Response smoothing estimators in binary regression

Gerhard Tutz
Ludwig-Maximilians-Universität München
Akademiestraße 1, 80799 München
tutz@stat.uni-muenchen.de

Abstract

A shrinkage type estimator is introduced which has favorable properties in binary regression. Although binary observations are never very far away from the underlying probability, in all interesting cases there is a non–zero distance between observation and underlying mean. The proposed response smoothing estimate is based on a smoothed version of the observed responses which is obtained by shifting the observation slightly towards the mean of the observations and therefore closer to the underlying probability. Estimates of this type are very easily computed by using common program packages and exist also when the number of predictors is very large. Moreover, they are robust against outliers. A combination of response smoothing estimators and Pregibon's resistant fitting procedure corrects for the overprediciton of the resistant fitting in a very simple way. Estimators are compared in simulation studies and applications.

KEYWORDS: Logit model, resistant fitting, response smoothing estimator, shrinkage, weighted estimation, data sharpening.

1 Introduction

The usual method of fitting binary regression models is maximum likelihood with favourable asymptotic properties but high sensitivity to 'outliers'. In particular if the number of covariates is high as compared to the number of observations, unstable estimates are to be expected. Along with the development of diagnostic tools for binary regression models (e.g. Pregibon (1981), Landwehr, Pregibon & Shoemaker (1984), Fowlkes (1987)) robust estimation procedures have been suggested. Pregibon's (1982) resistant fitting procedure is based on the downgrading of the influence of observations with high residuals. Copas (1988) considers the substantial bias of resistant fitting which yields numerically larger coefficients, yielding a more extreme fit, closer to 0 or 1. He considers a bias corrected version and proposes a misclassification model where transpositions between the possible outcomes 0 and 1 happen with a small probability. Carroll & Pederson (1993) study an estimate which is closely related to Copas' misclassification estimate but which is consistent for the logistic model.

Alternative approaches to estimation in binary regression which aim at high dimensional settings are based on penalized likelihood estimation. Marx, Eilers & Smith (1992), LeCessie & van Houwelingen (1992) and Segerstedt (1992) consider ridge regression within the framework of generalized linear models. In ridge regression a term is added to the likelihood which penalizes the squared length of the vector of regression parameters yielding shrinkage of the estimate. By avoiding the ill-conditioning of the information matrix ridge type estimators allow the fitting of models even when maximum likelihood estimates do not exist. More recently, Klinger (1997) proposed shrinkage methods based on soft thresholds which have similar properties. Tibshirani (1996) proposed a shrinkage type estimator called Lasso which is also connected to subset selection.

In the following a simple and easily implemented method is proposed to obtain improved estimates which are robust and well adapted to the case of many predictors. The influence of observations may be downgraded by choosing an influence function in the spirit of M-estimation (Hampel et al. (1986)). As Copas (1988) remarks, the choice of the influence function and its associated tuning constant is essentially arbitrary. Thus he arrived at resistant fitting by modelling random misclassification parts of transpositions between 0 and 1. The approach followed here also starts from the observation that in binary regression the only alternative to the observations $y_i \in \{0, 1\}$ is its counterpart $1 - y_i \in \{1, 0\}$. But instead of assuming transcription errors the effect of observations y_i is downgraded by introducing its counterpart $1 - y_i$ and putting different weights on these pairs of observation. In the extreme case where both observation have the same weight the effect is strong shrinkage of the estimate, an effect which is comparable to observing $y_i = 0.5$ which is not a recordable observation in binary regression. In contrast to other shrinkage type estimators like ridge regression and the Lasso, shrinkage is not obtained by restricting the range of the parameters but by explicitly exploiting the discreteness of binary observations. The essential point is that estimates are dramatically improved just by using common program packages which allow weighted estimation like e.g. S-PLUS or SAS.

In Section 2 the basic concept of response smoothed estimation is outlined. Section 3 is an illustration of the estimation in the simple setting of a binary covariable where estimation of parameters is strongly connected to the estimation of odds ratios. After considering methods how to choose the tuning parameter, in Section 5 simulation results show the performance of the estimator. In Section 6 the method is generalized by linking the transformation estimate to Pregibon's resistant estimate. Simulation results show the improvement of the resistant estimate also in the case of contaminated data. After deriving properties of the generalized estimate the method is illustrated by an application to a real data set.

2 Response smoothing estimators based on smoothed maximum likelihood estimation

The model investigated is the binary regression model

$$\pi_i = P(y_i = 1|x_i) = h(x_i'\beta)$$

where h is an appropriate response function, e.g. the logistic distribution function $h(\eta) = 1/(1+\exp(-\eta))$. Let the data be given by $(y_i, x_i), i = 1, \ldots, n$, where $y_i \in \{0, 1\}$ are the responses and x_i represents the covariates. The original data set of n observations is doubled by defining

$$y_{n+i} = 1 - y_i, x_{n+i} = x_i, \quad i = 1, \ldots n.$$

Thus in the enlarged data set one has for each observations y_i its counterpart $1 - y_i$ with identical covariate value x_i.

The weighting scheme used in estimation distinguishes between the original data $(y_i, x_i), i = 1, \ldots, n$, and the pseudo data $(y_i, x_i), i = n+1, \ldots, 2n$. The weighting scheme puts weight $1 - \alpha_i$ on the observations from the original data and α_i on the pseudo observations corresponding to y_i where $\alpha_i \in [0, 0.5]$. Instead of the usual log-likelihood one considers the weighted log-likelihood

$$l_w(\beta) = \sum_{i=1}^{2n} w_i l_i(y_i, \pi_i)$$

where $l_i(y_i, \pi_i) = y_i \log(\pi_i) + (1 - y_i)\log(1 - \pi_i)$ with $\pi_i = h(x_i'\beta)$ is the log-likelihood contribution of observation i and

$$w_i = \begin{cases} 1 - \alpha_i & i \leq n \\ \alpha_i & i > n. \end{cases}$$

The parameters α_i specify the amount of smoothing which is applied. For $\alpha_i = 0$ only the original observations are used, and usual maximum likelihood estimation is obtained. With increasing α_i the pseudo observations are increasingly influential. Since $\pi_i = \pi_{n+i} = h(x_i'\beta)$ the log-likelihood reduces to

$$\begin{aligned} l_w(\beta) &= \sum_{i=1}^{n} (1 - \alpha_i)\{y_i \log(\pi_i) + (1 - y_i)\log(1 - \pi_i)\} \\ &\quad + \alpha_i\{(1 - y_i)\log(\pi_i) + y_i \log(1 - \pi_i)\} \\ &= \sum_{i=1}^{n} (\alpha_i + y_i(1 - 2\alpha_i))\log(\pi_i) + (1 - \alpha_i - y_i(1 - 2\alpha_i))\log(1 - \pi_i). \end{aligned}$$

It is seen that if $\alpha_i = 1/2$ the contribution of observations y_i and $y_{n+i} = 1 - y_i$ reduces to

$$0.5\{\log(\pi_i) + \log(1 - \pi_i)\} = 0.5\log(\pi_i(1 - \pi_i)).$$

Thus the log-likelihood depends on the data only through π_i and the contribution takes its maximum value at $\pi_i = 0.5$ yielding shrinkage of $\hat{\beta}$. If $\alpha_i = 0, i = 1, \ldots, n$, one obtains the usual log-likelihood.

Alternative forms of the weighted log-likelihood are given by

$$\begin{aligned} l_w(\beta) &= \sum_{i=1}^{n} (1-\alpha_i)^{y_i} \alpha_i^{1-y_i} \log(\pi_i) + \alpha_i^{y_i} (1-\alpha_i)^{1-y_i} \log(1-\pi_i) \\ &= \sum_{i=1}^{n} (1-\alpha_i) \log(1 - |y_i - \pi_i|) + \alpha_i \log(|y_i - \pi_i|). \end{aligned}$$

The corresponding score function $s_w(\beta) = \partial l_w / \partial \beta$ has the form

$$\begin{aligned} s_w(\beta) &= \sum_{i=1}^{2n} w_i x_i (\partial h(\eta_i)/\partial \eta)(y_i - \pi_i)/\sigma_i^2 \\ &= \sum_{i=1}^{n} (1-\alpha_i) x_i (\partial h(\eta_i)/\partial \eta)(y_i - \pi_i)/\sigma_i^2 + \alpha_i x_i (\partial h(\eta_i)/\partial \eta)(1 - y_i - \pi_i)/\sigma_i^2 \\ &= \sum_{i=1}^{n} x_i \{y_i - \pi_i + \alpha_i (1 - 2y_i)\} (\partial h(\eta_i)/\partial \eta)/\sigma_i^2 \end{aligned}$$

where $\partial h(\eta_i)/\partial \eta$ is the derivative at $\eta_i = x_i' \beta$ and $\sigma_i^2 = \pi_i(1 - \pi_i)$ is the variance at η_i. For the logit model one obtains the simpler form

$$\begin{aligned} s_w(\beta) &= \sum_{i=1}^{n} x_i (y_i - \pi_i + \alpha_i (1 - 2y_i)) \\ &= \sum_{i=1}^{n} x_i (y_i - \pi_i + (-1)^{y_i} \alpha_i). \end{aligned}$$

From the estimation equation $s_w(\hat{\beta}_w) = 0$ one has with $\hat{\pi}_i = h(x_i' \hat{\beta}_w)$

$$\sum_{i=1}^{n} x_i (y_i + (-1)^{y_i} \alpha_i) = \sum_{i=1}^{n} x_i \hat{\pi}_i. \qquad (1)$$

The latter form shows the effect of smoothing. In the sufficient statistic

$$\sum_{i=1}^{n} x_i (y_i + (-1)^{y_i} \alpha_i)$$

the observation y_i is replaced by the smaller value $1-\alpha_i$ if $y_i = 1$ and replaced by α_i if $y_i = 0$. Thus the essential effect is that the observation y_i itself is shrinked towards 0.5. It is easily seen that for $\alpha_i = 0.5$ the flat estimate $\hat{\beta}_w = 0.0$ is a solution of the estimation equation. It is noteworthy that the specific weighting scheme with $1-\alpha_i$ and α_i on the pair of observations yields estimation equations which do not incorporate weights in the score function. For estimation equations of this type see Carroll & Pederson (1993) and Section 6. Instead the effect of the smoothing is seen from the term $y_i + (-1)^{y_i}\alpha_i$ which replaces the original observation by a smoothed observation closer to 0.5. This effect is the cause why the solution $\hat{\beta}_w$ of the equation $s_w(\hat{\beta}_w) = 0$ is called *response smoothing estimate*.

The consideration of pseudo-observations may be seen as a way of deriving the estimate and, more important, as a way how to simply obtain estimates from common program packages. As is seen from the score function, pseudo-observations are not needed. The actual modification is that the original observations y_i are transformed into $y_i + (-1)^{y_i}\alpha_i$, a transformation which is well motivated by the fact that observations $y_i \in \{0, 1\}$ are always a crude exaggeration of the underlying probability π_i, since y_i is never equal to π_i, except in trivial cases. With the focus on data transformation it may also be called data transformation estimate. The concept is similar in spirit to the more recently introduced methods of 'data sharpening'. For example Choi & Hall (1999) sharpen the data by making them slightly more clustered than before in order to reduce bias of density estimators. Although it is unusual to smooth across the response the method yields improved estimates which exist in cases where the maximum likelihood estimate fails and corrects for the overprediction of the resistant estimate.

3 Low dimensional case: log odds ratio

A simple case which is of interest of its own is given if the single covariate takes only two values, for example treatment or placebo. Then estimation of parameters corresponds to the estimation of odds ratios or transformations of log odds ratios. A vast body of literature exists for this case, see e.g. Parzen, Lipsitz, Ibrahim

& Klar (2002) who recently proposed an estimate that always exists. Here odds ratios are used as an example in low dimensions.

Let the binary covariate x given in effect coding $x \in \{1, -1\}$ and the two response probabilities $\pi_1 = P(y = 1|x = 1), \pi_2 = P(y = 1|x = -1)$ be specified by a logit model. Then the parameters are given by

$$\beta_0 = \frac{1}{2}(\text{logit}(\pi_1) + \text{logit}(\pi_2)),$$
$$\beta_1 = \frac{1}{2}(\text{logit}(\pi_1) - \text{logit}(\pi_2)),$$

where $\text{logit}(\pi_i) = \log(\pi_i/(1 - \pi_i))$ and $\beta_1 = \frac{1}{2}\log(\{\pi_1/(1 - \pi_1)\}/\{\pi_2/(1 - \pi_2)\})$ is (apart from the factor 1/2) the log odds ratio. Let the weights for the smoothed estimates be given by $\alpha_i = \alpha$, thus only distinguishing between original and pseudo observations. Since the model is saturated one may compute the smoothed estimates of π_1, π_2. With tuning parameter α one obtains after some calculation

$$\hat{\pi}_{1,\alpha} = p_1 + \alpha(1 - 2p_1), \hat{\pi}_{2,\alpha} = p_2 + \alpha(1 - 2p_2)$$

where $p_1(p_2)$ denotes the relative frequencies for $y = 1$ given $x = 1 (x = -1)$ in the original sample. For $\alpha = 0$ one obtains the relative frequencies $\hat{\pi}_{1,\alpha=0} = p_1, \hat{\pi}_{2,\alpha=0} = p_2$. For $0 < \alpha < 0.5$ the estimates are smoothed towards 0.5. This is easily seen from the reparameterization

$$\hat{\pi}_{1,\alpha} = \gamma p_1 + (1 - \gamma)0.5, \hat{\pi}_{2,\alpha} = \gamma p_2 + (1 - \gamma)0.5$$

where $\gamma = 1 - 2\alpha$. The shrinkage of the corresponding parameter estimate $\hat{\beta}_{1,\alpha}$ towards zero is seen from

$$\hat{\beta}_{1,\alpha} = \frac{1}{2}(\text{logit}(\hat{\pi}_{1,\alpha}) - \text{logit}(\hat{\pi}_{2,\alpha})).$$

Bias and variances of $\hat{\pi}_{i,\alpha}$ as estimators of π_i are given by

$$\begin{aligned} \text{bias}(\hat{\pi}_{i,\alpha}) &= E(\hat{\pi}_{i,\alpha} - \pi_i) = (1 - \gamma)(0.5 - \pi_i), \\ \text{var}(\hat{\pi}_{i,\alpha}) &= \gamma^2 \pi_i (1 - \pi_i)/n_i, \end{aligned}$$

where $n_1(n_2)$ is the number of observations at $x = 1 (x = -1)$. Since $0 \le \gamma \le 1$ smoothing decreases the variance but adds some bias which is positive for $\pi < 0.5$

and negative for $\pi > 0.5$. The MSE optimal estimator which minimizes the mean squared error MSE$= E\sum_i(\pi_i - \hat{\pi}_{i,\alpha})^2$ depends on the underlying probabilities. By usual minimization procedures one obtains the optimal smoothing parameter

$$\gamma_{\text{opt}} = \frac{(.5-\pi_1)^2 + (.5-\pi_2)^2}{\pi_1(1-\pi_1)/n_1 + (.5-\pi_1)^2 + \pi_2(1-\pi_2)/n_2 + (.5-\pi_2)^2}.$$

One has $\gamma_{\text{opt}} = 1$, i.e. the maximum likelihood estimator, if and only if $\pi_1, \pi_2 \in \{0,1\}$. In all other cases the maximum likelihood estimator may be improved by choosing an appropriate tuning parameter. Thus for appropriate choice of γ the maximum likelihood estimators can be improved in all cases of practical relevance.

The dramatic improvement of estimates in the finite sample case is demonstrated for $n_1 = n_2 = 10$ and $\pi_1 = 0.1, \pi_2 = 0.8$. Since β_1 corresponds to the log odds ratio the mean squared error for β_1 is shown in Figure 1 for varying parameter α. The usual maximum likelihood estimate

$$\hat{\beta}_1 = \frac{1}{2}\log\frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

can equal $\infty(p_1 = 1)$ and $-\infty(p_2 = 1)$ with positive probability. Thus for $\alpha \to 0$ the mean squared error tends to infinity. As is seen from Figure 1 there is a distinct minimum where mean squared error is minimized. For $\alpha > 0$ the estimate always exists. It is much less sensitive if α is larger than the optimal value than in the case where $\alpha \to 0$ and therefore closer to the maximum likelihood estimate.

In the even simpler case without covariates the estimator reduces to $\hat{\pi}_\alpha = (1-\alpha)p + \alpha(1-p) = \gamma p + (1-\gamma)0.5$ where p is the relative frequencies and $\gamma = 1 - 2\alpha$. Thus one obtains a convex combination of the mle p and $\mu = 0.5$. Estimators of this type may also derived from a Bayesian viewpoint assuming a beta binomial distribution with prior mean $\mu = 0.5$ (see Santner & Duffy (1989), p.25).

4 Choice of shrinkage

In the general case the weights α_i are connected to single observations. In analogy to shrinkage in ridge regression where one shrinkage parameter is used one wants the shrinkage to be determined by few tuning parameters which have to be chosen appropriately. The simplest case of a constant tuning parameter $\alpha_i = \alpha$ chosen from $[0, 0.5]$ implies for $\alpha > 0$ shrinkage of all parameters including the intercept. The estimates $\hat{\pi}_i$ are shrinked towards 0.5. However, this seems only adequate if the underlying probabilities are symmetrically distributed around 0.5.

If an intercept is included, which will be the case in the following, the usual maximum likelihood estimate for the logit model has the property

$$\frac{1}{n}\sum_i \hat{\pi}_i = \bar{y} \tag{2}$$

where $\bar{y} = n^{-1}\sum_i y_i$ is the mean over all observations. In order to retain this property for arbitrary $\bar{y}$ one has to use different weights for $y_i = 0$ and $y_i = 1$.

Let α_0 denote the weight for $y_i = 0$ and α_1 denote the weight for $y_i = 1$, i.e. $\alpha_i = \alpha_0$ if $y_i = 0$, $\alpha_i = \alpha_1$ if $y_i = 1, i = 1, \dots, n$. Then by considering the first equation of the system of equations (1) which corresponds to the intercept one obtains

$$\sum_i y_i - n_1\alpha_1 + n_0\alpha_0 = \sum_i \hat{\pi}_i$$

where n_j is the number of observations with $y_i = j$. Some calculations show that $\bar{y} = n^{-1}\sum \hat{\pi}_i$ is fulfilled if

$$\alpha_1 = \frac{1-\bar{y}}{\bar{y}}\alpha_0 \tag{3}$$

where $\alpha_0 \in [0, \bar{y}]$. Thus the pseudo observations in (1) are given by

$$y_i + (-1)^{y_i}\alpha_i = \begin{cases} 1 - \alpha_1 & \text{if} \quad y_i = 1 \\ \alpha_0 & \text{if} \quad y_i = 0. \end{cases}$$

In the extreme case $\alpha_0 = \bar{y}$, $\alpha_1 = 1 - \bar{y}$, one obtains $y_i + (-1)^{y_i}\alpha_i = \bar{y}$ and $\hat{\pi}_i = \bar{y}$ is a solution of (1). Then the parameter $\beta' = (\beta_0, \beta_1, \dots, \beta_p)$ which weights $x' = (1, x_1, \dots, x_p)$ is estimated by $\hat{\beta} = (\hat{\beta}_0, 0, \dots, 0)$ with $\hat{\beta}_0 = exp(\bar{y})/(1 + exp(\bar{y}))$. Thus reduction to just one tuning parameter, say $\alpha = \alpha_0$, should be based on using $\alpha_i = \alpha_0$ if $y_i = 0$ and $\alpha_i = \alpha_1$ if $y_i = 1$ where (3) has to be fulfilled.

The corresponding weighting scheme for observations and pseudo-observations is given by

$$w_i = \begin{cases} 1 - \{(1-\bar{y})/\bar{y}\}^{y_i}\alpha & i \le n \\ \{(1-\bar{y})/\bar{y}\}^{1-y_i}\alpha & i > n \end{cases}$$

where y_i, $i = 1, \ldots, 2n$ are observations and pseudo-observations.

Property (2) results from the symmetry of the logistic distribution and does not hold for models with asymmetrical response function, e.g. the log-log or the complementary log-log model which have response function $h(\eta) = 1 - \exp(-\exp(\eta))$ and $h(\eta) = \exp(-\exp(\eta))$ respectively. For asymmetric distribution functions it is suggested to reduce the number of tuning parameters by specifying the weight α_0 for $y_i = 0$ and α_1 for $y_i = 1$ and performing a grid search across the values of $(\alpha_0, \alpha_1) \in [0, 0.5]^2$.

A data driven choice of the tuning parameter may be based on similar concepts as in smoothing methods. There is a wide body of literature on bandwidth choice in nonparametric regression and density estimation. Classical methods aim at approximately unbiased estimations of mean average squared error or the expected Kullback-Leibler discrepancy.

Following Hurvich, Simonoff & Tsai (1998) we do not aim at the bandwidth which minimizes mean integrated squared error or similar measures but want to approximate the average squared error for the observed data set. Thus instead of a measure which is optimal for given sample size and design the choice is closer connected to the observed data set and therefore is more relevant to the data analyst. The criteria used are the averaged squared error

$$\mathrm{ASE} = \sum_i (\pi_i - \hat{\pi}_i)^2,$$

averaged Kullback-Leibler discrepancy

$$\mathrm{AKL} = \sum_i \pi_i \log\left(\frac{\pi_i}{\hat{\pi}_i}\right) + (1-\pi_i)\log\left(\frac{1-\pi_i}{1-\hat{\pi}_i}\right)$$

and averaged L_1-distance

$$\mathrm{AL1} = \sum_i \mid \pi_i - \hat{\pi}_i \mid$$

The latter is considered since it is strongly connected to classification.

In cross validation or leaving-one-out methods successively one observation y_i is left out and the estimate $\hat{\pi}_i$, based on the reduced data set, is understood as an estimate of π_i. If in the criterion π_i is replaced by y_i one obtains, e.g. for the squared error,

$$\mathrm{CV}_\alpha(SE) = \sum\nolimits_i (y_i - \hat{\pi}_i)^2.$$

In the following we will consider ASE, AKL, AL1 based on smoothing parameters which are optimized by cross-validation for each simulated data set.

5 Comparison of maximum likelihood and response smoothing estimators

In a simulation study the improvement of the estimates is investigated. The parameters for the underlying model for the number of predictors p ranging from 2 to 16 are given by

$$\begin{aligned}
p = {}&\ 2: && \beta' = (0, 1, 0.2),\\
p = {}&\ 4: && \beta' = (0, 1, 0.7, 0.4, 0.1),\\
p = {}&\ 8: && \beta' = (0, 1, 7/8, 6/8, \ldots, 1/8),\\
p = {}&16: && \beta' = (0, 1, 15/16, 14/16, \ldots, 1/16).
\end{aligned}$$

For binary predictors effect coding is used with $x \in \{-1, 1\}$, for continuous predictors the $x_i's$ have been drawn from a uniform distribution with support $[-2, 2]$. All simulations have been performed by using the fitting procedure for the logit model provided by R, allowing for 50 iterations and termination criterion 0.0001.

In the first column of Table 1 and Table 2 the squared error of the maximum likelihood estimates are given. These are contrasted to the response smoothing estimates where optimal smoothing parameters were chosen by cross validation based on an error measure. The resulting squared error losses are denoted by CV(KL), CV(SE) and CV(L1) for Kullback-Leibler, squared error and

L_1-distance. The individual improvements for simulation s may be measured by CV_s/ML_s where CV_s is the loss based on cross validated tuning parameter and ML_s is the loss for the maximum likelihood estimate. Since the distribution of CV_s/ML_s is skewed the mean across logarithms is considered. Since $n^{-1}\sum_s \ln(CV_s/ML_s) = \ln((\prod_s CV_s/ML_s)^{1/n})$ this corresponds to the logarithm of the geometric mean. In Table 1 the means of $\ln(CV_s/ML_s)$ are abbreviated by $lnCV(KL), lnCV(MSE), lnCV(L1)$ according to the loss function which is used in cross–validation.

It is seen from Table 1 that for two predictors already for sample size n=20 asymptotics is kicking in and there is not much space for improvement. However, in interesting cases, when the number of predictors is eight or higher, estimates are strongly improved. For example with eight variables in the predictor and n=30 the squared error loss is strongly decreased to about 53 percent. With termination criterion 0.0001 Fisher scoring stopped below the maximum of 50 iterations for all of the 200 simulations.

The effect is illustrated in Figure 2 where the Kullback-Leibler loss for the maximum likelihood estimates is plotted against the loss resulting for the smoothed estimate for 200 simulations. The underlying model is a logit model with p=8 binary predictors and sample sizes n=20 and 30. It is seen that in particular for data sets which produce estimates which are far from the true values the estimates are strongly improved. Only if maximum likelihood estimates are very good with losses around 0.05 there is not much space for improvement and the response smoothing estimate is not better than the ml estimate.

The consideration of losses which are defined for the estimation of the probability does not show the heavy instability of the maximum likelihood estimate of β. Therefore Figure 3 shows the deviations $\| \hat{\beta}_w - \beta \|$ against $\| \hat{\beta}_{ML} - \beta \|$ for the simulated data sets where $\hat{\beta}_w$ is the response smoothing estimate. The data are the same as in the lower panel of Figure 2 (binary covariables, $p = 8$, $n = 30$). The line in Figure 3 again shows the limit where maximum likelihood estimates have better fit than response smoothing estimates. It is seen that for response smoothing estimates $\| \hat{\beta}_w - \beta \|$ has maximal values around 4 whereas the deviations $\| \beta_{ML} - \beta \|$ take values up to 200, signaling that estimates deteriorate. Even in these cases the algorithm showed convergence within 50 iterations and 0.0001 as termination criterion. The strength of the effect on the estimates of β is seen from Table 3, where the averaged deviations $\| \hat{\beta} - \beta \|$ as well as the effect in individual data sets are given for 200 simulations.

6 Generalized response smoothing estimators

The robust-resistant estimates due to Pregibon (1982) use weights which depend on the design point x_i or/and the true probability. The weights are used to downgrade observations with high residuals. What looks intuitively like a sensible concept regrettably yields estimates with a strong tendency towards overprediction with the effect that estimates of β are numerically larger than the true values. In the following it is shown how the combination with the response smoothing estimator keeps the robustness of the estimation while correcting for the overprediction of Pregibon's estimate.

Let in the general case δ_i denote the weight on the original observations (y_i, x_i) and $\tilde{\delta}_i$ denote the weight on the pseudo data $(y_{n+i} = 1 - y_i, x_{n+i} = x_i), i = 1, \ldots, n$. The weighted log likelihood which uses original and pseudo data is given by

$$\begin{aligned} l_w(\beta) &= \sum_{i=1}^{n} \delta_i \{ y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i) \} + \\ &\quad + \tilde{\delta}_i \{ (1 - y_i) \log(\pi_i) + y_i \log(1 - \pi_i) \}. \end{aligned}$$

With $h_i = (\partial h(\eta_i)/\partial \eta)/\sigma_i^2$ the corresponding score function has the form

$$s_w(\beta) = \sum_{i=1}^{n} x_i h_i \{ \delta_i (y_i - \pi_i) + \tilde{\delta}_i (1 - y_i - \pi_i) \} = \sum_{i=1}^{n} x_i h_i \{ y_i (\delta_i - \tilde{\delta}_i) - \pi_i (\delta_i + \tilde{\delta}_i) + \tilde{\delta}_i \}.$$

In order to obtain a simpler form let the weights $\delta_i, \tilde{\delta}_i$ be transformed by

$$\varepsilon_i = \delta_i + \tilde{\delta}_i, \quad \alpha_i = \tilde{\delta}_i / \varepsilon_i$$

corresponding to

$$\tilde{\delta}_i = \alpha_i \varepsilon_i, \;\; \delta_i = (1-\alpha_i)\varepsilon_i. \tag{4}$$

This gives after some derivation the closed form

$$s_w(\beta) = \sum_{i=1}^{n} \varepsilon_i x_i h_i \{y_i + (-1)^{y_i}\alpha_i - \pi_i\}$$

yielding the estimation equation

$$\sum_{i=1}^{n} \varepsilon_i x_i h_i \{y_i - \pi_i + (-1)^{y_i}\alpha_i\} = 0. \tag{5}$$

The transformation of weights is chosen such that α_i plays the same role as in Section 2. But in addition now weights ϵ_i are included. Estimation equation (5) represents a general form which is usually used in robust estimation. If $\epsilon_i = 1, \alpha_i = 0$ equation (5) yields the usual maximum likelihood estimate. If $\alpha_i = 0$ and $\epsilon_i = \epsilon(x_i, \pi_i)$ one is in the so-called Mallows class (Mallows (1975)), for $\epsilon_i = \epsilon(x_i, \pi_i, y_i)$ one obtains the Schweppe class (Hampel et al. (1986), see also Carroll & Pederson (1993)). From considering the general form $0 = \sum \epsilon_i x_i h_i \{y_i - \pi_i - c(x_i, \beta)\}$ (Carroll & Pederson (1993)) one sees that in the response smoothing estimate the debiasing factor $c(x_i, \beta)$ is replaced by $(-1)^{y_i}\alpha_i$. Pregibon's (1982) estimate uses $\alpha_i = 0$ and the weight

$$\epsilon_i(x_i, \pi_i, y_i) = w(y_i - \pi_i)$$

where

$$w(u) = \begin{cases} 1, & \text{if } |u| \le 1 - \exp(-\gamma/2) \\ \{-\frac{1}{2}\gamma/\log(1-|u|)\}^{1/2}, & \text{otherwise} \end{cases}$$

with the tuning parameter γ which in Pregibon's examples is taken to be $(1.345)^2$ in order to obtain estimates with approximately 95% asymptotic relative efficiency. By the incorporation of estimates $(-1)^{y_i}\alpha_i$ the overprediction of Pregibon's estimate should be corrected while being resistant to outliers.

It should be noted that (5) can be considered as a starting point for the estimation procedure. The derivation from original and pseudo observation is chosen to show that program packages which allow for weighting may be used by plugging in the corresponding weights from equation (4).

If ϵ_i depends on β second derivatives of s_w are more difficult then in the usual case. In order to keep computation simple we used an iterative procedure by solving (5) with weighted Fisher scoring. For fixed α_i and ϵ_i given by $w(y_i - x_i'\beta_{ML})$ one obtains by weighted Fisher scoring $\hat{\beta}^{(1)}$. Replacing $w(y_i - x_i'\beta_{ML})$ by $w(y_i - x_i'\beta^{(1)})$ one obtains the next iterate $\hat{\beta}^{(2)}$, etc. Convergence was fast, below four cycles for $\hat{\beta}^{(i)}$. The only modification is that weighted Fisher scoring now uses two weights, α_i for the response smoothing and ϵ_i for the incorporation of Pregibon's weights.

Table 4 shows the results for $p = 8$ binary covariates with $n = 30$. For the estimation of probabilities the squared error averaged across simulations are given. The performance on individual data sets is measured by the log of the proportion between squared error loss for the ML estimate and the considered estimate. For the estimation of the parameter, the distance $\| \hat{\beta} - \beta \|$ is considered together with the log proportion where the maximum likelihood estimate β_{ML} is compared to the considered estimate, in the form $\log(\| \hat{\beta} - \beta \| / \| \beta_{ML} - \beta \|)$.

It is seen that, due to overprediction, Pregibon's resistant estimate is worse than the usual maximum likelihood estimate. Although the response smoothing improves estimates strongly, the combination of response smoothing and resistant estimation shows some additional improvement. In addition, the combined estimate is considered where both tuning constants, α and Pregibon's γ, are chosen by cross-validation. However, the effect is weak, in particular since the computational effort is much higher. Fig 4 shows the results for Pregibon's estimate with $\gamma = 1.345^2$ as compared to response smoothing combined with Pregibon's estimate. For clarity estimates where $\| \hat{\beta} - \beta \| > 50$ are denoted by a cross. It is seen that in these cases Pregibon's estimate is equivalent to the ml estimate, for all other data sets the resistent fit is worse than the ml estimate. In contrast, the combined estimate improves almost in every case.

Since resistant estimates are based on the concept of downweighting observations with high residuals it might be argued that the improvement is due to the lack of outliers in the data sets. However, 'a certain number of outliers are bound to occur even if the assumed model is correct' Copas (1988). Thus, outliers should be present in the simulated data sets. In order to strengthen the effect, a small simulation with contaminated data was performed (8 binary predictors, $n = 30$). From the 30 observations 5 randomly drawn observations have been transposed by replacing the observations y_i by $1 - y_i$. Table 5 shows the results for Kullback-Leibler loss as cross–validation criterion. It is seen that contamination stabilizes the ml estimate as well as the resistant estimate but response smoothing is an efficient tool to improve the estimates. For example the squared distance between the estimate and the true value is reduced from above 8.7 to 1.6.

7 Application

Data which have often been used in diagnostics of binary data are the vaso–constriction data from Finney (1947). For $n = 39$ observations it has been measured whether vaso-constriction of the skin of the digits occurs after the inspiration of air. The data vary across volume of air (VOL) and inspiration rate (RATE) which are used as explanatory variables in logarithmic form. Pregibon (1982) shows that two values are poorly accounted for by the logistic model (see also Atkinson & Riani (2000)). If these two data are left out the data may be completely separated by the fitting of the logistic model, an effect which is favourable in discriminant analysis.

Table 6 shows the estimates for various fitting methods, maximum likelihood, weighted fitting with cross-validation based on α, Pregibon's resistant estimate, resistant estimate with response smoothing and cross-validated choice of α and cross-validated choice of α and γ. Pregibon's estimate has the largest values due to the tendency to overprediction. While simple response smoothing yields rather small values the combination of Pregibon's resistant estimate and response smoothing yields values in the middle which still are smaller than the maximum likelihood estimates. Given that the existence of maximum likelihood estimates depends on two observations and therefore the data contain not much information about the slopes of volume and rate, this slightly damped estimates seem sensible.

The effect of damping by α is illustrated in Figure 5, where the simple weighted estimate and weighting with resistant fitting is shown.

8 Approximation results

Starting from the estimation equation $s_w(\beta) = 0$ first order approximation yields

$$\hat{\beta} - \beta \approx \left(\frac{-\partial s_w(\beta)}{\partial \beta'} \right)^{-1} s_w(\beta).$$

Based on this approximation one obtains the sandwich matrix

$$cov(\hat{\beta}) = F(\beta)^{-1} cov(s_w(\beta)) F(\beta)^{-1} \tag{6}$$

where

$$F(\beta) = \sum_{i=1}^{n} x_i {x_i}' \epsilon_i \sigma_i^{-2} \left(\frac{\partial h}{\partial \eta} \right)^2$$

is the usual Fisher matrix and

$$cov(s_w(\beta)) = \sum_{i=1}^{n} x_i {x_i}' \epsilon_i^2 \sigma_i^{-2} \left(\frac{\partial h}{\partial \eta} \right)^2 (1 - 2\alpha_i),$$

(see Appendix). If is seen that for $\alpha_i = 0 (\varepsilon_i = 1)$ the covariance takes the usual form $cov(\hat{\beta}) = F(\beta)^{-1}$. For the extreme value $\alpha_i = 0.5$ the information is taken out of the data and one has $cov(\hat{\beta}) = 0$. Thus (6) should only be used for very small values of α_i. In particular for larger values of α_i one has to take the bias into account.

The bias $b(\beta) = E(\hat{\beta}) - \beta$ may be approximated by

$$b(\beta) = F^{-1} \sum_{i=1}^{n} x_i \epsilon_i \sigma_i^{-2} \left(\frac{\partial h}{\partial \eta} \right) \alpha_i (1 - 2\pi_i)$$

where β has to be replaced by $\hat{\beta}$ and π_i by $\hat{\pi}_i = h(x_i' \hat{\beta}_{ML})$.

For the bias corrected estimate $\hat{\beta}_c - b(\hat{\beta})$ one obtains the approximation

$$cov(\hat{\beta}_c) = F^{-1} \left\{ \sum_{i=1}^{n} x_i {x_i}' \epsilon_i^2 \left(\sigma_i^{-2} \left(\frac{\partial h}{\partial \eta} \right)^2 + var(\hat{\pi}_i) \right) \right\} F^{-1}$$

where

$$var(\hat{\pi}_i) = \left(\frac{\partial h}{\partial \eta} \right)^2 x_i' F(\hat{\beta}_{ML})^{-1} x_i.$$

9 Concluding remarks

The considered estimates are based on smoothing across responses. Alternatively one may see them as data transformation estimates where in the spirit of data sharpening data are transformed. It is essential that pseudo observations are used only as a tool of computing the estimates, thus no 'fake data' are involved. The incorporation of $(-1)^{y_i}\alpha_i$ in the score function is an analogue to the incorporation of an debiasing factor in general estimation equations.

The essential advantages of the response smoothing estimates are: easy computation by use of pseudo observations, improvement of mean squared error (compared to mle), improved existence of estimates, correction of overfitting in resistant fitting procedures and robustness against contamination.

In order to obtain simple estimates only one or two tuning parameters have been used. Thus the potential of smoothing parameters α_i which could be adjusted to the position of x_i in the design space and the response is not fully exploited. Future research might evaluate the potential of locally adapted smoothing.

Appendix

The weighted score function is given by

$$s_w(\beta) = s(\beta) + s_\alpha(\beta)$$

where

$$s(\beta) = \sum_{i=1}^{n} x_i \epsilon_i \frac{\partial h(\eta_i)}{\partial \eta} \sigma_i^{-2}(y_i - \pi_i)$$

$$s_\alpha(\beta) = \sum_{i=1}^{n} x_i \epsilon_i \frac{\partial h(\eta_i)}{\partial \eta} \sigma_i^{-2}\alpha_i(1 - 2y_i)$$

For the first order approximation

$$\hat{\beta} - \beta = \left(-\frac{\partial s_w(\beta)}{\partial \beta'}\right)^{-1} s_w(\beta)$$

derivatives are needed. They have the form

$$-\frac{\partial s}{\partial \beta'} = F + \sum_{i=1}^{n} x_i {x_i}' \epsilon_i \frac{\partial^2 h}{\partial \eta^2} \{y_i - \pi_i\}$$

where F is the weighted Fisher matrix

$$F = \sum_{i=1}^{n} x_i {x_i}' \epsilon_i {\sigma_i}^{-2} \left(\frac{\partial h}{\partial \eta}\right)^2 . \tag{7}$$

and

$$-\frac{\partial s_\alpha}{\partial \beta'} = -R_\alpha = \sum_{i=1}^{n} x_i {x_i}' \epsilon_i {\sigma_i}^{-2} \frac{\partial^2 h}{\partial \eta} \alpha_i (1 - 2y_i).$$

Tedious derivation yields

$$cov(s_w(\beta)) = \sum x_i {x_i}' {\varepsilon_i}^2 {\sigma_i}^{-2} \left(\frac{\partial h}{\partial \eta}\right)^2 (1 - 2\alpha_i)^2.$$

Under usual assumptions including $s(\beta) = 0_p(n^{1/2}), \quad F^{-1} = 0(n^{-1})$, and $\alpha = \sup\{\alpha_i\}$ one obtains

$$-\frac{\partial s_w(\beta)}{\partial \beta'} = F - R_\alpha + 0_p(n^{-1/2}) + 0_p({n^{1/2}}_\alpha). \tag{8}$$

If $\alpha = o(n^{-1/2})$ one obtains

$$\left(-\frac{\partial s_w(\beta)}{\partial \beta'}\right)^{-1} = F^{-1} + 0_p(n^{-3/2})$$

yielding

$$\hat{\beta} - \beta = \left(-\frac{\partial s_w}{\partial \beta'}\right)^{-1} s_w(\beta) = F^{-1} s_w + 0_p(n^{-1}).$$

The approximation of the covariance is given by

$$cov(\hat{\beta}) = F^{-1} cov(s_w) F^{-1}$$

with F from (7) and $cov(\hat{\beta})$ from (8).

References

Atkinson, A. and Riani, M. (2000). *Robust Diagnostic Regression Analysis.* New York: Springer-Verlag.

Carroll, R. J. and Pederson, S. (1993). On robustness in the logistic regression model. *Journal of the Royal Statistical Society* **B 55**, 693–706.

Choi, E. and Hall, P. (1999). Data sharpening as a prelude to density estimation. *Biometrika* **86**, 941–947.

Copas, J. B. (1988). Binary regression models for contaminated data (with discussion). *Journal of the Royal Statistical Society* **B 50**, 225–265.

Finney, D. (1947). The estimation from individual records of the relationship between dose and quantal response. *Biometrika* **34**, 320–334.

Fowlkes, E. (1987). Some diagnosties for binary logistic regression via smoothing. *Biometrika* **74**, 503–515.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: the Approach Based on Influence Functions.* New York: Wiley.

Hurvich, C. M., Simonoff, J. S., and Tsai, C. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society* **B 60**, 271–293.

Klinger, A. (1997). Generalized soft-thresholding and varying-coefficient models. Discussion Paper 59, SFB 386, Institut für Statistik, Universität München.

Landwehr, J. M., Pregibon, D., and Shoemaker, A. C. (1984). Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association* **79**, 61–71.

LeCessie, S. and van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics* **41**, 191–201.

Mallows, C. L. (1975). *On some topics in robustness.* Technical Memorandum. Murray Hill: Bell Telephone Laboratories.

Marx, B., Eilers, P., and Smith, E. (1992). Ridge likelihood estimation for generalized linear regression. In R. van der Heijden, W. Jansen, B. Francis,

& G. Seeber (Eds.), *Statistical Modelling*, pp. 227–238. Amsterdam: North-Holland.

Parzen, A., Lipsitz, S., Ibrahim, J., and Klar, N. (2002). An estimate of the odds ratio that always exists. *Journal of Computational and Graphical Statistics* **2**, 420–436.

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

Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics* **38**, 485–498.

Santner, T. J. and Duffy, D. E. (1989). *The Statistical Analysis of Discrete Data*. New York: Springer-Verlag.

Segerstedt, B. (1992). On ordinary ridge regression in generalized linear models. *Commun. Statist. – Theory Meth.* **21**, 2227–2246.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society* **B 58**, 267–288.

p	n	ML	$CV(KL)$	$CV(SE)$	$CV(L1)$	$\ln CV(KL)$	$\ln CV(SE)$	$\ln CV(L1)$
2	10	0.0624	0.0509	0.0511	0.0669	-0.242	-0.222	0.144
	20	0.0341	0.0326	0.0332	0.0378	0.014	-0.026	0.112
4	10	0.1163	0.0662	0.0693	0.0940	-0.613	-0.581	-0.265
	20	0.0563	0.0401	0.0409	0.0591	-0.342	-0.343	0.087
	30	0.0366	0.0310	0.0315	0.0384	-0.144	-0.138	0.043
8	10	0.1496	0.1083	0.1079	0.1105	-0.341	-0.340	-0.336
	20	0.1194	0.0618	0.0650	0.0956	-0.695	-0.660	-0.286
	30	0.0793	0.0461	0.0486	0.0727	-0.530	-0.508	-0.104
16	20	0.1313	0.1002	0.0994	0.1018	-0.271	-0.277	-0.283
	30	0.1362	0.0706	0.0735	0.0954	-0.676	-0.639	-0.430
	40	0.1243	0.0550	0.0575	0.0918	-0.846	-0.813	-0.407

TABLE 1: *Simulation results of squared error loss for logit model with binary covariates with the tuning parameter chosen by cross validation*

p	n	ML	$CV(KL)$	$CV(SE)$	$CV(L1)$	$\ln CV(KL)$	$\ln CV(SE)$	$\ln CV(L1)$
2	10	0.0772	0.0570	0.0591	0.0749	-0.260	-0.246	0.008
	20	0.0315	0.0319	0.0329	0.0365	0.120	-0.131	0.168
4	10	0.1359	0.0746	0.0809	0.1307	-0.625	-0.565	-0.311
	20	0.0616	0.0428	0.0441	0.0590	-0.323	-0.303	0.005
	30	0.0339	0.0303	0.0305	0.0350	-0.117	-0.115	0.027
8	10	0.1447	0.1149	0.1134	0.1104	-0.221	-0.228	-0.281
	20	0.1191	0.0675	0.0682	0.0914	-0.609	-0.590	-0.312
	30	0.0741	0.0442	0.0453	0.0662	-0.471	-0.448	-0.109
16	20	0.1190	0.1050	0.1031	0.0957	-0.125	-0.140	-0.248
	30	0.1180	0.0691	0.0729	0.0894	-0.577	-0.526	-0.338
	40	0.1115	0.0501	0.0537	0.0803	-0.831	-0.773	-0.401

TABLE 2: *Simulation results for logit model with continuous covariates with the tuning parameter chosen by cross validation.*

		Kullback–Leibler		squared error	
n	$\parallel \hat{\beta}_{ML} - \beta \parallel$	$\parallel \hat{\beta} - \beta \parallel$	$\log \parallel \hat{\beta}_{ML} - \beta \parallel / \parallel \hat{\beta} - \beta \parallel$	$\parallel \hat{\beta}_{ML} - \beta \parallel$	$\log \parallel \hat{\beta}_{ML} - \beta \parallel / \parallel \hat{\beta} - \beta \parallel$
10	36.857	12.469	-1.958	12.545	-1.947
20	53.670	2.340	-2.975	4.793	-2.810
30	32.032	1.472	-1.941	3.972	-1.830

TABLE 3: *Simulation results for logit model with eight binary covariates and $n = 30$ with the tuning parameter chosen by cross validation.*

	Squared error	log proportion $\log(SE(\hat{\beta})/SE(\hat{\beta}_{ML}))$	$\parallel \hat{\beta} - \beta \parallel$	log proportion $\log(\parallel \hat{\beta} - \beta \parallel / \parallel \beta_{ML} - \beta \parallel)$
ml	0.079		32.023	
resistant, $\gamma = 1.345^2$	0.088	0.152	32.128	0.246
response smoothing estimate	0.046	-0.530	1.479	-1.941
response smoothing and resistant, $\gamma = 1.345^2$	0.043	-0.607	1.431	-1.974
response smoothing and resistant, γ cross validated	0.043	-0.603	1.466	-1.959

TABLE 4: *Squared error loss and distance to true parameter with mean improvement on individual data sets*

	Squared error	log proportion $\log(SE(\hat{\beta})/SE(\hat{\beta}_{ML}))$	$\parallel \hat{\beta} - \beta \parallel$	log proportion $\log(\parallel \hat{\beta} - \beta \parallel / \parallel \beta_{ML} - \beta \parallel)$
ml	0.084		8.761	
resistant, $\gamma = 1.345$	0.098	0.165	9.782	0.283
response smoothing estimate	0.069	-0.143	1.622	-0.683
response smoothing and resistant, $\gamma = 1.345^2$	0.063	-0.237	1.566	-0.721

TABLE 5: *Squared error loss and distance to true parameter with 5 contaminated data*

	β_0	$\log(Vol)$	$\log(Rate)$
Maximum likelihood	-2.922	5.218	4.629
response smoothing $\alpha_{CV} = 0.05$	-1.657	3.405	2.763
resistant estimate	-5.328	8.584	7.609
resistant with response smoothing $\alpha_{CV} = 0.05, \gamma = 1.345^2$	-1.947	3.768	3.074
resistant with response smoothing	-2.303	4.198	3.463

TABLE 6: *Estimates for vaso-constriction data based on logistic model*

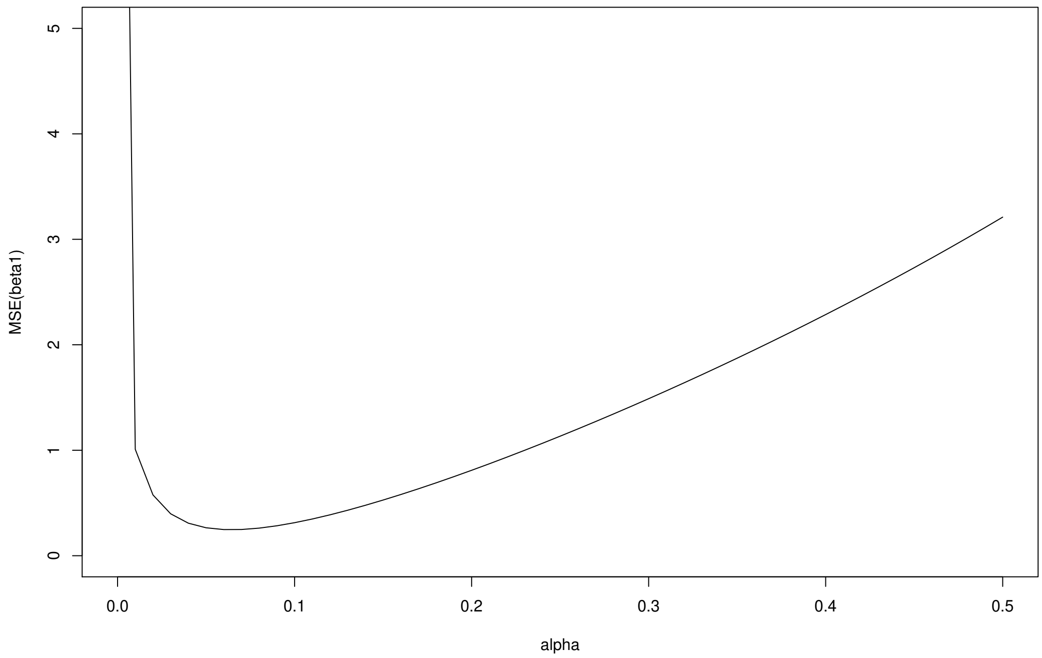

FIGURE 1: *Mean squared error of* β_1 *for* $n_1 = n_2 = 10$, $\pi_1 = 0.1$, $\pi_2 = 0.8$ *plotted against the tuning parameter* α

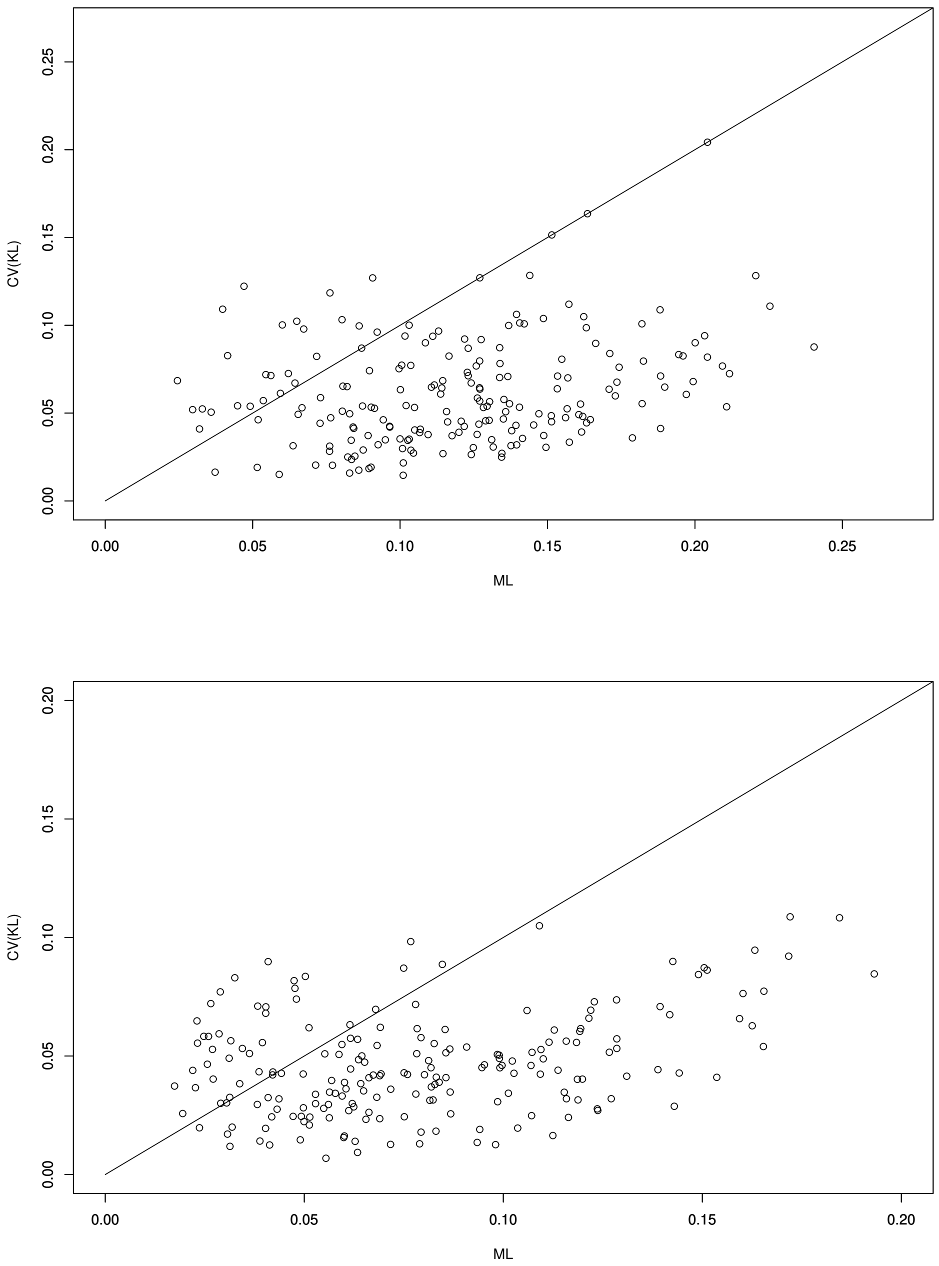

FIGURE 2: *Squared error loss of the maximum likelihood estimate against the response smoothing estimate (200 simulations, logit model with eight predictors; top: sample size n=20, bottom: sample size n=30)*

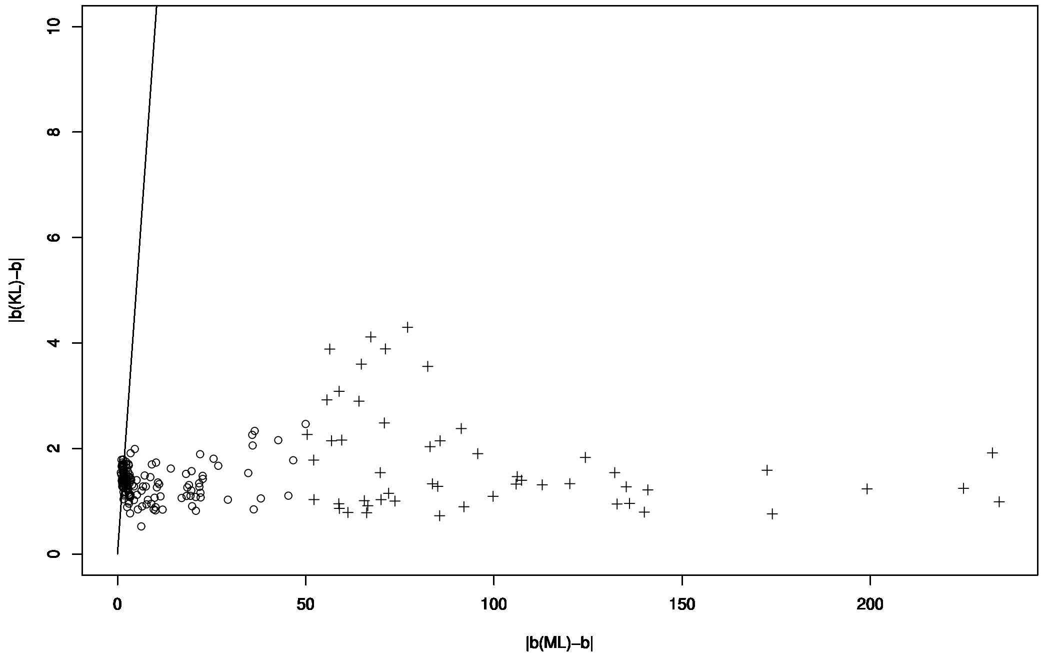

FIGURE 3: $\| \beta_w - \beta \|$ *against* $\| \beta_{ML} - \beta \|$ *for binary covariates* $(n = 30,\, p = 8)$

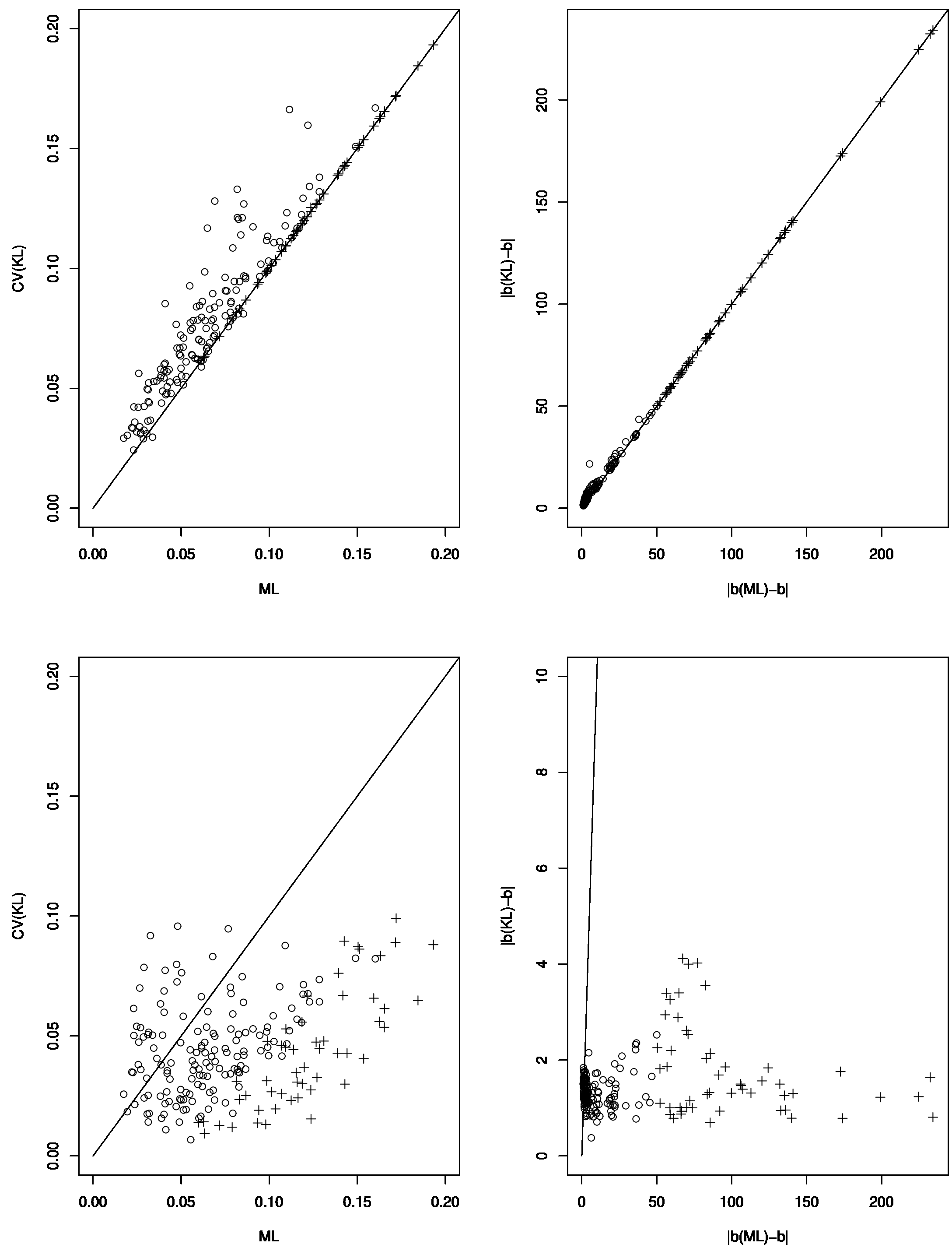

FIGURE 4: *Left panels correspond to squared error loss, right panels to* $\| \hat{\beta} - \beta \|$, *upper panels show the resistant estimate, lower panels show the resistant estimate combined with response smoothing.*

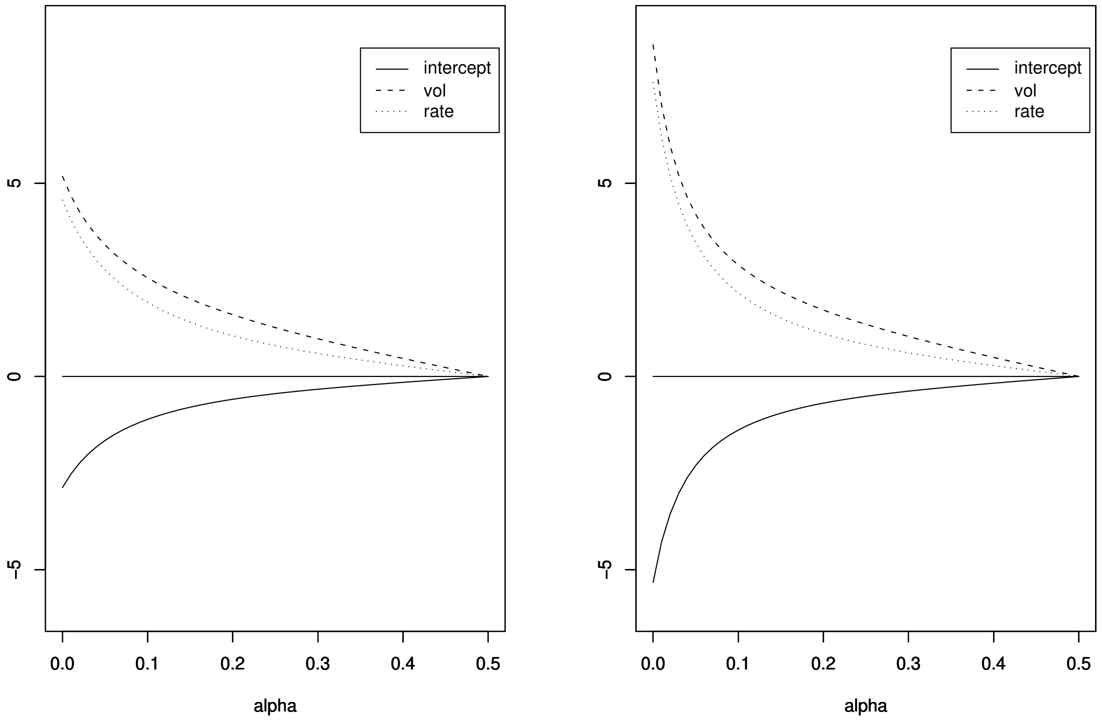

FIGURE 5: *Response smoothing estimate (left) and Pregibon's estimate with response smoothing (right) plotted against α for vaso–constriction data*