Shrinkage, cross validation calibration, and the construction of prognostic indices

Werner Vach

Center for Data Analysis and Model Building &
Institute of Medical Biometry and Medical Informatics,
University of Freiburg,
Albertstr. 26-28, D- 79104 Freiburg, Germany
phone: +49 761 202 1178
fax: +49 761 202 0503
email: wv@fdm.uni-freiburg.de

Summary

The use of shrinkage methods for the construction of prognostic indices has been paid increasing attention in the literature on medical statistics in the last years. One approach for the construction of a shrinkage factor is cross validation calibration as suggested by van Houwelingen and le Cessie (1990). We investigate this approach in more detail. First we try to clarify, why shrinkage factors constructed by cross validation calibration tends to be smaller than 1. Second we explain, why use of this shrinkage factor results in an improvement of the average prediction error. Third we argue that nevertheless such a shrinkage cannot be recommended for the construction of prognostic indices. Fourth we explain that selection bias resulting from the application of model building procedures is not necessarily an argument in favour of shrinkage methods. Finally we conclude that it is necessary to rethink the use of loss functions like the average prediction error in constructing prognostic indices.

Keywords: Average prediction error, Generalized linear models, Logistic regression, Model building, Prediction, Selection bias, Regression to the mean

1. Introduction

The construction of prognostic indices plays an important role in many fields of medical research. In the case of a binary event to be prognosed in a given and fixed time interval, e.g. recurrence of a disease or success of a therapy, estimates of the probability of this event given individual values of prognostic factors can provide a prognostic index for each patient. Such estimates can be obtained from a logistic regression model (Cox 1970). If numerical quantities like number of episodes, decrease of blood pressure or time until success should be prognosed, other members of the class of generalized linear models (McCullagh & Nelder 1989) can be used to define prognostic indices, for example Poisson, Gaussian or exponential models.

In statistical terms, the construction of a prognostic index corresponds to the construction of a predictor. In recent years the evaluation and construction of predictors have been paid increased attention in the literature on medical statistics. A key paper was written by van Houwelingen and le Cessie (1990) discussing a variety of topics. Among other considerations they proposed a class of shrinkage methods to improve predictors resulting from fitting generalized linear models with respect to the average prediction error. They suggest to estimate the necessary shrinkage factor by "cross validation calibration". We investigate this approach in more detail and try to explain why shrinkage factors determined by cross validation calibration tend to be smaller than 1 and why shrinkage methods tend to improve the average prediction error. In spite of this property we argue that the proposed shrinkage method is not adequate in constructing prognostic indices, because the improvement in the average has to be paid by large deteriorations for some patients. We further sketch out, why in using model building procedures the well known phenomenon of selection bias does not necessarily provides an argument supporting the use of shrinkage methods, if the aim is the construction of a prognostic index.

2. Notation

In this paper we focus on prediction, i.e. we want to predict an outcome variable Y based on a p-dimensional covariate vector X. We assume that we have a sample S of independent realizations $(Y_i, X_i)_{i \in S}$ to construct a predictor. Let us first forget this construction step, such that a predictor μ is nothing else than a real-valued function on the sample space of X. We are interested in the behaviour of the predictor μ for future observations, which we denote by (Y^*, X^*) to distinguish them from the observations in the sample, and for which we assume in the sequel that they have the same distribution as the observations in the sample. A first basic measure for the accuracy of the predictor μ is the expected squared deviation between the true Y^* and the "prognosis" $\mu(x)$ for a single patient with covariate vector $X^* = x$, i.e.

$$IPE(\mu, x) := E_F[(Y^* - \mu(x))^2 | X^* = x] \ .$$

IPE is the abbreviation for individual prediction error, and the subscript F should remind us that the expectation is computed with regard to future observations. A corresponding global measure is the average of the expected squared deviation over all future individuals, i.e.

$$APE(\mu) := \int IPE(\mu, x) dF^{X^*}(x) = E_F(Y^* - \mu(X^*))^2 \ .$$

Once we have decided for a specific rule to construct a predictor, we cannot continue to regard it as a fixed function, but we have to regard it as a random variable $\hat{\mu}$ depending on the sample S. To judge the quality of such a predictor $\hat{\mu}$ we are interested in the sampling properties with respect to the two statistics *IPE* and *APE* introduced above. Especially we may ask for the expected values of these statistics, i.e. for

$$EIPE(\hat{\mu}, x) := E_S IPE(\hat{\mu}, x) = E_S E_F[(Y^* - \hat{\mu}(X^*))^2 | X^* = x]$$

and

$$EAPE(\hat{\mu}) := E_S APE(\hat{\mu}) = E_S E_F(Y^* - \hat{\mu}(X^*))^2 \ ,$$

where the subscript S should remind us that the expectation is now taken with respect to the distribution of the sample S.

A well known measure for the quality of a numerical parameter estimate is its mean square error (MSE), i.e. the expected squared difference between the estimate and the true parameter. A predictor can be regarded as a parameter estimate for the conditional expectation of Y given $X = x$, which we denote in the following by $\mu_0(x)$. Then we can consider the mean square error at the individual level, i.e.

$$IMSE(\hat{\mu}, x) = E_S(\hat{\mu}(x) - \mu_0(x))^2$$

and in the average over future observations, i.e.,

$$AMSE(\hat{\mu}) = \int IMSE(\hat{\mu}, x) dF^{X^*}(x) = E_F E_S(\hat{\mu}(X^*) - \mu_0(X^*))^2 \ .$$

These quantities are closely related to the expected prediction errors, especially we have

$$\begin{array}{lcllll} EIPE(\hat{\mu}, x) & = & IMSE(\hat{\mu}, x) & + & \sigma^2(x) & \text{and} \\ EAPE(\hat{\mu}) & = & AMSE(\hat{\mu}) & + & \sigma^2 & \end{array}$$

where $\sigma^2(x) = Var_F(Y^* | X^* = x)$ and $\sigma^2 = E_F(\sigma^2(X^*))$, which both are independent of the choice of $\hat{\mu}$. Hence decreasing the expected individual prediction error (*EIPE*) by choosing an alternative predictor $\hat{\hat{\mu}}$ is equivalent to decreasing the individual mean square error (*IMSE*) and decreasing the expected average prediction error (*EAPE*) is equivalent to decreasing the average mean square error (*AMSE*).

In comparing two predictors $\hat{\hat{\mu}}$ and $\hat{\mu}$ one may not want to rely on a comparison of the single expectations, but may prefer a direct comparison. This can be done by considering the probabilities to obtain a better prediction using $\hat{\hat{\mu}}$ instead of $\hat{\mu}$. Such probabilities can be defined on the individual level by

$$P^{IPE}_{imp}(\hat{\hat{\mu}}, \hat{\mu}, x) = P_S(IPE(\hat{\hat{\mu}}, x) < IPE(\hat{\mu}, x)) \ .$$

On the population level we can ask, how often $\hat{\hat{\mu}}$ provides us with a predictor with a smaller average prediction error than $\hat{\mu}$, i.e. we can consider

$$P^{APE}_{imp}(\hat{\hat{\mu}}, \hat{\mu}) = P_S(APE(\hat{\hat{\mu}}) < APE(\hat{\mu})) \ .$$

We may also ask for the average probability of an individual improvement, which results in the measure

$$AP_{imp}(\hat{\hat{\mu}}, \hat{\mu}) := \int P^{IPE}_{imp}(\hat{\hat{\mu}}, \hat{\mu}, x) dF^{X^*}(x) = E_F P^{IPE}_{imp}(\hat{\hat{\mu}}, \hat{\mu}, X^*) \ .$$

Note that $P^{IPE}_{imp}(\hat{\hat{\mu}}, \hat{\mu}) = P_S(|\hat{\hat{\mu}}(x) - \mu_0(x)| < |\hat{\mu}(x) - \mu_0(x)|)$, i.e. the probability of an improvement in the prediction error at the individual level is nothing else than the probability to be closer to the true expectation for the single individual. Similar $AP_{imp}(\hat{\hat{\mu}}, \hat{\mu})$ is the probability that $\hat{\hat{\mu}}(X^*)$ is closer to $\mu_0(X^*)$ than $\hat{\mu}(X^*)$, i.e. that for a randomly chosen future individual the prediction based on $\hat{\hat{\mu}}$ is closer to the true conditional expectation of Y than the prediction based on $\hat{\mu}$.

As a particular class of predictors we consider those arising from fitting generalized linear models. Here the relationship between Y and X is described by the specification of conditional mean and variance; with a prespecified inverse link function ν and a variance function V we assume

$$\begin{aligned} E[Y|X=x] &= \mu_\beta(x) := \nu(\eta(\beta, x)) \text{ with } \eta(\beta, x) := \beta_0 + \sum_{j=1}^{p} \beta_j x_j \text{ and} \\ Var[Y|X=x] &= \phi V(\mu_\beta(x)) \ . \end{aligned}$$

The estimate $\hat{\beta}$ for the regression parameter β is obtained by solving the estimation equation

$$\sum_{i \in S} \frac{\partial}{\partial \beta} \mu_\beta(X_i) V^{-1}(\mu_\beta(X_i))(Y_i - \mu_\beta(X_i)) = 0 \ .$$

Then we can define for any possible value x of X the ordinary predictor

$$\hat{\mu}_{ORD}(x) := \mu_{\hat{\beta}}(x) = \nu(\hat{\eta}(x)) \text{ with } \hat{\eta}(x) := \eta(\hat{\beta}, x) \ .$$

As special cases we will refer to logistic regression for binary Y with $\nu(z) = (1 + exp(-z))^{-1}$, $V(z) = z(1-z)$, and $\mu_\beta(x) = P_\beta(Y = 1|X = x)$ and to Gaussian regression with $\nu(z) = z$ and $V(z) = 1$. In the latter case estimation of β corresponds to ordinary least squares regression.

In addition to some analytic results we will also report the results of some numerical investigations of the above quantities with respect to logistic regression. We always consider the situation of 10 binary, balanced and independent covariates and a sample size of 150. We consider only regression parameter vectors β where the intercept is chosen such that the overall expectation of Y is 0.5. We approximate the expectation and probabilities with respect to S defined above by averages and relative frequencies of a Monte Carlo simulation with 1000 repetitions. Integration with respect to the distribution of X^* is carried out exactly by summing up over all possible 1024 values.

3. Shrinking predictors by cross validation calibration

In the case of ordinary least squares regression we can express the ordinary predictor $\hat{\mu}_{ORD}$ as $\bar{Y} + \sum_{j=1}^{p}(x_j - \bar{X}_j)\hat{\beta}_j$. Here $\bar{Y}$ is the sample mean of the outcome variable and $\bar{X}_j$ is the sample mean of the jth covariate. Van Houwelingen and le Cessie (1990) suggest to shrink this ordinary predictor in dependence on a shrinkage factor c. The resulting predictor is then defined by

$$\hat{\mu}_c(x) = \bar{Y} + c\sum_{j=1}^{p}(x_j - \bar{X}_j)\hat{\beta}_j \ .$$

For generalized linear models they suggest to use the analogon

$$\hat{\mu}_c(x) = \nu(\beta_0 + \sum_{j=1}^{p}\bar{X}_j\hat{\beta}_j + c\sum_{j=1}^{p}(x_j - \bar{X}_j)\hat{\beta}_j)$$

which coincides with the first definition for ordinary least squares regression, because then $\nu(z) = z$ and $\bar{Y} = \beta_0 + \sum_{j=1}^{p}\bar{X}_j\hat{\beta}_j$. Furthermore they suggest to estimate the shrinkage factor c by "cross validation calibration" which is performed in the following manner: For each individual i of the sample S the estimate $\hat{\beta}_{-i}$ based on all individuals except i is computed. Next for each individual the linear combination $\eta_i := \eta(\hat{\beta}_{-i}, X_i)$ is computed. Then a generalized linear model is fitted to the observations $(Y_i, \eta_i)_{i\in S}$ with the values of η_i as a single covariate. The regression coefficient $\hat{c}$ obtained is then taken as the shrinkage factor, resulting in a new predictor

$$\hat{\mu}_{HC}(x) := \hat{\mu}_{\hat{c}}(x) \ .$$

Van Houwelingen and le Cessie (1990) conjecture that this predictor is preferable to the ordinary predictor with respect to the average prediction error. They also mentioned the related work of Copas (1983) considering the same type of predictors but determining shrinkage factors in a different manner.

4. Why shrinkage factors based on cross validation calibration tend to be smaller than 1

To answer this question one has to realize that in regressing Y_i on η_i one tries to estimate the function

$$g(h) = E_F(Y^*|\hat{\eta}(X^*) = h)$$

i.e. the conditional expectation of a future Y^* given $\hat{\eta}(X^*)$. In the example of a prognostic index based on logistic regression this would be the probability of the event of interest given the estimated value of the prognostic index without applying the transformation ν. Precisely in cross validation calibration we are estimating a slightly different function, because g depends on the sample size and the distribution of X^* and both are slightly changed in the cross validation loop, but this difference is negligible for our considerations.

At first sight one may expect that $g(h)$ is equal to $\nu(h)$. In reality, however, it is a function which tends to be smaller than $\nu(h)$ if h is large and which tends to be larger than $\nu(h)$ if h is small. This phenomenon is nothing else than the well known phenomenon of "regression to the mean" dating back to Galton (1886). In comparing the height of parents and their (adult) siblings he observed that parents larger than the population mean tend to have siblings smaller than their parents, and that parents smaller than the population mean tend to have siblings larger than their parents. Whereas Galton tried to use the height of the parent to predict the height of a sibling and observed regression to the mean, in cross validation calibration one tries to predict an outcome based on the value of a prognostic index and one observes some shrinkage.

To explain this relationship further, let us try to describe Galton's example by some formulas. Let the random variables H_p and H_s denote the height of the parent and the sibling. Furthermore there exists a latent variable H representing the genetic determination of height, which is shared by the parent and the sibling. Hence we have $H_p = H + \epsilon_p$ and $H_s = H + \epsilon_s$, with H, ϵ_p, and ϵ_s independent. If σ^2 is the variance of H in the total population, and σ_p^2 and σ_s^2 are the variances of ϵ_p and ϵ_s, we have $Var(H_p) = \sigma^2 + \sigma_p^2$, $Var(H_s) = \sigma^2 + \sigma_s^2$ and $Cov(H_p, H_s) = \sigma^2$. Now Galton considers the prediction of H_s by H_p, i.e. the function $f(h) = E[H_s|H_p = h]$. Now if we assume that H, ϵ_p, and ϵ_s are all Gaussian, the joint distribution of H_s and H_p is a bivariate normal distribution and $f(h)$ is a linear function. Especially we have $f(h) = h_0 + \alpha(h - h_0)$, where h_0 denotes the expectation of both H_s and H_p and α satisfies

$$\alpha = \frac{Cov(H_p, H_s)}{Var(H_p)} = \frac{\sigma^2}{\sigma^2 + \sigma_p^2} < 1 \ . \tag{1}$$

This implies that we have $E[H_s|H_p = h] < h$ if $h > h_0$ and $E[H_s|H_p = h] > h$ if $h < h_0$. If the assumption of Gaussian distributions cannot be maintained, $f(h)$ need not be linear. However, as far as the distribution of H_p is unimodal, it still holds (Das & Mulder 1983) that $f(h) - h$ is a function changing the sign at a value h_0, which is in general not the mean, but the mode of the distribution of H_p.

In our case of prognostic indices and future observations (Y^*, X^*) the role of H_p is played by $\hat{\mu}_{ORD}(X^*)$ and Y^* plays the role of H_s. The latent variable H is now the true conditional expectation $\mu_0(X^*)$, such that we have $Y^* = \mu_0(X^*) + \epsilon_{Y^*}$ and $\hat{\mu}_{ORD}(X^*) = \mu_0(X^*) + \epsilon_{\hat{\mu}_{ORD}}$. Here ϵ_{Y^*} is the error term in the true regression model and $\epsilon_{\hat{\mu}_{ORD}}$ is the error of the predictor resulting from the randomness of the sample S. Of course in general we cannot rely on the assumption of normality and hence have to refer to the general results of Das and Mulder (1983). These allow to conclude that we have $E_F[Y^*|\hat{\mu}_{ORD}(X^*) = m) < m$ if $m > m_0$ and $E_F[Y^*|\hat{\mu}_{ORD}(X^*) = m) > m$ if $m < m_0$, if we assume that the distribution of $\hat{\mu}_{ORD}(X^*)$, i.e. the distribution of the prognostic index in the total population, is unimodal with mode at m_0. Now it remains only to note that $\hat{\mu}_{ORD}(X^*) = \nu(\hat{\eta}(X^*))$ and that ν is strictly monotone, which implies $g(h) < \nu(h)$ if $\nu(h) > m_0$ and $g(h) > \nu(h)$ if $\nu(h) < m_0$. Coming back to cross validation calibration we try to fit the function $g(h)$ by a function of the form $\nu(c_0 + ch)$ and hence the resulting shrinkage factor $\hat{c}$ tends to be smaller than 1. Similar considerations can be found in the paper of Copas (1983).

Although we hope that the parallels drawn so far convince the reader, there may be left some irritation. The above results seem to imply that for example in the case of a binary outcome variable Y a patient with a known prognostic index of 0.8 has only perhaps a probability of 0.7 to show the event. Does this indicate a substantial flaw of our prognostic index? The answer is no. The truth is the following: If we see a patient with a covariate vector x, we are interested in his or her probability to show the event of interest given his or her individual covariate values. And the answer to this question is given by the prognostic index. Once we have this information, it is again true, that conditional only on the information provided by his or her value of the prognostic index, the probability of an event can be quite different. However, it makes no sense to ask this question for the individual patient, because we have the information on his or her covariates. Considering the conditional expectation of the outcome given only the value of the prognostic index gives an answer to the question, how all patients with this value of the prognostic index behave in the average in the future. This may be important if we want to predict the future behaviour of a predictor in a population, but not for the single patient (cf. Vach 1997).

The above considerations allow us to determine the two most important factors with impact on the magnitude of the shrinkage factor. Using (1) we can conclude that the shrinkage factor is closely related to the quantity $\sigma^2_{\mu_0}/(\sigma^2_{\mu_0}+\sigma^2_{\hat{\mu}_{ORD}})$, where $\sigma^2_{\mu_0}$ denotes the population variance of $\mu_0(X^*)$ and $\sigma^2_{\hat{\mu}_{ORD}}$ is the variance of the predictor. Neglecting that in reality $\sigma^2_{\hat{\mu}_{ORD(x)}}$ is varying for different x we can conclude that the shrinkage factor tends to be small if $\sigma^2_{\mu_0}$ is small or if $\sigma^2_{\hat{\mu}_{ORD}}$ is large. A small $\sigma^2_{\mu_0}$ corresponds to the situation that none of the covariates has a great influence. Large values of $\sigma^2_{\hat{\mu}_{ORD}}$ indicate that the predictions are not very stable. Hence one can conclude that a small value of a shrinkage factor computed by cross validation calibration in a particular application indeed indicates that the properties of our predictor may be not very satisfactory.

5. Why shrinkage methods tend to decrease the average prediction error

To give an answer to this question we first consider the simple situation of a scalar parameter θ and a (nearly) unbiased estimator $\hat{\theta}$. Then we can consider the estimates

$$\hat{\theta}_c := (1-c)k + c\hat{\theta} = k + c(\hat{\theta}-k)$$

where k is an arbitrary constant. The bias of $\hat{\theta}_c$ is equal to $(1-c)(k-\theta)$ and its variance is c^2V where V is the variance of $\hat{\theta}$. Hence

$$MSE(\hat{\theta}_c) := c^2V + (1-c)^2(\theta-k)^2 \ .$$

This function is always increasing at $c = 1$, i.e., if we choose a c slightly smaller than 1, the *MSE* decreases. The optimal c minimizing the *MSE* is $c_{opt} = \frac{(\theta-k)^2}{(\theta-k)^2+V}$ and the minimal value of $MSE(\hat{\theta}_c)/MSE(\hat{\theta})$ is equal to c_{opt}. This implies that we can improve any unbiased estimator by shrinkage, if we take the *MSE* as the risk function of interest. These arguments can be easily extended to the case where we have a second (biased) estimator $\hat{\theta}_2$ with a variance smaller than that of $\hat{\theta}$. We then consider $\hat{\theta}_c := (1-c)\hat{\theta}_2 + c\hat{\theta}$, where again at least values of c slightly smaller than 1 tends to decrease the *MSE*.

In the case of prediction the ordinary predictor $\hat{\mu}_{ORD}(x)$ based on fitting a generalized linear model is (nearly) unbiased for $\mu_0(x)$. A biased estimator with a small variance is the overall sample mean $\bar{Y}$. Hence predictors of the form

$$\hat{\mu}_c(x) = (1-c)\bar{Y} + c\hat{\mu}_{ORD}(x) = \bar{Y} + c(\hat{\mu}_{ORD}(x) - \bar{Y})$$

have the potential to decrease the *IMSE* for all x if c is chosen smaller than 1. In the case of ordinary least squares regression the proposal of van Houwelingen and le Cessie

(1990) is indeed exactly of this type, and in the general case similar arguments can be applied to $\hat{\eta}(x)$ and $\hat{\eta}(\bar{X})$. In their proposal c is chosen as a constant; if we look at the above formulas for c_{opt} we have to conclude, that the optimal c depends on x and is approximatively equal to

$$c(x) := \frac{(\mu_o(x) - EY)^2}{(\mu_o(x) - EY)^2 + Var(\hat{\mu}_{ORD}(x))} \ .$$

The values of $c(x)$ may highly vary, especially $c(x)$ is equal to 0 for patients with $\mu_0(x)$ equal to the population mean of Y and it tends to 1 for patients with $\mu_0(x)$ far away from the population mean. Hence with c fixed we can decrease the *IMSE* for all patients only for values of c very close to 1. However, to decrease the *AMSE* it may be sufficient to decrease the *IMSE* for a majority of patients. Hence we have a wide range of values of $c < 1$ with an decrease of the *AMSE* comparing $\hat{\mu}_c$ with $\hat{\mu}_{ORD}$. In the last section we have shown that cross validation calibration tends to produce shrinkage factors smaller than 1, which have hence the potential to decrease the *AMSE*, and decreasing the *AMSE* implies decreasing the expected average prediction error.

We do not want to state that it is only a random coincidence that cross validation calibration tends to produce shrinkage factors which are smaller than 1 as well as are able to improve the average prediction error. In general there are possible (small) values of a potential shrinkage factor which lead to an increase of the *AMSE*, and hence not any method producing numbers smaller than 1 defines a good shrinkage procedure. Some numerical computations presented in the next section will show that shrinkage factors determined by cross validation calibration have the potential to decrease the *AMSE*, and hence cross validation calibration is indeed useful. Additionally, at an intuitive level cross validation calibration seems to be related to the solution of a reasonable optimization problem, namely to find that c which minimizes the *AMSE* or *EAPE* within the class $(\hat{\mu}_c)_{c\in[0,1]}$ of predictors. To understand this, let us first consider the situation of Gaussian regression. It is well known that the *EAPE* of $\hat{\mu}_c$ can be estimated using cross validation by

$$\widehat{EAPE}(\hat{\mu}_c) = \sum_{i\in S}(Y_i - \hat{\mu}_c^{-i}(X_i))^2 \ ,$$

where $\hat{\mu}_c^{-i} := \bar{Y}_{-i} + c(\eta_i - \bar{Y}_{-i})$ and $\bar{Y}_{-i}$ is the sample mean of Y with the i-th observation left. Hence a naive proposal is to chose that c minimizing $\widehat{EAPE}(\hat{\mu}_c)$, as already suggested by Stone (1974). Now minimizing $\widehat{EAPE}(\hat{\mu}_c)$ means to minimize

$$\sum_{i\in S}(Y_i - \bar{Y}_{-i} - c(\eta_i - \bar{Y}_{-i}))^2 \ ,$$

i.e. to perform a regression of the differences between the outcome and the cross validated global mean on the difference between the cross validated prediction and the cross validated global mean. Van Houwelingen and le Cessie (1990) propose a direct regression of the outcome on the cross validated prediction, i.e. minimizing

$$\sum_{i \in S} (Y_i - c_o - c\eta_i)^2 \ .$$

If we slightly change the first problem by replacing the first $\bar{Y}_{-i}$ by $\bar{Y}$ and the second one by $\bar{\eta}$, the two problems would have the same solution for c; without this change the problems have different solutions for c, but they are still of comparable magnitude.

In the general case one may still try to minimize

$$\sum_{i \in S} (Y_i - \hat{\mu}_c^{-i}(X_i))^2 \ . \tag{2}$$

Cross validation calibration is instead related to minimizing

$$\sum_{i \in S} l(Y_i, \hat{\mu}_c^{-i}(X_i)) \tag{3}$$

where l denotes a (quasi)-loglikelihood. However, the behaviour of $l(y,m)$ is usually very similar to $(y-m)^2$, and hence minimizing (3) can be regarded as a close approximation to minimizing (2). For example for logistic regression we have $l(y,m) = -\log(1 - |y-m|)$, and the comparison with $(y-m)^2$ in Figure 1 shows, that there is only a great difference for large deviations between y and m. One advantage of using (3) or cross validation calibration, respectively, is that it can be generalized to the construction of prognostic indices derived from the Cox proportional hazard model allowing the use of censored survival data (Verweij and van Houwelingen 1993).

6. Shrinkage, average prediction error and individual prediction error

In the last two sections we have tried to explain why $\hat{\mu}_{HC}$ has some potential to reduce the expected average prediction error in comparison to the ordinary predictor $\hat{\mu}_{ORD}$. Indeed, in Table 1 we demonstrate that the expected average prediction error can

Table 1 about here

be improved by using $\hat{\mu}_{HC}$ in the special case of logistic regression. In the computations we made use of the proposal of van Houwelingen and le Cessie (1990) to use a first order approximation of β_{-i} derived by Pregibon (1981) to facilitate cross validation calibration. We observed for a situation with 10 independent, binary and balanced covariates and a sample size of 150 for four different choices of the regression parameters shrinkage factors with an expectation about 0.6 and an improvement with respect to the average mean square error, the probability of improving the average prediction error as well as the average probability of an individual improvement.

With respect to the construction of prognostic indices this implies that shrinkage based on cross validation calibration can be useful to define indices with a better behaviour in the average over all future patients. However, this improvement in the average does not coincide with an improvement for each patient. This we try to illustrate in Table 2. We

Table 2 about here

selected some arbitrary covariate vectors x representing the prognostic factor values of different patients. We consider two choices of the regression parameter vector and for both $\mu_0(x)$, the true conditional probability of an event for the patient, varies between 0.5 and 0.9; for covariate vectors x with $\mu_0(x)$ smaller than the population mode $m_0 = 0.5$ results can be deduced by symmetry. We first investigate the bias of $\hat{\mu}_{ORD}(x)$ and $\hat{\mu}_{HC}(x)$. The ordinary predictor is always nearly unbiased, whereas – due to the shrinkage – the absolute value of the bias of $\hat{\mu}_{HC}(x)$ increases with increasing distance between $\mu_0(x)$ and m_0. (Due to the curvature of the logistic link function the bias decreases for extreme values of $\mu_0(x)$). As an example for the possible bias we look at one result for the first parameter constellation considered: A patient with a covariate vector $x = (1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1)$ and a true conditional probability of 0.864 for an event has to expect a prognostic index of 0.761. This large bias also implies for this patient an increase of the *IMSE* by a factor of 3 and consequently the probability to achieve a worse prediction error in using $\hat{\mu}_{HC}$ than $\hat{\mu}_{ORD}$ is nearly 70%. Overall we observe that the *IMSE* is increased for patients with $\mu_0(x) > 0.7$ or $\mu_0(x) < 0.3$ and that a decrease occurs only for patients with $0.3 < \mu_0(x) < 0.7$. As for the considered constellations the vast majority of patients belongs to this interval we observe a decrease in the average. But this improvement has to been paid by some patients with rather distinct deteriorations, resulting from a distinct bias. These are just those patients with a relatively low or high risk (compared to the population mode m_0), for which in some applications a precise estimation of a success probability is of special importance.

7. Model building, selection bias and shrinkage

The concept of shrinkage can applied not only to predictors based on fitting a single regression model. Van Houwelingen and le Cessie (1990) already suggest to compute shrinkage factors also for prognostic indices based on the results of a model building procedure. In practical applications it is difficult to fix the complete model building procedure in advance, because it involves many steps of selecting variables, categorising variables, selecting interactions or selecting higher order terms. Hence application of cross validation calibration is not feasible and van Houwelingen and le Cessie (1990) therefore advise the use of a split sample approach.

Recently Schumacher, Holländer and Sauerbrei (1997) discuss an additional argument in favour for the use of shrinkage methods following the application of model building procedures. It is well known that regression parameter estimates resulting from variable selection procedures suffer from selection bias (Miller 1984). This means that once a variable is selected in a model, the regression parameter estimate tends to overestimate the true parameter in absolute value. Now one can regard shrinkage methods like that considered in this paper also as shrinkage methods for regression parameter estimates, because they are equivalent to shrinking each regression coefficient $\hat{\beta}_j$ by the shrinkage factor $\hat{c}$. From this point of view application of shrinkage may contribute to the reduction of selection bias.

Even if this line of arguments may be true, can we conclude that shrinkage may also contribute to a reduction of the bias of prognostic indices? In order to give an answer first we have to investigate the bias of predictors resulting from the application of model building procedures. In a small investigation we consider the widespread used method of backward elimination for logistic regression with nominal selection levels of $\alpha = 5\%$ and $\alpha = 15\%$. We denote the resulting regression coefficients with $\hat{\beta}_j^{BE,\alpha}$ (which are set to zero if the variable X_j is not selected), and the resulting predictor with $\hat{\mu}_{BE,\alpha}$. For some selected values of the covariate vector x the bias of $\hat{\mu}_{BE,\alpha}(x)$ is shown in Table 3. We observe that for patients with a covariate vector x such that $\mu_0(x)$ is larger than m_0

Table 3 about here

the predictor almost always underestimates the true conditional probability of an event. For example the patient considered already above with a true conditional probability of 0.864 has to expect a value of the prognostic index of 0.741 if backward elimination with a nominal level of 5% is used, and of 0.809 if the level is 15%.

To understand this phenomenon let us assume that any X_j is centred around zero, such that selection of variables has little impact on the intercept. Now $\mu_0(x) > m_0$ implies usually that many of the terms $\beta_j x_j$ are positive and few negative. Variable selection procedures tend to set some of the β_j to zero, and hence we have more positive than negative terms $\beta_j x_j$ and this results in an decrease of the predictor compared to the ordinary predictor. The reality is of course a little bit more complex, because the size of β_j is related to the probability of selecting the jth covariate. This is illustrated by the contradiction in the lower part of Table 3. If $\beta = (-2.25, 1.5, 1.2, 0.9, 0.6, 0.3, 0, 0, 0, 0, 0)$ and $x = (1, 1, 0, 0, 0, 0, 0, 0, 0, 0)$ then in spite of $\mu_0(x) = 0.611 > 0.5 = m_0$ the predictions $\hat{\mu}_{BE,\alpha}(x)$ overestimate $\mu_0(x)$. Here the covariates X_3, X_4 and X_5 with $\beta_j(x_j - \bar{X}_j)$ negative are much more often removed from the model than the covariates X_1 and X_2 with $\beta_j(x_j - \bar{X}_j)$ positive. In spite of such contradictions we have the approximative rule that $\hat{\mu}_{BE,\alpha}(x)$ underestimates $\mu_0(x)$ if $\mu_0(x) > m_0$ and vice versa. Hence the bias of the predictors suggest to explode the single regression coefficients rather than to shrink them. This is related to the fact, that selection bias considers the conditional expectation of $\hat{\beta}_j^{BE,\alpha}$ given the jth covariate is selected, whereas the bias of the predictors relies on the unconditional expectation of $\hat{\beta}_j^{BE,\alpha}$. Both are compared in Table 4, and indeed, the conditional expectations indicate overestimation of the absolute value, whereas the unconditional expectations indicate underestimation.

Table 4 about here

One may argue, that the bias of the predictors will imply that shrinkage factors will tend to be larger than 1 and hence correct for this bias. This is not true, we have computed in our simulations also the shrinkage factors determined by cross validation calibration for $\hat{\mu}_{BE,\alpha}$ with repeating the model building process within each step of the cross validation loop. For the first parameter constellation considered in Table 4 the expected values are 0.47 and 0.51 for $\alpha = 5\%$ and $\alpha = 15\%$, respectively, for the second constellation the corresponding values are 0.70 and 0.67. These small values, especially for the first constellation, can be explained by the large variability of $\hat{\mu}_{BE,\alpha}(x)$ shown also in Table 3, which is especially much higher than that of $\hat{\mu}_{ORD}(x)$, and shrinkage factors mainly reflect this variance.

8. Conclusions

We have tried to clarify, why the shrinkage method proposed by van Houwelingen and le Cessie (1990) has the potential to decrease the expected average prediction error in comparison to the ordinary predictors derived from fitting generalized linear models. First, due to the phenomenon of regression to the mean, the suggested cross validation calibration procedure results in shrinkage factors tending to values smaller than 1. Second, any unbiased estimator can be improved with respect to the the expected squared deviation from the true parameter by taking a weighted combination with another biased, but more stable estimate, if the weight of the unbiased estimate is from a range of values smaller than 1. In their proposal the second argument can be applied to the majority of future patients regarding predictors as estimates of the conditional expectation of the outcome variable. This results in an improvement of the expected squared deviation in the average, and hence in an improvement of the average prediction error.

For the application in the construction of prognostic indices it is important to note that this improvement in the average has to be paid for by some patients with a substantial bias, which implies a substantial increase of the expected deviation and a high probability of a worse prediction. It is highly questionable whether this is acceptable in medical applications where it must be the aim to make good decisions for each patient. One may object that it is impossible to decrease the average prediction error without increasing the individual prediction error for some patients. This question needs further clarifications and the objection may be acceptable if one cannot identify those patients for which prediction suffers from the change to an alternative of $\hat{\mu}_{ORD}$. However, for the approach of van Houwelingen and le Cessie (1990) it is obvious that patients with an expected outcome far away from the population mean are most affected, and we can identify them by their values of the prognostic index.

As a general consequence in the construction of prognostic indices one should be cautious to rely on an alternative to the ordinary predictor $\hat{\mu}_{ORD}$, if it is only justified by an improvement of the average prediction error. We always need additional investigations to check whether such an improvement has to be paid for by some patients with an unacceptable deterioration. For example, two recently proposed alternative shrinkage methods (Breiman 1995, Tibshirani 1996) seem to be not very adequate for the construction of prognostic indices from this point of view. Both approaches result in variable specific shrinkage factors $\hat{c}_j$ to be multiplied with the ordinary estimates $\hat{\beta}_j$ from fitting a generalized linear model. Each $\hat{c}_j$ tends to be between 0 and 1, hence again we can conclude that the largest change of the predictor occurs for patients with a covariate vector x such that $\mu_0(x)$ is far away from the population mean, because for this patients either the majority of the values $\beta_j(x_j - \bar{X}_j)$ is positive or the majority is negative. In contrast to the approach of van Houwelingen and le Cessie both approaches have the potential to set some effect estimates to 0, hence they make also a contribution to the variable selection problem and perhaps to reduction of selection bias. However, our considerations with respect to variable selection have shown, that correction of selection bias and construction of good predictors simultaneously is somewhat a contradiction. Hence further investigations have to clarify, to which kind of problems these proposals can really contribute.

There exists also a further argument in favour for using $\hat{\mu}_{ORD}$ compared to alternatives promising a better average prediction error. In constructing prognostic indices usually we use data from clinical investigations, and the patients selected in these studies are a small sample from a more or less well defined total population. Nevertheless we want to use the predictor for all patients in the total population. Now inclusion and exclusion criteria of clinical studies often depend on important prognostic factors like age or performance status, such that the total population and the sample differ with respect to the distribution of the covariates. The basic property of $\hat{\mu}_{ORD}$ to give a nearly unbiased estimate of the conditional expectation does not suffer from such a covariate-dependent selection, so far we can assume that the regression model in the total population is roughly the same as in the sample. However, superiority of an alternative predictor with respect to the average prediction error for future patients behaving like the sample need not hold with respect to the average prediction error for future patients from the total population. For example, the superiority of the predictor proposed by van Houwelingen and le Cessie (1990) can be destroyed in the total population, if we have here a higher fraction of patients with a conditional expectation far away from the sample mean of the conditional expectation. This will be typically the case if patients with unfavourable prognostic factor values like high age or low performance status are excluded from the study.

So far our critique of the average prediction error concerns the averaging over patients, which is contradictory to the general aim of medical science to provide optimal decisions for each patient. Another point not investigated in this paper is the question, whether a squared deviation is an appropriate loss function in the construction of prognostic indices. Prognostic indices are often used to classify patients to identify high risk or low risk patients with important implications like choice of therapy. In such a setting control of classification errors seems to be the main goal. It is worth to note that decreasing the expected average prediction error need not to decrease classification errors. A simple example is given by the predictor proposed by van Houwelingen and le Cessie (1990), if we consider a classification scheme based on the prognostic index with cut-points depending on the sample, e.g. a classification into a lower, medium and upper third. Then, in spite of the improvement in the average prediction error, there is no change in the classification error, because the new predictor is a monotone transformation of the old one, and hence the classification of patients does not change.

There has been a variety of proposals for new methods to construct prognostic indices in the last years; besides shrinkage methods we may mention generalized additive models (Hastie and Tibshirani 1990), neural nets (Hecht-Nielsen 1990, Hertz et al. 1991) or classification and regression trees (Breiman et al. 1984). Their investigations on usefulness in medical settings (Liestøl et al. 1994, Faraggi and Simon 1995, Marshall et al. 1994, Schmoor et al. 1993) and even their widespread application, especially of neural nets (e.g. Baxt 1991, Kippenhan et al. 1992, Ravdin et al. 1992, Kappen and Neijt 1993), indicate that there is some discontent with existing methods. On the other side, there has been little discussion on loss functions suitable in medical applications. The main measures for the quality of prognostic indices in theoretical considerations as well as in simulations are measures like classification error rates and average prediction errors, allowing deteriorations for some patients to be compensated by improvements for other patients. Here a further rethinking on the aims of prognostic indices in medical applications seems to be necessary.

Acknowledgments

The author is indebted to his colleagues at the Institute of Medical Biometry and Medical Informatics in Freiburg for a lot of discussions on the pros and cons of shrinkage methods. Especially he thanks Kirstin Schulz and Martin Schumacher for comments on an earlier draft of this paper.

References

Baxt, W. G. (1991). Use of an artificial neural network for the diagnosis of myocardial infarction. *Annals of Internal Medicine* 115, 843.848

Breiman, L. (1995). Better subset regression using the nonnegative garotte. *Technometrics* 37, 373-384.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and regression trees*, Monterey: Wadsworth and Brooks/Cole.

Copas, J.B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society B* 45, 311-354.

Cox, D.R. (1970). *Analysis of binary data*, London: Methuen.

Das, P. and Mulder, P.G.H. (1983). Regression to the mode. *Statistica Neerlandica* 37, 15-20.

Faraggi, D. and Simon, R. (1995). A neural network for survival data. *Statistics in Medicine* 14, 73-82.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *J. Anthrop. Inst.* 15, 246-263.

Hastie, T.J., Tibshirani, R.J. (1990). *Generalized additive models*, London: Chapman and Hall.

Hecht-Nielsen, R. (1990). *Neurocomputing*, Reading, MA: Addison-Wesley.

Hertz, J.A., Krogh, A., Palmer, R.G. (1991). *Introduction to the theory of neural computation*, Redwood City, U.S.A.: Addison Wesley Publishing Company.

van Houwelingen, J.C. and le Cessie, S. (1990). Predictive value of statistical models. *Statistics in Medicine* 9, 1303-1325.

Kappen, H.J., Neijt, J.P. (1993). Neural network analysis to predict treatment outcome. *Annals of Oncology* 4, Suppl, 31-34.

Kippenhan, J.S., Barker, W.W., Pascal, S., Nagel, J., Duara, R. (1992). Evaluation of a neural-network classifier for PET scans of normal and Alzheimer disease subjects. *Journal of Nuclear Medicine* 33, 1459-1467.

Liestøl, K., Andersen, P.K., Andersen, U. (1994). Survival analysis and neural nets. *Statistics in Medicine* 13, 1189-1200.

Marshall, G., Grover, F. L., Henderson, W. G., Hammermeister, K. E. (1994). Assessment of predictive models for binary outcomes: An empirical approach using operative death from cardiac surgery. *Statistics in Medicine* 13, 1501-1511.

McCullagh, P. and Nelder, J. A (1989). *Generalized linear models*, Second edition, Chapman and Hall: London.

Miller, A.J. (1984). Selection of subsets of regression variables (with Discussion). *Journal of the Royal Statistical Society A* 147, 389-425.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics* 9, 705-724.

Ravdin, P.M., Clark, G.M., Hilsenbeck, S.G., Owens, M.A., Vendely, P., Pandian, M.R. (1992). A demonstration that breast cancer recurrence can be predicted by neural network analysis. *Breast Cancer Research and Treatment* 21, 47-53.

Schmoor, C., Ulm, K., Schumacher, M. (1993). Comparison of the Cox model and the regression tree procedure in analysing a randomized clinical trial. *Statistics in Medicine* 12, 2351-2366.

Schumacher, M., Holländer, N., Sauerbrei, W. (1997). Resampling and cross-validation techniques: A tool to reduce bias caused by model building?. *Statistics in Medicine* (to appear)

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B* 36, 111-147.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society B* 58, 267-268.

Vach, W. (1997). On the relation between the shrinkage effect and a shrinkage method. *Computational Statistics* (to appear)

Verweij, P.J.M. and van Houwelingen, H.C. (1993). Cross-validation in survival analysis. *Statistics in Medicine* 12, 2305-2314.

β	$E\hat{c}$	$AMSE(\hat{\mu})$ ORD	HC	P_{imp}^{APE} $(\hat{\mu}_{HC}, \hat{\mu}_{ORD})$	AP_{imp}
(-2.75,1.0, 0.9 ,0.8, 0.7 ,0.6, 0.5 ,0.4, 0.3 ,0.2,0.1)	0.59	0.0168	0.0148	0.703	0.563
(-2.70,1.2,1.05,0.9,0.75,0.6,0.45,0.3,0.15,0.0,0.0)	0.63	0.0167	0.0148	0.694	0.560
(-2.80,1.4, 1.2 ,1.0, 0.8 ,0.6, 0.4 ,0.2, 0.0 ,0.0,0.0)	0.66	0.0164	0.0147	0.691	0.554
(-2.25,1.5, 1.2 ,0.9, 0.6 ,0.3, 0.0 ,0.0, 0.0 ,0.0,0.0)	0.64	0.0162	0.0146	0.674	0.550

Table 1: Performance of the ordinary predictor $\hat{\mu}_{ORD}$ and the predictor $\hat{\mu}_{HC}$ with respect to the average behaviour for future individuals. Additionally the expectation of the shrinkage factors are given

$\beta = (-2.75, 1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1)$

x	$\mu_0(x)$	bias$(\hat{\mu}(x))$ ORD	HC	$IMSE(\hat{\mu}(x))$ ORD	HC	P_{imp}^{IPE} $(\hat{\mu}_{HC}, \hat{\mu}_{ORD}, x))$
1111111111	0.940	0.001	-0.096	0.002	0.017	0.23
1111110110	0.905	-0.001	-0.104	0.004	0.019	0.27
1011111111	0.864	0.001	-0.103	0.006	0.019	0.31
0111111011	0.810	-0.001	-0.098	0.010	0.019	0.33
1110010111	0.741	0.002	-0.078	0.014	0.016	0.43
0101111101	0.679	0.001	-0.061	0.017	0.014	0.50
0011111010	0.611	-0.004	-0.042	0.021	0.013	0.67
1001101011	0.562	0.003	-0.021	0.021	0.011	0.78
0110011010	0.513	-0.003	-0.006	0.022	0.011	0.86

$\beta = (-2.25, 1.5, 1.2, 0.9, 0.6, 0.3, 0.0, 0.0, 0.0, 0.0, 0.0)$

x	$\mu(x)$	bias$(\hat{\mu}(x))$ ORD	HC	MSE$(\hat{\mu}(x))$ ORD	HC	P_{imp}^{IPE} $(\hat{\mu}_{HC}, \hat{\mu}_{ORD}, x))$
1111100000	0.905	-0.001	-0.089	0.003	0.015	0.28
1111000000	0.875	-0.004	-0.093	0.006	0.016	0.30
1110100000	0.839	-0.002	-0.089	0.008	0.017	0.34
1101100000	0.794	-0.002	-0.082	0.011	0.016	0.36
1011100000	0.741	-0.002	-0.072	0.014	0.015	0.45
0111100000	0.679	-0.005	-0.058	0.018	0.014	0.52
1100000000	0.611	-0.004	-0.037	0.022	0.014	0.68
0110100000	0.537	-0.001	-0.013	0.024	0.013	0.85

Table 2: Bias and mean square error of the ordinary predictor $\hat{\mu}_{ORD}$ and the predictor $\hat{\mu}_{HC}$ for patients with specific values of the covariate vector x. Additionally the probability of improvement is given for each patient.

$\beta = (-2.75, 1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1)$

		bias($\hat{\mu}(x)$)			sdv($\hat{\mu}(x)$)		
x	$\mu_0(x)$	ORD	$BE, 5\%$	$BE, 15\%$	ORD	$BE, 5\%$	$BE, 15\%$
1111111111	0.940	0.001	-0.091	-0.035	0.041	0.099	0.067
1111110110	0.905	-0.001	-0.082	-0.032	0.062	0.111	0.085
1011111111	0.864	0.001	-0.123	-0.055	0.076	0.134	0.108
0111111011	0.810	-0.001	-0.121	-0.056	0.099	0.146	0.127
1110010111	0.741	0.002	-0.050	-0.021	0.120	0.145	0.133
0101111101	0.679	0.001	-0.102	-0.051	0.130	0.159	0.152
0011111010	0.611	-0.004	-0.091	-0.049	0.142	0.163	0.159
1001101011	0.562	0.003	-0.034	-0.018	0.146	0.160	0.159
0110011010	0.513	-0.003	-0.017	-0.008	0.149	0.167	0.161

$\beta = (-2.25, 1.5, 1.2, 0.9, 0.6, 0.3, 0.0, 0.0, 0.0, 0.0, 0.0)$

		bias($\hat{\mu}(x)$)			sdv($\hat{\mu}(x)$)		
x	$\mu_0(x)$	ORD	$BE, 5\%$	$BE, 15\%$	ORD	$BE, 5\%$	$BE, 15\%$
1111100000	0.905	-0.001	-0.042	-0.018	0.059	0.076	0.066
1111000000	0.875	-0.004	-0.027	-0.010	0.075	0.084	0.075
1110100000	0.839	-0.002	-0.021	-0.010	0.090	0.093	0.090
1101100000	0.794	-0.002	-0.036	-0.019	0.104	0.108	0.106
1011100000	0.741	-0.002	-0.062	-0.029	0.119	0.124	0.121
0111100000	0.679	-0.005	-0.074	-0.035	0.135	0.131	0.132
1100000000	0.611	-0.004	0.054	0.027	0.147	0.128	0.138
0111100000	0.537	-0.001	-0.016	-0.007	0.154	0.139	0.147

Table 3: Bias and standard deviations of the ordinary predictor $\hat{\mu}_{ORD}$ and two predictors based on backward elimination variable selection for some patients with specific values of the covariate vector x.

	β_j	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
$\alpha = 5\%$	$P(\hat{\beta}_j^{BE,\alpha} \neq 0)$	0.77	0.69	0.59	0.48	0.38	0.30	0.19	0.16	0.09	0.06
	$E(\hat{\beta}_j^{BE,\alpha} \| \hat{\beta}_j^{BE,\alpha} \neq 0)$	1.20	1.13	1.08	1.05	1.00	0.98	0.95	0.92	0.75	0.54
	$E(\hat{\beta}_j^{BE,\alpha})$	0.92	0.78	0.63	0.51	0.39	0.29	0.18	0.15	0.07	0.04
$\alpha = 15\%$	$P(\hat{\beta}_j^{BE,\alpha} \neq 0)$	0.90	0.86	0.77	0.68	0.59	0.48	0.38	0.32	0.22	0.18
	$E(\hat{\beta}_j^{BE,\alpha} \| \hat{\beta}_j^{BE,\alpha} \neq 0)$	1.15	1.05	1.00	0.95	0.89	0.86	0.79	0.74	0.65	0.36
	$E(\hat{\beta}_j^{BE,\alpha})$	1.04	0.92	0.76	0.64	0.53	0.42	0.30	0.23	0.14	0.07
	β_j	1.5	1.2	0.9	0.6	0.3	0.0	0.0	0.0	0.0	0.0
$\alpha = 5\%$	$P(\hat{\beta}_j^{BE,\alpha} \neq 0)$	0.98	0.91	0.68	0.36	0.14	0.05	0.05	0.05	0.06	0.05
	$E(\hat{\beta}_j^{BE,\alpha} \| \hat{\beta}_j^{BE,\alpha} \neq 0)$	1.58	1.30	1.12	1.02	0.93	*	*	*	*	*
	$E(\hat{\beta}_j^{BE,\alpha})$	1.54	1.19	0.76	0.37	0.13	-0.01	0.00	0.04	0.00	0.05
$\alpha = 15\%$	$P(\hat{\beta}_j^{BE,\alpha} \neq 0)$	0.99	0.97	0.84	0.57	0.29	0.16	0.17	0.14	0.19	0.16
	$E(\hat{\beta}_j^{BE,\alpha} \| \hat{\beta}_j^{BE,\alpha} \neq 0)$	1.60	1.30	1.05	0.91	0.76	*	*	*	*	*
	$E(\hat{\beta}_j^{BE,\alpha})$	1.59	1.25	0.89	0.51	0.22	-0.02	0.00	0.00	0.02	0.06

Table 4: Evaluation of variable selection in logistic regression based on backward elimination with nominal levels of $\alpha = 5\%$ and $\alpha = 15\%$. For two parameter constellations we report for each variable the probability to be selected, the conditional expectation of the regression coefficient estimate given the selection of the variable and the ordinary expectation. A * indicates a theoretical value of 0; the approximation in the Monte Carlo study was very poor because we can use only about 50 or 150 samples, respectively.

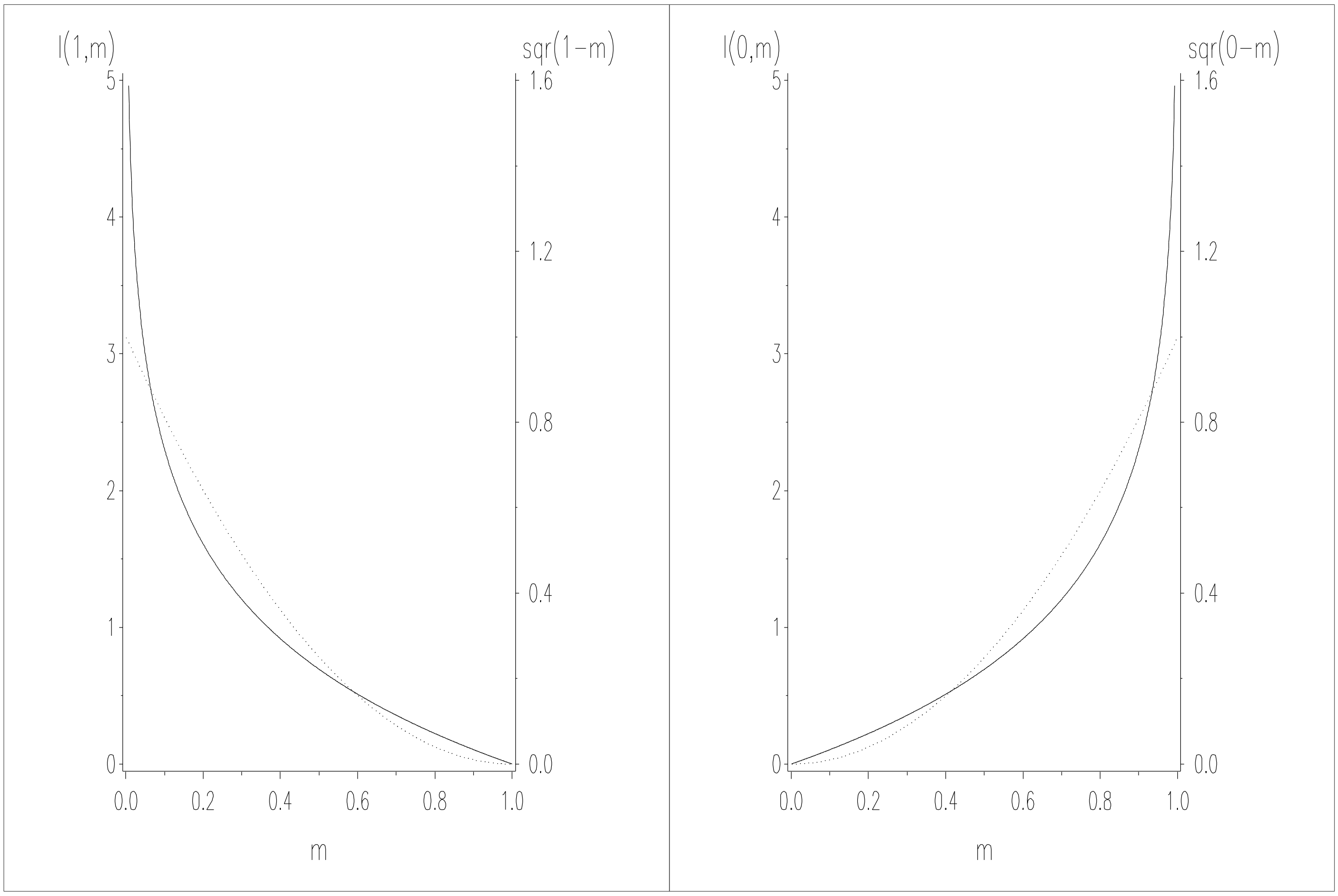

Figure 1: The functions $m \mapsto l(y, m) = -\log(1 - |y - m|)$ (solid line) and $m \mapsto sqr(m) = (y - m)^2$ (dotted line) for the two possible values $y = 1$ (left side) and $y = 0$ (right side).