; Stone, 1974). This report relates cross-validation to the theory of U-statistics (Hoeffding, 1948). This connection permits use of cross-validation both to produce estimates of mean-squared error of prediction and to assess the variability of these estimates of mean-squared error. In the section Cross-Validation Procedure Under Study, the cross-validation approach under study is developed and illustrated with a few simple examples. In the section Use of U-Statistics, application of U-statistics to cross-validation is introduced and illustrated. The section Conclusions includes concluding remarks.

Cross-Validation Procedure Under Study

In the problems under study, a simple random sample is used to construct a prediction of a real predicted random variable Y_0 with values in a set $\mathcal{Y}$ by a predicting random vector $\mathbf{X}_0$ of finite dimension $p \geq 1$ with values in a set $\mathcal{X}$. The procedures under study are designed to estimate the accuracy of the prediction when neither $\mathbf{X}_0$ nor Y_0 has been observed. The methods considered make minimal assumptions concerning the joint distribution of $\mathbf{X}_0$ and Y_0, and they are designed to apply both to small samples and to large samples. Thus they do not rely on large-sample approximations. To develop the desired analysis of cross-validation, consider mutually independent and identically distributed pairs $(\mathbf{X}_i, Y_i)$, $i \geq 0$. Because mean-squared error is emphasized in this report to simplify analysis, assume that Y_0 has finite and positive variance $\sigma^2(Y_0)$. Because illustrations often are related to linear regression, assume that $\mathbf{X}_0$ has a finite and positive-definite covariance matrix $\mathrm{Cov}(\mathbf{X}_0)$. To avoid degenerate cases, assume that no constant p-dimensional vector β and real constant α exists such that $Y_0 = \alpha + \beta'\mathbf{X}_0$ with probability 1. Here $\beta'\mathbf{X}_0 = \sum_{j=1}^{p} \beta_j X_{j0}$ if β_j is element j of β and X_{j0} is element j of $\mathbf{X}_0$ for $1 \leq j \leq p$.

For sample sizes n at least as great as some minimum sample size n^*, the desired prediction based on $\mathbf{X}_i$, $0 \leq i \leq n$, and Y_i, $1 \leq i \leq n$, relies on a real function $g_n(\mathbf{T}, \mathbf{u}, \mathbf{t})$ defined for p by n matrices $\mathbf{T}$ with columns $\mathbf{T}_i$ in $\mathcal{X}$ for $1 \leq i \leq n$, n-dimensional vectors $\mathbf{u}$ with elements u_i in $\mathcal{Y}$ for $1 \leq i \leq n$, and p-dimensional vectors $\mathbf{t}$ in $\mathcal{X}$. The function g_n is assumed to satisfy the symmetry condition that for any permutation ω on the set $\overline{n}$ of positive integers no greater than n, $g_n(\mathbf{T}_\omega, \mathbf{u}_\omega, \mathbf{t}) = g_n(\mathbf{T}, \mathbf{u}, \mathbf{t})$, where $\mathbf{T}_\omega$ is the $p \times n$ matrix with column i equal to $\mathbf{T}_{\omega(i)}$ for $1 \leq i \leq n$ and $\mathbf{u}_\omega$ is the n-dimensional vector with element i equal to $u_{\omega(i)}$ for $1 \leq i \leq n$. The function g_n satisfies the basic regularity condition that $g_n(\mathbf{U}, \mathbf{V}, \mathbf{W})$ is a random variable whenever $\mathbf{U}$ is a $p \times n$ random matrix with columns in $\mathcal{X}$, $\mathbf{V}$ is an n-dimensional random vector with elements in $\mathcal{Y}$, and $\mathbf{V}$ is a p-dimensional random vector in $\mathcal{X}$. Let $\widetilde{\mathbf{X}}_n$ be the $p \times n$ matrix with columns $\mathbf{X}_i$, $1 \leq i \leq n$, and let $\mathbf{Y}_n$ be the n-dimensional vector with elements Y_i, $1 \leq i \leq n$. Then $\widehat{E}_{n0} = g_n\left(\widetilde{\mathbf{X}}_n, \mathbf{Y}_n, \mathbf{X}_0\right)$ predicts Y_0. The error of prediction is then $r_{n0} = Y_0 - \widehat{E}_{n0}$, and the mean-squared error is $\mathrm{MSE}\left(Y_0, \widehat{E}_{n0}\right) = E\left(r_{n0}^2\right)$. The symmetry assumption implies that the prediction that uses the sample $(\mathbf{X}_i, Y_i)$ for $1 \leq i \leq n$ is the same as the prediction that uses the sample $(\mathbf{X}_{\omega(i)}, Y_{\omega(i)})$ for $1 \leq i \leq n$. Thus the order of observation does not matter.

Corresponding author: S. J. Haberman, E-mail: haberman.statistics@gmail.com

The mean-squared error may be decomposed into components by observing that Y_0 and $\widehat{E}_{n0}$ are conditionally independent given $\mathbf{X}_0$ and $E(Z|\,\mathbf{X}_0)$ and $Z-E(Z|\,\mathbf{X}_0)$ are uncorrelated if the random variable Z has a finite variance (Blackwell, 1947). Here the random variable $E(Z|\,\mathbf{X}_0)$ is the conditional expectation of Z given $\mathbf{X}_0$. It has expectation 0 and variance $\sigma^2(Z|\,\mathbf{X}_0)$. It follows that

$$\mathrm{MSE}\left(Y_0,\widehat{E}_{n0}\right)=\left[E\left(Y_0\right)-E\left(\widehat{E}_{n0}\right)\right]^2+\sigma^2\left(E\left(r_0|\mathbf{X}_0\right)\right)+\sigma^2\left(\widehat{E}_{n0}|\mathbf{X}_0\right)+\sigma^2\left(Y_0|\mathbf{X}_0\right).$$

Thus $\mathrm{MSE}\left(Y_0,\widehat{E}_{n0}\right)\geq\sigma^2\left(Y|\mathbf{X}_0\right)>0$. The mean-squared error $\mathrm{MSE}\left(Y_0,\widehat{E}_{n0}\right)$ does not approach 0 no matter how large the sample may be or how well $\widehat{E}_{n0}$ is chosen.

The problem treated with cross-validation is the estimation of $\mathrm{MSE}\left(Y_0,\widehat{E}_{n0}\right)$ when the sample observations $(\mathbf{X}_i, Y_i)$, $1\le i\le n$, are observed but the predicting variable $\mathbf{X}_0$ and the predicted variable Y_0 are not available. In some very elementary cases, this estimation can be accomplished without difficulty by use of $\widehat{E}_{nk}=g_n\left(\widetilde{\mathbf{X}}_n,\mathbf{Y}_n,\mathbf{X}_k\right)$ and $r_{nk}=Y_k-\widehat{E}_{nk},\ \ 1\leq k\leq n$.

Example 1. Let $n^*=2$, and let $g_n\left(\mathbf{T},\mathbf{u},\mathbf{t}\right)=n^{-1}\sum_{i=1}^{n}u_i$ for $n\ge n^*$ so that, for $0\leq k\leq n$, $\widehat{E}_{nk}$ is the sample mean $\overline{Y}_n$ of the Y_i, $1\le i\le n$. In this case, the vectors $\mathbf{X}_i$, $0\le i\le n$, are ignored. Let $s_n^2=(n-1)^{-1}\sum_{i=1}^{n}r_{nk}^2$ be the sample variance of the Y_i, $1\le i\le n$. Then $\mathrm{MSE}\left(Y_0,\widehat{E}_{n0}\right)=\left(1+n^{-1}\right)\sigma^2\left(Y_0\right)$ has the unbiased estimate $\overline{\mathrm{MSE}}_n\left(Y_0,\widehat{E}_{n0}\right)=\left(1+n^{-1}\right)s_n^2$. As the sample size n increases, the strong law of large numbers implies that $\overline{\mathrm{MSE}}_n\left(Y_0,\widehat{E}_{n0}\right)$ converges with probability 1 to $\sigma^2(Y_0)$.

Example 2. Consider prediction based on linear regression when the joint distribution of $(\mathbf{X}_0, Y_0)$ is multivariate normal. Let $n\ge n^*=p+3$. For p-dimensional vector $\mathbf{b}$ with elements b_j for $1\le j\le p$ and a p-dimensional vector $\mathbf{d}$ with elements d_j for $1\le j\le p$, let the inner product $\mathbf{b}'\,\mathbf{d}$ of $\mathbf{b}$ and $\mathbf{d}$ be $\sum_{i=1}^{p}b_jd_j$. Let $S_{n-}(\mathbf{T},\mathbf{u})$ be the minimum value of the sum of squares

$$S_n\left(c,\mathbf{d},\mathbf{T},\mathbf{u}\right)=\sum_{i=1}^{n}\left(u_i-c-\mathbf{d}'\mathbf{T}_i\right)^2$$

for real c and p-dimensional vectors $\mathbf{d}$. Let $M_n(\mathbf{T},\mathbf{u})$ be the set of pairs $(c,\mathbf{d})$ such that $S_n(c,\mathbf{d},\mathbf{T},\mathbf{u})=S_{n-}(\mathbf{T},\mathbf{u})$. If no real number c and p-dimensional vector $\mathbf{d}$ exist such that $c+\mathbf{T}_i{}'\,\mathbf{d}=0$ for $1\le i\le n$ and some element of $\mathbf{d}$ is not 0, then $M_n(\mathbf{T},\mathbf{u})$ has only one element. Under the assumption of multivariate normality, the probability is 1 that $M_n\left(\widetilde{\mathbf{X}}_n,\mathbf{Y}_n\right)$ has only one element. Nonetheless, for completeness, let $\|\mathbf{d}\|^2=\mathbf{d}'\,\mathbf{d}$, and let $\mu_n(\mathbf{T},\mathbf{u})$ be the minimum value of $c^2+\|\mathbf{d}\|^2$ for $(c,\mathbf{d})$ in $M_n(\mathbf{T},\mathbf{u})$. Then there is a unique real number $a_n(\mathbf{T},\mathbf{u})$ and a unique p-dimensional vector $\mathbf{b}_n(\mathbf{T},\mathbf{u})$ such that $(a_n(\mathbf{T},\mathbf{u}),\mathbf{b}_n(\mathbf{T},\mathbf{u}))$ is in $M_n(\mathbf{T},\mathbf{u})$ and

$$\left[a_n\left(\mathbf{T},\mathbf{u}\right)\right]^2+\left\|\mathbf{b}_n\left(\mathbf{T},\mathbf{u}\right)\right\|^2=\mu_n\left(\mathbf{T},\mathbf{u}\right)$$

(Rao & Mitra, 1972). Let $g_n(\mathbf{T},\mathbf{u},\mathbf{t})=a_n(\mathbf{T},\mathbf{u})+[\mathbf{b}_n(\mathbf{T},\mathbf{u})]'\,\mathbf{t}$ so that

$$\widehat{E}_{nk}=a_n\left(\widetilde{\mathbf{X}}_n,\mathbf{Y}_n\right)+\left[\mathbf{b}_n\left(\widetilde{\mathbf{X}}_n,\mathbf{Y}_n\right)\right]'\mathbf{X}_k$$

for $0\le k\le n$. Let $s_n^2=\left(n-p-1\right)^{-1}\sum_{k=1}^{n}r_{nk}^2$ be the residual mean square error for the linear regression of Y_i on $\mathbf{X}_i$ for $1\le i\le n$ so that $E\left(s_n^2\right)=\sigma^2\left(Y_0|\mathbf{X}_0\right)$. Results for the T^2 statistic (Hotelling, 1931) and the standard result that a random variable with an F distribution with p and $n-p$ degrees of freedom has expectation $(n-p)/(n-p-2)$ lead to

$$\mathrm{MSE}\left(Y_0,\widehat{E}_{n0}\right)=\sigma^2\left(Y_0|\mathbf{X}_0\right)\left(1+\frac{1}{n}+\frac{p}{n-p-2}\right).$$

It follows that

$$\overline{\mathrm{MSE}}_n\left(Y_0,\widehat{E}_{n0}\right)=s_n^2\left(1+\frac{1}{n}+\frac{p}{n-p-2}\right)$$

is an unbiased estimate of the mean-squared error $\mathrm{MSE}\left(Y_0,\widehat{E}_{n0}\right)$. If n becomes large and p remains constant, then $\mathrm{MSE}\left(Y_0,\widehat{E}_{n0}\right)$ converges to $\sigma^2(Y_0,\mathbf{X}_0)$ and $\overline{\mathrm{MSE}}_n\left(Y_0,\widehat{E}_{n0}\right)$ converges to $\sigma^2(Y_0,\mathbf{X}_0)$ with probability 1. Nonetheless, it is important to note that $\mathrm{MSE}\left(Y_0, \widehat{E}_{n0}\right)/\sigma^2\left(Y_0|\mathbf{X}_0\right)$ is approximately $n/(n-p)$ if p and n are both large. In this case, if $p\ /\ n$ is not close to 0, then a substantial fraction of the mean-squared error is due to statistical error associated with estimation of the predictor rather than error inherent in prediction of Y from $\mathbf{X}$.

Despite Examples 1 and 2, very few cases exist in which mean-squared error is readily estimated without either large-sample approximations or strong restrictions on the joint distribution of the predictor $\mathbf{X}_0$ and the predicted variable Y_0. The following example for ridge regression can be helpful.

Example 3. Consider prediction based on ridge regression for $n^* = 1$. For each integer $n \geq 1$, let λ_n be a positive real number. Let $S_{n-}(\mathbf{T}, \mathbf{u})$ be the minimum value of the sum of squares

$$S_n\left(c, \mathbf{d}, \mathbf{T}, \mathbf{u}\right) = \lambda_n \left\|\mathbf{d}\right\|^2 + \sum_{i=1}^{n}\left(u_i - c - \mathbf{d}'\mathbf{T}_i\right)^2$$

over real c and p-dimensional vectors d. There is a unique real number $a_n(\mathbf{T}, \mathbf{u})$ and a unique p-dimensional vector $\mathbf{b}_n(\mathbf{T}, \mathbf{u})$ such that

$$S_n\left(a_n\left(\mathbf{T}, \mathbf{u}\right), \mathbf{b}_n\left(\mathbf{T}, \mathbf{u}\right), \mathbf{T}, \mathbf{u}\right)) = S_{n-}\left(\mathbf{T}, \mathbf{u}\right).$$

As in Example 2, let $g_n(\mathbf{T}, \mathbf{u}, \mathbf{t}) = a_n(\mathbf{T}, \mathbf{u}) + [\mathbf{b}_n(\mathbf{T}, \mathbf{u})]'\,\mathbf{t}$ so that

$$\widehat{E}_{nk} = a_n\left(\widetilde{\mathbf{X}}_n, \mathbf{Y}_n\right) + \left[\mathbf{b}_n\left(\widetilde{\mathbf{X}}_n, \mathbf{Y}_n\right)\right]'\mathbf{X}_k$$

for $0 \leq k \leq n$. In this case, no simple expression for $\mathrm{MSE}\left(Y_0, \widehat{E}_{n0}\right)$ appears to be available even under multivariate normality.

Cross-validation procedures seek to provide estimates of the mean-squared error $\mathrm{MSE}\left(Y_0, \widehat{E}_{n0}\right)$ that avoid large-sample approximations and restrictive assumptions.

In elementary cross-validation, the available observations $(\mathbf{X}_i, Y_i)$, $1 \leq i \leq n$, are divided into two groups sometimes termed a training sample $(\mathbf{X}_i, Y_i)$, i in A, and an evaluation sample $(\mathbf{X}_i, Y_i)$, i in B, where A and B are nonempty disjoint subsets of the set $\overline{n}$ of positive integers no greater than n such that A and B have union n and A has $m \geq n^*$ elements. For an integer k in B, the pairs $(\mathbf{X}_i, Y_i)$, i in A, and $\mathbf{X}_k$ are used to provide a prediction of Y_k. Let $\Pi(m, A)$ be the set of one-to-one functions from $\overline{m}$ onto A. For π in $\Pi(m, A)$, let $\widetilde{\mathbf{X}}_\pi$ be the $p \times m$ matrix with columns $\mathbf{X}_{\pi(i)}$ for $1 \leq i \leq m$, and let $\mathbf{Y}_\pi$ be the m-dimensional vector with elements $Y_{\nu(i)}$ for $1 \leq i \leq m$. Then $g_m\left(\widetilde{\mathbf{X}}_\pi, \mathbf{Y}_\pi, \mathbf{X}_k\right)$ has the same value $\widehat{E}_{mk}(A)$ for all π in $\Pi(m, A)$, and $\widehat{E}_{mk}(A)$ has the same distribution as $\widehat{E}_{m0}$. In addition, $r_{mk}(A) = Y_k - \widehat{E}_{mk}(A)$ has the same distribution as r_{m0}. It follows that the mean-squared error $\mathrm{MSE}\left(Y_0, \widehat{E}_{m0}\right)$ has the unbiased estimate $[r_{nk}(A)]^2$. If $\overline{\mathrm{MSE}}_n\left(Y_0, \widehat{E}_{m0}; A\right)$ is the average of $[r_{mk}(A)]^2$ over integers k in B, then $\overline{\mathrm{MSE}}_n\left(Y_0, \widehat{E}_{m0}; A\right)$ is also an unbiased estimate of $\mathrm{MSE}\left(Y_0, \widehat{E}_{m0}\right)$.

Unfortunately, $\mathrm{MSE}\left(Y_0, \widehat{E}_{n0}\right)$ is not normally the same as $\mathrm{MSE}\left(Y_0, \widehat{E}_{m0}\right)$. Consider the following examples.

Example 4. In Example 1, $\mathrm{MSE}\left(Y_0, \widehat{E}_{m0}\right) = \sigma^2\left(Y_0\right)\left(1 + m^{-1}\right)$ is larger than $\mathrm{MSE}\left(Y_0, \widehat{E}_{n0}\right)$. Nonetheless, if m is large, then $\mathrm{MSE}\left(Y_0, \widehat{E}_{m0}\right) - \mathrm{MSE}\left(Y_0, \widehat{E}_{n0}\right)$ is small.

Example 5. In Example 2,

$$\mathrm{MSE}\left(Y_0, \widehat{E}_{m0}\right) = \sigma^2\left(Y_0|\mathbf{X}_0\right)\left(1 + \frac{1}{m} + \frac{p}{m-p-2}\right) > \mathrm{MSE}\left(Y_0, \widehat{E}_{n0}\right).$$

For fixed p, if m is large, $\mathrm{MSE}\left(Y_0, \widehat{E}_{m0}\right) - \mathrm{MSE}\left(Y_0, \widehat{E}_{n0}\right)$ is small; however, it should be noted that $\left[\mathrm{MSE}\left(Y_0, \widehat{E}_{m0}\right) - \mathrm{MSE}\left(Y_0, \widehat{E}_{n0}\right)\right]/\sigma^2\left(Y_0|\mathbf{X}_0\right)$ can be made arbitrarily large by letting $m-p-2$ not increase while letting p increase.

A practical complication generally encountered in the choice of m and n is the typical trade-off between a decreasing value of $\mathrm{MSE}\left(Y_0, \widehat{E}_{m0}\right)$ and an increasing variance of $\overline{\mathrm{MSE}}_n\left(Y_0, \widehat{E}_{m0}; A\right)$ as m approaches n. Assume that Y_0 and $\widehat{E}_{m0}$ have finite fourth moments, so that r_{m0}^2 has a finite variance. The variance of $r_{mk}^2(A)$ is $\sigma^2\left(r_{m0}^2\right)$ for k in B, and the covariance of $r^2_{mk}(A)$ and $r^2_{mk\prime}(A)$ is the variance $\sigma^2(E\left(r^2_{m0}|\widetilde{\mathbf{X}}_m, \mathbf{Y}_m\right)$ of the conditional expectation $E\left(r^2_{m0}|\widetilde{\mathbf{X}}_m, \mathbf{Y}_m\right)$ of r^2_{m0} given $(\mathbf{X}_i, Y_i)$, $1 \le i \le m$. In addition,

$$\sigma^2\left(r^2_{m0}\right) = \sigma^2\left(E\left(r^2_{m0}|\widetilde{\mathbf{X}}_m, \mathbf{Y}_m\right)\right) + \sigma^2\left(r^2_{m0}|\widetilde{\mathbf{X}}_m, \mathbf{Y}_m\right)),$$

where $\sigma^2\left(r^2_{m0}|\widetilde{\mathbf{X}}_m\right)$ is the variance of $r^2_{m0} - E\left(r^2_{m0}|\widetilde{\mathbf{X}}_m, \mathbf{Y}_m\right)$ (Blackwell, 1947). It follows that

$$\begin{aligned}\sigma^2\left(\overline{\mathrm{MSE}}_n\left(Y_0, \widehat{E}_{m0}; A\right)\right) &= (n-m)^{-1}\left[\sigma^2\left(r^2_{m0}\right) + (n-m-1)\,\sigma^2\left(E\left(r^2_{m0}|\widetilde{\mathbf{X}}_m, \mathbf{Y}_m\right)\right)\right] \\ &= (n-m)^{-1}\,\sigma^2\left(r^2_{m0}|\widetilde{\mathbf{X}}_m, \mathbf{Y}_m\right) + \sigma^2\left(E\left(r^2_{m0}|\widetilde{\mathbf{X}}_m, \mathbf{Y}_m\right)\right).\end{aligned}$$

For a given m, the variance $\sigma^2\left(\overline{\mathrm{MSE}}_n\left(Y_0, \widehat{E}_{m0}; A\right)\right)$ decreases to $\sigma^2\left(E\left(r^2_{m0}|\widetilde{\mathbf{X}}_m, \mathbf{Y}_m\right)\right)$ as n increases.

Example 6. Consider Example 1 for Y with a finite fourth moment. Here $r_{mk}(A) = Y_k - \widehat{E}_{mk}(A)$, where $\widehat{E}_{mk}(A)$ is the sample mean of Y_i for i in A. Let $\kappa_4(Y_0)$ be the fourth cumulant of Y_0. Let $\sigma^4(Y_0)$ be $[\sigma^2(Y_0)]^2$. Let

$$D_{m0} = \widehat{E}_{m0} - E(Y) = m^{-1}\sum_{i=1}^{m}\left[Y_i - E(Y)\right].$$

The conditional expectation of r^2_{m0} given $(\mathbf{X}_i, Y_i)$, $1 \le i \le m$, is $\sigma^2\left(Y_0\right) + D^2_{m0}$ so that

$$\sigma^2\left(E\left(r^2_{m0}|\widetilde{\mathbf{X}}_m, \mathbf{Y}_m\right)\right) = \sigma^2\left(D^2_{m0}\right) = \frac{2\sigma^4\left(Y_0\right)}{m} + \frac{\kappa_4\left(Y_0\right)}{m^3}$$

(Fisher, 1930). The difference $r^2_{m0} - E\left(r^2_{m0}|\widetilde{\mathbf{X}}_m, \mathbf{Y}_m\right)$ is $[Y_0 - E(Y_0)]^2 - \sigma^2(Y_0)$ so that $\sigma^2\left(r^2_{m0}|\widetilde{\mathbf{X}}_m, \mathbf{Y}_m\right) = 2\sigma^4\left(Y_0\right) + \kappa_4\left(Y_0\right)$. It follows that

$$\sigma^2\left(\overline{\mathrm{MSE}}_n\left(Y_0, \widehat{E}_{m0}; A\right)\right) = 2\frac{n}{m\,(n-m)}\sigma^4(Y) + \frac{m^3 + n - m}{(n-m)\,m^3}\kappa_4(Y).$$

This variance only approaches 0 if both $n - m$ and m approach ∞. If Y has a normal distribution, then $\kappa_4(Y) = 0$, so that

$$\sigma^2\left(\overline{\mathrm{MSE}}_n\left(Y_0, \widehat{E}_{m0}; A\right)\right) = 2\frac{n}{m\,(n-m)}\sigma^4(Y).$$

With or without normality, if n goes to ∞ and m / n approaches a positive constant $f < 1$, then $\sigma^2\left(\overline{\mathrm{MSE}}_n\left(Y_0, \widehat{E}_{m0}; A\right)\right)$ approaches

$$\frac{2\sigma^4(Y)}{f\,(1-f)} + \frac{\kappa_4(Y)}{1-f}.$$

In the normal case, the limit is smallest if $f = 1/2$.

Example 7. In Example 2, an argument similar to that in Example 6 together with use of the variance formula for the F distribution shows that for $m - p > 4$,

$$\sigma^2\left(\overline{\mathrm{MSE}}_n\left(Y_0, \widehat{E}_{m0}; A\right)\right) = 2\frac{n}{m\,(n-m)}\left[1 + \frac{(m-2)\,p}{(m-p-2)^2\,(m-p-4)}\right]\sigma^4\left(Y_0|\mathbf{X}_0\right),$$

where $\sigma^4(Y_0|\,\mathbf{X}_0)$ is the square of $\sigma^2(Y_0|\,\mathbf{X}_0)$. For fixed dimension p, if n approaches ∞ and m / n approaches a positive $f < 1$, then $n\sigma^2\left(\overline{\mathrm{MSE}}_n\left(Y_0, \widehat{E}_{m0}; A\right)\right)$ approaches $2\sigma^4(Y_0|\,\mathbf{X}_0)/[f(1-f)]$. Nonetheless, $\sigma^2\left(\overline{\mathrm{MSE}}_n\left(Y_0, \widehat{E}_{m0}; A\right)\right)/\sigma^4\left(Y_0|\mathbf{X}_0\right)$ becomes arbitrarily large if m / n approaches $f < 1$, n approaches ∞, and $m - p$ remains bounded.

Use of *U*-Statistics

Greater efficiency in estimation of $\text{MSE}\left(Y_0, \widehat{E}_{m0}\right)$ can be achieved by use of U-statistics (Hoeffding, 1948). These statistics provide a general approach to construction of unbiased estimates, determination of their variances, and estimation of their variances. For some integer $C \geq 1$, let the real function $h(\mathbf{T}, \mathbf{u})$ be defined for $p \times C$ matrices $\mathbf{T}$ with columns in $\mathcal{X}$ and C-dimensional vectors $\mathbf{u}$ with elements in $\mathcal{Y}$. Assume that $h(\mathbf{W}, \mathbf{Z})$ is a random variable whenever $\mathbf{W}$ is a $p \times C$ random matrix with columns in $\mathcal{X}$ and $\mathbf{Z}$ is a C-dimensional random variable with elements in $\mathcal{Y}$. For $n \geq C$, let the U-statistic U_n defined by h be the average of $h\left(\widetilde{\mathbf{X}}_\pi, \mathbf{Y}_\pi\right)$ over functions π in $\Pi\left(C, \overline{n}\right)$ from $\overline{C}$ to $\overline{n}$. Alternatively, let $\mathcal{A}\left(C, n\right)$ be the class of subsets A of $\overline{n}$ with C elements, and let U_A, A in $\mathcal{A}\left(C, n\right)$ be the average of $h\left(\widetilde{\mathbf{X}}_\pi, \mathbf{Y}_\pi\right)$ over functions π in $\Pi(C, A)$. Then U_n is the average of U_A for A in $\mathcal{A}\left(C, n\right)$. Assume that $h\left(\widetilde{\mathbf{X}}_C, \mathbf{Y}_C\right)$ has a finite expectation. For A in $\mathcal{A}\left(C, n\right)$, U_n and U_A have the finite expectation $E\left(U_n\right) = E\left(U_A\right) = E\left(U_C\right) = E\left(h\left(\widetilde{\mathbf{X}}_C, \mathbf{Y}_C\right)\right)$. If $h\left(\widetilde{\mathbf{Z}}_m\right)$ has a finite variance, then U_n has the finite variance

$$\sigma^2\left(U_n\right) = \binom{n}{C}^{-1} \sum_{c=1}^{C} \binom{C}{c} \binom{n-C}{m-C} \sigma^2\left(U_C | \widetilde{\mathbf{X}}_{C-c}, \mathbf{Y}_{C-c}\right).$$

Here, for nonnegative integers j and k,

$$\binom{k}{j} = \begin{cases} \frac{k!}{j!(k-j)!}, & j \leq k, \\ 0, & j > k. \end{cases}$$

For $c < C$, $\sigma^2\left(U_C | \widetilde{\mathbf{X}}_{C-c}, \mathbf{Y}_{C-c}\right)$ is the conditional variance of U_C given $(\mathbf{X}_i, Y_i)$ for $1 \leq i \leq C - c$. For $c = C$, $\sigma^2\left(U_C | \widetilde{\mathbf{X}}_{C-c}, \mathbf{Y}_{C-c}\right)$ is the variance $\sigma^2(U_C)$. The $\sigma^2\left(U_C | \widetilde{\mathbf{X}}_{C-c}, \mathbf{Y}_{C-c}\right)$ are nondecreasing in c, and the variance $\sigma^2(U_n)$ satisfies the inequality

$$\frac{C^2}{n} \sigma^2\left(U_C | \widetilde{\mathbf{X}}_1, \mathbf{Y}_1\right) \leq \sigma^2\left(U_n\right) \leq \frac{m}{n} \sigma^2\left(U_C\right).$$

The variance $\sigma^2(U_n)$ is nonincreasing in n. If $2C \leq n$, then $\sigma^2(U_n)$ has the unbiased estimate

$$s^2\left(U_n\right) = U_n^2 - V\left(U_n\right),$$

where $V(U_n)$ is the average of $U_A U_B$ over all pairs of disjoint sets A and B in $\mathcal{A}\left(C, n\right)$.

One basic use of U-statistics in cross-validation involves use of multiple training sets. Consider the average $U_n = \overline{\text{MSE}}_n\left(Y_0, \widehat{E}_{m0}\right)$ of $\left[r_{mk}^2\left(A\right)\right]^2$ over pairs (A, k) such that A is in $\mathcal{A}\left(m, n\right)$, k is in $\overline{n}$, and k is not in A. Here $\overline{\text{MSE}}_n\left(Y_0, \widehat{E}_{m0}\right)$ is the average of $\overline{\text{MSE}}_n\left(Y_0, \widehat{E}_{m0}; A\right)$ for A in $\mathcal{A}\left(m, n\right)$. If $m = n - 1$, then $U_{m+1} = \overline{\text{MSE}}_n\left(Y_0, \widehat{E}_{m0}\right)$ is the estimate of $\text{MSE}\left(Y_0, \widehat{E}_{m0}\right)$ provided by n-fold cross-validation (Geisser, 1975; Stone, 1974). This case is relatively simple in terms of computation because $\overline{\text{MSE}}_n\left(Y_0, \widehat{E}_{m0}\right)$ is the average of $r_{A(k,n)k}^2$ for $1 \leq k \leq n$ and $A(k, n)$ is the set consisting of all positive integers no greater than n that do not equal k. If r_{m0}^2 has finite variance and $m = n - 1$, then

$$\sigma^2\left(\overline{\text{MSE}}_n\left(Y_0, \widehat{E}_{m0}\right)\right) = n^{-1}\left[\sigma^2\left(r_{m0}^2\right) + (n-1)\,\text{Cov}\left(r_{A(1,n)1}^2, r_{A(2)2}^2\right)\right].$$

The general formula for $\sigma^2\left(U_n\right) = \sigma^2\left(\overline{\text{MSE}}_n\left(Y_0, \widehat{E}_{m0}\right)\right)$ for $n - 1 > m$ applies with $C = m + 1$. This case is usually much more difficult, although occasional exceptions do exist.

Example 8. In Example 6, U_{m+1} is the average $\left(1 + m^{-1}\right) s_{m+1}^2$ of

$$\left\{Y_k - m^{-1}\left[(m+1)\,\overline{Y}_{m+1} - Y_k\right]\right\}^2 = \left(\frac{m+1}{m}\right)^2 \left(Y_k - \overline{Y}_{m+1}\right)^2$$

over positive integers $k \leq m + 1$. It follows that $\overline{\text{MSE}}_n\left(Y_0, \widehat{E}_{m0}\right) = \left(1 + m^{-1}\right) s_n^2$ (Hoeffding, 1948). If $n \geq 3$, then

$$\sigma^2\left(\overline{\text{MSE}}_n\left(Y_0, \widehat{E}_{m0}\right)\right) = \left(\frac{m+1}{m}\right)^2 \left[\frac{2\sigma^4\left(Y\right)}{n-1} + \frac{\kappa_4\left(Y\right)}{n}\right]$$

(Fisher, 1930).

Example 9. In Example 2, the case of $m = n - 1$ leads to the PRESS statistic for the regression of Y_i on $\mathbf{X}_i$ for $1 \le i \le n$ in B (Draper & Smith, 1981, p. 325). In this case, computations are unusually easy (Cook & Weisberg, 1982, p. 33) as long as for no positive integer $k \le n$ does there exist a real number c and p-dimensional vector $\mathbf{d}$ with some nonzero element such that $c + \mathbf{T}_i{}'\mathbf{d} = 0$ for $1 \le i \le n$ and $i \ne k$. This condition holds with probability 1. Let $\overline{\mathbf{X}}_n$ be the sample mean of the $\mathbf{X}_i$ for $1 \le i \le n$, and let $\overline{\mathrm{Cov}}_n(\mathbf{X})$ be the sample covariance matrix of $\mathbf{X}_i$, $1 \le i \le n$. Let the leverage measure

$$h_{ni} = \frac{1}{n} + \frac{1}{n-1}\left(\mathbf{X}_i - \overline{\mathbf{X}}_n\right)'\left[\overline{\mathrm{Cov}}_n(\mathbf{X})\right]^{-1}\left(\mathbf{X}_i - \overline{\mathbf{X}}_n\right) < 1.$$

Then

$$\overline{\mathrm{MSE}}_n\left(Y_0, \hat{E}_{m0}\right) = n^{-1}\sum_{i=1}^{n}\left[r_{ni}/\left(1 - h_{ni}\right)\right]^2$$

is the PRESS estimate of the mean-squared error $\mathrm{MSE}\left(Y_0, \hat{E}_{m0}\right)$.

The preceding examples involve a common but not universal phenomenon. In numerous statistical problems, removal of one observation is quite straightforward. Nonetheless, one difficulty that arises in general is that the variance of $\overline{\mathrm{MSE}}_n\left(Y_0, \hat{E}_{m0}\right)$ does not have an unbiased estimate if $m = n - 1$.

A second use of U-statistics in cross-validation involves construction of a new estimation function $\overline{g}_{nm}$ equal to the average of $g_m(\mathbf{T}_\pi, \mathbf{u}_\pi, \mathbf{t})$ for π in $\Pi\left(m, \overline{n}\right)$ for the $p \times n$ matrix $\mathbf{T}$ with columns $\mathbf{T}_i$ in $\mathcal{X}$ for $1 \le i \le n$, the n-dimensional vector $\mathbf{u}$ with elements u_i in $\mathcal{Y}$ for $1 \le i \le n$, and the p-dimensional vector $\mathbf{t}$ in $\mathcal{X}$. Here $\mathbf{T}_\pi$ is the $p \times m$ matrix with columns $\mathbf{T}_{\pi(i)}$ for $1 \le i \le m$, and $\mathbf{u}_\pi$ is the m-dimensional vector with elements $u_{\pi(i)}$ for $1 \le i \le m$. The estimate $\overline{E}_{nmk}$ is then $\overline{g}_{nm}\left(\widetilde{\mathbf{X}}_n, \mathbf{Y}_n, \mathbf{X}_k\right)$ for $k \ge 0$, and $\overline{r}_{nmk} = Y_k - \overline{E}_{nmk}$. The estimate $\overline{E}_{nmk}$ is also the average of $\hat{E}_{mk}(A)$ for A in $\mathcal{A}(m, n)$, while $\overline{r}_{nmk}$ is the average of $r_{mk}(A)$ for A in $\mathcal{A}(m, n)$.

The mean-squared error

$$\mathrm{MSE}\left(Y_0, \overline{E}_{nm0}\right) = \binom{n}{m}^{-1}\sum_{c=0}^{m}\binom{m}{c}\binom{n-m}{m-c}E\left(\left[E\left(r_{m0}|\widetilde{\mathbf{X}}_{0c}, \mathbf{Y}_{0c}\right)\right]^2\right) \le \mathrm{MSE}\left(Y_0, \hat{E}_{n0}\right),$$

where $\widetilde{\mathbf{X}}_{0c}$ is the $p \times (c+1)$ matrix with columns $\mathbf{X}_i$, $0 \le i \le c$ and $\mathbf{Y}_{0c}$ is the $(c+1)$-dimensional vector with elements Y_i, $0 \le i \le c$. Thus $E\left(r_{m0}|\widetilde{\mathbf{X}}_{0c}, \mathbf{Y}_{0c}\right)$ is the conditional expectation of r_{m0} given $(\mathbf{X}_i, Y_i)$, $0 \le i \le c$. The mean-squared error for prediction of Y_0 by $\overline{E}_{nm0}$ is only equal to the mean-squared error for prediction of Y_0 by $\hat{E}_{m0}$ in the trivial case in which $\hat{E}_{m0}$ is constant with probability 1.

If $2m + 1 \le n$, then $\mathrm{MSE}\left(Y_0, \overline{E}_{nm0}\right)$ has the unbiased estimate $\overline{\mathrm{MSE}}_n\left(Y_0, \overline{E}_{nm0}\right)$ equal to the average of $r_{mk}(A)r_{mk}(B)/N(n, A, B)$ for sets A and B in $\mathcal{A}(m, n)$ and positive integers $k \le n$ in neither A nor B, where $N(n, A, B)$ is the number of positive integers no greater than n that are in neither A nor B. The variance of $\overline{\mathrm{MSE}}_n\left(Y_0, \overline{E}_{nm0}\right)$ can be computed if Y_0 and $\hat{E}_{m0}$ have finite fourth moments, but the formula is somewhat complex in general cases. Unbiased estimation of the variance requires that $4(m+1)$ not exceed n, and computations are far from straightforward.

If $m + 1 = n$, then $\overline{E}_{nm0}$ is associated to some extent with the simplest case of jackknifing (Miller, 1964; Quenouille, 1956). In this case, $\overline{E}_{nm0}$ is the average of $\hat{E}_{m0}(A)$ for the n sets A in $\mathcal{A}(m, n)$. In general, $\overline{E}_{nm0}$ is associated with delete-$(n - m)$ jackknifing. Applications of jackknifing depend on large-sample conditions related to the condition that $n\mathrm{MSE}\left(\hat{E}_{n0}, \overline{E}_{nm0}\right)$ approaches 0 (Shao & Wu, 1989). Nonetheless, it should be emphasized that traditional uses of jackknifing involve evaluation of parameter estimates rather than accuracy of prediction.

Example 10. In Example 1, for any choice of the positive integer $m < n$, $\overline{E}_{nm0} = \hat{E}_{n0}$, so that $\mathrm{MSE}\left(Y_0, \overline{E}_{nm0}\right) = \left(1 + n^{-1}\right)\sigma^2\left(Y_0\right)$ and $\overline{\mathrm{MSE}}_n\left(Y_0, \overline{E}_{nm0}\right) = \left(1 + n^{-1}\right)s_n^2$. Thus $\overline{E}_{nm0}$ recovers the original prediction $\hat{E}_{n0}$. In this example, the variance of $\overline{\mathrm{MSE}}_n\left(Y_0, \overline{E}_{nm0}\right)$ can be obtained from Example 8.

Example 11. In Example 9, $m = n - 1$ and $\hat{E}_{m0} - \hat{E}_{n0}$ is

$$-\frac{r_{nn}}{1 - h_{nn}}\left[\frac{1}{n} + \frac{1}{n-1}\left(\mathbf{X}_0 - \overline{\mathbf{X}}_n\right)'\left[\overline{\mathrm{Cov}}_n(\mathbf{X})\right]^{-1}\left(\mathbf{X}_n - \overline{\mathbf{X}}_n\right)\right].$$

For fixed dimension p, the difference is of order $1/n$. Because $\sum_{i=1}^{n} r_{ni} = 0$ and $\sum_{i=1}^{n} r_{ni}\mathbf{X}_i$ equals the vector $\mathbf{0}_p$ with all elements 0, $\overline{E}_{nm0} - \widehat{E}_{n0}$ is the average of

$$\frac{r_{ni}h_{ni}}{1-h_{ni}}\left[\frac{1}{n} + \frac{1}{n-1}\left(\mathbf{X}_0 - \overline{\mathbf{X}}_n\right)' \left[\overline{\mathrm{Cov}}_n\left(\mathbf{X}\right)\right]^{-1}\left(\mathbf{X}_i - \overline{\mathbf{X}}_n\right)\right]$$

for $1 \le i \le n$ (Cook, 1977). For fixed p, this difference is of order n^{-2}. Results are much more complex and less satisfactory if p increases as the sample size n increases.

Incomplete *U*-Statistics

Because the number of sets in $\mathcal{A}\left(m,n\right)$ can be extremely large for m neither near n nor near 1, computations in many cases can be sufficiently tedious that sampling is needed. This problem can be especially severe if computation of $r_{mk}(A)$ is difficult for even one set A in $\mathcal{A}\left(m,n\right)$. One remedy involves selection of a sample $\mathcal{B}$ of distinct sets in $\mathcal{A}\left(m,n\right)$. The sample is not necessarily random. In this discussion, inferences are conditional on the sample selected. The mean-squared error MSE $\left(Y_0, \widehat{E}_{m0}\right)$ is estimated by the average $\overline{\mathrm{MSE}}_n\left(Y_0, \widehat{E}_{m0}; \mathcal{B}\right)$ of $\overline{\mathrm{MSE}}_n\left(Y_0, \widehat{E}_{m0}; A\right)$ for A in $\mathcal{B}$ (Blom, 1976). If $\mathcal{B} = \mathcal{A}\left(m,n\right)$, then $\overline{\mathrm{MSE}}_n\left(Y_0, \widehat{E}_{m0}; \mathcal{B}\right)$ is $\overline{\mathrm{MSE}}\left(Y_0, \widehat{E}_{m0}; \mathcal{B}\right)$. If $\mathcal{B}$ has a single element A, then $\overline{\mathrm{MSE}}_n\left(Y_0, \widehat{E}_{m0}; \mathcal{B}\right)$ is $\overline{\mathrm{MSE}}_n\left(Y_0, \widehat{E}_{m0}; A\right)$. In all cases, $\overline{\mathrm{MSE}}_n\left(Y_0, \widehat{E}_{m0}; \mathcal{B}\right)$ is an unbiased estimate of MSE $\left(Y_0, \widehat{E}_{m0}\right)$. Determination of $\sigma^2\left(\overline{\mathrm{MSE}}_n\left(Y_0, \widehat{E}_{m0}; \mathcal{B}\right)\right)$ requires that Y_0 and $\widehat{E}_{m0}$ have finite fourth moments. Unbiased estimation of this variance requires that $2(m+1)$ not exceed n. This estimation can entail substantial computational labor.

A good illustration of incomplete U-statistics involves a generalization of K-fold replication (Geisser, 1975; Stone, 1974). Let T be the smallest positive integer such that $Tm\ /\ n$ is a positive integer. For positive integers i and j such that $j > 1$, let $\mathrm{mod}\left(i,\ j\right)$ be the smallest nonnegative integers such that $i - \mathrm{mod}\left(i,\ j\right)$ is an integer multiple of j. For example, $\mathrm{mod}\left(10, 7\right) = 3$. Let $\mathcal{B}$ consist of the T sets B_t, $1 \le t \le T$ of integers $\mathrm{mod}\left((t-1)m + i, n\right)$ such that $1 \le i \le m$. The case of $m + 1 = n$ leads to n-fold cross-validation with $T = n$ and $\mathcal{B} = \mathcal{A}\left(n-1, n\right)$. In traditional K-fold replication, $T = K$ and $n(K-1) = mK$. Estimation of the variance of $\overline{\mathrm{MSE}}_n\left(Y_0, \widehat{E}_{m0}; \mathcal{B}\right)$ is not possible for traditional K-fold replication. On the other hand, this estimation is quite possible if n is a multiple of 5, $m = 2n/5$, and $T = 5$.

A second application of incomplete U-statistics involves the average $\overline{E}_{mk}\left(\mathcal{B}\right)$ of $\widehat{E}_{mk}\left(A\right)$ over A in $\mathcal{B}$ for a nonnegative integer k. Let $N(\mathcal{B})$ be the number of sets in $\mathcal{B}$. The mean-squared error MSE $\left(Y_0, \overline{E}_{m0}\left(\mathcal{B}\right)\right)$ is the average of $E(r_{m0}(A)r_{m0}(B)]]$ for ordered pairs $(A,\ B)$ such that A and B are in $\mathcal{B}$. If A and B have an intersection with c members, then

$$E\left(r_{m0}\left(A\right) r_{m0}\left(B\right)\right) = E\left[E\left(r_{m0}|\widetilde{\mathbf{X}}_{0c}, \mathbf{Y}_{0c}\right)\right]^2.$$

If $2m + 1 \le n$, then MSE $\left(Y_0, \overline{E}_{m0}\left(\mathcal{B}\right)\right)$ has the unbiased estimate $\overline{\mathrm{MSE}}_n\left(Y_0, \overline{E}_{m0}\left(\mathbf{B}\right)\right)$ equal to the average of $r_{mk}(A)r_{mk}(B)/N(n,\ A,\ B)$ for sets A and B in $\mathcal{B}$ and positive integers $k \le n$ in neither A nor B, where $N(n,\ A,\ B)$ is the number of positive integers no greater than n that are in neither A nor B. The variance of $\overline{\mathrm{MSE}}_n\left(Y_0, \overline{E}_{nm0}\right)$ and its associated estimate are usually difficult to compute even if Y_0 and $\widehat{E}_{m0}$ have finite fourth moments.

Example 12. In Example 1, as long as each i from 1 to n appears in the same number of sets in $\mathcal{B}$, $\overline{E}_{nmk}$ is the sample mean $\overline{Y}_n$ of Y_i, $1 \le i \le n$. This result applies in K-fold replication.

Conclusions

A basic challenge in cross-validation is verification that results are reproducible. In the sense that the variance of estimated mean square error can itself be estimated, the answer is affirmative with caveats. In general, the estimation procedure based on m observations needs to employ less than half the total sample of size n, and the estimation procedure examined must be obtained with multiple choices of m observations from the sample of size n. Multiple selections of m observations can be applied to produce an improved estimation procedure based on all n observations, but evaluation of the variance of mean-squared error for the improved estimation procedure is often not feasible.

Consideration of cross-validation requires some perspective. In most routine statistical problems that involve large samples, cross-validation involves rather small corrections. For example, a crude but biased estimate of $\mathrm{MSE}\left(Y_0, \widehat{E}_{n0}\right)$ is the average $\overline{\mathrm{MSE}}_{an}\left(Y_0, \widehat{E}_{n0}\right)$ of r_{nk}^2 over positive integers $k \leq n$. The bias arises because $\widehat{E}_{n0}$ depends on Y_k and $\mathbf{X}_k$ for $1 \leq k \leq n$. The difference between $\mathrm{MSE}\left(Y_0, \widehat{E}_{n0}\right)$ and the expectation $E\left(\overline{\mathrm{MSE}}_{an}\left(Y_0, \widehat{E}_{n0}\right)\right)$ is often of order n^{-1}, and the standard deviation $\sigma\left(\overline{\mathrm{MSE}}_{n}\left(Y_0, \widehat{E}_{n0}\right)\right)$ of $\overline{\mathrm{MSE}}_n(Y_0, \widehat{E}_{n0}$ is often of order $1/n^{1/2}$, so the bias is relatively small.

Examples 1 and 2 illustrate the issue. For the sample mean of Example 1, $\overline{\mathrm{MSE}}_{an}\left(Y_0, \widehat{E}_{n0}\right) = \left(1 - n^{-1}\right) s_n^2$ has expectation $(1 - n^{-1})\sigma^2(Y_0)$, whereas $\mathrm{MSE}\left(Y_0, \widehat{E}_{n0}\right)$ is $(1 + n^{-1})\sigma^2(Y_0)$. If Y_0 has a finite fourth moment, then $\overline{\mathrm{MSE}}_{n}\left(Y_0, \widehat{E}_{n0}\right)$ has the standard deviation $(1 + n^{-1})[2\sigma^4(Y_0)/(n-1) + \kappa_4(Y)/n]^{1/2}$. In Example 2, $\overline{\mathrm{MSE}}_{an}\left(Y_0, \widehat{E}_{n0}\right) = \left[1 - \left(p+1\right)/n\right] s_n^2$ has expectation $[1 - (p+1)/n]\sigma^2(Y_0|\,\mathbf{X}_0)$, whereas

$$\mathrm{MSE}\left(Y_0, \widehat{E}_{n0}\right) = \sigma^2\left(Y_0|\mathbf{X}_0\right)\left(1 + \frac{1}{n} + \frac{p}{n-p-2}\right).$$

For fixed p, the bias is of order n^{-1}, and $n^{1/2}\sigma(\left(\overline{\mathrm{MSE}}_{n}\left(Y_0, \widehat{E}_{n0}\right)\right)$ converges to $2^{1/2}\sigma^2(Y_0|\,\mathbf{X}_0)$. In this example, the situation does change if $p\ /\ n$ is far from 0. Similar issues arise in variance estimation. A crude but biased estimate of $\sigma^2\left(\overline{\mathrm{MSE}}_{n}\left(Y_0, \widehat{E}_{n0}\right)\right)$ is n^{-1} times the sample variance of the r_{ni}^2 for $1 \leq i \leq n$. The bias is of order n^{-2} in typical cases.

A further matter in terms of perspective is that much simpler statistical methods exist for approximation of bias in estimated mean-squared error (Mallows, 1973). These methods apply to Examples 1 and 2. Given that cross-validation involves significant effort and can have some costs in terms of efficiency of estimation, it is important to consider when cross-validation should be used at all. In the end, it appears that applications of cross-validation are most appropriate in samples that are not very large and in cases in which estimates are not well-behaved differentiable functions of the original data. Relevant cases include use of stepwise regression and other methods of model selection, use of inequality constraints, and cases in which the dimension p of $\mathbf{X}$ is large relative to the sample size n. In numerous routine statistical problems, cross-validation has limited value.

References

Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *Annals of Mathematical Statistics*, *18*, 105–110. https://doi.org/10.1214/aoms/1177730497

Blom, G. (1976). Some properties of incomplete *U*-statistics. *Biometrika*, *63*, 573–580. https://doi.org/10.2307/2335738

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*, 15–18. https://doi.org/10.2307/1268249

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York, NY: Chapman and Hall.

Draper, N. R., & Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York, NY: John Wiley.

Fisher, R. A. (1930). Moments and product moments of sampling distributions. *Proceedings of the London Mathematical Society*, *30*(S2), 199–238. https://doi.org/10.1112/plms/s2-30.1.199

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, *70*, 320–328. https://doi.org/10.2307/2285815

Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, *21*, 215–223. https://doi.org/10.2307/1268518

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, *19*, 293–325. https://doi.org/10.1214/aoms/1177730196

Hotelling, H. (1931). The generalization of Student's ratio. *Annals of Mathematical Statistics*, *2*, 360–378. https://doi.org/10.1214/aoms/1177732979

Mallows, C. L. (1973). Some comments on C_P. *Technometrics*, *15*, 661–675. https://doi.org/10.2307/1267380

Miller, R. G. (1964). A trustworthy jackknife. *Annals of Mathematical Statistics*, *35*, 1594–1605. https://doi.org/10.2307/2238296

Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, *79*, 575–583. https://doi.org/10.2307/2288403

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, *43*, 353–360. https://doi.org/10.2307/2332914

Rao, C. R., & Mitra, S. K. (1972). Generalized inverse of a matrix and its applications. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability: Vol. 1. Theory of statistics* (pp. 601–620). Berkeley, CA: University of California Press.

Shao, J., & Wu, C. F. J. (1989). A general theory for jackknife variance estimation. *Annals of Statistics*, *17*, 1176 – 1197. https://doi.org/10.1214/aos/1176347263

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, *36*, 111 – 147. https://doi.org/10.2307/2984809

Suggested citation:

Haberman, S. J. (2019). *Cross-validation and U-statistics* (Research Report No. RR-19-27). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12263

Action Editor: Rebecca Zwick

Reviewers: J. R. Lockwood and Sandip Sinharay