KATHOLIEKE
UNIVERSITEIT
LEUVEN

DEPARTEMENT TOEGEPASTE ECONOMISCHE WETENSCHAPPEN

RESEARCH REPORT 0202

THE BREAKDOWN BEHAVIOR OF THE MAXIMUM LIKELIHOOD ESTIMATOR IN THE LOGISTIC REGRESSION MODEL

by
C. CROUX
C. FLANDRE
G. HAESBROECK

D/2002/2376/02

The Breakdown Behavior of the Maximum Likelihood Estimator in the Logistic Regression Model

Christophe Croux* Cécile Flandre† Gentiane Haesbroeck‡

Abstract: In this note we discuss the breakdown behavior of the Maximum Likelihood (ML) estimator in the logistic regression model. We formally prove that the ML-estimator never explodes to infinity, but rather breaks down to zero when adding severe outliers to a data set. Numerical experiments confirm this behavior. As a more robust alternative, a Weighted Maximum Likelihood (WML) estimator will be considered.

Keywords: Breakdown Point, Logistic Regression, Maximum Likelihood, Robust Estimation, Weighted Maximum Likelihood.

AMS subject classification: 62F35, 62G35.

*Dept. of Applied Economics, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium, Email: Christophe.Croux@econ.kuleuven.ac.be.

†Cardif Vie S.A., Chaussée de la Hulpe 150, B-1170 Bruxelles, Belgium, Email: Cecile.Flandre@cardif.be.

‡Dept. of Mathematics, University of Liège (B37), Grande Traverse 12, B-4000 Liège, Belgium, Email: G.Haesbroeck@ulg.ac.be.

1 Introduction

One aim in robust statistics is to build high breakdown point estimators. The breakdown point of an estimator tells us which percentage of the data may be corrupted before the estimator becomes completely unreliable. In linear regression models, the breakdown points of many robust estimators have been calculated. Robust estimators have also been introduced for the logistic regression model, but their breakdown points are not well established. In fact, even the study of the breakdown behavior of the classical Maximum Likelihood estimator has not been completed yet.

Christmann (1994) showed that any sensible estimator in the logistic model, robust or not, will tend to infinity if one *replaces* a certain number of observations to well chosen positions. The *replacement* breakdown point of Donoho and Huber (1983) seems therefore not to be appropriate for measuring robustness of estimators in logistic regression [1]. This has also been noticed by Künsch, Stefanski and Carroll (1989, section 4) who therefore proposed to investigate what happens when outliers are *added* to a sample.

First, we prove in Section 2 that the classical Maximum Likelihood estimator (ML) stays uniformly bounded if one adds outliers to the original sample. This contradicts the assertion of Künsch et al (1989, Section 4), who claimed that ML-estimator could tend to infinity when extreme outliers are added. On the other hand, it is shown in Section 3 that the norm of the ML-estimator always tends to zero, when adding only a few badly placed outlying observations. These results motivated a new definition of the finite sample breakdown point for an estimator in the logistic regression model.

Section 4 considers a robustified version of the ML method based on reweighting. The weighting step is based on detection and deletion of leverage points by the Minimum Covariance Determinant estimator of Rousseeuw (1985). It can be easily added to the

[1] An exception is the logistic regression model with large strata where replacement breakdown points can still be computed, e.g. see Müller and Neykov (2001).

classical ML procedure, yielding a highly robust estimator. An example and a small scale simulation study compare the breakdown behavior of these two estimators (Section 5).

2 Explosion Robustness of the ML-Estimator

Let $z_i^* = (x_i^t, y_i^*)^t \in I\!R^{p-1} \times I\!R$ $(i = 1, \dots, n)$ be realizations of independent p-dimensional random vectors $Z_i^* = (X_i^t, Y_i^*)^t$, following the model

$$Y_i^* = \alpha + X_i^t \beta + \varepsilon_i \tag{2.1}$$

where ε_i follows a symmetric distribution with a strictly increasing cumulative distribution function F. Taking $F(u) = 1/(1 + \exp(-u))$ results in the logit model, while the probit is obtained using the normal cumulative distribution function for F. Typically, in the logistic model with binary data, the underlying dependent variable Y^* is non observable, and only the dummy variable Y obtained by taking

$$y_i = \begin{cases} 0 \text{ if } y_i^* \leq 0 \\ 1 \text{ if } y_i^* > 0 \end{cases} \tag{2.2}$$

can be recorded. Therefore, we get

$$P(Y_i = y_i \mid X_i = x_i) = F(\alpha + x_i^t \beta)^{y_i} \left\{1 - F(\alpha + x_i^t \beta)\right\}^{1 - y_i} \quad \text{for } y_i = 0, 1. \tag{2.3}$$

In what follows, $Z_n = \{z_1, \dots, z_n\}$ denotes the observed sample, and we will use the notations $\gamma = (\alpha, \beta^t)^t$ and $\tilde{x}_i = (1, x_i{}^t)^t$ for all $1 \leq i \leq n$. An estimator for γ computed from the sample Z_n is denoted by $\hat{\gamma}(Z_n)$ or simply $\hat{\gamma}_n$. The ML-estimator $\hat{\gamma}_n^{ML}$ is defined as

$$\hat{\gamma}_n^{ML} = \operatorname*{argmax}_{\gamma} \log L(\gamma; Z_n) = \operatorname*{argmin}_{\gamma} \sum_{i=1}^{n} d(\gamma; z_i)$$

where $\log L(\gamma; Z_n)$ is the log-likelihood function calculated in γ and $d(\gamma; z_i) = -y_i \log F(\tilde{x}_i^t\gamma) - (1 - y_i)\log\{1 - F(\tilde{x}_i^t\gamma)\}$ stands for the deviance at observation i.

We will assume throughout the paper the existence of the ML-estimator at the observed sample, yielding a finite $\|\hat{\gamma}_n^{ML}\|$, where $\|.\|$ denotes the Euclidean norm. The latter condition leads to the overlap situation described by Albert and Anderson (1984) and Santner and Duffy (1986), excluding complete or quasi-complete separation between the observations with $y_i = 0$ and $y_i = 1$. This means that if we denote $I^1 = \{i \in \{1, ..., n\} | y_i = 1\}$ and its complement $I^0 = \{i \in \{1, ..., n\} | y_i = 0\}$, we cannot find any $\gamma \in I\!R^p$ such that

$$\tilde{x}_i^t\gamma \geq 0 \quad \forall i \in I^1 \quad \text{and} \quad \tilde{x}_i^t\gamma \leq 0 \quad \forall i \in I^0. \tag{2.4}$$

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

To study the robustness of estimators, we will introduce data contamination by adding m potential outliers to the original data set Z_n. These added observations $z_i = (x_i^t, y_i)^t$ may have completely arbitrary values for the explicative variables, meaning that we allow for leverage points in the contaminated sample. The y_i values are of course restricted to be one or zero, otherwise they are immediately identifiable as typing errors. In the following, $\hat{\gamma}(Z'_{n+m})$ denotes the estimator computed from the contaminated sample $Z'_{n+m} = \{z_1, ..., z_n, z_{n+1}, ..., z_{n+m}\}$. The *explosion breakdown point* $\varepsilon^+(\hat{\gamma}_n; Z_n)$ of the estimator $\hat{\gamma}_n$ at the sample Z_n is then defined as the minimal fraction of outliers that need to be added to the original sample before the estimator tends to infinity:

$$\varepsilon^+(\hat{\gamma}_n; Z_n) = m^+/(n + m^+) \text{ with } m^+ = \min\{m | \sup_{z_{n+1}, \dots, z_{n+m}} \|\hat{\gamma}(Z'_{n+m})\| = \infty\}.$$

(If the set over which we take the minimum is empty, then we set $\varepsilon^+(\hat{\gamma}_n; Z_n) = 1$.)

If we add outliers to Z_n, then the contaminated data set Z'_{n+m} remains in the overlap situation, so every $\hat{\gamma}(Z'_{n+m})$ remains finite. The next Theorem shows that the ML-estimator even remains uniformly bounded when adding outliers (The proof is given in the Appendix).

Theorem 1. *Suppose that* $\|\hat{\gamma}^{ML}(Z_n)\| < \infty$. *For every finite number* m *of outliers, there exists a real positive constant* $M(Z_n, m)$ *such that*

$$\sup_{z_{n+1},\dots,z_{n+m}} \|\hat{\gamma}^{ML}(Z'_{n+m})\| \leq M(Z_n, m).$$

As a corrolary we have $\varepsilon^{+}(\hat{\gamma}_n^{ML}; Z_n) = 1$. We will call this property the *explosion robustness* of the ML-estimator in logistic regression. This is quite different from the behavior of the classical estimators in linear regression, which can become arbitrarily large just by adding one single outlier.

Instead of *adding* outliers, one could also think of *replacing* good observations by contaminants. Christmann (1994) showed that the minimal number of observations that need to be replaced before the estimator tends to infinity equals the number of observations in "overlap." This number depends only on the sample and is the same for every sensible estimator. The effect of replacing good observations by outliers is quite different from the impact of adding outliers, which distinguishes the logistic regression model from the usual linear regression model. In the next section we will motivate a new definition of breakdown point for the logistic regression model.

3 Breakdown Point in Logistic Regression

We will focus on the slope parameter β. This parameter can be written as $\beta = \frac{\beta}{\|\beta\|}\|\beta\| = \theta/\sigma$ with $\|\theta\| = 1$ and $\sigma = 1/\|\beta\|$. We interpret the vector θ as the direction in which we move the "fastest" from the observations in I^0 to these from I^1, whereas σ measures this "fastness". Since the parameter θ belongs to $S^{p-2} = \{\theta \in I\!R^{p-1} | \quad \|\theta\| = 1\}$ which has no border, an estimator of θ never breaks down. On the contrary, the parameter σ

belongs to $[0, +\infty]$, including two types of possible breakdown for an estimator $\hat{\sigma}_n$. We will say that an estimator $\hat{\sigma}_n$ of σ implodes if it tends to 0 and explodes if it becomes infinite. This corresponds to an explosion of $\hat{\beta}_n$, respectively an implosion of (the norm of) $\hat{\beta}_n$. A discussion of these two extremal cases is presented below.

Case 1: If $\hat{\beta}_n$ explodes, then only the sign of $\tilde{x}_i^t \hat{\gamma}_n$ matters. The fitted probabilities will all be zero or one. We can therefore say, as in Stromberg and Ruppert (1992), that the fitted values break down.

Case 2: If $\|\hat{\beta}_n\|$ decreases to 0, the error term in (2.1) dominates. Explanatory variables have then no influence on the dummy variable y_i, so the model becomes obviously senseless. The fitted probabilities are all equal.

The addition breakdown point of $\hat{\beta}_n$ is now defined as the smallest proportion of contamination that can cause the estimator to grow to infinity or to vanish into zero.

Definition 1. *The breakdown point of an estimator $\hat{\beta}_n$ for the logistic regression model (2.3) at the sample Z_n is given by $\varepsilon^*(\hat{\beta}_n; Z_n) = m^*/(n+m^*)$ with $m^* = \min(m^+, m^-)$,*

$$\begin{aligned} m^+ &= \min\{m \in \mathbb{N}_0 | \sup_{z_{n+1},\cdots,z_{n+m}} \|\hat{\beta}(Z'_{n+m})\| = \infty\} \\ m^- &= \min\{m \in \mathbb{N}_0 | \inf_{z_{n+1},\cdots,z_{n+m}} \|\hat{\beta}(Z'_{n+m})\| = 0\}, \end{aligned}$$

where Z'_{n+m} is obtained by adding m arbitrary points to Z_n.

In the previous section it was shown that the ML-estimator never explodes, but the next theorem shows that it is always possible to find $2(p-1)$ outliers such that the ML slope estimator tends to zero while adding these well chosen points (The proof can be found in the Appendix).

Theorem 2. *At any sample Z_n, the breakdown point of the ML-estimator satisfies*

$$\varepsilon^*(\hat{\beta}^{ML}; Z_n) \leq \frac{2(p-1)}{n+2(p-1)}.$$

It follows that the asymptotic breakdown point $\lim_n \varepsilon^*(\hat{\beta}^{ML}; Z_n)$ equals zero. The above theorem formally shows the non robustness of the ML-estimator. Not because it explodes to infinity (as is often believed), but because it can implode to zero when adding outliers to the data set. It can be checked that the standard errors of the ML-estimator explode together with the estimator, but this is not true for implosion to zero. The latter type of breakdown is therefore harder to detect. The most dangerous outliers, as can be seen from the proof of Theorem 2, are misclassified observations (meaning that $\hat{\alpha}_n + x_i^t \hat{\beta}_n$ and y_i have different signs) being at the same time outlying in the space of explicative variables. We will call them *bad leverage* points.

It might be a bit strange to speak of breakdown when the estimator tends to a central point in the parameter space. A similar phenomenom is seen in the autoregressive model of order one, where the Least Squares estimator is driven to zero in presence of badly placed outliers. This example motivated Genton and Lucas (2000) to introduce a very general notion of breakdown point, which depends on the type of outlier constellation one considers and on a certain badness measure (measuring how bad an estimated parameter fits the data). When applying their definition to the logistic regression model, using bad leverage points as outlier constellation and the sum of deviances as badness measure, we obtain an expression equivalent to the implosion breakdown point considered above.

While we were able to obtain a result for the ML-estimator, the computation of the breakdown point for other estimators is much harder and will depend heavily on the sample Z_n. In the next section, we will compare the breakdown behavior of the ML-estimator with a robust estimator by means of a numerical experiment and a simulation study.

Remark: Theorem 1 implies that the intercept estimator is explosion robust. On the other hand if the slope estimator tends to zero, $\hat{\alpha}_n^{ML}$ will return $F^{-1}(\hat{p}_{n+m})$, where $0 < \hat{p}_{n+m} < 1$ is the frequency of observations in Z'_{n+m} with $y_i = 1$, which will in general be different from 0.

4 A Weighted Maximum Likelihood estimator

In applications, it is important to know whether a parameter estimate reflects the general structure in the data cloud and that the fitted model has not been corrupted by a few influential data points. Therefore, many authors have proposed robust procedures for the logistic regression model, e.g. Pregibon (1982), Copas (1988), Künsch et al. (1989), Morgenthaler (1992), Carroll and Pederson (1993), Bianco and Yohai (1996). In this section we focus on a very simple Weighted Maximum Likelihood (WML) procedure.

As we saw in Section 3, the most influential points on the ML-estimator are bad leverage points. We will try to detect these points to give them less weight. The classical approach for identifying points outlying in the space of the explanatory variables is to compute the Mahalanobis distances $MD_i = \{(x_i - T(X))^t C(X)^{-1}(x_i - T(X))\}^{\frac{1}{2}}$ based on the arithmetic mean $T(X)$ and on the covariance estimator $C(X)$. As this approach is not robust since $T(X)$ and $C(X)$ are extremely sensitive to outliers, Rousseeuw and van Zomeren (1990) suggest to replace them by robust estimators of multivariate location and scale. The resulting "robust Mahalanobis" distances will then be denoted by RD_i $(1 \leq i \leq n)$. Herefore, we will use the Minimum Covariance Determinant (MCD) estimator (Rousseeuw 1985). This estimator selects the subset of h observations out of n minimizing the determinant of the covariance matrix computed from these h points. Then, the usual average and sample covariance matrix computed from this optimal subset give the multivariate location and scale MCD estimators. It has become standard to take $h \approx [3n/4]$, yielding a 25% breakdown point estimator of multivariate location and scatter. The MCD estimator is, using the algorithm of Rousseeuw and Van Driessen (1999) fast to compute and implemented in some of the major software packages. Moreover, it has a reasonable efficiency (Croux and Haesbroeck, 1999).

A weighted maximum likelihood estimator can now be defined as

$$\hat{\gamma}_n^{WML} = \underset{\gamma}{\text{argmin}} \sum_{i=1}^{n} w_i d_i(\gamma; z_i) \tag{4.1}$$

where the weights w_i are derived as follows:

$$w_i = \begin{cases} 1 & \text{if} \quad RD_i^2 \leq \chi^2_{(p-1),0.975} \\ 0 & \text{else} \end{cases} \tag{4.2}$$

Note that this procedure also downweights the good leverage points, which are well classified observations being outlying in x-space. Since these good leverage points yield very small deviances $d(\gamma, z_i)$, they have a negligible influence on the ML-estimator. Downweighting them seems therefore not necessary. However, this may discard some computational problems arising when computing the ML estimator. Indeed, most algorithms divide at a certain point by $(1 - F(\tilde{x}_i{}^t\gamma))F(\tilde{x}_i{}^t\gamma)$, which may give numerical problems for *all* leverage points.

We do not claim that this WML-estimator has any optimality properties. Its main attractivity is that it can easily be computed using existing software. For example, in the S-plus language it is sufficient to type, with obvious names for the objects

```
robustcov<cov.mcd(x,quan=floor(3*n/4)+1)
rdsquared<mahalanobis(x,center=robustcov$center,cov=robustcov$cov)
weights<(rdsquared<=qchisq(0.975,p-1)))
wml<glm(y~x,family=binomial,subset=weights,data=data)
```

Moreover, an expression for the covariance matrix of the estimator is immediately available as

$$\text{Cov}(\hat{\gamma}_n^{WML}) \approx \left(\sum_{i=1}^{n} w_i (1 - F(\tilde{x}_i{}^t\gamma)) F(\tilde{x}_i{}^t\gamma) \tilde{x}_i \tilde{x}_i^t \right)^{-1}.$$

5 Numerical Experiments

Consider the well-known Vaso Constriction data set of Finney (1947), see also Pregibon (1982). 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). Figure 1a gives the scatter plot of these data in the covariate space, together with the y-value. To assess the effect of contamination on the estimators, we added one outlier with $(x_1, x_2, y) = (s, s, 1)$ to the $n = 39$ observations of the sample, and computed an estimator $\hat{\beta}(s)$ based on these 40 data points. By letting s move along the real line, the outlier follows the dotted line of Figure 1a. We see from the figure that for large values of s the added observation will be correctly classified and will therefore be a good leverage point. For large negative values of s we get a bad leverage point.

To visualize the influence of the contaminant $(s, s, 1)$ on the estimates, we plotted the values of $\hat{\beta}(s)$ with respect to s for the ML- and the WML-estimators in Figures 1b and 1c. Since $\hat{\beta}_n^{ML} = (5.220, 4.631)$, we see that good leverage points do not perturb the fit obtained by the ML procedure (reason why we call them "good"). On the other hand, for s tending to $-\infty$, a bad leverage point breaks the slope estimator towards zero. For the WML estimator, both good and bad leverage points have hardly any influence, illustrating its robustness. This is not surprising, since leverage points have received a zero weight. The WML is only sensible to the added point if it is close to the original sample, so looking like a regular observation. If we look at the robustness in terms of the percentage of correctly classified observations (Figure 1d), the estimator WML is better than ML for this example. In presence of a bad leverage point, the percentage of well classified observations can even get close to 50%, which is the same success rate as a random classification rule can attain.

Instead of adding only one outlier to a real data set, we will now look at the effect of adding multiple outliers by means of a modest simulation study. We simulated 1000 samples of 100 observations, following the model equations (2.1) and (2.2) with $p = 3$. The explanatory variables were generated from a normal $N(0, 10\mathbf{I}_2)$, and the error terms ε_i are according to a logistic distribution. The parameters α and β were set to 0.2 and $(0.1414, 0.1414)^t$.

Introduce now a cloud of 11 contaminants, which leads to a sample with 10% additional contamination. These 11 outliers have values for the explanatory variables coming from a $N(0, 100\mathbf{I}_2)$, and the dependent variable is first generated according to the model, but then its sign is reversed. The outliers are therefore generated as bad leverage points. The mean values as well as the mean squared errors of the ML- and WML estimators were computed over the 1000 simulation runs, once in the uncontaminated case, and once in presence of 10% of bad leverage points. Results in Table 1 show that the two procedures are almost equivalent when there is no contamination, even if the weighted estimator has somewhat bigger MSEs than ML. Under 10% of contamination, the plain ML fails in terms of bias while the weighted estimator remains almost unbiased. Under contamination, the MSE measures of WML are stable whereas the MSEs of ML increases significantly. Finally, the average number of original observations which are well classified by ML deteriorates, while this number is unchanged for WML. The WML-estimator, as well as other robust estimators introduced in the literature, is thus not only useful for estimation in contaminated samples, in which the ML-estimator becomes completely unreliable, but it also has good properties when no outliers are present.

As a conclusion, we may say that the numerical experiments confirmed the theoretical results. Moreover, a straightforward and feasible robust method is available in logistic regression models.

Acknowledgment: We wish to thank Andreas Christmann for very helpful remarks.

Appendix

Proof of Theorem 1:

For every γ, define

$$\delta(\gamma, Z_n) = \inf\left\{\rho > 0 \mid \; \exists i \in I^1 \text{ such that } \tilde{x}_i^t\gamma < -\rho \text{ or } \exists i \in I^0 \text{ such that } \tilde{x}_i^t\gamma > \rho\right\}.$$

Due to (2.4), $0 < \delta(\gamma, Z_n) < +\infty$. Indeed, if $\delta(\gamma, Z_n)$ is not finite we would have $\tilde{x}_i^t\gamma \geq 0$ $\forall i \in I^1$ and $\tilde{x}_i^t\gamma \leq 0$ $\forall i \in I^0$, which contradicts the overlap supposition. Consider the compact set $S^{p-1} = \{\gamma \in I\!R^p \mid \; \|\gamma\| = 1\}$. Since the application $\gamma \to \delta(\gamma, Z_n)$ is continuous in γ, we have

$$\delta^*(Z_n) = \inf_{\gamma \in S^{p-1}} \quad \delta(\gamma, Z_n) > 0.$$

Denote $\hat{\gamma}_{n+m}$ the ML-estimator in the logistic regression based on a contaminated sample Z'_{n+m} where arbitrary points $z_{n+1}, \ldots, z_{n+m}$ have been added. Since $\hat{\gamma}_{n+m}$ minimizes the sum of the deviances $d(\gamma; z_i)$ of the sample points, we set

$$D(\hat{\gamma}_{n+m}; Z'_{n+m}) := \min_{\gamma} \sum_{i=1}^{n+m} d(\gamma; z_i).$$

Putting D_0 the total deviance for $\gamma = 0$, and using symmetry of F, we have that

$$D_0 := D(0, Z'_{n+m}) = \sum_{i=1}^{n+m} d(0; z_i) = (n+m)\log 2.$$

Take $\tilde{z} = \exp(-D_0)$ and define

$$M(Z_n, m) = \frac{F^{-1}(1 - \tilde{z})}{\delta^*(Z_n)}, \tag{5.1}$$

which is a constant only depending on the original sample Z_n and on the number m of observations added to Z_n. Suppose now that $\hat{\gamma}_{n+m}$ satisfies

$$\|\hat{\gamma}_{n+m}\| > M(Z_n, m). \tag{5.2}$$

First of all, for each $\hat{\gamma}_{n+m} \in I\!R^p$, we know that there exists at least one $1 \leq i_0 \leq n$ such that

$$i_0 \in I^0 \text{ and } \tilde{x}_{i_0}^t \frac{\hat{\gamma}_{n+m}}{\|\hat{\gamma}_{n+m}\|} \geq \delta \left(\frac{\hat{\gamma}_{n+m}}{\|\hat{\gamma}_{n+m}\|}, Z_n \right) \geq \delta^*(Z_n) > 0. \tag{5.3}$$

or

$$i_0 \in I^1 \text{ and } \tilde{x}_{i_0}^t \frac{\hat{\gamma}_{n+m}}{\|\hat{\gamma}_{n+m}\|} \leq -\delta \left(\frac{\hat{\gamma}_{n+m}}{\|\hat{\gamma}_{n+m}\|}, Z_n \right) \leq -\delta^*(Z_n) < 0. \tag{5.4}$$

These two cases have to be studied separately:

Case 1: For i_0 verifying (5.3), it follows from (5.1) and (5.2) that

$$\begin{aligned}
D(\hat{\gamma}_{n+m}; Z'_{n+m}) &= \sum_{i=1}^{n+m} d(\hat{\gamma}_{n+m}, z_i) \\
&\geq d(\hat{\gamma}_{n+m}; z_{i_0}) \\
&= -\log \left[1 - F \left(\|\hat{\gamma}_{n+m}\| \tilde{x}_{i_0}^t \frac{\hat{\gamma}_{n+m}}{\|\hat{\gamma}_{n+m}\|} \right) \right] \\
&\geq -\log \left[1 - F \left(\|\hat{\gamma}_{n+m}\| \delta^*(Z_n) \right) \right] \\
&> -\log \left[1 - F \left(M(Z_n, m) \delta^*(Z_n) \right) \right] \\
&= -\log(\tilde{z}) = D_0.
\end{aligned}$$

Case 2: For i_0 satisfying (5.4), we obtain in a similar way

$$\begin{aligned}
D(\hat{\gamma}_{n+m}, Z'_{n+m}) &= \sum_{i=1}^{n+m} d(\hat{\gamma}_{n+m}, z_i) \\
&\geq d(\hat{\gamma}_{n+m}, z_{i_0}) \\
&= -\log \left[F \left(\|\hat{\gamma}_{n+m}\| \tilde{x}_{i_0}^t \frac{\hat{\gamma}_{n+m}}{\|\hat{\gamma}_{n+m}\|} \right) \right] \\
&= -\log \left[1 - F \left(-\|\hat{\gamma}_{n+m}\| \tilde{x}_{i_0}^t \frac{\hat{\gamma}_{n+m}}{\|\hat{\gamma}_{n+m}\|} \right) \right] \\
&\geq -\log \left[1 - F \left(\|\hat{\gamma}_{n+m}\| \delta^*(Z_n) \right) \right] \\
&> -\log \left[1 - F \left(M(Z_n, m) \delta^*(Z_n) \right) \right] \\
&= -\log(\tilde{z}) = D_0.
\end{aligned}$$

We conclude that $D(\hat{\gamma}_{n+m}, Z'_{n+m}) > D_0 = D(0, Z'_{n+m})$ implying that $\hat{\gamma}_{n+m}$ cannot be the ML-estimator. Therefore, equation (5.2) does not hold which proves the theorem. $\square$

Proof of Theorem 2:

Let $\delta > 0$ be fixed and denote $Z_n = \{(1, x_i, y_i) | 1 \leq i \leq n\}$ the observed sample. It is always possible to find a positive constant ξ such that $-\log F(-\xi) = D_0 = (n+m)\log 2$. Furthermore, set $M = \max_{1 \leq i \leq n} \|x_i\|$, $N = \frac{\xi}{\delta}$, $A = (p-1)^{\frac{1}{2}}(2N + M)$ and $m = 2(p-1)$. Take $\{e_1, \ldots, e_{p-1}\}$ the canonical basis of $I\!R^{p-1}$ and add the set of m outliers

$$\left\{z_i^0 = (1, v_i, 0), z_i^1 = (1, v_i, 1), \text{ with } v_i = Ae_i, \text{ for } i = 1, \ldots, p-1\right\}$$

to Z_n. We will prove that for all β with $\|\beta\| > \delta$ and every α

$$D\left((\alpha, \beta), Z'_{n+m}\right) > D_0 = D\left((0,0), Z'_{n+m}\right) \tag{5.5}$$

yielding that the ML-estimator verifies

$$\|\hat{\beta}^{ML}_{n+m}\| < \delta. \tag{5.6}$$

Since (5.6) will hold for every $\delta > 0$, we have proven the theorem, since it implies that we can make $\|\hat{\beta}^{ML}_{n+m}\|$ arbitrary small by adding $m = 2(p-1)$ outliers.

In order to prove (5.5), take $\|\beta\| > \delta$ and α arbitrarily, and define the $(p-2)$ dimensional hyperplane $H_\delta = \left\{x \in I\!R^{p-1}; \alpha + x^t\beta = 0\right\}$. The Euclidean distance between a vector $x \in I\!R^{p-1}$ and H_δ equals $dist(x, H_\delta) = \left|x^t \frac{\beta}{\|\beta\|} + \frac{\alpha}{\|\beta\|}\right|$. First, suppose that there exists an $1 \leq i_0 \leq p-1$ such that $dist(v_{i_0}, H_\delta) > N$. If $\beta^t v_{i_0} + \alpha > 0$, consider the outlier $z^0_{i_0}$. We obtain readily that $\beta^t v_{i_0} + \alpha > N\|\beta\| > N\delta = \xi$ and

$$\begin{aligned} d\left((\alpha, \beta), z^0_{i_0}\right) &= -\log\left(1 - F(\beta^t v_{i_0} + \alpha)\right) \\ &> -\log\left(1 - F(\xi)\right) \\ &= -\log F(-\xi) = D_0. \end{aligned} \tag{5.7}$$

For $\beta^t v_{i_0} + \alpha < 0$, the outlier $z_{i_0}^1$ will verify

$$\begin{aligned} d\left((\alpha, \beta), z_{i_0}^1\right) &= -\log\left(F(\beta^t v_{i_0} + \alpha)\right) \\ &> -\log F(-\xi) = D_0 \end{aligned} \tag{5.8}$$

since $-(\beta^t v_{i_0} + \alpha) > \xi$.

On the other hand, suppose that $dist(v_j, H_\delta) \leq N$ for all $1 \leq j \leq p-1$. Denote j_0 the index such that $|\beta_{j_0}| = \max_{1\leq j\leq p-1} |\beta_j|$. We have $(p-1)^{\frac{1}{2}}|\beta_{j_0}| \geq \|\beta\|$. First suppose that $\beta_{j_0} > 0$. Then,

$$dist(v_{j_0}, H_\delta) = \frac{|\beta^t v_{j_0} + \alpha|}{\|\beta\|} = \frac{|\alpha + \beta_{j_0} A|}{\|\beta\|} \leq N,$$

yielding $\alpha \leq N\|\beta\| - \beta_{j_0} A$ and therefore

$$-\alpha \geq \beta_{j_0} A - N\|\beta\| \geq \left(\frac{A}{(p-1)^{\frac{1}{2}}} - N\right) \|\beta\| = (M+N)\|\beta\|.$$

Take now an observation z_{i_0} from Z_n with $y_{i_0} = 1$. Then we obtain

$$\alpha + \beta^t x_{i_0} \leq \alpha + \|x_{i_0}\|\|\beta\| \leq -(M+N)\|\beta\| + M\|\beta\| = -N\|\beta\| < -N\delta = -\xi.$$

The latter inequality implies as above that

$$\begin{aligned} d\left((\alpha, \beta), z_{i_0}\right) &= -\log\left(F(\alpha + \beta^t x_{i_0})\right) \\ &> -\log F(-\xi) = D_0. \end{aligned} \tag{5.9}$$

For $\beta_{j_0} < 0$, we can prove in a similar way that there exists an observation z_{i_0} satisfying $d\left((\alpha, \beta), z_{i_0}\right) > D_0$.

From (5.7), (5.8), and (5.9), we conclude that we can always find an observation in Z'_{n+m} which contributes at least D_0 to the total deviance. This proves (5.5) and ends the proof. □

6 References

Albert, A., and Anderson, J.A. (1984). On the Existence of Maximum Likelihood Estimates in Logistic Regression Models, *Biometrika,* **71**, 1–10.

Bianco, A. M., and Yohai, V. J. (1996). Robust Estimation in the Logistic Regression Model, in *Robust Statistics, Data Analysis, and Computer Intensive Methods*, 17–34; *Lecture Notes in Statistics* **109**, Springer Verlag, Ed. H. Rieder. New York.

Carroll, R. J., and Pederson, S. (1993). On Robust Estimation in the Logistic Regression Model, *Journal of the Royal Statistical Society B,* **55**, 693–706.

Christmann, A. (1994). Least Median of Weighted Squares in Logistic Regression with Large Strata, *Biometrika,* **81**, 413–417.

Copas, J.B. (1988). Binary Regression Models for Contaminated Data, *Journal of the Royal Statistical Society B,* **50**, 225–265.

Croux, C., and Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator, *Journal of Multivariate Analysis,* **71**, 161–190.

Donoho, D.L., and Huber, P.J. (1983). The notion of breakdown point. In *A Festschrift for Erich Lehmann*, P. Bickel, K. Doksum, and J.L. Hodges, eds. Wadsworth, Belmont, CA.

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

Genton, M.G., and Lucas, A. (2000). Comprehensive Definitions of Breakdown Points for Independent and Dependent Observations. Tinbergen Institute, Discussion paper TI 2000-40/2. Under Revision for publication.

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, *Journal of the American Statistical Association,* **84**, 460–466.

Morgenthaler, S. (1992). Least Absolute Deviations fits for generalized linear models. *Biometrika*, **79**, 747–754.

Müller, C.H., and Neykov, N. (2001). Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. To appear in the *Journal of Statistical Planning and Inference.*

Pregibon, D. (1982). Resistant Fits for some commonly used Logistic Models with Medical Applications, *Biometrics,* **38**, 485–498.

Rousseeuw, P.J. (1985), Multivariate Estimation with High Breakdown Point, in: W. Grossmann, G. Pflug, I. Vincze, and W. Wertz (Ed.), *Mathematical Statistics and Applications, Vol. B*, Dordrecht: Reidel, 283–297.

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

Rousseeuw, P.J., and van Zomeren, B.C. (1990). Unmasking Multivariate Outliers and Leverage Points, *Journal of the American Statistical Association,* **85**, 633–639.

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, *Biometrika,* **73**, 755–758.

Stromberg, A., and Ruppert, D. (1992). Breakdown in nonlinear regression, *Journal of the American Statistical Association,* **87**, 991–997.

Table 1: *Simulated Bias, Mean Squared Errors, and average % of well classified observations for the ML and WML estimators over 1000 runs, once according to the model and once in the presence of 10% contamination.*

		no contamination		10% contamination	
estimates		ML	WML	ML	WML
$\hat{\alpha}$	Bias	0.0047	0.0091	-0.0411	-0.0041
	MSE	0.0453	0.0590	0.0400	0.0525
$\hat{\beta}_1$	Bias	0.0033	0.0059	-0.1188	-0.0053
	MSE	0.0055	0.0103	0.0169	0.0083
$\hat{\beta}_2$	Bias	0.0071	0.0096	-0.1164	-0.0063
	MSE	0.0048	0.0094	0.0161	0.0078
% well classified		63.76	63.39	59.35	63.46

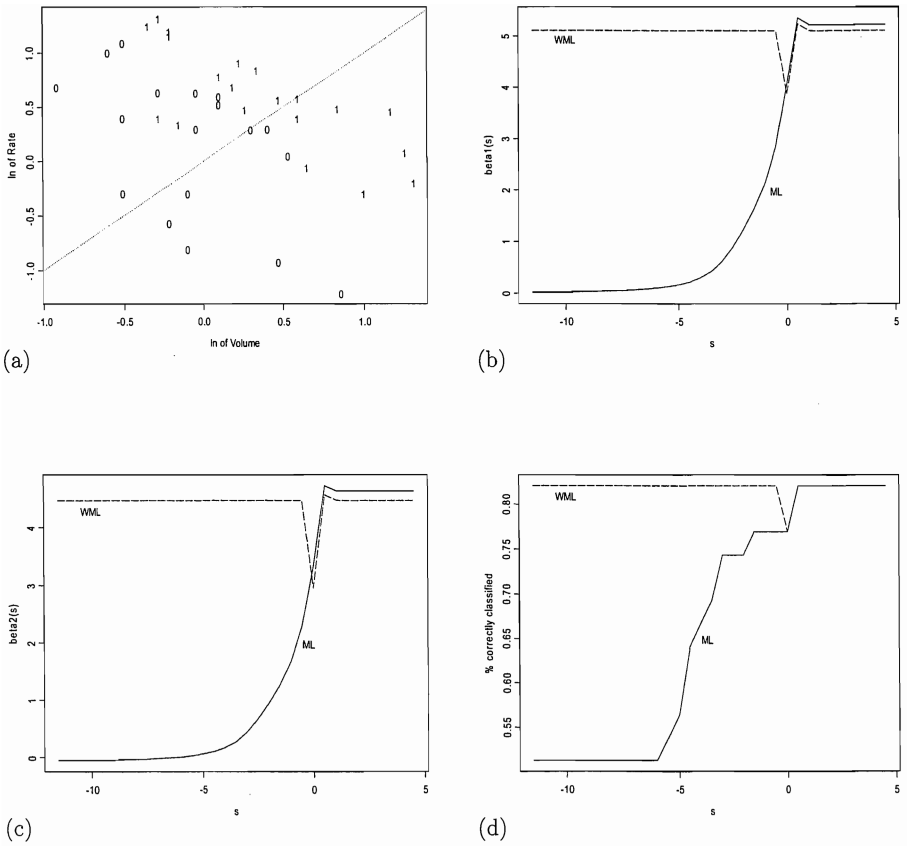

Figure 1: Stability experiment for the "Vaso Constriction" data : (a) Scatterplot of the observations (x_{1i}, x_{2i}), indicated by their y_i value. (b) Estimates of the first slope parameter, (c) estimates of the second slope parameter, (d) % of correctly classified observations, when adding $(s, s, 1)$ to the data set for the ML-estimator (solid line) and the WML-estimator (dashed line), as a function of s.