COMSTA 2620
pp: 1–18 (col.fig.: nil)
PROD. TYPE: COM
ED: Jolly
PAGN: Mamatha – SCAN: Raj

ARTICLE IN PRESS

ELSEVIER Computational Statistics & Data Analysis ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Robustness against separation and outliers in logistic regression

Peter J. Rousseeuw[a], Andreas Christmann[b,*]

[a] *Department of Mathematics and Computer Science, Universitaire Instelling Antwerpen (UIA), Universiteitsplein 1, B-2610 Wilrijk, Belgium*

[b] *University of Dortmund, HRZ, Abteilung A1 University of Dortmund, D-44221 Dortmund, Germany*

Received 30 September 2002; received in revised form 30 September 2002

Abstract

The logistic regression model is commonly used to describe the effect of one or several explanatory variables on a binary response variable. It suffers from the problem that its parameters are not identifiable when there is separation in the space of the explanatory variables. In that case, existing fitting techniques fail to converge or give the wrong answer. To remedy this, a slightly more general model is proposed under which the observed response is strongly related but not equal to the unobservable true response. This model will be called the hidden logistic regression model because the unobservable true responses are comparable to a hidden layer in a feedforward neural net. The maximum estimated likelihood estimator is proposed in this model. It is robust against separation, always exists, and is easy to compute. Outlier-robust estimation is also studied in this setting, yielding the weighted maximum estimated likelihood estimator.

Keywords: Binary regression; Hidden layer; Neural net; Overlap; Robustness

1. Introduction

The logistic regression model assumes independent Bernoulli distributed response variables with success probabilities $\Lambda(x_i'\theta)$ where Λ is the logistic distribution function, $x_i \in \mathbb{R}^p$ are vectors of explanatory variables, $1 \leqslant i \leqslant n$, and $\theta \in \mathbb{R}^p$ is unknown. Under these assumptions, the classical maximum likelihood (ML) estimator has certain asymptotic optimality properties. However, even if the logistic regression assumptions

* Corresponding author. Tel.: +49-231-755-2763; fax: +49-231-755-2731.
E-mail address: a.christmann@hrz.uni-dortmund.de (A. Christmann).

0167-9473/02/$ - see front matter © 2002 Published by Elsevier Science B.V.
PII: S0167-9473(02)00304-3

are satisfied there are data sets for which the ML estimate does not exist. This occurs for exactly those data sets in which there is no overlap between successes and failures, cf. Albert and Anderson (1984) and Santner and Duffy (1986). This identification problem is not limited to the ML estimator but is shared by all estimators for logistic regression, such as that of Künsch et al. (1989).

One way to approach this problem is to measure the amount of overlap. This can be done by exploiting a connection between the notion of overlap and the notion of regression depth proposed by Rousseeuw and Hubert (1999a), leading to the algorithm of Christmann and Rousseeuw (2001). A comparison between this approach and the support vector machine is given in Christmann et al. (2002).

Of course, finding that there is no overlap in the data set does not imply that the underlying population distributions have no overlap, and the practitioner often needs to obtain regression estimates and odds ratios anyway, e.g. in a comparative study.

In Section 2 we adopt a different approach, based on a slight extension of the logistic regression model. This model assumes that due to an additional stochastic mechanism the true response of a logistic regression model is unobservable, but that there exists an observable variable which is strongly related to the true response. E.g., in a medical context there is often no perfect laboratory test procedure to detect whether a specific illness is present or not (i.e., misclassification errors may sometimes occur). In that case, the true response (whether the disease is present) is not observable, but the result of the laboratory test is.

It can be argued that the true unobservable responses are comparable to a hidden layer in a feedforward neural network model, which is why we call this the hidden logistic regression (HLR) model. In Section 3 we propose the maximum estimated likelihood (MEL) technique in this model, and show that it is immune to the identification problem described above. In Section 4 we consider outlier–robust estimation in this setting. The MEL estimator and its robustification are studied by simulations (Section 5) and on real data sets (Section 6). Section 7 provides a discussion and an outlook to further research.

2. The hidden logistic regression model

The classical logistic regression model assumes n observable independent responses Y_i with Bernoulli distributions $\mathrm{Bi}(1, \Lambda(x_i'\theta))$, where $i = 1, \ldots, n$ and $\theta \in \mathbb{R}^p$. Throughout this paper we assume that there is an intercept, so we put $x_{i,1} = 1$ for all i, and thus $p \geqslant 2$.

The new model assumes that the true responses are unobservable (latent) due to an additional stochastic mechanism. In medical diagnosis there is typically no test procedure (e.g. a blood test) which is completely free of misclassification errors. Another possible cause of misclassifications is the occurrence of clerical errors, which could be made when registering the response variable or (perhaps more often) one of the explanatory variables.

To clarify the model, let us first consider a medical application with only $n=1$ patient. His/her true status (e.g. presence or absence of the disease) has two possible values,

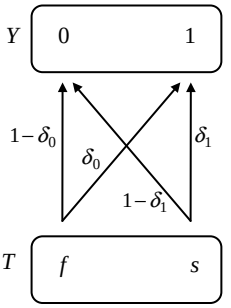

Fig. 1. Unobservable truth T and observable response Y.

typically denoted as success (s) and failure (f). We assume that the true status T is unobservable. However, we can observe the variable Y which is strongly related to T as in Fig. 1. If the true status is $T=s$ we observe $Y=1$ with probability $P(Y=1|T=s)=\delta_1$, hence a misclassification occurs with probability $P(Y=0|T=s)=1-\delta_1$. Analogously, if the true status is f we observe $Y=1$ with probability $P(Y=1|T=f)=\delta_0$ and we obtain $Y=0$ with probability $P(Y=0|T=f)=1-\delta_0$. We of course assume that the probability of observing the true status is higher than 50%, i.e. $0<\delta_0<0.5<\delta_1<1$.

Ekholm and Palmgren (1982) considered the general case with n observations. In our notation, there are n unobservable independent random variables T_i resulting from a classical logistic regression model with finite parameter vector $\theta=(\theta_1,\ldots,\theta_p)'=(\alpha,\beta_1,\ldots,\beta_{p-1})'$. Hence T_i has a Bernoulli distribution with success probability $\pi_i=\Lambda(x_i'\theta)$, where $\Lambda(z)=1/[1+\exp(-z)]$ and $x_i\in\mathbb{R}^p$. Furthermore, they assume that the observable responses Y_i are related to T_i as in Fig. 1. For instance, when $T_i=s$ we obtain $Y_i=1$ with probability $P(Y_i=1|T_i=s)=\delta_1$ whereas $Y_i=0$ occurs with the complementary probability $P(Y_i=0|T_i=s)=1-\delta_1$. (The plain logistic model assumes $\delta_0=0$ and $\delta_1=1$.) We call the entire mechanism in Fig. 2 as hidden logistic regression model because the true status T_i is hidden by the stochastic structure in the top part of Fig. 2. This model can be interpreted as a special kind of neural net, with a single hidden layer that corresponds to the latent variable T.

3. The maximum estimated likelihood (MEL) method

3.1. Construction

We now need a way to fit data sets arising from the hidden logistic model. Two approaches already exist, by Ekholm and Palmgren (1982) and by Copas (1988), but here we will deliberately proceed in a different way.

Let us start by looking only at Fig. 1, where Y is observed but T is not. How can we then estimate T? This is actually the smallest nontrivial estimation problem, because any such problem needs more than one possible value of the parameter and more than one possible outcome. Here, we have exactly two values for both, and the

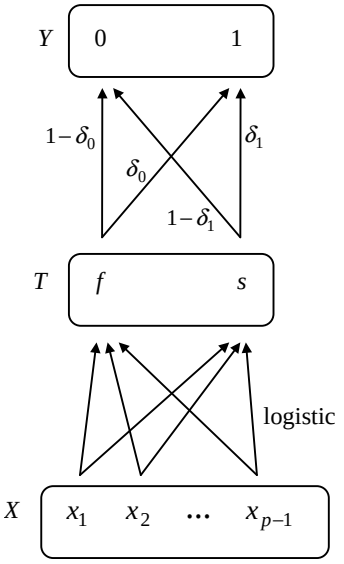

Fig. 2. Hidden logistic regression model.

only distributions on two possible outcomes are the Bernoulli distributions. Under f the likelihood of $Y=0$ exceeds that of $Y=1$, and under s the opposite holds. Therefore, the maximum likelihood estimator of T given $(Y=y)$ becomes simply

$$\hat{T}_{\mathrm{ML}}(Y=0)=f,$$
$$\hat{T}_{\mathrm{ML}}(Y=1)=s, \tag{1}$$

which conforms with intuition.

Let us now consider the conditional probability that $Y=1$ given $\hat{T}_{\mathrm{ML}}$, yielding

$$P(Y=1|\hat{T}_{\mathrm{ML}})=\delta_0 \quad \text{if } y=0,$$
$$=\delta_1 \quad \text{if } y=1, \tag{2}$$

where y is the observed value of Y. Denoting (2) by $\tilde{Y}$, we can rewrite it as

$$\tilde{Y}=\delta_0+(\delta_1-\delta_0)Y=(1-Y)\delta_0+Y\delta_1,$$

which is a weighted average of δ_0 and δ_1 with weights $1-Y$ and Y.

In the model with n observations y_i we obtain analogously

$$\tilde{y}_i=(1-y_i)\delta_0+y_i\delta_1, \tag{3}$$

which we will call the pseudo-observations. In words, the pseudo-observation $\tilde{y}_i$ is the success probability conditional on the most likely estimate of the true status t_i. Note that the pseudo-observations are the result of a deterministic transformation of the y_i so we are not adding any noise to the data.

We now want to fit a logistic regression to the pseudo-observations $\tilde{y}_i$. (In the classical case, $\tilde{y}_i = y_i$.) There are several estimation methods, but here we will apply the maximum likelihood formula. The goal is thus to maximize

$$L(\theta|(\tilde{y}_1,\ldots,\tilde{y}_n)) = \prod_{i=1}^{n} [\Lambda(x_i'\theta)]^{\tilde{y}_i} [1-\Lambda(x_i'\theta)]^{1-\tilde{y}_i} \tag{4}$$

over $\theta \in \mathbb{R}^p$. We call (4) the estimated likelihood because we do not know the true likelihood, which depends on the unobservable $t_1,\ldots,t_n$. (We only know the true likelihood when $\delta_0 = 0$ and $\delta_1 = 1$.) The maximizer $\hat{\theta}$ of (3) can thus be called the MEL estimator.

Note that the MEL approach does not shrink the cumulative distribution function Λ of the logistic distribution (as in Copas, 1988), but instead the responses are transformed to values lying in a narrower interval than $(0,1)$. To illustrate this, Fig. 3 plots pseudo-observations $\tilde{y}_i$ (from a data set with a Bernoulli response variable) versus the linear predictor $x_i'\theta$. The curve is the actual logistic cdf with limits 0 and 1, but the pseudo-observations $\tilde{y}_i$ lie strictly between 0 and 1.

In order to compute the MEL estimator we can take the logarithm of (4), yielding

$$\sum_{i=1}^{n} \tilde{y}_i \ln(\Lambda(x_i'\theta)) + (1-\tilde{y}_i)\ln(1-\Lambda(x_i'\theta)), \tag{5}$$

which always exists since θ is finite. Differentiating with respect to θ yields the (p-variate) score function

$$s(\theta|(\tilde{y}_1,\ldots,\tilde{y}_n)) = \sum_{i=1}^{n} (\tilde{y}_i - \Lambda(x_i'\theta))\, x_i \tag{6}$$

for all $\theta \in \mathbb{R}^p$. Setting (6) equal to zero yields the desired estimate.

3.2. Properties of the MEL estimator

Unlike the classical ML estimator, the MEL estimator always exists.

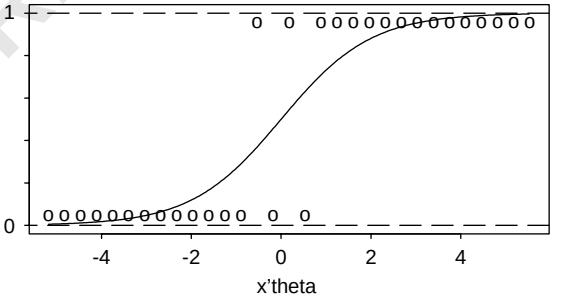

Fig. 3. Plot of pseudo-observations $\tilde{y}_i$ versus $x'\theta$.

Property 1. When $0 < \delta_0 < \delta_1 < 1$ and the data set has a design matrix of full column rank, the MEL estimator always exists and is unique.

(Note that when the design matrix is not of full column rank, we can first reduce the dimension of the x_i by means of principal component analysis.)

Proof. The Hessian matrix of (5) equals

$$\frac{\partial}{\partial\theta}s(\theta) = -\sum_{i=1}^{n} \Lambda(x_i'\theta)(1-\Lambda(x_i'\theta))\,x_i x_i' \tag{7}$$

and is thus negative definite because the design matrix has rank p. Therefore, the differentiable function (5) is strictly concave. Now let us take any $\theta \neq 0$ and replace θ in (5) by $\lambda\theta$. If we let $\lambda \to +\infty$ then (5) always tends to $-\infty$ because there is at least one x_i in the data set with $x_i'\theta \neq 0$ due to full rank, and neither $\tilde{y}_i$ or $(1-\tilde{y}_i)$ can be zero. Therefore, there must be a finite maximizer $\hat{\theta}_{\mathrm{MEL}}$ of (5), which is unique because the concavity is strict. □

This implies that the MEL estimator exists even when the data set has no overlap. Therefore, the resulting odds ratios $\mathrm{OR}_j = \exp(\hat{\theta}_j)$ also always exist, i.e. they are never zero or $+\infty$.

A property shared by all logistic regression estimators is x-affine equivariance. This says that when the x_i are replaced by $x_i^* = Ax_i$, where A is a nonsingular $p \times p$ matrix, then the regression coefficients transform accordingly.

Property 2. The MEL estimator is x-affine equivariant.

Proof. From (6) it follows that $\hat{\theta}^*_{\mathrm{MEL}} = (A')^{-1}\hat{\theta}_{\mathrm{MEL}}$, hence $(x_i^*)'\hat{\theta}^*_{\mathrm{MEL}} = x_i'A'(A')^{-1}\hat{\theta}_{\mathrm{MEL}} = x_i'\hat{\theta}_{\mathrm{MEL}}$. This also yields the same predicted values. □

In linear regression there exist two other types of equivariance: one about adding a linear function to the response ('regression equivariance') and one about multiplying the response by a constant factor ('y-scale equivariance'), but these obviously do not apply to logistic regression.

3.3. Choice of δ_0 and δ_1

If δ_0 and δ_1 are known from the context (e.g. from the type I and type II error probabilities of a blood test) then we can use these values. But in many cases, δ_0 and δ_1 are not given in advance. Copas (1988, p. 241) found that accurate estimation of δ_0 and δ_1 from the data itself is very difficult, if not impossible unless n is extremely large. He essentially considers them as tuning constants that can be chosen, as do we. The 'symmetric' approach used by Copas is to choose a single constant $\gamma > 0$ and to set

$$\delta_0 = \gamma \quad \text{and} \quad \delta_1 = 1-\gamma. \tag{8}$$

Table 1
Bias and standard error of some estimators: ML, MEL with $\delta = 0.01$, WEMEL with $\delta = 0.01$ based on MCD, and KSC with tuning constant $u = 3.2\sqrt{p}$

		ML		MEL		WEMEL		KSC	
n		Bias	SE	Bias	SE	Bias	SE	Bias	SE
Case A with $\theta = (1,0,0)'$ and $x_{i,j}$ from $N(0,1)$									
20	α	0.291	0.032	0.272	0.028	0.270	0.029	0.312	0.034
	β_1	0.010	0.031	0.009	0.029	0.012	0.031	0.061	0.042
	β_2	−0.014	0.035	−0.004	0.030	0.004	0.031	0.012	0.035
50	α	0.097	0.012	0.095	0.012	0.088	0.012	0.096	0.012
	β_1	−0.015	0.011	−0.015	0.011	−0.024	0.012	−0.017	0.011
	β_2	−0.021	0.012	−0.021	0.012	−0.022	0.012	−0.020	0.012
100	α	0.053	0.008	0.052	0.008	0.052	0.008	0.054	0.008
	β_1	0.004	0.008	0.004	0.008	0.000	0.008	0.004	0.008
	β_2	−0.004	0.008	−0.004	0.008	−0.001	0.008	−0.004	0.008
Case B with $\theta = (1,1,2)'$ and $x_{i,j}$ from $N(0,1)$									
20	α	0.586	0.067	0.360	0.039	0.373	0.040	0.492	0.057
	β_1	0.652	0.083	0.364	0.045	0.386	0.048	0.550	0.067
	β_2	1.372	0.159	0.780	0.057	0.792	0.062	1.198	0.107
50	α	0.133	0.019	0.097	0.017	0.101	0.017	0.167	0.023
	β_1	0.156	0.022	0.104	0.019	0.108	0.020	0.198	0.026
	β_2	0.350	0.030	0.247	0.025	0.249	0.026	0.419	0.037
100	α	0.061	0.011	0.038	0.010	0.042	0.010	0.064	0.011
	β_1	0.085	0.012	0.050	0.011	0.050	0.011	0.088	0.012
	β_2	0.154	0.016	0.084	0.015	0.095	0.015	0.167	0.017
Case C with $\theta = (1,1,2)'$ and $x_{i,j}$ from Student's t_3									
20	α	0.621	0.090	0.291	0.041	0.323	0.045	0.487	0.054
	β_1	0.699	0.084	0.324	0.034	0.364	0.036	0.589	0.065
	β_2	1.486	0.207	0.628	0.053	0.719	0.061	1.117	0.090
50	α	0.227	0.030	0.121	0.018	0.132	0.019	0.292	0.035
	β_1	0.250	0.035	0.116	0.017	0.136	0.018	0.307	0.037
	β_2	0.525	0.071	0.241	0.027	0.275	0.029	0.661	0.078
100	α	0.081	0.012	0.038	0.011	0.047	0.012	0.090	0.013
	β_1	0.107	0.012	0.044	0.011	0.057	0.011	0.117	0.013
	β_2	0.186	0.018	0.061	0.015	0.094	0.016	0.205	0.020

His computations require that γ be small enough so that terms in γ^2 can be ignored. In his Table 1 the values $\gamma = 0.01$ and $\gamma = 0.02$ occur, whereas he considers $\gamma = 0.05$ to be unreasonably high (p. 238). In most of Copas' examples $\gamma = 0.01$ performs well, and this turns out to be true also for our MEL method, so we could use $\gamma = 0.01$ as the default choice. This approach has the advantage of simplicity.

On the other hand, there is something to be said for an 'asymmetric' choice which takes into account how many y_i's are 0 and 1 in the data set. Let us consider the marginal distribution of the y_i (that is, unconditional on the x_i) from which we construct some estimate $\hat{\pi}$ of the marginal success probability $P(Y = 1)$. It seems reasonable to

constrain δ_0 and δ_1 such that the average of the pseudo-observations $\tilde{y}_i$ corresponds to $\hat{\pi}$. This yields

$$\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n}\tilde{y}_i = (1-\hat{\pi})\delta_0 + \hat{\pi}\delta_1,$$

$$\hat{\pi} - \hat{\pi}\delta_1 = \delta_0 - \hat{\pi}\delta_0,$$

$$\frac{1-\delta_1}{\delta_1 - \hat{\pi}} = \frac{\delta_0}{\hat{\pi} - \delta_0}.$$

Since it is natural to assume that $\delta_0 < \hat{\pi} < \delta_1$, the latter ratios equal a (small) positive number which we will denote by δ. Consequently, we can write both δ_0 and δ_1 as functions of δ, as

$$\delta_0 = \frac{\hat{\pi}\delta}{1+\delta} \quad \text{and} \quad \delta_1 = \frac{1+\hat{\pi}\delta}{1+\delta}. \tag{9}$$

However, since we have assumed that $\delta_0 < \hat{\pi} < \delta_1$ we have to construct $\hat{\pi}$ accordingly. We cannot take the standard estimate $\bar{\pi} = 1/n\sum_{i=1}^{n} y_i = (\text{number of } y_i = 1)/n$ because $\bar{\pi}$ can become 0 or 1. A natural idea is to bound $\bar{\pi}$ away from 0 and 1 by putting

$$\hat{\pi} = \max(\delta, \min(1-\delta, \bar{\pi})), \tag{10}$$

which means truncation at δ and $1-\delta$. This is sufficient because always

$$\delta_0 = \frac{\hat{\pi}\delta}{1+\delta} < \frac{\hat{\pi}+\hat{\pi}\delta}{1+\delta} = \hat{\pi}$$

and

$$\delta_1 = \frac{1+\hat{\pi}\delta}{1+\delta} > \frac{\hat{\pi}+\hat{\pi}\delta}{1+\delta} = \hat{\pi},$$

hence $\delta_0 < \hat{\pi} < \delta_1$. Note that both misclassification probabilities in Fig. 1 are less than δ because

$$\delta_0 = \frac{\hat{\pi}\delta}{1+\delta} < \frac{\delta}{1+\delta} < \delta$$

and

$$1-\delta_1 = \frac{1+\delta-1-\hat{\pi}\delta}{1+\delta} = \frac{(1-\hat{\pi})\delta}{1+\delta} < \frac{\delta}{1+\delta} < \delta.$$

Our default choice will be $\delta = 0.01$, which implies smaller classification errors than by putting $\gamma = 0.01$ in formula (8).

When the data are 'balanced' in the sense that there are as many $y_i = 1$ as $y_i = 0$, expression (10) yields $\hat{\pi} = 0.5$, hence $\delta_0 = 1-\delta_1$ by (9), yielding identical misclassification probabilities, as in the symmetric formulas (8). In all other 'unbalanced' cases, our asymmetric approach yields less biased predictions. An extreme case is when all $y_i = 1$. (This is a situation where the classical ML estimator does not exist.) The MEL estimator will put all $\tilde{y}_i = \delta_1$ yielding a fit with all slopes $\hat{\beta}_1 = \cdots = \hat{\beta}_{p-1} = 0$ and with intercept $\alpha = \Lambda^{-1}(\delta_1)$. Using the symmetric approach (8) yields $\delta_1 = 0.99$

hence $\hat{\alpha} = \text{logit}(0.99) = \ln(99) \approx 4.595$ and so the fitted values are constant and equal to 0.99. On the other hand, the asymmetric approach yields $\hat{\pi} = 0.99$ and $\delta_1 = (1 + (0.99)(0.01))/(1 + 0.01) = 1.0099/1.01 = 0.999901$. This again yields zero slopes but a larger intercept $\hat{\alpha} = \text{logit}(0.999901) \approx \ln(10099) \approx 9.22$, so the fitted values are 0.9999 which is much closer to 1.

Our recommendation is to compute $\hat{\pi}$, δ_0, and δ_1 as in (9) and (10) with $\delta = 0.01$, to compute the pseudo-observations $\tilde{y}_i$ according to (3) and to carry out the resulting MEL method.

Our S-PLUS code for this method and its robustification described in Section 4 can be downloaded from

http://win-www.uia.ac.be/u/statis/Robustn.htm or
http://www.statistik.uni-dortmund.de/sfb475/berichte/rouschr2.zip.

The ML estimator has the nice property under the logistic regression model that if $\hat{\theta}$ is the ML estimate for the data set $\{(x_i', y_i), 1 \leqslant i \leqslant n\}$, then $-\hat{\theta}$ is the ML estimate for the data set $\{(x_i', 1 - y_i), 1 \leqslant i \leqslant n\}$. Hence, recoding all response variables Y_i to $1 - Y_i$ affects the ML estimator only in the way that it changes the signs of the regression coefficients, and the odds ratios become $\exp(-\hat{\theta}_j) = 1/\text{OR}_j$. We call this equivariance with respect to recoding the response variable. The MEL estimator has the same property, whether δ_0 and δ_1 are given by (8) or (9).

Property 3. The MEL estimator is equivariant with respect to recoding the response variable.

Proof. Writing $y_i^* = 1 - y_i$ and recomputing (10) and (9) (or (8)) yields $\tilde{y}_i^* = 1 - \tilde{y}_i$ by (3) .Applying the ML estimator to the $(x_i', \tilde{y}_i^*)$ yields the desired result. □

4. Outlier–robust estimation

Like the ML method, the MEL estimator still has the disadvantage that it is not robust to outliers because the impact of bad leverage points is unbounded in (6). In this section we consider a robustification of the MEL estimator.

The score function (6) is similar to an M-estimator equation. Since the (pseudo-) residuals are always bounded in binary regression models due to

$$|\tilde{y}_i - \Lambda(x_i'\theta)| < 1,$$

the main problem comes from the factor x_i which need not be bounded. This corresponds to the leverage point issue. We propose to downweight leverage points in a straightforward robust manner, yielding the weighted maximum estimated likelihood (WEMEL) estimator defined as the solution $\hat{\theta}$ of

$$\sum_{i=1}^{n} (\tilde{y}_i - \Lambda(x_i'\theta)) w_i x_i = 0, \tag{11}$$

where the weights w_i only depend on how far away x_i is from the bulk of the data. We use

$$w_i = \frac{M}{\max\{\mathrm{RD}^2(x_i^*), M\}}, \tag{12}$$

where $x_i^* = (x_{i,2}, \ldots, x_{i,p})' \in \mathbb{R}^{p-1}, \mathrm{RD}(x_i^*)$ is its robust distance, and M is the 75th percentile of all $\mathrm{RD}^2(x_j^*)$, $j = 1, \ldots, n$. This means that we give a weight less than 1 to the 25% most extreme design points.

When all regressor variables are continuous and there are not more than (say) 30 of them, we can use the robust distances that come out of the minimum covariance determinant (MCD) estimator of Rousseeuw (1984), for which the fast algorithm of Rousseeuw and Van Driessen (1999b) is available. This algorithm has been incorporated in the packages S-Plus (as the function cov.mcd) and SAS/IML (as the routine MCD), and both provide the robust distances in their output. We propose to compute the MCD with the default value of $q = 0.75$ which gives a breakdown point of approximately 25%. In case that not all regressor variables are continuous or there are many of them (even more than 1000), we can use the robust distances provided by the robust principal components algorithm of Hubert et al. (2002).

The WEMEL estimate is easy to compute because most GLM algorithms (including the one in S-Plus) allow the user to input prior weights w_i. After computing the weights w_i the computation time is approximately equal to the computing time for the ML estimator. Hence, the WEMEL estimate can be computed even for large high-dimensional data sets.

Obviously, Properties 1–3 are also valid for the WEMEL estimator.

Before we investigate the robustness properties of the WEMEL estimator in more detail in the next section let us consider four artificial data sets. All these data sets have the following 10 observations in common:

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{pmatrix}' \quad \text{and}$$

$$\mathbf{y} = (0,0,0,0,0,1,1,1,1,1)'.$$

In cases (a) and (b) there is one additional observation $(x_{11,2}, y_{11}) = (11, 1)$, and $(x_{11,2}, y_{11}) = (11, 0)$, respectively. In case (c) there is one bad leverage point $(x_{11,2}, y_{11}) = (15, 0)$, and in case (d) there are two bad leverage points $(x_{11,2}, y_{11}) = (15, 0)$ and $(x_{12,2}, y_{12}) = (-5, 1)$. The estimated success probability curves with respect to the ML, MEL, WEMEL estimator both with our default value of $\delta = 0.01$ and the M-estimator (denoted by KSC) proposed by Künsch et al. (1989) with tuning constant $u = 3.2\sqrt{p}$ are given in Fig. 4.

In case (a) the ML and the M-estimate do not exist due to complete separation, so Fig. 4a only shows the fitted curves of the MEL and WEMEL estimates which exist for all data sets. All four estimates behave similarly in Fig. 4b, because there are no bad leverage points in case (b). In Figs. 4c and d the ML and MEL estimates are highly influenced by one or two bad leverage points. The M-estimator shows reasonable

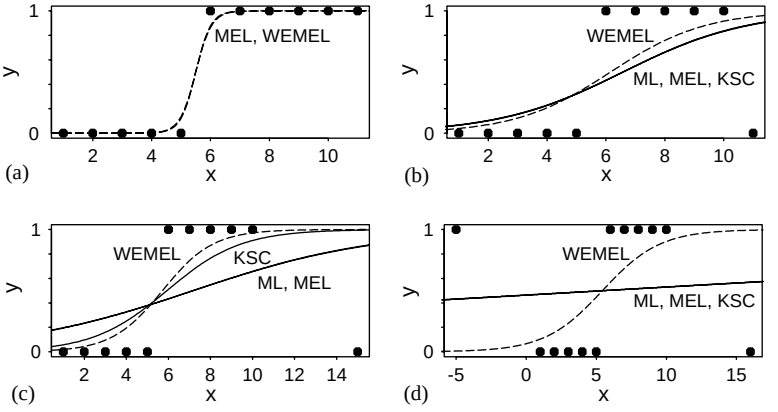

Fig. 4. Estimated success probability curves for four estimators.

robustness in case (c) but not if there are two bad leverage points. The WEMEL estimator shows good robustness even in case (d).

5. Simulations

In this section we carry out some simulations to compare the bias and the standard error of the usual ML estimator, a conditionally unbiased bounded influence M-estimator proposed by Künsch et al. (1989), and the estimators MEL and WEMEL both with $\delta = 0.01$ using our asymmetric choice of δ_0 and δ_1 under the assumptions of the logistic regression model. We will estimate $p = 3$ coefficients, including the intercept term. Both explanatory variables are generated from the standard normal distribution or from Student's t distribution with 3 degrees of freedom to model the case of moderate leverage points. We also consider the case that one or two data points or even 10% of the observations are bad leverage points located at $(1, 10, 10)'$. As true parameter vectors we use $\theta_A = (1, 0, 0)'$ and $\theta_B = (1, 1, 2)'$. The sample sizes n will be 20, 50, and 100. For each situation 1000 samples are generated.

We use the depth-based algorithm (Christmann and Rousseeuw, 2001) to check whether the data set has overlap, i.e. whether the ML estimate exists. It turned out that not all data sets had overlap. This occurred 12 times in case A for $n = 20$, 129 times in case B for $n = 20$, 178 times in case C for $n = 20$, four times in case C for $n = 50$, three times in case D for $n = 20$, two times in case E for $n = 20$, and 3 times in case F for $n = 20$. The M-estimate proposed by Künsch et al. (1989) does not exist for these data sets, too. For two data sets in case B with $n = 20$ and six data sets in case C the algorithm to compute the M-estimate had convergence problems. These special data sets were not considered in summarizing the results for the M-estimator. This contrasts sharply to the MEL and WEMEL estimates, which existed for all data sets.

Table 2
Bias and standard error of some estimators: ML, MEL with $\delta = 0.01$, WEMEL with $\delta = 0.01$ based on MCD, and KSC with tuning constant $u = 3.2\sqrt{p}$

		ML		MEL		WEMEL		KSC	
n		Bias	SE	Bias	SE	Bias	SE	Bias	SE
Case D with $\theta = (1,1,2)'$ and $x_{i,j}$ from $N(0,1)$, 10% bad leverage points									
20	α	−0.353	0.040	−0.379	0.025	0.185	0.040	−0.373	0.027
	β_1	−1.510	0.023	−1.502	0.019	−0.034	0.034	−1.501	0.026
	β_2	−1.718	0.025	−1.720	0.019	0.191	0.041	−1.687	0.027
50	α	−0.485	0.012	−0.488	0.012	−0.096	0.014	−0.484	0.012
	β_1	−1.426	0.009	−1.420	0.009	−0.234	0.014	−1.411	0.009
	β_2	−1.774	0.009	−1.777	0.009	−0.261	0.017	−1.760	0.009
100	α	−0.495	0.008	−0.497	0.008	−0.127	0.009	−0.494	0.008
	β_1	−1.404	0.006	−1.399	0.006	−0.255	0.009	−1.392	0.006
	β_2	−1.791	0.005	−1.793	0.005	−0.359	0.011	−1.779	0.006
Case E with $\theta = (1,1,2)'$ and $x_{i,j}$ from $N(0,1)$, 1 bad leverage point									
20	α	−0.374	0.022	−0.379	0.022	0.403	0.040	−0.365	0.022
	β_1	−1.461	0.018	−1.452	0.018	0.288	0.044	−1.423	0.019
	β_2	−1.666	0.017	−1.673	0.017	0.752	0.057	−1.631	0.022
50	α	−0.467	0.010	−0.468	0.010	0.059	0.017	−0.061	0.015
	β_1	−1.248	0.009	−1.245	0.009	0.040	0.018	−0.170	0.017
	β_2	−1.598	0.008	−1.604	0.007	0.142	0.023	−0.174	0.022
100	α	−0.388	0.006	−0.391	0.006	0.027	0.010	−0.026	0.010
	β_1	−0.864	0.006	−0.867	0.006	0.023	0.011	−0.065	0.011
	β_2	−1.195	0.006	−1.208	0.006	0.054	0.014	−0.074	0.014
Case F with $\theta = (1,1,2)'$ and $x_{i,j}$ from $N(0,1)$, 2 bad leverage points									
20	α	−0.353	0.040	−0.379	0.025	0.185	0.040	−0.373	0.027
	β_1	−1.510	0.023	−1.502	0.019	−0.034	0.034	−1.501	0.026
	β_2	−1.718	0.025	−1.720	0.019	0.191	0.041	−1.687	0.027
50	α	−0.439	0.089	−0.440	0.089	0.017	0.099	−0.393	0.093
	β_1	−1.375	0.066	−1.370	0.065	−0.094	0.108	−1.165	0.171
	β_2	−1.705	0.066	−1.709	0.065	−0.063	0.136	−1.437	0.138
100	α	−0.474	0.007	−0.475	0.007	0.010	0.010	−0.110	0.009
	β_1	−1.228	0.006	−1.226	0.005	−0.006	0.011	−0.217	0.010
	β_2	−1.615	0.005	−1.620	0.005	0.011	0.014	−0.306	0.012

Tables 1 and 2 compare bias and standard error of the estimators for data sets with overlap. In case A, where the true slopes are zero, there is not much difference between the estimators. But in cases B and C in which the explanatory variables have an impact on the success probabilities, the MEL and the WEMEL estimator have a substantially smaller bias and standard error than the ML and the M-estimator. This can be explained by the well-known phenomenon that ML estimation in logistic regression tends to overestimate the magnitude of nonzero coefficients for finite sample sizes. For instance, Firth (1993, p. 30) mentions that the ML estimator is biased away from 0 and that a bias correction for this estimator requires some 'shrinkage' towards 0. The

MEL and WEMEL estimators exhibit this kind of shrinkage behavior. The differences between the estimators are largest in case C, where there are moderate good leverage points generated by Student's t_3 distribution.

In situations D to F there are one or more bad leverage points. The only estimator under consideration which works well also in these situations turns out to be the WEMEL estimator. Somewhat surprisingly, the bounded-influence M-estimator proposed by Künsch et al. (1989) is not robust enough in most of these cases. Even if there is only one bad leverage point the M-estimator can behave as bad as the nonrobust ML and MEL estimators for $n = 20$. This is probably a consequence of the fact that Künsch et al. proposed to use an M-estimator to estimate the scatter matrix of the explanatory variables instead of a positive-breakdown estimator.

6. Examples

In this section we consider some benchmark data sets. The banknotes data set (Riedwyl, 1997) has no overlap, hence the ML estimate does not exist. The vaso constriction data (Finney, 1947; Pregibon, 1981) and the food stamp data (Künsch

Table 3
Comparison of estimates: ML, MEL with $\delta = 0.01$, WEMEL with $\delta = 0.01$, and KSC with tuning constant $u = 3.2\sqrt{p}$ or $u^* = 3.7\sqrt{p}$

Data set (n, p)	Method	$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
Banknotes	ML			— Does not exist—				
(200,7)	MEL	147.09	0.46	−1.02	1.33	2.20	2.32	−2.37
	WEMEL	251.86	−0.24	−1.60	2.06	2.09	2.37	−2.16
	KSC, u			— Does not exist—				
Vaso constriction	ML	−2.92	5.22	4.63				
(39,3)	MEL	−2.77	4.98	4.41				
	WEMEL	−2.72	4.94	4.34				
	KSC, u^*	−2.98	5.27	4.67				
	KSC, u	−6.34	9.86	8.77				
Food stamp	ML	0.93	−1.85	0.90	−0.33			
(150,4)	MEL	0.89	−1.83	0.88	−0.33			
	WEMEL[a]	5.24	−1.73	0.60	−1.04			
	WEMEL[b]	4.76	−1.79	0.63	−0.96			
	KSC, u	4.51	−1.78	0.74	−0.93			
Toxoplasmosis	ML	0.10	−0.45	−0.19	0.21			
(697,4)	MEL	0.10	−0.44	−0.19	0.21			
	WEMEL[a]	0.09	−0.40	−0.14	0.19			
	KSC, u	0.10	−0.45	−0.19	0.21			

[a]Weights are computed w.r.t. to the single continuous explanatory variable.
[b]Weights are computed w.r.t. to all explanatory variables using the PCA method.

Table 4
Comparison of odds ratios with respect to: ML, MEL with $\delta = 0.01$, WEMEL with $\delta = 0.01$, and KSC with tuning constant $u = 3.2\sqrt{p}$ or $u^* = 3.7\sqrt{p}$

Data set (n, p)	Method	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
Banknotes	ML	— Does not exist—					
	MEL	1.58	0.36	3.78	9.03	10.18	0.09
	WEMEL	0.79	0.20	7.85	8.09	10.70	0.12
	KSC, u	— Does not exist—					
Vaso constriction	ML	185.03	102.64				
	MEL	146.13	81.97				
	WEMEL	139.77	76.71				
	KSC, u^*	194.42	106.70				
	KSC, u	19 148.89	6438.17				
Food stamp	ML	0.16	2.45	0.72			
	MEL	0.16	2.42	0.72			
	WEMEL[1]	0.18	1.82	0.35			
	WEMEL[2]	0.17	1.88	0.38			
	KSC, u	0.17	2.10	0.39			
Toxoplasmosis	ML	0.64	0.83	1.23			
	MEL	0.64	0.83	1.23			
	WEMEL[1]	0.67	0.87	1.21			
	KSC, u	0.64	0.83	1.23			

[1]Weights are computed w.r.t. to the single continuous explanatory variable.
[2]Weights are computed w.r.t. to all explanatory variables using the PCA method.

et al., 1989) are well known in the literature on outlier detection and robust logistic regression. They both have little overlap: it suffices to delete 3 (resp. 6) observations in these data sets to make the ML estimate nonexistent (see Christmann and Rousseeuw, 2001). Some of these observations are considered as outliers in Künsch et al. (1989). The toxoplasmosis data set (Efron, 1986) is chosen because the ratio n/p is high.

Table 3 shows that the MEL estimates with $\delta = 0.01$ were quite similar to the ML estimates for the data sets with overlap. For the vaso constriction data set, the WEMEL estimate and the M-estimate with tuning constant $u = 3.7\sqrt{p}$ are similar to the ML and the MEL estimates, but the M-estimate with a smaller tuning constant behaves differently because some observations received smaller weights. For the food stamp data set, the WEMEL and the M-estimates differ from the ML and the MEL estimates because there are outliers and at least one leverage point in the data set. The considered estimators behave very similarly for Efron's toxoplasmosis data set. The corresponding results for the odds ratios are given in Table 4.

Fig. 5 shows that the choice of δ has relatively little impact on the MEL estimates for the food stamp data set, which has overlap. In contrast to that, Fig. 6 shows the effect of δ for the banknotes data. Because this data set has no overlap we know that $\|\hat{\theta}\|$ tends to $+\infty$ as δ goes to 0 (since $\delta = 0$ corresponds to the ML estimator).

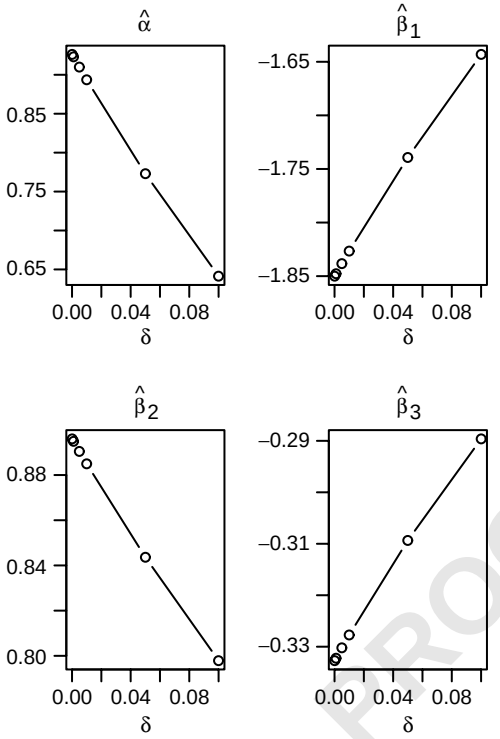

Fig. 5. Graphs of the MEL coefficients versus δ for the food stamp data set, for $\delta = 0.0001$, $0.001, 0.005, 0.01, 0.05$, and 0.1.

One could therefore choose the tuning constant δ as if it were a 'ridge parameter' in Fig. 6.

7. Discussion and outlook

The main problem addressed in this paper is that the coefficients of the binary regression model (with logistic or probit link function) cannot be estimated when the x_i's of successes and failures do not overlap. This is a deficiency of the model itself, because the fit can be made perfect by letting $\|\theta\|$ tend to infinity. Therefore, this problem is shared by all reasonable estimators that operate under the logistic model.

Our approach to resolve this problem is to work with a generalized model, which we call the hidden logistic model. Here we compute the pseudo-observations $\tilde{y}_i$, defined as the probability that $y_i = 1$ conditional on the maximum likelihood estimate of the true status t_i. The resulting MEL estimator always exists and is unique, even though the hypothetical misclassification probabilities (based on our default setting $\delta = 1\%$) are so small that they would not be visible in the observed data.

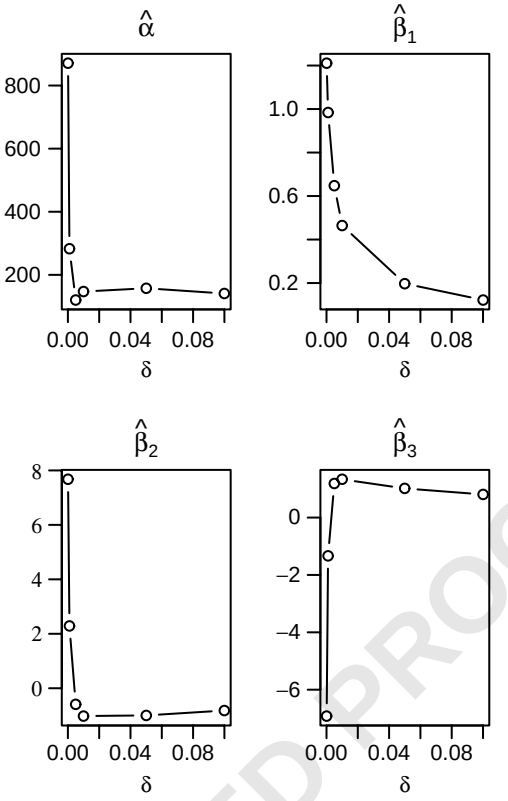

Fig. 6. Graphs of the first four MEL coefficients versus δ for the banknotes data set, for $\delta = 0.0001, 0.001, 0.005, 0.01, 0.05,$ and 0.1.

The hidden logistic model was previously used (under a different name) in an important paper by Copas (1988). However, his approach and ours are almost diametrically opposite. Copas' motivation is to reduce the effect of the outliers that matter, which are the observations (x_i, y_i) where x_i is far away from the bulk of the data and y_i has the value which is very unlikely under the logistic model. In the terminology of Rousseeuw and Van Zomeren (1990) these are bad leverage points. In logistic regression their effect is always to flatten the fit, i.e. to bring the estimated slopes closer to zero. Copas' approach shrinks the logistic distribution function Λ away from 0 and 1 (by letting it range between γ and $1 - \gamma$), so that bad leverage points are no longer that unlikely under his model, which greatly reduces their effect. On the other hand, his approach aggravates the problems that arise when there is little overlap between successes and failures, as in his analysis of the vaso constriction data.

Our approach goes into the other direction: rather than shrinking Λ while leaving the responses y_i unchanged, we leave Λ unchanged and shrink the y_i to the pseudo-observations $\tilde{y}_i$ which are slightly larger than zero or slightly less than 1.

This completely eliminates the overlap problem. It does not help at all for the problem of bad leverage points, but for that problem we can use existing techniques from the robustness literature, yielding the WEMEL method.

We recommend that in general δ and γ should not depend on the sample size n and that these quantities should not converge to zero, because the proportion of typing errors may not decrease to zero. The data quality in very large data sets is not always very good; in fact, often the opposite is true in data mining projects. Actually estimating δ and γ with reasonable accuracy precision seems only possible for huge data sets.

We have not yet addressed the computation of influence functions and breakdown values. As one referee pointed out, it will also be worthwhile to investigate maxbias curves and breakdown functions (He et al., 1990). It would be interesting to connect our work in the hidden logistic model with the existing body of literature on outlier detection, robust estimation and estimation of the median effective dose (ED50) in the classical logistic model, including the work of Pregibon (1982), Künsch et al. (1989), Christmann (1994, 1998), Huang (2001), and Müller and Neykov (2002).

The unobservable true responses in the hidden logistic regression model are comparable to a hidden layer in a feedforward neural net. Recently, Intrator and Intrator (2001) investigated the behavior of artificial neural networks as an alternative to logistic regression in a simulation study.

Acknowledgements

The authors thank the editor and two anonymous referees for helpful comments. The second author was supported by the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures").

References

Albert, A., Anderson, J.A., 1984. On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71, 1–10.

Christmann, A., 1994. Least median of weighted squares in logistic regression with large strata. Biometrika 81, 413–417.

Christmann, A., 1998. On positive breakdown point estimators in regression models with discrete response variables. Habilitation Thesis, Department of Statistics, University of Dortmund.

Christmann, A., Rousseeuw, P.J., 2001. Measuring overlap in logistic regression. Comput. Statist. Data Anal. 37, 65–75.

Christmann, A., Fischer, P., Joachims, T., 2002. Comparison between various regression depth methods and the support vector machine to approximate the minimum number of misclassifications. Comput. Statist. 17, 273–287.

Copas, J.B., 1988. Binary regression models for contaminated data. With discussion. J. Roy. Statist. Soc. B 50, 225–265.

Efron, B., 1986. Double exponential families and their use in generalized linear regression. J. Amer. Statist. Assoc. 81, 709–721.

Ekholm, A., Palmgren, J., 1982. A model for binary response with misclassification. In: Gil-christ, R. (Ed.), GLIM-82, Proceedings of the International Conference on Generalized Linear Models. Springer, Heidelberg, pp. 128–143.

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

Firth, D., 1993. Bias reduction of maximum likelihood estimates. Biometrika 80, 27–38.

He, X., Simpson, D.G., Portnoy, S.L., 1990. Breakdown robustness of tests. J. Amer. Statist. Assoc. 85, 446–452.

Huang, Y., 2001. Interval estimation of the ED50 when a logistic dose–response curve is incorrectly assumed. Comput. Statist. Data Anal. 36, 525–537.

Hubert, M., Rousseeuw, P.J., Verboven, S., 2002. A fast method for robust principal components with applications to chemometrics. Chemometrics Intell. Lab. Systems 60, 101–111.

Intrator, O., Intrator, N., 2001. Interpreting neural-network results: a simulation study. Comput. Statist. Data Anal. 37, 373–393.

Künsch, H.R., Stefanski, L.A., Carroll, R.J., 1989. Conditionally unbiased bounded influence estimation in general regression models, with applications to generalized linear models. J. Amer. Statist. Assoc. 84, 460–466.

Müller, C., Neykov, C., 2002. Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. J. Statist. Plann. Inference, to appear.

Pregibon, D., 1981. Logistic regression diagnostics. Ann. Statist. 9, 705–724.

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

Riedwyl, H., 1997. Lineare Regression und Verwandtes. Birkhäuser, Basel.

Rousseeuw, P.J., 1984. Least median of squares regression. J. Amer. Statist. Assoc. 79, 871–880.

Rousseeuw, P.J., Hubert, M., 1999a. Regression depth. J. Amer. Statist. Assoc. 94, 388–433.

Rousseeuw, P.J., Van Driessen, K., 1999b. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223.

Rousseeuw, P.J., Van Zomeren, B.C., 1990. Unmasking multivariate outliers and leverage points. J. Amer. Statist. Assoc. 85, 651–663.

Santner, T.J., Duffy, D.E., 1986. A note on A. Albert and J.A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. Biometrika 73, 755–758.