*Biometrika* (2007), **94**, 2, *pp*. 297–311
© 2007 Biometrika Trust
*Printed in Great Britain*

doi:10.1093/biomet/asm036
Advance Access publication 14 May 2007

# Model evaluation based on the sampling distribution of estimated absolute prediction error

BY LU TIAN

*Department of Preventive Medicine, Northwestern University Medical School, 680 N. Lake Shore Drive, Chicago, Illinois 60611, U.S.A.*
lutian@northwestern.edu

TIANXI CAI

*Department of Biostatistics, Harvard University, 655 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.*
tcai@hsph.harvard.edu

ELS GOETGHEBEUR

*Department of Applied Mathematics and Computer Science, Ghent University, Krijgsloan 281-S9 9000, Ghent, Belgium*
els.goetghebeur@ugent.be

AND L. J. WEI

*Department of Biostatistics, Harvard University, 655 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.*
wei@hsph.harvard.edu

## SUMMARY

The construction of a reliable, practically useful prediction rule for future responses is heavily dependent on the ‘adequacy’ of the fitted regression model. In this article, we consider the absolute prediction error, the expected value of the absolute difference between the future and predicted responses, as the model evaluation criterion. This prediction error is easier to interpret than the average squared error and is equivalent to the misclassification error for a binary outcome. We show that the prediction error can be consistently estimated via the resubstitution and crossvalidation methods even when the fitted model is not correctly specified. Furthermore, we show that the resulting estimators are asymptotically normal. When the prediction rule is ‘nonsmooth’, the variance of the above normal distribution can be estimated well with a perturbation-resampling method. With two real examples and an extensive simulation study, we demonstrate that the interval estimates obtained from the above normal approximation for the prediction errors provide much more information about model adequacy than their point-estimate counterparts.

*Some key words*: 0.632 bootstrap; Bootstrap; $K$-fold crossvalidation; Model and variable selection; Perturbation-resampling; Prediction.

## 1. Introduction

Often aggregate prediction errors, which measure the 'distance' between the future and predicted outcomes, are used to evaluate the adequacy of a fitted model or compare competing models (Davison & Hinkley, 1997, §6.4). Methods for estimating prediction errors are mainly based on the apparent or resubstitution error, crossvalidation, bootstrap and covariance penalties (Akaike, 1973; Mallows, 1973; Stein, 1981; Efron, 1983, 1986, 2004; Breiman, 1992; Shao, 1993, 1996; Efron & Tibshirani, 1997; Ye, 1998; Tibshirani & Knight, 1999). Recent research in this area was mostly devoted to reducing bias of the apparent error when the sample size is not large with respect to the number of unknown parameters in the fitted model (Molinaro et al., 2005).

For the case with a continuous response variable, generally the prediction error considered in the literature is the average squared error. This choice is driven by mathematical convenience rather than physical relevance. Moreover, little effort has been made to study the distributional properties of the estimated prediction error.

In this article, we assume that the sample size is relatively large with respect to the dimension of the vector of regression parameters. Furthermore, instead of using the $L_2$ norm, we consider the average absolute prediction error, the expected value of the absolute difference between the future and predicted responses. For binary response, this prediction error is the misclassification error.

## 2. Approximations to the distribution of estimated prediction error

Let $Y$ be a continuous or binary response variable and let $X$ be the vector of its predictors. Let $Z$, a $p \times 1$ bounded vector, be a function of $X$. Also, let $\{(Y_i, Z_i), i = 1, \ldots, n\}$ be $n$ independent copies of $(Y, Z)$. Denote by $Z^0$ the $Z$-value of a future observation, drawn independently from the population generating $(Y, Z)$. Let a prediction for its response $Y^0$ be derived from a regression model assuming that the conditional mean $E(Y|Z)$ has a parametric form $g(\beta' Z)$, where $g(\cdot)$ is a known, strictly increasing, differentiable function and $\beta$ is the vector of unknown parameters. Let $\hat{\beta}$ be an estimator of $\beta$ based on the entire dataset $\{(Y_i, Z_i), i = 1, \ldots, n\}$ and let $\hat{Y}(\hat{\beta}' Z^0)$ be the predicted value for $Y^0$. For instance, if $Y$ is a continuous variable, one may let $\hat{Y}(\hat{\beta}' Z^0) = g(\hat{\beta}' Z^0)$. If $Y$ is a binary variable, one may let $\hat{Y}(\hat{\beta}' Z^0) = I\{g(\hat{\beta}' Z^0) \geqslant 0{\cdot}5\}$, a commonly used binary prediction rule, where $I(\cdot)$ is the indicator function.

To evaluate how well the fitted model predicts this future response $Y^0$, we consider the absolute prediction error $D_0$ or a function thereof, where

$$D_0 = E|Y^0 - \hat{Y}(\hat{\beta}' Z^0)| \qquad (1)$$

and the expectation $E$ is with respect to $\{(Y_i, Z_i), i = 1, \ldots, n\}$ and $(Y^0, Z^0)$. Note that $D_0$ depends on the sample size $n$. To estimate $D_0$, we first consider the so-called 'apparent or resubstitution error' $\hat{D}(\hat{\beta})$, where

$$\hat{D}(\beta) = n^{-1} \sum_{i=1}^{n} |Y_i - \hat{Y}(\beta' Z_i)| \qquad (2)$$

(Davison & Hinkley, 1997, §6.4).

To approximate the large sample distribution of $\{\hat{D}(\hat{\beta}) - D_0\}$, we need to show that $\hat{\beta}$ is stabilized as $n$ increases when the fitted model may not be correctly specified; that is, $\hat{\beta}$ converges to a constant vector in probability, as $n \to \infty$. If we use a parametric likelihood score function $S^{\dagger}(\beta)$ to estimate $\beta$, under the strong assumption that the equation $E\{S^{\dagger}(\beta)\} = 0$ has a unique root, generally $\hat{\beta}$ converges to this root in probability (White, 1982). Unfortunately, it is rather difficult to verify the above uniqueness condition even when the estimator $\hat{\beta}$ exists and is unique for any finite sample size $n$ under the fitted model (Silvapulle, 1981; Jacobsen, 1989).

We propose to estimate $\beta$ via the simple estimating function

$$S(\beta) = n^{-1} \sum_{i=1}^{n} Z_i\{Y_i - g(\beta' Z_i)\}. \tag{3}$$

First, we assume that, if $J$ is the support of $Y$, $J \subseteq [g(-\infty), g(+\infty)]$, $E(Y) < \infty$, and both the matrix $n^{-1} \sum_{i=1}^{n} Z_i Z_i'$ and its limit are positive definite. Furthermore, when $Y$ is a binary outcome, we assume an additional condition that one cannot find a vector $b$ such that $I(Y_1 > Y_2) = I(b' Z_1 > b' Z_2)$ almost surely. Note that the above mild conditions are needed for consistency of $\hat{\beta}$ even when $g(\beta' Z)$ is the correct form of the true conditional mean of $Y$ given $Z$. In Appendix 1, without assuming that $g(\beta' Z)$ is the correct form of the conditional mean of $Y$ given $Z$, we show that there is a unique root $\hat{\beta}$ to $S(\beta) = 0$, almost surely, and also a unique root $\beta_0$ to $E\{S(\beta)\} = 0$. We then show that $\hat{\beta}$ converges to $\beta_0$ in probability, as $n \to \infty$. When there exists a $\beta^{\dagger}$ such that $E(Y|Z) = g(Z'\beta^{\dagger})$, it follows that $\beta^{\dagger} = \beta_0$, because $\beta_0$ is the unique solution to the equation $E[Z\{Y - g(\beta' Z)\}] = 0$.

We now assume that the conditional density or probability mass function of $Y$ given $Z$ is continuously differentiable. In Appendix 2, we show that $\hat{D}(\hat{\beta})$ is a good estimator for $D_0$ in the sense that $\{\hat{D}(\hat{\beta}) - D_0\}$ converges to zero in probability, as $n \to \infty$. To draw inferences about $D_0$, one needs a good approximation to the distribution of $\hat{D}(\hat{\beta})$. Although $\hat{D}(\beta)$ is not differentiable with respect to $\beta$, we show in Appendix 2 that the distribution of

$$W = n^{1/2}\{\hat{D}(\hat{\beta}) - D_0\} \tag{4}$$

is asymptotically Gaussian with mean 0.

If $\hat{Y}(\hat{\beta}' Z^0) = g(\hat{\beta}' Z^0)$, then $D_0$ in (1) becomes $E|Y^0 - g(\hat{\beta}' Z^0)|$. For this case, the variance of $W$ in (4) can be consistently estimated by

$$n^{-1} \sum_{i=1}^{n} \hat{\eta}_i^2, \tag{5}$$

where

$$\hat{\eta}_i = |Y_i - g(\hat{\beta}' Z_i)| - \hat{D}(\hat{\beta}) + d(\hat{\beta}) A^{-1}(\hat{\beta}) Z_i\{Y_i - g(\hat{\beta}' Z_i)\},$$

$$A(\beta) = n^{-1} \sum_{i=1}^{n} \dot{g}(\beta' Z_i) Z_i Z_i', \tag{6}$$

$\dot{g}(x) = dg(x)/dx$, and

$$d(\beta) = -n^{-1} \sum_{i=1}^{n} \text{sign}\{Y_i - g(\beta' Z_i)\} \dot{g}(\beta' Z_i) Z_i', \tag{7}$$

the quasi-derivative of $\hat{D}(\beta)$. The justification of consistency of (5) is given in Appendix 2.

If $\hat{Y}(\hat{\beta}' Z^0)$ is not $g(\hat{\beta}' Z^0)$, for example when $Y$ is binary and $\hat{Y}(\hat{\beta}' Z^0) = I\{g(\hat{\beta}' Z^0) \geqslant c\}$, where $c$ is a pre-specified constant, the variance of $W$ may involve the unknown conditional density or probability mass function of $Y$ given $Z$, efficient estimation of which is difficult nonparametrically, especially when the dimension of $Z$ is large. In general, one may use a perturbation-resampling technique to obtain an approximation to the distribution of $W$. To be specific, let $y$ and $z$ be the observed values of $Y$ and $Z$, and let $G_i, i = 1, \ldots, n$, be independent and identically distributed random variables with a known distribution whose mean and variance are one. Furthermore, let $\beta^*$ be the solution to the equation

$$S^*(\beta) = n^{-1} \sum_{i=1}^{n} z_i\{y_i - g(\beta' z_i)\}G_i = 0. \tag{8}$$

Note that the only random quantities in $S^*(\beta)$ are the $G_i$'s. Next, let $\tilde{D}(\beta)$ and $\tilde{\beta}$ be the observed $\hat{D}(\beta)$ and $\hat{\beta}$, respectively, and let

$$W^* = n^{-1/2} \sum_{i=1}^{n} \{|y_i - \hat{Y}(z_i'\beta^*)| - \tilde{D}(\tilde{\beta})\}(G_i - 1). \tag{9}$$

It is straightforward to show that, for large $n$, the unconditional distribution of $W$ in (4) can be approximated well by the conditional distribution of $W^*$ given the data. This perturbation-resampling technique has been used successfully, for example by Park & Wei (2003) and Cai et al. (2005).

Note that the distribution of $W^*$ can be easily approximated via a large number, $M$ say, of realizations from $\{G_i, i = 1, \ldots, n\}$. For each realized sample, we compute the corresponding realized $W^*$. The distribution of $W$ can then be approximated based on these $M$ independent realizations of $W^*$, and interval estimates for $D_0$ can be constructed accordingly. The length of such an interval, coupled with the observed point estimate $\tilde{D}(\tilde{\beta})$ and the scale of the response variable $Y$, provides an easily interpretable metric for assessing the adequacy of the prediction rule based on the fitted model.

It is interesting to note that, if $(G_1, \ldots, G_n)$ is a multinomial random vector with size $n$ and marginal cell probabilities of $n^{-1}$, the resulting $W^*$ obtained by replacing $G_i - 1$ in (9) by $G_i$ is essentially the bootstrap counterpart of $W$. It is not clear, however, how to justify analytically whether or not the bootstrapping provides a good approximation to the distribution of $W$ under the current setting.

For a small or moderate sample size $n$ with respect to the dimension $p$ of $\beta$, $\hat{D}(\hat{\beta})$ may underestimate $D_0$. One way of reducing such bias is to use crossvalidation procedures to estimate $D_0$. To this end, first consider the popular $K$-fold crossvalidation method. We randomly split the data into $K$ disjoint subsets of about equal sizes and label them as $\mathcal{I}_k, k = 1, \ldots, K$. For each $k$, we use all observations which are not in $\mathcal{I}_k$ to obtain an estimate $\hat{\beta}_{(-k)}$ for $\beta$ via (3), and then compute the predicted error estimator $\hat{D}_{(k)}\{\hat{\beta}_{(-k)}\}$ via (2) based on observations in $\mathcal{I}_k$. Then, an average prediction error estimator for $D_0$ is

$$\hat{\mathcal{D}} = K^{-1} \sum_{k=1}^{K} \hat{D}_{(k)}\{\hat{\beta}_{(-k)}\}. \tag{10}$$

When $K$ is fixed and relatively small with respect to $n$, for each $k = 1, \ldots, K$, the sizes of training and validation sets are of order $n$ and $\{\hat{\mathcal{D}} - D_0\}$ converges to 0 in probability. Furthermore, each $\hat{D}_{(k)}(\beta)$ is locally linear around $\beta_0$. It follows from the multivariate central limit theorem that

$$\mathcal{W} = n^{1/2}(\hat{\mathcal{D}} - D_0) \tag{11}$$

is asymptotically normal. In Appendix 3, we show that the limiting distribution of $\mathcal{W}$ is the same as that of $W$. Therefore, the point estimates $\hat{D}(\hat{\beta})$ and $\hat{\mathcal{D}}$ may be different, but, for large $n$, a confidence interval for the absolute predicted error based on the $K$-fold crossvalidation method has approximately the same length as that based on the apparent error.

For a more general random crossvalidation scheme, let $n_t$ and $n_v$ be the 'average' sizes of the training and validation sets, where $n_t/n$ and $n_v/n$ converge to nonzero constants, as $n \to \infty$. The assignments of the datapoints to the training set are determined by independent tossings of a coin each time with a 'success probability' of $n_t/n$. We then use all observations in the training set to obtain an estimate $\hat{\beta}_t$ for $\beta$ via (3) and compute the corresponding $\hat{D}(\hat{\beta}_t)$ in (2) based on the validation set. We repeat this process by taking a fresh random training set at each stage. Let $\hat{\mathbb{D}}$ be the average $\hat{D}(\hat{\beta}_t)$ defined as in (10), but where the summation is over the entire set of possible training-validation splits. In Appendix 4, we show that $n^{1/2}(\hat{\mathbb{D}} - D_0)$ has the same limiting distribution as that of $W$ in (4). In practice, one may generate a large number of random splits to approximate $\hat{\mathbb{D}}$. Note that the conventional leave-one-out method does not belong to the above class of crossvalidation procedures.

An interesting hybrid of crossvalidation and apparent error, the 0·632 bootstrap estimator, for estimating the prediction error was proposed by Efron & Tibshirani (1997). This estimator is essentially a linear combination of the apparent error and a crossvalidation counterpart. If the crossvalidation component belongs to the class discussed in the last paragraph, this combination has the same large sample distribution as $W$. However, since Efron and Tibshirani's estimator uses a smooth version of leave-one-out crossvalidation, it is not clear how to justify its large sample approximation.

## 3. Comparing models based on estimated prediction errors

Suppose that, for a fixed vector $X$ of predictors, there are two competing regression models, $g_j(\hat{\beta}_j' Z_{(j)})$, $j = 1, 2$, say, where the $p_j$-dimensional vector $Z_{(j)}$ is a function of $X$ and $\hat{\beta}_j$ is the estimator from (3) with the data $\{(Y_i, Z_{(j)i}), i = 1, \ldots, n\}$. The theoretical and empirical prediction errors $D_{0j}$ and $\hat{D}_j(\beta_j)$ are defined by (1) and (2) accordingly, for $j = 1, 2$. We are interested in drawing inferences about, for example, $\Delta = D_{02} - D_{01}$ to assess how much improvement Model 1 offers over Model 2.

A consistent estimator for $\Delta$ is $\hat{\Delta} = \hat{D}_2(\hat{\beta}_2) - \hat{D}_1(\hat{\beta}_1)$. It follows from the arguments in §2 that

$$W_\Delta = n^{1/2}(\hat{\Delta} - \Delta) \tag{12}$$

is asymptotically normal with mean 0. To approximate this normal distribution, one may use the analytical or perturbation method discussed in §2. For the resampling method, let $\beta_j^*$ be the solution of

$$S_j^*(\beta_j) = \sum_{i=1}^{n} z_{(j)i}\{y_i - g_j(\beta_j' z_{(j)i})\} G_i = 0,$$

$j = 1, 2$. Also, let

$$W_j^* = n^{-1/2} \sum_{i=1}^{n} \{|y_i - \hat{Y}(z_{(j)i}' \beta_j^*)| - \tilde{D}_j(\tilde{\beta}_j)\}(G_i - 1),$$

where $\tilde{D}_j$ and $\tilde{\beta}_j$ are the observed values of $\hat{D}_j$ and $\hat{\beta}_j$, respectively. Then the distribution of $W_\Delta$ can be approximated by the conditional distribution of $W_\Delta^* = W_2^* - W_1^*$. Confidence intervals for $\Delta$ based on this normal approximation can then be constructed.

For the $K$-fold crossvalidation method, the estimator $\hat{\mathcal{D}}_2 - \hat{\mathcal{D}}_1$, where $\hat{\mathcal{D}}_j$ is defined by (10) for Model $j$, $j = 1, 2$, may be less biased than $\hat{\Delta}$ for small sample cases. On the other hand, let $\mathcal{W}_j$ be defined by (11) based on Model $j$. The limiting distribution of $(\mathcal{W}_2 - \mathcal{W}_1)$ is the same as that of $W_\Delta$. Similarly, for general crossvalidation or its hybrids as discussed in §2, the distribution of the corresponding $n^{1/2}(\hat{\mathbb{D}}_2 - \hat{\mathbb{D}}_1 - \Delta)$ can also be approximated by that of $W_\Delta$.

## 4. Examples

We use two examples to illustrate the proposed procedures, one with a continuous response and one with a binary outcome.

The data of the first example are from the clinical trial, ACTG 320, conducted by the AIDS Clinical Trials Group to evaluate the benefit of using a three-drug combination therapy, which includes indinarvir, zidovudine and lamivudine, for treating HIV-infected patients (Hammer et al., 1997). There were 583 patients randomly assigned to this treatment group. Even with the relatively potent therapy, a significant proportion of patients will not respond to treatment. It is therefore important to identify early biomarkers, which can predict treatment failure effectively, for future patients' care and management.

In this example, we let the response variable $Y$ be the change of CD4 cell counts from Week 0 to week 24, an important measure of the patient's immune response. This variable is still one of the major endpoints for modern clinical studies of HIV diseases, especially those conducted in resource-limited countries. Based on ACTG 320, the potential early predictors $X$ for such $Y$ are age, baseline HIV-1 RNA, baseline CD4, and the changes of RNA and CD4 from Week 0 to Week 8. Since it is relatively expensive to obtain the RNA measure, an important and interesting question is whether or not early RNA observations are needed to make a 'meaningfully better' prediction of the change of CD4 counts from baseline to Week 24 for a patient treated by this combination drug. For our analysis, 392 patients in ACTG 320 had baseline and Week 8 RNA and CD4 measures. The observed $Y$ from these patients ranges from $-100$ to 734. To evaluate the added value from early RNA marker values, we first assume that the conditional mean of $Y$ given $Z$ has a parametric form of $g(\beta' Z) = \beta' Z$, where $Z = (1, X')'$, a $6 \times 1$ vector. Based on the estimating function (3), the point estimate of each component of $\hat{\beta}$ with its estimated standard error and corresponding $p$-value for testing zero covariate effect are given in Table 1. Note that early changes of RNA values and CD4 counts are highly statistically significant. For this model, the apparent error $\hat{D}(\hat{\beta})$ is 51 with an estimated standard error of 2·7 based on (5). The 0·95 confidence interval of $D_0$ is (46, 56), a rather tight interval from a clinical point of view.

Table 1. *AIDS example. Estimates of the regression parameters with their standard errors and corresponding p-values for testing zero covariate effects*

| | Age | Baseline RNA | RNA change | Baseline CD4 | CD4 change |
|---|---|---|---|---|---|
| Estimate | −0·55 | 0·08 | −12·06 | 0·02 | 0·68 |
| Std error | 0·35 | 5·53 | 2·80 | 0·07 | 0·10 |
| $p$-value | 0·12 | 0·99 | 0·00 | 0·72 | 0·00 |

Next, consider another linear additive model whose $Z$ does not include the baseline or early change of RNA. For this case, $\hat{D}(\hat{\beta}) = 52$ with the estimated standard error of 2·7. The corresponding confidence interval for the prediction error is (47, 57), which is practically identical to the previous interval estimate. Moreover, the 0·95 confidence interval for the difference between the prediction error for the full model and the one without RNA values via $W_{\Delta}$ in (12) is

$$(-2{\cdot}0, 0{\cdot}4). \tag{13}$$

Since this interval estimate is quite tight around 0, it suggests that there is no clinically meaningful improvement from a model which contains RNA information over the model without RNA values.

With the 10-fold crossvalidation method, the less-biased point estimates for predicted errors are 52 and 53 for models with and without RNA measures, respectively. The corresponding 0·95 confidence intervals are (47, 57) and (48, 58) based on the variance estimate (5) for the apparent error. Moreover, these intervals are almost identical to those estimated by standard nonparametric bootstrap methods via 500 bootstrap samples. For comparison, we also used a random crossvalidation as discussed in §2 with $n_t = 2n/3$ and the 0·632-bootstrap method proposed by Efron & Tibshirani (1997) to evaluate the above two fitted models. The estimates for the difference of the prediction errors of these models are $-0{\cdot}50$ and $-0{\cdot}63$, respectively. To construct interval estimates based on the 0·632-bootstrap method, we generated 500 by 500 double-bootstrap samples to estimate the variance. The resulting 95% confidence interval for the difference of the prediction errors for the aforementioned two models is $(-1{\cdot}8, 0{\cdot}6)$, which is practically identical to interval (13).

For the full model with predictors listed in Table 1, the point estimate of the predicted error is about 51, which is relatively large from a clinical point of view. This, coupled with very tight interval estimates, suggests that further research may be needed to identify potentially more important predictors in addition to early changes in CD4 count. However, it seems clear that early RNA measures do not add much value for predicting the patient's immune response.

The second example comes from a prostate cancer study, which examines whether or not certain 'baseline' bio- and clinical markers are helpful for predicting tumour penetration, a binary response variable (Hosmer & Lemeshow, 2000, Ch. 1). Potential predictors include age, race, digital rectal exam (no nodule, left nodule, right nodule or bilobar nodule), detection of capsular involvement in rectal exam (DCI), prostate-specific antigen (PSA), tumour volume obtained from ultrasound (TV) and total Gleason Score (GS). A total of 376 subjects with complete data are included in this analysis.

For a binary $Y$, the estimating function (3) is the likelihood score function from the standard logistic regression model. First we fitted the data with an additive logistic regression based on the above potential predictors. In Table 2, we present the point estimates of regression coefficients and their standard error estimates. Note that the coefficient for prostate-specific antigen is highly significant. With this model and the binary decision rule, $I\{g(\hat{\beta}'Z^0) \geqslant 0{\cdot}5\}$, the apparent error for estimating the misclassification rate is 0·24. With $M = 1000$ perturbation samples $\{G_i\}$ in (9), the 0·95 confidence interval for the error rate is (0·19, 0·29). The corresponding point estimate and 0·95 interval based on 10-fold crossvalidation are 0·27 and (0·21, 0·33), respectively. The estimates from random crossvalidation with $n_t = 2n/3$ are 0·26 and (0·20, 0·31). With 500 by 500 double bootstrap

Table 2. *Prostate cancer example. Estimates of the regression parameters with their standard errors and corresponding p-values for testing zero covariate effects*

| | | | | | | | Nodule in rectal exam | | |
|---|---|---|---|---|---|---|---|---|---|
| | Age | Race | DCI | PSA | TV | GS | Left | Right | Bilobar |
| Estimate | −0·01 | −0·65 | 0·49 | 0·03 | −0·01 | 0·96 | 0·73 | 1·51 | 1·39 |
| Std error | 0·02 | 0·47 | 0·45 | 0·01 | 0·01 | 0·17 | 0·34 | 0·37 | 0·47 |
| $p$-value | 0·56 | 0·17 | 0·27 | 0·00 | 0·13 | 0·00 | 0·03 | 0·00 | 0·00 |

samples, the 0·632-bootstrap method gives an estimate of 0·25 and a 0·95 confidence interval of (0·21, 0·30).

Since prostate-specific antigen is a routinely used, but controversial, biomarker for diagnosis of prostate cancer, it is interesting to examine how much accuracy the prostate-specific antigen would add for predicting tumour penetration. To this end, we fitted the data with another logistic model, which is identical to the first model but does not include PSA. With the apparent error, the estimate $\hat{\Delta}$ in (12) for the difference of prediction errors between the second and first models is 0·021 with 0·95 confidence interval (−0·020, 0·062). The 10-fold crossvalidation estimate is 0·018 with a 0·95 interval of (−0·023, 0·059), while the 0·632-bootstrap estimate is 0·017 and its 0·95 interval is (−0·012, 0·045). These indicate that PSA adds rather modest value, if any, on top of other variables, for predicting tumour penetration.

## 5. Evaluating finite-sample properties of the interval estimators

To examine finite-sample performance of the proposed inference procedures based on the apparent error and its crossvalidation counterparts, we conducted an extensive simulation study under various scenarios mimicking the above two examples. For the first part of the simulation study, we let the response $Y$ be the CD4 count change from Week 0 to 24 and let $X$ be the five predictors discussed in the aforementioned ACTG 320 study. We assume that $X$ is normal with mean and covariance matrix estimated from the observed data. Given $X$, the response $Y$ is generated from a linear model with covariate vector $(1, X')'$ and a zero-mean normal error. The model parameters, the regression coefficients and the variance of the error were chosen using the least squares estimates from the observed data, as shown in Table 1.

We generated 1000 independent sets of $\{(Y_i, X_i), i = 1, \ldots, 392\}$, and fitted four different working models to each realized dataset. These models are linear, additive models involving different vectors $Z$ of covariates. The covariate vector $Z$ of the first model is $(1, X')'$, which is the deterministic part of the true model. The second model has three covariates, namely age, baseline RNA and baseline CD4. The third model has age, CD4 count at baseline and early change in CD4 count as its covariates. The fourth model, an overfitted one, involves the above five predictors and all their squares and first-order interactions. For each model, we obtained the empirical absolute prediction errors via the apparent error, 10-fold crossvalidation and two random crossvalidations with $n_v = n/3$ and $n_v = n/10$. For random crossvalidation, we generated 200 random training and validation sets for each realized dataset. The variance estimators of the empirical prediction errors were obtained via 500 sets of perturbations.

Table 3. *Empirical bias and coverage probability for the apparent error* (AE)*, the* 10*-fold crossvalidation* (CV-10) *and random crossvalidations* ($\text{RCV}_{1/3}$ *and* $\text{RCV}_{1/10}$) *with* $n = 392$

| | Bias | | | | Coverage probability | | | |
|---|---|---|---|---|---|---|---|---|
| Model | AE | CV-10 | $\text{RCV}_{1/3}$ | $\text{RCV}_{1/10}$ | AE | CV-10 | $\text{RCV}_{1/3}$ | $\text{RCV}_{1/10}$ |
| I | −0·86 | 0·06 | 0·24 | 0·07 | 0·92 | 0·95 | 0·95 | 0·95 |
| II | −0·71 | 0·06 | 0·21 | 0·06 | 0·94 | 0·94 | 0·94 | 0·94 |
| III | −0·54 | 0·09 | 0·21 | 0·09 | 0·93 | 0·94 | 0·94 | 0·95 |
| IV | −3·17 | 0·29 | 1·02 | 0·28 | 0·66 | 0·93 | 0·91 | 0·93 |

Model I: $E(Y|X) =$ intercept + age + baseline RNA + RNA change + baseline CD4 + CD4 change; Model II: $E(Y|X) =$ intercept + age + baseline RNA + baseline CD4; Model III: $E(Y|X) =$ intercept + age + baseline CD4 + CD4 change; Model IV: $E(Y|X) =$ intercept + five predictors from Model I + all 2nd-order terms.

In Table 3, for each case we report the empirical bias and coverage probability of the confidence interval with nominal level of 0·95. Note that the true prediction error for each model is approximated with 5000 fresh, independent sets of $\{(Y^0, Z^0), (Y_i, Z_i), i = 1, \ldots, 392\}$. The coverage probability of the 0·95 interval estimate based on the apparent error is only 0·66 for Model IV. On the other hand, the crossvalidation point estimates and the corresponding interval estimates perform quite well. For comparison, we also implemented the 0·632-bootstrap method for each dataset. For all cases studied here, the bias and mean squared error of the 0·632-bootstrap estimator are very similar to those for 10-fold crossvalidation.

For the second part of the simulation study, we mimicked the prostate cancer example and generated the binary outcome $Y$ based on a logistic regression model with conditional probability $g(-7{\cdot}50 + 0{\cdot}96\text{GS} + 0{\cdot}73\text{DE}_l + 1{\cdot}51\text{DE}_r + 1{\cdot}39\text{DE}_b)$, where $g(s) = \{1 + \exp(-s)\}^{-1}$, GS is the Gleason score and $\text{DE}_l$, $\text{DE}_r$ and $\text{DE}_b$ are indicators for the nodule of the rectal exam being left, right and bilobar, respectively. The regression coefficients for the above model are obtained by fitting the observed data from the prostate cancer study with this specific logistic regression model via maximum likelihood. The vector $X$ of predictors was generated repeatedly by sampling with replacement from the original dataset. We generated 2000 sets of datasets $\{(Y^0, Z^0), (Y_i, X_i), i = 1, \ldots, n\}$, and fitted four different working models to each dataset . Model I is the true model. Model II has only PSA and $\text{PSA}^2$ as its covariates. Model III has the probit link and PSA and GS as its covariates. Model IV is a logistic model which contains Age, PSA, TV, GS, Rectal Exam, all quadratic terms of the continuous covariates and all the first-order interactions between the continuous covariates. Here, the binary prediction rule is $I\{g(\hat{\beta}'Z) \geqslant 0{\cdot}5\}$. We computed the empirical prediction errors for the apparent error, 10-fold crossvalidation and two random crossvalidations with $n_v = n/3$ and $n_v = n/10$. As in the first part of the simulation study, for random crossvalidation we generated 200 random training and validation sets for each realized dataset. The variance estimators of the empirical prediction errors were obtained via 500 sets of perturbations.

In Table 4, we report the empirical bias and the coverage probability of the 0·95 confidence interval for each model and estimation procedure with $n = 200$. It is interesting to note that, with respect to bias, the apparent error is only slightly worse than its crossvalidation counterparts. Moreover, all confidence interval estimates appear to have reliable coverage probabilities even for the extremely overfitted model.

Table 4. *Empirical bias and coverage probability for the apparent error* (AE)*, the* 10*-fold crossvalidation* (CV-10) *and random crossvalidations* ($\text{RCV}_{1/3}$ *and* $\text{RCV}_{1/10}$) *with* $n = 200$

| | Bias | | | | Coverage probability | | | |
|---|---|---|---|---|---|---|---|---|
| Model | AE | CV-10 | $\text{RCV}_{1/3}$ | $\text{RCV}_{1/10}$ | AE | CV-10 | $\text{RCV}_{1/3}$ | $\text{RCV}_{1/10}$ |
| I | −0·007 | 0·004 | 0·005 | 0·004 | 0·97 | 0·97 | 0·98 | 0·97 |
| II | −0·022 | 0·018 | 0·026 | 0·019 | 0·95 | 0·95 | 0·95 | 0·96 |
| III | −0·011 | 0·006 | 0·009 | 0·006 | 0·97 | 0·95 | 0·98 | 0·96 |
| IV | −0·009 | 0·005 | 0·006 | 0·005 | 0·97 | 0·95 | 0·97 | 0·96 |

Model I: logit$\{\text{pr}(Y = 1|X)\}$ = intercept + GS + three rectal exam indicators; Model II: logit$\{\text{pr}(Y = 1|X)\}$ = intercept + PSA + $\text{PSA}^2$; Model III: probit$\{\text{pr}(Y = 1|X)\}$ = intercept + PSA + GS; Model IV: logit$\{\text{pr}(Y = 1|X)\}$ = intercept + Age + PSA + TV + GS + rectal exam indicators + all 2nd-order terms for the continuous covariates.

## 6. Remarks

We have derived model evaluation procedures for continuous and binary responses for which the $L_1$ prediction error is a meaningful, physically interpretable metric. Based on the results of our numerical studies, we recommend the use of interval estimates centred at the crossvalidation point estimates for the prediction error. For a nominal or ordinal discrete response variable, other distance functions between the predicted and observed may be more appropriate.

We have used the simple estimating function in (3) to estimate the parameters of the fitted model, but have used $\hat{D}(\beta)$ in (2) for model evaluation. It would be ideal to use the same criterion for both stages, estimating $\beta$ by $\hat{\beta}$, which minimizes $\hat{D}(\beta)$ based on the training set, and then estimating the prediction error with $\hat{D}(\hat{\beta})$ based on the validation set. Unfortunately, it is not clear that the resulting $\hat{\beta}$ would converge to a constant vector, as $n$ increases, so as to justify the large-sample distribution of $\hat{D}(\hat{\beta})$. Moreover, when $Y$ is binary, we find that such a minimizer $\hat{\beta}$ may not exist.

When $Y_i, i = 1, \ldots, n$, are continuous event times, but possibly censored, it is not clear how to estimate the prediction error $D_0$ in (1), especially when the support of the censoring is much shorter than that of the event time. On the other hand, if one is interested in predicting certain $t$-year survival probabilities, it seems possible to develop model evaluation procedures using approaches similar to those taken here for the case with a binary outcome.

Suppose that there are two predictors, an inexpensive $X^{(1)}$ and an expensive or invasive $X^{(2)}$. An important and interesting question is whether and when $X^{(2)}$ could improve the prediction of a future $Y^0$ after observing $X^{(1)}$. Further research on this topic is highly warranted.

## Acknowledgement

The authors are grateful to the referee and the editor for helpful comments on the paper. The work is partially supported by the I2B2 and AIDS grants from the U.S. National Institutes of Health.

## Appendix 1

### *Existence and uniqueness of the root to the estimating function*

First, we show that, under the mild conditions imposed on $g(\cdot)$, $Y$ and $Z$ in §2, there is a unique root to the equation $E\{S(\beta)\} = 0$. To this end, for a given $p$-dimensional unit vector $b$, let $d(t; b) = b'E\{S(tb)\}$, a function in $t \in R$. Here, we show that if, for any given $b$,

$$d(+\infty; b)d(-\infty; b) < 0, \tag{A1}$$

then the estimating equation $E\{S(\beta)\} = 0$ has at least one solution. If $\beta = 0$ is not a solution, we show first that $d(t; b)$ always has a unique solution in $t$ for any given unit vector $b$. Since $\dot{d}(t; b) = -E\{\dot{g}(b'Zt)(b'Z)^2\} < 0$, $d(t; b)$ is a strictly decreasing function in $t$. It follows that $d(+\infty, b) < 0$ and $d(-\infty, b) > 0$, and $d(t; b) = 0$ has a unique solution, $t_0(b)$ say, by the continuity of $d(t; b)$. We then define a map $H$ from the unit sphere $S^{p-1} = \{b \mid \|b\| = 1\}$ to $R^p$: $H(b) = E[S\{t_0(b)b\}]$. Since $b'H(b) = d\{t_0(b); b\} = 0$, $H(b)$ induces a continuous vector field on the unit sphere $S^{p-1}$.

When $p = 3, 5, \ldots$, it follows from the Hairy Ball Theorem (Hatcher, 2002, Theorem 2.28) that there exists a vector $b_0$ such that $H(b_0) = 0$, that is, $H\{t_0(b_0)b_0\} = 0$, and $t_0(b_0)b_0$ is a solution to the equation $E\{S(\beta)\} = 0$.

Now consider the case in which $p$ is an even number. Note that $H(b) = H(-b)$ because $d(t; b) = -d(-t; -b) \Rightarrow t_0(b) = -t_0(-b)$. When $p = 2$, it is trivial to show that there is a $b_0 \in S^1$ such that $H(b_0) = 0$. When $p = 4, 6, \ldots$, consider all vectors $b = (0, b_2, \ldots, b_p)$ on the $(p-1)$-dimensional unit sphere. They form a $(p-2)$-dimensional unit sphere $S^{p-2}$. For any given $b = (0, b_2, \ldots, b_p)$, construct a circle $S_b^1 = \{e = (b_1, rb_2, \ldots, rb_p) \mid r \in [-1, 1], \|e\| = 1\}$, containing $b$. For a given $e = (b_1, rb_2, \ldots, rb_p)' \in S_b^1$, we decompose $H(e)$ into a sum of two orthogonal vectors, $\mathcal{H}_{p-2}(e)$ and $\mathcal{H}_1(e)$, where $\mathcal{H}_{p-2}(e) = (0, h_2(e) - d(e)b_2, \ldots, h_p(e) - d(e)b_p)'$, $\mathcal{H}_1(e) = (h_1(e), d(e)b_2, \ldots, d(e)b_p)'$, $d(e) = b_2h_2(e) + \ldots + b_ph_p(e)$, and $H(e) = (h_1(e), \ldots, h_p(e))'$. Note that $\mathcal{H}_1(e)$ is a continuous vector field on $S_b^1$ and satisfies $\mathcal{H}_1(e) = \mathcal{H}_1(-e)$. Therefore, for any $b = (0, b_2, \ldots, b_p)$, there exists a unit vector $e_0(b) \in S_b^1$ such that $\mathcal{H}_1\{e_0(b)\} = 0$. Also, note that $\mathcal{H}_{p-2} : b \to \mathcal{H}_{p-2}\{e_0(b)\}$ induces a continuous vector field on $S^{p-2}$. Since $p - 2 = 2, 4, \ldots$, it follows from the Hairy Ball Theorem that there exists a unit vector $b^*$ such that $\mathcal{H}_{p-2}\{e_0(b^*)\} = 0$. Therefore, $H\{e_0(b^*)\} = \mathcal{H}_{p-2}\{e_0(b^*)\} + \mathcal{H}_1\{e_0(b^*)\} = 0$. Lastly, since $g(\cdot)$ is strictly increasing and $E(ZZ')$ is strictly positive definite, the root is unique.

One now needs to check (A1). For a continuous response variable $Y$, if $E(ZY) < \infty$ and $J \subset [g(-\infty), g(+\infty)]$, then $d(+\infty, b) = \lim_{t\to\infty} E[b'Z\{Y - g(b'Zt)\}] = E[I(b'Z > 0)b'Z\{Y - g(+\infty)\}] + E[I(b'Z < 0)b'Z\{Y - g(-\infty)\}] < 0$. Similarly, $d(-\infty, b) > 0$. For a binary $Y$, if $\lim_{t\to\infty} g(t) = 1$, $\lim_{t\to\infty} g(-t) = 0$ and $\text{pr}(Y_1 > Y_2 | b'Z_1 > b'Z_2) < 1$ for all $b$, then $d(+\infty; b) = E\{I(b'Z > 0)b'Z(Y - 1) + I(b'Z < 0)b'ZY\} < 0$ and $d(-\infty; b) > 0$.

To show that there is a unique solution to the estimating equation $S(\beta) = 0$, almost surely, one simply replaces the expectation $E$ in $d(t; b) = b'E\{S(tb)\}$ with the expected value taken under the empirical distribution generated by $\{(Y_i, Z_i), i = 1, \ldots, n\}$.

Lastly, since $S(\beta)$ is monotone in $\beta$, it converges to $E\{S(\beta)\}$ uniformly in any compact set of $\beta_0$ in probability. It follows that $\hat{\beta}$ converges to $\beta_0$ in probability, as $n \to \infty$.

## Appendix 2

### *Large sample properties of $\hat{D}(\hat{\beta})$*

First, we show that $\hat{D}(\hat{\beta}) - D_0$ converges to 0 in probability, as $n \to \infty$. Since the conditional density or probability mass function of $Y^0$ given $Z^0$ is continuously differentiable, $E|Y^0 - \hat{Y}(\beta'Z^0)|$ is continuously differentiable in $\beta$. Moreover, since $g(\cdot)$ is strictly increasing and $Z$ is bounded, it follows from a uniform law of large numbers (Pollard, 1990, Ch. 8) that $\sup_{\beta\in\Omega} \left|\hat{D}(\beta) - E|Y^0 - \hat{Y}(\beta'Z^0)|\right|$

goes to 0, where $\Omega$ is a compact parameter set containing $\beta_0$. This, coupled with the convergence of $\hat{\beta}$ to $\beta_0$, implies that $\{\hat{D}(\hat{\beta}) - E|Y^0 - \hat{Y}(\hat{\beta}' Z^0)|\}$ converges to 0 in probability.

To derive the large sample distribution of $\hat{D}(\hat{\beta})$, first, since $g(\cdot)$ is differentiable,

$$n^{1/2}(\hat{\beta} - \beta_0) \doteq n^{-1/2} \sum_{i=1}^{n} [E\{A(\beta_0)\}]^{-1} Z_i\{Y_i - g(\beta_0' Z_i)\}, \tag{A2}$$

where $A(\beta)$ is defined by (6). Secondly, we show that the class of functions indexed by $\beta$, $\mathcal{G} = \{|y - \hat{Y}(\beta' z)| : \|\beta - \beta_0\| \leqslant \delta\}$, is a Donsker class for both continuous and binary $Y$ with appropriate $\hat{Y}(x)$, where $\delta$ is a given positive number. When $Y$ is continuous and $\hat{Y}(x) = g(x)$, the function $|y - g(\beta' z)|$ is Lipschitz in $\beta$, which implies that $\mathcal{G}$ is a Donsker class (van der Vaart & Wellner, 1996, Theorem 2.7.11). When $Y$ is binary and $\hat{Y}(\beta' z) = I\{\beta' z \geqslant g^{-1}(c)\}$, the class of functions $\{\hat{Y}(\beta' z) : \|\beta - \beta_0\| \leqslant \delta\}$ is a Vapnik–Chervonenkis class (van der Vaart & Wellner, 1996, Lemma 2.6.15) and $\mathcal{G} = \{I(y = 0)\hat{Y}(\beta' z) + I(y = 1)\{1 - \hat{Y}(\beta' z)\} : \|\beta - \beta_0\| \leqslant \delta\}$ is therefore a Donsker class. It follows that in both cases the process $n^{1/2}[\hat{D}(\beta) - E\{\hat{D}(\beta)\}]$ is stochastic equicontinuous at $\beta_0$. This, coupled with the fact that $E\{\hat{D}(\beta)\}$ is continuously differentiable at $\beta_0$ and (A2), implies that

$$n^{1/2}\{\hat{D}(\hat{\beta}) - D_0\} \doteq n^{1/2}\{\hat{D}(\beta_0) - D_0\} + E\{d(\beta_0)\} n^{1/2}(\hat{\beta} - \beta_0) \doteq n^{-1/2} \sum_{i=1}^{n} \eta_i,$$

where $d(\beta)$ is defined by (7) and

$$\eta_i = |Y_i - \hat{Y}(\beta_0' Z_i)| - D_0 + E\{d(\beta_0)\}[E\{A(\beta_0)\}]^{-1} Z_i\{Y_i - g(\beta_0' Z_i)\}. \tag{A3}$$

Thus, $n^{1/2}\{\hat{D}(\hat{\beta}) - D_0\}$ converges in distribution to a zero-mean normal random variable. Moreover, $\hat{\eta}_i$ for the variance estimator (5) is obtained by replacing all the theoretical quantities for $\eta_i$ in (A2) with their empirical counterparts.

## APPENDIX 3

### *Large sample properties of $\hat{\mathcal{D}}$*

For each partition $\mathcal{I}_k$, $n^{1/2}\{\hat{D}_{(k)}(\hat{\beta}_{(-k)}) - D_0\}$ is asymptotically equivalent to

$$n^{-1/2} K \sum_{i=1}^{n} I(\xi_i = k)\big\{|Y_i - \hat{Y}(\hat{\beta}_{(-k)}' Z_i)| - D_0\big\},$$

where $\{\xi_i; i = 1, \ldots, n\}$ are $n$ exchangeable discrete random variables uniformly distributed over $\{1, 2, \ldots, K\}$, independent of the data, which satisfy $\sum_{i=1}^{n} I(\xi_i = k) \doteq n/K, k = 1, \ldots, K$. Conditional on $\{\xi_i; i = 1, \ldots, n\}$, it follows from the standard large-sample expansion of a smooth estimating function that

$$\hat{\beta}_{(-k)} - \beta_0 = \frac{K}{n(K-1)} [E\{A(\beta_0)\}]^{-1} \sum_{i=1}^{n} I(\xi_i \neq k) Z_i\{Y_i - g(\beta_0' Z_i)\} + o_p(n^{-1/2}).$$

Then, using the same argument as in Appendix 2, one can show that $n^{1/2}\{\hat{D}_{(k)}(\hat{\beta}_{(-k)}) - D_0\}$ is asymptotically equivalent to

$$n^{1/2}\{\hat{D}_{(k)}(\beta_0) - D_0\} + E\{d(\beta_0)\} n^{1/2}(\hat{\beta}_{(-k)} - \beta_0) \doteq n^{-1/2} \sum_{i=1}^{n} \eta_{ki},$$

where

$$\begin{aligned}\eta_{ki} = {} & I(\xi_i = k) K\{|Y_i - \hat{Y}(\beta_0' Z_i)| - D_0\} \\ & + I(\xi_i \neq k) \frac{K}{K-1} E\{d(\beta_0)\}[E\{A(\beta_0)\}]^{-1} Z_i\{Y_i - g(\beta_0' Z_i)\}.\end{aligned}$$

It follows that $n^{1/2}(\hat{\mathcal{D}} - D_0) \doteq n^{-1/2}\sum_{i=1}^{n}(\sum_{k=1}^{K} K^{-1}\eta_{ki}) = n^{-1/2}\sum_{i=1}^{n} \eta_i$ is asymptotically equivalent to $n^{1/2}\{\hat{D}(\hat{\beta}) - D_0\}$.

## Appendix 4

### *Large Sample Properties of* $\hat{\mathbb{D}}$

For a 'random' crossvalidation, suppose that $n_v = n(1-\rho)$ and $n_t = n\rho$, and let $\zeta_i = 1$ and 0 indicate that the $i$th observation belongs to the training and validation sample, respectively. Then $\{\zeta_i, i = 1, \ldots, n\}$ are independent and identically distributed and are independent of the data with $\text{pr}(\zeta_i = 1) = \rho$. Let $\hat{\beta}_t$ denote the estimator of $\beta_0$ obtained from the training sample. It follows from the standard large sample expansion around $\beta_0$ that

$$n^{1/2}(\hat{\beta}_t - \beta_0) = n^{-1/2}[\rho E\{A(\beta_0)\}]^{-1}\sum_{i=1}^{n}\zeta_i Z_i\{Y_i - g(\beta_0' Z_i)\} + \epsilon,$$

where $\epsilon = o_p(1)$. Here and in the sequel, except when indicated otherwise, the probability and expectation are with respect to the product probability measure generated under $\{\zeta_i, i = 1, \ldots, n\}$ and $\{(Y_i, Z_i), i = 1, \ldots, n\}$. Furthermore, with arguments similar to those used in Newey & Smith (2004, p. 225) and Barndorff-Nielsen & Cox (1994, p. 304), it can be shown that $E|\epsilon| = o(1)$ and $E(n\|\hat{\beta}_t - \beta_0\|^2) < \infty$.

In addition, one can show that $n^{1/2}\{\hat{D}_v(\hat{\beta}_t) - D_0\}$ is asymptotically equivalent to

$$n^{1/2}\{\hat{D}_v(\beta_0) - D_0\} + E\{d(\beta_0)\}n^{1/2}(\hat{\beta}_t - \beta_0) \doteq n^{-1/2}\sum_{i=1}^{n}\eta_i^*,$$

where

$$\hat{D}_v(\beta) = \frac{1}{n(1-\rho)}\sum_{i=1}^{n}(1-\zeta_i)|Y_i - \hat{Y}(\beta' Z_i)|$$

is the estimated prediction error conditional on $\{\zeta_i; i = 1, \ldots, n\}$, and

$$\eta_i^* = \left\{\frac{1-\zeta_i}{1-\rho}|Y_i - \hat{Y}(\beta_0' Z_i)| - D_0\right\} + \frac{\zeta_i}{\rho}E\{d(\beta_0)\}[E\{A(\beta_0)\}]^{-1}Z_i\{Y_i - g(\beta_0' Z_i)\}.$$

Let

$$\mathcal{E} = n^{1/2}\{\hat{D}_v(\hat{\beta}_t) - D_0\} - n^{-1/2}\sum_{i=1}^{n}\eta_i^*, \quad \tilde{\mathbb{D}}_n(\beta) = n^{-1}\sum_{i=1}^{n}\left|\frac{1-\zeta_i}{1-\rho}|Y_i - \hat{Y}(\beta' Z_i)| - D_0(\beta)\right|,$$

and $\mathcal{E}_0 = \sup_{|\beta-\beta_0|\leqslant\varepsilon_0} n^{1/2}|\tilde{\mathbb{D}}_n(\beta) - \tilde{\mathbb{D}}_n(\beta_0)|$. It can then be shown that, for any sufficiently small $\varepsilon_0 > 0$,

$$E|\mathcal{E}| \lesssim E|\mathcal{E}_0| + n^{-1/2}\varepsilon_0^{-2}E(n\|\hat{\beta}_t - \beta_0\|^2) + E|\epsilon|, \qquad \text{(A4)}$$

where the notation $\lesssim$ means being bounded above up to a universal constant. It follows from Theorems 2.14.1 and 2.14.2 of van der Vaart & Wellner (1996) that $E|\mathcal{E}_0|$ can be bounded using the bracketing or uniform entropy integral of the class of functions $\mathcal{F} = \{(1-\zeta)(|y - \hat{Y}(\beta' z)| - |y - \hat{Y}(\beta_0' z)|) : \|\beta - \beta_0\| \leqslant \varepsilon_0\}$.

In what follows, we show that, both when $Y$ is continuous with $\hat{Y}(x) = g(x)$ and when $Y$ is binary with $\hat{Y}(x) = I\{g(x) \geqslant c\}$, there exists $\varepsilon_0$ such that $E|\mathcal{E}_0|$ is arbitrarily small. When $Y$ is continuous and $\hat{Y}(x) = g(x)$, $\mathcal{F} = \{(1-\zeta)(|y - g(\beta' z)| - |y - g(\beta_0' z)|) : \|\beta - \beta_0\| \leqslant \varepsilon_0\}$ and has an envelope function

$$M_{\varepsilon_0}(z) = \varepsilon_0\|z\| \sup_{\|\beta-\beta_0\|\leqslant\varepsilon_0} \dot{g}(\beta' z).$$

Let $J_{[\,]}\{\delta, \mathcal{F}, L_2(P)\}$ denote the bracketing integral defined in van der Vaart & Wellner (1996, p. 240). It follows from Theorems 2.7.11 and 2.14.2 of van der Vaart & Wellner (1996) that $J_{[\,]}\{1, \mathcal{F}, L_2(P)\} \lesssim \int_0^1 \{-\log(x)\}^{1/2} dx < \infty$ and

$$E|\mathcal{E}_0| \lesssim J_{[\,]}\{1, \mathcal{F}, L_2(P)\}[E\{M_{\varepsilon_0}(Z)^2\}]^{1/2} \lesssim \varepsilon_0,$$

When $Y$ is binary and $\hat{Y}(x) = I\{g(x) \geqslant c\}$,

$$\mathcal{F} = \{(1-\zeta)(2y-1)[I\{\beta' z \geqslant g^{-1}(c)\} - I\{\beta_0' z \geqslant g^{-1}(c)\}] : \|\beta - \beta_0\| \leqslant \varepsilon_0\}$$

is a Vapnik–Chervonenkis class and its uniform entropy integral satisfies $J(1, \mathcal{F}) \lesssim \int_0^1 \{-\log(x)\}^{1/2} dx < \infty$ (van der Vaart & Wellner, 1996, p. 239). This, coupled with Theorem 2.14.1 of van der Vaart & Wellner (1996), implies that

$$E|\mathcal{E}_0| \lesssim [E\left\{M_{\varepsilon_0}(Z)^2\right\}]^{1/2},$$

where the envelope function $M_{\varepsilon_0}(z) = I\{|\beta_0' z - g^{-1}(c)| \leqslant \varepsilon_0 K\}$ and $K$ is a constant depending only on the support of $Z$. Therefore, if $Z$ contains at least one continuous component and $g^{-1}(c)$ is an interior point of the support of $\beta_0' Z$, then $E\{M_{\varepsilon_0}(Z)^2\}^{1/2} \lesssim \varepsilon_0^{1/2}$. If all the components of $Z$ are discrete and $g^{-1}(c)$ is not equal to any $\beta_0' Z$, then $E|\mathcal{E}_0| = 0$ for sufficiently small $\varepsilon_0$. We have therefore shown that $E|\mathcal{E}_0| = o(1)$ under suitable conditions. This, together with (A4), implies that $E|\mathcal{E}| \to 0$. Then, by Markov's inequality,

$$\begin{aligned} &\mathrm{pr}_{Y,Z}\left(\left| n^{1/2}\left[E_\zeta\{\hat{D}_v(\hat{\beta}_t)\} - D_0\right] - n^{-1/2}\sum_{i=1}^n \eta_i \right| > \varepsilon_0\right) \\ &= \mathrm{pr}_{Y,Z}\left\{\left|E_\zeta(\mathcal{E})\right| > \varepsilon_0\right\} \leqslant \mathrm{pr}_{Y,Z}\left(E_\zeta|\mathcal{E}| > \varepsilon_0\right) \leqslant \varepsilon_0^{-1} E|\mathcal{E}| \to 0, \end{aligned}$$

for any $\varepsilon_0 > 0$, where $E_\zeta$ is the expectation with respect to $\{\zeta_i; i = 1, \ldots, n\}$ and $\mathrm{pr}_{Y,Z}$ is the probability measure with respect to $\{(Y_i, Z_i); i = 1, \ldots, n\}$. It follows that $n^{1/2}(\hat{\mathbb{D}} - D_0) = n^{1/2}[E_\zeta\{\hat{D}_v(\hat{\beta}_t)\} - D_0] \doteq n^{-1/2}\sum_{i=1}^n \eta_i$ is asymptotically equivalent to $n^{1/2}\{\hat{D}(\hat{\beta}) - D_0\}$.

## REFERENCES

AKAIKE, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* **60**, 255–65.

BARNDORFF-NIELSEN, O. E. & COX, D. R. (1994). *Inference and Asymptotics*. London: Chapman and Hall.

BREIMAN, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: $X$-fixed prediction error. *J. Am. Statist. Assoc.* **87**, 738–54.

CAI, T., TIAN, L. & WEI, L. (2005). Semiparametric Box-Cox power transformation models for censored survival observations. *Biometrika* **92**, 619–32.

DAVISON, A. C. & HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

EFRON, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Am. Statist. Assoc.* **78**, 316–31.

EFRON, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Am. Statist. Assoc.* **81**, 461–70.

EFRON, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *J. Am. Statist. Assoc.* **99**, 619–32.

EFRON, B. & TIBSHIRANI, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *J. Am. Statist. Assoc.* **92**, 548–60.

HAMMER, S., SQUIRES, K., HUGHES, M., GRIMES, J., DEMETER, L., CURRIER, J., ERON, J., FEINBERG, J., BALFOUR, H., DEYTON, L., CHODAKEWITZ, J., FISCHL, M., PHAIR, J., SPREEN, W., PEDNEAULT, L., NGUYEN, B., COOK, J. & ACTG 320 STUDY TEAM. (1997). A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. *N. Engl. J. Med.* **337**, 725–33.

HATCHER, A. (2002). *Algebraic Topology*. Cambridge: Cambridge University Press.

HOSMER, D. W. & LEMESHOW, S. (2000). *Applied Logistic Regression*. New York: John Wiley and Sons.

JACOBSEN, M. (1989). Existence and unicity of MLEs in discrete exponential family distributions. *Scand. J. Statist.* **16**, 335–49.

Mallows, C. L. (1973). Some comments on $C_p$. *Technometrics* **15**, 661–75.

Molinaro, A., Simon, R. & Pfeiffer, R. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics* **21**, 3301–7.

Newey, W. K. & Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* **72**, 219–55.

Park, Y. & Wei, L. J. (2003). Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika* **90**, 717–23.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*. Hayward, CA: Institute of Mathematical Statistics.

Shao, J. (1993). Linear model selection by cross-validation. *J. Am. Statist. Assoc.* **88**, 486–94.

Shao, J. (1996). Bootstrap model selection. *J. Am. Statist. Assoc.* **91**, 655–65.

Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *J. R. Statist. Soc.* B **43**, 310–3.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9**, 1135–51.

Tibshirani, R. & Knight, K. (1999). Model search by bootstrap "bumping". *J. Comp. Graph. Statist.* **8**, 671–86.

Van der vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. New York: Springer-Verlag Inc.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–26.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *J. Am. Statist. Assoc.* **93**, 120–31.

[*Received November* 2005. *Revised October* 2006]