STATISTICS IN MEDICINE, VOL. 9, 1303–1325 (1990)

PREDICTIVE VALUE OF STATISTICAL MODELS

J. C. VAN HOUWELINGEN AND S. LE CESSIE

Department of Medical Statistics, Leiden University, P.O. Box 9512, 2300 RA Leiden, The Netherlands

SUMMARY

A review is given of different ways of estimating the error rate of a prediction rule based on a statistical model. A distinction is drawn between apparent, optimum and actual error rates. Moreover it is shown how cross-validation can be used to obtain an adjusted predictor with smaller error rate. A detailed discussion is given for ordinary least squares, logistic regression and Cox regression in survival analysis. Finally, the split-sample approach is discussed and demonstrated on two data sets.

1. INTRODUCTION

In clinical biostatistics there is a slow but steady development in the direction of more complicated statistical modelling. Not all medical researchers are satisfied with mere hypothesis testing and *P*-values; some want to quantify the effect of explanatory variables on the outcome variable in which they are interested, and there is a growing demand for multivariate analysis. Apart from giving insight into the role of explanatory variables, a multivariate analysis can also be used to predict the outcome of new observations (see Van Houwelingen *et al.*[1] for an example).

When using a statistical model as a tool for prediction, questions arise about generalization of the model to new observations from the same population, and to other populations. Generally speaking, the validity of the prediction model has to be assessed, and this gives rise to the problems of how to quantify the error rate of a prediction, how to estimate the error rate and how to adjust the prediction rule (if necessary).

In this paper, all these aspects will be discussed in settings ranging from simple normal distributions to the much more complicated proportional hazards model.

2. A SIMPLE EXAMPLE

We start with a sequence of independent random variables, $Y_1, \ldots, Y_n, Y_{n+1}, \ldots$ with common mean μ and variance σ^2. They may, for example, all be normally distributed, $N(\mu, \sigma^2)$, but normality is not essential here.

We suppose that the first n variables $Y_1, \ldots, Y_n$ are observed, and we want to predict future observations, starting with Y_{n+1}, by a predictor denoted by $\hat{Y}$. A generally accepted and mathematically very convenient error rate for the predictor is the mean squared error,

$$MSE = E(Y_{\text{new}} - \hat{Y})^2 = (\mu - \hat{Y})^2 + \sigma^2. \tag{1}$$

0277–6715/90/111303–23$11.50

If the common mean μ is known, the best predictor would be $\hat{Y} = \mu$ and we would obtain the optimum error rate

$$MSE_{\text{OPT}} = \sigma^2. \tag{2}$$

Generally, the *optimum error rate* is the error rate that can be obtained if the parameters of the statistical model are known and the optimal predictor is used.

A natural predictor of future observations is

$$\hat{Y} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{3}$$

If we use this predictor, the actual error rate is given by

$$MSE_{\text{ACT}} = E(Y_{\text{new}} - \hat{Y})^2 = (\mu - \bar{Y})^2 + \sigma^2. \tag{4}$$

Generally, the *actual error rate* of a prediction rule is the error rate obtained by averaging over the distribution of future observations.

Notice that the actual error is a random variable, and in this example it depends on the value of $\bar{Y}$. If we are lucky $\bar{Y}$ is close to μ and our prediction is good, but if we are unlucky $\bar{Y}$ may be far from μ and our prediction quite bad. We note that $MSE_{\text{ACT}} \geqslant MSE_{\text{OPT}}$ and that MSE_{ACT} approaches MSE_{OPT} as n increases. These observations apply also to the general definitions above.

Unfortunately, MSE_{ACT} depends on the unknown values of μ and σ^2. So MSE_{ACT} is unknown and we are confronted with the second problem mentioned in the introduction, the estimation of MSE_{ACT}. A simple intuitive way of estimating MSE_{ACT} is to apply the predictor to the observed values $Y_1, \ldots, Y_n$ retrospectively, and to measure the average value of the error rate. This leads to the apparent error rate

$$MSE_{\text{APP}} = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y})^2 = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \frac{n-1}{n} s^2. \tag{5}$$

Generally, the *apparent error rate* is the average error rate when the predictor is applied to the available observations retrospectively.

For a discussion of the concepts of optimum, actual and apparent error rate see the report by the Panel on Discriminant Analysis.[2]

Since the data are used twice, the apparent error rate tends to underestimate the actual error rate and to be too optimistic about the predictive power of the prediction rule. Following Efron[3,4] we define the *optimism* O by

$$O = MSE_{\text{ACT}} - MSE_{\text{APP}}. \tag{6}$$

If we could estimate O, then we could use it to adjust the apparent error rate to obtain an estimate of the actual error rate MSE_{ACT}. It is not easy to estimate O itself. Therefore we consider the expected value of O averaged over the distribution of $Y_1, \ldots, Y_n$, that is

$$\begin{aligned} \Omega = E(O) &= E(MSE_{\text{ACT}}) - E(MSE_{\text{APP}}) \\ &= \left(1 + \frac{1}{n}\right)\sigma^2 - \left(\frac{n-1}{n}\right)\sigma^2 = \frac{2}{n}\sigma^2. \end{aligned} \tag{7}$$

A natural estimate of Ω is given by

$$\hat{\Omega} = \frac{2}{n} s^2, \tag{8}$$

and this leads to the following estimate of MSE_{ACT}:

$$\widehat{MSE}_{ACT} = MSE_{APP} + \hat{\Omega} = \left(1 + \frac{1}{n}\right)s^2. \tag{9}$$

This method of correcting the apparent error rate is essentially Mallows's C_p method.[5] It should be emphasized that $\widehat{MSE}_{ACT}$ is only an (unbiased) estimator of MSE_{ACT}, and can be quite inaccurate when n is small. Observe that the variances of $\widehat{MSE}_{ACT}$ and MSE_{ACT} are of order $1/n$ and $1/n^2$ respectively, showing that the estimated error rate varies much more than the actual error rate.

Quite a different method of estimating the actual error rate is by means of cross-validation.[6] Instead of estimating the expected optimism Ω, the error rate is estimated by mimicking the prediction situation. Generally, the *cross-validation error rate* of a prediction rule is obtained by averaging the error made by 'predicting' a single observation by means of the predictor based on the other observations. In this situation, the prediction of Y_i using the other observations is

$$\bar{Y}_{(-i)} = \frac{1}{n-1}\sum_{j \neq i} Y_j = \bar{Y} - \frac{1}{n-1}(Y_i - \bar{Y}). \tag{10}$$

This leads to

$$MSE_{CV} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y}_{(-i)})^2 = \frac{n}{n-1}s^2. \tag{11}$$

Essentially this is Allen's PRESS.[7] Observe that MSE_{CV} differs very little from $\widehat{MSE}_{ACT}$, the difference being of order $1/n^2$, so that cross-validation also yields an estimate of MSE_{ACT} that can be quite inaccurate.

3. A LESS SIMPLE EXAMPLE

We now let $Y_1, \ldots, Y_n, Y_{n+1}, \ldots$ be independent binary (0/1) variables with success probability π. We let p be an estimate of π. The problem that makes this example more complicated than that of the previous section is how to define the error rate of p in predicting future observations. We will discuss three measures of prediction error that can also be found in Efron.[4]

Classification error

First, let us predict $\hat{Y} = 1$ if $p > \frac{1}{2}$ and $\hat{Y} = 0$ if $p < \frac{1}{2}$, and randomize between $\hat{Y} = 1$ and $\hat{Y} = 0$ if $p = \frac{1}{2}$. The expected misclassification error of this rule is given by

$$CER = \begin{cases} \pi & \text{if } p < \frac{1}{2} \\ \frac{1}{2} & \text{if } p = \frac{1}{2} \\ 1 - \pi & \text{if } p > \frac{1}{2}. \end{cases} \tag{12}$$

This error rate is the natural one in discriminant analysis.[8] (It is equal to $E|\hat{Y} - Y_{new}|$.) Obviously

$$CER_{OPT} = \min(\pi, 1 - \pi). \tag{13}$$

Let $\hat{\pi} = \Sigma\, Y_i/n$ be the natural estimate of π. Its actual error rate CER_{ACT} is given by (12) with p replaced by $\hat{\pi}$. The apparent error rate equals

$$CER_{APP} = \min(\hat{\pi}, 1 - \hat{\pi}). \tag{14}$$

Hence the optimism of the apparent error rate is given by

$$O = \begin{cases} \pi - \hat{\pi} & \text{if} \quad \hat{\pi} < \frac{1}{2} \\ \hat{\pi} - \pi & \text{if} \quad \hat{\pi} > \frac{1}{2}. \end{cases}$$

Using a normal approximation for the distribution of $\hat{\pi}$, we find that approximately

$$\Omega = E_\pi O = 2c_n \varphi\left(\frac{0{\cdot}5 - \pi}{c_n}\right), \tag{15}$$

where $c_n = \sqrt{[\pi(1-\pi)/n]}$, the standard deviation of $\hat{\pi}$, and φ is the standard normal density. An estimate of Ω is given by

$$\hat{\Omega} = 2\hat{c}_n \varphi\left(\frac{0{\cdot}5 - \hat{\pi}}{\hat{c}_n}\right) \quad \text{with } \hat{c}_n = \sqrt{[\hat{\pi}(1-\hat{\pi})/n]}. \tag{16}$$

This is in line with Efron[4] formula (2.4).

The cross-validation approach for this situation is equivalent to the leaving one out (LOO) method.[8] If $Y_i = 1$, the estimate of π when leaving out that observation is given by

$$\hat{\pi}_1 = \frac{n\hat{\pi} - 1}{n - 1}. \tag{17}$$

Similarly, if $Y_i = 0$, π is estimated by

$$\hat{\pi}_0 = \frac{n\hat{\pi}}{n - 1}. \tag{18}$$

The cross-validation error rate is given by

$$CER_{CV} = \hat{\pi}\{[\hat{\pi}_1 < \tfrac{1}{2}] + \tfrac{1}{2}[\hat{\pi}_1 = \tfrac{1}{2}]\} + (1 - \hat{\pi})\{[\hat{\pi}_0 > \tfrac{1}{2}] + \tfrac{1}{2}[\hat{\pi}_0 = \tfrac{1}{2}]\}, \tag{19}$$

where $[\cdot]$ represents the indicator function, so that for example $[\hat{\pi}_1 < \frac{1}{2}]$ is equal to one if $\hat{\pi}_1 < \frac{1}{2}$ and zero otherwise.

It can be shown that

$$CER_{CV} = \begin{cases} \min(\hat{\pi}, 1 - \hat{\pi}) = CER_{APP} & \text{if} \quad |\hat{\pi} - \tfrac{1}{2}| > 1/2n \\ 1 & \text{if} \quad n \text{ is even and } \hat{\pi} = \tfrac{1}{2} \\ \tfrac{1}{2} + \tfrac{1}{2}\min(\hat{\pi}, 1 - \hat{\pi}) & \text{if} \quad n \text{ is odd and } \hat{\pi} = \tfrac{1}{2} \pm 1/2n. \end{cases} \tag{20}$$

This seems a rather peculiar estimate of CER_{ACT}; it corrects CER_{APP} for $\hat{\pi} = \frac{1}{2}$ or $\hat{\pi} = \frac{1}{2} \pm 1/2n$ only. However, it can be shown that $E(CER_{CV} - CER_{ACT})$ is zero for n even and negligibly small for n odd. So CER_{CV} appears a poor estimator of the right thing. The estimator $\widehat{CER}_{ACT} = CER_{APP} + \hat{\Omega}$ appears more precise, but slightly biased.

In Figures 1 and 2 the mean CERs and their spread as measured by $\sqrt{[E(CER - CER_{ACT})^2]}$ are given for different values of π with $n = 25$. The conclusion is that CER_{CV} is indeed a poor estimator of CER_{ACT}, while CER_{APP} and $\widehat{CER}_{ACT}$ appear to perform equally well. Apparently, the correction by $\hat{\Omega}$ decreases the bias but increases the variance.

Mean squared error

The expected squared error of p itself as a predictor of Y_{new} is given by

$$MSE = E_\pi(Y - p)^2 = \pi(1 - \pi) + (p - \pi)^2. \tag{21}$$

Since this is completely in line with the previous section with $\mu = \pi$ and $\sigma^2 = \pi(1 - \pi)$, we mention it without further discussion. All the results of Section 2 apply.

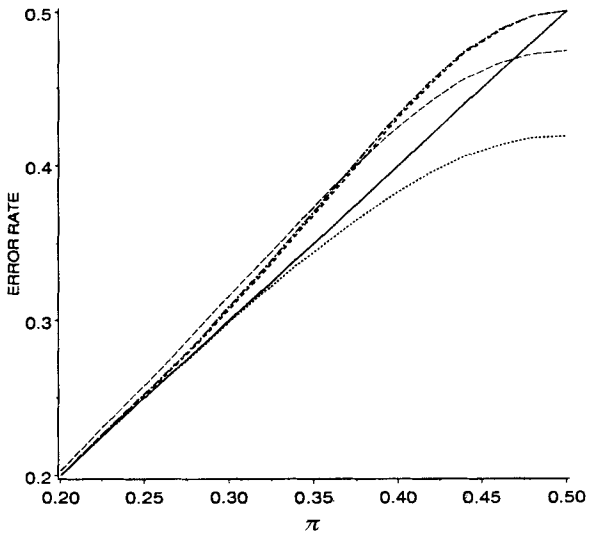

Figure 1. Mean classification error rates with $n = 25$

—— CER_{OPT} (13) .... CER_{APP}(14)
---- CER_{ACT} (12 with $p = \hat{\pi}$) ----- CER_{CV} (19)
--- $\widehat{CER}_{ACT} = CER_{APP} + \hat{\Omega}$ (16)

Numbers in brackets refer to equations in the text

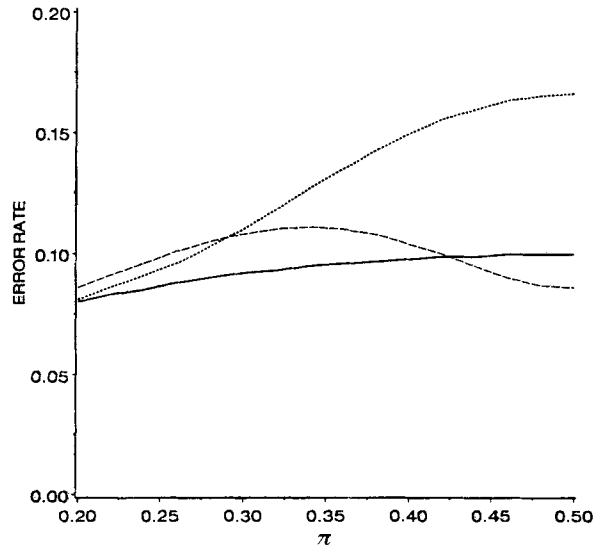

Figure 2. Mean deviation of estimated classification error rates from the actual error rate CER_{ACT} with $n = 25$

—— CER_{APP} .... CER_{CV} ---- $\widehat{CER}_{ACT}$

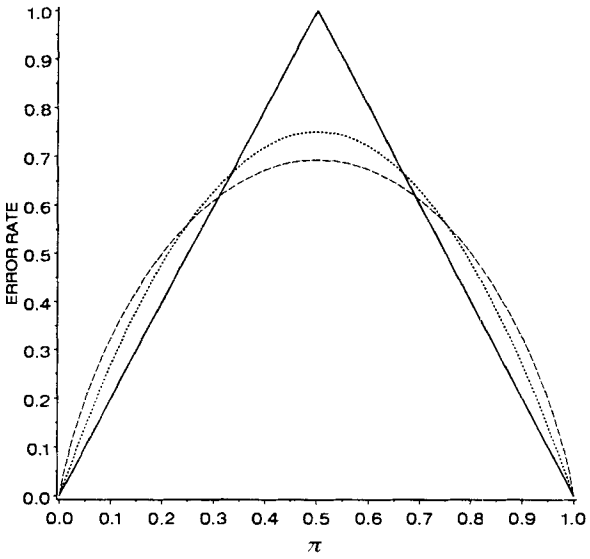

Figure 3. Rescaled optimum error rates for different definitions of error rate. Scaling is such that the integral under the curve equals 0·5 for all three curves

—— $2CER_{OPT}$ .... $3MSE_{OPT}$ – – – – MML_{OPT}

Kullback–Leibler error rate

A third type of error rate measure is defined by

$$MML_{ACT} = E_\pi\{-Y\log(p) - (1-Y)\log(1-p)\}. \tag{22}$$

MML stands for mean value of minus log-likelihood.

At first sight this error rate seems a bit strange, but it has many useful features as we shall now see. The first observation is that

$$MML_{OPT} = -\pi\log\pi - (1-\pi)\log(1-\pi). \tag{23}$$

This is just the entropy of the binomial distribution. It is maximal at $\pi = \frac{1}{2}$ and minimal at $\pi = 0$ or $\pi = 1$. This is in accordance with intuition that Y is harder to predict for π about 0·5 than for π near 0 or 1. Figure 3 shows the rescaled optimum error rates for all three types considered here, showing that all three measures are more or less the same.

Secondly,

$$MML_{ACT} - MML_{OPT} = \pi\log\left(\frac{\pi}{p}\right) + (1-\pi)\log\left(\frac{1-\pi}{1-p}\right) \tag{24}$$

is strictly positive for $p \neq \pi$. Actually, it is just the Kullback–Leibler distance of the estimated binomial distribution (parameter p) from the true binomial distribution (parameter π). So, the further p is away from π, the larger is the error rate.

The apparent error rate of $\hat{\pi} = \bar{Y}$ is given by

$$\begin{aligned} MML_{APP} &= -\hat{\pi}\log(\hat{\pi}) - (1-\hat{\pi})\log(1-\hat{\pi}) \\ &= -\frac{1}{n}\max_p l(Y_1, \ldots, Y_n, p), \end{aligned} \tag{25}$$

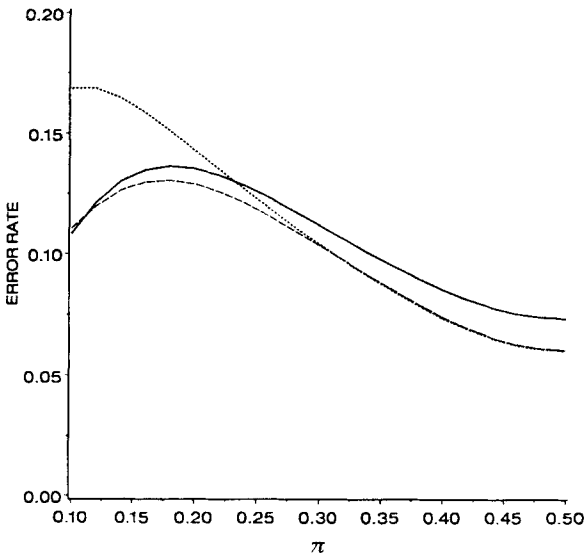

Figure 4. Mean deviation of estimated MML error rates from the actual error rate MML_{ACT} with $n = 25$
—— MML_{APP} (25) .... MML_{CV} ---- $\widehat{MML}_{ACT}$ (28)
Numbers in brackets refer to equations in the text.

where l stands for log-likelihood. The optimism of MML_{APP} with respect to MML_{ACT} is

$$O = MML_{ACT} - MML_{APP} = (\hat{\pi} - \pi)\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right). \tag{26}$$

A Taylor series expansion of $\log[\hat{\pi}/(1-\hat{\pi})]$ yields

$$O = (\hat{\pi} - \pi)\log\frac{\pi}{1-\pi} + \frac{(\hat{\pi}-\pi)^2}{\pi(1-\pi)} + o((\hat{\pi}-\pi)^2). \tag{27}$$

The expected value of the first two terms is $1/n$; this leads to a very simple non-stochastic correction term $\hat{\Omega} = 1/n$, and

$$\widehat{MML}_{ACT} = MML_{APP} + 1/n. \tag{28}$$

Essentially this is Akaike's information principle, which states that the observed maximum log-likelihood has to be reduced by the dimension of the parameter to obtain an (unbiased) estimate of the predictive value of the model.[9,10]

By similar reasoning, it can be shown that approximately

$$MML_{CV} - MML_{APP} = \frac{1}{n-1}.$$

So, MML_{CV} and $\widehat{MML}_{ACT}$ virtually coincide. The same phenomenon was observed for MSE in Section 2. However, both methods only produce an estimate of MML_{ACT}. Since the bias correction is (almost) non-stochastic, MML_{CV} and $\widehat{MML}_{ACT}$ are better estimators of MML_{ACT} than MML_{APP}.

Figure 4 shows $\sqrt{[E_\pi(MML - MML_{ACT})^2]}$ as a function of π for $n = 25$. Apparently, $\widehat{MML}_{ACT}$ can still be inaccurate.

4. LEAST SQUARES REGRESSION

Our next step is to study regression. We let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be a sequence of observations, where Y is the dependent outcome variable and X is a p-dimensional covariate vector. Our purpose is to predict the value of a future observation Y_{new} given the value of the corresponding covariates X_{new}. We assume that the pairs $(Y_1, X_1), \ldots, (Y_n, X_n)$ are independent and that the Y_i are homoskedastic, so that

$$\text{var}(Y_i | X_i) = \sigma^2. \tag{29}$$

Let μ_i be the expected value of Y_i given X_i. The linear regression model

$$\mu_i = X_i' \beta, \tag{30}$$

where β represents a vector of regression coefficients, serves as a plausible model for μ_i. Observe that at this stage we do not make the assumption of normal distributions. Obviously, the optimum mean squared error rate is given by

$$MSE_{\text{OPT}} = \sigma^2, \tag{31}$$

which can be achieved if the regression coefficients are known and the linear model (30) holds true.

The usual ordinary least squares (OLS) estimator of β is given by

$$\hat{\beta} = (\sum X_i X_i')^{-1} \sum Y_i X_i. \tag{32}$$

Using $\hat{\beta}$ as an estimate of β, we obtain

$$\hat{Y}_{\text{new}} = X_{\text{new}}' \hat{\beta} \tag{33}$$

as a predictor for Y_{new}. The actual error rate of this predictor is

$$MSE_{\text{ACT}}(X_{\text{new}}) = (X_{\text{new}}' \hat{\beta} - \mu_{\text{new}})^2 + \sigma^2. \tag{34}$$

To get rid of the dependence on X_{new}, we want to take the average value of $MSE_{\text{ACT}}(X_{\text{new}})$. There are two different approaches, which arise from different views of the design of the experiment. The first approach is to consider the whole design as replicated, yielding the same X_i and the same unknown μ_i as in the observed sample. The second approach is to consider the X_i as random sample from some unknown distribution. We will discuss both approaches in more detail.

Replicated design

This is the analysis of variance situation where the X_i correspond to factors for stratification, treatment and so on. If the design is replicated, and the Y_i are to be predicted, the actual mean squared error is given by

$$MSE_{\text{ACT}} = \frac{1}{n} \sum (X_i' \hat{\beta} - \mu_i)^2 + \sigma^2, \tag{35}$$

while the apparent error rate is

$$MSE_{\text{APP}} = \frac{1}{n} \sum (Y_i - X_i' \hat{\beta})^2. \tag{36}$$

Using the property that $\text{cov}(\hat{\beta}) = \sigma^2 (\sum X_i X_i')^{-1}$, it can be shown that

$$E(MSE_{\text{ACT}} - MSE_{\text{APP}}) = \frac{2p}{n} \sigma^2. \tag{37}$$

This result does not depend on the validity of model (30). However, it does assume that the μ_i do not change. Let s^2 be any estimate of σ^2; then

$$\widehat{MSE}_{\text{ACT}} = MSE_{\text{APP}} + \frac{2p}{n} s^2 \tag{38}$$

can serve as an estimate of MSE_{ACT}. Essentially this is Mallows's C_P.[5] Mallows suggests the use of a high-dimensional (large p) model to estimate σ^2, and $\widehat{MSE}_{\text{ACT}}$ as a tool to select the best predictive model. If we assume that model (30) is correct, we can use $[n/(n-p)]\,MSE_{\text{APP}}$ as an estimate of σ^2, giving

$$\widehat{MSE}_{\text{ACT}} = \frac{n+p}{n-p} MSE_{\text{APP}}. \tag{39}$$

Observe that $p = 1$ and $X_i \equiv 1$ brings us back to the situation of Section 2.

Random X

This corresponds to the observational study, where X and Y are both random variables. For convenience we assume that the linear model (30) holds true for some β_0. Roughly speaking, we consider the deviation from (30), the so-called model error, as a random variable that is incorporated in the total error term (see Breiman and Freedman[11] for a more rigorous argument based on the assumption of a multivariate normal distribution). In this situation, the actual error rate is obtained by averaging over the distribution of (Y, X). That yields

$$MSE_{\text{ACT}} = (\hat{\beta} - \beta_0)' C_X (\hat{\beta} - \beta_0) + \sigma^2, \tag{40}$$

where

$$C_X = E(XX'). \tag{41}$$

The expected value of MSE_{ACT} with respect to the distribution of $X_1, \ldots, X_n$ is given by

$$E(MSE_{\text{ACT}}) = \sigma^2 \left(1 + \frac{1}{n} E\{\text{trace}(C_X \hat{C}_X^{-1})\} \right),$$

where $\hat{C}_X = (1/n)\sum X_i X_i'$. When X has a multivariate normal distribution,

$$E\{\text{trace}(C_X \hat{C}_X^{-1})\} = \frac{np}{n-p-1},$$

resulting in

$$E(MSE_{\text{ACT}}) = \sigma^2 \left(1 + \frac{p}{n-p-1} \right). \tag{42}$$

This is the S_P criterion for the selection of a regression model.[11-14] Observe that (42) depends on the assumption of X having a multivariate normal distribution.

One way of avoiding unnecessary assumptions is to use cross-validation to assess predictive value. We let $\hat{\beta}_{(-i)}$ be the estimate of β based on the observations $Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n$. It can be shown (see Cook and Weisberg,[15] Section 2.2.3) that

$$Y_i - X_i' \hat{\beta}_{(-i)} = (Y_i - X_i' \hat{\beta})/(1 - h_{ii}), \tag{43}$$

where $h_{ii} = X_i'(\sum X_j X_j')^{-1} X_i$. This leads to

$$MSE_{CV} = \frac{1}{n}\sum (Y_i - X_i'\hat{\beta})^2/(1 - h_{ii})^2, \tag{44}$$

which is Allen's PRESS.[7,16] Strictly speaking, MSE_{CV} estimates MSE_{ACT} for the case of $(n-1)$ instead of n observations and has the nice feature of mimicking the prediction problem exactly. However, it should be emphasized that MSE_{CV} is only an estimator of MSE_{ACT} and that this estimate can be quite inaccurate for moderately large n. Since $\sum h_{ii} = p$, MSE_{CV} can be simplified by replacing h_{ii} by the average value p/n, leading to

$$MSE_{GCV} = (1 - p/n)^{-2} MSE_{APP} = \frac{n}{n-p} s^2. \tag{45}$$

This procedure was coined 'generalized cross-validation' (GCV) by Golub, Heath and Wahba.[17] Their argument for introducing GCV is based on invariance under rotation in the model space, which is not very appealing. From our point of view, GCV can be seen as a simplification of CV. It clearly shows the link between cross-validation and the S_P criterion because MSE_{GCV} differs very little from S_P. If the ratio p/n is not too large ($p/n \leqslant 0{\cdot}1$), all estimates of MSE_{ACT}, including Mallows's C_P, are very similar. However, they may all be quite wrong.

Cross-validation can also be used in the replicated design, although it has less intuitive appeal.

Summarizing, there are several estimates available for the actual error rate which all correct the apparent error rate in more or less the same way for the estimation of the unknown β. All are based on an average with respect to the distribution of $\hat{\beta}$ and all use s^2 as an estimate of σ^2. In that respect they must all be handled with care, because in the true prediction problem $\hat{\beta}$ is not a random variable but fixed. Moreover σ^2 is unknown as well and s^2 might be a poor estimate of it. It should be kept in mind that in the prediction problem past observations are given and only future observations are unknown. Averaging over past observations is only a technical trick to obtain at least an estimate of the future error rate.

The estimators of MSE_{ACT} presented here can all be found in the vast literature on selection of models in linear regression.[6,12-14,18,19] Bootstrapping[20] might be a competitor for cross-validation, but is very time-consuming.

5. LOGISTIC REGRESSION

Logistic regression as a way of modelling a binary outcome variable is still gaining popularity in clinical biostatistics. On the one hand, clinicians tend to dichotomize variables to make things easier to understand; on the other, important outcome variables are of an intrinsic dichotomous nature, with survival status (dead/alive) the most obvious one. We consider a sequence of observations $(Y_1, X_1), \ldots, (Y_n, X_n)$ as in Section 4, but now Y_i is a binary (0/1) variable with

$$P(Y_i = 1 \mid X_i) = \pi(X_i), \tag{46}$$

which we model in the usual logistic form as

$$\pi(X_i) = e^{X_i'\beta}/(1 + e^{X_i'\beta}). \tag{47}$$

To measure the error rate of a prediction of Y_{new} based on an estimate p_{new} of $P(Y_{new} = 1)$ we could use all three error rates of Section 3. In this section we focus on CER and MML. Let $\hat{\beta}$ be the maximum likelihood estimator of β and let $\hat{\pi}_i$ be the estimate of $\pi(X_i)$ based on $\hat{\beta}$ using (47).

Then the apparent classification error rate is given by

$$CER_{\text{APP}} = \frac{1}{n}\sum\{Y_i[\hat{\pi}_i < \tfrac{1}{2}] + (1 - Y_i)[\hat{\pi}_i > \tfrac{1}{2}] + \tfrac{1}{2}[\hat{\pi}_i = \tfrac{1}{2}]\}, \tag{48}$$

where $[\cdot]$ stands for the indicator function (see (19)). The apparent MML error rate is given by

$$\begin{aligned} MML_{\text{APP}} &= -\frac{1}{n}\sum\{Y_i \log(\hat{\pi}_i) + (1 - Y_i)\log(1 - \hat{\pi}_i)\} \\ &= -\frac{1}{n} l(\hat{\beta}), \end{aligned} \tag{49}$$

with l being the log-likelihood of the observations. The actual error rates for the replicated design are given by

$$CER_{\text{ACT}} = \frac{1}{n}\sum\{\pi_i[\hat{\pi}_i < \tfrac{1}{2}] + (1 - \pi_i)[\hat{\pi}_i > \tfrac{1}{2}] + \tfrac{1}{2}[\hat{\pi}_i = \tfrac{1}{2}]\}$$

$$MML_{\text{ACT}} = -\frac{1}{n}\sum\{\pi_i \log(\hat{\pi}_i) + (1 - \pi_i)\log(1 - \hat{\pi}_i)\}. \tag{50}$$

In Efron[4] an approximation for $E(CER_{\text{ACT}} - CER_{\text{APP}})$ is given which can be used for correcting the optimism bias of the apparent error rate. It is a generalization of the correction (16) given in Section 3. The interested reader is referred to Efron's paper.

The mathematics are much easier for the MML error rate. A first order approximation for the optimism O is given by

$$\begin{aligned} O &= MML_{\text{ACT}} - MML_{\text{APP}} \\ &= \frac{1}{n}(\hat{\beta} - \hat{\beta}_0)'\sum \pi_i(1 - \pi_i)X_i X_i'\hat{\beta}. \end{aligned} \tag{51}$$

(See Van Houwelingen and Le Cessie[21] for more details.) Since $\text{cov}(\hat{\beta}) = (\sum \pi_i(1 - \pi_i)X_i X_i')^{-1}$, it follows that

$$E(O) = p/n \tag{52}$$

to a first approximation. So the optimism correction for the apparent error rate is simply p/n. Once again this is essentially Akaike's information criterion.[9] It must be emphasized that the analysis as given above is only valid under the assumption that the logistic regression model (47) holds true.

In the 'random X' case, cross-validation can be used to assess the predictive power of a prediction rule. Let $\tilde{\pi}_i$ be the estimate of π_i obtained by substituting $\hat{\beta}_{(-i)}$ in (47), that is the estimate based on all observations except the ith. The CV error rate can be found by replacing $\hat{\pi}_i$ by $\tilde{\pi}_i$ in (48) and (49), yielding

$$CER_{\text{CV}} = \frac{1}{n}\sum\{Y_i[\tilde{\pi}_i < \tfrac{1}{2}] + (1 - Y_i)[\tilde{\pi}_i > \tfrac{1}{2}] + \tfrac{1}{2}[\tilde{\pi}_i = \tfrac{1}{2}]\}$$

$$MML_{\text{CV}} = -\frac{1}{n}\sum\{Y_i \log(\tilde{\pi}_i) + (1 - Y_i)\log(1 - \tilde{\pi}_i)\}. \tag{53}$$

It can be shown that, approximately,

$$MML_{CV} - MML_{APP} = \frac{1}{n}\sum \frac{(Y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)} \frac{h_{ii}}{(1-h_{ii})}, \tag{54}$$

where $h_{ii} = \hat{\pi}_i(1-\hat{\pi}_i)X_i'(\sum \hat{\pi}_j(1-\hat{\pi}_j)X_jX_j')^{-1}X_i$. Bearing in mind that the average value of $h_{ii} = p/n$, the right hand side of (54) has an expected value of order p/n. So the difference between MML_{CV} and $\widehat{MML}_{ACT} = MML_{APP} + p/n$ is expected to be very small if the logistic model holds true.

As an example we consider data from a Dutch study on infants who are preterm and/or small for gestational age.[22] The effect of weight at birth and gestational age on morbidity after two years were studied using logistic regression. The results are given in Table I and show that the predictive power of the models is rather poor (use Figure 3 as a gauge for the *MML* error rate). The classification error rate (*CER*) is quite constant over the models and is hardly improved upon by model extension. This could be an argument against the use of *CER*. Cross-validation agrees quite well with Efron's model-based optimism corrections in the first two models. In the third model there is less agreement. The difference between MML_{CV} and $\widehat{MML}_{ACT}$ for the third model may be an indication of an imperfect fit or instability of the model.

Generally, cross-validation as discussed above is not well suited to picking up lack of fit. The reason is that deviation from the model can be detected much better by grouping of the data in larger cells, for example the Hosmer–Lemeshow statistic.[23] One way of mimicking that situation is to group observations according to the values of the covariate vector and to leave out subgroups instead of single observations in the cross-validation process. That comes close to our general goodness-of-fit statistic for the logistic model.[24]

The results from this method are given in Table II and indicate lack of fit for the first model and less so for the second model. This could be expected, since the second model is a clearly significant extension of the first one.

6. SURVIVAL ANALYSIS

Here the methods are much more complicated than with the (logistic) regression models considered so far. Error rates are much harder to define and, if defined, are harder to interpret intuitively. We concentrate on giving some methods for cross-validation that can be used with survival data.

We let $(Y_1, \delta_1, X_1), \ldots, (Y_n, \delta_n, X_n)$ denote our data set, where Y_i is the observed survival time, δ_i is a censoring indicator (0 if the survival time is censored and 1 otherwise) and X_i represents a vector of covariates. Let $S_i(t \mid \beta) = S(t \mid X_i, \beta)$ be a model for the survivor function and $h_i(t \mid \beta)$ be the corresponding hazard rate. The full log-likelihood of the data is given by

$$l(\beta) = \sum \{\log(S_i(Y_i \mid \beta)) + \delta_i \log(h_i(Y_i \mid \beta))\}. \tag{55}$$

Let $\hat{\beta}_{(-i)}$ be the maximum likelihood estimate (*MLE*) of β based on all but the ith observation. A generalization of the *MML* error rate for this case is given by

$$MML_{CV} = -\frac{1}{n}\sum \{\log(S_i(Y_i \mid \hat{\beta}_{(-i)})) + \delta_i \log(h_i(Y_i \mid \hat{\beta}_{(-i)}))\}. \tag{56}$$

For example, a simple exponential survival time distribution $h_i(t) \equiv e^{\beta}$ and no censoring yields $\hat{\beta} = -\log(\bar{Y})$ and

$$MML_{CV} = \frac{1}{n}\sum \{Y_i/\bar{Y}_{(-i)} + \log(\bar{Y}_{(-i)})\}. \tag{57}$$

Table I. Error rates for several logistic regression models applied to the Dutch POPS data on 1310 preterm infants[22]

	Linear terms	Linear + quadratic terms	Linear + logarithmic terms
No. of covariates*	3	5	5
CER			
Apparent (48)†	0·2473	0·2458	0·2481
Model-based optimism correction	0·0010	0·0016	0·0028
Estimated actual	0·2483	0·2474	0·2509
Cross-validation error rate (53)	0·2481	0·2466	0·2496
MML			
Apparent (49)	0·5356	0·5180	0·5167
Model-based optimism correction (52)	0·0023	0·0038	0·0038
Estimated actual	0·5379	0·5218	0·5205
Cross-validation error rate (53)	0·5382	0·5216	0·5248

* Constant term included.
† Numbers in brackets refer to equations in the text.

Table II. Values of $MML_{CV} - MML_{APP}$ at different stages of grouping

Average size of subgroup	Linear terms	Linear + quadratic terms
1·0	0·0026	0·0036
1·1	0·0026	0·0036
5·0	0·0028	0·0037
15·1	0·0041	0·0044
28·5	0·0053	0·0042
Nominal difference (p/n)	0·0023	0·0038

This expression measures the deviation of Y_i with respect to $\bar{Y}_{(-i)}$, since $x/\alpha + \log(\alpha)$ is minimal at $\alpha = x$.

The MML_{CV} cross-validation criterion can be used in parametric and non-parametric models as well. However, with a non-parametric model it is very time-consuming to re-estimate survival and hazard function for all subsets obtained by leaving out a single observation. Therefore we also consider the partial likelihood in the proportional hazard model, that is

$$L(\beta) = \prod_{\delta_i = 1} \frac{e^{X_i'\beta}}{\sum_{j \in R_i} e^{X_i'\beta}}, \tag{58}$$

where the product is over all non-censored survival times and R_i is the set of individuals at risk at time Y_i. (For convenience it is assumed that no two survival times coincide.)

Let $L_{(-i)}(\beta)$ be the likelihood if the ith observation is left out and $\hat{\beta}_{(-i)}$ the MLE in that case. Given $\hat{\beta}_{(-i)}$ and the pattern of (censored) survival times, the probability that the ith

individual lives to time Y_i and dies at time Y_i (if $\delta_i = 1$) is given by

$$PR(Y_i, \delta_i) = \prod_{\substack{Y_j < Y_i \\ \delta_j = 1}} \left[1 - \frac{e^{X_i'\hat{\beta}_{(-i)}}}{\sum_{k \in R_j} e^{X_k'\hat{\beta}_{(-i)}}} \right] \left[\frac{e^{X_i'\hat{\beta}_{(-i)}}}{\sum_{k \in R_i} e^{X_k'\hat{\beta}_{(-i)}}} \right]^{\delta_i} \tag{59}$$

It is straightforward to check that

$$PR(Y_i, \delta_i) = \frac{L(\hat{\beta}_{(-i)})}{L_{(-i)}(\hat{\beta}_{(-i)})}. \tag{60}$$

Hence a cross-validation criterion is given by

$$MMLP_{CV} = -\frac{1}{n} \sum \{ l(\hat{\beta}_{(-i)}) - l_{(-i)}(\hat{\beta}_{(-i)}) \}, \tag{61}$$

where l stands for log-likelihood ($\log(L)$) and the P in $MMLP$ refers to 'partial'. This has the intuitive interpretation of an error rate, because it equals zero if the survival of all individuals is correctly 'predicted', and it is positive if not.

To illustrate the use of this criterion by a simple example, we consider two (treatment) groups A and B and let the survival pattern of the observations be A, B, A, B, A, B+, A+, B+, so that the patient who died first belonged to group A, the second to group B and so on. (The + symbol denotes censoring).

Consider the model $h_A(t) = e^{\beta} h_B(t)$. The MLE of β is 0·55, showing that survival in group A is on average shorter than in B.

The results of cross-validation are given in Table III. The estimate of β when leaving out the first individual equals 0·1732. Using this value, the probability that the first individual dies first is $e^{0\cdot1732}/(4e^{0\cdot1732} + 4) = 0\cdot1358$, yielding a contribution to $MMLP_{CV}$ of $-\log(0\cdot1358) = 1\cdot9966$. If there were no difference between A and B, the contribution would be $-\log(1/8) = 2\cdot08$. If the difference between A and B were maximal, that is all As die first, the contribution would be $-\log(1/4) = 1\cdot39$. By this kind of reasoning, some insight can be gained into the meaning of the observed value of $MMLP_{CV}$. The same calculation on the pattern A, A, B, B, A, A+, B+, B+ yields $\hat{\beta} = 0\cdot6891$ and $MMLP_{CV} = 1\cdot8017$. Although the estimate of β is higher, the value of $MMLP_{CV}$ indicates that the predictive information in both situations is about the same.

We have used $MMLP_{CV}$ as a criterion for model selection in survival analysis and were quite satisfied with the results. Details of this research will be published elsewhere.

A drawback of $MMLP_{CV}$ is its lack of intuitive appeal. Another possibility is to consider cross-validation by pairs instead of by single observations. Let $b = \hat{\beta}_{(-i,-j)}$ be the estimate of β after leaving out observations i and j. The probability that $Y_i < Y_j$ is given by

$$P(Y_i < Y_j \mid b) = \frac{e^{X_i'b}}{e^{X_i'b} + e^{X_j'b}}. \tag{62}$$

An error rate based on this property of the proportional hazard model could be

$$-\frac{2}{n(n-1)} \sum_{Y_i < Y_j} \log \left[\frac{e^{X_i'\hat{\beta}(-i,-j)}}{e^{X_i'\hat{\beta}(-i,-j)} + e^{X_j'\hat{\beta}(-i,-j)}} \right] \tag{63}$$

with minor adjustments for censored observations. Although this error rate has much intuitive appeal, computation is very time-consuming.

Summarizing, cross-validation error rates for survival analysis can be defined and used in practice, but there is still much work to be done in this field.

Table III. Cross-validation on the pattern A, B, A, B, A, B+, A+, B+ for survival observations in two groups A and B

Observation left out	Estimate of β	Contribution to $MMLP_{CV}$
1	0·1732	1·9966
2	1·1424	2·6578
3	0·3466	1·9800
4	0·9331	2·3519
5	0·6200	2·0850
6	0·1354	0·9122
7	1·0483	1·6053
8	0·1354	0·9122
Mean value		1·8126

7. IMPROVING THE PREDICTORS

So far we have based our predictions upon ordinary least squares or *MLE* estimates. There are, however, ways of improving these predictors. To give an idea of how this can be achieved, we present a simple example.

Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be a sequence of independent pairs with $E(Y_i|X_i) = \beta X_i$ and $\text{var}(Y_i|X_i) = \sigma^2$, where X_i represents a single covariate. For the time being we assume that there is no constant term in the regression model. Let

$$\hat{\beta} = \frac{\sum X_i Y_i}{\sum X_i^2} \tag{64}$$

be the OLS estimator of β, and consider the predictor

$$\hat{Y}_i(c) = c\hat{\beta}X_i. \tag{65}$$

The mean squared error of this predictor in the replicated design case is given by

$$MSE(c) = \frac{1}{n}\sum X_i^2(\beta - c\hat{\beta})^2 + \sigma^2. \tag{66}$$

Its expected value with respect to the distribution of $\hat{\beta}$ is given by

$$E\{MSE(c)\} = (1-c)^2\beta^2\sum X_i^2/n + \left(1 + \frac{c^2}{n}\right)\sigma^2, \tag{67}$$

which is minimal for

$$c = \frac{\beta^2\sum X_i^2}{\beta^2\sum X_i^2 + \sigma^2} = \frac{\beta^2/\text{var}(\hat{\beta})}{(\beta^2/\text{var}(\hat{\beta})) + 1}. \tag{68}$$

So there is an optimal value of c, between 0 and 1, for which $E(MSE(c)) < E(MSE(1))$, the expected actual error rate based upon the OLS estimator $\hat{\beta}$.

A heuristic estimate of this optimal value is given by

$$\hat{c}_{\text{heur}} = \frac{(\hat{\beta}^2/\text{var}(\hat{\beta})) - 1}{\hat{\beta}^2/\text{var}(\hat{\beta})}. \tag{69}$$

The generalization of $\hat{c}_{\text{heur}}$ to multiple regression with k covariates and a constant term is given by

$$\hat{c}_{\text{heur}} = \frac{SS_{\text{exp}} - ks^2}{SS_{\text{exp}}} \tag{70}$$

where $SS_{\text{exp}} = \sum(\hat{Y} - \bar{Y})^2 = \sum((X_i - \bar{X})'\hat{\beta})^2$ is the explained sum of squares. The adjusted predictor is

$$\hat{Y}_{\text{new}} = \bar{Y} + \hat{c}_{\text{heur}}(X_{\text{new}} - \bar{X})'\hat{\beta}. \tag{71}$$

This resembles very much the shrinkage predictors of Copas[25] and Jones and Copas.[26]

There are two cross-validation techniques for obtaining the optimal value of c: ridge regression and calibration.

Ridge regression

If $\hat{Y}_{(-i)}(c)$ is the predictor for Y_i based on the other observations, then $\hat{c}_{\text{RR}}$ is the value of c which minimizes[27,28]

$$MSE_{\text{CV}} = \frac{1}{n}\sum(Y_i - \hat{Y}_{(-i)}(c))^2. \tag{72}$$

This can be extended to multiple regression, where Hoerl and Kennard[27,28] advocate an estimator of β of type

$$\hat{\beta}(c) = (\sum X_i X_i' + cI)^{-1}\sum Y_i X_i. \tag{73}$$

This can be interpreted as a Bayes estimator with respect to a $N(0, c^{-1}\sigma^2 I)$ prior on β, or as the OLS estimator under restriction upon $|\beta|^2$ (see Draper and Smith,[29] Section 6.7). Determination of the optimal value of c by means of cross-validation is not easy, because there is no explicit expression for $\hat{c}_{\text{RR}}$.

Calibration

As far as we are aware, this approach is fairly new, although something similar is given by Rao.[30] The idea is to start with ordinary least squares and obtain $\hat{Y}_{(-i)}$. Then estimate the optimal $\hat{c}_{\text{cal}}$ by regression of Y_i on $\hat{Y}_{(-i)}$, in this case by minimization of $(1/n)\sum(Y_i - c\hat{Y}_{(-i)})^2$. Use $\hat{c}_{\text{cal}}Y_i$ as a predictor for future observations, hoping that it yields improved predictions.

In this example, ridge regression and calibration coincide because $\hat{Y}_{(-i)}(c) = c\hat{Y}_{(-i)}$. Unlike ridge regression, the generalization of the calibration approach to multiple regression is easy to carry out.

General *cross-validation calibration* for regression with a constant term is defined by the following procedure:

1. Obtain $\hat{Y}_{(-i)}$ based upon OLS. (Software packages like SPSS produce this quantity automatically.)
2. Perform regression of Y_i on $\hat{Y}_{(-i)}$ to obtain $\hat{c}_{\text{cal}}$. (If $\hat{c}_{\text{cal}} < 0$, truncate at 0.)
3. Use $\hat{Y}_{\text{new}} = \bar{Y} + \hat{c}_{\text{cal}}(X_{\text{new}} - \bar{X})'\hat{\beta}$ as the predictor.

The main difference between this predictor and the shrinkage predictor of Copas[25] or the related predictor (71) is the way the shrinkage factor is obtained. The latter predictors are based on certain model assumptions, while cross-validation calibration is assumption-free and can even be used if, for example, the assumption of homoskedasticity does not hold.

Table IV. Monte Carlo simulations for $n = 50$, $k = 4$ and $\beta = 1$ (500 replications). The shrinkage factor $\hat{c}_{cal}$ is obtained by cross-validation calibration

	Mean	Standard deviation
s^2	1·000	0·211
MSE_{CV} (44)*	1·116	0·235
MSE_{ACT} (40)	1·118	0·074
$\hat{c}_{cal}$	0·884	0·056
$MSE(\hat{c}_{cal})_{ACT}$	1·118	0·076
$\hat{c}_{heur}$ (70)	0·918	0·036
$MSE(\hat{c}_{heur})_{ACT}$	1·115	0·073

Fraction of samples for which $MSE(\hat{c}_{cal})_{ACT} < MSE_{ACT}$: 0·570
Fraction of samples for which $MSE(\hat{c}_{heur})_{ACT} < MSE_{ACT}$: 0·620

* Numbers in brackets refer to equations in the text.

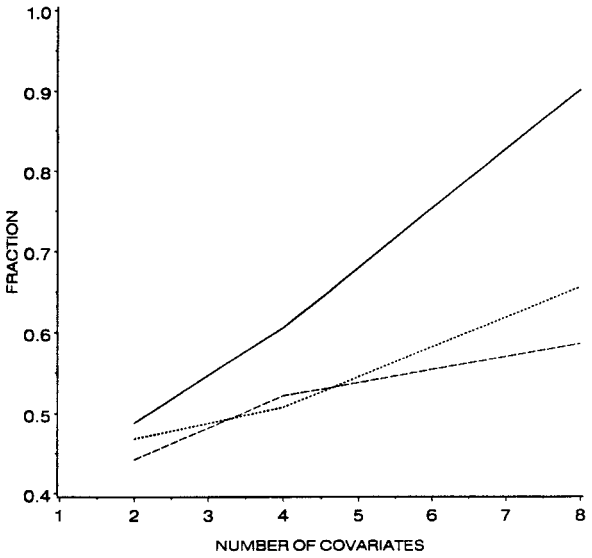

Figure 5. Fraction of Monte Carlo samples for which the calibration predictor performed better than the OLS predictor ($n = 25$)

—— $\beta = 0{\cdot}5$ .... $\beta = 1$ ---- $\beta = 2$

We studied the merits of the calibration approach by Monte Carlo simulation with the model $Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + e$, where $X_1, \ldots, X_k$ and e are independent N(0, 1). Without loss of generality, we took $\beta_i = 0$ for $i = 0, 2, \ldots, k$ and varied the value of $\beta_1 = \beta$. We studied sample sizes of $n = 25$ and $n = 50$, $k = 2, 4$ and 8, $\beta = 0{\cdot}5$, 1 and 2, and took 500 replications. The detailed results for $n = 50$, $k = 4$ and $\beta = 1$ are given in Table IV. The table shows that application of $\hat{c}_{cal}$ gives slightly better results than OLS, while $\hat{c}_{heur}$ performed even better. It is also striking that the estimate of MSE_{ACT} is nearly unbiased, but has rather a large standard deviation.

In Figure 5 and Figure 6 plots are given of the fraction of samples for which the calibrated prediction performs better than the OLS predictor. The heuristic shrinkage factor $\hat{c}_{heur}$ performed

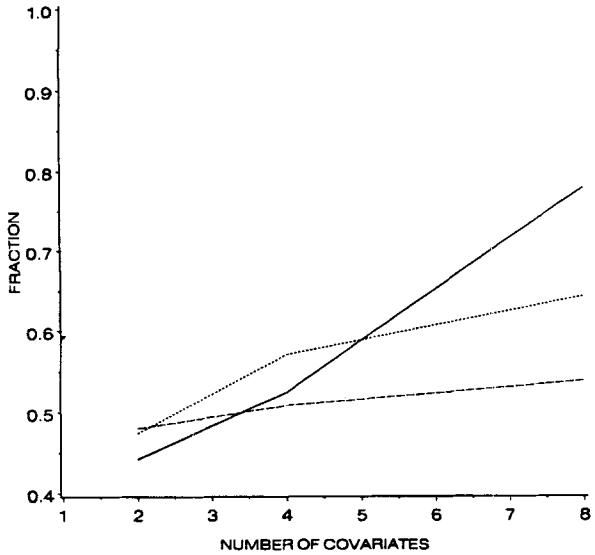

Figure 6. Fraction of Monte Carlo samples for which the calibration predictor performed better than the OLS predictor ($n = 50$)

—— $\beta = 0{\cdot}5$ .... $\beta = 1$ ---- $\beta = 2$

slightly better. However, it is based on the assumption of homoskedasticity, while $\hat{c}_{cal}$ is also valid in the heteroskedastic case. Therefore we recommend $\hat{c}_{cal}$ as a more general adjustment than $\hat{c}_{heur}$. The striking feature of the figures is that the improvement increases with the dimension of the model. An explanation might be that the optimal c is better estimated for higher dimensions.

Although the improvement does not appear very big, cross-validation can be useful in high-dimensional models. (This is also true for ridge regression.) The nice feature of the approach is that it can easily be extended to generalized linear models like logistic regression and Cox regression for survival data. The general procedure is first to obtain the estimate $X_i'\hat{\beta}_{(-i)}$ for all sample points. Approximations can be found in Pregibon[31] for logistic regression and in Storer and Crowley[32] for Cox regression. The next step is to fit a model with $X_i'\hat{\beta}_{(-i)}$ as a single covariate, yielding the shrinkage factor $\hat{c}_{cal}$ as the regression coefficient. Finally the adjusted predictions are obtained by using an adjusted covariate vector $X^*_{new} = \bar{X} + \hat{c}_{cal}(X_{new} - \bar{X})$ in the model based on all data.

It is also possible to give a generalization of $\hat{c}_{heur}(71)$. A marked difference of $\hat{c}_{cal}$ from $\hat{c}_{heur}$ gives an indication of lack of fit of the model. We come back to this in the next section.

8. THE SPLIT-SAMPLE APPROACH

In the situations we have studied so far, we assumed that the set of covariates that was used in the statistical model was fixed. We estimated the regression coefficients and applied cross-validation to assess predictive power. Section 7 considered adjustment of the predictor by a single shrinkage parameter that was obtained either by a heuristic argument or by calibration on the cross-validation predictions. Strictly speaking, a second round of cross-validation would be needed to assess such a procedure. Stone[6] discusses such multi-stage cross-validation procedures. Gong[33] studies the assessment of forward selection procedures by several methods. However, assessing

the whole variable selection and model building process by cross-validation or similar techniques gets very complicated, since we have to repeat the whole process as many times as we have observations in our sample. Apart from being very (computer) time consuming, such a validation process asks for very strict rules for the model building process. However, it is hard to give such explicit rules because model building often proceeds by trial and error. Aspects of the process such as the combination of outcome categories of a categorial covariate, the inclusion of quadratic and high order polynomial terms and the inclusion of interaction terms between covariates are very hard to prescribe by strict rules.

It is clear that it is very difficult to get information about the real predictive value of a statistical model. Nevertheless this is very important, because the endless fine-tuning of statistical models by which every statistician is tempted, especially when a scientifically or clinically very interesting data set is involved, might result in an unexpectedly large amount of over-fitting which leads to a far too optimistic apparent error rate of the prediction rule.

As we have seen before, the optimism correction is of order $p/(n-p)$, where p is the number of covariates and n is the number of observations. If we take p as the number of covariates in the final model we might be far too optimistic, and perhaps it is more realistic to take p as the number of covariates at the start of the model building process. A way out of all these problems is the split-sample technique,[34,35] where the observations are divided randomly into two subsamples, labelled the training sample and the validation sample. The statistical model is fitted to the training sample, using everything that is good statistical practice, but being aware of the danger of over-fitting. When a satisfactory model is fitted to the training sample and a prediction rule is formulated by means of the prognostic index $X'\hat{\beta}$, the latter can be tested on the validation set. Unfortunately, the result is often disappointing since the error rate tends to be much larger than expected.

A simple way of checking the validity of the prediction rule is by regressing the outcome variable on the prognostic index in the validation sample. As we have seen in Section 7, a slight amount of shrinkage can be expected. The amount of shrinkage can even be estimated beforehand by the methods of Section 7. If there is much more shrinkage than expected, so that the observed $\hat{c}$ is much smaller that $\hat{c}_{\text{heur}}$ or $\hat{c}_{\text{cal}}$ as estimated from the training sample, then the observed $\hat{c}$ can be used to adjust the predictor. This process must be handled with care. There is a great temptation to modify the prediction rule to obtain a better fit to the validation sample, for example by adding or deleting covariates. That procedure invalidates the validation sample and endangers the whole operation. Repeated use of the validation sample to adjust the predictor might result in the same kind of optimism as using the whole sample as the training sample. Ideally, the data set should be split into three parts: one to select covariates, a second to estimate the regression coefficients and a third to assess the prediction rule. In this section we restrict attention to the two-part split sample. First we make some mathematical observations about regression on the prognostic index in the validation sample, and then we discuss two examples of practical application.

Let $(Y_1, X_1), \ldots, (Y_n, X_n), (Y_{n+1}, X_{n+1}), \ldots, (Y_{n+m}, X_{n+m})$ represent the data set, where the first n observations constitute the training sample and the last m observations form the validation sample. We will only discuss the OLS case in detail, but results carry over to generalized linear models. (For theoretical background see Cox and Hinkley.[36])

The estimate of the regression parameter in the training sample is given by

$$\hat{\beta}_{\text{T}} = S_{\text{T}}^{-1} \sum Y_i (X_i - \bar{X}_{\text{T}}), \tag{74}$$

where $S_{\text{T}} = \sum_{i=1}^{n} (X_i - \bar{X}_{\text{T}})(X_i - \bar{X}_{\text{T}})'$ and $\bar{X}_{\text{T}} = \sum_{i=1}^{n} X_i / n$. Observe that the regression parameter corresponding with the constant term is not involved here. As in ridge regression, we prefer to treat the constant term separately.

The coefficient for the regression of observations Y on predictions $(X - \bar{X}_T)'\hat{\beta}_T$ in the validation sample is given by

$$\hat{c} = \frac{\hat{\beta}_T' S_V \hat{\beta}_T}{\hat{\beta}_T' S_V \hat{\beta}_T}, \tag{75}$$

where $S_V = \sum_{i=n+1}^{n+m} (X_i - \bar{X}_V)(X_i - \bar{X}_V)'$, analogously to S_T. A heuristic 'estimate' of $\hat{c}$ based on the training sample is given by

$$\hat{c}_{heur} = \frac{\hat{\beta}_T' S_T \hat{\beta}_T - p s_T^2}{\hat{\beta}_T' S_T \hat{\beta}_T}, \tag{76}$$

where $s_T^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2/(n - p - 1)$ and p is the number of covariates.

Some justifications for $\hat{c}_{heur}$ are that:

(a) $E(\hat{c}|\hat{\beta}_T) = \beta_0' S_V \hat{\beta}_T / \beta_T' S_V \hat{\beta}_T$, where β_0 is the true value of the parameter;
(b) $S_V/(m-1) \approx S_T/(n-1)$, since both estimate the covariance matrix of the covariates;
(c) $E\{\hat{\beta}_T' S_T \hat{\beta}_T - \beta_0' S_T \hat{\beta}_T\} = p\sigma^2$, since $\text{cov}(\hat{\beta}_T) = \sigma^2 S_T^{-1}$; this yields $\hat{\beta}_T' S_V \hat{\beta}_T - p s^2$ as an estimate of $\beta_0' S_V \hat{\beta}_T$.

Observe that $\hat{c}_{heur}$ as given by (76) coincides with $\hat{c}_{heur}$ as given by (70). An informal goodness-of-fit test can be obtained by comparing $\hat{c}$ with $\hat{c}_{heur}$. If $\hat{c}$ is markedly smaller than $\hat{c}_{heur}$, there is an indication of lack of fit. The argument is that $E(\hat{\beta} - \beta_0)' S_T (\hat{\beta} - \beta_0)$ tends to increase if the linear model does not fit either because there is some bias in $\hat{\beta}$ or because $\text{cov}(\hat{\beta}) > \sigma^2 S_T^{-1}$. In generalized linear models the equivalent formula is

$$\hat{c}_{heur} = \frac{\text{model } \chi^2 - p}{\text{model } \chi^2}, \tag{77}$$

where model $\chi^2 = -2\log(L(\hat{\beta}_T)) + 2\log(L(0))$ in the training sample. (This allows for a constant term in the 'null model'.)

Graft survival

Our data set consists of graft survival for 6620 kidneys transplanted in the period 1984–1987. The whole sample of size 6620 was randomly split into a training sample of size 4253 and a validation set of size 2367. A prognostic index was constructed from the training sample using 16 covariates including HLA-DR and HLA-B match between recipient and donor, age of recipient, age of donor, blood group, sex and transplantation centre.

Using the SAS-PHGLM procedure for Cox regression analysis, a model χ^2 of 274·18 was obtained; so the heuristic estimate of the shrinkage coefficient $\hat{c}_{heur}$ is 0·942. Cox regression of graft survival on the prognostic index gave $\hat{c} = 0{\cdot}64$. An explanation of this marked difference is that the model building process started with many more (75) covariates, corresponding to many different transplantation centres (50) and a division of recipient and donor age into a number of age groups.

To visualize the shrinkage effect, the prognostic index was used to divide the training sample into three subgroups: low risk (2541), medium risk (859) and high risk (853). Kaplan–Meier survival curves were obtained for each group. Next, the patients in the validation sample were classified into corresponding groups using the same cut points for the prognostic index, and Kaplan–Meier curves were obtained for these groups. The six estimated survival are shown in Figure 7.

The figure clearly shows the shrinkage effect. The survival curves of the validation sample are markedly closer to each other than in the training set. An explanation in this example might be

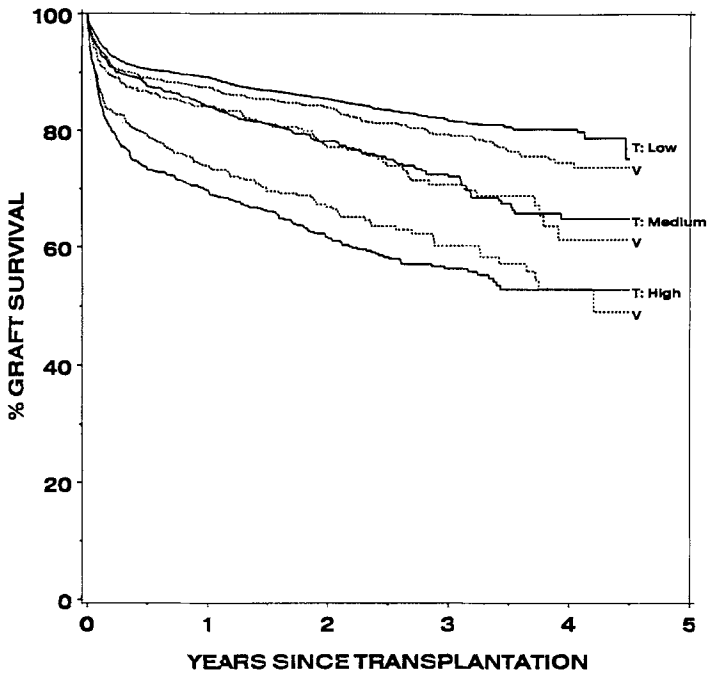

Figure 7. Observed kidney graft survival in three risk groups in the training sample (T) and the validation sample (V)

the large number of different centres. Presumably better results could have been obtained if the centre effects had been estimated by empirical Bayes methodology as in Gilks.[37]

More results on a prognostic index will be given in a forthcoming paper by Thorogood *et al.*[38]

Ovarian cancer

In a recent paper by Van Houwelingen *et al.*[1] a prognostic index was presented for the survival of 268 patients with advanced ovarian cancer. After minor modification this prognostic index was applied to a similar Danish data set of 301 patients. Cox regression on the Dutch data set gave a model with $\chi^2 = 67{\cdot}0, p = 7$ leading to $\hat{c}_{heur} = 0{\cdot}90$. Cox regression on this prognostic index in the Danish data set gave $\hat{c} = 0{\cdot}83$. The conclusion is that the Dutch index did well on the Danish data.

On the other hand, the Danish researchers developed a simpler index with $p = 6$ (fewer covariates but more categories) for which they obtained model $\chi^2 = 103{\cdot}3$ and $\hat{c}_{heur} = 0{\cdot}94$. Using it on the Dutch data gave $\hat{c} = 0{\cdot}62$. This indicates that the Danish prognostic index did not perform as well on the Dutch data as the Dutch data performed on the Danish data. An explanation might be that too few covariates were included in the Danish prognostic index.

More about this research will be published in a forthcoming paper by Lund *et al.*[39]

9. DISCUSSION

The first issue raised in this paper is that the apparent error rate is too optimistic in assessing the predictive power of a statistical model. A better assessment can be obtained by using optimism corrections. In ordinary least squares Mallows's C_P ((38), (39)) can be used, and in logistic

regression Akaike's information criterion ((49), (52)) in combination with the MML error rate. In Cox regression there is no simple correction. Instead of using model-based corrections, cross-validation can serve as a general tool for the assessment of predictive value. The formulae are given in (44), (53) and (61) for ordinary least squares, logistic regression and Cox regression respectively.

A second issue is the construction of improved predictors based on shrinkage. In Section 7 it is shown how the results of cross-validation can be used to calibrate the model. Fitting a model with the cross-validation 'prediction' $X_i'\hat{\beta}_{(-i)}$ as a single covariate yields a shrinkage factor $\hat{c}_{\text{cal}}$. The adjusted predictor is obtained by using an adjusted covariate vector $X^*_{\text{new}} = \bar{X} + \hat{c}_{\text{cal}}(X_{\text{new}} - \bar{X})$ in the full data model.

The last issue is the effect of statistical model building. The amount of covariate selection undoubtedly influences the predictive value of the final model. Cross-validation in the final model does not take this into account. Therefore it is advised in Section 8 to use the split-sample approach whenever extensive statistical model building is contemplated. The sample is randomly divided into a training sample and a validation sample. A (linear) prognostic index can be derived from the training sample. By regression on this prognostic index in the validation sample, the predictive value of the model can be assessed and a shrinkage factor can be obtained which can be used to obtain an improved predictor in the manner of Section 7.

REFERENCES

1. Van Houwelingen, J. C., Ten Bokkel Huinink, W. W., Van der Burg, M. E. L., Van Oosterom, A. T. and Neijt, J. P. 'Predictability of the survival of patients with advanced ovarian cancer', *Journal of Clinical Oncology*, **7**, 769–773 (1989).
2. Panel on Discriminant Analysis, Classification and Clustering. 'Discriminant analysis and clustering', *Statistical Science*, **4**, 34–69 (1989).
3. Efron, B. 'Estimating the error rate of a prediction rule: improvements on cross-validation', *Journal of the American Statistical Association*, **78**, 316–331 (1983).
4. Efron, B. 'How biased is the apparent error rate of a prediction rule?', *Journal of the American Statistical Association*, **81**, 461–470 (1986).
5. Mallows, C. L. 'Some comments on C_p', *Technometrics*, **15**, 661–675 (1973).
6. Stone, M. 'Cross-validatory choice and assessment of statistical predictions', *Journal of the Royal Statistical Society, Series B*, **36**, 111–147 (1974).
7. Allen, D. M. 'The relation between variable selection and data augmentation and a method for prediction', *Technometrics*, **16**, 125–127 (1974).
8. Lachenbruch, P. and Mickey, M. 'Estimation of error rates in discriminant analysis', *Technometrics*, **10**, 1–11 (1968).
9. Stone, M. 'An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion', *Journal of the Royal Statistical Society, Series B*, **39**, 44–47 (1977).
10. Atkinson, A. C. 'A note on the generalized information criterion for choice of a model', *Biometrika*, **67**, 413–418 (1980).
11. Breiman, L. and Freedman, D. 'How many variables should be entered in a regression equation?', *Journal of the American Statistical Association*, **78**, 131–136 (1983).
12. Hocking, R. R. 'The analysis and selection of variables in linear regression', *Biometrics*, 1–49 (1976).
13. Thompson, M. L. 'Selection of variables in multiple regression, Part I. A review and evaluation', *International Statistical Review*, **46**, 1–49 (1978).
14. Thompson, M. L. 'Selection of variables in multiple regression, Part II. Chosen procedures, computations and examples', *International Statistical Review*, **46**, 129–146 (1978).
15. Cook, D. R. and Weisberg, S. *Residuals and Influence in Regression*, Chapman and Hall, London, New York, 1982.
16. Geisser, S. 'The predictive sample reuse method with applications', *Journal of the American Statistical Association*, **70**, 320–328 (1975).

17. Golub, G., Heath, M. and Wahba, G. 'Generalized cross-validation as a method for choosing a good ridge parameter', *Technometrics*, **21**, 215–223 (1979).
18. Allen, D. M. 'Mean square error of prediction as a criterion for selecting variables', *Technometrics*, **13**, 469–475 (1971).
19. Shibata, R. 'An optimal selection of regression variables', *Biometrika*, **68**, 45–54 (1981).
20. Bunke, O. and Droge, B. 'Bootstrap and cross-validation estimates of the prediction error for linear regression models', *Annals of Statistics*, **12**, 1400–1424 (1984).
21. Van Houwelingen, J. C. and Le Cessie, S. 'Logistic regression, a review', *Statistica Neerlandica*, **42**, 215–232 (1988).
22. Verloove, S. P. and Verwey, R. Y. 'Project on pre-term and small-for-gestational-age infants in the Netherlands', thesis, 1983; Ann Arbor, UMI, 8807276, 1988.
23. Hosmer, D. W. and Lemeshow, S. 'Goodness-of-fit tests for the multiple logistic regression model', *Communications in Statistics – Theory and Methods*, **9**, 1043–1069 (1980).
24. Le Cessie, S. and Van Houwelingen, J. C. 'A goodness of fit test for binary regression models, based on smoothing methods', Technical Report 7, Department of Medical Statistics, Leiden University, 1989 (*Biometrics* in press).
25. Copas, J. B. 'Regression, prediction and shrinkage' (with discussion), *Journal of the Royal Statistical Society, Series B*, **45**, 311–354 (1983).
26. Jones, M. C. and Copas, J. B. 'On the robustness of shrinkage predictors in regression to differences between past and future data', *Journal of the Royal Statistical Society, Series B*, **48**, 223–237 (1986).
27. Hoerl, A. E. and Kennard, R. W. 'Ridge regression. Biased estimation for nonorthogonal problems', *Technometrics*, **12**, 55–67 (1970).
28. Hoerl, A. E. and Kennard, R. W. 'Ridge regression. Applications to nonorthogonal problems', *Technometrics*, **12**, 69–82 (1970).
29. Draper, N. R. and Smith, H. *Applied Regression Analysis*, 2nd edn, Wiley, New York, 1981.
30. Rao, C. R. 'Prediction of future observations in growth curve models' (with discussion), *Statistical Science*, **2**, 434–471 (1987).
31. Pregibon, D. 'Logistic regression diagnostics', *Annals of Statistics*, **9**, 705–724 (1981).
32. Storer, B. E. and Crowley, J. 'A diagnostic for Cox regression and general conditional likelihoods', *Journal of the American Statistical Association*, **80**, 139–147 (1985).
33. Gong, G. 'Cross-validation, the jackknife, and the bootstrap: excess error rate estimation in forward logistic regression', *Journal of the American Statistical Association*, **81**, 108–113 (1986).
34. Mosteller, F. and Tukey, J. W. *Data Analysis and Linear Regression*, Addison-Wesley, Reading, Mass., 1977.
35. McCarthy, P. J. 'The use of balanced half-sample replication in cross-validation studies', *Journal of the American Statistical Association*, **44**, 596–604 (1976).
36. Cox, D. R. and Hinkley, D. W. *Theoretical Statistics*, Chapman and Hall, London, 1974.
37. Gilks, W. R. 'Some applications of hierarchical models in kidney transplantation', *The Statistician*, **36**, 127–136 (1987).
38. Thorogood, J., Van Houwelingen, J. C., Persijn, G. G., Zantroort, F. A., Schreüder, G. M. Th. and Van Rood, J. J. 'Prognostic indices for prediction of survival of first and second kidney grafts', (submitted for publication).
39. Lund, B., Williamson, P., Van Houwelingen, H. C. and Neijt, J. P. 'A comparison of the predictive power of different prognostic indices for overall survival in patients with advanced ovarian carcinoma', (*Cancer Research* in press).