Robust adaptive estimators for binary regression models

DANIEL GERVINI
Department of Biostatistics, University of Zürich
Sumatrastrasse 30, 8006 Zürich, Switzerland
gervini@ifspm.unizh.ch

February 19, 2004

Abstract

This article introduces adaptive weighted maximum likelihood estimators for binary regression models. The asymptotic distribution under the model is established, and asymptotic confidence intervals are derived. Finite-sample properties are studied by simulation. For clean datasets, the proposed adaptive estimators are more efficient than the non-adaptive ones even for moderate sample sizes, and for outlier-contaminated datasets they show a comparable robustness. As for the asymptotic confidence intervals, the actual coverage levels under the model are very close to the nominal levels (even for moderate sample sizes), and they are reasonably stable under contamination.

Key Words: Efficient estimation; Logistic regression; Robust confidence intervals; Weighted maximum likelihood

1 Introduction

Binary regression models are very common in statistical applications. They are used in situations where a dichotomous response variable y_i and a vector of covariates x_i are observed for each individual. The probability of positive response is modeled as a function of a linear combination of the covariates: $P(y_i = 1|x_i) = \pi(x_i^\top \beta)$. When π is known (as we will assume in this paper) only the regression parameter β needs to be estimated, which is usually done by maximum likelihood.

The maximum likelihood estimator attains the minimum asymptotic variance under the model and then it is optimal, but it is very sensitive to atypical data. Observations with extreme covariates, in particular, have a large influence on the estimator, and if they are accompanied by misclassified responses, the resulting estimates can be seriously biased. Pregibon (1981, 1982) made the earliest systematic attempts to fix this problem; he proposed methods to unmask influential observations and robust estimators for the logistic model. Later robust proposals in this area include Stefanski, Carroll and Ruppert (1986), Copas (1988), Künsch, Stefanski and Carroll (1989), Morgenthaler (1992), Carroll and Pederson (1993), Bianco and Yohai (1996), Kordzakhia, Mishra and Reiersølmoen (2001), and Müller and Neykov (2003).

All these estimators differ greatly in terms of outlier resistance and efficiency under the model. We have studied asymptotic and finite-sample behavior of some of these estimators and found, for example, that suitably tuned Mallows-type estimators (Carroll and Pederson, 1993) are very robust to outlier contamination but inefficient under the model, while Schweppe-type estimators (Künsch et al., 1989) are very efficient under the model but show a poor outlier resistance (see Sections 5 and 6 below). Although a trade-off between robustness and efficiency is inevitable, in this article we propose estimators that can be as robust as Mallows-type estimators under contamination but are much more efficient under the model (in fact, 100% efficient in some situations). This is achieved by an adaptive weighting scheme, similar to that of Gervini (2002, 2003).

The new adaptive estimators are introduced in Section 3, after a brief revision of some existing estimators. Their asymptotic distribution is established in Section 4, and asymptotic confidence ellipsoids are derived. Finite-sample properties are studied in Section 6 by simulation.

2 Preliminaries: the binary regression model and existing estimators

In binary regression models we have a sample of response variables $y_i \in \{0,1\}$ and covariates $x_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, and it is assumed that the probability of positive response $\pi_i = P(y_i = 1|x_i)$ is linked with the covariates via the relationship $g(\pi_i) = x_i^\top \beta$. The most common link function g is the logit $g(\pi_i) = \log\{\pi_i/(1-\pi_i)\}$, and the corresponding model is known as logistic regression. Another common link function is $g = \Phi^{-1}$, the inverse of the normal distribution function, and the resulting model is called probit regression. In this paper we are not concerned with the choice of link function; we will focus on estimating the regression parameter β for a given g.

Let $\pi = g^{-1}$ be the inverse link function and $d(y_i, x_i, \beta) = -2y_i \log \pi(x_i^\top \beta) - 2(1-y_i)\log\{1-\pi(x_i^\top\beta)\}$ be the deviance. The maximum likelihood estimator of β is

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} d(y_i, x_i, \beta), \tag{1}$$

which satisfies the first-order condition

$$\sum_{i=1}^{n} \frac{\{y_i - \pi(x_i^\top \hat{\beta})\}\pi'(x_i^\top \hat{\beta})}{\pi(x_i^\top \hat{\beta})\{1-\pi(x_i^\top \hat{\beta})\}} x_i = 0. \tag{2}$$

For logistic and probit models the objective function in (1) is convex, so that if a finite minimizer exists, it is the unique solution of (2). In some situations (particularly for small samples) a finite minimizer may not exist. Albert and Anderson (1984) show that $\hat{\beta}$ exists if and only if the data overlap, in the sense that no hyperplane in the covariate space separates responses from non-responses —this condition can be easily verified with a linear program proposed by Santner and Duffy (1986).

Equations (1) and (2) suggest two ways of robustifying the maximum likelihood estimator. One way is to minimize the sum of tapered deviances

$\sum_{i=1}^{n} \rho(d(y_i, x_i, \beta))$ for an appropriate function ρ. The estimators of Pregibon (1982) and Bianco and Yohai (1996) are of this form (the latter with an additional bias-correction term). Bianco and Yohai's estimators are preferable to Pregibon's for theoretical reasons, but their computation is complicated, so they will not be considered in this article.

The second way to obtain robust estimators is to robustify the estimating equation (2). In this category we have two types of estimators, the so-called Schweppe's and Mallows' types. The former were introduced by Künsch et al. (1989): the regression estimator and a companion scatter estimator of the covariates ($\hat{B}$) are defined as solutions of the system

$$\begin{aligned} \sum_{i=1}^{n} \psi_b\big(r(y_i, x_i, \hat{\beta}, \hat{B})(x_i^\top \hat{B}^{-1} x_i)^{\frac{1}{2}}\big)\,(x_i^\top \hat{B}^{-1} x_i)^{-\frac{1}{2}}\, x_i &= 0 \qquad (3) \\ n^{-1} \sum_{i=1}^{n} v\big(x_i^\top \hat{\beta}, b/(x_i^\top \hat{B}^{-1} x_i)^{\frac{1}{2}}\big)\, x_i x_i^\top &= \hat{B}, \end{aligned}$$

where $r(y_i, x_i, \hat{\beta}, \hat{B}) = y_i - \pi(x_i^\top \hat{\beta}) - c\big(x^\top \hat{\beta}, b/(x_i^\top \hat{B}^{-1} x_i)^{\frac{1}{2}}\big)$, $c(t, b)$ is a bias-correcting function, and $v(t, b) = \psi_b^2\big(1 - \pi(t) - c(t, b)\big)\pi(t) + \psi_b^2\big(- \pi(t) - c(t, b)\big)\{1 - \pi(t)\}$. Usually ψ_b is taken as Huber's function $\psi_b(t) = (-b) \vee (t \wedge b)$, for which

$$c(t, b) = \begin{cases} b\pi(t)/\{1 - \pi(t)\} - \pi(t) & \text{if } t < 0,\ b < 1 - \pi(t) \\ 1 - \pi(t) - b\{1 - \pi(t)\}/\pi(t) & \text{if } t > 0,\ b < \pi(t) \\ 0 & \text{otherwise.} \end{cases}$$

Mallows-type estimators were also suggested by Künsch et al. (1989) but analyzed more deeply by Carroll and Pederson (1993). They are defined as solutions of

$$\sum_{i=1}^{n} w(x_i; \hat{\eta})\psi_b\big(y_i - \pi(x_i^\top \hat{\beta}) - c(x_i^\top \hat{\beta}, b)\big)x_i = 0, \qquad (4)$$

where $\hat{\eta}$ is a vector of nuisance parameters (typically, location and scatter estimators of the covariates). The most important difference between Mallows-type and Schweppe-type estimators is that covariates and residuals are downweighted independently of each other in (4), but not in (3).

The weights $w(x_i;\hat{\eta})$ in (4) do not necessarily depend on all p explanatory variables; most regression models have an intercept and some categorical covariates, which are not downweighted. If we write $x_i^\top = (u_i^\top, z_i^\top)$, where $u_i \in \mathbb{R}^{p-q}$ are the categorical covariates and $z_i \in \mathbb{R}^q$ are the continuous covariates, then the weights are typically of the form $w(x_i;\hat{\eta}) = \omega\big((z_i - \hat{\mu})^\top \hat{\Sigma}^{-1}(z_i - \hat{\mu})/t\big)$, with $\omega : \mathbb{R}_+ \to \mathbb{R}_+$ a non-increasing function, $\hat{\mu}$ and $\hat{\Sigma}$ robust estimators of location and scatter of the z_i's, and t a threshold value (usually $t = \chi^2_{q,1-\alpha}$ for some $\alpha \in (0,1)$). In this paper we will use the following weight functions:

- Hard-rejection: $\omega(u) = I\{u < 1\}$.
- Huber: $\omega(u) = \min\{1, 1/u\}$.
- Gaussian: $\omega(u) = \exp(0.5 - 0.5u)$.

The most important difference between these functions is that hard-rejection weights completely eliminate observations beyond the threshold, while the other two weights decrease to zero beyond the threshold but never vanish completely.

The independence between covariate weights and deviance weights in (4) is the reason why Mallows-type estimators are in general less efficient than Schweppe-type estimators: observations with extreme covariates are downweighted even if they are well-fitted. Obviously, the efficiency of Mallows-type estimators could be improved by taking smaller thresholding proportions, but then the estimator might be less robust. To attain high efficiency and high robustness at the same time, it is necessary to use adaptive thresholds, as introduced in the next section.

3 Adaptive estimators

Adaptive Mallows-type estimators can be constructed as follows. Let $\hat{\mu}^{(0)}$ and $\hat{\Sigma}^{(0)}$ be initial location and scatter estimators of the z_i's, and let F_n be the empirical distribution function of the squared Mahalanobis distances $m_i^2 = (z_i - \hat{\mu}^{(0)})^\top (\hat{\Sigma}^{(0)})^{-1}(z_i - \hat{\mu}^{(0)})$. Since F_n converges to $F_{\chi^2_q}$ (the χ^2_q distribution function) when the z_i's are normally distributed, but has heavier tails when there are outliers, the proportion of outliers in the covariates can be estimated by

$$\begin{aligned}\alpha_n &= \sup_{s \geq F_{\chi_q^2}^{-1}(1-\delta)} \{F_{\chi_q^2}(s) - F_n(s)\}_+ \\ &= \max_{i \geq i_0} \left\{ F_{\chi_q^2}(m_{(i)}^2) - \frac{i-1}{n} \right\}_+ ,\end{aligned}$$

where $\{\cdot\}_+$ denotes the positive part, δ is a constant that determines the length of the tails ($\delta = 0.25$ is a reasonable choice), and $i_0 = \min\{i : m_{(i)}^2 \geq F_{\chi_q^2}^{-1}(1-\delta)\}$. The adaptive threshold is then defined as

$$\begin{aligned}t_n &= F_n^{-1}(1-\alpha_n) \\ &= m_{(n-[n\alpha_n])}^2 .\end{aligned} \tag{5}$$

Mallows-type estimators with weights $w(x_i; \hat{\eta}) = \omega(m_i^2/t_n)$ would already be more efficient than those with fixed-threshold weights, but a further improvement is possible. Let $w_i = I\{m_i^2 < t_n\}$, and define reweighted estimators of location and scatter $\hat{\mu}^{(1)} = \sum_{i=1}^n w_i x_i / \sum_{i=1}^n w_i$ and $\hat{\Sigma}^{(1)} = \sum_{i=1}^n w_i (x_i - \hat{\mu})(x_i - \hat{\mu})^\top / \sum_{i=1}^n w_i$. What we propose is to use weights

$$w(x_i; \hat{\eta}) = \omega\big((z_i - \hat{\mu}^{(1)})^\top (\hat{\Sigma}^{(1)})^{-1} (z_i - \hat{\mu}^{(1)})/t_n\big) \tag{6}$$

in (4). The reweighted multivariate estimators $\hat{\mu}^{(1)}$ and $\hat{\Sigma}^{(1)}$ were proposed by Gervini (2003). These estimators have breakdown points not smaller than those of $\hat{\mu}^{(0)}$ and $\hat{\Sigma}^{(0)}$, and are asymptotically Normal if the z_i's follow an elliptical distribution. Their asymptotic variance depends on $t_0 = \operatorname{p\,lim} t_n$. Gervini (2003) shows that $t_0 = F^{-1}(1-\alpha_0)$, where $\alpha_0 = \sup_{s \geq F_{\chi_q^2}^{-1}(1-\delta)} \{F_{\chi_q^2}(s) - F(s)\}_+$ and F is the distribution of $(z-\mu_0)^\top \Sigma_0^{-1} (z-\mu_0)$. If the z_i's are normally distributed, then $t_0 = \infty$ and the reweighted estimators are asymptotically equivalent to the sample mean and the sample covariance matrix, and therefore fully efficient. This efficiency carries over to the adaptive Mallows' estimators, as shown in the next section.

Remark. For certain datasets a finite solution to (4) may not exist, even if the maximum likelihood estimator does exist. In general, it is difficult to find conditions for existence of robust estimators. However, for the particular choice $\psi_b(u) = u$ (for which $c(t, b) = 0$), $\hat{\beta}$ is just a weighted maximum likelihood estimator, and a necessary and sufficient condition for existence is that observations with $w(x_i; \hat{\eta}) > 0$ overlap (that is, there is no hyperplane separating responses from non-responses among those observations with non-null weight). This condition can be verified with a simple modification of the linear program of Santner and Duffy (1986), but as pointed out by Albert and Anderson (1984, p. 8), in practice this step is not necessary. The reason is that the weighted maximum likelihood estimator is the minimizer of $\sum_{i=1}^{n} w(x_i; \hat{\eta}) d(y_i, x_i, \beta)$, which is a strictly convex objective function. Therefore, if the minimization is carried out with an algorithm that reduces the objective function at each iteration (such as Newton–Raphson), two and only two outcomes are possible: *(i)* the algorithm converges to the unique minimizer $\hat{\beta}$, when a finite minimizer exists, or *(ii)* the algorithm diverges, when a finite minimizer does not exist. Non-existence of $\hat{\beta}$ is then automatically revealed by non-convergence of the algorithm.

4 Asymptotics for Mallows-type estimators

The estimating equation (4) can be written $\sum_{i=1}^{n} \Psi(x_i, y_i; \hat{\beta}; \hat{\eta}) = 0$, with $\Psi(x, y; \beta; \eta) = w(x; \eta)\psi_b(y - \pi(x^\top \beta) - c(x^\top \beta, b))x$. Under appropriate regularity conditions, the classical asymptotics of M-estimators hold (see Van der Vaart, 1998, ch. 5). Let β_0 be the model parameter and E_0 denote expectation under the model; define $M_0(\beta) = E_0\{\Psi(x, y; \beta; \eta_0)\}$, with $\eta_0 = \operatorname{p\,lim} \hat{\eta}$. If $C_0 = \mathsf{D} M_0(\beta)|_{\beta=\beta_0}$ (where D denotes the differential) and $A_0 = E_0\{\Psi(x, y; \beta_0; \eta_0)\Psi(x, y; \beta_0; \eta_0)^\top\}$, then

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{D} \mathcal{N}_p(0, C_0^{-1} A_0 C_0^{-1}).$$

This result is valid for non-adaptive and adaptive weights alike, as long as $\hat{\eta}$ converges to η_0 in probability. For the adaptive weights (6), $\hat{\eta} = (\hat{\mu}^{(1)}, \hat{\Sigma}^{(1)}, t_n)$ and $\eta_0 = (\mu_0, \Sigma_0, t_0)$.

For the weighted maximum likelihood estimator given by $\psi_b(u) = u$, A_0 and

C_0 have simple expressions:

$$
\begin{aligned}
C_0 &= -E\{w(x;\eta_0)\pi'(x^\top\beta_0)xx^\top\}, && (7)\\
A_0 &= E\{w^2(x;\eta_0)\pi(x^\top\beta_0)(1-\pi(x^\top\beta_0))xx^\top\}. && (8)
\end{aligned}
$$

For adaptive weights (6) we have $w(x;\eta_0) = \omega((z-\mu_0)^\top\Sigma_0^{-1}(z-\mu_0)/t_0)$. If $z \sim \mathcal{N}_p(\mu_0, \Sigma_0)$ (or in general, if F has lighter tails than $F_{\chi^2_q}$), then $t_0 = \infty$ and $w(x;\eta_0) \equiv \omega(0)$, which implies that $\hat\beta$ is equivalent to the maximum likelihood estimator and thus fully efficient. If F has heavier tails than $F_{\chi^2_q}$ (for example, if z is multivariate Cauchy or Student), then t_0 is finite and $\hat\beta$ is not fully efficient, although it is still asymptotically Normal (note that A_0 and C_0 are finite when t_0 is finite, even if the covariates do not have finite second moments).

Using the asymptotic normality of $\hat\beta$ it is possible to construct confidence ellipsoids for β_0. First, we estimate the matrices (7) and (8) with

$$
\begin{aligned}
\hat C &= -\frac{1}{n}\sum_{i=1}^n w(x_i;\hat\eta)\pi'(x_i^\top\hat\beta)x_ix_i^\top,\\
\hat A &= \frac{1}{n}\sum_{i=1}^n w(x_i;\hat\eta)\pi(x_i^\top\hat\beta)(1-\pi(x_i^\top\hat\beta))x_ix_i^\top.
\end{aligned}
$$

Then the estimated asymptotic variance of $\sqrt{n}(\hat\beta-\beta_0)$ is $\hat V = \hat C^{-1}\hat A\hat C^{-1}$. The asymptotic confidence ellipsoid of level $1-\alpha$ for β_0 is given by $\mathcal{E}(\beta_0) = \{\beta \in \mathbb{R}^p : n(\hat\beta-\beta)^\top\hat V^{-1}(\hat\beta-\beta) \leq \chi^2_{p,1-\alpha}\}$. This can be generalized to linear transformations of β_0. If T is a $r \times p$ matrix of full rank r and $\theta_0 = T\beta_0$, then $\hat\theta = T\hat\beta$ and $\mathcal{E}(\theta_0) = \{\theta \in \mathbb{R}^r : n(\hat\theta-\theta)^\top(T\hat VT^\top)^{-1}(\hat\theta-\theta) \leq \chi^2_{r,1-\alpha}\}$. This generalization includes, for example, confidence intervals for the individual coefficients β_{0j}, $j = 1,\ldots,p$, and for contrasts between them.

5 Comparative asymptotic behavior of non-adaptive estimators

Using the asymptotic results of the previous section, we made some comparisons among different weighted maximum likelihood estimators. We used hard-

Table 1: Asymptotic relative efficiencies of weighted maximum likelihood estimators of β_{02} for clean logistic models (α: threshold proportion; HR: hard rejection weights; HU: Huber's weights; GA: Gaussian weights).

	$\beta_0 = (0, 1.79)$			$\beta_0 = (-1, 1.18)$		
α	HR	HU	GA	HR	HU	GA
0.50	0.21	0.72	0.63	0.12	0.63	0.50
0.25	0.58	0.92	0.89	0.41	0.86	0.82
0.10	0.86	0.99	0.96	0.72	0.96	0.94
0.05	0.94	1.00	0.98	0.85	0.99	0.96

rejection, Huber's and Gaussian weights, with threshold values $t = \chi^2_{q,1-\alpha}$. For non-contaminated models we compared the asymptotic efficiencies of the estimators, and for contaminated models we compared their asymptotic biases. Only fixed-threshold estimators were considered for this section; adaptive estimators will be studied later by simulation.

5.1 Asymptotic efficiency

Our target model was logistic regression with intercept and a $\mathcal{N}(0, 1)$ covariate. Since there is no model equivariance as in linear regression, the behavior of the estimators depends strongly on the regression parameter, so β_0 must be chosen with care. We took $\beta_0 = (0, 1.79)$, for which $P_0(y = 1) = 0.5$ and the distribution of $\pi(x^\top \beta_0)$ is symmetric, and $\beta_0 = (-1, 1.18)$, for which $P_0(y = 1) \approx 0.3$ and the distribution of $\pi(x^\top \beta_0)$ is skewed. These parameters provide well-conditioned models, in the sense that $P_0(\pi(x^\top \beta_0) \leq 0.05) = 0.05$.

Asymptotic relative efficiencies are given in Table 1. To save space, we show results for $\hat{\beta}_{02}$ only; intercept estimates exhibit a similar pattern. Table 1 shows that hard-rejection weights are in general very inefficient, while Gaussian and Huber's weights have a comparable performance for $\alpha \leq 0.25$.

5.2 Asymptotic bias under contaminations

To study the (asymptotic) effect of outliers on the estimators, we considered point-mass contamination models: given $\varepsilon \in [0, 0.5)$ and $\tilde{x} \in \mathbb{R}^p$,

$$P_*(y = 1|x) \;\; = \;\; (1 - \varepsilon)\pi(x^\top \beta_0) + \varepsilon I\{\pi(\tilde{x}^\top \beta_0) \leq 0.5\}, \qquad (9)$$

where β_0 is the target parameter. This model generates a proportion ε of misclassified observations with possibly outlying covariates $\tilde{x} = (u, \tilde{z})$ (note that the categorical variables remain unchanged). The asymptotic bias of weighted maximum likelihood estimators under P_* can be numerically computed, assuming that η_0 is known. Note that under P_* the estimator $\hat{\beta}$ defined by $\sum_{i=1}^n \Psi(x_i, y_i; \hat{\beta}; \eta_0) = 0$ converges to the solution β_* of the equation $E_*\{\Psi(x, y; \beta; \eta_0)\} = 0$. For weighted maximum likelihood estimators, β_* solves

$$(1 - \varepsilon)M_0(\beta_*) + \varepsilon w(\tilde{x}; \eta_0)\tilde{x}[I\{\pi(\tilde{x}^\top \beta_0) \leq 0.5\} - \pi(\tilde{x}^\top \beta_*)] = 0$$

and can be easily computed. The asymptotic squared bias of $\hat{\beta}$ under (9) is $B(\varepsilon, \tilde{x}) = \|\beta_* - \beta_0\|^2$. An overall measure of robustness for a given proportion of contamination is the maximum bias $\mathcal{B}(\varepsilon) = \sup_{\tilde{x} \in \mathbb{R}^p} B(\varepsilon, \tilde{x})$.

We have computed $\mathcal{B}(0.10)$ and $\mathcal{B}(0.20)$ for the same estimators and the same target models as in Section 5.1. The contaminated covariates $\tilde{x}$ were of the form $\tilde{x}(k) = (1, k)$ with $k \in [-4, 4]$ (all robust estimators under consideration realized their maximum bias within this range). The results are given in Table 2. The first conclusion we draw from the table is that smaller thresholds do not necessarily lead to smaller biases. For hard-rejection weights, $\alpha = 0.10$ or perhaps $\alpha = 0.05$ provide the best thresholds, while $\alpha = 0.25$ is the best choice for Gaussian weights. The latter is the most robust estimator among the ones considered in this section.

The maximum bias, however, tells only part of the story. For better insight, we have plotted $B(\varepsilon, \tilde{x}(k))$ as a function of k for $\varepsilon = 0.10$ and for two different α's (Figure 1). Hard-rejection-weight estimators have zero bias after a certain k, while the biases of the other estimators decrease to zero more slowly. The biases of Gaussian-weight estimators, though, are practically zero after a certain point.

Table 2: Maximum asymptotic biases of weighted maximum likelihood estimators under point-mass contaminations (ε: contamination proportion; α: threshold proportion; HR: hard rejection weights; HU: Huber's weights; GA: Gaussian weights).

	$\beta_0 = (0, 1.79)$			$\beta_0 = (-1, 1.18)$		
α	HR	HU	GA	HR	HU	GA
			$\varepsilon = 0.10$			
0.50	2.92	0.86	0.82	4.66	1.16	1.14
0.25	1.66	0.98	0.83	1.99	1.04	0.85
0.10	1.57	1.30	1.05	1.51	1.17	0.90
0.05	1.67	1.55	1.22	1.46	1.28	0.98
			$\varepsilon = 0.20$			
0.50	6.30	2.05	2.05	9.43	2.67	2.78
0.25	3.40	2.12	1.89	3.95	2.27	1.99
0.10	2.96	2.50	2.17	2.87	2.32	1.96
0.05	2.98	2.76	2.37	2.66	2.39	2.01

Again, Gaussian weights appear as the most recommendable.

6 Comparative finite-sample behavior of adaptive and non-adaptive estimators

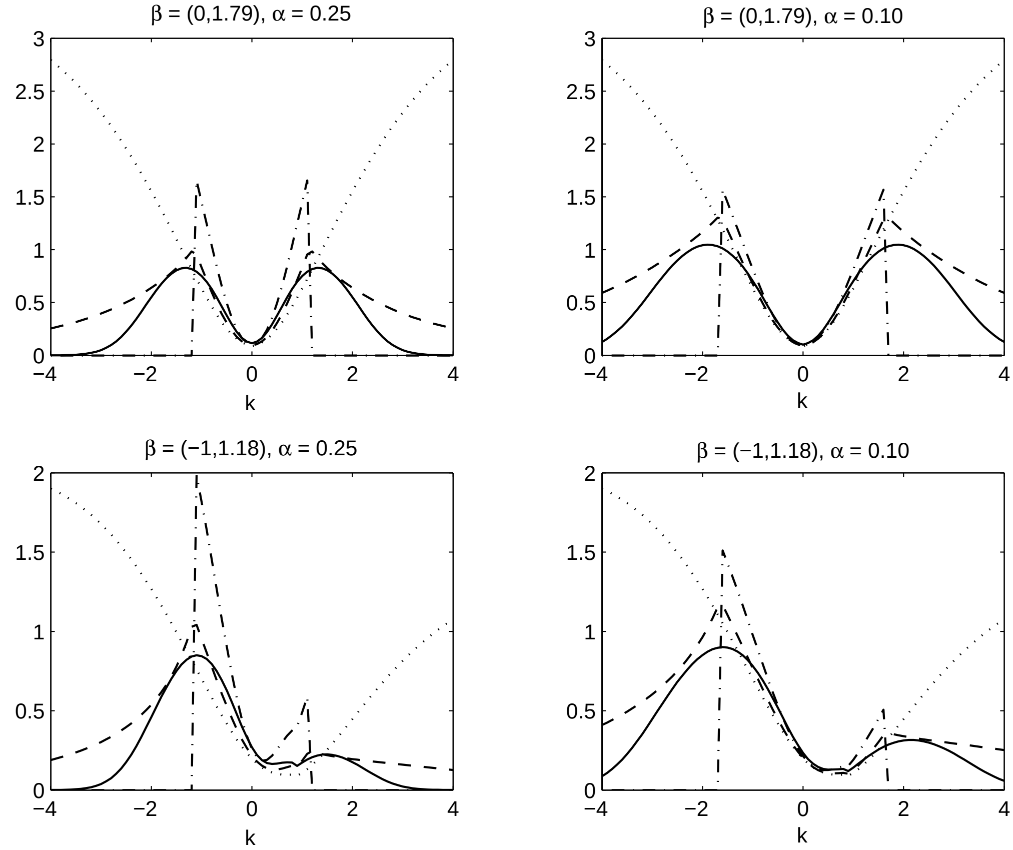
This section has two aims. First, to evaluate the finite-sample performance of point estimators of the regression parameter under different models, with and without outliers; this is done in Section 6.1. Second, to determine if the asymptotic confidence intervals proposed in Section 4 are approximately valid for small or moderate sample sizes, and to what extent coverage level and interval length are affected by outliers; this is investigated in Section 6.2.

6.1 Point estimators

Eight estimators were included in the Monte Carlo study:

- Maximum likelihood estimator.

Figure 1: Bias curves of weighted maximum likelihood estimators under 10%-point-mass contaminations at $\tilde{x} = (1, k)$ (solid line, Gaussian weights; dashed line, Huber weights; dash-dot line, hard-rejection weights; dotted line, maximum likelihood estimator).



- Schweppe-type estimator with threshold $b = 1.5\sqrt{p}$ (we also tried larger multiples of $\sqrt{p}$, such as those used by Künsch et al. (1989), but they did not provide reasonable outlier resistance).

- Fixed-threshold weighted maximum likelihood estimator, with hard-rejection weights ($\alpha = 0.25$ and $\alpha = 0.10$) and Gaussian weights ($\alpha = 0.25$).

- Adaptive weighted maximum likelihood estimator, with hard-rejection weights ($\delta = 0.25$ and $\delta = 0.10$) and Gaussian weights ($\delta = 0.25$).

The last two classes of estimators require location and scatter estimates of the covariates. We used the minimum covariance determinant estimator, computed with the fast algorithm of Rousseeuw and van Driessen (1999). The regression estimators themselves were computed by Newton-Raphson, using zero as starting point —there was no need to try with different starting points because the estimating equations have at most one finite solution.

The following 24 models were simulated (with 5000 replications for each):

- Clean logistic models with intercept and $\mathcal{N}_{p-1}(0, I)$ covariates, sample sizes $n = 100$ and $n = 500$, and parameters $\beta_0 = (0, 1.79)$, $\beta_0 = (-1, 1.18)$, $\beta_0 = (0, 1.79, 0, 0, 0)$ and $\beta_0 = (-1, 1.18, 0, 0, 0)$.

- Point-mass contaminated models with the same β_0's as above, $\tilde{x}(k) = (1, k, 0_{p-2})$ with $k = 2$ and $k = 5$, $\varepsilon = 0.10$ and $\varepsilon = 0.20$, and sample size $n = 100$.

None of the replications resulted in samples for which the estimates did not exist. This is not surprising because the simulated models are well-balanced and the sample sizes are large, and the probability of obtaining non-overlapping datasets goes to zero as n goes to infinity. However, for any fixed n the probability of obtaining a sample for which some of the estimators do not exist *might* be non-zero, and then the bias of these estimators, properly speaking, will be infinite. Therefore, although we did not find problems of non-existence in these simulations, the reader should keep in mind that the biases, variances and mean squared errors reported below must be interpreted as conditional quantities given the existence of the estimators.

Table 3: Bias and variance of estimators of β_{02} for clean logistic models.

	$\beta_0 = (0, 1.79)$		$\beta_0 = (-1, 1.18)$	
Estimators	$\sqrt{n} \times \text{bias}$	$n \times \text{var}$	$\sqrt{n} \times \text{bias}$	$n \times \text{var}$
		$n = 100$		
Max. Lik.	0.86 (0.06)	17.02 (0.44)	0.49 (0.05)	10.81 (0.28)
Schweppe	1.14 (0.07)	23.69 (0.65)	0.67 (0.05)	14.21 (0.45)
F, h-r, 0.25†	0.70 (0.08)	29.45 (0.74)	0.45 (0.08)	29.27 (0.72)
F, h-r, 0.10	0.81 (0.06)	20.12 (0.50)	0.49 (0.06)	16.25 (0.40)
F, Ga, 0.25	0.74 (0.06)	18.87 (0.48)	0.43 (0.05)	13.54 (0.33)
A, h-r, 0.25‡	0.83 (0.06)	18.58 (0.47)	0.49 (0.05)	14.04 (0.36)
A, h-r, 0.10	0.83 (0.06)	17.72 (0.45)	0.48 (0.05)	12.73 (0.32)
A, Ga, 0.25	0.80 (0.06)	17.28 (0.44)	0.47 (0.05)	11.42 (0.29)
		$n = 500$		
Max. Lik.	0.43 (0.05)	14.53 (0.31)	0.23 (0.04)	9.14 (0.19)
Schweppe	0.52 (0.06)	19.02 (0.42)	0.30 (0.05)	11.62 (0.24)
F, h-r, 0.25	0.38 (0.07)	25.24 (0.55)	0.17 (0.07)	23.69 (0.48)
F, h-r, 0.10	0.40 (0.06)	17.28 (0.38)	0.21 (0.05)	13.16 (0.27)
F, Ga, 0.25	0.38 (0.06)	16.49 (0.36)	0.20 (0.05)	11.43 (0.23)
A, h-r, 0.25	0.42 (0.05)	14.98 (0.32)	0.22 (0.04)	10.08 (0.21)
A, h-r, 0.10	0.42 (0.05)	14.82 (0.32)	0.22 (0.04)	9.81 (0.20)
A, Ga, 0.25	0.41 (0.05)	14.75 (0.32)	0.21 (0.04)	9.40 (0.19)

† "F, h-r, 0.25" stands for fixed-threshold estimator, hard-rejection weight, $\alpha = 0.25$; "Ga" stands for Gaussian weights.

‡ "A, h-r, 0.25" stands for adaptive-threshold estimator, hard-rejection weight, $\delta = 0.25$.

Table 4: Bias and variance of estimators of β_{02} for clean logistic models.

	$\beta_0 = (0, 1.79, 0, 0, 0)$		$\beta_0 = (-1, 1.18, 0, 0, 0)$	
Estimators	$\sqrt{n} \times$ bias	$n \times$ var	$\sqrt{n} \times$ bias	$n \times$ var
		$n = 100$		
Max. Lik.	1.68 (0.06)	20.70 (0.57)	1.03 (0.05)	12.25 (0.31)
Schweppe	2.10 (0.07)	26.10 (0.69)	1.27 (0.05)	14.68 (0.36)
F, h-r, 0.25†	2.70 (0.11)	55.42 (1.79)	1.76 (0.09)	40.69 (1.23)
F, h-r, 0.10	2.22 (0.08)	35.27 (1.18)	1.42 (0.07)	24.35 (0.66)
F, Ga, 0.25	1.70 (0.07)	23.55 (0.64)	1.04 (0.05)	14.03 (0.34)
A, h-r, 0.25‡	2.01 (0.08)	28.90 (0.89)	1.26 (0.06)	18.54 (0.47)
A, h-r, 0.10	1.92 (0.07)	26.73 (0.86)	1.20 (0.06)	16.89 (0.41)
A, Ga, 0.25	1.63 (0.07)	21.46 (0.58)	1.00 (0.05)	12.63 (0.31)
		$n = 500$		
Max. Lik.	0.80 (0.05)	14.70 (0.31)	0.46 (0.04)	9.56 (0.19)
Schweppe	0.93 (0.06)	17.12 (0.37)	0.55 (0.05)	11.01 (0.23)
F, h-r, 0.25	1.01 (0.07)	22.68 (0.49)	0.68 (0.06)	17.62 (0.37)
F, h-r, 0.10	0.86 (0.06)	17.48 (0.36)	0.52 (0.05)	12.33 (0.25)
F, Ga, 0.25	0.78 (0.06)	15.47 (0.32)	0.47 (0.05)	10.28 (0.21)
A, h-r, 0.25	0.81 (0.06)	15.45 (0.32)	0.49 (0.05)	10.48 (0.21)
A, h-r, 0.10	0.81 (0.06)	15.23 (0.31)	0.48 (0.05)	10.15 (0.21)
A, Ga, 0.25	0.78 (0.05)	14.85 (0.31)	0.46 (0.04)	9.75 (0.20)

† "F, h-r, 0.25" stands for fixed-threshold estimator, hard-rejection weight, $\alpha = 0.25$; "Ga" stands for Gaussian weights.

‡ "A, h-r, 0.25" stands for adaptive-threshold estimator, hard-rejection weight, $\delta = 0.25$.

Table 5: Simulated mean squared errors ×10 of estimators of β_{02} under point-mass contaminations at $\tilde{x} = (1, k)$. Sample size $n = 100$.

	$\beta_0 = (0, 1.79)$		$\beta_0 = (-1, 1.18)$	
Estimators	$k = 2$	$k = 5$	$k = 2$	$k = 5$
	$\varepsilon = 0.10$			
Max. Lik.	14.50 (0.04)	30.64 (0.03)	4.58 (0.03)	12.61 (0.02)
Schweppe	8.16 (0.06)	8.49 (0.06)	5.17 (0.03)	6.54 (0.05)
F, h-r, 0.25†	3.03 (0.07)	2.94 (0.07)	2.50 (0.06)	2.48 (0.06)
F, h-r, 0.10	9.21 (0.11)	2.20 (0.06)	3.58 (0.05)	1.57 (0.04)
F, Ga, 0.25	8.57 (0.05)	2.04 (0.05)	3.00 (0.03)	1.40 (0.04)
A, h-r, 0.25‡	5.88 (0.10)	2.09 (0.06)	2.83 (0.05)	1.36 (0.04)
A, h-r, 0.10	8.29 (0.10)	2.08 (0.06)	3.21 (0.05)	1.34 (0.04)
A, Ga, 0.25	10.54 (0.05)	14.13 (0.13)	3.51 (0.03)	5.95 (0.05)
	$\varepsilon = 0.20$			
Max. Lik.	25.30 (0.05)	38.06 (0.03)	9.22 (0.04)	16.67 (0.02)
Schweppe	26.86 (0.05)	35.09 (0.03)	10.09 (0.03)	14.23 (0.01)
F, h-r, 0.25	12.97 (0.21)	3.00 (0.08)	5.72 (0.09)	2.15 (0.06)
F, h-r, 0.10	26.99 (0.09)	2.48 (0.07)	10.17 (0.05)	1.58 (0.04)
F, Ga, 0.25	23.53 (0.06)	2.68 (0.05)	8.51 (0.04)	1.65 (0.03)
A, h-r, 0.25	12.87 (0.18)	2.42 (0.07)	5.82 (0.09)	1.46 (0.04)
A, h-r, 0.10	24.73 (0.08)	2.42 (0.07)	9.06 (0.04)	1.46 (0.04)
A, Ga, 0.25	20.40 (0.10)	26.71 (0.07)	7.36 (0.05)	10.93 (0.03)

† "F, h-r, 0.25" stands for fixed-threshold estimator, hard-rejection weight, $\alpha = 0.25$; "Ga" stands for Gaussian weights.

‡ "A, h-r, 0.25" stands for adaptive-threshold estimator, hard-rejection weight, $\delta = 0.25$.

Table 6: Simulated mean squared errors ×10 of estimators of β_{02} under point-mass contaminations at $\tilde{x} = (1, k, 0, 0, 0)$. Sample size $n = 100$.

	$\beta_0 = (0, 1.79, 0, 0, 0)$		$\beta_0 = (-1, 1.18, 0, 0, 0)$	
Estimators	$k = 2$	$k = 5$	$k = 2$	$k = 5$
		$\varepsilon = 0.10$		
Max. Lik.	14.16 (0.05)	30.67 (0.03)	4.37 (0.03)	12.58 (0.02)
Schweppe	12.20 (0.07)	21.88 (0.14)	4.89 (0.03)	11.55 (0.02)
F, h-r, 0.25†	10.54 (0.14)	6.36 (0.28)	5.17 (0.74)	4.70 (0.75)
F, h-r, 0.10	11.41 (0.09)	4.17 (0.14)	4.01 (0.05)	2.59 (0.08)
F, Ga, 0.25	12.37 (0.06)	5.16 (0.07)	3.94 (0.03)	2.59 (0.04)
A, h-r, 0.25‡	12.77 (0.07)	3.30 (0.10)	4.18 (0.04)	2.00 (0.06)
A, h-r, 0.10	13.10 (0.07)	3.22 (0.10)	4.22 (0.04)	1.92 (0.06)
A, Ga, 0.25	13.29 (0.05)	11.25 (0.07)	4.16 (0.03)	4.92 (0.03)
		$\varepsilon = 0.20$		
Max. Lik.	25.19 (0.05)	38.39 (0.03)	9.09 (0.04)	16.83 (0.02)
Schweppe	25.83 (0.05)	36.83 (0.03)	9.77 (0.04)	15.25 (0.02)
F, h-r, 0.25	25.72 (0.16)	6.36 (0.27)	9.91 (0.11)	4.55 (0.64)
F, h-r, 0.10	25.32 (0.10)	4.49 (0.15)	9.35 (0.07)	2.66 (0.09)
F, Ga, 0.25	25.33 (0.07)	14.51 (0.12)	9.17 (0.05)	6.20 (0.05)
A, h-r, 0.25	25.75 (0.08)	3.94 (0.13)	9.38 (0.05)	2.24 (0.08)
A, h-r, 0.10	25.66 (0.07)	3.94 (0.13)	9.33 (0.05)	2.24 (0.08)
A, Ga, 0.25	25.32 (0.06)	27.45 (0.06)	9.16 (0.04)	11.17 (0.03)

† "F, h-r, 0.25" stands for fixed-threshold estimator, hard-rejection weight, $\alpha = 0.25$; "Ga" stands for Gaussian weights.

‡ "A, h-r, 0.25" stands for adaptive-threshold estimator, hard-rejection weight, $\delta = 0.25$.

Tables 3 and 4 give $\sqrt{n}\times$ bias and $n\times$ variance of the estimators of β_{02} for non-contaminated models; Tables 5 and 6 give $10\times$ mean squared error of the same estimators under contaminated models. Monte Carlo standard errors are given in parenthesis. We see that for each type of weight function, adaptive estimators are always more efficient than their non-adaptive counterparts; the improvements are most remarkable for hard-rejections weights. The robustness is not affected when hard-rejection weights are used, but adaptive Gaussian weights do not perform well. Comparisons across different types of weights are less clear-cut. For clean models, the most efficient non-adaptive robust estimator is the weighted maximum likelihood estimator with Gaussian weights (in agreement with Table 1). This estimator also ranks well for contaminated models if $p=2$, but for $p=5$ the hard-rejection-weight estimator with $\alpha=0.10$ is more robust. The latter, however, is outperformed by the adaptive estimator with hard-rejection weights and $\delta=0.25$, which is much more efficient and of comparable robustness. All things considered, the adaptive estimator with hard-rejection weights and $\delta=0.25$ seems to be the most recommendable, followed by the non-adaptive estimator with Gaussian weights and $\alpha=0.25$.

6.2 Confidence intervals

The analysis of coverage level and length of confidence intervals derived from robust estimators was restricted to the best estimators of each type, as found in Section 6.1: the fixed-threshold estimator with Gaussian weights and $\alpha=0.25$, and the adaptive-threshold estimator with hard-rejection weights and $\delta=0.25$. These confidence intervals were compared with those derived from the maximum likelihood estimator. The sampling models were similar to those in Section 6.1 and the number of replications was again 5000 for each model. We also ran some simulations with heavier-tailed covariates (Student's *t* with 2 degrees of freedom) but the results were similar, so they are not reported.

Tables 7 and 8 give coverage percentage and median length of the confidence intervals for the regression coefficient β_{02}. Monte Carlo standard errors are not given to save space, but they are small enough to make all differences in median length significant. We see that for clean models the actual coverage is very close

Table 7: Coverage and median length of nominal 95% confidence intervals for β_{02}. Clean logistic model with parameters $\beta_0^{\mathrm{A}} = (0, 1.79)$, $\beta_0^{\mathrm{B}} = (-1, 1.18)$, $\beta_0^{\mathrm{C}} = (0, 1.79, 0, 0, 0)$ and $\beta_0^{\mathrm{D}} = (-1, 1.18, 0, 0, 0)$.

	β_0^{A}		β_0^{B}		β_0^{C}		β_0^{D}	
Estimators	cov.	len.	cov.	len.	cov.	len.	cov.	len.
				$n = 50$				
Max. Lik.	96.9	2.17	96.0	1.74	96.8	2.42	96.2	1.93
Fixed†	96.7	2.32	96.0	1.94	97.3	2.75	96.9	2.24
Adaptive‡	97.0	2.40	96.1	2.11	97.1	3.37	97.1	2.92
				$n = 100$				
Max. Lik.	95.7	1.49	95.3	1.20	95.4	1.57	95.4	1.26
Fixed	95.8	1.58	94.9	1.33	95.6	1.67	95.6	1.35
Adaptive	96.1	1.57	95.4	1.35	95.3	1.80	95.3	1.51
				$n = 500$				
Max. Lik.	95.2	0.66	95.5	0.53	95.0	0.66	95.3	0.53
Fixed	95.1	0.70	95.0	0.58	95.1	0.68	95.3	0.55
Adaptive	95.3	0.66	95.2	0.55	95.1	0.68	95.1	0.56

† Fixed-threshold weighted maximum likelihood estimator with Gaussian weights and $\alpha = 0.25$.

‡ Adaptive weighted maximum likelihood estimator with hard-rejection weights and $\delta = 0.25$.

Table 8: Coverage and median length of nominal 95% confidence intervals for β_{02} under point-mass contaminations at $\tilde{x} = (1, k, 0_{p-2})$, with $k = 2$. Sample size $n = 100$, parameters $\beta_0^{\mathrm{A}} = (0, 1.79)$, $\beta_0^{\mathrm{B}} = (-1, 1.18)$, $\beta_0^{\mathrm{C}} = (0, 1.79, 0, 0, 0)$ and $\beta_0^{\mathrm{D}} = (-1, 1.18, 0, 0, 0)$.

	β_0^{A}		β_0^{B}		β_0^{C}		β_0^{D}	
Estimators	cov.	len.	cov.	len.	cov.	len.	cov.	len.
				$\varepsilon = 0.01$				
Max. Lik.	90.3	1.33	93.8	1.13	92.1	1.39	95.2	1.18
Fixed†	94.5	1.52	94.4	1.29	93.8	1.53	95.1	1.28
Adaptive‡	93.5	1.50	94.8	1.31	93.1	1.66	95.1	1.44
				$\varepsilon = 0.05$				
Max. Lik.	6.5	0.98	60.4	0.95	11.1	1.01	68.0	1.00
Fixed	67.3	1.27	83.7	1.12	38.7	1.14	77.0	1.09
Adaptive	66.9	1.52	83.6	1.35	44.1	1.25	80.3	1.22
				$\varepsilon = 0.10$				
Max. Lik.	0	0.80	10.8	0.85	0	0.83	17.0	0.89
Fixed	7.9	0.98	43.6	0.95	1.0	0.91	30.5	0.96
Adaptive	71.4	1.67	73.9	1.52	5.0	0.99	39.2	1.06

† Fixed-threshold weighted maximum likelihood estimator with Gaussian weights and $\alpha = 0.25$.

‡ Adaptive weighted maximum likelihood estimator with hard-rejection weights and $\delta = 0.25$.

to the nominal one in all cases; for small samples the intervals tend to err on the conservative side, which is good. Non-adaptive confidence intervals are shorter than adaptive ones for $n = 50$, but for $n = 500$ the situation reverses, as expected. For point-mass contaminations, the coverage level deteriorates as the contamination proportion increases, due to the bias of the estimators (for $n = 100$ the variance is relatively small). Both robust estimators are comparable for small ε, but for $\varepsilon = 0.10$ the adaptive one is more stable in terms of coverage level.

7 Conclusions

This paper has proposed robust estimators for binary regression models that combine high outlier resistance and high efficiency under the model. Among existing estimators, we found that the Mallows-type weighted maximum likelihood with Gaussian weights is reasonably robust and efficient. However, for large sample sizes this estimator is outperformed in efficiency by some of the adaptive estimators introduced in this article, which at the same time maintain high outlier resistance. The computational complexity of the new estimators is not greater than that of fixed-threshold Mallows' estimators. Therefore, given the gains in efficiency and the comparable robustness provided by adaptive-threshold estimators, we think that our proposal is a practical improvement over existing methods.

References

[1] Albert, A. and Anderson, J.A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71**, 1-10.

[2] Bianco, A.M. and Yohai, V.J. (1996). Robust estimation in the logistic regression model. In: *Robust Statistics, Data Analysis and Computer Intensive Methods*, Ed. H. Rieder, pp. 17-34. New York: Springer-Verlag.

[3] Carroll, R.J. and Pederson, S. (1993). On robustness in the logistic regression model. *J. R. Statist. Soc.* B **55**, 693-706.

[4] Copas, J.B. (1988). Binary regression models for contaminated data (with discussion). *J. R. Statist. Soc.* B **50**, 225-265.

[5] Gervini, D. (2002). A class of robust and fully efficient regression estimators. *Ann. Statist.* **30**, 583-616.

[6] Gervini, D. (2003). A robust and efficient adaptive reweighted estimator of multivariate location and scatter. *J. Multivar. Annal.* **84**, 116-144.

[7] Kordzakhia, N., Mishra, G. D., and Reiersølmoen, L. (2001). Robust estimation in the logistic regression model. *Journal of Statistical Planning and Inference* **98**, 211-223.

[8] Künsch, H.R., Stefanski, L.A. and 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.

[9] Morgenthaler, S. (1992). Least-absolute-deviations fits for generalized linear models. *Biometrika* **79**, 747-754.

[10] Müller, C. H. and Neykov, N. (2003). Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. *Journal of Statistical Planning and Inference* **116**, 503-519.

[11] Pregibon, D. (1981). Logistic regression diagnostics. *Ann. Statist.* **9**, 705-724.

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

[13] Rousseeuw, P.J. and Van Driessen, K. (1999). A fast algorithm for the Minimum Covariance Determinant estimator. *Technometrics* **41**, 212-223.

[14] Santner, T.J. and 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.

[15] Stefanski, L.A., Carroll, R.J. and Ruppert, D. (1986). Optimally bounded score functions for generalized linear models with applications to logistic regression. *Biometrika* **73**, 413-424.

[16] Van der Vaart, A.W. (1998). *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.