Implementing the Bianco and Yohai estimator for logistic regression

Christophe Croux * Gentiane Haesbroeck†

Abstract: In 1996, Bianco and Yohai proposed a highly robust estimation procedure in the logistic regression model. The theoretical results they obtain were very promising. In this paper we complement their study by providing a fast and stable algorithm to compute this estimator. Moreover, we derive criteria for the existence of the estimator at finite samples and discuss the problem of the selection of the loss-function. Many other robust estimators have been introduced in logistic regression, some of these being standard available in some statistical softwares. We compare the performance of these estimators with respect to the estimator of Bianco and Yohai. We also look into the advantages of an extra weighting step.

Keywords: Robust Estimation, Influence Function, Logistic Regression, Maximum Likelihood.


*Dept. of Applied Economics, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium, Email: Christophe.Croux@econ.kuleuven.ac.be.
†Dept. of Mathematics, University of Liège (B37), Grande Traverse 12, B-4000 Liège, Belgium, Email:G.Haesbroeck@ulg.ac.be.
1 Introduction

Let $Y_i, 1 \leq i \leq n,$ be independent Bernoulli variables whose success probabilities depend on the values of $p$-dimensional explanatory variables $X_1, \ldots, X_n$ through the relation

$$P(Y_i = 1|X_i = x_i) = F(\alpha + \beta^t x_i)$$

where $F$ is a strictly increasing cumulative distribution function. Taking $F(u) = 1/(1 + \exp(-u))$ results in the logit model, which is the model we will consider in this paper.

To simplify notation, we will use $\gamma = (\alpha, \beta^t)$ and $z_i = (1, x_i^t)$ for all $1 \leq i \leq n$. An estimator for $\gamma$ computed from the sample $X_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is denoted $\hat{\gamma}_n$. The Maximum Likelihood (ML) estimator $\hat{\gamma}_n^{ML}$ is defined as

$$\hat{\gamma}_n^{ML} = \arg \max_\gamma \log L(\gamma; X_n) = \arg \min_\gamma \sum_{i=1}^n d(z_i^t \gamma; y_i) \quad (1.1)$$

where $\log L(\gamma; Z_n)$ is the log-likelihood function calculated in $\gamma$ and $d(z_i^t \gamma; y_i)$ is the deviance function given by

$$d(z_i^t \gamma; y_i) = -y_i \log F(z_i^t \gamma) - (1 - y_i) \log \{1 - F(z_i^t \gamma)\}.$$

The ML estimator will be the most efficient estimator, but it may behave very poorly in presence of outliers. Therefore robust alternatives need to be constructed.

In this paper, focus is on a generalization of equation (1.1) which consists of replacing the deviance function by another one. The estimator $\hat{\gamma}_n$ of interest is therefore defined by

$$\hat{\gamma}_n = \arg \min_\gamma \sum_{i=1}^n \varphi(z_i^t \gamma; y_i), \quad (1.2)$$

where $\varphi$ is a positive, continuous and almost everywhere differentiable function. It needs to satisfy $\varphi(s; 0) = \varphi(-s; 1)$ for any score $s$, where a score value $s_i = z_i^t \gamma$ is obtained as a linear combination of a given parameter vector $\gamma$. Instead of working with $\varphi$, we will most of the time use the univariate function $\phi(s) = \varphi(s; 0)$. The value $\phi(s)$ gives for an observation corresponding to a null $y$, the impact of a particular score $s$ on the value of the objective function in (1.2). This function is assumed to be non decreasing since a large positive score $s$ should not be attributed to observations having a null $y$ value and should therefore receive a larger weight in the function to minimize. We further require that $\lim_{s \to -\infty} \phi(s) = 0$, implying that a large negative score, what we expect for a value $y = 0$, is not contributing to the objective function in (1.2).
As equation (1.2) shows, the estimator \( \hat{\gamma}_n \) belongs to the class of M-type estimators and the associated first order condition is given by

\[
\frac{1}{n} \sum_{i=1}^{n} \Psi(z_i^t \gamma; y_i) z_i = 0
\]

(1.3)

where \( \Psi(s; 0) = \frac{\partial \phi(s;0)}{\partial s} \) and \( \Psi(s; 1) = -\Psi(-s; 0) \). Notice that these first-order conditions may have multiple solutions and we need to use the value of the objective function in (1.2) to select the final estimator. Throughout the paper we will use the notation \( \psi(s) = \Psi(s; 0) = \phi'(s) \).

Particular cases of equation (1.2) have already been considered in the literature. First, the ML estimator obviously belongs to this class of M-estimators with \( \phi_{ML}(s) = -\ln(1 - F(s)) \).

A more robust proposal is due to Pregibon (1982) who suggested an estimator defined by

\[
\hat{\gamma}_n = \arg \min_{\gamma} \sum_{i=1}^{n} \lambda(d(z_i^t \gamma; y_i))
\]

where \( \lambda \) is a strictly increasing Huber’s type function. This estimator was designed to give less weight to observations poorly accounted for by the model but did not downweight influential observations in the design space. Moreover, it was not consistent. Then, Bianco and Yohai (1996) constructed a consistent and more robust version of Pregibon’s estimator by working with a bounded function \( \rho \), and defining

\[
\hat{\gamma}_n = \arg \min_{\gamma} \sum_{i=1}^{n} \left\{ \rho(d(z_i^t \gamma; y_i)) + C(z_i^t \gamma; y_i) \right\}
\]

(1.4)

with \( C(\gamma^t z_i; y_i) \) a bias correction term given by

\[
C(s, y) = G(F(s)) + G(1 - F(s))
\]

where

\[
G(t) = \int_{0}^{t} \rho'(-\ln u)du.
\]

The estimator of Bianco and Yohai, simply denoted as BY from now on, is a particular case of (1.2) and corresponds to the univariate function \( \phi \) given by

\[
\phi_{BY}(s) = \rho(-\ln(1 - F(s))) + G(F(s)) + G(1 - F(s)) - G(1).
\]

(1.5)

The construction of the BY estimator shows that the resulting \( \phi_{BY} \) function only depends on the choice of the \( \rho \) function. In their 1996 paper, Bianco and Yohai suggested using the following function

\[
\rho(t) = \begin{cases} 
  t - \frac{t^2}{2c} & \text{if } t \leq c \\
  \frac{c}{2} & \text{otherwise}
\end{cases}
\]

(1.6)
where $c$ is a tuning parameter, but stressed that other choices are possible. For this particular case, the right panel of Figure 1 represents the corresponding $\phi_{BY}$ function. The left panel is for the classical ML. While $\phi_{BY}$ yields large but bounded values for large positive scores (corresponding to misclassified observations), the $\phi_{ML}$ function returns extremely large values for these.

![Figure 1: $\phi$ functions for the Maximum Likelihood (left) and Bianco and Yohai (right) estimators](image)

This paper investigates some properties of estimators defined by an optimization problem of the type (1.2), where we focus on the Bianco and Yohai estimator. The latter estimator has been shown to be consistent, asymptotically normal, and to have good bias and robustness properties, while still being sufficiently efficient (see Bianco and Yohai 1996). Numerical experiments we carried out confirmed these findings, but we did find one major inconvenience: when working with the loss function (1.6) it occurred frequently that the BY estimator was not existing. Even for samples without any outliers it might happen that the BY-estimator “explodes to infinity”, meaning that the minimum in (1.2) is attained at the edge of the real vector space. The contribution of the paper is twofold: (i) we propose a $\rho$ function guaranteeing existence of the BY estimator as soon as the ML estimator exists; (ii) we propose a stable and quite fast algorithm to compute the BY estimator.

The paper is organized as follows: Section 2 derives criteria for the existence of the
estimator $\hat{\gamma}_n$ at finite samples. This leads to the overlap condition of Albert and Anderson (1984) for the existence of the ML estimator and to a short discussion in Section 3 of influence functions of M-type estimators in logistic regression. Section 4 outlines a fast and stable algorithm for the computation of the estimator. Finally, in Section 5, a simulation experiment and an example compare the BY-estimator with other robust estimators in logistic regression.

## 2 Existence of the estimator

Let us look at the problem of existence of M-type estimators. Albert and Anderson (1984) proved that the ML-estimator exists when there is overlap between the observations with $y_i = 0$ and $y_i = 1$. Overlap means that we cannot separate in the space of the explicative variables the group of observations with $y_i = 1$ from the group with $y_i = 0$ by means of a hyperplane. Formally, if we denote by $I^1 = \{i \in \{1, \ldots, n\} | y_i = 1\}$ the set of observations for which $y_i = 1$ and by $I^0 = \{i \in \{1, \ldots, n\} | y_i = 0\}$ its complement, we cannot find any $\gamma \in R^p$ such that

$$z_i^T \gamma \geq 0 \quad \forall i \in I^1 \quad \text{and} \quad z_i^T \gamma \leq 0 \quad \forall i \in I^0. \quad (2.1)$$

In particular, this condition excludes the situation where all $y_i$ are equal.

For the broader class of estimators that we investigate, Proposition 1 (proved in the Appendix) shows that the same existence criterion holds, subject to some additional constraints on the function $\psi = \phi'$ appearing in the first order condition. Note that, since $\phi$ is non-decreasing, $\psi$ needs to be a positive function.

**Proposition 1.** If the following conditions hold

1. There is overlap in the sample.

2. Either

   $$\exists L_0 > 0 \quad \text{such that} \quad \psi \quad \text{is increasing on} \quad ] - \infty, L_0] \quad \text{and decreasing on} \quad [L_0, +\infty[,$$

   or $\psi$ is increasing on $] - \infty, +\infty[.$

3. $$\lim_{s \to \infty} \frac{\psi(st)}{\psi(-s)} = \infty \quad \forall \ t > 0.$$

then $\hat{\gamma}_n$ exists and is finite in norm.
The first condition is, as shown in Albert and Anderson (1984), also needed for the existence of the ML estimator. Condition 2 corresponds to two different behaviors of the $\psi$ function. It may be increasing, as for the ML estimator, or it may be redescending, as is the case for the BY estimator. Both cases are represented on Figure 2. The form of the $\psi$ function will determine the form of the influence function, as we will further explore in Section 3.

Figure 2: $\psi$ functions for the Maximum Likelihood and Bianco and Yohai estimators

Let us now turn to condition 3, which is obviously satisfied for increasing $\psi$ functions. For the redescending ones it says that the $\psi$ function should redescend to zero faster on the side of well classified observations ($s < 0$) than on the side of misclassified observations ($s > 0$). The current BY estimator, using (1.6) as loss function, does not meet condition 3. This can be checked by easy calculation, and is also suggested by Figure 2. This is mainly due to the constant behavior of the $\rho$ function for large arguments. Its derivative, given by

$$
\rho'(t) = \begin{cases} 
1 - \frac{t}{c} & \text{if } t \leq c \\
0 & \text{otherwise}
\end{cases}
$$

vanishes for large values of its argument. Misclassified observations tend to be downweighted too severely, while these observations contain most of the information in logistic regression. Another $\rho$ function, still being bounded to ensure robustness, will be more suitable for that
The idea is to choose a function whose derivative is given by

\[ \rho'(t) = \begin{cases} e^{-\sqrt{d}} & \text{if } t \leq d \\ e^{-\sqrt{t}} & \text{otherwise} \end{cases} \]  

(2.2)

for a given constant \( d \). The constant \( d \) should be chosen to reach a compromise between robustness and efficiency: as \( d \) increases, the procedure gains in efficiency but loses in robustness and conversely. As a compromise, we took \( d = 0.5 \). The functions \( \rho \) and \( G \) that appear in (1.5) have analytical forms:

\[ \rho(t) = \begin{cases} t e^{-\sqrt{d}} & \text{if } t \leq d \\ -2e^{-\sqrt{t}(1 + \sqrt{t})} + e^{-\sqrt{d}}(2(1 + \sqrt{d}) + d) & \text{otherwise} \end{cases} \]  

(2.3)

and

\[ G(t) = \begin{cases} t e^{-\sqrt{-\ln t}} + e^{\frac{1}{4} \sqrt{\pi}}\Phi(\sqrt{2\left(\frac{1}{2} + \sqrt{-\ln t}\right)}) - e^{-\frac{1}{4} \sqrt{\pi}} & \text{if } t \leq e^{-d} \\ e^{-\sqrt{d}t} - e^{-\frac{1}{4} \sqrt{d} + e^{\frac{1}{4} \sqrt{\pi}}\Phi(\sqrt{2\left(\frac{1}{2} + \sqrt{d}\right)})} & \text{otherwise} \end{cases} \]

where \( \Phi \) is the normal cumulative distribution function.

Figure 3: \( \phi \) and \( \psi \) functions for the Bianco and Yohai estimator with the newly proposed \( \rho \) function (2.3)

Figure 3 gives the plot of the \( \phi \) and \( \psi \) functions resulting from the newly introduced function \( \rho \). We see that the curves resemble those of the original BY estimator but with tails in the \( \psi \) function satisfying Condition 3 of Proposition 1. From now on, we will work with
this version of the Bianco and Yohai estimator, which has the advantage of always existing at samples in “overlap”, and we will also denote it by BY.

3 Influence functions

In Section 2, it was shown that the behavior of the tails of the ψ function is of importance for the existence of the M-type estimator. This function is also determining the form of the influence function. Recall that the influence function of an estimator $T$ at $H_0$, where $H_0$ is the model distribution, i.e. $P_{H_0}(Y = 1|X = x) = F(z^t \gamma_0)$, is defined by

$$IF(w, T, H_0) = \lim_{\varepsilon \to 0} \frac{T((1 - \varepsilon)H_0 + \varepsilon \Delta_w) - T(H_0)}{\varepsilon}$$

where $\Delta_w$ is the Dirac distribution having all its mass at $w = (1, x^t, y)^t$ with $x \in \mathbb{R}^p$ and $y \in \{0, 1\}$. Contamination on $y$ is restricted to the values 0 and 1 since any other choice for $y$ would be detected before starting the estimation process. For M-type estimators, as defined by (1.2) and verifying the first order condition (1.3), one easily gets the following expression (see also Hampel et al, 1986, page 316)

$$IF(w, T, H_0) = -E_{H_0} \left[ \frac{\partial^2 \varphi(t; Y)}{\partial t^2} \bigg|_{t = Z^t \gamma_0} ZZ^t \right]^{-1} \Psi(z^t \gamma_0; y)z. \quad (3.1)$$

It is worth noting some particular features of expression (3.1). Aside from a constant matrix depending only on the estimator, we see that (3.1) factors into two parts. The first part $\Psi(z^t \gamma_0; y)$ depends only on the value of the score $s = z^t \gamma_0$ and the dependent variable. The second part is the covariate value $z$. The right panel of Figure 2 and the left panel of Figure 3 give the plots of the ψ functions of the ML and BY estimators respectively. The plot of the classical estimator shows a strictly increasing function, while the ψ function of the BY estimator is redescending, indicating that the BY estimator is more robust.

Before detailing more the influence functions of the classical and BY estimators, let us distinguish between the different cases of outliers in logistic regression. By analogy with linear regression, an observation $w = (z, y)$ is called a leverage point when $z$ is outlying in the covariate space. A vertical outlier is an observation which is not a leverage point but whose residual with respect to the fitted value, given by $y - F(z^t \gamma_0)$ in this context, is large. Moreover, leverage points can be good or bad, depending on whether they have a small or large residual with respect to the imposed fit.
For the classical estimator, the function $\Psi$ is given by $\Psi(s, y) = F(s) - y$. The influence of vertical outliers on the ML estimator is therefore bounded since we always have $|F(z^t\gamma_0) - y| \leq 1$. A good leverage point corresponds to an outlying $z$ value, but with a small value for $F(z^t\gamma_0) - y$. When $z$ increases, the value of the residual $F(z^t\gamma_0) - y$ will typically tend much faster to zero than $z$ to infinity, resulting in a small influence. We can therefore say that good leverage points are harmless for the ML-estimator. But a bad leverage point will induce large values for the IF. This occurs for example by taking a misclassified observation, so with $z^t\gamma_0(y_0 - 0.5) < 0$, and letting $\|z\|$ tend to infinity, leading to an infinite value of the influence function.

In Figure 4 (a), the influence function of the estimator $\hat{\beta}_1^{\text{ML}}$ is represented at a logistic regression model with two independent normally distributed covariates, and $\gamma_0 = (0, 1, 1)^t$. Without loss of generality, we fixed $y = 0$. We see that, when the covariates both tend to $-\infty$, the influence becomes zero. These points are good leverage points. On the other hand, when they go to $\infty$, they yield bad leverage points since they correspond to misclassified points, and they give a very large influence.

![Figure 4](image_url)

Figure 4: Influence function for (a) the Maximum Likelihood estimator and (b) the Bianco and Yohai estimator of the first slope parameter at a logistic regression model with regressors from a standard bivariate normal distribution.

Figure 4 (b) pictures the IF of the BY estimator. The main difference now is that bad leverage points will also have a very small influence. The reason is that for a misclassified
observation with \( \|z\| \) tending to infinity, the value of \( \Psi(z^t \gamma_0, y) \) will become very small due to the redescending character of \( \psi \), yielding a small value for the IF in (3.1). However, despite the redescending behavior of this influence function, some outliers in the covariate space can still have a large influence on the estimator \( \text{BY} \). They correspond to bounded values of \( z^t \gamma_0 \) with \( \|z\| \) going to infinity. This occurs in the neighborhood of the hyperplane orthogonal to the parameter vector \( \gamma_0 \), which is the hyperplane trying to “separate” the groups of observations with \( y_i = 1 \) and \( y_i = 0 \). These leverage points, which are neither good or bad, may still be highly influential. So although the \( \text{BY} \) estimator is much more resistant than the ML estimator, we cannot say that it has a bounded influence function.

To obtain an overall bounded influence function, one can add a weighting step to downweight leverage points. Leverage dependant weights have already been used by several authors to render robust regression estimators more robust, see e.g. Carroll and Pederson (1993). We first have to identify points outlying in the space of the explanatory variables. The classical approach is to compute for each observation the Mahalanobis distance, based on the arithmetic mean \( \bar{x} \) and on the covariance estimator \( C \) yielding

\[
MD_i = \left\{ (x_i - \bar{x})^t C^{-1} (x_i - \bar{x}) \right\}^{\frac{1}{2}}.
\]

As this approach is not robust since \( \bar{x} \) and \( C \) are extremely sensitive to outliers, Rousseeuw and van Zomeren (1990) proposed to replace them by robust estimators of location and scatter. We will use the Minimum CovarianceDeterminant (MCD) estimator (Rousseeuw 1985) for this. The MCD estimator selects that subset of \( h \) observations out of \( n \) which minimizes the determinant of the covariance matrix corresponding to these \( h \) points. Then, the classical estimators computed solely from this optimal subset give the location and scatter MCD estimators. In this paper, \( h \) will be set at \( \left\lfloor \frac{3n}{4} \right\rfloor \) yielding a 25% breakdown point estimator. The MCD estimator is, using the algorithm FASTMCD of Rousseeuw and Van Driessen (1999), quite fast to compute. Moreover, it has a normal rate of convergence and a reasonable efficiency (Croux and Haesbroeck, 1999). The resulting “robust Mahalanobis” distances will be denoted by \( RD_i \) \((1 \leq i \leq n)\).

A weighted \( \text{BY} \) estimator (WBY) can then be defined as

\[
\hat{\gamma}_n = \arg \min_{\gamma} \sum_{i=1}^{n} \omega_i \varphi_{\text{BY}}(z_i^t \gamma; y_i)
\]  

(3.2)

where the function \( \varphi_{\text{BY}} \) is the same as for the \( \text{BY} \) estimator and where the weights are taken
\[ \omega_i = \begin{cases} 1 & \text{if } RD_i^2 \leq \chi^2_{n,0.975} \\ 0 & \text{else} \end{cases} \quad (3.3) \]

This WBY estimator remains consistent, since the weighting is only based on the \( x \)-variables, for which no distributional assumptions are being made. The influence function of this weighted estimator can easily be shown to be given by

\[ IF(w, T_W, H_0) = -E_{H_0} \left[ W(X) \frac{\partial^2 \varphi_{BY}(t; Y)}{\partial t^2} \bigg|_{t=z_0^\gamma_0} ZZ^t \right]^{-1} W(x) \Psi(z_0^\gamma_0; y) z, \]

where \( W \) is the weight function and we recall that \( w = (z^t, y)^t \) and \( z = (1, x^t)^t \). Since the weight function corresponding to the weighting scheme (3.3) equals zero for large leverage points, the influence function of the weighted Bianco and Yohai estimator is bounded. Of course, such a weighting scheme will also reduce the weight of the good leverage points, which is not necessary, and which will lead to a loss of efficiency.

Note that the same weighting step can also be applied to the classical ML estimator, yielding the Weighted Maximum Likelihood (WML) estimator. This estimator is easy to compute and has a bounded influence function. The main difference is that the WML is less resistant to vertical outliers than the BY or WBY estimator.

4 Algorithm

Computation of M-estimators is usually done by iterative algorithms. In this section, such an algorithm is outlined for the estimators defined by equation (1.2). An important feature of this algorithm is its capacity to detect “explosion” of the estimator. When using the ML or the BY estimator with \( \rho \) function as in (2.3), we know from Section 2 that this cannot happen if there is overlap in the sample. Note that there exist algorithms for computing the number of observations in overlap in a sample (see Christmann and Rousseeuw, 2001). For other choices of the loss function, as in (1.6), it might happen that we have this explosion even if there is some overlap in the data. Whether explosion might occur or not is not easy to detect in advance, but the algorithm will give a sign in presence of non-existence of the estimator. Note that even for the computation of the ML-estimator, most software programs will not warn you that \( \hat{\gamma}_{n, ML}^\rho \) explodes to infinity, but will return after a certain number of iterations an outcome having a more or less arbitrary magnitude. This algorithm
avoids these kinds of numerical problems, by rewriting the parameter vector $\gamma$ as $\gamma = \frac{\xi}{\sigma}$ with $\|\xi\| = 1$ and $\sigma = 1/\|\gamma\| \geq 0$. The optimization problem defining the M-estimator becomes, in terms of the two variables $\sigma$ and $\xi$:

$$\hat{(\sigma, \xi)} = \arg \min_{(\sigma, \xi) \in \mathbb{R}^+ \times S^{p-1}} \frac{1}{n} \sum_{i=1}^{n} \varphi\left(\frac{z^t_i \xi}{\sigma}; y_i\right).$$

(4.1)

We will minimize the above objective function by alternating between minimizations over $\sigma$ and $\xi$ separately.

- When $\xi$ is known, then (4.1) reduces to a one-dimensional optimization problem in $\sigma$, which can be solved with simple routines. We used the Splus implementation of Brent’s safeguarded polynomial interpolation procedure for univariate minimization (Brent 1973), which renders a minimum $\hat{\sigma}_1$. The parameter $\sigma$ being related to the inverse of the norm of $\gamma$ can be used to detect explosion of the parameter of interest. When $\hat{\sigma}_1$ is numerically non distinguishable from zero (so smaller than a specified cut off value of 0.0001 say), then the algorithm stops and we report an explosion of the estimation procedure. The nice thing is that explosion corresponds to $\hat{\sigma} = 0$, which is much easier to check than $\hat{\sigma} = \infty$.\(^1\)

- When $\sigma$ is known, (4.1) becomes a minimization problem under a constraint:

$$\min f(\xi) = \frac{1}{n} \sum_{i=1}^{n} \varphi\left(\frac{z^t_i \xi}{\sigma}; y_i\right) \text{ under } g(\xi) = \xi^t \xi - 1 = 0.$$  

(4.2)

If $\hat{\xi}_0$ is an initial solution for the problem (4.2), then, close to $\hat{\xi}_0$ we can approximate the function $f$ by

$$f(\hat{\xi}_0 + h) \approx f(\hat{\xi}_0) + \text{grad} f(\hat{\xi}_0)^t h,$$

which shows that the fastest decrease is in the opposite direction of the gradient of $f$ at $\hat{\xi}_0$. A classical gradient algorithm will therefore take a step $\varepsilon h$, with $h = -\text{grad} f(\hat{\xi}_0)$, for a certain (small) value of the scalar $\varepsilon$. However, $\hat{\xi}_0 + \varepsilon h$ must also satisfy the constraint. For that, let us approximate the surface $S = \{\xi \in \mathbb{R}^p | g(\xi) = 0\}$ by the tangent hyperplane at $\hat{\xi}_0$, which is given by

$$S(\hat{\xi}_0) = \{t \in \mathbb{R}^p | t = \hat{\xi}_0 + v \text{ with } v^t \text{grad} g(\hat{\xi}_0) = 0\}.$$  

\(^1\)In principle we cannot exclude that $\hat{\gamma}_n$ is exactly equal to zero, yielding an infinite value of $\hat{\sigma}$. But this situation is much more unlikely than non-existence of the estimator. Moreover, it is easy to check before starting the procedure whether 0 satisfies the first order condition or not. If this is not the case, the problem addressed here will not take place.
The idea is now to find a step $h$ such that $\hat{\xi}_0 + \varepsilon h$ satisfies approximately the constraint (i.e. belongs to $S(\hat{\xi}_0)$) while achieving the smallest value for $\varepsilon \nabla f(\hat{\xi}_0)'h$. This is obtained by projecting $-\nabla f(\hat{\xi}_0)$ onto $S(\hat{\xi}_0)$, yielding

$$h = -\nabla f(\hat{\xi}_0) + \frac{[\nabla g(\hat{\xi}_0)'\nabla f(\hat{\xi}_0)][\nabla g(\hat{\xi}_0)]}{\|\nabla g(\hat{\xi}_0)\|^2}.$$

Since $\nabla g(\hat{\xi}_0) = 2\hat{\xi}_0$, we get as step direction

$$h = -\nabla f(\hat{\xi}_0) + [\hat{\xi}_0'\nabla f(\hat{\xi}_0)]\hat{\xi}_0. \quad (4.3)$$

The updated estimate will then be given by $\hat{\xi}_1 = \hat{\xi}_0 + \varepsilon h/\|h\|$, with $h$ given by (4.3) instead of by the steepest descent direction. For a value of $\varepsilon$ small enough, we will always find a decrease of the objective function (unless if $\hat{\xi}_0$ is a local minimum). A step halving procedure helps us in selecting $\varepsilon$. Start with $\varepsilon = 1$: if $f(\hat{\xi}_1) < f(\hat{\xi}_0)$, we keep this $\hat{\xi}_1$. If the objective function has not decreased, we check whether $\hat{\xi}_i = \hat{\xi}_0 + \frac{h}{2^t\|h\|}$ with $t = 1, 2, ..., \text{maxhalf}$ will yield a decrease of the objective function. As soon as a decrease is observed, keep that value of $\hat{\xi}_1$. If no decrease is obtained after a given number of step halving, i.e. $\text{maxhalf}$ which we have set equal to 10, the procedure will report that a local minimum has been reached at $(\sigma, \hat{\xi}_1)$.

The outline of the algorithm is now clear. Starting from an initial solution $\hat{\gamma}_0 = \frac{\hat{\xi}_0}{\hat{\sigma}_0}$, we iterate towards better solutions for the global optimization search by solving alternatively the two sub-problems outlined above. As an initial solution $\hat{\gamma}_0$ we take the Weighted ML-estimator, being robust and quite fast to compute at the same time. In the next Section we will see that the BY and WBY estimator can still improve on the WML estimator in terms of efficiency, certainly in presence of vertical outliers. When no extreme leverage points are present in the sample, the choice of the starting value is not crucial, and we can even take a Least Squares fit as initial value. With $\hat{\gamma}_0$, we get $\hat{\xi}_0$ by taking the unit vector in the direction of $\hat{\gamma}_0$ as a starting value for the procedure. Then we get successively $\hat{\sigma}_0, \hat{\xi}_1, \hat{\sigma}_1, \hat{\xi}_2, ..., \text{ and by putting } \hat{\gamma}_k = \hat{\xi}_k/\hat{\sigma}_k \text{ also a sequence of approximations of the estimator } \hat{\gamma}.$

The procedure stops either when a local minimum has been reached in the step halving part of the algorithm, or when convergence has failed. In the latter case, a comment specifies the reason of the non convergence: either an explosion of the estimator has been reported, or no local minimum has been reached after a predefined number of steps (the latter in fact never happened to us with a number of steps set at 1000). Let us point out again
that using an estimator under the conditions of Proposition 1 guarantees convergence. This convergence will be to a local minimum close to the initial value. Since we worked with a good starting value, we believe that there is a high probability that this local minimum is also a global one. In case of doubt, several initial values could be tried out.

This algorithm has been implemented in Splus and is available upon request or on the authors webpages. A last remark is worth mentioning: when huge outliers occur in the logistic regression setting, often numerical imprecision occurs in the computation of the deviances given by 

\[ d(s; y_i) = -y_i \log F(s) - (1 - y_i) \log \{1 - F(s)\} \] . Instead of directly computing this expression, it can be seen that a numerically more stable and accurate formula is given by

\[ \log(1 + \exp(-\text{abs}(s))) + \text{abs}(s) \times ((y - 0.5) \times s < 0), \]

in which the second term equals abs(s) if the observation is misclassified, 0 otherwise.

5 Simulation and example

In order to test the performance of the algorithm described in Section 4, we conducted a simulation experiment. The estimator of interest is the estimator of Bianco and Yohai, as well as its reweighted version, defined with the \( \rho \) function assuring its existence as discussed in Section 2, and with value of the tuning constant set at \( d = 1/2 \). In order to measure the statistical performance (robustness, efficiency) of this estimator, the maximum likelihood estimator and its reweighted version were also included in the experiment as well as two other well-known robust estimators: the conditionally unbiased bounded influence function (CUBIF) estimator of Künsch et al. (1989) and a weighted MLE estimator with Mallows-type leverage dependent weights (Carroll and Pederson 1993). The CUBIF estimator is a consistent M-estimator which is optimal in the sense that it achieves the minimum trace of the asymptotic covariance matrix subject to an upper bound on its sensitivity. The sensitivity measures in a way the maximum influence an observation can have on a linear combination of the parameters. The MALLOWS estimator is an ML estimator computed with weights depending on a robust estimation of the center and scatter matrix of the covariates and on a particular weight function. These two robust estimators can be computed by standard available routines of the Robust Package of Splus.
Three data configurations were implemented for a dimension $p$ equal to 2 or 10 and a sample size $n = 100$: in the first one, denoted as I in the tables, the explanatory variables are distributed according to a standard normal distribution $N(0, I)$, the error terms $\varepsilon_i$ according to a logistic distribution, and the dependent variable is generated according to the following model equations

$$y_i = \begin{cases} 
0 & \text{if } z_i^t \gamma + \varepsilon_i \leq 0 \\
1 & \text{if } z_i^t \gamma + \varepsilon_i > 0
\end{cases}$$

where the true parameter values are set to $\gamma = (\alpha, \beta_1, \beta_2)^t = (0, 2, 2)^t$ when $p = 2$ and $\gamma = (0, 0, \ldots, 0)^t$ for $p = 10$. In the second and third situations, 6 outliers were added to the 100 original observations. The intermediate contamination case (II) consists of adding misclassified observations on a hyperplane parallel to the true discriminating hyperplane $y = z^t \gamma$, the shift between the two lines being equal to $1.5 \times \sqrt{p}$. The third case is concerned with bad leverage points obtained by moving the contaminating line from the true line by a shift equal to $5 \times \sqrt{p}$. A better vision of the simulation setup can be obtained by looking at Figure 5 where a typical sample contaminated by extreme outliers is represented for $p = 2$.

Figure 5: Simulated sample of size $n = 100$, where $p = 2$, and with 6 extreme bad leverage points added. The explicative variables $(x_{i1}, x_{i2})$ are indicated by the corresponding value of $y_i$, and the solid line gives the discriminating hyperplane.

The estimators were computed over $m=1000$ runs and Table 1 contains summary mea-
Table 1: Biases and Mean Squared Errors of the estimators ML, WML, BY, WBY, CUBIF and MALLOWS over 1000 simulations, in an uncontaminated situation (I), with 5% of intermediate contamination (II) and with 5% of extreme contamination (III)

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias</td>
<td>MSE</td>
<td>Bias</td>
</tr>
<tr>
<td>$p = 2$</td>
<td>ML</td>
<td>0.199</td>
<td>0.640</td>
</tr>
<tr>
<td></td>
<td>WML</td>
<td>0.229</td>
<td>1.365</td>
</tr>
<tr>
<td></td>
<td>BY</td>
<td>0.285</td>
<td>0.807</td>
</tr>
<tr>
<td></td>
<td>WBY</td>
<td>0.312</td>
<td>0.954</td>
</tr>
<tr>
<td></td>
<td>CUBIF</td>
<td>0.190</td>
<td>0.634</td>
</tr>
<tr>
<td></td>
<td>MALLOWS</td>
<td>0.191</td>
<td>0.634</td>
</tr>
<tr>
<td>$p = 10$</td>
<td>ML</td>
<td>0.018</td>
<td>0.582</td>
</tr>
<tr>
<td></td>
<td>WML</td>
<td>0.032</td>
<td>1.504</td>
</tr>
<tr>
<td></td>
<td>BY</td>
<td>0.018</td>
<td>0.576</td>
</tr>
<tr>
<td></td>
<td>WBY</td>
<td>0.036</td>
<td>1.640</td>
</tr>
<tr>
<td></td>
<td>CUBIF</td>
<td>0.016</td>
<td>0.633</td>
</tr>
<tr>
<td></td>
<td>MALLOWS</td>
<td>0.016</td>
<td>0.580</td>
</tr>
</tbody>
</table>

Biases combining the individual results for the different components of the slope estimators (results for the intercept are similar). Bias and Mean Squared Error (MSE) measures were computed as follows:

$$\text{Bias} = \frac{1}{m} \sum_{i=1}^{m} \hat{\beta}_i - \beta$$ and $$\text{MSE} = \frac{1}{m} \sum_{i=1}^{m} \| \hat{\beta}_i - \beta \|^2,$$

where $\| \cdot \|$ indicates the Euclidean norm.

When there is no contamination in the samples, the estimators behave not too differently. There is some loss in precision (increased MSE) for the estimators based on weighting, certainly for $p = 10$.

Under 5% of intermediate contamination and for $p = 2$, the estimators which perform best in terms of bias and MSE are the BY estimator and its reweighted version. If the leverage points are further away from the main data cloud, as in simulation scheme III, then we see that the weighting step becomes more advantageous. Note that the BY outperforms
here its competitors CUBIF and MALLOWS.

When $p = 10$, the reweighted estimators WML and WBY achieve smaller biases than all the others (behaving all very similarly) but also, as in the uncontaminated case, larger MSE’s. There does not seem to be much difference here between the two different types of contamination.

More generally we can say that to protect against huge leverage points, weighting the covariates pays off. The weighting step can be seen as a way of cleaning the data set before applying the estimation procedure of interest. It will lead to a small loss in efficiency, apparently getting bigger when the dimension increases. The advantage here is clearly in favor of WBY or WML (which does better than the estimator based on Mallows weights). The estimators WBY and WML are quite comparable, but for $p = 2$ and sampling schemes I and III there is a gain in efficiency when using WBY. Notice also that the CUBIF and MALLOWS estimators, often used as benchmark estimators in robust logistic regression, are nowhere superior to the other considered estimators.

We also investigated a sampling scheme with a dummy and a continuous explanatory variable. The results (not included) are very similar to those obtained in Table 1. As a conclusion from this modest simulation study, we may say that the estimator BY, or its reweighted version when necessary, have good overall behavior.

To conclude this section, let us consider a well-known data set in order to compare the estimators on a real data example. The data set, referred to as skin data, was introduced by Finney (1947) and was studied by Pregibon (1982) to point out the impact of potential influential observations in logistic regression. The binary outcomes (presence or absence of vaso-constriction of the skin of the digits after air inspiration) are explained by two explanatory variables: $x_1$ the volume of air inspired and $x_2$ the inspiration rate (both in logarithms). Table 2 gives the estimated coefficients as well as their standard errors using the estimators included in the above simulation study. Standard errors for the Bianco and Yohai estimator are obtained by estimating the asymptotic variance of the normal limiting distribution of the BY estimator, as derived by Bianco and Yohai (1996). The skin data set is difficult to handle. Indeed, two “central” points (observations 4 and 18) have been detected as being influential for the classical estimators, but deleting these points leads to a data set whose overlap relies only on one observation. Most robust procedures applied to skin data merely consist in downweighting in one way or another these two central observations.
Table 2: Estimated parameters, standard errors (se), and goodness-of-fit measures for the skin data set, for the ML, WML, BY, WBY, CUBIF and MALLOWS estimator.

<table>
<thead>
<tr>
<th></th>
<th>ML</th>
<th>WML</th>
<th>BY</th>
<th>WBY</th>
<th>CUBIF</th>
<th>MALLOWS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>$\alpha$</strong></td>
<td>Value</td>
<td>-2.875</td>
<td>-2.837</td>
<td>-6.854</td>
<td>-6.824</td>
<td>-2.847</td>
</tr>
<tr>
<td></td>
<td>se</td>
<td>1.320</td>
<td>1.352</td>
<td>10.047</td>
<td>9.850</td>
<td>1.311</td>
</tr>
<tr>
<td><strong>$\beta_1$</strong></td>
<td>Value</td>
<td>5.179</td>
<td>5.139</td>
<td>10.738</td>
<td>10.709</td>
<td>5.141</td>
</tr>
<tr>
<td></td>
<td>se</td>
<td>1.863</td>
<td>1.882</td>
<td>15.307</td>
<td>15.113</td>
<td>1.850</td>
</tr>
<tr>
<td><strong>$\beta_2$</strong></td>
<td>Value</td>
<td>4.562</td>
<td>4.508</td>
<td>9.367</td>
<td>9.335</td>
<td>4.523</td>
</tr>
<tr>
<td></td>
<td>se</td>
<td>1.847</td>
<td>1.882</td>
<td>12.779</td>
<td>12.540</td>
<td>1.824</td>
</tr>
<tr>
<td>$\chi^2_{arc}$</td>
<td></td>
<td>48.39</td>
<td>48.48</td>
<td>40.87</td>
<td>40.91</td>
<td>48.47</td>
</tr>
</tbody>
</table>

As goodness-of-fit measure, we follow the approach of Kordzakhia et al (2001) who used the chi-square statistics based on the arcsin transformation $\chi^2_{arc}$ defined as

$$
\chi^2_{arc} = \sum_{i=1}^{n} 4[\arcsin \sqrt{y_i} - \arcsin \sqrt{\pi_i}]
$$

with $\pi_i = F(z_i^\hat{\gamma})$ the fitted probabilities for $i = 1, \ldots, n$. The lower $\chi^2_{arc}$, the better the goodness-of-fit.

First of all, since the explanatory variables do not contain any outliers, the weighting step does not alter the estimator much. Then, while the two robust estimators CUBIF and MALLOWS yield estimations, standard errors and $\chi^2_{arc}$ measures really close to the classical ones, the results based on the BY procedure are much different. In fact, according to BY and WBY the explicative varaibles are not significant, while they are for the other considered estimators. As for ML applied to the data set without the two influential points, the standard errors of the BY estimators are quite big. We see also that the goodness-of-fit measure is minimized when the BY procedure is used. We therefore believe that the BY estimator yields the best results on this data set.

In Kordzakhia et al (2001), it was argued that the chisquare statistics could be exploited to select the optimal value of the tuning parameters. Table 3 gives the estimations and $\chi^2_{arc}$ measures for different values of the constant $d$. As the constant $d$ increases, the estimations based on BY get closer and closer to the classical results. The value of $d$ which minimizes the $\chi^2_{arc}$ statistics is given by 0.095 but the difference with $d = 0.5$ is not really big. This confirms that the choice $d = 0.5$ is a nice compromise between robustness and efficiency.
Table 3: Estimated parameters for the skin data set based on the Bianco and Yohai estimator for different values of the tuning constant d

<table>
<thead>
<tr>
<th></th>
<th>d = 0.095</th>
<th>d = 0.5</th>
<th>d = 1.25</th>
<th>d = 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Int</td>
<td>-7.651</td>
<td>-6.854</td>
<td>-5.271</td>
<td>-3.370</td>
</tr>
<tr>
<td>$\hat{\beta}_1$</td>
<td>11.984</td>
<td>10.738</td>
<td>8.393</td>
<td>5.764</td>
</tr>
<tr>
<td>$\hat{\beta}_2$</td>
<td>10.405</td>
<td>9.367</td>
<td>7.418</td>
<td>5.123</td>
</tr>
<tr>
<td>$\chi^2_{\text{arc}}$</td>
<td>39.793</td>
<td>40.874</td>
<td>43.309</td>
<td>47.149</td>
</tr>
</tbody>
</table>

To conclude this paper, we can state that the Bianco and Yohai estimator for logistic regression can without any doubt compete with other existing robust estimators. A fast and reliable algorithm is available. This paper focused on the estimation part for the logistic regression problem, for the testing part there is still work to be done, but the work of Cantoni and Ronchetti (2001) can be helpful here.

6 Appendix

Before proving Proposition 1, an intermediate result is needed.

**Lemma 1:** Assume that the conditions of Proposition 1 are satisfied and let $\xi \in S^{p-1}$ be fixed. Define the function $K_\xi$ by

$$K_\xi : [0, +\infty[ \rightarrow \mathbb{R} : \lambda \rightarrow \frac{1}{n} \sum_{i=1}^{n} \varphi(\lambda \xi^t z_i; y_i).$$

Then,

$$\lambda(\xi) = \inf \{ \lambda \geq 0 : \frac{dK_\xi(\lambda)}{d\lambda} > 0 \ \forall \ \tilde{\lambda} \geq \lambda \}$$

(6.1)

is finite. Moreover, the function which, to any $\xi \in S^{p-1}$, associates $\lambda(\xi)$ is continuous.

**Proof of Lemma 1:** Let $I = \{ i \in \{1, \ldots, n\} | \xi^t z_i \neq 0 \}$. By the overlap condition, $I$ is not empty. Let $M = \sup_i ||z_i||$ and denote $s_i = \xi^t z_i$, for all $1 \leq i \leq n$. Put $\delta = \frac{1}{2} \inf_{i \in I} |\xi^t z_i|$, denote

$$A_B = \{ i \in I | \text{sign}((y_i - \frac{1}{2}) s_i) < 0 \},$$

and $A_W = I \setminus A_B$. The overlap condition ensures that $A_B$ and $A_W$ are not empty. By condition 3, there exists a $L_1 > 0$ such that

$$\forall \lambda > \frac{L_1}{\delta} : \frac{\psi(\delta \lambda t)}{\psi(-\delta \lambda)} > \frac{\frac{1}{2} A_W M}{\frac{1}{2} A_B \delta},$$

18
for all $t > 0$. The symbol $\sharp$ indicates the number of elements of a set. In particular, for $t = \frac{M}{\delta}$, we get
\[
\frac{\psi(\lambda M)}{\psi(-\delta \lambda)} > \frac{\sharp A_W M}{\sharp A_B \delta},
\]  
(6.2)
and for $t = 1$,
\[
\frac{\psi(\delta \lambda)}{\psi(-\delta \lambda)} > \frac{\sharp A_W M}{\sharp A_B \delta}.
\]  
(6.3)

By definition of $K_\xi$, and recalling that $\varphi(s) = \phi(s, 0)$ and $\psi(s) = \varphi'(s)$ with $\phi(s, 0) = \phi(-s, 1)$ for all $s$, we have
\[
dK_\xi(\lambda) \frac{d}{d\lambda} = \frac{1}{n} \sum_{i \in A_B} \psi(\lambda |s_i| |s_i|) - \frac{1}{n} \sum_{i \in A_W} \psi(-\lambda |s_i| |s_i|).
\]  
(6.4)

Assuming first that $\exists L_0 > 0$ such that $\psi$ is increasing on $]-\infty, L_0]$ and decreasing on $[L_0, +\infty[$, take any $\lambda > L = \sup \{ \frac{L_0}{\delta} \}$, Since for all $i \in I$, $\delta < |s_i| < M$, it follows
\[
\frac{dK_\xi(\lambda)}{d\lambda} \geq \frac{\sharp A_B}{n} \psi(\lambda M) \delta - \frac{\sharp A_W}{n} \psi(-\lambda \delta) M > 0,
\]
where the inequality follows from (6.2). On the other hand, for $\psi$ increasing on $]-\infty, +\infty[$, take any $\lambda > L = \frac{L_0}{\delta}$. Since $-\delta < s_i < \delta$, for all $i \in I$, we have
\[
\frac{dK_\xi(\lambda)}{d\lambda} \geq \frac{\sharp A_B}{n} \psi(\lambda \delta) \delta - \frac{\sharp A_W}{n} \psi(-\lambda \delta) M > 0,
\]
where the inequality follows from (6.3). It follows from these expressions that in both cases the infimum in (6.1) exists as a finite number.

For the continuity of the function $\lambda$, consider the function
\[
g : S^{p-1} \times \mathbb{R}^+ \to \mathbb{R} : (u, \lambda) \to \frac{dK_u(\lambda)}{d\lambda} = \frac{1}{n} \sum_{i=1}^{n} \Psi(\lambda u^t z_i; y_i) u^t z_i.
\]
This function is continuous and satisfies $g(u, \lambda(u)) = 0$ for all $u \in S^{p-1}$. The implicit function theorem yields then the result.

\[\square\]

Proof of Proposition 1: Let us show that the function
\[
S : \gamma \to \frac{1}{n} \sum_{i=1}^{n} \varphi(\gamma^t z_i; y_i)
\]
has at least one minimum included in a ball of finite radius. Following the notations of Lemma 1, put
\[
\lambda^* = \max_{\xi \in S^{p-1}} \lambda(\xi),
\]
19
which is well defined since the function $\xi \to \lambda(\xi)$ is continuous. By definition of $\lambda(\xi)$ we have that the minimum of $S$ can never be outside the compact set $\{\gamma : \|\gamma\| \leq \lambda^*\}$. Since the function $S$ is continuous, its minimum will be attained for a certain $\gamma_n$ with $\|\gamma_n\| < \lambda^*$. \hfill \(\square\)

7 References


