Harvard University
Harvard University Biostatistics Working Paper Series

Year 2005 *Paper* 35

Model Evaluation Based on the Distribution of Estimated Absolute Prediction Error

Lu Tian[*] Tianxi Cai[†]
Els Goetghebeur[‡] L. J. Wei[**]

[*]Northwestern University, lutian@northwestern.edu
[†]Harvard University, tcai@hsph.harvard.edu
[‡]Harvard School of Public Health, egoetghe@hsph.harvard.edu
[**]Harvard University, wei@hsph.harvard.edu

http://biostats.bepress.com/harvardbiostat/paper35

Model Evaluation Based on the Distribution of Estimated Absolute Prediction Error

Lu Tian, Tianxi Cai, Els Goetghebeur, and L. J. Wei

Abstract

The construction of a reliable, practically useful prediction rule for future response 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 mis-classification error for the binary outcome. We show that the distributions of the apparent error and its cross-validation counterparts are approximately normal even under a misspecified fitted model. When the prediction rule is "unsmooth", the variance of the above normal distribution can be estimated well via a perturbation-resampling method. We also show how to approximate the distribution of the difference of the estimated prediction errors from two competing models. With two real examples, we demonstrate that the resulting interval estimates for prediction errors provide much more information about model adequacy than the point estimates alone.

MODEL EVALUATION BASED ON THE DISTRIBUTION OF ESTIMATED ABSOLUTE PREDICTION ERROR

Lu Tian
Department of Preventive Medicine, Northwestern University Medical School, Chicago, IL 60611, U.S.A.
Tianxi Cai, Els Goetghebeur and L. J. Wei
Department of Biostatistics, Harvard University, 655 Huntington Ave, Boston, MA 02115, U.S.A.

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 mis-classification error for the binary outcome. We show that the distributions of the apparent error and its cross-validation counterparts are approximately normal even under a misspecified fitted model. When the prediction rule is "unsmooth", the variance of the above normal distribution can be estimated well via a perturbation-resampling method. We also show how to approximate the distribution of the difference of the estimated prediction errors from two competing models. With two real examples, we demonstrate that the resulting interval estimates for prediction errors provide much more information about model adequacy than the point estimates alone.

Keywords: 0.632 resampling; Bootstrap; K-fold cross-validation; Model and variable selections; Perturbation-resampling; Prediction.

1. Introduction

One of main goals for fitting data with regression models is to construct reliable, parsimonious prediction rules for future responses. Often aggregate prediction errors, which measure the "distance" between the future and predicted outcomes, are utilised to evaluate the adequacy of a fitted model or compare competing models (Davison & Hinkley, 1997, Section 6.4). Methods to estimate prediction errors are mainly based on the apparent or re-substitution error, cross-validation, bootstrap and covariance penalties (Mallows, 1973; Akaike, 1973; Stein, 1981; Efron, 1983, 1986; Breiman, 1992; Shao, 1993, 1996; Efron & Tibshirani, 1997; Ye, 1998; Tibshirani & Knight, 1999; Efron, 2004). 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 (Efron & Tibshirani, 1995, Section 5).

In this article, we consider the case that the sample size is relatively large with respect to the dimension of the vector of regression parameters. Furthermore, instead of using L_2 norm, we consider the average absolute prediction error, the expected value of the absolute difference between the future and predicted responses to assess model adequacy. For binary response, this prediction error is the misclassification error. Without assuming that the fitted model is the true model, we show that the apparent error consistently estimates the prediction error and the distribution of the standardised apparent error is approximately normal. We then show that this normal can be approximated well via a perturbation-

resampling method, especially for unsmooth prediction rules. Based on the above normal approximation, confidence intervals for the prediction errors are constructed accordingly, which provide more information about model adequacy than point estimates alone.

In this paper, we also show that the limiting distributions of various cross-validation estimators for such a prediction error are the same as that of the apparent error. Moreover, empirically we find that the bias issue of the apparent error even with modest sample sizes is not alarming. Lastly, we show how to construct interval estimates for the difference of the prediction errors of two competing fitted models. All the proposals are illustrated with two real examples.

2. Approximations to the Distribution of Estimated Prediction Error

Let Y be a continuous or binary response variable and X be the vector of its predictors. Let Z, a $p \times 1$ bounded vector, be a function of X. Also, let $\{(Y_i, Z_i), i = 1, \cdots, n\}$ be n independent copies of (Y, Z). For a future, independent subject from the same population of (Y, Z), suppose that its $Z = Z^0$ and its response Y^0 is predicted based on 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 β is the vector of unknown parameters. Let $\hat{\beta}$ be an estimate of β based on the entire data set $\{(Y_i, Z_i)\}$ 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 $\hat{Y}(\hat{\beta}' Z^0) = I(g(\hat{\beta}' Z^0) \geq 0.5)$, 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

$$D_0 = E|Y^0 - \hat{Y}(\hat{\beta}' Z^0)| \tag{2.1}$$

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

re-substitution error" $\hat{D}(\hat{\beta})$, where

$$\hat{D}(\beta) = n^{-1} \sum_{i=1}^{n} |Y_i - \hat{Y}(\beta' Z_i)| \tag{2.2}$$

(Davison and Hinkley, 1997, Section 6.4).

To approximate the large sample distribution of $\{\hat{D}(\hat{\beta}) - D_0\}$, we need to show that $\hat{\beta}$ is stablised 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 the parametric likelihood score function $S^{\dagger}(\beta)$ to estimate β, 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, the above uniqueness condition is rather difficult to verify even when the estimator $\hat{\beta}$ exists and is unique for any finite sample size n under the fitted model (Silvapulle, 1981; Jacobsen, 1989).

In this article, we propose to estimate β via the following simple estimating function

$$S(\beta) = n^{-1} \sum_{i=1}^{n} Z_i \{Y_i - g(\beta' Z_i)\}. \tag{2.3}$$

We assume that if J is the support of Y, $J \subseteq [g(-\infty), g(+\infty)]$, $E(Y) < \infty$, Z is uniformly bounded 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 these 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 A, 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 β_0 to $E\{S(\beta)\} = 0$. We then show that $\hat{\beta}$ converges to β_0 in probability, as $n \to \infty$.

Now, assume that the conditional density or probability mass function of Y given Z is continuously differentiable. In Appendix B, 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 make 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 β, in Appendix B, we show that the distribution of

$$W = n^{1/2}\{\hat{D}(\hat{\beta}) - D_0\} \tag{2.4}$$

is asymptotically Gaussian with mean 0.

Now, if $\hat{Y}(\hat{\beta}'Z^0) = g(\hat{\beta}'Z^0)$, D_0 in (2.1) becomes $E|Y^0 - g(\hat{\beta}'Z^0)|$. For this case, the variance of W in (2.4) can be consistently estimated by

$$n^{-1}\sum_{i=1}^{n}\hat{\eta}_i^2, \tag{2.5}$$

where

$$\hat{\eta}_i = |Y_i - g(\hat{\beta}'Z_i)| - \hat{D}(\hat{\beta}) + d(\hat{\beta})A^{-1}(\hat{\beta})Z_i\{Y_i - g(\hat{\beta}'Z_i)\},$$
$$A(\beta) = n^{-1}\sum_{i=1}^{n}\dot{g}(\beta'Z_i)Z_iZ_i', \tag{2.6}$$

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

$$d(\beta) = -n^{-1}\sum_{i=1}^{n}\text{sign}\{Y_i - g(\beta'Z_i)\}\dot{g}(\beta'Z_i)Z_i, \tag{2.7}$$

the quasi-derivative of $\hat{D}(\beta)$. The justification of consistency of (2.5) is given in Appendix B.

If $\hat{Y}(\hat{\beta}'Z^0)$ is not $g(\hat{\beta}'Z^0)$, for example, when Y is binary and $\hat{Y}(\hat{\beta}'Z^0) = I(g(\hat{\beta}'Z^0) \geq c)$, 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, which is difficult to estimate well 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 $G_i, i = 1, \cdots, n$, be independent and identically distributed random variables with a known distribution whose

mean and variance are one. Furthermore, let β^* be the solution to the equation

$$S^*(\beta) = n^{-1} \sum_{i=1}^{n} \{y_i - g(\beta' z_i)\} G_i = 0. \tag{2.8}$$

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

$$W^* = n^{-1/2} \sum_{i=1}^{n} \{|y_i - \hat{Y}(z_i'\beta^*)| - \tilde{D}(\tilde{\beta})\}(G_i - 1). \tag{2.9}$$

It is straightforward to show that for large n, the unconditional distribution of W in (2.4) can be approximated well by the conditional distribution of W^* given the data. This perturbation-resampling technique has been utilised 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, say, M of realizations from $\{G_i, i = 1, \cdots, n\}$. For each realized sample $\{G_i, i = 1, \cdots, n\}$, 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 fitted model.

It is interesting to note that if $(G_1, \cdots, G_n)$ is a multinomial random vector with size n and marginal cell probabilities of n^{-1}, the resulting W^* by replacing $G_i - 1$ in (2.9) by G_i is essentially the bootstrap counterpart of W. It is not clear, however, how to justify analytically whether 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 β, $\hat{D}(\hat{\beta})$ may underestimate D_0. One remedy to reduce such bias is to use cross-validation procedures to estimate D_0. To this end, first, let us consider the popular K-fold cross-validation. Specifically, we randomly split the data into K disjoint subsets of about equal size and label them

as $\mathcal{I}_k, k = 1, \cdots, K$. For each k, we use all observations which are not in $\mathcal{I}_k$ to obtain an estimate $\hat{\beta}_{(-k)}$ for β via (2.3), and then compute the predicted error estimate $\hat{D}_{(k)}\{\hat{\beta}_{(-k)}\}$ via (2.2) based on observations in $\mathcal{I}_k$. Then, an average prediction error estimate for D_0 is

$$\hat{\mathcal{D}} = K^{-1} \sum_{k=1}^{K} \hat{D}_{(k)}\{\hat{\beta}_{(-k)}\}. \tag{2.10}$$

When K is fixed and relatively small with respect to n, for each $k = 1, \cdots, K$, 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 β_0. It follows from the multivariate central limit theorem that

$$\mathcal{W} = n^{1/2}\{\hat{\mathcal{D}} - D_0\} \tag{2.11}$$

is asymptotically normal. In Appendix C, 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 slightly different, but, for large n, a confidence interval for the absolute predicted error based on the K-fold cross-validation method approximately has the same length as that based on the apparent error.

Now, for a more general cross-validation scheme, let n_t and n_v be the sizes of the training and validation sets, where n/n_v is approximately a positive integer, and n_t and $n_v \to \infty$, as $n \to \infty$. Given the data, we randomly choose a training set, use those observations in this set to estimate β via (2.3), then compute the corresponding $\hat{D}(\hat{\beta})$ in (2.2) 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 $\hat{D}(\hat{\beta})$ defined as (2.10), but the summation is over the entire set of possible training-validation splits. In Appendix D, we show that $n^{1/2}(\hat{\mathbb{D}} - D_0)$ has the same limiting distribution as that of W in (2.4). 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 cross-validation procedures.

An interesting hybrid of cross-validation and apparent error, the 0.632 bootstrap estimator, for estimating the prediction error was proposed by Efron and Tibshirani (1997). This estimator is essentially a linear combination of the apparent error and a cross-validation counterpart. If the cross-validation component belongs to the class discussed in the last paragraph, this combination has the same large sample distribution as W does. However, since Efron and Tibshirani's estimator utilises a smooth version of the leave-one-out cross-validation, 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, say, $g_j(\hat{\beta}_j' Z_{(j)}), j = 1, 2$, where the p_j-dimensional vector $Z_{(j)}$ is a functions of X and $\hat{\beta}_j$ is the estimator via (2.3) with the data $\{(Y_i, Z_{(j)i}), i = 1, \cdots, n\}$. The theoretical and empirical prediction errors D_{0j} and $\hat{D}_j(\beta_j)$ are defined by (2.1) and (2.2) accordingly, $j = 1, 2$. We are interested in making inferences about, for example, $\Delta = D_{02} - D_{01}$ to assess how much improvement Model 1 is over Model 2.

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

$$W_\Delta = n^{1/2}\{\hat{\Delta} - \Delta\} \tag{3.1}$$

is asymptotically normal with mean 0. To approximate this normal distribution, one may use the analytical or perturbation method discussed in Section 2. Specifically, for resampling method, let β_j^* be the solution of

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

$j = 1, 2$. 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_Δ can be approximated by the conditional distribution of $W^*_\Delta = W^*_2 - W^*_1$. Confidence intervals for Δ based on this normal approximation can then be constructed.

For the K-fold cross validation method, the estimator $\hat{\mathcal{D}}_2 - \hat{\mathcal{D}}_1$, where $\hat{\mathcal{D}}_j$ is defined by (2.10) 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 (2.11) based on Model j. The limiting distribution of $\{\mathcal{W}_2 - \mathcal{W}_1\}$ is the same as that of W_Δ. Similarly, for a general cross-validation or its hybrids discussed in Section 2, the distribution of the corresponding $n^{1/2}\{\hat{\mathbb{D}}_2 - \hat{\mathbb{D}}_1\}$ can also be approximated by that of W_Δ.

4. Examples and Numerical Studies

We use two examples to illustrate the proposed procedures. One is with a continuous response and the other is with a binary dependent variable. 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 this relatively potent therapy, a significant proportion of patients will not respond to the treatment. Therefore, it is 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 24, an important measure of the patient's immune response. This variable is still one of the major endpoints for modern clinical studies on HIV diseases, especially conducted in resource-limited countries. Based on ACTG 320, the potential early predictors X for such Y are age, baseline CD4, baseline HIV-1 RNA, and the changes of CD4 and RNA from Week 0 to Week 8. Since RNA measure is relatively expensive to obtain, an important and

interesting question is whether 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 range from -100 to 734. To evaluate the added value from early RNA marker values, first, we assume that the conditional mean of Y given Z has a parametric form of $g(\beta' Z) = \beta' Z$, where $Z = (1, X)$, a 6×1 vector. With the estimating function (2.3), 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 $\hat{D}(\hat{\beta}) = 51$ with an estimated standard error of 2.7 based on (2.5). The 0.95 confidence interval of D_0 is $(46, 56)$, a rather tight interval from a clinical point of view.

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 of the prediction error for the full model with and the one without RNA values via W_Δ in (3.1) is

$$(-2.0, 0.4). \tag{4.1}$$

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 involved.

With the 10-fold cross-validation method, the 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 (2.5) 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" cross-validation discussed in Section 2 with $n_t = 2n/3$ and the 0.632-bootstrap method proposed by Efron and Tibshirani (1997) to evaluate the above two fitted models. The estimates for the difference of the prediction errors of these models are -0.50 and -0.63, respectively. To construct interval estimates based on the 0.632 method, we generated 500 by 500 double bootstrap samples to estimate the variance. The resulting 0.95 confidence interval for the difference of the prediction errors for the aforementioned two models is $(-1.8, 0.6)$, which is practically identical to interval (4.1).

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 more potentially important predictors on the top of early CD4 count change. However, it seems clear that early RNA measures do not add much value for predicting the patient's immune response.

We also conducted an extensive simulation study to examine the performance of the proposed inference procedures based on the apparent error and cross-validation counterparts under various scenarios. Specifically, we mimicked the above HIV example and generated data $\{(Y_i, Z_i), i = 1, \cdots, n\}$ from two linear regression models. The first model relates the response variable, the CD4 count change from Week 0 to 24, to a linear combination of five predictors in the aforementioned HIV study with a mean-zero, normal random error term. The deterministic part of the second model consists of all linear and also quadratic terms of these predictors. The true values of the model parameters, regression coefficients and error variances, of these two models are chosen using the least squares estimates with the observed data from ACTG 320. To generate the data from these two models, first we assume that the vector X of five predictors is jointly normal whose mean and covariance matrix are

estimated with the data from the HIV study. Then, for each model, we generated 1000 sets of $\{(Y_i, Z_i), i = 1, \cdots, n\}$, where Z_i was generated from the above multivariate normal. For each realized data set $\{(Y_i, Z_i), i = 1, \cdots, n\}$, we fitted the data with two working models. The first one is a linear, additive regression model with five predictors. The second one is also a linear, additive model, but with three predictors only: age, CD4 count baseline and early CD4 count change. For each case, we computed the empirical absolute prediction errors obtained by the apparent error, 10-fold cross-validation, a "random" cross-validation with $n_t = 2n/3$ and 0.632 bootstrap method. In Table 2, for each scenario, we report the estimates of bias and square root mean square error based on 1000 data sets. Note that for the 0.632 bootstrap and "random" cross-validation methods, we generated 100 bootstrap samples and 100 random training and validation sets for each realized data set, respectively. Also, for each case in Table 2, the "true" value of the prediction error is estimated by another 5000 independent sets $\{(Y^0, Z^0), (Y_i, Z_i), i = 1, \cdots, n\}$, where (Y^0, Z^0) was used to estimate the prediction error of the model based on $\{(Y_i, Z_i), i = 1, \cdots, n\}$. Based on all cases studied, we find that the apparent error tends to have slightly larger bias and square root mean square error than the other three procedures. However, it appears that there are no differences among these methods statistically or clinically.

The second example for illustration is from a prostate cancer study, which examines whether certain "baseline" bio- and clinical markers are helpful for predicting tumour penetration, a binary response variable (Hosmer & Lemeshow, 2000, Chapter 1). For this study, potential predictors include age, race, digital rectal exam (no nodule, left nodule, right nodule, and bilobar nodule), detection of capsular involvement in rectal exam (DCI), prostate specific antigen (PSA), tumour volume obtained from ultra sound (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 (2.3) 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 3, we present the point estimates of regression coefficients and their standard error estimates. Note that the PSA is highly statistically significant. With this model and the binary decision rule: $I(g(\hat{\beta}'Z^0) \geq 0.5)$, the apparent error for estimating the misclassification rate is 0.24. With $M = 1000$ perturbation samples $\{G_i\}$ in (2.9), the 0.95 confidence interval for the error rate is $(0.19, 0.29)$. The corresponding point estimate and 0.95 interval based on the 10-fold cross validation method is 0.27 and $(0.21, 0.33)$, respectively. The estimates from the "random" cross-validation with $n_t = 2n/3$ are 0.26 and $(0.20, 0.31)$. With 500 by 500 double bootstrap samples, the 0.632-bootstrap method gives an estimate of 0.25 and a 0.95 confidence interval of $(0.21, 0.30)$.

Since PSA is a routinely used, but controversial biomarker for diagnosis of prostate cancer, it is interesting to examine how much accuracy the PSA 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 (3.1) for the difference of prediction errors between the second and first models is 0.021 with 0.95 confidence interval $(-0.02, 0.062)$. The 10-fold cross-validation 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 there is any, on top of other variables, for predicting tumour penetration.

5. Remarks

In this paper, we derived model evaluation procedures for continuous and binary responses for which the L_1 prediction error is a meaningful, physically interpretable metric. For a response such as the nominal or ordinal discrete variable, other distance functions between the predicted and observed may be more appropriate.

In this article, we use a simple estimating function (2.3) to estimate the parameters of the fitted model, but utilize $\hat{D}(\beta)$ in (2.2) for model evaluation. It would be ideal to use

the same criterion for both stages, that is, we estimate β by $\hat{\beta}$, which minimizes $\hat{D}(\beta)$ with the training set, and then estimate the prediction error with $\hat{D}(\hat{\beta})$ with the validation set. Unfortunately, it is not clear that the resulting $\hat{\beta}$ would converge to a constant vector, as n increases, to justify the large sample approximation distribution of $\hat{D}(\hat{\beta})$. Moreover, when Y is binary, we find that such a minimizer $\hat{\beta}$ may not exist.

When $Y_i, i = 1, \cdots, n$, are continuous event times, but possibly censored, it is not clear how to estimate the prediction error D_0 in (2.1), 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 probability, it seems possible to develop model evaluation procedures using similar approaches taken in this article for handling the case with binary outcome.

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

Appendix A: 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 Section 2, there is a unique root to the equation $E\{S(\beta)\} = 0$. To this end, for a given $p-$dimensional unit vector b ($\|b\| = 1$), let $d(t; b) = b'E\{S(tb)\}$, a function in $t \in R$. Here, we show that if for any given b,

$$d(\infty, b)d(-\infty, b) < 0, \tag{A.1}$$

then the estimating equation $E\{S(\beta)\} = 0$ has at least one solution. Now, 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, say, $t_0(b)$ 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 $p = 3, 5, \cdots$, it follows from Hairy Ball Theorem (Hatcher, 2002, Theorem 2.28, Sec. 2.2) 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 that p is an even number. Note that $H(b) = H(-b)$ due to the fact that $d(t;b) = -d(-t;-b) \Rightarrow t_0(b) = -t_0(-b)$. When $p = 2$, it is trivial to show that there is a $b_0 \in S^1$ such that $H(b_0) = 0$. When $p = 4, 6, \cdots$, consider all vectors $b = (0, b_2, \cdots, b_p)$ on the $p-1$ dimensional unit sphere. They form a $p-2$ dimensional unit sphere S^{p-2}. For any given $b = (0, b_2, \cdots, b_p)$, construct a circle $S_b^1 = \{e = (b_1, rb_2, \cdots, rb_p) \mid r \in [-1, 1], \|e\| = 1\}$, containing b. For a given $e = (b_1, rb_2, \cdots, rb_p)' \in S_b^1$, we decompose $H(e)$ into a sum of two orthogonal vectors, $\mathcal{H}_{p-2}(e)$ and $\mathcal{H}_1(e)$, where $\mathcal{H}_{p-2}(e) = (0, h_2(e) - d(e)b_2, \cdots, h_p(e) - d(e)b_p)'$, $\mathcal{H}_1(e) = (h_1(e), d(e)b_2, \cdots, d(e)b_p)'$, $d(e) = b_2h_2(e) + \cdots + b_ph_p(e)$, and $H(e) = (h_1(e), \cdots, h_p(e))'$. 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 $b = (0, b_2, \cdots, b_p)$, there exists a unit vector $e_0(b) \in S_b^1$ such that $\mathcal{H}_1\{y_0(b)\} = 0$. 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 $p-2 = 2, 4 \cdots$, it follows from 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 to the equation.

Now, one needs to check Condition (A.1). For a continuous response variable Y, if $EZY < \infty$ 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 $\lim_{t\to\infty} g(\pm t) = \pm 1$ and $\mathrm{pr}(Y_1 > Y_2 | b'Z_1 > b'Z_2) < 1$ 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 $d(t; b) = b'E[Z\{Y - g(tb'Z)\}]$ defined in the beginning of this section with the expected value taken under the empirical distribution generated by the data $\{(Y_i, Z_i), i = 1, \cdots, n\}$.

Lastly, since $S(\beta)$ is monotone in β, it converges to $E\{S(\beta)\}$ uniformly in any compact set of β_0 in probability. It follows that $\hat{\beta}$ converges to β_0 in probability, as $n \to \infty$.

Appendix B: 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 β. Moreover, since $g(\cdot)$ is strictly increasing and Z is bounded, it follows from a uniform law of large numbers (Pollard, 1990, Chapter 8) that $\sup_{\beta \in \Omega} \left| \hat{D}(\beta) - E|Y^0 - \hat{Y}(\beta' Z^0)| \right|$ goes to 0, where Ω is a compact parameter set containing β_0. This, coupled with the convergence of $\hat{\beta}$ to β_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,

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

where $A(\beta)$ is defined by (2.6). Furthermore, it follows from a functional central limit theorem (Pollard, 1990, Chapter 10) that $n^{1/2}[\hat{D}(\beta) - E\{\hat{D}(\beta)\}]$, a process in β, converges weakly to a zero mean Gaussian process and thus is stochastic equi-continuous in β. This, coupled with the fact that $E\{\hat{D}(\beta)\}$ is differentiable in β and (B.1), implies that $n^{1/2}\{\hat{D}(\hat{\beta}) -$

$D_0\}$ is asymptotically equivalent to

$$
\begin{aligned}
& n^{1/2}\{\hat{D}(\beta_0) - D_0\} + E\{d(\beta_0)\}n^{1/2}(\hat{\beta} - \beta_0) \\
& \simeq n^{-1/2}\sum_{i=1}^{n}\Big(|Y_i - \hat{Y}(\beta_0' Z_i)| - D_0 + E\{d(\beta_0)\}[E\{A(\beta_0)\}]^{-1} Z_i\{Y_i - g(\beta_0' Z_i)\}\Big) \\
& = n^{-1/2}\sum_{i=1}^{n}\eta_i, \qquad \text{(B.2)}
\end{aligned}
$$

where $d(\beta)$ is defined by (2.7). Thus, $n^{1/2}\{\hat{D}(\hat{\beta}) - D_0\}$ converges in distribution to a zero-mean normal random variable. Moreover, $\hat{\eta}$ for the variance estimate (2.5) is obtained by replacing all the theoretical quantities for η in (B.2) with their empirical counterparts.

Appendix C: 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 $n^{-1/2}\sum_{i=1}^{n} I(\xi_i = k)\big\{|Y_i - \hat{Y}(\hat{\beta}'_{(-k)} Z_i)| - D_0\big\}$, where $\{\xi_i; i = 1, \cdots, n\}$ are n exchangeable discrete random variables uniformly distributed over $\{1, 2, \cdots, K\}$, independent of the data, and satisfy that $\sum_{i=1}^{n} I(\xi_i = k) = n/K, k = 1, \cdots, K$. It follows from Lemma 4.2 of Wellner & Zhan (1996) and the standard large sample expansion of a smooth estimating function that

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

Here and in the sequel, $o_p(\cdot)$ is with respect to the probability measure generated under $\{\xi_i, i = 1, \cdots, n\}$ and $\{(Y_i, Z_i), i = 1, \cdots, n\}$. Then using the same argument in Appendix B, one can show that $n^{1/2}\{\hat{D}_{(k)}(\hat{\beta}_{(-k)}) - D_0\}$ is asymptotically equivalent to

$$
n^{1/2}\{\hat{D}_{(k)}(\beta_0) - D_0\} + E\{d(\beta_0)\}n^{1/2}(\hat{\beta}_{(-k)} - \beta_0),
$$

which is asymptotically equivalent to $n^{-1/2}\sum_{i=1}^{n}\eta_{ki}$, where

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

It follows that $n^{1/2}(\hat{\mathcal{D}} - D_0) \simeq n^{-1/2}\sum_{i=1}^{n}(\sum_{k=1}^{K} K^{-1}\eta_{ki})$. Since $\sum_{k=1}^{K} I(\xi_i = k) = 1$ and $\sum_{k=1}^{K} I(\xi_i \neq k) = K - 1$, it is straightforward to show that

$$n^{-1/2}\sum_{i=1}^{n}\left(\sum_{k=1}^{K} K^{-1}\eta_{ki}\right) = n^{-1/2}\sum_{i=1}^{n}\Big(|Y_i - \hat{Y}(\beta_0' Z_i)| - D_0 + E\{d(\beta_0)\}[E\{A(\beta_0)\}]^{-1} Z_i\{Y_i - g(\beta_0' Z_i)\}\Big).$$

This implies that $n^{1/2}(\hat{\mathcal{D}} - D_0)$ is asymptotically equivalent to $n^{1/2}\{\hat{D}(\hat{\beta}) - D_0\}$.

Appendix D: Large Sample Properties of $\hat{\mathbb{D}}$

For a "random" cross-validation with $n_v = n/K$ and $n_t = n(K-1)/K$, where K is a positive integer, consider the random vector $\xi = (\xi_1, \cdots, \xi_n)'$ defined in Appendix C. It is straightforward to show that the random variable $\hat{\mathbb{D}}$ is equivalent to $E_\xi(\hat{\mathcal{D}})$, where the expectation is with respect to ξ only. Then, for the corresponding K-fold cross validation, it follows from Lemma 4.2 of Wellner & Zhan (1996) and the standard large sample expansion around $\hat{\beta}$ that

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

Furthermore, one can also show that $n^{1/2}\{\hat{D}_{(k)}(\hat{\beta}_{(-k)}) - \hat{D}(\hat{\beta})\}$ is asymptotically equivalent to

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

where

$$\eta_{ki}^* = I(\xi_i = k)K\left\{|Y_i - \hat{Y}(\hat{\beta}' Z_i)| - \hat{D}(\hat{\beta})\right\} + I(\xi_i \neq k)\frac{K}{K-1}E\{d(\beta_0)\}[E\{A(\beta_0)\}]^{-1} Z_i\{Y_i - g(\hat{\beta}' Z_i)\}.$$

It follows that $n^{1/2}\{\hat{\mathcal{D}} - \hat{D}(\hat{\beta})\} \simeq n^{-1/2}\sum_{i=1}^{n}(\sum_{k=1}^{K} K^{-1}\eta_{ki}^*)$. Since $\sum_{k=1}^{K} I(\xi_i = k) = 1$ and $\sum_{k=1}^{K} I(\xi_i \neq k) = K - 1$, $n^{-1/2}\sum_{i=1}^{n}\left(\sum_{k=1}^{K} K^{-1}\eta_{ki}^*\right) = 0$. This implies that $E\{n^{1/2}|\hat{\mathcal{D}} -$

$\hat{D}(\hat{\beta})|\} \to 0$. Therefore, for any $\epsilon > 0$,

$$\begin{aligned}\text{pr}\left(n^{1/2}|E_\xi\hat{\mathcal{D}} - \hat{D}(\hat{\beta})| > \epsilon\right) &\leq \text{pr}\left(E_\xi\left\{n^{1/2}|\hat{\mathcal{D}} - \hat{D}(\hat{\beta})|\right\} > \epsilon\right) \\ &\leq \epsilon^{-1}E_{(Y,Z)}\left[E_\xi\left\{n^{1/2}|\hat{\mathcal{D}} - \hat{D}(\hat{\beta})|\right\}\right] \to 0,\end{aligned}$$

where $E_{(Y,Z)}$ in the last term is the expectation with respect to $\{(Y_i, Z_i), i = 1, \cdots, n\}$. It follows that $n^{1/2}(\hat{\mathbb{D}} - D_0)$ 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**, 255–265.

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association* **87**, 738–754.

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 University Press.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association* **78**, 316–331.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association* **81**, 461–470.

Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association* **99**, 619–632.

Efron, B. & Tibshirani, R. (1995). Cross-validation and the bootstrap: Estimating the error rate of a prediction rule. Technical report, Stanford University. (http://www.stat.stanford.edu/ tibs/research.html).

Efron, B. & Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association* **92**, 548–560.

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 University Press.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression*. John Wiley & Sons.

Jacobsen, M. (1989). Existence and unicity of MLEs in discrete exponential family distributions. *Scandinavian Journal of Statistics* **16**, 335–349.

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

Molinaro, A., Simon, R. & Pfeiffer, R. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics* **21**, 3301–3307.

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. *Journal of the American Statistical Association* **88**, 486–494.

Shao, J. (1996). Bootstrap model selection. *Journal of the American Statistical Association* **91**, 655–665.

Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society, Series B: Methodological* **43**, 310–313.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics* **9**, 1135–1151.

Tibshirani, R. & Knight, K. (1999). Model search by bootstrap "bumping". *Journal of Computational and Graphical Statistics* **8**, 671–686.

Wellner, J. A. & Zhan, Y. (1996). Bootstrapping Z-estimators. Technical report, University of Washington.

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. *Journal of the American Statistical Association* **93**, 120–131.

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

	Age	Baseline RNA	RNA Change	Baseline CD4	CD4 Change
Estimate	-0.55	0.08	-12.06	0.03	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

Table 2. *Bias and square root mean square error (SMSE) for apparent error (AE), 10-fold cross-validation (CV_{10}), "random" cross-validation with $n_t = 2n/3$ (CV-1/3), and* 0.632 *bootstrap method*

		Bias				SMSE			
n	True error	AE	CV_{10}	CV-1/3	0.632	AE	CV_{10}	CV-1/3	0.632
				True model I & Fitted model I					
200	57.61	-1.62	0.23	0.57	0.02	3.45	3.16	3.19	3.11
400	57.23	-0.91	-0.01	0.17	-0.10	2.32	2.18	2.18	2.16
600	57.04	-0.60	0.00	0.12	-0.06	1.84	1.77	1.77	1.76
				True model II & Fitted model I					
200	58.35	-1.27	-0.04	0.19	-0.17	3.40	3.22	3.24	3.21
400	58.08	-0.64	-0.04	0.08	-0.10	2.34	2.27	2.27	2.26
600	58.04	-0.47	-0.06	0.01	-0.10	1.81	1.76	1.76	1.76
				True model I & Fitted Model II					
200	57.71	-1.61	0.27	0.63	0.07	3.47	3.20	3.25	3.15
400	57.31	-0.91	0.02	0.20	-0.07	2.35	2.22	2.21	2.19
600	57.13	-0.51	0.11	0.22	0.04	1.84	1.80	1.80	1.78
				True model II & Fitted model II					
200	57.81	-1.15	0.11	0.34	-0.03	3.26	3.13	3.14	3.10
400	57.49	-0.57	0.06	0.17	-0.01	2.28	2.24	2.24	2.22
600	57.47	-0.39	0.02	0.10	-0.02	1.82	1.80	1.79	1.79

Table 3. *Estimates of the regression parameters with their standard errors and corresponding p-values for testing zero covariate effects for the prostate cancer example*

	Age	Race	DCI	PSA	TV	GS	Nodule in Rectal Exam		
							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