How Biased Is the Apparent Error Rate of a Prediction Rule?

BRADLEY EFRON*

A regression model is fitted to an observed set of data. How accurate is the model for predicting future observations? The apparent error rate tends to underestimate the true error rate because the data have been used twice, both to fit the model and to check its accuracy. We provide simple estimates for the downward bias of the apparent error rate. The theory applies to general exponential family linear models and general measures of prediction error. Special attention is given to the case of logistic regression on binary data, with error rates measured by the proportion of misclassified cases. Several connected ideas are compared: Mallows's C_p, cross-validation, generalized cross-validation, the bootstrap, and Akaike's information criterion.

KEY WORDS: Mallows's C_p; Cross-validation; AIC; Bootstrap methods; Logistic regression; Generalized linear models.

1. INTRODUCTION

Suppose the statistician fits a regression model to an observed set of data. How accurate, or inaccurate, is the model for predicting future observations? An obvious first guess is the *apparent error rate*, which is the observed inaccuracy of the fitted model applied to the original data points. However, the apparent error rate usually underestimates the true error rate. The reason is simple: the model is selected to lie near the observed points, which is what *fitting* means, so these points give a falsely optimistic picture of the model's true accuracy.

This article concerns estimating the bias of the apparent error rate. Here is a simple example of our results. A professional football player had the following field-goal kicking record over the 1969–1972 seasons:

Yards:	55	45	35	25	12	[Total]	
Successes/Attempts:	$\frac{1}{4}$	$\frac{8}{27}$	$\frac{15}{32}$	$\frac{22}{25}$	$\frac{10}{12}$	$[\frac{56}{100}]$	(1.1)

Yards indicates the distance of the kick, discretized into the five categories shown. A standard linear logistic regression,

$$\Pr\{\text{success}\} = 1/[1 + \exp - \{\alpha_0 + \alpha_1 \cdot \text{Yards}\}], \quad (1.2)$$

was fitted to data (1.1) by maximum likelihood [see Efron (1982) for more details].

The fitted logistic regression estimates the probability of a successful kick as a function of the distance, $\hat{\Pr}\{\text{success}\} = 1/[1 + \exp - \{\hat{\alpha}_0 + \hat{\alpha}_1 \cdot \text{Yards}\}]$. We can consider this to be a prediction rule for future kicks, predicting success or failure as $\hat{\Pr}\{\text{success}\}$ is greater or less than .5. This rule has apparent error rate .310; that is, it mispredicts 31 of the 100 original data points (1.1). How optimistic is the value .310?

The theory that follows, in particular formula (2.4), estimates the downward bias of the apparent error rate to be only .012, indicating that bias is not a serious problem in this case. The reason for the small bias is the large ratio of data points to fitted parameters, 100 to 2. A random subset of 20 data points was selected from the 100 shown in (1.1). The logistic regression (1.2) fitted to just these 20 points had apparent error rate .400; that is, it misclassified 8 of the 20 points. Bias estimate (2.4) was now much larger, equaling .066.

Sections 2–5 concentrate on the special but important case of binary data and logistic regression. Error rates are usually measured as in the football example, by counting mispredictions. However, the theory allows more general measures of prediction error, for instance the Deviance (twice the Kullback–Leibler distance). Considering the prediction error of the Deviance leads to a nice corroboration of Akaike's information criterion, as shown in Section 6.

Section 6 extends the theory to linear models for general exponential families, as discussed for instance in McCullagh and Nelder (1983). Theorem 2 of Section 6 states the general result. This includes the most famous special case of all, ordinary linear regression with prediction error measured by squared Euclidean distance, where our bias estimate is equivalent to Mallows's C_p statistic (1973).

This relationship is examined in Section 7. It is easier to see the connection of our results with other methods, such as cross-validation, in the ordinary linear regression setting. Section 7 gives a comparative discussion of several closely related ideas: cross-validation, generalized cross-validation, bootstrap estimates of prediction error, Mallows's C_p, and Akaike's information criterion.

2. LOGISTIC REGRESSION

Logistic regression fits a model of the form

$$\pi_i = \frac{1}{1 + \exp(-t_i'\alpha)}, \qquad i = 1, 2, \ldots, n, \quad (2.1)$$

to an observed vector of binary data $y = (y_1, y_2, \ldots, y_n)$. Here the y_i independently equal 1 or 0 with probabilities π_i or $(1 - \pi_i)$; the t_i are observed p-dimensional covariate vectors; and α is an unknown p-dimensional vector of parameters.

The maximum likelihood estimate (MLE) of α, say $\hat{\alpha}$, gives estimates $\hat{\pi}_i$ by substitution in (2.1). We can think of the $\hat{\pi}_i$ as predicting whether a future observation with covariate vector t_i will be a 1 or a 0. For example, the predictions $\hat{\eta}_i$ might be given by the rule

$$\begin{aligned} \hat{\eta}_i &= 1 \quad \text{if } \hat{\pi}_i > C_0 \\ &= 0 \quad \text{if } \hat{\pi}_i \le C_0, \end{aligned} \quad (2.2)$$

for some cutoff point C_0. The choice $C_0 = .5$ is common.

* Bradley Efron is Professor of Statistics and Biostatistics, Stanford University, Sequoia Hall, Stanford, CA 94305. The author is grateful to Robert Tibshirani for suggesting the close connection of Section 5 to Akaike's information criterion.

Journal of the American Statistical Association
June 1986, Vol. 81, No. 394, Theory and Methods

How accurate is prediction rule (2.2)? The apparent error rate

$$\bar{\text{err}} = \#\{y_i \neq \hat{\eta}_i\}/n \tag{2.3}$$

is the proportion of cases in the original data set y incorrectly predicted by $\hat{\eta}$. However, since y was used to construct $\hat{\eta}$, $\bar{\text{err}}$ will usually be biased downward: a new data vector generated according to (2.1) might not be predicted nearly as accurately by the old $\hat{\eta}$.

We will derive estimates for $\omega(\pi)$, the expected downward bias of the apparent error rate as an estimator of the true error rate. The bias estimate for the logistic regression situation (2.1)–(2.3) is

$$\omega(\hat{\pi}) = \frac{2}{n}\sum_{i=1}^{n} \hat{\pi}_i(1 - \hat{\pi}_i)\phi\left(\frac{\hat{c}_i}{\sqrt{\hat{d}_i}}\right)\sqrt{\hat{d}_i}. \tag{2.4}$$

Here $\phi(z) = (2\pi)^{-1/2}\exp(-\frac{1}{2}z^2)$,

$$\hat{c}_i = \log\left(\frac{C_0}{1 - C_0}\right) - t_i'\hat{\alpha}, \tag{2.5}$$

and

$$\hat{d}_i = t_i'\hat{\Sigma}^{-1}t_i, \qquad \hat{\Sigma} \equiv \sum_{j=1}^{n} \hat{\pi}_j(1 - \hat{\pi}_j)t_jt_j'. \tag{2.6}$$

The matrix $\hat{\Sigma}^{-1}$ is the usual estimate for the covariance matrix of $\hat{\alpha}$, so $\hat{d}_i = \hat{\text{var}}(t_i'\hat{\alpha})$ is a quantity available in the output of most logistic regression programs. Formula (2.4) gave the estimates of bias for the football data quoted in the Introduction.

3. OPTIMISM OF THE APPARENT ERROR RATE

This section considers estimating the *expected optimism* of the apparent error rate, in other words the downward bias of $\bar{\text{err}}$ as an estimate of the true error rate. A simple bias formula is derived applying to general prediction rules and general measures of prediction error. Section 4 specializes this formula to the case of logistic regression and counting error, (2.1)–(2.3), obtaining (2.4). The results here, applying to binary data, are extended to general exponential families in Section 6.

Suppose then that $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ are probabilities, giving the data vector $y = (y_1, y_2, \ldots, y_n)$ by independent binary sampling,

$$\begin{aligned} y_i &= 1 \quad \text{with probability} \quad \pi_i \\ &= 0 \quad \text{with probability} \quad 1 - \pi_i, \end{aligned} \tag{3.1}$$

abbreviated $y \sim B(\pi)$. From y we form a vector of predictions $\hat{\eta} = (\hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_n)$, each $\hat{\eta}_i$ in the range [0, 1]. For now we do not have to specify the rule $y \to \hat{\eta}$.

Given y_i and $\hat{\eta}_i$, we have some measure $Q[y_i, \hat{\eta}_i]$ of prediction error, for instance the "counting error" of Section 1 (where the $\hat{\eta}_i$ equalled 0 or 1)

$$\begin{aligned} Q[y_i, \hat{\eta}_i] &= 1 \quad \text{if} \quad y_i \neq \hat{\eta}_i \\ &= 0 \quad \text{if} \quad y_i = \hat{\eta}_i. \end{aligned} \tag{3.2}$$

The average prediction error for the vector $\hat{\eta}$ is defined to be

$$Q[y, \hat{\eta}] \equiv \frac{1}{n}\sum_{i=1}^{n} Q[y_i, \hat{\eta}_i] \equiv \bar{\text{err}}. \tag{3.3}$$

For counting error, $\bar{\text{err}}$ in (3.3) is the apparent error rate $\bar{\text{err}}$ defined at (2.3).

A wide class of error measures Q can be generated in the following way: let $q(\hat{\eta}_i)$ be a concave function of $\hat{\eta}_i \in [0, 1]$ satisfying $q(0) = q(1) = 0$. (It is convenient but not necessary to have $q(0)$ and $q(1)$ equal 0, as the more general development in Section 6 shows.) Then define the prediction error to be

$$Q[y_i, \hat{\eta}_i] = q(\hat{\eta}_i) + \dot{q}(\hat{\eta}_i)(y_i - \hat{\eta}_i). \tag{3.4}$$

Here y_i equals 0 or 1, and $\dot{q}(\hat{\eta}_i)$ is the derivative of $q(\hat{\eta}_i)$, uniquely defined by left continuity at sharp corners of the concave function q. In other words $Q[y_i, \hat{\eta}_i]$ is the height at y_i of the tangent line to q through the point $(\hat{\eta}_i, q(\hat{\eta}_i))$. See Efron (1978b) for an extensive discussion of such functions.

Three examples of error measures $Q[y_i, \hat{\eta}_i]$ are shown in Table 1. Example 1, counting error, is (3.2) extended in the obvious way for predictions $\hat{\eta}_i$ possibly intermediate between 0 and 1. Notice that Example 2, squared error, agrees with counting error when $\hat{\eta}_i$ equals 0 or 1, but is different for intermediate values $\hat{\eta}_i \in (0, 1)$. In Example 3, the average prediction error $Q[y, \hat{\eta}]$, (3.3), can be expressed as

$$Q[y, \hat{\eta}] = -(2/n)\log f_{\hat{\eta}}(y), \tag{3.5}$$

where $\log f_\pi(y)$ is the log-likelihood of $y \sim B(\pi)$; then $n\,Q[y, \hat{\eta}]$ equals the deviance, twice the Kullback–Leibler distance (see Section 6).

The *true error rate* $\text{Err}(y, \pi)$ of a prediction vector $\hat{\eta}$ is defined to be

$$\text{Err}(y, \pi) \equiv E_{\text{NEW}}\{Q[y^{\text{NEW}}, \hat{\eta}]\}. \tag{3.6}$$

Here y^{NEW} is a hypothetical new data vector, with the same distribution but independent of the original data vector y, which gave $\hat{\eta}$. The notation in (3.6) indicates expectation over $y^{\text{NEW}} \sim B(\pi)$, with $\hat{\eta}$ held fixed. In the case of counting error (3.2), Err is the expected proportion of incorrect predictions $y_i^{\text{NEW}} \neq \hat{\eta}_i$.

The difference between Err and $\bar{\text{err}}$ is the *optimism*

$$\text{op}(y, \pi) \equiv \text{Err} - \bar{\text{err}}. \tag{3.7}$$

The expectation of $\text{op}(y, \pi)$ over $y \sim B(\pi)$,

$$\omega(\pi) \equiv E_\pi\{\text{op}(y, \pi)\}, \tag{3.8}$$

is the expected optimism for the rule $y \to \hat{\eta}$, the quantity we wish to estimate.

Theorem 1. Let $\hat{\zeta} = (\hat{\zeta}_1, \hat{\zeta}_2, \ldots, \hat{\zeta}_n)$ be the vector with ith component

$$\hat{\zeta}_i \equiv -\dot{q}(\hat{\eta}_i). \tag{3.9}$$

Table 1. Three Measures of Prediction Error for Binary Data: The Measure $Q[y_i, \hat{\eta}_i]$ Is Derived From the Concave Function $q(\hat{\eta}_i)$ According to (3.4)

Name	$q(\hat{\eta}_i)$	$Q[y_i, \hat{\eta}_i]$
1. Counting Error	$\min(\hat{\eta}_i, 1 - \hat{\eta}_i)$	1 if $y_i = 1$, $\hat{\eta}_i < \frac{1}{2}$ or if $y_i = 0$, $\hat{\eta}_i > \frac{1}{2}$ 0 otherwise
2. Squared Error	$\hat{\eta}_i(1 - \hat{\eta}_i)$	$(y_i - \hat{\eta}_i)^2$
3. Deviance (twice Kullback–Leibler)	$-2[\hat{\eta}_i\log(\hat{\eta}_i) + (1 - \hat{\eta}_i)\log(1 - \hat{\eta}_i)]$	$-2\log \hat{\eta}_i^{y_i}(1 - \hat{\eta}_i)^{1-y_i}$

Then the expected optimism is

$$\omega(\pi) = \frac{1}{n} E_\pi\left\{\sum_{i=1}^{n} \hat{\zeta}_i \cdot (y_i - \pi_i)\right\}. \tag{3.10}$$

Proof. From definition (3.4),

$$Q[y_i^{\text{NEW}}, \hat{\eta}_i] - Q[y_i, \hat{\eta}_i] = \hat{\zeta}_i \cdot (y_i - y_i^{\text{NEW}}), \tag{3.11}$$

so

$$\text{op}(y, \pi) = \frac{1}{n}\sum_{i=1}^{n} \hat{\zeta}_i \cdot (y_i - \pi_i). \tag{3.12}$$

The theorem follows from definition (3.8).

Remark A. For the three error measures in Table 1, $\hat{\zeta}_i$ equals (a) $\text{sign}(2\hat{\eta}_i - 1)$, (b) $2\hat{\eta}_i - 1$, and (c) $2\log[\hat{\eta}_i/(1 - \hat{\eta}_i)]$, respectively.

Remark B. Another expression for $\omega(\pi)$ is

$$\omega(\pi) = \frac{1}{n}\sum_{i=1}^{n} \text{cov}_\pi(y_i, \hat{\zeta}_i). \tag{3.13}$$

In the case of counting error, where $\hat{\zeta}_i = \text{sign}(2\hat{\eta}_i - 1) = 2\hat{\eta}_i - 1$ for $\hat{\eta}_i = 0$ or 1 as in (2.2),

$$\omega(\pi) = \frac{2}{n}\sum_{i=1}^{n} \text{cov}_\pi(y_i, \hat{\eta}_i). \tag{3.14}$$

Equation (3.14) is a quantitative statement of the fact that the expected bias of the apparent error rate depends on how much each y_i affects its own prediction $\hat{\eta}_i$.

Remark C. Formula (3.10) can also be expressed as

$$\omega(\pi) = \frac{1}{n}\sum_{i=1}^{n} \pi_i(1 - \pi_i)\Delta_i, \tag{3.15}$$

where

$$\Delta_i \equiv E_\pi\{\hat{\zeta}_i \mid y_i = 1\} - E_\pi\{\hat{\zeta}_i \mid y_i = 0\}. \tag{3.16}$$

This is another statement showing how $\omega(\pi)$ depends on the effect of y_i on its own prediction. Expressions (3.10), (3.13), and (3.15) are numerically identical of course, but (3.15) is slightly more convenient for the theoretical calculations of Section 4.

Remark D. The optimism $\text{op}(y, \pi)$, (3.7), refers to the error-rate bias for a given vector $\hat{\eta}$. The expected optimism $\omega(\pi)$, (3.8), is the expected bias for the *rule* $y \to \hat{\eta}$. We would like to estimate $\text{op}(y, \pi)$, but must settle for estimating $\omega(\pi)$, as briefly discussed in Section 5.

Remark E. Section 5 also considers the problem of estimating the true error rate $\text{Err} = \bar{\text{err}} + \text{op}(y, \pi)$. Constructing estimates of Err better than $\bar{\text{err}}$ is the most obvious purpose of estimating the bias $\omega(\pi)$.

Remark F. The notation Err and $\bar{\text{err}}$ is taken from Efron (1983). However, the definition of Err has been changed in a way that makes the problem easier. The difference has to do with working in the framework for Mallows's C_p statistic, rather than that appropriate to cross-validation calculations. This distinction, which relates to questions of conditionality, is discussed in Section 7.

Remark G. Often the observations y_i occur in groups, for instance, the five groups in the football example (2.7). If $\hat{\eta}_g$ is the common prediction in group g, then we might wish to measure prediction error according to a grouped measure $Q[p_g, \hat{\eta}_g]$, where p_g is the observed proportion of y_i's equaling 1 in group g. It turns out that formula (3.10), which ignores grouping, still gives reasonable answers in the grouped case. Starting with squared error at (3.3), for example, formula (3.10) is the expected optimism for the grouped measure $Q[p_g, \hat{\eta}_g] = (p_g - \hat{\eta}_g)^2$, as well as for the ungrouped measure $Q[y_i, \hat{\eta}_i] = (y_i - \hat{\eta}_i)^2$. See Efron (1978b), in particular the last column of Table 3.

Remark H. The proof of Theorem 1 uses only $E_{\text{NEW}}\{y^{\text{NEW}} \mid y, \pi\} = \pi$, and not the full distributional assumptions y, y^{NEW} independently $\sim B(\pi)$. In particular, (3.10) applies to the case where the y_i are correlated binary variables.

4. DERIVATION OF FORMULA (2.4)

This section specializes Theorem 1 to the case of logistic regression and counting error (2.1)–(2.3), leading to estimate (2.4) for the expected optimism $\omega(\pi)$. Except for the end remarks, the discussion in this section is mainly technical.

Suppose then that the binary data vector is distributed as $y \sim B(\pi)$ as in (3.1), that $\pi = (\pi_1, \ldots, \pi_n)$ is given by the logistic formula (2.1), and that the prediction rule is (2.2). For convenience define

$$\chi_i = \pi_i(1 - \pi_i). \tag{4.1}$$

Then (3.15), (3.16) can be expressed as

$$\omega(\pi) = \frac{1}{n}\sum_{i=1}^{n} \chi_i\Delta_i, \tag{4.2}$$

where, since $\hat{\zeta}_i = 2\hat{\eta}_i - 1$ in this case,

$$\Delta_i = 2[E_\pi\{\hat{\eta}_i \mid y_i = 1\} - E_\pi\{\hat{\eta}_i \mid y_i = 0\}]. \tag{4.3}$$

The derivation of (2.4) consists of finding a simple approximation for (4.3).

Under model (2.1), the p-dimensional vector

$$z = \sum_{i=1}^{n} t_i y_i \tag{4.4}$$

is sufficient for α, having an exponential family of density functions

$$f_\alpha(z) = \exp(\alpha' z - \phi(\alpha)),$$

$$\phi(\alpha) = \sum_{i=1}^{n} \log(1 + \exp(t_i'\alpha)). \tag{4.5}$$

The vector α is the natural parameter of this family, whereas the expectation parameter is the vector

$$\beta \equiv E_\alpha\{z\} = \sum_{i=1}^{n} t_i\pi_i. \tag{4.6}$$

The MLE $\hat{\beta}$ of β equals the observed vector z, with covariance matrix

$$\Sigma = \sum_{i=1}^{n} \chi_i t_i t_i'. \tag{4.7}$$

Notice that according to (2.1), (2.2), $\hat{\eta}_i$ equals 1 or 0 as

$t_i'(\hat{\alpha} - \alpha)$ exceeds or is less than

$$c_i = \log(C_0/(1 - C_0)) - t_i'\alpha. \tag{4.8}$$

Then (4.3) can be written as

$$\Delta_i = 2[\Pr_\pi\{t_i'(\hat{\alpha} - \alpha) > c_i \mid y_i = 1\} - \Pr_\pi\{t_i'(\hat{\alpha} - \alpha) > c_i \mid y_i = 0\}]. \tag{4.9}$$

Standard exponential family theory gives the approximation

$$t_i'(\hat{\alpha} - \alpha) \doteq t_i'\dot{\Sigma}^{-1}(\hat{\beta} - \beta) \tag{4.10}$$

(see Efron 1978a, eq. 2.4), so

$$\Delta_i \doteq 2[\Pr_\pi\{t_i'\dot{\Sigma}^{-1}(\hat{\beta} - \beta) > c_i \mid y_i = 1\} - \Pr\{t_i'\dot{\Sigma}^{-1}(\hat{\beta} - \beta) > c_i \mid y_i = 0\}]. \tag{4.11}$$

Now let $\beta_{(i)} \equiv \sum_{j\neq i} t_j\pi_j$ and $\hat{\beta}_{(i)} \equiv \sum_{j\neq i} t_j y_j$, so

$$\hat{\beta} - \beta = (\hat{\beta}_{(i)} - \beta_{(i)}) + t_i(y_i - \pi_i). \tag{4.12}$$

Likewise define $\dot{\Sigma}_{(i)} \equiv \sum_{j\neq i} \chi_j t_j t_j'$. A standard matrix identity gives

$$\dot{\Sigma}^{-1} = \left(I - \frac{\chi_i\dot{\Sigma}_{(i)}^{-1}t_i t_i'}{1 + \chi_i t_i'\dot{\Sigma}_{(i)}^{-1}t_i}\right)\dot{\Sigma}_{(i)}^{-1}. \tag{4.13}$$

Finally, letting $d_{(i)i} = t_i'\dot{\Sigma}_{(i)}^{-1}t_i$, we have

$$d_i = \frac{d_{(i)i}}{1 + \chi_i d_{(i)i}} \quad \text{or} \quad d_{(i)i} = \frac{d_i}{1 - \chi_i d_i}, \tag{4.14}$$

where $d_i = t_i'\dot{\Sigma}^{-1}t_i$.

From (4.12), (4.13) we get

$$t_i'\dot{\Sigma}^{-1}(\hat{\beta} - \beta) = (1 - \chi_i d_i)t_i'\dot{\Sigma}_{(i)}^{-1}(\hat{\beta}_{(i)} - \beta_{(i)}) + d_i(y_i - \pi_i). \tag{4.15}$$

This is a convenient formula for use in (4.11), since it separates out the dependence of $t_i'\dot{\Sigma}^{-1}(\hat{\beta} - \beta)$ on y_i.

Standard asymptotic theory gives a limiting normal distribution for $\hat{\beta}_{(i)} - \beta_{(i)}$ as the matrix $\dot{\Sigma}_{(i)}$ grows large, $\hat{\beta}_{(i)} - \beta_{(i)} \to N_p(0, \dot{\Sigma}_{(i)})$, so

$$t_i'\dot{\Sigma}_{(i)}^{-1}(\hat{\beta}_{(i)} - \beta_{(i)}) \to N(0, d_{(i)i}). \tag{4.16}$$

The mean 0 and variance $d_{(i)i}$ are exact in (4.16), only the normality being asymptotic. Notice that $\hat{\beta}_{(i)}$ is independent of y_i. We can now write (4.15) as

$$t_i'\dot{\Sigma}^{-1}(\hat{\beta} - \beta) = (1 - \chi_i d_i)\sqrt{d_{(i)i}}Z + d_i(y_i - \pi_i) = [(1 - \chi_i d_i)d_i]^{1/2}Z + d_i(y_i - \pi_i), \tag{4.17}$$

where $Z \to N(0, 1)$ is independent of y_i.

Using the last expression of (4.17) in (4.11) gives

$$\Delta_i \doteq 2\left\{\Phi\left(\frac{c_i + d_i\pi_i}{[d_i(1 - \chi_i d_i)]^{1/2}}\right) - \Phi\left(\frac{c_i - d_i(1 - \pi_i)}{[d_i(1 - \chi_i d_i)]^{1/2}}\right)\right\}, \tag{4.18}$$

$\Phi(z) \equiv \int_{-\infty}^{z} \phi(x)dx$. The factor $1 - \chi_i d_i$ is asymptotically negligible, leading to the slightly cruder approximations

$$\Delta_i \doteq 2\left\{\Phi\left(\frac{c_i}{\sqrt{d_i}} + \sqrt{d_i}\pi_i\right) - \Phi\left(\frac{c_i}{\sqrt{d_i}} - \sqrt{d_i}(1 - \pi_i)\right)\right\}, \tag{4.19}$$

$$\doteq 2\phi\left(\frac{c_i}{\sqrt{d_i}}\right)\sqrt{d_i}. \tag{4.20}$$

Going back to (4.2),

$$\omega(\pi) \doteq \frac{2}{n}\sum_{i=1}^{n} \chi_i\phi\left(\frac{c_i}{\sqrt{d_i}}\right)\sqrt{d_i}. \tag{4.21}$$

Formula (2.4) is (4.21) with

$$\hat{\pi}_i = 1/(1 + \exp(-t_i'\hat{\alpha})) \tag{4.22}$$

substituted everywhere for π_i. In other words, $\omega(\hat{\pi})$ is the MLE for $\omega(\pi)$, or at least the MLE for approximation (4.21) to $\omega(\pi)$. Formula (2.4) has the usual asymptotic optimality properties of maximum likelihood estimates. The favorable finite-sample properties of maximum likelihood, approximate median unbiasedness and high efficiency among nearly unbiased estimators, should also hold. Section 5 describes the performance of (2.4) in a sampling experiment.

Remark I. Theorem 1 combined with (4.15) makes it easy to derive bias expressions like (2.4) applying to deviance and squared error, rather than to counting error,

$$\text{squared error:} \quad \omega(\pi) \doteq \frac{2}{n}\sum_{i=1}^{n} \chi_i^2 d_i$$

$$\text{deviance:} \quad \omega(\pi) \doteq \frac{2p}{n}. \tag{4.23}$$

This last formula is an expression of Akaike's information criterion (AIC) (see the corollary in Sec. 6).

Remark J. Bootstrap methods can be used to approximate the bias estimate $\omega(\hat{\pi})$ for any prediction rule $y \to \hat{\eta}$, not necessarily involving logistic regression, and for any error measure $Q[y, \hat{\eta}]$. Parametric bootstrap data vectors are generated according to $y^* \sim B(\hat{\pi})$, giving bootstrap prediction vectors $y^* \to \hat{\eta}^*$; B such bootstrap replications give the estimate

$$\omega(\hat{\pi}) = \frac{1}{B}\sum_{b=1}^{B}\left[\frac{\sum_{i=1}^{n} \hat{\zeta}_i^*(b)\{y_i^*(b) - \hat{\pi}_i\}}{n}\right]. \tag{4.24}$$

As $B \to \infty$, (4.24) approaches $\omega(\hat{\pi})$, the MLE of $\omega(\pi)$.

Remark K. For the football example described in the Introduction, the approximate MLE (2.4) agreed well with actual MLE $\omega(\hat{\pi})$ evaluated by Monte Carlo, (4.24). Table 2 shows the comparison. The difference between (4.24) and (2.4) are small compared with the statistical variability in $\omega(\hat{\pi})$ (coming from the variability of $\hat{\pi}$ as an estimate of π).

The statistical variability is indicated by reevaluating the approximation based on (4.18) at vectors π moderately distant from the MLE $\hat{\pi}$. For example, the symbol $+\ -$ refers to (4.18) evaluated at the vector π obtained by substituting $(\hat{\alpha}_1 + \hat{s}d_1,$

Table 2. Estimates of $\omega(\pi)$ for the Football Data, All $n = 100$ Kicks, and for a Randomly Chosen Subset of $n = 20$ Kicks, as Described in the Introduction: The Last Four Rows Indicate the Statistical Variability in the Estimate $\omega(\hat{\pi})$

	$n = 100$	$n = 20$
Bootstrap (4.24):	.0120 (±.0011, $B = 4{,}000$)	.059 (±.004, $B = 1{,}600$)
Approximation (2.4):	.0119	.066
Approximation (4.18):	.0121	.075
+ +	.0143	.029
+ −	.0155	.108
− +	.0076	.063
− −	.0100	.065

$\hat{\alpha}_2 - \hat{sd}_2)'$ in (2.1), where $\hat{sd}_j$ is the original estimated standard deviation of $\hat{\alpha}_j$.

Large numbers of bootstrap replications B were taken in order to make the comparisons in Table 2 more informative. In fact, $B = 200$ bootstraps gave reasonable estimates of $\omega(\hat{\pi})$ in both cases. An advantage of the bootstrap method is that quantities of interest besides $\omega(\pi)$ can be estimated from the same replications, for example the variability in the prediction vector $\hat{\eta}$.

Remark L. Formula (2.4) makes double use of $\hat{\pi}$; as the vector that defines the predictions $\hat{\eta}$, via (2.2), and as the point in the space of possible π vectors at which $\omega(\pi)$ is evaluated. These two uses can be separated. In some cases the prediction vector $\hat{\eta}$ might not be obtained from $\hat{\pi}$, the MLE of π.

Here is an important example: suppose that in (2.1) α is partitioned into (α_0, α_1), and likewise $t_i' = (t_{0i}', t_{1i}')$; that $\hat{\pi}_i^0$ is the MLE of π_i in the lower-dimensional model where α_1 is assumed to be zero; and that $\hat{\eta}_i$ equals 1 or 0 as $\hat{\pi}_i^0$ is greater or less than C_0. In this case we are using a prediction rule based on a possibly inadequate parametric model.

The bias estimate for this situation turns out to be

$$\omega(\hat{\pi}) = \frac{2}{n}\sum_{i=1}^{n} \hat{\pi}_i(1 - \hat{\pi}_i)\phi\left(\frac{\hat{c}_i}{\sqrt{\hat{D}_i}}\right)\frac{\hat{d}_i^0}{\sqrt{\hat{D}_i}}, \qquad (4.25)$$

where $\hat{\pi}_i$ is obtained from (2.1) by substitution of $\hat{\alpha}$, the MLE of α in the full model; $\hat{c}_i$ is as given in (2.5); $\hat{d}_i^0 = t_{0i}'\hat{\Sigma}^{0-1}t_{0i}$, where $\hat{\Sigma}^0 \equiv \sum_{j=1}^n \hat{\pi}_j^0(1 - \hat{\pi}_j^0)t_{0j}t_{0j}'$; and $\hat{D}_i = t_{0i}'\hat{\Sigma}^{0-1}\hat{\Sigma}\hat{\Sigma}^{0-1}t_{0i}$, where $\hat{\Sigma} \equiv \sum_{j=1}^n \hat{\pi}_i(1 - \hat{\pi}_i)t_{0j}t_{0j}'$.

5. A SAMPLING EXPERIMENT

Table 3 reports the results of a sampling experiment on the performance of estimate (2.4) in a small-sample situation. The data for each trial of the sampling experiment consist of 20 independent vectors (y_i, s_i), where

$$y_i = 1, \quad \text{probability } \tfrac{1}{2}$$
$$= 0, \quad \text{probability } \tfrac{1}{2} \qquad (5.1)$$

and

$$s_i \mid y_i \sim N_2((y_i - \tfrac{1}{2}, 0), I) \qquad (5.2)$$

for $i = 1, 2, \ldots, 20$. Conditioning on s_i, model (2.1) applies to $\pi_i = P\{y_i = 1 \mid s_i\}$, with $t_i' = (1, s_i)$, $\alpha = (0, 1, 0)'$ [see Sec. 1 of Efron (1975)]. The sampling experiment comprised 100 trials, with 20 observations (y_i, s_i) for each trial.

For each trial, prediction rule (2.2) based on the logistic regression maximum likelihood estimates $\hat{\pi}_i$ was calculated from the data $\{(y_i, s_i), i = 1, 2, \ldots, 20\}$, $C_0 = .5$. The first two columns of Table 3 show the true error rate Err, (3.6), and the apparent error rate $\overline{\text{err}}$, (2.3). We see that the expected optimism in this situation is substantial, $\omega(\pi) = .342 - .254 = .088$. Four hundred more trials verified this value to within .001.

Column 4 shows the approximate MLE of the bias $\omega(\hat{\pi})$, (2.4); $\omega(\hat{\pi})$ is nearly unbiased for $\omega(\pi)$, with quite small standard deviation. For comparison, column 6 shows the cross-validation estimate of bias, $\hat{\omega}^{CV} \equiv \widehat{\text{Err}}^{CV} - \overline{\text{err}}$, where

$$\widehat{\text{Err}}^{CV} = \frac{1}{n}\sum_{i=1}^{n} Q[y_i, \hat{\eta}_{(i)i}], \qquad (5.3)$$

$\hat{\eta}_{(i)i}$ indicating the prediction (2.2) for case i based on the 19 observations (y_j, s_j), $j \neq i$. The estimate $\hat{\omega}^{CV}$ is also nearly unbiased, but has standard deviation more than four times larger than that of the MLE $\omega(\hat{\pi})$. This comparison of $\omega(\hat{\pi})$ with $\hat{\omega}^{CV}$ is somewhat unfair, as discussed in Section 7. See also Remark T.

Table 3. First 10 Trials of the Sampling Experiment, and Summary Statistics for 100 Trials

			Maximum Likelihood		Cross-Validation	
Trial	Err	$\overline{\text{err}}$	$\widehat{\text{Err}}$	$\omega(\hat{\pi})$	$\widehat{\text{Err}}^{CV}$	$\hat{\omega}^{CV}$
1	.364	.300	.409	.109	.500	.200
2	.302	.300	.405	.105	.400	.100
3	.378	.250	.324	.074	.300	.050
4	.276	.200	.289	.089	.300	.100
5	.320	.250	.335	.085	.300	.050
6	.369	.200	.284	.084	.250	.050
7	.296	.200	.278	.078	.300	.100
8	.437	.100	.177	.077	.100	.000
9	.336	.350	.450	.100	.450	.100
10	.354	.150	.234	.084	.250	.100
100 trials						
Mean	.342	.254	.346	.093	.349	.096
(Sd)	(.055)	(.094)	(.105)	(.015)	(.117)	(.065)
Coeff. of Variation	.16	.37	.30	.16	.34	.68

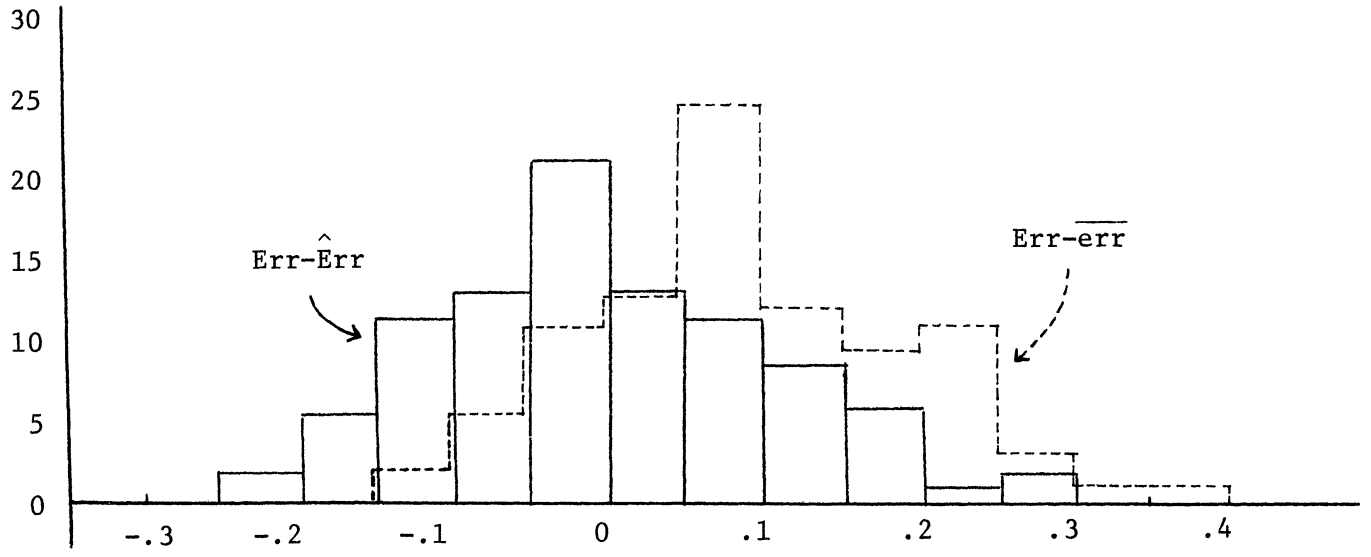

Figure 1. Errors in Estimating Err. Err − Êrr, solid line, compared with Err − ērr, dashed line.

The most obvious use of the bias estimate $\omega(\hat{\pi})$ is to correct ērr as an estimate of Err, say to

$$\hat{\text{Err}} \equiv \bar{\text{err}} + \omega(\hat{\pi}). \tag{5.4}$$

Êrr performs well in Table 3, removing the bias in ērr as an estimate of Err, while decreasing the coefficient of variation of the estimate from .37 to .30. Figure 1 compares the distributions of Err − Êrr and Err − ērr for the 100 trials. Notice that Err − ērr is positive 80% of the time, while Err − Êrr exceeds zero only 54% of the time.

The cross-validation estimate $\hat{\text{Err}}^{\text{CV}}$ is also nearly unbiased for Err, also with smaller coefficient of variation than ērr. Let MSE(Ẽrr) indicate the mean squared error of an estimate Ẽrr for Err,

$$\text{MSE}(\tilde{\text{Err}}) \equiv E[\text{Err} - \tilde{\text{Err}}]^2. \tag{5.5}$$

In the sampling experiment,

$$\text{MSE}(\bar{\text{err}}) = .0174, \quad \text{MSE}(\hat{\text{Err}}) = .0115, \quad \text{MSE}(\hat{\text{Err}}^{\text{CV}}) = .0135. \tag{5.6}$$

These values can be compared with the MSE for the *ideal constant* estimator $\hat{\text{Err}}^{\text{IC}} = \bar{\text{err}} + \omega(\pi) = \bar{\text{err}} + .088$: $\text{MSE}(\hat{\text{Err}}^{\text{IC}}) = .0096$. ($\hat{\text{Err}}^{\text{IC}}$ is the preferred estimate of Err if $\omega(\pi)$ is known, which of course is not so in most real problems.) The relative inefficiency of Ẽrr is defined as in Efron (1983) to be

$$\text{REL}(\tilde{\text{Err}}) \equiv [\text{MSE}(\tilde{\text{Err}}) - \text{MSE}(\hat{\text{Err}}^{\text{IC}})] \div [\text{MSE}(\bar{\text{err}}) - \text{MSE}(\hat{\text{Err}}^{\text{IC}})]. \tag{5.7}$$

For our sampling experiment,

$$\text{REL}(\hat{\text{Err}}) = .24, \quad \text{REL}(\hat{\text{Err}}^{\text{CV}}) = .50. \tag{5.8}$$

To summarize the results of the experiment, $\hat{\text{Err}} = \bar{\text{err}} + \omega(\hat{\pi})$ quite effectively improves ērr as an estimate of Err, and $\omega(\hat{\pi})$ is an excellent estimator of $\omega(\pi)$. Cross-validation is hopelessly inefficient for estimating $\omega(\pi)$, but less bad for estimating Err.

Why would we want to estimate $\omega(\pi)$? In the author's opinion, $\omega(\pi)$ is an interesting measure of how vulnerable a prediction rule is to overfitting. A large value of $\omega(\hat{\pi})$, or perhaps of $\omega(\hat{\pi})/\bar{\text{err}}$, suggests retreating to a more parsimonious prediction rule. However, no quantitative guidelines have been investigated.

Remark M. In the sampling experiment, the correlation between $\omega(\hat{\pi})$ and $\text{op}(y, \pi)$ was $\text{cor}(\omega(\hat{\pi}), \text{op}) = -.84$. This confirms Remark D, that $\omega(\hat{\pi})$ is not estimating the random variable $\text{op}(y, \pi)$, but rather its expectation $\omega(\pi)$.

Remark N. For any estimator $\tilde{\text{Err}} = \bar{\text{err}} + \tilde{\omega}$, the MSE (5.5) is

$$\text{MSE}(\tilde{\text{Err}}) = E[(\bar{\text{err}} + \text{op}) - (\bar{\text{err}} + \tilde{\omega})]^2 = E[\text{op} - \tilde{\omega}]^2. \tag{5.9}$$

In this context $\tilde{\omega}$ is judged by how well it estimates op, no matter what it is supposed to be estimating. In the sampling experiment $\text{cor}(\hat{\omega}^{\text{CV}}, \text{op}) = .03$. This makes $\hat{\omega}^{\text{CV}}$ a relatively less bad estimate of op than of ω, compared with the MLE $\omega(\hat{\pi})$, which has much smaller variance but a substantial negative correlation [see (3.1) of Efron (1983)].

Remark O. In the setting of Efron (1983) it was possible to find a compromise between cross-validation and maximum likelihood that had small variance *and* nonnegative correlation with op. This compromise, the ".632 estimator," was the clear winner in the sampling experiments of the 1983 paper. It is plausible, but so far unverified, that a similar compromise is possible here.

Remark P. Our sampling experiment differs from experiment (2, 20) of Efron (1983) in the choice of prediction rule, logistic regression rather than linear discrimination, and in the definition of Err, as discussed in Section 7. That is why the numbers in Table 3 differ from those in Table 2 of Efron (1983).

6. EXPONENTIAL FAMILIES AND GENERAL LINEAR MODELS

All of our calculations so far have concerned binary data. Similar results hold when the y_i come from a general linear model, as described in McCullagh and Nelder (1983). This section gives a brief discussion of the theory, mostly without proofs.

We suppose that the independent observations y_i are members of a one-parameter exponential family with density functions

$$f_{\mu_i}(y_i) = \exp(\lambda_i y_i - \psi(\lambda_i)), \tag{6.1}$$

where $\mu_i = E\{y_i\}$ is the expectation parameter of the family, λ_i is the natural (or canonical) parameter, and $\psi(\lambda_i)$ is the normalizing function. The two parameters are related by the differential formula $\mu_i = d\psi(\lambda_i)/d\lambda_i$. In the binary case, μ_i equals π_i, λ_i equals the logit $\log\{\pi_i/(1 - \pi_i)\}$, and $\psi(\lambda_i) = \log(1 + e^{\lambda_i})$.

In a general linear model the natural parameters λ_i are expressed as linear combinations of known p-dimensional covariate vectors t_i and an unknown p-dimensional parameter vector α,

$$\lambda_i = t_i'\alpha, \qquad i = 1, 2, \ldots, n. \tag{6.2}$$

Model (2.1) is a special case of (6.2).

Having observed $y = (y_1, y_2, \ldots, y_n)$, we compute the MLE $\hat{\alpha}$, then $\hat{\lambda}_i = t_i'\hat{\alpha}$, and finally

$$\hat{\mu}_i = \frac{d\psi(\lambda_i)}{d\lambda_i}\bigg|_{\hat{\lambda}_i}.$$

Think of $\hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n)$ as a prediction vector for y, with apparent prediction error

$$Q[y, \hat{\mu}] = \frac{1}{n}\sum Q[y_i, \hat{\mu}_i] \equiv \bar{\text{err}}, \tag{6.3}$$

for some error measure $Q[y_i, \hat{\mu}_i]$. (We could let $\hat{\mu}$ further determine a prediction vector $\hat{\eta}$, as in (2.2), but for the discussion here it is sufficient to take $\hat{\eta} = \hat{\mu}$.) Following definitions (3.6)–(3.8), let $\text{Err}(y, \pi) \equiv E_{\text{NEW}}\{Q[y^{\text{NEW}}, \hat{\mu}]\}$ be the true error rate of $\hat{\mu}$, where y^{NEW} is an independent replication of y; and $\omega(\mu) \equiv E_\mu\{\text{Err} - \bar{\text{err}}\}$, the expected optimism of the rule $y \to \hat{\mu}$. We wish to estimate $\omega(\mu)$.

As in (3.4), we consider error measures $Q[y_i, \hat{\mu}_i]$, which are the difference between a concave function $q(y_i)$ and its tangent through a point $(\hat{\mu}_i, q(\hat{\mu}_i))$,

$$Q[y_i, \hat{\mu}_i] = q(\hat{\mu}_i) + \dot{q}(\hat{\mu}_i)(y_i - \hat{\mu}_i) - q(y_i). \tag{6.4}$$

Example. The concave function

$$q(y_i) = 2\{\psi(\lambda(y_i)) - y_i\lambda(y_i)\}, \tag{6.5}$$

where $\lambda(y_i)$ is the value of λ_i corresponding to $\mu_i = y_i$, makes $Q[y_i, \hat{\mu}_i]$ equal the *deviance*, that is, twice the Kullback–Leibler distance $I(y_i, \hat{\mu}_i)$. In this case

$$Q[y, \hat{\mu}] = \frac{1}{n}\sum_{i=1}^{n} Q[y_i, \hat{\mu}_i] = \frac{2}{n}[\log f_y(y) - \log f_{\hat{\mu}}(y)], \tag{6.6}$$

where $f_\mu(y) \equiv \prod_{i=1}^n f_{\mu_i}(y_i)$ [see Efron (1978a)].

A generalized version of (3.10) follows easily from (6.4) and the definition of $\omega(\mu)$:

Theorem 2. Let $\hat{\zeta}$ be the vector with ith component $\hat{\zeta}_i \equiv -\dot{q}(\hat{\mu}_i)$. Then

$$\omega(\mu) = \frac{1}{n}E_\mu\left\{\sum_{i=1}^{n}\hat{\zeta}_i \cdot (y_i - \mu_i)\right\}. \tag{6.7}$$

Theorem 2 applies to any prediction rule $y \to \hat{\mu}$, not necessarily one based on maximum likelihood or general linear models, and to any measure of prediction error of form (6.4).

Corollary. In the special case where $\hat{\mu}$ is the MLE of μ in a general linear model (6.2), and prediction error is based on the deviance (6.6), then

$$\omega(\mu) \doteq 2p/n. \tag{6.8}$$

Proof. Letting $T = (t_1, t_2, \ldots, t_n)$, (6.7) becomes

$$\omega(\mu) = (2/n)E_\mu\{(\hat{\alpha} - \alpha)'T(y - \mu)\}, \tag{6.9}$$

where we have used $\hat{\zeta}_i = 2t_i'\hat{\alpha}$. The exponential family approximation (4.10) then gives

$$\omega(\mu) \doteq (2/n)E_\mu\{(y - \mu)'T'\dot{\Sigma}^{-1}T(y - \mu)\}, \tag{6.10}$$

where

$$\dot{\Sigma} \equiv \sum_{i=1}^{n}\chi_i t_i t_i', \qquad \chi_i \equiv \text{var}_{\mu_i}(y_i). \tag{6.11}$$

However, the last expression in (6.10) equals

$$\frac{2}{n}\sum_{i=1}^{n}\chi_i(t_i'\dot{\Sigma}^{-1}t_i) = \frac{2}{n}\,\text{tr}\,\dot{\Sigma}^{-1}\left\{\sum_{i=1}^{n}t_i t_i'\chi_i\right\} = \frac{2}{n}\,\text{tr}\,\dot{\Sigma}^{-1}\dot{\Sigma} = \frac{2}{n}\,\text{tr}\,I_p = \frac{2p}{n}. \tag{6.12}$$

Remark Q. The corollary extends to the case where $\hat{\mu}$ is the MLE of μ in the p_0-dimensional subfamily of (6.2) obtained as in Remark L: by assuming, possibly incorrectly, that the last $p - p_0$ coordinates of α are zero. Then (6.8) becomes $\omega(\mu) \doteq 2p_0/n$.

Remark R. Akaike's information criterion (AIC) suggests penalizing the maximized log-likelihood $\log f_{\hat{\mu}}(y)$ by p in assessing the goodness of fit of a p-parameter model; that is, a choice between competing models containing different numbers of parameters is made by maximizing $\log f_{\hat{\mu}}(y) - p$ [see Atkinson (1980) and Stone (1977)]. Our corollary supports the AIC. Using formula (6.6) for $\bar{\text{err}}$, the corrected estimate $\hat{\text{Err}} = \bar{\text{err}} + \omega(\mu)$ is

$$\hat{\text{Err}} \doteq \frac{2}{n}\log f_y(y) - \frac{2}{n}\{\log f_{\hat{\mu}}(y) - p\}, \tag{6.13}$$

so AIC amounts to selecting the model with the minimum estimate of Err.

The only approximation in the proof of the corollary comes from (4.10). In the case of normally distributed observations, $y_i \sim N(\mu_i, \sigma^2)$, (4.10) is exact and so is (6.8); $\hat{\text{Err}} = \bar{\text{err}} + \omega(\mu)$ equals $\{\|y - \hat{\mu}\|^2/\sigma^2 + 2p\}/n$, Mallows's C_p statistic (see Sec. 7).

Remark S. The proof of Theorem 2 does not use the full distributional assumptions that y_i^{NEW} and y_i are independent observations from f_{μ_i}, independently for $i = 1, 2, \ldots, n$. All we need is $E_{\text{NEW}}\{y_i^{\text{NEW}} \mid y, \mu\} = \mu_i$ and $E_{\text{NEW}}\{q(y_i^{\text{NEW}})\} = E_\mu\{q(y_i)\}$ for $i = 1, 2, \ldots, n$. This last condition was automatically fulfilled for binary data because there we took $q(y_i) = 0$ for $y_i = 0$ or 1. In particular, (6.6) applies to situations where the y_i are correlated.

7. MALLOWS'S C_p AND CROSS-VALIDATION

In the ordinary least squares (OLS) situation, with normally distributed observations, linear models, and squared error prediction assessment, our theory coincides with Mallows's C_p approach (1973). It is easier to pinpoint the differences between Mallows's approach, which is the framework for our results, and cross-validation methods in the OLS context of this section.

The data vector y is now assumed to have an n-dimensional normal distribution with mean vector μ and covariance matrix $\sigma^2 I$, where μ is known to lie in a p-dimensional linear subspace $\mathcal{L}$,

$$y \sim N_n(\mu, \sigma^2 I), \qquad \mu \in \mathcal{L}. \tag{7.1}$$

Contained in $\mathcal{L}$ is a p_0-dimensional subspace $\mathcal{L}_0$, $p_0 \leq p$. Let $\hat{\mu}$ and $\hat{\mu}_0$ denote the projections of y into $\mathcal{L}$ and $\mathcal{L}_0$, respectively, with corresponding estimates of σ^2,

$$\hat{\sigma}^2 = \|y - \hat{\mu}\|^2/(n - p), \quad \hat{\sigma}_0^2 = \|y - \hat{\mu}_0\|^2/(n - p_0). \tag{7.2}$$

The statistician is interested in the estimator $\hat{\mu}_0$, perhaps for prediction purposes, despite the possibility that $\mu \notin \mathcal{L}_0$. (Normality is not actually needed here; it is enough for the mean vector and covariance matrix to be as described in (7.1).)

Mallows's C_p is an estimate of prediction error for $\hat{\mu}_0$. Using error measure $Q[y, \hat{\mu}_0] = \|y - \hat{\mu}_0\|^2/n$, the true error rate of $\hat{\mu}_0$ is

$$\text{Err} \equiv E_{\text{NEW}} Q[y^{\text{NEW}}, \hat{\mu}_0] = \frac{1}{n}\{\|\mu - \hat{\mu}_0\|^2 + n\sigma^2\}, \tag{7.3}$$

where $y^{\text{NEW}} \sim N_n(\mu, \sigma^2 I)$ is independent of y. The statistic

$$C_p(\mathcal{L}_0, \mathcal{L}) \equiv \frac{1}{n}\{\|y - \hat{\mu}_0\|^2 + 2p_0\hat{\sigma}^2\} \tag{7.4}$$

is an unbiased estimator of Err, in the sense that both have the same expectation under model (7.1),

$$E_{\mu,\sigma^2}\{\text{Err}\} = E_{\mu,\sigma^2}\{C_p(\mathcal{L}_0, \mathcal{L})\}$$
$$= \frac{1}{n}\{\|\mu - \mu_0\|^2 + (n + p_0)\sigma^2\}, \tag{7.5}$$

μ_0 being the projection of μ into $\mathcal{L}_0$. Our statistic $C_p(\mathcal{L}_0, \mathcal{L})$ differs slightly from the usual definition of C_p (see Remark Q).

For the OLS situation, Theorem 2 gives

$$\omega(\mu, \sigma^2) = (2p_0/n)\sigma^2. \tag{7.6}$$

In this case, $\bar{\text{err}} = Q[y, \hat{\mu}_0] = \|y - \hat{\mu}_0\|^2/n$. The obvious estimate of Err, as in (5.4), is

$$\hat{\text{Err}} = \bar{\text{err}} + \omega(\hat{\mu}, \hat{\sigma}^2) = \frac{1}{n}\{\|y - \hat{\mu}_0\|^2 + 2p_0\hat{\sigma}^2\}. \tag{7.7}$$

We see that the approach of the earlier sections results in $C_p(\mathcal{L}_0, \mathcal{L})$ when applied to the situation considered by Mallows.

Table 4 summarizes four different estimates of prediction error for the OLS situation. The naive C_p estimate $C_p(\mathcal{L}_0, \mathcal{L}_0)$ is just $C_p(\mathcal{L}_0, \mathcal{L})$ with $\mathcal{L} = \mathcal{L}_0$. In other words, we use $\mathcal{L}_0$ twice, both to define the prediction vector $\hat{\mu}_0$ and to estimate the true error rate Err for the rule $y \to \hat{\mu}_0$. In this sense $C_p(\mathcal{L}_0, \mathcal{L}_0)$ is similar to formula (2.4), or more exactly to (2.4) plus $\bar{\text{err}}$, while $C_p(\mathcal{L}_0, \mathcal{L})$ is similar to (4.25) plus $\bar{\text{err}}$ (see Remark L).

It is interesting that

$$E_{\mu,\sigma^2}\{C_p(\mathcal{L}_0, \mathcal{L}_0)\} \geq E_{\mu,\sigma^2}\{\text{Err}\}, \tag{7.8}$$

with equality if and only if $\mu \in \mathcal{L}_0$. The naive estimator $C_p(\mathcal{L}_0, \mathcal{L}_0)$ tends to overestimate Err when the assumption $\mu \in \mathcal{L}_0$ is false. The *generalized cross-validation* estimate

$$\text{GCV}(\mathcal{L}_0) \equiv \frac{1}{n}\left\{\frac{\|y - \hat{\mu}_0\|^2}{(1 - p_0/n)^2}\right\} = \frac{1}{1 - (p_0/n)^2} C_p(\mathcal{L}_0, \mathcal{L}_0) \tag{7.9}$$

introduced by Craven and Wahba (1979), overestimates Err slightly more.

Generalized cross-validation is a rotationally invariant form of the *cross-validation* estimate

$$\text{CV}(\mathcal{L}_0) \equiv \frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{\mu}_{0i}^{(i)}\}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{\mu}_{0i}}{1 - P_{ii}^0}\right)^2, \tag{7.10}$$

as motivated in Golub, Heath, and Wahba (1979). Here $\hat{\mu}_{0i}^{(i)}$ is the prediction for y_i calculated from the reduced data set, which omits y_i and its corresponding covariate vector, described more carefully below; and P_{ii}^0 is the iith diagonal element of the projection matrix p^0 into $\mathcal{L}_0$. The term "hat matrix" is often given to P_0.

The average value of P_{ii}^0 is

$$1/n \text{ tr } P^0 = p_0/n, \tag{7.11}$$

so $\text{GCV}(\mathcal{L}_0)$ is just $\text{CV}(\mathcal{L}_0)$ with the denominator $1 - P_{ii}^0$ replaced by its average value $1 - p_0/n$. This results in $E\{\text{CV}(\mathcal{L}_0)\}$ usually exceeding $E\{\text{GCV}(\mathcal{L}_0)\}$, because of Jenson's inequality. The inequality $E\{\text{CV}\} \geq E\{\text{GCV}\}$ is always true when $\mu \in \mathcal{L}_0$, and true in an average sense, averaging over spheres of constant $\|\mu - \mu_0\|$ value, when $\mu \notin \mathcal{L}_0$.

Table 4. Four Different Estimates of Prediction Error for the Ordinary Least Squares Situation

Name	Notation	Formula	Expectation (7.1)
1. C_p	$C_p(\mathcal{L}_0, \mathcal{L})$	$\frac{1}{n}\{\|y - \hat{\mu}_0\|^2 + 2p_0\hat{\sigma}^2\}$	$E\{\text{Err}\} = \frac{1}{n}\{\|\mu - \mu_0\|^2 + (n + p_0)\sigma^2\}$
2. Naive C_p	$C_p(\mathcal{L}_0, \mathcal{L}_0)$	$\frac{1}{n}\{\|y - \hat{\mu}_0\|^2 + 2p_0\hat{\sigma}_0^2\} = \frac{n + p_0}{n}\hat{\sigma}_0^2$	$E\{\text{Err}\} + \frac{2p_0}{n - p_0}\frac{\|\mu - \mu_0\|^2}{n}$
3. Generalized cross-validation	$\text{GCV}(\mathcal{L}_0)$	$\frac{1}{n}\left\{\frac{\|y - \hat{\mu}_0\|^2}{(1 - p_0/n)^2}\right\} = \frac{n}{n - p_0}\hat{\sigma}_0^2$	$\frac{E\{C_p(\mathcal{L}_0, \mathcal{L}_0)\}}{1 - p_0^2/n^2}$
4. Cross-validation	$\text{CV}(\mathcal{L}_0)$	$\frac{1}{n}\sum(y_i - \hat{\mu}_{0i}^{(i)})^2 = \frac{1}{n}\sum\left(\frac{y_i - \hat{\mu}_{0i}}{1 - P_{ii}^0}\right)^2$	$\doteq E\{\text{Err}_+\}$ (6.13)

Table 5. First 10 Trials of a Sampling Experiment Comparing 6 Different Estimators of Err and Err_+ in an OLS Situation, and Summary Statistics for 20 Trials: The Bootstrap Estimate of Err_+ is Defined in Section 2 of Efron (1983); $C_p(\mathcal{L}_0, \mathcal{L}_5)$ is (7.4) With $\mathcal{L}_5$ the Space of Fifth Degree Polynomials in x

Trial	Err (7.3)	$\overline{err}$ (7.7)	$C_p(\mathcal{L}_0, \mathcal{L}_0) \doteq GCV$ (7.9)	$C_p(\mathcal{L}_0, \mathcal{L})$ (7.4)	$C_p(\mathcal{L}_0, \mathcal{L}_5)$ (Quintic)	CV (7.10)	$\hat{Err}_+^{(BOOT)}$ (B = 400)	Err_+ (7.14)
1	2.66	2.41	2.95	2.54	2.49	3.36	3.19	3.42
2	1.93	1.98	2.43	2.18	2.13	3.84	2.97	3.21
3	3.70	3.27	3.99	3.50	3.52	4.73	4.48	2.79
4	2.37	2.73	3.34	2.91	2.91	4.09	3.66	3.42
5	1.73	2.01	2.45	2.20	2.24	2.84	2.67	3.11
6	2.46	3.03	3.71	3.36	3.26	4.80	4.22	2.86
7	2.35	1.65	2.01	1.79	1.83	2.31	2.17	3.46
8	2.02	1.86	2.27	1.99	2.00	3.01	2.54	2.72
9	3.39	2.21	2.70	2.46	2.49	3.67	3.16	4.91
10	2.81	2.85	3.48	3.02	2.95	4.13	3.84	3.72
20 Trials AVE:	2.32	2.21	2.71	2.42	2.40	3.26	2.97	3.39
20 Trials (SD):	(.64)	(.67)	(.81)	(.69)	(.67)	(1.15)	(.98)	(.54)

It may seem strange that $CV(\mathcal{L}_0)$ is biased upward for $E\{Err\}$, given how plausible $CV(\mathcal{L}_0) = (1/n)\sum_{i=1}^n \{y_i - \hat{\mu}_{0i}^{(i)}\}^2$ looks as an estimator of Err. In fact $CV(\mathcal{L}_0)$ is estimating a somewhat different quantity, which we now describe.

Suppose as in (6.2) that $\mu_i = t_i'\alpha$ for $i = 1, 2, \ldots, n$. The $p \times n$ matrix $T = (t_1, t_2, \ldots, t_n)$ must have row space $\mathcal{L}_{row}(T) = \mathcal{L}$, in accordance with (7.1). Suppose also that we can partition t_i and α into $t_i' = (t_{0i}', t_{1i}')$ and $\alpha' = (\alpha_0', \alpha_1')$ as in Remark L, where t_{0i} and α_0 are of dimension p_0, and that the $p_0 \times n$ matrix $T_0 = (t_{01}, t_{02}, \ldots, t_{0n})$ has $\mathcal{L}_{row}(T_0) = \mathcal{L}_0$. Then the projection $\hat{\mu}_0$ of y into $\mathcal{L}_0$ is given by

$$\hat{\mu}_0 = P^0 y, \qquad P^0 = T_0'(T_0T_0')^{-1}T_0. \tag{7.12}$$

Equivalently we can describe the prediction rule $\hat{\mu}_0$ by

$$\hat{\mu}_{0i} = t_{0i}'\hat{\alpha}_0, \qquad \hat{\alpha}_0 = (T_0T_0')^{-1}T_0 y. \tag{7.13}$$

The appropriate context for cross-validation is that where the pairs (t_i, y_i), $i = 1, 2, \ldots, n$, are independently selected according to some joint probability distribution F on $(p + 1)$-dimensional space. Now suppose that one more independent pair is obtained from F, say (t_+, y_+). The predicted value for y_+ based on the original rule (7.13) is $\hat{\mu}_0(t_+) = t_{0+}'\hat{\alpha}_0$. The expected squared error of prediction is

$$Err_+ \equiv E_+\{y_{0+} - t_{0+}'\hat{\alpha}_0\}^2, \tag{7.14}$$

E_+ indicating expectation over (t_+, y_+), with $\hat{\alpha}_0$ fixed.

The expected value of $CV(\mathcal{L}_0)$ tends toward $E_{\mu,\sigma^2}\{Err_+\}$ rather than $E_{\mu,\sigma^2}\{Err\}$. The predictor $\hat{\mu}_{0i}^{(i)}$ appearing in (7.10) equals $t_{0i}'\hat{\alpha}_0^{(i)}$, where $\hat{\alpha}_0^{(i)} = (T_{0(i)}T_{0(i)}')^{-1}T_{0(i)}y_{(i)}$, $T_{0(i)} = (t_{01}, \ldots, t_{0,i-1}, t_{0,i+1}, \ldots, t_{0,n})$, and $y_{(i)} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$. We see that $E_{\mu,\sigma^2}\{CV(\mathcal{L}_0)\}$ equals the expected value of Err_+, for sample size $n - 1$ rather than n.

The results of a small sampling experiment are reported in Tables 5 and 6. The data for each trial of the experiment comprised 20 pairs (x_i, y_i), generated as follows:

$$x_i \sim N(0, 10^2), \quad \mu_i = x_i + .01x_i^2, \quad y_i \sim N(\mu_i, 1), \quad i = 1, 2, \ldots, 20. \tag{7.15}$$

The spaces $\mathcal{L}$ and $\mathcal{L}_0$ were taken to be those associated with quadratic and linear regression, respectively. In other words, $\hat{\mu}_{0i} = \hat{\alpha}_{00} + \hat{\alpha}_{01}x_{i1}$, the simple linear regression of y_i on x_i based on the data (x_i, y_i), $i = 1, \ldots, 20$; whereas $\hat{\sigma}^2$ in (7.2) was based on a quadratic regression for $\hat{\mu}$, dimension $p = 3$.

Twenty trials of (7.15) were run. Notice in Table 5 that Err_+ is usually larger than Err. This is no surprise; Err_+ is the prediction error for a completely new pair (x_+, y_+), whereas Err is the average prediction error for a new pair having x^{NEW} equal to one of the 20 original x_i values. It is easier to predict in the latter case because the (x, y) pairs are nearer the training set $\{(x_i, y_i), i = 1, \ldots, 20\}$. See Section 6 of Efron (1983).

Table 6 shows how well the six estimators performed in the sampling experiment. Two mean squared errors are shown, MSE (5.5) and MSE_+, which is (5.5) with Err_+ replacing Err.

Several facts are worth mentioning: it is much easier to estimate Err than Err_+; cross-validation is a better estimator of Err_+ than Err, though not wonderful in either case; the bootstrap estimation for Err_+ described in Efron (1983) does somewhat better in both cases; $C_p(\mathcal{L}_0, \mathcal{L})$ does very well in estimating Err, considering it is an unbiased estimator; $\overline{err}$ does even better in this case, but the MSE criterion favors estimators that are biased downwards; $C_p(\mathcal{L}_0, \mathcal{L}_5)$, based on an overly large choice of $\mathcal{L}$ in (7.4), performs just as well as $C_p(\mathcal{L}_0, \mathcal{L})$ using the correct choice of $\mathcal{L}$.

Table 6. How Well the Six Estimators in Table 5 Estimated Err and Err_+ in the 20 Trials of the Sampling Experiment: MSE is Defined at (5.5); MSE_+ Is the Corresponding Mean Squared Error for Err_+; It Is Much Easier to Estimate Err

	$\overline{err}$ (7.7)	$C_p(\mathcal{L}_0, \mathcal{L}_0) \doteq GCV$ (7.9)	$C_p(\mathcal{L}_0, \mathcal{L})$ (7.4)	$C_p(\mathcal{L}_0, \mathcal{L}_5)$ (Quintic)	CV (7.10)	$\hat{Err}_+^{(BOOT)}$ (B = 400)
MSE:	.31	.54	.35	.33	1.60	.92
MSE_+:	2.09	1.38	1.71	1.70	1.47	1.33

The main point is that it is easier to estimate Err than Err_+, and that $C_p(\hat{\mathcal{L}}_0, \mathcal{L})$ is the estimator of choice for Err. The logistic regression experiment in Section 5 reached a similar conclusion, with $C_p(\hat{\mathcal{L}}_0, \mathcal{L})$ replaced by its binary data analogue (5.4).

Remark T. The difference $\text{Err}_+ - \text{Err}$ is only about .005 in the experiment of Section 5, too small to be apparent in Table 3.

Remark U. Unlike $C_p(\hat{\mathcal{L}}_0, \mathcal{L})$, neither cross-validation nor $\hat{\text{Err}}_+^{(\text{BOOT})}$ require the statistician to name a space $\mathcal{L}$ guaranteed to contain the mean vector μ. However, they estimate a different quantity from $C_p(\hat{\mathcal{L}}_0, \mathcal{L})$, Err_+ rather than Err, and with less efficiency.

Remark V. Which quantity is more relevant, Err or Err_+? Arguments can be made both ways, depending on the context, but for comparing different possible models $\hat{\mathcal{L}}_0$, efficiency of the error estimation is the primary consideration. This offers some pragmatic ground for preferring C_p to cross-validation, though the evidence so far is by no means overwhelming.

Remark W. Despite its name, $\text{GCV}(\hat{\mathcal{L}}_0)$ is (nearly) a member of the C_p family of estimates.

Remark X. The cross-validation estimate $\text{CV}(\hat{\mathcal{L}}_0)$ depends on the coordinate system in which $\hat{\mathcal{L}}_0$ and y are expressed. We can get an invariant version of $\text{CV}(\hat{\mathcal{L}}_0)$ by averaging (7.10) over a uniform choice among all possible orthogonal coordinate systems. The invariant version of $\text{CV}(\hat{\mathcal{L}}_0)$ turns out to equal

$$((n - p_0)/(n - p_0 - 2))\text{GCV}(\hat{\mathcal{L}}_0). \qquad (7.16)$$

This calculation is close to the one in Golub, Heath, and Wahba (1979), which gives exactly $\text{GCV}(\hat{\mathcal{L}}_0)$, except that they average over a group that includes complex-valued rotations.

Remark Y. The bootstrap estimate $\hat{\text{Err}}_+^{(\text{BOOT})}$ in Tables 5 and 6 is not the analogue of the bootstrap method for binary data described in Remark J. The bootstrap argument of Remark J, applied to the OLS situation of this section, gives exactly $C_p(\hat{\mathcal{L}}_0, \mathcal{L})$.

[*Received April 1985. Revised July 1985.*]

REFERENCES

Atkinson, A. C. (1980), "A Note on the Generalized Information Criterion for Choice of a Model," *Biometrika*, 67, 413–418.

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data With Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Numerische Mathematik*, 31, 377–403.

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

——— (1978a), "The Geometry of Exponential Families," *Annals of Statistics*, 6, 362–376.

——— (1978b), "Regression and ANOVA With Zero–One Data: Measures of Residual Variation," *Journal of the American Statistical Association*, 73, 113–121.

——— (1982), "Maximum Likelihood and Decision Theory," *Annals of Statistics*, 10, 340–356.

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

Golub, G., Heath, M., and Wahba, G. (1979), "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter," *Technometrics*, 21, 215–223.

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

McCullagh, P., and Nelder, J. (1983), *Generalized Linear Models*, New York: Chapman & Hall.

Stone, M. (1977), "An Asymptotic Equivalence of Choice of Model by Cross-Validation of Akaike's Criterion," *Journal of the Royal Statistical Society*, Ser. B, 39, 44–47.