Published in final edited form as:

Stat Biosci. 2016 October ; 8(2): 333–350. doi:10.1007/s12561-016-9147-y.

Locally Weighted Score Estimation for Quantile Classification in Binary Regression Models

John D. Rice and Jeremy M. G. Taylor

University of Michigan, Department of Biostatistics, 1415 Washington Heights, Ann Arbor, MI 48104, USA

Abstract

One common use of binary response regression methods is classification based on an arbitrary probability threshold dictated by the particular application. Since this is given to us *a priori*, it is sensible to incorporate the threshold into our estimation procedure. Specifically, for the linear logistic model, we solve a set of locally weighted score equations, using a kernel-like weight function centered at the threshold. The bandwidth for the weight function is selected by cross validation of a novel hybrid loss function that combines classification error and a continuous measure of divergence between observed and fitted values; other possible cross-validation functions based on more common binary classification metrics are also examined. This work has much in common with robust estimation, but differs from previous approaches in this area in its focus on prediction, specifically classification into high- and low-risk groups. Simulation results are given showing the reduction in error rates that can be obtained with this method when compared with maximum likelihood estimation, especially under certain forms of model misspecification. Analysis of a melanoma data set is presented to illustrate the use of the method in practice.

Keywords

binary classification; local likelihood; logistic regression; asymmetric loss; robust estimation

1 Introduction

This paper develops a class of robust estimators in binary regression models with the goal of increasing predictive accuracy at a given classification threshold. The motivation for this work is a situation arising in the practice of oncology in which it must be decided on the basis of some clinical variables whether or not to perform a sentinel lymph node biopsy in melanoma patients (Mocellin et al., 2009). Given a classification threshold $p^* \in (0, 1)$, the decision is made based on whether the predicted probability of metastasis in the lymph node is greater than p^* (perform biopsy) or less than p^* (do not perform biopsy). There are two types of errors possible in this setting: missing a metastatic cancer by not performing the biopsy (false negative) or performing an unnecessary biopsy (false positive). We assume throughout this paper that the ultimate objective of the analysis is classification of future subjects into high- and low-risk groups based on the threshold p^* , and that p^* is dictated by considerations specific to the application.

1.1 Background

Although a great deal of research has been done on the problem of predicting future binary outcomes in a population based on a sample from that population, most authors dealing primarily with classification (as opposed to regression) have focused on median classification, in which a positive response is predicted if the estimated response probability exceeds $p^* = 0.5$. In one of the few articles dealing with general quantile classification, Mease et al. (2007) propose an over- and undersampling method for boosted classification trees, where the relative amount of over- or undersampling is determined by the classification threshold. (It is from Mease et al., 2007, that we borrow the term “quantile” as used in this paper.) In other research, classification thresholds play a secondary role: Kordas (2006) discusses the use of the smoothed maximum score estimator of Horowitz (1992) for the binary choice model at quantiles other than the median. Both authors assume a latent continuous response variable that is dichotomized as the observed binary outcome. Wang et al. (2008) advocate estimation using an asymmetric version of the hinge loss function (Hastie et al., 2009), for which losses incurred by a false positive and false negative differ, in order to provide interval estimates of response probabilities. Support vector machines (SVM) are also a popular method for classification of binary responses. Dmochowski et al. (2010) point out that the case of unequal costs for false negatives and false positives may be dealt with by assigning class weights in the SVM procedure in accordance with these costs.

To introduce notation, suppose we observe a sample $(y_i, \mathbf{x}_i')$, $y_i \in \{0, 1\}$, $i = 1, \dots, n$, and assume a model in which each response is related to a vector of covariates $\mathbf{x}_i$ by

$P(Y=1|\mathbf{x}_i) = G(\mathbf{x}_i'\boldsymbol{\beta})$, $P(Y=0|\mathbf{x}_i) = 1 - G(\mathbf{x}_i'\boldsymbol{\beta}) \equiv \overline{G}(\mathbf{x}_i'\boldsymbol{\beta})$, where $G(\cdot)$ is a known link

function. We want to find an estimate $\hat{\boldsymbol{\beta}}$ such that $G(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$ is most accurate for $P(Y=1|\mathbf{x}_i)$ near p^* , but is not necessarily as accurate when $P(Y=1|\mathbf{x}_i)$ is not near p^* . In other words, we sacrifice global goodness of fit for improved local goodness of fit around p^* , where “local goodness of fit” is defined in equation (3) below.

1.2 Locally weighted score equation approach

Since we are interested in robustness to model misspecification away from p^* , we propose obtaining regression estimates by solving a set of locally weighted score equations, using a kernel-like weight function centered about p^* . There has been some research on robust estimation in binary models, but it has not generally focused on prediction. Pregibon (1982) recommends maximizing a modified log-likelihood function that “tapers” contributions with large deviance residuals. A thorough development of M-estimation in the context of logistic regression is given by Carroll and Pederson (1993). Ruckstuhl and Welsh (2001) analyze a form of misspecification familiar to the literature on robust estimation for continuous responses, in which the assumed probability mass function is “contaminated” by an unknown pmf.

In this paper, we consider the following form of misspecification: we fit an assumed model

$$P(Y=1|\mathbf{x}_i)=G(\mathbf{x}_i'\boldsymbol{\beta}), \quad (1)$$

while the true model is instead

$$p(\mathbf{x}_i) \equiv P(Y=1|\mathbf{x}_i)=G[\mathbf{x}_i'\boldsymbol{\beta}_0+Q(\mathbf{x}_i,\boldsymbol{\beta}_0)] \quad (2)$$

where Q is an unknown function of the covariates and $\boldsymbol{\beta}_0$. We make two assumptions regarding Q :

- i. $Q(\mathbf{x}_i, \boldsymbol{\beta}_0) \equiv Q_i = 0$ when $\mathbf{x}_i'\boldsymbol{\beta}_0 = \eta^*$, where $\eta^* = G^{-1}(p^*)$
- ii. $|Q_i|$ is increasing in $|\mathbf{x}_i'\boldsymbol{\beta}_0 - \eta^*|$.

In other words, we are supposing that at least around the threshold of interest, the assumed model is correctly specified, while it may differ from the true model when the linear predictor is not close to η^* .

It is important to note that the definition of model misspecification we give in equation (2) is more of a mathematical convenience than a reflection of what we believe to be the data-generating mechanism. We are then interested in achieving the best linear (in the sense that we are modeling $\mathbf{x}_i'\boldsymbol{\beta}$ rather than allowing nonparametric transformations of the predictor variables as in a generalized additive model) approximation of the model at the specified threshold, regardless of the true data-generating mechanism.

1.3 Metrics of local accuracy

We define a function that measures *local* accuracy,

$$\varepsilon(\mathbf{x}_i) = \mathbf{1}[p(\mathbf{x}_i) \geq G(\eta^*)] \mathbf{1}(\mathbf{x}_i'\hat{\boldsymbol{\beta}} < \eta^*) + \mathbf{1}[p(\mathbf{x}_i) < G(\eta^*)] \mathbf{1}(\mathbf{x}_i'\hat{\boldsymbol{\beta}} \geq \eta^*), \quad (3)$$

where $\mathbf{1}(\cdot)$ is the indicator function. We refer to $E[\varepsilon(\mathbf{X})]$ as the *local error rate* (LER), where an error occurs if the underlying true response probability $p(\mathbf{x}_i)$ is not on the same side of p^* that $G(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$ is. This does not correspond to the expectation of any loss function, implying that it cannot be estimated in practice, at least in our scenario with ungrouped binary data. It may be possible, however, to estimate (3) in the setting of Ruckstuhl and Welsh (2001) with grouped data. In this case, where $Y|\mathbf{x}_i$ is distributed binomially with probability $p(\mathbf{x}_i)$ and m_i is the number of trials, we would have two estimates of $p(\mathbf{x}_i)$: y_i/m_i and $G(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$. An estimate of $E[\varepsilon(\mathbf{X})]$ could then be constructed as

$$\hat{\epsilon} = \frac{1}{n} \sum_{i=1}^n \left\{ \mathbf{1} [y_i/m_i \geq G(\eta^*)] \mathbf{1} (\mathbf{x}_i' \hat{\boldsymbol{\beta}} < \eta^*) + \mathbf{1} [y_i/m_i < G(\eta^*)] \mathbf{1} (\mathbf{x}_i' \hat{\boldsymbol{\beta}} \geq \eta^*) \right\}. \quad (4)$$

For this estimator to be sensible, we would need both $n \rightarrow \infty$ and $m_i \rightarrow \infty$ for all $i = 1, \dots, n$. In this paper, however, we deal only with the case $m_i = 1$ for all observations, as this was the case for the data set motivating this work. The function in equation (3) is of use to us in that it constitutes a risk measure that we will attempt to minimize indirectly and by which we may evaluate the performance of our proposed method in simulation studies.

The structure of this paper is as follows. In Section 2, we describe the weighted score estimator and some of its properties. In Section 3, we suggest a method for choosing the degree of locality for the weight function. Section 4 presents a simulation study. Section 5 applies the method to a melanoma data set.

2 Weighted score estimation

2.1 Estimating equations

As stated previously, we wish to find locally accurate estimates of $\boldsymbol{\beta}_0$. As in Carroll and Pederson (1993), we obtain estimates of $\boldsymbol{\beta}$ by solving the weighted score equations

$$\mathbf{0} = \sum_{i=1}^n w_h(\mathbf{x}_i' \boldsymbol{\beta}) \mathbf{x}_i [y_i - G(\mathbf{x}_i' \boldsymbol{\beta})]. \quad (5)$$

Here, $w_h(\cdot)$ is some unimodal function, symmetric about η^* , that attains a maximum of 1 at $\eta = \eta^*$, as in Copas (1995). The parameter h determines the degree of locality of the model fit, and may be thought of as a bandwidth parameter (see Kordas, 2006, for a closely related use of the term “bandwidth”). We use the Gaussian kernel weight function

$$w_h(\eta) = \exp \left[-\frac{(\eta - \eta^*)^2}{2h^2} \right]. \quad (6)$$

Define $\hat{\boldsymbol{\beta}}_h$ as the solution to equation (5), which may be obtained using the Newton-Raphson algorithm. We refer to this as the weighted score estimator (WSE).

Intuitively, for an appropriate choice of h , the solution to equation (5) should be a more accurate estimate of the true parameter vector under the misspecified model (2), since for small values of h we will be downweighting the regions of the data that are not reflective of the linear part of (2), $\mathbf{x}_i' \boldsymbol{\beta}_0$. When $h \rightarrow \infty$, the solution to equation (5) corresponds to the maximum likelihood estimator (MLE) for the linear logistic model, since in this case $w_h(\eta) \rightarrow 1$ for all η .

The idea of downweighting observations inconsistent with an assumed model is not new, and has been considered in the context of binary response models by Pregibon (1982), Carroll and Pederson (1993), Copas (1988), and Ruckstuhl and Welsh (2001). However, none of these articles has addressed the robustness issue from the perspective of what we might call *locally correct* model specification, as measured by equation (3).

Supposing the model is correctly specified (i.e., $Q_i = 0$ for all i), consistency follows immediately for fixed h , as this estimator falls in the Mallows class (Carroll and Pederson, 1993) in which weights do not depend on the outcome. Moreover, we know from an application of standard robust estimation theory given by Carroll and Pederson (1993) that the asymptotic variance of the weighted score estimators is approximately

$$\mathcal{J}_n^{-1}(\beta_0) \mathbf{V}_n(\beta_0) \mathcal{J}_n^{-1}(\beta_0), \quad (7)$$

where $\mathcal{J}(\beta)$ and $\mathbf{V}(\beta)$ are, respectively,

$$\begin{aligned} \mathcal{J}_n(\beta) &= \frac{1}{n} \sum_{i=1}^n w_h(\mathbf{x}_i' \beta) \mathbf{x}_i \mathbf{x}_i' G^{(1)}(\mathbf{x}_i' \beta) \\ \mathbf{V}_n(\beta) &= \frac{1}{n} \sum_{i=1}^n w_h^2(\mathbf{x}_i' \beta) \mathbf{x}_i \mathbf{x}_i' G(\mathbf{x}_i' \beta) \overline{G}(\mathbf{x}_i' \beta) \end{aligned}$$

where a superscript (j) denotes the j th derivative of a function with respect to its argument. (See the Appendix for a development of the asymptotic distributions of the proposed estimators when $Q_i \neq 0$.)

2.2 Robustness and rationale of the method

Our method may be thought of as a generalization of the robust estimators of Carroll and Pederson (1993) in that our weight function $w_h(\cdot)$ depends on the classification threshold, allowing for improved accuracy at any p^* , while the weighting schemes they consider implicitly assume an interest in conventional median classification since they downweight observations with very high or low fitted probabilities (i.e., probabilities far from $p^* = 0.5$). We make the weight a function of the linear predictor rather than of the predicted probability so as to avoid issues regarding the restricted range of the mean of $Y|\mathbf{x}_i$, which for binary response models is (0, 1).

The motivation for our specific approach comes from Copas (1995), who maximizes a form of local likelihood for continuous models where locality is determined by proximity to some value t in the outcome space. However, in our binary setting, this is not appropriate, as it would result in only two weights, one for $y_i = 0$ and another for $y_i = 1$. Directly adapting this method to weight individual contributions to the log-likelihood instead by some function of $\mathbf{x}_i' \beta \in \mathbb{R}$ is also problematic, as (combined with the requirement for consistency of the resultant estimator) it leads to estimating equations formally identical to (5), but with a weight function that may be negative and is not maximized at η^* .

Alternatively, we may think of the contribution of a single observation to the objective function implied by the weighted score (5), given by

$$\tilde{l}_h(y_i, \mathbf{x}_i' \boldsymbol{\beta}) = \int_{y_i}^{G(\mathbf{x}_i' \boldsymbol{\beta})} w_h \left(G^{-1}(\pi) \right) \frac{y_i - \pi}{G^{(1)}(G^{-1}(\pi))} d\pi, \quad (8)$$

where G is assumed to be monotone, so that $\frac{d}{d\pi} G^{-1}(\pi) = 1/G^{(1)}(G^{-1}(\pi))$. Minimization of $-\sum_{i=1}^n \tilde{l}_h(y_i, \mathbf{x}_i' \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ results in the same estimators obtained by solving (5). The gain in robustness due to the WSE approach is apparent from Figure 1, which is a plot of $-\tilde{l}_h(y_i, \mathbf{x}_i' \boldsymbol{\beta}) / -\tilde{l}_h(1, \eta^*)$ versus $G(\mathbf{x}_i' \boldsymbol{\beta})$; in the usual maximum likelihood case ($h = \infty$) the loss is unbounded. By contrast, we see from this figure that for $h < \infty$, the maximum possible loss is finite (see Carroll and Pederson, 1993, for a discussion of the conditions for bounded influence functions in this class of estimators).

Furthermore,

$$\lim_{h \rightarrow 0} \frac{-\tilde{l}_h(y_i, \mathbf{x}_i' \boldsymbol{\beta})}{h} \propto y_i \mathbf{1}(\mathbf{x}_i' \boldsymbol{\beta} < \eta^*) + \frac{p^*}{1-p^*} (1-y_i) \mathbf{1}(\mathbf{x}_i' \boldsymbol{\beta} \geq \eta^*), \quad (9)$$

which is a weighted misclassification error. Thus the WSE is a compromise between the efficiency of the MLE and the robustness of a classification approach. This expression also provides some intuition as to the meaning of the threshold p^* . Specifically, the weight assigned to a false positive in (9) is $p^*/(1-p^*)$, reflective of the fact that as $p^* \rightarrow 0$, we are increasingly willing to accept false positives in order to avoid false negatives. This also implies that the relative cost of a false positive to a false negative is $p^*/(1-p^*)$, which may provide some guidance to practitioners in understanding the meaning of the particular threshold.

The objective function corresponding to the smoothed maximum score estimator of Horowitz (1992) and Kordas (2006) is shown in Figure 2 for purposes of comparison (these authors use $\tau = 1 - p^*$ to define the quantile of interest). Their objective function is

$$\frac{1}{n} \sum_{i=1}^n (y_i - p^*) K \left(\frac{\mathbf{x}_i' \boldsymbol{\beta}}{\sigma_n} \right), \quad (10)$$

where $K(\cdot)$ is an integrated kernel function (i.e., a cdf), $\sigma_n \sim n^{-1/5}$ is a bandwidth, and the maximizer of (10) gives the smoothed binary regression quantile estimator. As Kordas (2006) notes, maximizing (10) is equivalent to minimizing

$$\frac{1}{n} \sum_{i=1}^n \rho_{p^*} \left(y_i - K \left(\frac{\mathbf{x}_i' \boldsymbol{\beta}}{\sigma_n} \right) \right), \quad (11)$$

which is the loss depicted in Figure 2. Here, $\rho_{p^*}(u) = [\mathbf{1}(u \geq 0) - p^*]u$. To place this in our context, this means that a false positive ($\mathbf{x}_i' \boldsymbol{\beta} \geq 0, y_i = 0$) implies a loss of approximately p^* , while a false negative ($\mathbf{x}_i' \boldsymbol{\beta} < 0, y_i = 1$) implies a loss of approximately $1 - p^*$, with the approximation becoming exact as $\sigma_n \rightarrow 0$. This can be seen in the relative heights of the curves in Figure 2, which echo the patterns depicted for the objective functions associated with the weighted score estimators in Figure 1.

We point out here that there is an interesting connection between our method and the work of Kordas (2006), in which the objective function is smoothed by replacing an indicator function with a cdf, given here in equations (10) and (11). Clearly, the corresponding gradient will involve a pdf, which is similar in form to our weight functions apart from a scaling factor. Specifically, to maximize equation (10) (using the normal cdf as the kernel function K), we must solve

$$\begin{aligned} \mathbf{0} &= \frac{\partial}{\partial \boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^n (y_i - p^*) \Phi \left(\frac{\mathbf{x}_i' \boldsymbol{\beta}}{\sigma_n} \right) \\ &= \frac{1}{n \sigma_n} \sum_{i=1}^n \phi \left(\frac{\mathbf{x}_i' \boldsymbol{\beta}}{\sigma_n} \right) \mathbf{x}_i (y_i - p^*). \end{aligned} \quad (12)$$

Compare (12) with the weighted score equations (5), which we may rewrite as

$$\mathbf{0} = \sqrt{2\pi} \sum_{i=1}^n \phi \left(\frac{\mathbf{x}_i' \boldsymbol{\beta} - \eta^*}{h} \right) \mathbf{x}_i [y_i - G(\mathbf{x}_i' \boldsymbol{\beta})], \quad (13)$$

since we are using the Gaussian kernel given by (6).

3 Selection of bandwidth

A key element of the proposed method is the value of the parameter h , which is similar to the bandwidth of the smoothed indicator function in Kordas (2006) and Horowitz (1992). We may think of the selection of the bandwidth as a way of dealing with model misspecification locally: if, for a given threshold, the specified model is consistent with the truth globally, then h should be large, as we will be able to gain information from all observations regardless of distance from p^* . On the other hand, if the specified model is *not* consistent with the truth everywhere, then h should be small, in order to minimize the contamination of the model fit resulting from misspecification.

Likelihood cross validation is recommended by Eguchi and Copas (1998) as a way to choose h in local likelihood estimation. However, in their setting, the outcome is continuous and their weights are a function of the outcome itself rather than a monotone transformation of its estimated conditional mean (as in our case). Irizarry (2001) deals with a similar issue, and bases his selection on modified information criteria, i.e., functions that are based on the negative maximized log-likelihood plus a penalty term. However, his weights are solely functions of the covariates, with the unknown parameter playing no role.

These methods are primarily aimed at assessing overall goodness of fit, as opposed to the local goodness of fit in which we are interested. Likewise, alternative measures such as AUC, while commonly used for the evaluation of the predictive ability of binary response models, are not useful in our situation because they are designed to assess performance at all possible thresholds simultaneously. Although ideally we would like to select h to minimize

$$\frac{1}{n} \sum_{i=1}^n \varepsilon(\mathbf{x}_i), \quad (14)$$

where $\varepsilon(\mathbf{x}_i)$ is defined as in equation (3), we have seen that this quantity is not estimable from ungrouped binary data. Therefore, we shall attempt to minimize (14) indirectly. To do this, we recommend minimizing a cross-validated version of a hybrid error rate (see Wahba, 2002, for a discussion of a similar cost function designed to compromise between SVM and logistic regression).

Defining $\hat{\beta}_h^{(-i)}$ as the estimate of β obtained by leaving out the i th observation and solving (5), we choose h as

$$\hat{h} = \arg \min_{h > 0} -2 \sum_{i=1}^n \left[y_i \log G \left(\mathbf{x}_i' \hat{\beta}_h^{(-i)} \right) \mathbf{1} \left(\mathbf{x}_i' \hat{\beta}_h^{(-i)} < \eta^* \right) + (1 - y_i) \log \bar{G} \left(\mathbf{x}_i' \hat{\beta}_h^{(-i)} \right) \mathbf{1} \left(\mathbf{x}_i' \hat{\beta}_h^{(-i)} \geq \eta^* \right) \right]. \quad (15)$$

This error is a combination of classification error, from the terms $\mathbf{1} \left(\mathbf{x}_i' \hat{\beta}_h^{(-i)} < \eta^* \right)$ and $\mathbf{1} \left(\mathbf{x}_i' \hat{\beta}_h^{(-i)} \geq \eta^* \right)$; and deviance loss, from the terms $\log G \left(\mathbf{x}_i' \hat{\beta}_h^{(-i)} \right)$ and $\log \bar{G} \left(\mathbf{x}_i' \hat{\beta}_h^{(-i)} \right)$. Note that this is equal to the deviance (or minus twice the log likelihood) for all subjects with $\mathbf{1} \left(\mathbf{x}_i' \hat{\beta}_h^{(-i)} \geq \eta^* \right) \neq y_i$. Due to the presence of indicators this will be a piecewise constant function with jump discontinuities. However, as we will be using a grid search to find the optimal value of h , these discontinuities will present no numerical problems.

The flexibility of the method is such that we might also choose h to optimize any desired predictive metric, such as the negative predictive value (NPV),

$$\hat{h} = \arg \max_{h>0} \frac{\sum_{i=1}^n (1-y_i) \mathbf{1} \left(\mathbf{x}_i' \hat{\boldsymbol{\beta}}_h^{(-i)} < \eta^* \right)}{\sum_{i=1}^n \mathbf{1} \left(\mathbf{x}_i' \hat{\boldsymbol{\beta}}_h^{(-i)} < \eta^* \right)}, \quad (16)$$

or specificity,

$$\hat{h} = \arg \max_{h>0} \frac{\sum_{i=1}^n (1-y_i) \mathbf{1} \left(\mathbf{x}_i' \hat{\boldsymbol{\beta}}_h^{(-i)} < \eta^* \right)}{\sum_{i=1}^n (1-y_i)}. \quad (17)$$

In particular applications, these may be of greater interest than the form of local accuracy defined by (3).

4 Simulation study

4.1 Design

We examine the performance of the proposed method by generating data based on model (2). First, covariates X_{i1} , X_{i2} are generated iid standard normal. Then we obtain Z_{i1} , Z_{i2} as follows:

$$Z_{i1} = \frac{X_{i1}}{\sqrt{2}} - \frac{X_{i2}}{\sqrt{2}}, \quad Z_{i2} = \frac{X_{i1}}{\sqrt{2}} + \frac{X_{i2}}{\sqrt{2}}.$$

The linear predictor is

$$\eta_i = \eta^* + \omega_2 Z_{i2} \left[1 + \gamma \left(e^{-\omega_1 \text{sgn}(Z_{i2}) Z_{i1}} - 1 \right) \right], \quad (18)$$

the response is then generated as Bernoulli with mean $e^{\eta_i} / (1 + e^{\eta_i})$. The parameter γ is used to control the degree of departure from the linear logistic model outside of the region around η^* , with larger values leading to more nonlinearity in the true model (see Figure 3). For these simulations, $\gamma \in \{0, 0.1, 0.2, 0.3\}$, $\omega_1 = 1$, $\omega_2 = 2$.

This simulation design was chosen in part so that for a given p^* , changes in γ have minimal effect on the marginal mean of the outcome. At the same time, we believe that the resulting contours of constant probability shown in Figure 3 are realistic in the sense that we might encounter shapes of this sort in real data.

Two sample sizes were examined, $n \in \{200, 500\}$. In each case, one training sample was used for fitting the weighted score models and selecting h , while a separate validation sample of the same size was held out for calculation of error rates.

To find the optimal value of h for each data set, we use a grid search of (15), (16), and (17) at 50 equally spaced points of $h^{-1} \in [0, 2]$. We expect these metrics to give better performance than the cross-validated deviance (which we also include for purposes of comparison) because they explicitly depend on the threshold. Because of the computational expense of true leave-one-out cross validation, we instead use 5-fold cross validation. To further increase computational efficiency as well as numerical stability, when performing the cross validation, we do not fit a full model for each fold, but rather use a one-step Newton approximation.

We compare the proposed method with two existing methods which produce as fitted values binary estimates of $\mathbf{1}[E(Y|\mathbf{x}_i) \geq p^*]$. The weighted SVM approach of Dmochowski et al. (2010) assigns class weights as $1-p^*$ for the $y_i=1$ class and p^* for the $y_i=0$ class. We refer to this method as the weighted SVM (wSVM) method. The smoothed binary regression quantiles (SBRQ) method of Kordas (2006) with $\tau = 1 - p^*$ is a semiparametric regression procedure that makes minimal assumptions regarding the latent error distribution of the responses. Both methods produce fitted values of either 1 or 0, corresponding to estimates of whether or not the response probability exceeds p^* .

4.2 Results

Performance was evaluated by computing the error rate (14) on the validation samples for $h = \hat{h}$ and $h = \infty$, corresponding to the particular WSE model minimizing (15) and the MLE, respectively. Their averages across 1000 simulated data sets for each γ are displayed in Table 1.

We see from this table that using the hybrid error rate to select h leads to substantial improvement over the MLE in the LER, assuming that there is some degree of misspecification ($\gamma > 0$). There is an even greater improvement in performance of the WSE method relative to the wSVM, likely due to the fact that the true model is close to the specified parametric model in this case. This implies that the hybrid error rate is a fair surrogate for the LER.

There is relatively little difference, indeed, between the LER for models obtained using each of the CV functions we examined: use of the hybrid error rate, the NPV, or the specificity seems to result in similar performance gains over the MLE, to the extent that no single CV function can be recommended over the others. This points to another strength of our method: the investigators may choose a cross-validation function that best reflects their interests in a particular classification metric and expect to see similar improvements in the local error rate relative to competing methods. This is in contrast to using likelihood cross validation to select the bandwidth, which clearly does not provide any advantage over the MLE, as we would expect.

Also apparent from Table 1 is that for large degrees of misspecification, the SBRQ method slightly outperforms the WSE method; this difference becomes more pronounced with larger sample sizes. One possible explanation for its superior performance with larger γ is that SBRQ makes fewer parametric assumptions about the data. The wSVM method, by contrast, is always the worst performing of the methods we have examined.

Recall, however, that SBRQ and wSVM do not produce estimated probabilities, while the WSE method does. This means that it is possible with our proposed method to examine how close to the classification threshold any individual subject might be, ultimately providing more information to the investigator than do the competing methods. Additionally, when there is no misspecification, the SBRQ and wSVM methods perform very poorly relative to the MLE, while the WSE method loses little in comparison. This is a further advantage of our approach if we assume that in practice we are unlikely to encounter extremely misspecified models.

Table 2 gives a summary of the distribution of selected h values across the simulated data sets. The harmonic mean is shown because $h = \infty$ has a nonzero probability of being chosen. Using the cross-validated deviance to choose h leads to increasing choices of h with increasing γ , regardless of the classification threshold, which is likely the explanation for its failure to show improvement over the MLE in Table 1. This is in contrast to the case with each of the metrics we recommend for the selection of h , which do not exhibit this pattern. Similarly, the selected values of the bandwidth parameter for the method of Kordas (2006) show little variability with respect to changes in γ .

Additionally, it is clear from Table 2 that for the CV functions given by equations (15), (16), and (17), the chosen bandwidth is decreasing with sample size. This is a desirable feature of the methodology, as with larger samples we are in effect able to focus on a smaller region about the threshold, and thereby obtain larger gains in performance relative to the MLE.

5 Melanoma data analysis

In order to illustrate the use of the proposed method in practice, we apply it to a melanoma data set originally analyzed by Sabel et al. (2012). The data consist of the binary outcome and covariate values on $n = 2244$ melanoma patients who had undergone a sentinel lymph node biopsy (SLNB). The outcome is equal to 1 if metastasis was found on SLNB, 0 otherwise. The covariates are age, mitotic rate, Breslow depth, ulceration, regression, and high Clark level (defined as IV or V); the last three are binary, the first three continuous. Mitotic rate and Breslow depth were log transformed prior to model fitting.

One of the principal goals of the original analysis was to evaluate the predictive models of Mocellin et al. (2009). Two cutoff values were considered in Sabel et al. (2012) in the application of the logistic regression model estimated by Mocellin et al. (2009) to classification on a new data set, $p^* = \{0.1, 0.2\}$. In the current analysis, we apply the WSE methodology to this data at these same two thresholds. We chose $h^{-1} \in [0, 2]$ from a grid of 1000 equally spaced points according to either (15), (16), or (17) with 10-fold cross validation. In general, we recommend an equally spaced grid on $h^{-1} \in [0, 2]$ as a good

practical range for the grid search of h . This includes $h = \infty$ (i.e., the MLE) at the upper end of the range of possible bandwidths. It is harder to suggest bounds for the lower end, but we have found that it is useful to continue the grid search for smaller values of h until numerical problems are encountered.

We do not present a comparison of our results from this analysis with those of the competing methods examined in the simulation studies. In the case of SBRQ, this is due to numerical problems with fitting the model for our quantiles of interest. The simulation study suggested that we would not find much improvement with respect to LER by fitting the wSVM models, so they were omitted as well.

Table 3 gives a summary of the chosen h values and the effect of the WSE method on classification for this data set. We see improvements with the WSE fits versus the MLE in each of the metrics for which we have conducted cross validation, especially in the case of $p^* = 0.1$, where we are able to improve the negative predictive value from 0.935 to 0.947 and the specificity from 0.093 to 0.140. For $p^* = 0.2$ we still see improvements over the MLE, although the gains in this case are less pronounced.

Another aspect of the analysis in which we are able to improve on the MLE is in the number of negative predictions (NNP), which Mocellin et al. (2009) cited as an important part of the intended use of their predictive models. Again, this seems to be much more dramatic for $p^* = 0.1$ than $p^* = 0.2$. Of course, by itself, NNP is fairly useless as an indicator of a model's usefulness, but given that we are also improving on a certain predictive metric simultaneously, an increase in NNP means fewer patients are subjected to unnecessary biopsies, which was one goal of the original analysis.

The NNP also suggests a reason for the difference in the analysis between the two thresholds considered. Specifically, nearly half of the data is classified into the low-risk group at this threshold regardless of fit (MLE or WSE). This indicates that the “center” of the data in some sense lies in this region. Indeed, the marginal probability of a positive biopsy result for this data set is approximately 25%, so that the WSE methodology, when used with thresholds in the neighborhood of 0.25, will not have much room for improvement over the MLE.

6 Discussion

In this paper, we have developed the weighted score estimation framework, which constitutes a robust estimation method drawing inspiration from the local likelihood approach of Copas (1995), in order to improve quantile classification for parametric binary response models. Although semiparametric methods such as that of Kordas (2006) are available for such problems, we have found that in practical situations, their computational complexity can make them unattractive, so there seems to be a need for a more numerically stable (if somewhat less flexible) approach such as ours.

We have dealt theoretically with model misspecification of a particular type, namely data generated according to (2) while fitting a model of form (1). This has allowed us to demonstrate the ability of the method to reduce contamination of the model fit due to

deviations from the assumed model away from the threshold of interest p^* . We have also proposed a novel loss function in equation (15) as a way to assess predictive accuracy when interest is in classification at an arbitrary threshold p^* , and have shown through simulation studies the increased accuracy with respect to the LER metric that is possible when an *a priori* classification threshold is incorporated into the parameter estimation procedure.

It might not be immediately apparent why we are not bypassing the bandwidth selection problem altogether and simply minimizing the hybrid error rate to estimate β . First, as we see from equation (9), as $h \rightarrow 0$, solving the locally weighted score equations is in fact minimizing a weighted misclassification error rate. Second, many loss functions commonly used in the evaluation of binary response models result in predicted probabilities being forced to 0 or 1 when optimized directly (Mease et al., 2007). One advantage of our approach is that virtually any measure of predictive accuracy can be used to select h while maintaining the interpretability of the model parameters, so that, for example, a researcher interested in maximizing the NPV of a model (e.g., Mocellin et al., 2009) may use that as the cross-validation function instead of the hybrid error rate that we have introduced in this paper.

Resolution of numerical instability occurring when attempting to fit WSE models with small h is one possible avenue for future research. In many of the simulated data sets, the cross-validation function was still decreasing at the smallest h for which the fitting algorithm converged. This indicates that the chosen values of h might be biased upwards in some cases, and this in turn could explain the lack of a decreasing trend in the chosen values of h with increasing degrees of model misspecification. We have conjectured that some values of h for which the model fitting procedure fails might correspond essentially to eliminating too much of the data, so a weight function with heavier tails (such as the Cauchy kernel) could be helpful in this regard.

Extensions to non-binary data could prove to be an interesting direction for future research. With grouped binary data, for example (i.e., with $m_i > 1$), we would be able to use equation (4) for estimation of the local error rate and selection of the bandwidth h and thereby examine issues such as the effect of group sizes in evaluating the accuracy of the LER or the performance of the LER as a criterion for bandwidth selection. Alternatively, with, say, count data, the work of Copas (1995) might be more directly applicable. What distinguishes our method is that the weights are based on proximity of the predicted mean of the outcome variable to p^* , while for Copas (1995), weights are based on proximity of the observed outcome to the point of interest. The issue is the degree of discreteness in the data: with only two possible outcomes, binary data would offer only two weights in the framework of Copas (1995), while our method allows the weights to vary continuously.

However, as Copas (1995) considers primarily continuous data, only alluding to the discrete case in passing, it might be interesting to pursue such an extension for, say, Poisson data. Especially if the Poisson mean is small, then the degree of discreteness mentioned above would be high, and it would be necessary to use a method based on proximity of the linear predictor (rather than the observed outcome) to η^* . Such an extension could be relatively simple, as all that would need to be changed is the form of the estimating equations (5). A

more difficult element of the adaptation to another outcome distribution, however, would be the selection of the bandwidth. However, with non-binary data, it may be easier to measure misspecification of the model, perhaps using the methods of Konishi and Kitagawa (1996) to derive an information criterion appropriate to our interest in local estimation.

Acknowledgments

This research was supported in part by National Institutes of Health grant T32 CA-83654. The authors also thank Michael Sabel for providing the melanoma sentinel lymph node biopsy data.

References

- Carroll RJ, Pederson S. On robustness in the logistic regression model. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1993; 55:693–706.
- Copas JB. Binary regression models for contaminated data. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1988; 50:225–265.
- Copas JB. Local likelihood based on kernel censoring. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1995; 57:221–235.
- Dmochowski JP, Sajda P, Parra LC. Maximum likelihood in cost-sensitive learning: Model specification, approximations, and upper bounds. *Journal of Machine Learning Research*. 2010; 11:3313–3332.
- Eguchi S, Copas JB. A class of local likelihood methods and near-parametric asymptotics. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1998; 60:709–724.
- Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer; 2009.
- Horowitz JL. A smoothed maximum score estimator for the binary response model. *Econometrica*. 1992; 60:505–531.
- Irizarry RA. Information and posterior probability criteria for model selection in local likelihood. *Journal of the American Statistical Association*. 2001; 96:303–315.
- Konishi S, Kitagawa G. Generalised information criteria in model selection. *Biometrika*. 1996; 83:875–890.
- Kordas G. Smoothed binary regression quantiles. *Journal of Applied Econometrics*. 2006; 21:387–407.
- Lehmann, EL. *Elements of Large Sample Theory*. Springer; 2004. corrected edition
- Mease D, Wyner AJ, Buja A. Boosted classification trees and class probability/quantile estimation. *Journal of Machine Learning Research*. 2007; 8:409–439.
- Mocellin S, Thompson JF, Pasquali S, Montesco MC, Pilati P, Nitti D, Saw RP, Scolyer RA, Stretch JR, Rossi CR. Sentinel node status prediction by four statistical models: results from a large bi-institutional series ($n = 1132$). *Annals of Surgery*. 2009; 250:964–969. [PubMed: 19953714]
- Pregibon D. Resistant fits for some commonly used logistic models with medical applications. *Biometrics*. 1982; 38:485–498. [PubMed: 7115876]
- Ruckstuhl AF, Welsh AH. Robust fitting of the binomial model. *Annals of Statistics*. 2001; 29:1117–1136.
- Sabel MS, Rice JD, Griffith KA, Lowe L, Wong SL, Chang AE, Johnson TM, Taylor JMG. Validation of statistical predictive models meant to select melanoma patients for sentinel lymph node biopsy. *Annals of Surgical Oncology*. 2012; 19:287–293. [PubMed: 21822550]
- Wahba G. Soft and hard classification by reproducing kernel Hilbert space methods. *Proceedings of the National Academy of Sciences*. 2002; 99:16524–16530.
- Wang J, Shen X, Liu Y. Probability estimation for large-margin classifiers. *Biometrika*. 2008; 95:149–167.

Appendix: Asymptotic distribution of WSEs under misspecification

This Appendix presents our derivation of the asymptotic distribution of the WSE under this form of model misspecification. We assume the usual regularity conditions involving existence of derivatives, expectations, and limits in probability (for details, see, e.g., Lehmann, 2004). In order to render the true model easier to work with on the probability scale, we approximate $P(Y=1|\mathbf{x}_i)$ by a Taylor expansion about the model we are fitting:

$$\begin{aligned} P(Y=1|\mathbf{x}_i) &= G(\mathbf{x}_i'\boldsymbol{\beta}_0 + Q_i) \\ &= G(\mathbf{x}_i'\boldsymbol{\beta}_0) + Q_i G^{(1)}(\mathbf{x}_i'\boldsymbol{\beta}_0) + o(Q_i^2) \\ &\approx G(\mathbf{x}_i'\boldsymbol{\beta}_0) + Q_i G^{(1)}(\mathbf{x}_i'\boldsymbol{\beta}_0). \end{aligned} \quad (19)$$

The expectation of the score will not be $\mathbf{0}$ when $Q_i \neq 0$ for all i , implying inconsistency of the estimators.

After using a Taylor expansion about $\boldsymbol{\beta}_0$ to approximate the estimating equations, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \approx \left[-\frac{1}{n} \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^n w_h(\mathbf{x}_i'\boldsymbol{\beta}_0) \mathbf{U}_i(\boldsymbol{\beta}_0) \right]^{-1} \left[\sqrt{n} \frac{1}{n} \sum_{i=1}^n w_h(\mathbf{x}_i'\boldsymbol{\beta}_0) \mathbf{U}_i(\boldsymbol{\beta}_0) \right]. \quad (20)$$

We note that

$$\left[-\frac{1}{n} \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^n w_h(\mathbf{x}_i'\boldsymbol{\beta}_0) \mathbf{U}_i(\boldsymbol{\beta}_0) \right]^{-1} \xrightarrow{p} \mathcal{J}^{-1}(\boldsymbol{\beta}_0) \quad (21)$$

by the weak law of large numbers and continuous mapping;

$$\mathcal{J}_n(\boldsymbol{\beta}) \approx \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' G^{(1)}(\mathbf{x}_i'\boldsymbol{\beta}) \left[w_h(\mathbf{x}_i'\boldsymbol{\beta}) - Q_i w_h^{(i)}(\mathbf{x}_i'\boldsymbol{\beta}) \right]$$

is the finite-sample analogue of $\mathbf{J}(\boldsymbol{\beta})$. Likewise, by the central limit theorem,

$$\sqrt{n} \left[\frac{1}{n} \sum_{i=1}^n w_h(\mathbf{x}_i'\boldsymbol{\beta}_0) \mathbf{U}_i(\boldsymbol{\beta}_0) - \mathbf{S}_n(\boldsymbol{\beta}_0) \right] \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{T}(\boldsymbol{\beta}_0)), \quad (22)$$

where

$$\begin{aligned}\mathbf{S}_n(\boldsymbol{\beta}) &\approx \frac{1}{n} \sum_{i=1}^n w_h(\mathbf{x}_i' \boldsymbol{\beta}) \mathbf{x}_i Q_i G^{(1)}(\mathbf{x}_i' \boldsymbol{\beta}) \xrightarrow{P} \mathbf{S}(\boldsymbol{\beta}), \\ \mathbf{T}_n(\boldsymbol{\beta}) &\approx \frac{1}{n} \sum_{i=1}^n w_h^2(\mathbf{x}_i' \boldsymbol{\beta}) \mathbf{x}_i \mathbf{x}_i' \left[G^{(1)}(\mathbf{x}_i' \boldsymbol{\beta}) + Q_i G^{(2)}(\mathbf{x}_i' \boldsymbol{\beta}) \right] \xrightarrow{P} \mathbf{T}(\boldsymbol{\beta}).\end{aligned}$$

Here we have made use of a further Taylor expansion to approximate the variance of $Y|\mathbf{x}_i$ under misspecification. Therefore, by Slutsky's theorem, we have from (21) and (22) that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{L}} N\left(\mathcal{J}^{-1}(\boldsymbol{\beta}_0)\mathbf{S}(\boldsymbol{\beta}_0), \mathcal{J}^{-1}(\boldsymbol{\beta}_0)\mathbf{T}(\boldsymbol{\beta}_0)\mathcal{J}^{-1}(\boldsymbol{\beta}_0)\right). \quad (23)$$

Note that due to the approximation in (19) as well as the further approximation for $\text{Var}(Y|\mathbf{x}_i)$, (23) is also approximate, even asymptotically. The accuracy of these approximations of course depends on the magnitude of Q_i ; that is, on the degree to which the true model deviates from the assumed model. The advantage to the use of these approximations is that they allow us to give very simple expressions for the quantities involved in the asymptotic distribution of the WSEs under misspecification.

Specifically, for example, the approximations for $\mathbf{J}_n(\boldsymbol{\beta})$ and $\mathbf{S}_n(\boldsymbol{\beta})$ show that the magnitude of the bias will depend on the proportion of observations with $p(\mathbf{x}_i)$ near p^* : since we have assumed that $|Q_i|$ increases with $|\mathbf{x}_i' \boldsymbol{\beta}_0 - \eta^*|$, the bias will be greater when the distribution of the covariates is such that many observations have a true response probability far from p^* . This makes intuitive sense, but these asymptotic formulae give analytical support to such intuition and provide explicit quantitative descriptions of the local accuracy of the weighted score estimators.

Figure 1.

This figure depicts the contribution of a single observation to the implicit objective functions used in weighted score estimation, scaled such that the loss is 1 at $p^* = 0.2$, $y = 1$.

Specifically, the vertical axis is $-\tilde{l}_h(y_i, \mathbf{x}_i'\boldsymbol{\beta}) / -\tilde{l}_h(1, \eta^*)$. These are obtained by numerical evaluation of equation (8). The x -axis of each panel corresponds to a predicted probability for one observation. The curves shown represent the loss or cost associated with that predicted probability for either $y = 1$ or $y = 0$ when using equation (5) to obtain estimates of $\boldsymbol{\beta}$.

Figure 2.

This figure depicts the contributions of a single observation to the objective functions of Kordas (2006) for $\tau = 1 - p^* = 0.8$. These curves correspond to equation (11), the minimizers of which are known as the smoothed binary regression quantile (SBRQ) estimators, and are scaled such that the loss is 1 at $p^* = 0.2$, $y = 1$.

Figure 3.

Contours of constant probability for simulations, $p^* = 0.25$. Solid lines are the true model; dashed lines represent the linear logistic model for reference, although it is only correct at the thick solid line, $E(Y|x_j) = p^*$.

Table 1

Simulation results: LER (14) averaged over 1000 simulations; for a single observation this is given by equation (3). With the exception of the column labeled Dev., which refers to likelihood cross validation, the cross-validation functions to choose h are based on equations (15), or the hybrid error rate (HER); (16), the negative predictive value (NPV); and (17), the specificity (Spec.).

n	p^*	γ	CV functions (WSE)						wSYM
			MLE	Dev.	HER	NPV	Spec.	SBRQ	
200	0.10	0.0	0.058	0.060	0.062	0.062	0.061	0.103	0.101
		0.1	0.062	0.062	0.062	0.062	0.062	0.088	0.101
		0.2	0.077	0.074	0.073	0.070	0.070	0.082	0.103
		0.3	0.087	0.083	0.080	0.078	0.078	0.082	0.104
	0.25	0.0	0.046	0.046	0.048	0.048	0.048	0.072	0.074
		0.1	0.047	0.046	0.048	0.048	0.048	0.064	0.069
		0.2	0.056	0.054	0.054	0.053	0.053	0.063	0.073
		0.3	0.065	0.061	0.060	0.057	0.057	0.060	0.075
	0.50	0.0	0.039	0.040	0.042	0.042	0.043	0.066	0.065
		0.1	0.042	0.042	0.043	0.043	0.043	0.057	0.062
		0.2	0.048	0.047	0.046	0.046	0.046	0.052	0.064
		0.3	0.053	0.051	0.050	0.049	0.048	0.051	0.065
500	0.10	0.0	0.039	0.040	0.042	0.042	0.041	0.065	0.072
		0.1	0.047	0.045	0.046	0.045	0.045	0.056	0.070
		0.2	0.064	0.059	0.055	0.054	0.055	0.054	0.072
		0.3	0.080	0.075	0.066	0.063	0.066	0.053	0.074
	0.25	0.0	0.029	0.030	0.032	0.032	0.031	0.047	0.052
		0.1	0.034	0.033	0.033	0.033	0.032	0.039	0.051
		0.2	0.046	0.043	0.041	0.040	0.039	0.038	0.052
		0.3	0.058	0.055	0.049	0.047	0.048	0.038	0.055
	0.50	0.0	0.024	0.025	0.026	0.027	0.026	0.039	0.043
		0.1	0.029	0.028	0.029	0.028	0.028	0.034	0.042
		0.2	0.040	0.038	0.036	0.035	0.034	0.031	0.043
		0.3	0.049	0.046	0.042	0.041	0.040	0.030	0.045

Simulation results: optimal bandwidths, the harmonic mean of the chosen h across 1000 simulations. For the WSE method, each column heading indicates the cross-validation function optimized to choose h with the exception of the column labeled Dev., which refers to likelihood cross validation, the cross-validation functions to choose h are based on equations (15), or the hybrid error rate (HER); (16), the negative predictive value (NPV); and (17), the specificity (Spec.). For the SBRQ method, this is calculated based on Theorem 2 in Horowitz (1992).

Table 2

n	p*	γ	CV functions (WSE)					SBRQ
			Dev.	HER	NPV	Spec.		
200	0.10	0.0	2.587	2.102	1.855	1.891	0.561	
		0.1	2.699	1.986	1.938	1.911	0.553	
		0.2	3.019	1.926	1.833	1.999	0.554	
	0.25	0.3	3.343	1.963	1.900	2.073	0.547	
		0.0	2.441	1.632	1.457	1.482	0.562	
		0.1	2.580	1.633	1.450	1.456	0.566	
	0.50	0.2	2.740	1.654	1.519	1.570	0.572	
		0.3	2.965	1.650	1.533	1.530	0.584	
		0.0	2.462	1.626	1.443	1.244	0.562	
	0.10	0.1	2.690	1.532	1.401	1.229	0.582	
		0.2	2.936	1.509	1.455	1.243	0.588	
		0.3	2.890	1.589	1.468	1.288	0.595	
500	0.10	0.0	2.299	1.504	1.541	1.709	0.472	
		0.1	2.664	1.550	1.617	1.808	0.467	
		0.2	3.280	1.476	1.614	1.808	0.461	
	0.25	0.3	4.003	1.473	1.598	1.861	0.462	
		0.0	2.347	1.335	1.419	1.361	0.456	
		0.1	2.684	1.447	1.423	1.411	0.461	
	0.50	0.2	3.065	1.393	1.405	1.423	0.469	
		0.3	3.849	1.403	1.394	1.462	0.475	
		0.0	2.526	1.354	1.368	1.184	0.454	
	0.10	0.1	2.802	1.399	1.366	1.262	0.467	
		0.2	3.030	1.339	1.362	1.244	0.480	
		0.3	3.745	1.407	1.438	1.266	0.493	

This table summarizes the results of the WSE analysis of the melanoma data set. The two values of p^* given here correspond to the values considered by Sabel et al. (2012). Three cross-validation functions were examined: the negative predictive value (NPV), specificity, and the hybrid error rate (although these results are not given in this table as h was chosen to be ∞ ; that is, for that metric, the MLE was deemed to be the best fitting model). The table gives the value of each of these metrics for both the linear logistic model ($h = \infty$) as well as the optimal model (where h is given by the “Chosen bandwidth” column). The final two columns give the number of negative predictions (NNP) associated with each model fit.

Table 3

CV function	p^*	Chosen bandwidth	Metric		NNP		
			$h = \infty$	$h = \hat{h}$	$h = \infty$	$h = \hat{h}$	$h = \hat{h}$
NPV	0.1	1.44	0.935	0.947	168	168	208
	0.2	0.76	0.850	0.853	999	999	995
Specificity	0.1	1.07	0.093	0.140	168	168	252
	0.2	1.13	0.505	0.512	999	999	1013