Robust Logistic Regression for Binomial Responses

Victoria-Feser, Maria-Pia

Robust Logistic Regression for Binomial Responses

Maria-Pia Victoria-Feser
University of Geneva, CH-1211 Geneva 4

August 2000

Abstract

In this paper robustness properties of the maximum likelihood estimator (MLE) and several robust estimators for the logistic regression model when the responses are binary are analysed analytically by means of the Influence Function (IF) and empirically by means of simulations. It is found that the MLE and the classical Rao's score test can be misleading in the presence of model misspecification which in the context of logistic regression means either misclassification errors in the responses or extreme data points in the design space. A general framework for robust estimation and testing is presented and a robust estimator as well as a robust testing procedure are presented. It is shown that they are less influenced by model misspecifications than their classical counterparts and they are applied to the analysis of binary data from a study on breastfeeding.

Keywords: logistic regression, misclassification, robust statistics, M-estimators, Rao's score test, influence function, breastfeeding.

1 Introduction

In many fields of social research such as psychology, binary data are quite common. They usually result from either planned experiments where the response is of the type success/failure or from observational studies where the response is yes/no for each subject in a survey. In all cases it is important that the data are analyzed properly such that inference from the sample to the population can be made.

Here we are interested in a robust approach to statistical inference for these type of data when the model is the logistic regression model. In particular, questions are addressed such as "what is the influence of data misclassification (a yes mistaken from a no) on the value of the parameter's estimates" or "what is the influence of a singular subject in the survey on the results of significance tests". We also state a general framework for robust inference and propose to the field researcher robust estimators and testing procedures.

The general robust theory is developed in Huber (1981) and Hampel, Ronchetti, Rousseeuw, and Stahel (1986), but the work of Wilcox (see Wilcox 1998 and the references therein) has opened the way for more systematic use of robust methods in psychology in particular and in the social sciences in general. It is often truly argued that robust methods are difficult to compute and software is not always available. Therefore, for the calculations made in this paper, Splus functions are available from the author upon request. Robust methods are built to deal with model misspecification. This includes for example cases such as the problem of heteroscedasticity in the linear (or ANOVA) model, as well departures from the normality of the error assumptions. In particular heavy tailed distributions, gross errors or even extreme data are dealt with by robust procedures. A question that might arise is what type of misspecification could hamper the analysis of binary data? As it will be discussed below, we can see here two types of problems: a misclassification problem whereby a yes response is mistaken by a no (or indeed the contrary) or the presence of leverage points in the independent variables. Both problems induce a misspecification in the assumption of the model that can have a dramatic effect in the results of a classical analysis.

Our analysis is based upon the Influence Function (IF) (Hampel et al. 1986) a mathematical tool which allows to investigate the robustness properties of estimators and testing procedures for a given model. It will be accompanied by illustrations based on simulations and on a real example.

The paper is organized as follows. In Section 2, the theoretical framework is set in which first the logistic model and its MLE is shortly presented, then a general framework for robust estimation is given and links are made with other results for the logistic model, a finally robust testing is developed. A robust estimator is proposed and in Section 3 it is compared to the MLE and other robust estimators through an extensive simulation study involving different parameters and contaminated samples. In Section 4, the results are applied on real data from a study on breastfeeding which is thoroughly analyzed. Finally, Section 5 concludes.

2 Theoretical results for robust inference

Several authors have studied the logistic regression model in terms of the robustness properties of the MLE (see e.g. Markatou, Basu, and Lindsay 1997, Carroll and Pederson 1993, Kuensch, Stefanski, and Carroll 1989, Copas 1988, Pregibon 1982). After setting the framework for the logistic model and shortly reviewing the main results about the robustness properties of the MLE, a general framework for robust estimation for the logistic model is introduced in which we propose a robust estimator and compare it to others through a simulation study.

2.1 The logistic model and the MLE

A very common model for the analysis of binary data is the logistic regression model. Let Y be a binary response variable (for example $Y = 1$ when the answer is yes). It is supposed that Y has the binomial distribution (B) with parameter $\mu = \mathrm{E}[Y] = P(Y = 1)$ and $n = 1$ (Y is also called a Bernoulli trial). When independent variables $\mathbf{X} = [X_1, X_2, \ldots, X_p]$ are observed, they are "linked" to the expectation of Y by means of a link function $g(\mu) = \mathbf{X}\beta$ such that $g^{-1}(\mathbf{X}\beta)$ gives values in $(0, 1)$, the definition interval of μ. $\mathbf{X}$ denotes here the design variables which usually include a constant and are supposed to be fixed. There are different possible choices for the link function, but we will consider here the canonical link (McCullagh and Nelder 1989), i.e.

$$g^{-1}(\mathbf{X}\beta) = \mu(\mathbf{X}\beta) = \frac{\exp(\mathbf{X}\beta)}{1 + \exp(\mathbf{X}\beta)}$$

With a sample (of size n) each response y_i is supposed to have expectation μ_i linked to the vector of observed independent variables $\mathbf{x}_i^{'}$ though $\mu_i = \mu(\mathbf{x}_i\beta) = \frac{\exp(\mathbf{x}_i\beta)}{1+\exp(\mathbf{x}_i\beta)}$. The n unknown expectations can therefore be determined by the specification of $p + 1$ parameters β_0, β_1, $\ldots$, β_p.

To estimate the parameter $\beta = [\beta_0, \beta_1, \ldots, \beta_p]'$ classically one uses the maximum likelihood estimator (MLE) defined as the solution in β of the score function

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_i)\mathbf{x}_i^{'} = 0$$

This is usually done by means of an iterative procedure based on the scoring method (see McCullagh and Nelder 1989).

The question we would like to address now is how is the MLE of β influenced by model misspecification? One way to study the question is by means of the IF (see Hampel et al. 1986). The IF is a function of the point which might be misspecified (or contaminated) and depends on the estimator and the model. It carries most of the information about the robustness properties of a statistic T since it measures a first-order approximation of the (asymptotic) bias of T due to an infinitesimal deviation (of any type) from the assumed model (for an illustration see Hampel et al. 1986, Figure 1, page 42). A statistic T with an unbounded or large IF is non robust since an infinitesimal model deviation can make the bias arbitrarily large. In general, the IF for the MLE is proportional to the score function, and for the logistic model, the IF is given by

$$IF((y, \mathbf{x}); \hat{\beta}, B) = J(\beta)^{-1}(y - \mu(\mathbf{x}\beta))\mathbf{x}^{'} \tag{1}$$

where $J(\beta)$ is the Fisher information matrix $E\left[(y-\mu)^2\mathbf{x}^{'}\mathbf{x}\right] = \frac{1}{n}\sum \mu_i(1-\mu_i)\mathbf{x}_i^{'}\mathbf{x}_i$. It is therefore unbounded in $\mathbf{x}$ and bounded in y. However, although the effect of extreme values in the design space is clearly dangerous, the contrary cannot be said about errors in the responses. These errors are in fact misclassification errors and have been studied by e.g. Copas (1988) and Pregibon (1982). Their results and ours (see below) show that misclassification errors can lead to a biased MLE. It is therefore important to use robust estimators which are little influenced by deviations in the $\mathbf{x}$'s as well as in the responses.

2.2 A general framework for robust estimation

In light of what was said above, we now turn to possible robust estimators for the logistic regression model. The problem comes from misclassification of the responses and also from extreme data in the design space. It is therefore important to bound both types of influences.

A question might arise at this point: what about finding outlying or extreme data by using the tools proposed by e.g. Pregibon (1982) and then remove them from the analysis? We see several drawbacks with this approach. First, removing completely one observation might be a too drastic approach in that some observations might be just at the limit of outlyingness. Moreover, it can be a tedious procedure in that successive steps are needed before a sample could be declared outlier free, often leaving the analyst with a considerably reduced sample. It is also a procedure that does not guarantee uniformity since two analysts might make different decisions. Second, as Kuensch et al. (1989) argue, a diagnostic approach might lead to situations where the estimators are numerically very unstable. Third, if data have been removed, then inference becomes very difficult if not impossible because of the dependence between the “good” and the “bad” data (see Wilcox 1998). Finally, it should be stressed that robust methods of estimation and testing are very general because they are not only robust to extreme or outlying data but also to any kind of (slight) model specifications. They do not lead to the removal of observations but instead downweight their influence upon the resulting estimator or upon the level of the test. Moreover, they provide weights which can be used as diagnostics tools to detect the suspicious data. Having said that, it should be however added that diagnostic tools can still be used in a preliminary analysis of the data.

A general formulation for a consistent robust estimator for the logistic regression model is given by the solution in β of (see also Carroll and Pederson 1993, equation 2.1)

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu(\mathbf{x}_i\beta))w_i\mathbf{x}_i^{'} - \frac{1}{n}\sum_{i=1}^{n} b(\mathbf{x}_i, \beta) = 0 \tag{2}$$

where w_i are weights that might depend on $\mathbf{x}_i$, y_i or both, and $b(\mathbf{x}_i, \beta)$ is defined to ensure consistency (see below). If $w_i = 1$ and $b(\mathbf{x}_i, \beta) = 0\ \forall i$, then (2) yields the MLE. It should be noted that (2) defines an M-estimator for which asymptotic properties are now well known (see e.g. Huber 1981, Hampel et al. 1986). For example, the asymptotic covariance matrix of the robust estimator is given by

$$V(\beta) = M^{-1}Q(M^{'})^{-1} \tag{3}$$

where

$$M = \frac{1}{n}\sum \mu_i(1-\mu_i)\mathbf{x}_i^{'}\mathbf{x}_i\left(\mu_i w_i|_{y=0} + (1-\mu_i)w_i|_{y=1}\right) \tag{4}$$

and

$$Q = \frac{1}{n}\sum \mu_i(1-\mu_i)\mathbf{x}_i^{'}\mathbf{x}_i\left(\mu_i w_i|_{y=0} + (1-\mu_i)w_i|_{y=1}\right)^2 \tag{5}$$

We consider here consistency as defined in Kuensch et al. (1989), namely conditional Fisher consistency. This means that given $\mathbf{x}_i$,

$$E\left[(y - \mu(\mathbf{x}_i\beta))w_i\mathbf{x}_i^{'}\right] - b(\mathbf{x}_i, \beta) = 0$$

$\forall i$ so that

$$b(\mathbf{x}_i, \beta) = E\left[(y - \mu(\mathbf{x}_i\beta))w_i\mathbf{x}_i^{'}\right]$$

If the weights do not depend on the response then $b(\mathbf{x}_i, \beta) = 0$.

Three robust estimators are considered in this paper. The simplest one is given by taking weights depending on the standardised residuals or Pearson residuals $\frac{y_i-\mu_i}{[\mu_i(1-\mu_i)]^{1/2}}$ (see McCullagh and Nelder 1989) as proposed in Cantoni

(1999) for generalized linear models. The weights are simply Huber-type weights given by

$$w_i = wy_i = \min\left\{1; c\left|\frac{y_i - \mu_i}{[\mu_i(1-\mu_i)]^{1/2}}\right|^{-1}\right\} \tag{6}$$

where c is a tuning constant which controls the degree of robustness (see e.g. Hampel et al. 1986). To ease the notation, we will also use $wy_i^0 = \min\left\{1; c\left|\frac{\mu_i}{[\mu_i(1-\mu_i)]^{1/2}}\right|^{-1}\right\}$ and $wy_i^1 = \min\left\{1; c\left|\frac{1-\mu_i}{[\mu_i(1-\mu_i)]^{1/2}}\right|^{-1}\right\}$. This Huber-type estimator does not consider simultaneously the problem of misclassification and extreme data in the design space. This problem could be solved by also considering a weighting scheme in the $\mathbf{x}$'s. This would lead to the a weight function of the type $w_i = wy_i \cdot wx_i$ which separates the weights on extreme residuals (wy_i) for the misclassification errors and the weights on extreme data in the design space (wx_i). One possibility would be to base wx_i on the diagonal elements of the hat matrix $H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. The hat diagonals lie between $1/n$ and 1 and their average value is p/n. Belsley, Kuh, and Welsch (1980) suggest that points with a hat diagonal greater than $2p/n$ be considered high leverage points. Therefore a choice for wx_i is given by the Huber weights

$$wx_i = \min\left\{1, \frac{2p/n}{h_{ii}}\right\}$$

We prefer, however, to choose a weighting function for the $\mathbf{x}$'s based on the IF. (2) can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\left[wy_i(y_i - \mu_i) - a_i\right]\mathbf{x}_i' wx_i = 0 \tag{7}$$

where the constants a_i ensuring conditional Fisher consistency are given by

$$a_i = E\left[wy_i(y - \mu_i)\right] = \mu_i(1-\mu_i)(wy_i^1 - wy_i^0)$$

This estimator belongs to the so-called Mallows class of estimators (Mallows 1975). Its IF is given by

$$IF((y,\mathbf{x}), \hat{\beta}, B) = M^{-1}\left[wy(y-\mu) - a\right]\mathbf{x}' wx$$

where M is given in (4) with $w_i|_{y=0} = wx_i \cdot wy_i^0$ and $w_i|_{y=1} = wx_i \cdot wy_i^1$. Actually there exists several ways to bound the IF (see Hampel et al. 1986), one of them being the standardized version given by

$$\left|IF((y,\mathbf{x}), \hat{\beta}, B)' V(\beta)^{-1} IF((y,\mathbf{x}), \hat{\beta}, B)\right|^{1/2} \leq c$$

or equivalently

$$|wy(y-\mu)-a|\,wx\left[\mathbf{x}Q^{-1}\mathbf{x}^{'}\right]^{1/2}\leq c$$

where Q is given in (5). We therefore propose here the following weighting system: Huber weights on the response's standardized residuals given by (6) and Huber weights on $\left[\mathbf{x}Q^{-1}\mathbf{x}^{'}\right]^{1/2}$, i.e.

$$wx_i=\min\left\{1;\frac{c_x}{\left[\mathbf{x}_iQ^{-1}\mathbf{x}_i^{'}\right]^{1/2}}\right\} \tag{8}$$

A referee suggested to note that the weights wx_i are a variation of Mahalanobis distances with the exception that the matrix Q is not the covariance matrix of $\mathbf{x}$. To compute the estimator one needs an iterative algorithm, whereby given a current value for the estimates one computes the weights and then a Newton-Raphson step for (7)[1]. Alternatively, by using a scoring method, these estimators can be seen as reweighted least squares estimators. Indeed, the iterative steps for the estimates are

$$\beta^{(k+1)}=\beta^{(k)}+M^{-1}\frac{1}{n}\sum\left[wy_i(y_i-\mu_i)-a_i\right]\mathbf{x}_i^{'}wx_i$$

or equivalently

$$M\beta^{(k+1)}=M\beta^{(k)}+\frac{1}{n}\sum\left[wy_i(y_i-\mu_i)-a_i\right]\mathbf{x}_i^{'}wx_i$$

However, we have that

$$M=\frac{1}{n}\mathbf{X}^{'}\mathbf{W}\mathbf{X}$$

where

$$\mathbf{W}=\operatorname{diag}\left(\left[\mu_iwy_i^0+(1-\mu_i)wy_i^1\right]\mu_i(1-\mu_i)wx_i\right)$$

Similarly

$$\frac{1}{n}\sum\left[wy_i(y_i-\mu_i)-a_i\right]\mathbf{x}_i^{'}wx_i=\frac{1}{n}\mathbf{X}^{'}\mathbf{W}\mathbf{v}$$

where

$$\mathbf{v}=\operatorname{vec}\left(\frac{wy_i(y_i-\mu_i)-a_i}{\left[\mu_iwy_i^0+(1-\mu_i)wy_i^1\right]\mu_i(1-\mu_i)}\right)=\operatorname{vec}(v_i)$$

so that we can write

$$\mathbf{X}^{'}\mathbf{W}\mathbf{X}\beta^{(k+1)}=\mathbf{X}^{'}\mathbf{W}\mathbf{z}$$

where

$$\mathbf{z}=\operatorname{vec}\left(\mathbf{x}_i\beta^{(k)}+v_i\right)$$

[1] In order to simplify the estimation, the weights in (5) are taken to be only the weights on the response, i.e. $w_i|_{y=0}=wy_i^0$ and $w_i|_{y=1}=wy_i^1$.

We note here that Markatou, Basu, and Lindsay (1997) proposed also a weighted likelihood estimating equations for the logistic regression model. It can be implicitly written as (2) (see their equation 3.5) with $b(x_i,\beta)=0,\forall i$ and a weight function w_i given (in their notation) by

$$w^k(y)=\left[\frac{A(\delta_y)+1}{\delta_y+1}\right]^k$$

with k a tuning constant, $\delta_y=\frac{y-\mu_i}{\mu_i-(1-y)}$ and A is a residual adjustment function (Lindsay 1994). Markatou, Basu, and Lindsay (1997) choose in particular $A(\delta)=2[\sqrt{\delta+1}-1]$. By using a little algebra, one can find that $E\left[w^k(y)(y-\mu)\right]\neq 0,\forall k\neq 0$, $k=0$ being the MLE. This means that this estimator is not consistent. Moreover, like the Huber-type estimator, it does not consider simultaneously the problem of misclassification and extreme data in the design space. For these reasons, we will not consider it for the simulations and the real data example.

Finally another estimator with a weight function which depends both on the design and the response has been proposed by Kuensch et al. (1989). Its weight function is given by

$$w_i=\min\left\{1;\frac{c}{|y_i-\mu_i-d_i|\left[\mathbf{x}_i\mathbf{A}^{-1}\mathbf{x}_i'\right]^{1/2}}\right\} \tag{9}$$

where $\mathbf{A}$ is defined implicitly by $Q=I$, c is a tuning constant and d_i is given by (see Kuensch et al. 1989)

$$d_i=\begin{cases}\frac{c}{\left[\mathbf{x}_i\mathbf{A}^{-1}\mathbf{x}_i'\right]^{1/2}}\frac{\mu_i}{(1-\mu_i)}-\mu_i & \text{if } \mathbf{x}_i\beta<0,\ \frac{c}{\left[\mathbf{x}_i\mathbf{A}^{-1}\mathbf{x}_i'\right]^{1/2}}<1-\mu_i\\ (1-\mu_i)-\frac{c}{\left[\mathbf{x}_i\mathbf{A}^{-1}\mathbf{x}_i'\right]^{1/2}}\frac{(1-\mu_i)}{\mu_i} & \text{if } \mathbf{x}_i\beta>0,\ \frac{c}{\left[\mathbf{x}_i\mathbf{A}^{-1}\mathbf{x}_i'\right]^{1/2}}<\mu_i\\ 0 & \text{otherwise}\end{cases}$$

The procedure to compute this estimator is rather complicated because of the implicit calculation of $\mathbf{A}$. Nevertheless, it should be stressed that this estimator not only has a bounded IF (1) with a bound controlled by a unique tuning constant c, but also that it is the most efficient estimator in the whole class of consistent M-estimators with bounded IF in which robust estimators of the type given in (2) are included. This estimator is actually the Optimal B-robust Estimator (OBRE) defined for general parametric models by Hampel et al. (1986). By means of some simulation studies and through several examples, Carroll and Pederson (1993) conclude that the OBRE has the overall best performance in terms of robustness and efficiency with reasonable sample sizes compared to the robust estimator they propose (with weights depending on $\mathbf{x}_i\beta$ through μ_i) and other (non consistent) ones.

In section 3 we present the results of a simulation study in which the MLE, the OBRE, the Huber type and Mallows type estimators are compared, from which it will be concluded that the Mallows type estimator has the best performance overall.

2.3 Testing in logistic regression

As Wilcox (1998) stresses, robustness becomes really appealing when it comes to testing. Robust theory actually started with testing procedures where the problem is to control the probability of type I error in the presence of model misspecification (see Box 1953). To investigate the robustness properties of a testing procedure one works with the asymptotic bias on the level of the test due to an infinitesimal model deviation (see Heritier and Ronchetti 1994). In the logistic regression model, as with any model from the generalized linear models, one is interested in testing the null hypothesis that a group of regressors or a linear combination of regressors is equal to 0. As a special case, one can test the significance of factors on more than two levels at the time. Several classical procedures allow this type of inference, and because we will consider their robust analogue, we concentrate here on the Rao's score test. Suppose that β is split into two parts $\beta_{(1)}$ and $\beta_{(2)}$ (and correspondingly $\mathbf{x}_{(1)}$ and $\mathbf{x}_{(2)}$) and we want to test the null hypothesis that $\beta_{(2)} = 0$. Rao's score test statistic is given by

$$R^2 = U(\dot{\beta})' J(\dot{\beta})^{-1} U(\dot{\beta})$$

where

$$U(\dot{\beta}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu(\mathbf{x}\dot{\beta}))\mathbf{x}_i'$$

and $\dot{\beta}$ is the MLE under the constraint $\beta_{(2)} = 0$. Rao's score test statistic can also be written as

$$R^2 = Z' C^{-1} Z \tag{10}$$

where

$$Z = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu(\mathbf{x}\dot{\beta}))\mathbf{x}_{(2)i}'$$

$$C = J(\dot{\beta})_{(22)} - J(\dot{\beta})_{(21)} J(\dot{\beta})_{(11)}^{-1} J(\dot{\beta})_{(12)}$$

and $\hat{\beta}_{(1)}$ is defined implicitly by

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \mu(\mathbf{x}\dot{\beta}))\mathbf{x}_{(1)i}' = 0$$

nR^2 can then be compared to the χ_q^2 distribution with $q = \dim(\beta_{(2)})$. It follows from the results of Heritier and Ronchetti (1994) that the asymptotic level of Rao's score test under a slight model misspecification in logistic regression (of amount ε) is given by

$$\alpha_0 + \varepsilon^2 \delta (y - \mu(\mathbf{x}\dot{\beta}))^2 \mathbf{x}_{(2)} J(\dot{\beta})^{-1}_{(22)} \mathbf{x}'_{(2)} \qquad (11)$$

where α_0 is the nominal level and δ is a quantity that depends on α_0 and q. By looking at (11), one can see that the level can become arbitrarily biased either because of misclassification in the response (especially when we observe $y = 0$ when μ is near 1 or when we observe $y = 1$ when μ is near 0) or when there are extreme points in the design subspace $X_{(2)}$. Heritier and Victoria-Feser (1997) examine an example of logistic regression and confirm that the level of the classical score test can be seriously biased by data contamination. It is therefore important to use a robust procedure which downweights extreme data so that the significance level is really the postulated one.

In this subsection we present the results of Heritier and Ronchetti (1994) on robust testing. In order to make the classical score tests statistic R^2 robust, the idea is to replace the score function Z in (10) by the scores function of an M-estimator, i.e.

$$Z_M = \frac{1}{n} \sum_{i=1}^{n} \psi((\mathbf{x}_i, y_i), \hat{\beta})_{(2)}$$

and $\hat{\beta}$ is obtained by solving

$$\frac{1}{n} \sum_{i=1}^{n} \psi((\mathbf{x}_i, y_i), \hat{\beta})_{(1)} = 0$$

with $\hat{\beta}_{(2)} = 0$. The robust test statistic is then obtained by standardizing Z_M by its asymptotic covariance matrix (see below). Heritier and Ronchetti (1994) show that under some fairly general conditions on the function ψ (n times) the resulting test statistic is asymptotically χ^2_q. If ψ is the score function one gets the classical score test statistic. Heritier and Ronchetti (1994) argue that the choice of ψ must be made such that the (standardized) IF of $\hat{\beta}_{(1)}$ is bounded. A choice for ψ as given in (7) with weights given by (6) and (8) satisfies this requirement. It should be stressed that it is not the same ψ function than the one proposed by Heritier and Ronchetti (1994) but a rather simpler one.

For the logistic model the robust score test statistic based on a Mallows-type estimator is given by

$$R^2_M = Z'_M C^{-1} Z_M$$

where

$$Z_M = \frac{1}{n} \sum_{i=1}^{n} \left[wy_i (y_i - \mu(\mathbf{x}_i \hat{\beta})) - a_i \right] \mathbf{x}'_{(2)i} wx_i$$

with the weights given in (6) and (8), $\hat{\beta}_{(2)} = 0$ and $\hat{\beta}_{(1)}$ defined implicitly by

$$\frac{1}{n}\sum_{i=1}^{n}\left[wy_i(y_i - \mu(\mathbf{x}_i\hat{\beta})) - a_i\right]\mathbf{x}'_{(1)i}wx_i = 0$$

The standardization matrix is given by

$$C = M_{(22.1)}V_{(22)}M'_{(22.1)}$$

where $M_{(22.1)} = M_{(22)} - M_{(21)}M^{-1}_{(11)}M_{(12)}$ and $V_{(22)}$ are obtained by computing respectively (3) and (4) at $\hat{\beta}_{(2)} = 0$ and $\hat{\beta}_{(1)}$. nR^2_M is then compared to a χ^2_q. We use this robust test statistic for the analysis of the data from the breastfeeding study.

3 Simulation study

In order to compare the different robust estimators in different settings by taking the OBRE as a benchmark, we performed a simulation study. The design matrix is made of a constant and two simulated standard normal variables. It is purposely simple because robust estimators result in high computational time in simulations. Three different sample sizes were considered, namely $n = 100$, 50, and 25. The sample size of 25 was not a good choice because all estimators where very unstable, even without contamination. For example, for the MLE without contaminated data, we got values of the bias up to 150! We also considered two arbitrary parameter sets, $\beta = (2, 3, 1)$ and $(-2, 1, 3)$ for which the true means $\mu_i = x_i^T\beta$ have mean's and standard deviation's values of respectively $(0.72, 035)$ and $(0.4, 0.4)$. We then contaminated the samples in five different ways. First we took proportions ε of responses chosen randomly and changed them from 0 to 1 or 1 to 0. This constitutes the misclassification-type error. Second we took proportions ε of x_2 and replaced them by the value of 10. This constitutes a systematic misspecification in one of the explanatory variables (which is also called leverage). Third we took proportions ε of one of x_1 or x_2 (chosen randomly) and replaced them by the value of 10. It should be stressed that only the value of one of the regressors should be contaminated, otherwise the misspecification error is confounded with a misclassification error. The aim is to create leverages in both explanatory variables. Finally, misclassification and misspecification errors where mixed, in that the first contamination type was combined either with the second or with the third. The proportions ε were of $0\%, 1\%, 2\%, \ldots, 10\%$ for $n = 100$ and of $0\%, 2\%, \ldots, 10\%$ for $n = 50$.

The simulation results are presented in the form of boxplots of the distribution of the estimates for one of the parameters. The line in the center of the box represents the median value and its distance to the horizontal line in the picture indicates the bias. The vertical size of the box in the interquartile range and gives an indication of the variability of the estimators. It is used to compare efficiencies. We believe that the boxplots give the best information about the distribution of the estimates, since it appears that the logistic model for binary data can be very unstable (whatever the estimator) in that some times the estimates are far away from the parameters's values and therefore summary statistics such as means and variances wouldn't be appropriate.

3.1 Computing the tuning constants

To compute the robust estimators, one has to choose first the tuning constant(s). In order to be fair in the comparisons, the tuning constants were chosen so that each of the robust estimator achieves the same degree of efficiency at the model compared to the MLE. The efficiency is computed here as the ratio of the traces of the asymptotic covariance matrices of the MLE and the robust estimator (see equations (3), (4) and (5)) for a given design matrix and a parameter's value. We chose an efficiency ratio of 0.85 which is for example the default value for robust regression based on M-estimators in the Splus 4.5 statistical software. To obtain the value of the tuning constant c for the Huber estimator and the OBRE, the efficiency is computed for several values of c, then a plot of c versus the efficiency ratio is produced in which the appropriate c is chosen. This has led to the choices of $c = 2.4\sqrt{3}$ (OBRE) and $c = 1.8\sqrt{3}$ (Huber) and $c = 2.75\sqrt{3}$ (OBRE) and $c = 3.4\sqrt{3}$ (Huber) for respectively $n = 50$ and $n = 100$, whatever the parameters's value. For the Mallows-type estimator, the procedure is slightly more complicated since two tuning constants c and c_x need to be chosen. One can proceed as for the other estimators and add the variation of c_x, and then produce contour plots to choose the combination of c and c_x giving the appropriate efficiency ratio. We did this for the two parameter's values and the two sample sizes. The parameter's values didn't lead to different contour plots, whereas the sample size made them slightly change. For example in Figure 1 is presented the contour plot for the case $n = 100$ and $\beta = (-2, 1, 3)$. On can then choose the tuning constants along the 0.85 contour. We chose the values $c = 2.4\sqrt{3}$ and $c_x = 4$. For $n = 50$, we chose in the same way the values of $c = 1.8\sqrt{3}$ and $c_x = 4$. It should be stressed that with real data, the true parameters are unknown, but as we already noticed, the parameters's value doesn't seem to change the efficiency ratio, whereas the sample size (and the design matrix) does, so that contour plots for a given dataset can be used to chose the tuning constants.

3.2 Estimates distributions

With misclassification errors, the simulations results depend on the sample size, the parameter which is estimated and the true parameters's values. For example, with $\beta = (2, 3, 1)$, all 4 estimators behave in the same manner, and depending on the parameter which is estimated they are (almost) unbiased for ε up to 5%: see Figure 2 for β_1. This behaviour is however not always observed. With $n = 50$ and/or with $\beta = (-2, 1, 3)$, the MLE becomes biased with already 1% of contaminated data, whereas the robust estimators remain stable with up to 4% contaminated data. For example, with $n = 100$ and $\beta = (-2, 1, 3)$, one can see in Figure 3 that for β_2 Mallows estimator performs well even with 4% of contaminated data. All simulations results cannot be presented here but the following conclusions can be drawn: with misclassification errors, the MLE can become biased with just one misclassified response whereas the robust estimators can withstand up to 4% contaminated data. However, it should be stressed that in some cases the breakdown point of the robust estimators (i.e. the smallest proportion of contaminated data they cannot anymore withstand) doesn't exceed 3%. This is the worst situation we have found.

When the contaminated data are leverage points, the simulation results are different. With leverage systematically in x_2 it is the estimates of β_2 that are most dramatically affected. However, Mallows estimator is very resistant with contaminated data at least up to 3%. For example, in Figure 4 and 5 are presented the β_2 estimates distributions for respectively $n = 100$ and $n = 50$. Only the Mallows estimator remains unbiased. Moreover, a bias can also be present for the other parameters as presented in Figure 6 for β_1. With $\beta = (2, 3, 1)$, the bias on all estimators are smaller than with $\beta = (-2, 1, 3)$, except for β_2, for which both the OBRE and Mallows estimators are resistant up to 3% of contaminated data. It therefore seems that the effect of contamination strongly depends on the parameter's value and we don't have an explanation for that. In general, with $n = 50$, the simulation results are similar to those with $n = 100$. With leverages in x_1 and/or x_2, the MLE of all three parameters can become biased with only one leverage. This is especially the case with $\beta = (-2, 1, 3)$, as is shown in Figure 7 for β_0 with $n = 50$. Among the robust estimators, once again it is Mallows's estimator which is the most resistant overall. The conclusion is that it is probable that with leverages, the MLE becomes biased with only 1% of contaminated data, that Huber estimator and the OBRE are more resistant but Mallows estimator is overall the most resistant, with at least 3% of contaminated data.

Finally, with both types of errors (misclassification and leverages), the simulation results show again that the MLE can be biased with only 1% of misclassified response and 1% leverages (this makes in reality 2% of contaminated data), whereas robust estimators are more resistant, with Mallows estimator having the best performance overall. For example, in Figure 8 are presented the distributions of β_2 estimates and one can see that only Mallows's estimator is resistant to up to 3% of both types of errors. The smallest breakdown point for Mallows's estimator we have found is of 3% of misclassification errors and 3% of leverages, i.e. 6% of contaminated data all together.

To summarize, the simulation results have shown that the MLE for the logistic model with binary data can become biased with only one either misclassified response or leverage. Huber estimator and the OBRE are more resistant than the MLE but in general to smaller amounts of contaminated data (of any type) than Mallows's estimator which in the worst situation we have found has a breakdown point of 3%. When analysing real data, it is neither possible to know where the errors might be (misclassification or leverages or both) nor their amount. Therefore it is safer to use Mallows's estimator which has overall the best performance.

3.3 Estimating standard errors

Parameters estimation is just the start to statistical inference. It is also important to be able to judge of their significance through a t-test. In order to do that, one needs to compute standard errors for the estimator and one could wonder what the effect of contaminated data is on the estimated standard errors. In the same simulation study as the one for the estimators, we also computed standard errors for the estimators using the diagonal elements of (3) with the estimated parameters for each estimator. In general we found that contamination has the effect of lowering standard errors, that the MLE can be affected by only 1% contamination, that the robust estimators are more resistant with the Mallows estimator being the most resistant overall. When $n = 50$, we also found that the OBRE estimated standard errors can be very large for some samples, thus showing some instability. This is not surprising because Kuensch et al. (1989) remarked that the effect of estimating the matrix A in (9) might have an effect on asymptotic results in small samples. The cases in which the standard errors are underestimated are the same as these when the estimators are biased, for all estimators, so that the same conclusions for all types of contaminations can be drawn for estimated standard errors. As an example, consider the case of β_2 with leverages in x_2 presented in Figure 9. The triangle in the boxplots are the standard errors computed using the true β. One can see that for the MLE the standard errors are systematically underestimated with only 1% leverages and that both the OBRE and the Mallows estimator give standard errors comparable to the 0% contamination case with up to 5% leverages. It should be stressed that underestimating standard errors means that the chance of finding significant parameters is increased or in other words that the significance tests are not made at the usual 5% significance level, but at a much larger one.

3.4 Computational time

Finally, a referee asked to compare the computational times of the different estimators. As expected, the computational times (as measures by the function unix.time in Splus which gives the cpu time needed in seconds to run a function) increase with the complexity of the estimator. The smaller computational times are for the MLE, followed by Huber estimator, the Mallows estimator and finally the OBRE. The comparisons of computational times were similar across all the simulation setting, and as an example we present in Figure 10 the different computational times for $n = 100$ and $\beta = (-2, 1, 3)$. It is clear that time is gained by using Mallows estimator compared to the OBRE.

4 Example: breastfeeding study

In this section we apply robust estimation and testing procedures on real data. Moustaki, Victoria-Feser, and Hyams (1998) conducted a study in a UK hospital on the decision of pregnant women to breastfeed their babies or not. 135 expecting mothers were asked what kind of feeding method they would use for their coming baby. Their responses where classified in two categories, one which included breastfeeding, try to breastfeed and mixed breast- and bottlefeeding and another which was only bottlefeeding. One of the aim of the study was to determine the factors which are important for a woman to choose to at least try to breastfeed and then use the results to promote breastfeeding to women with a lower probability of choosing it. The factors (variables) that were considered were the advancement of their pregnancy (beginning or end) (X_1), how they were fed as babies (only bottle- or some breastfeeding) (X_2), how their friends fed their babies (only bottle- or some breastfeeding) (X_3), if they had a partner (X_4), their age (X_5), the age at which they left full time education (X_6), their ethnic group (white or non white) (X_7) and if they smoked, stopped smoking or never smoked (X_8).

4.1 Choosing the tuning constant

Before using the robust estimators, one has to choose the tuning constants. As for the simulation study, given the design matrix, we computed several efficiency ration for different tuning constants or combinations of them. Since the parameters values doesn't seem to change very much these relations, we simply took the MLE's estimates. For the OBRE and Huber estimator, we found respectively $c = 2.5\sqrt{10}$ and $c = 0.6\sqrt{10}$ for an efficiency ratio of (approximately) 85%. For Mallows estimator, the contour plot is given in Figure 11 and one can see that $c = 0.6\sqrt{10}$ and $c_x = 4$ is a possible choice for an efficiency ratio of 85%.

4.2 Robust estimation

The different estimates using the tuning constants found above and their standard errors for the MLE and the robust estimators are given in Table 1[2]. On the whole, the estimates are of a similar order across methods. However, we can notice some substantial differences. The intercept is very large (and significant) when using the Mallows-type estimator. $\hat{\beta}_3$ (for the way friends fed their babies) is stable but becomes not significant with the Mallows-type estimator. $\hat{\beta}_6$ (for the age at the end of full time education) is 5 times higher and significant with the Mallows-type estimator compared to the MLE. Finally $\hat{\beta}_7$ (for the ethnic group) is substantially larger (in absolute value) for the Mallows-type estimator compared to the MLE.

This estimates differences means that the interpretations about the factors determining the choices of expecting mothers are also different. If one takes a classical approach, then the age at which expecting mothers leave full time education (meaning their educational level) is not important, whereas it is with a robust (Mallows) approach. On the other hand, how friends feed their babies is a significant factor with a classical approach and is not with a robust approach. Moreover, if one computes the odds ratios from the estimated parameters ($\exp(\hat{\beta})$), one finds that for a white expecting mother they are considerably smaller with a robust approach (0.045 compared to 0.141) meaning that it is considerably less probable that a white expecting mother chooses to at least try to breastfeed her baby. When comparing the three robust estimators, one also notices some differences. Huber estimator, compared to Mallows leads to a significant parameter for the way the expecting mother's friends feed their babies. The OBRE on the other side produces only two significant parameters, namely those for the smoking habit. So one might ask which result to trust? There is in our opinion no definite answer, but by construction of the estimators and from the simulations results, it is our opinion that Mallows estimator should be preferred. The only doubt would be about the significance of the factor how the expecting mother's friends feed their babies, since with Mallows estimator one can see that it is just not significant at the 5% level but would be at the 10% significance level. The gain with a robust approach with this particular data set is thus the significance of the educational level, and the different odds for a white

[2] Variables were coded as dummy, with ones for the first category of each factor. Bold estimates denote significant parameters at the 5% level.

expecting mother.

4.3 Diagnostics analysis

When performing a diagnostic check using the tools proposed by Pregibon (1982), we find after a few successive steps that three observations are extremes (namely number 75,89 and 90). If one looks at their weights using the robust estimators, we find respectively $w_{75} = 0.272$, $w_{89} = 1$, $w_{90} = 0.498$ for the OBRE, $w_{75} = 0.069$, $w_{89} = 1$, $w_{90} = 0.250$ for the Huber-type and $wy_{75} = 0.003$, $wy_{89} = 1$, $wy_{90} = 0.250$ and $wx_{75} = 0.049$, $wx_{89} = 0.127$, $wx_{90} = 0.160$ for the Mallows-type estimators. It is interesting to see that observation 89 is not downweighted by the OBRE and the Huber-type estimator, but it is by the Mallows-type estimator.

If one looks at the other weights, the picture becomes broader. In Figure 12 are plotted the weights for the OBRE and the Huber-type estimators. Both estimators found 6 either misclassified responses or leverages (or both), but not all the same. In Figure 13 are represented the weights on Y and on X for the Mallows-type estimator. The weights on the response are similar to those of the OBRE and the Huber-type estimator, except for observation 12 which is found outlying by the Mallows-type estimator. The weights on X reveal leverages that were previously not detected, for example observation 18.

4.4 Testing

With this data set, a few hypothesis are of interest. They are presented in Table 2 where they are tested classically and robustly using the results of subsection 2.3 with the tuning constants used to compute the estimates. The first hypothesis concerns the influence of the expectant mother's mother and friend in the way they fed their babies. If one uses a classical score test, we find that this influence is significant whereas a robust test fails to reject the null hypothesis. If one evaluates the influence on the social background as measured by the presence or not of a partner and the age of full time education's leave, both tests fails to reject the null hypothesis at the 5% level, but the robust one is at the borderline of significance. Finally, the factor smoking (with three levels) is clearly significant with both the classical and the robust score test statistic. These results confirm similar results by Heritier and Ronchetti (1994) on another dataset.

5 Conclusion

In this paper we have presented a general framework for robust estimation and inference based on the IF, applied to the logistic regression for the analysis of binary data. We have proposed a Mallows-type estimator and compared it with the MLE and other robust estimators all belonging to the general class of M-estimators. The findings show that the MLE can be biased in the presence of misclassification errors and extreme data in the design space, whereas the robust estimators are stable with reasonable amounts of contamination. The Mallows-type estimator is however preferred since it is more robust than the Huber estimator when there are leverages, more resistant, less complicated to compute and faster than the OBRE with reasonable amounts of contamination. For testing, a robust score test statistic is proposed that is stable under model misspecification. It is used and compared to the classical one on the breastfeeding data and it is found that the conclusions about some hypothesis can be different. We would therefore recommend to the applied researcher to at least try a robust procedure when analysing binary data. Finally, it should be stressed that the theoretical results can be extended to any model of the family of generalized linear models, but this will be the subject of other papers.

Acknowledgments

The author is partially supported by the Swiss National Science Foundation. She would like to thank Rand Wilcox, Eva Cantoni and Elvezio Ronchetti for their helpful comments on earlier versions of the paper, as well as Stephane Heritier for providing the routine to compute the OBRE.

References

Belsley, D. A., E. Kuh, and R. E. Welsch (1980). *Regression Diagnostics.* New York: Wiley.

Box, G. E. P. (1953). Non-normality and tests of variances. *Biometrika 40*, 318–335.

Cantoni, E. (1999). *Resistant Techniques for Non Parametric Regression, Generalized Linear and Additive Models.* Ph. D. thesis, University of Geneva, Switzerland.

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

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

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

Heritier, S. and E. Ronchetti (1994). Robust bounded-influence tests in general parametric models. *Journal of the American Statistical Association 89*(427), 897–904.

Heritier, S. and M.-P. Victoria-Feser (1997). Some practical applications of bounded-influence tests. In G. S. Maddala and C. Rao (Eds.), *Handbook of Statistics Vol 15: Robust Inference*, pp. 77–100. New York: Elsevier Science.

Huber, P. J. (1981). *Robust Statistics.* New York: John Wiley.

Kuensch, H. R., L. A. Stefanski, and R. J. Carroll (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.

Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *The Annals of Statistics 22*, 1081–1114.

Mallows, C. L. (1975). On some topics in robustness. Technical report, Bell Telephone Laboratories, Murray Hill, NJ.

Markatou, M., A. Basu, and B. Lindsay (1997). Weighted likelihood estimating equations: The discrete case with applications to logistic regression. *Journal of Statistical Planning and Inference 57*, 215–232.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models.* London: Chapman and Hall. Second edition.

Moustaki, I., M.-P. Victoria-Feser, and H. Hyams (1998). A UK study on the effect of socioeconomic background of pregnant women and hospital practice on the decision to breastfeed and the initiation and duration of breastfeeding. Working paper, Department of Statistics, London School of Economics.

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

Wilcox, R. R. (1998). The goals and stategies of robust methods. *British Journal of Mathematical and Statistical Psychology 51*, 1–39.

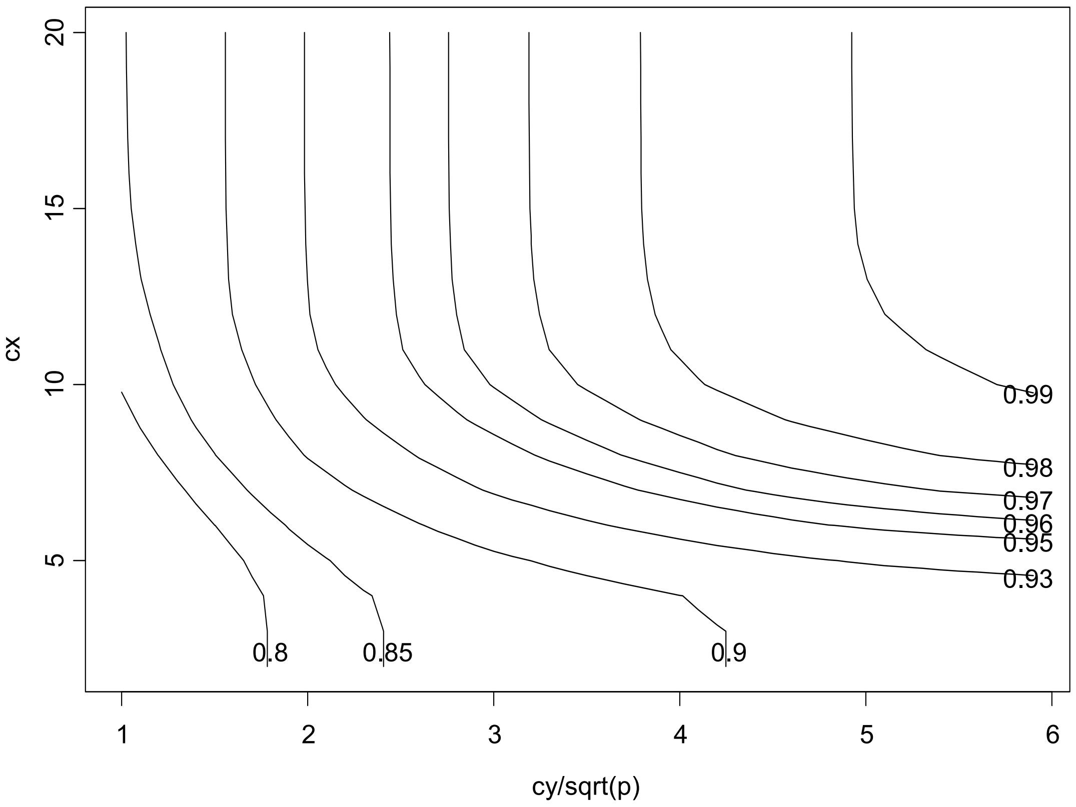

Figure 1: **Contour plot of efficiency ratio for the Mallows-type estimator**

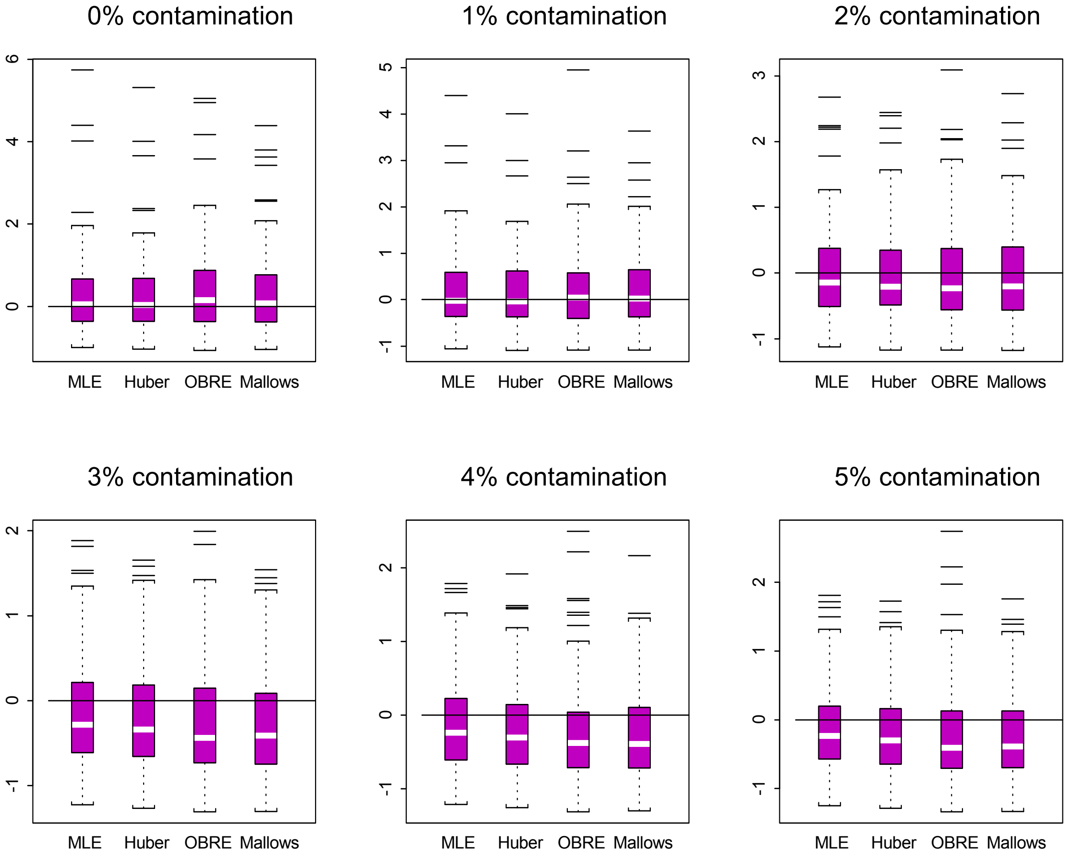

Figure 2: β_1 **estimates distributions with missclassification errors ($n = 100$, $\beta = (2, 3, 1)$)**

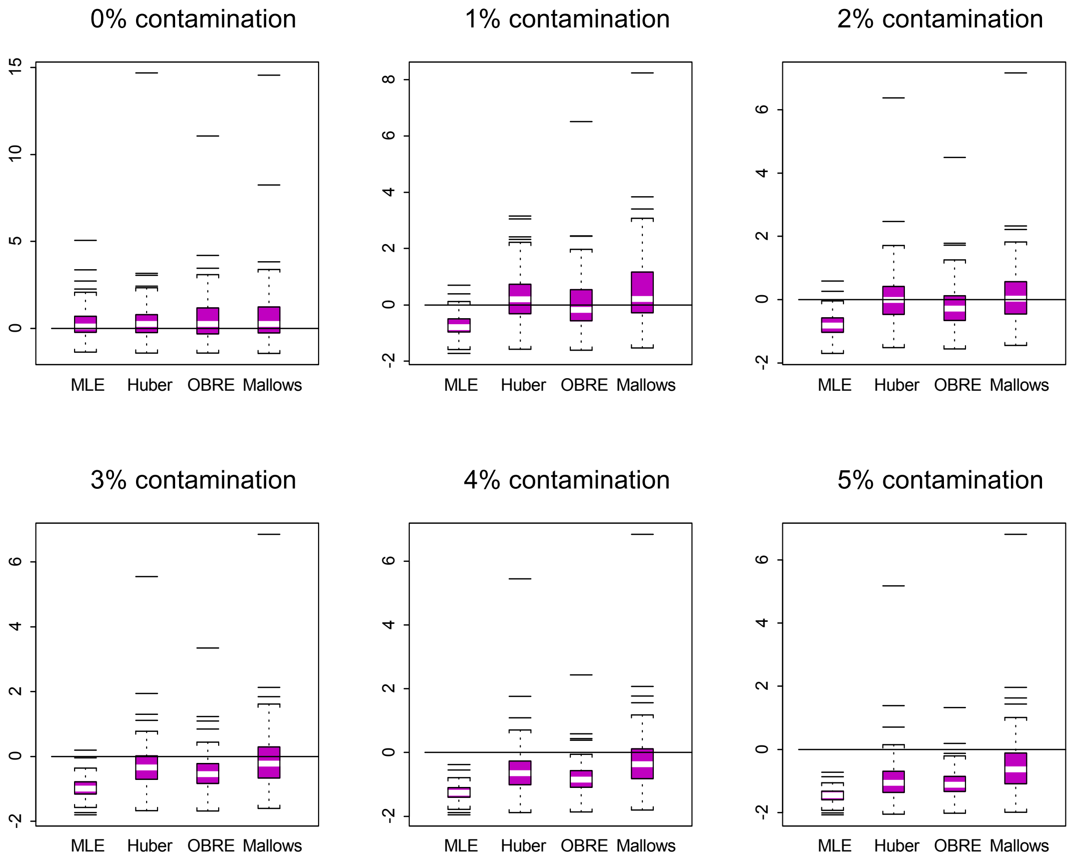

Figure 3: β_2 **estimates distribution with missclassification errors** ($n =$ 100**,** $\beta = (-2, 1, 3)$)

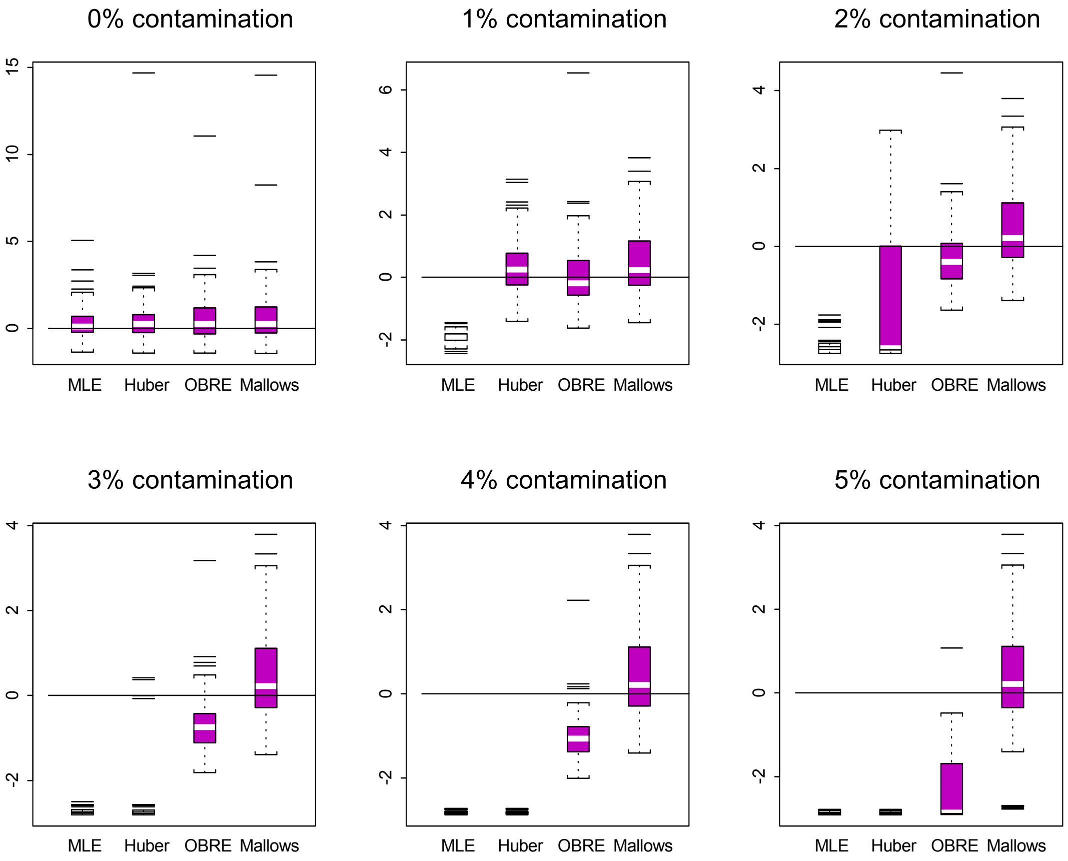

Figure 4: β_2 **estimates distribution with leverages in** x_2 ($n = 100$, $\beta = (-2, 1, 3)$)

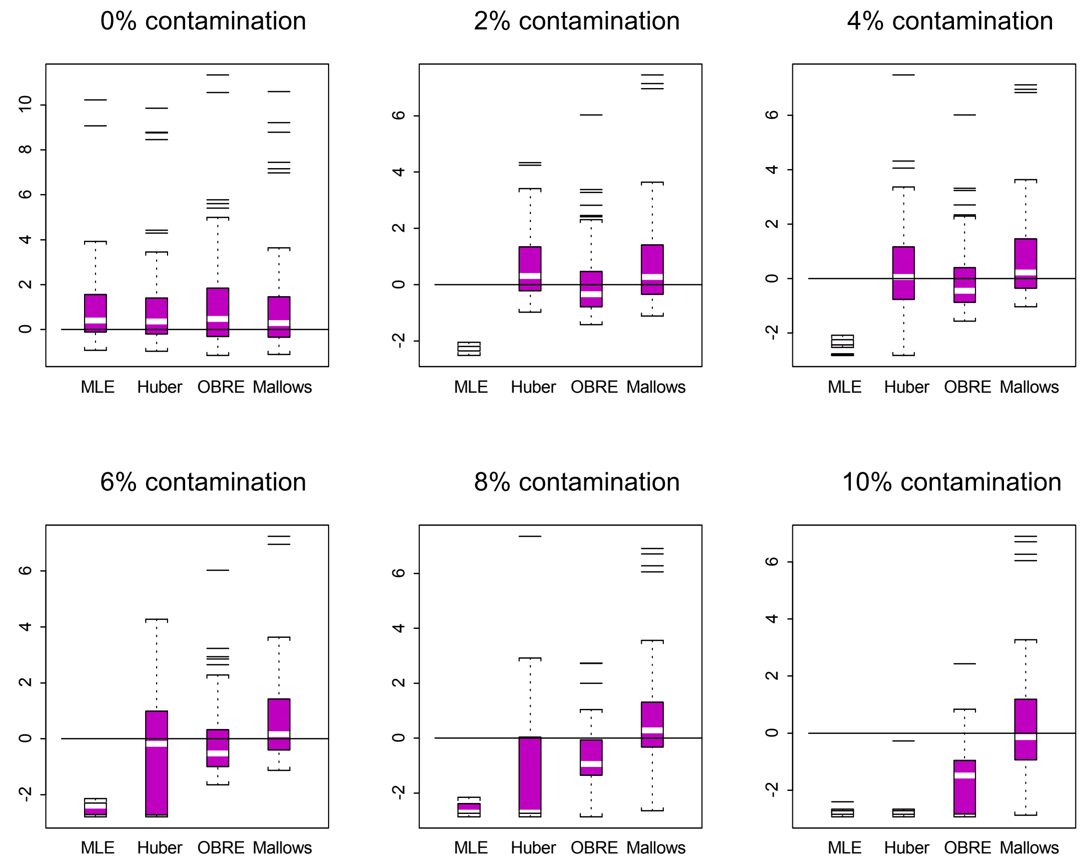

Figure 5: β_2 **estimates distribution with leverages in** x_2 **(**$n = 50$**,** $\beta = (-2, 1, 3)$**)**

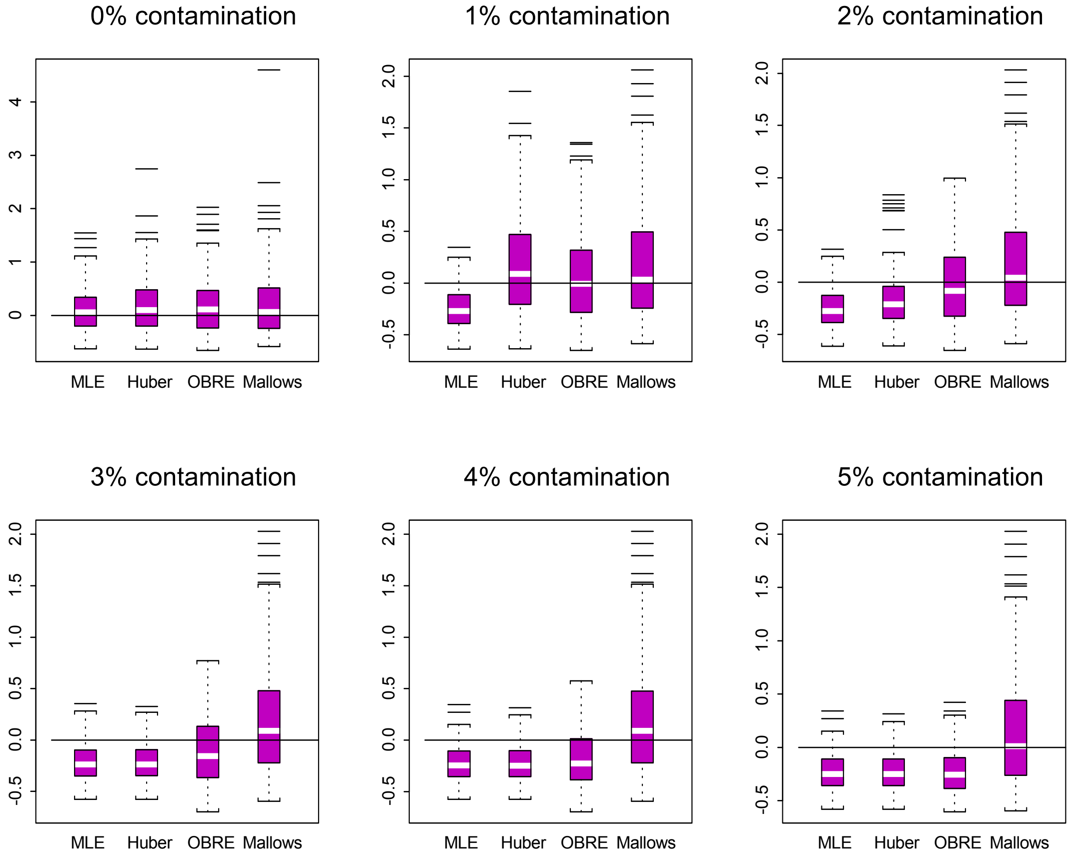

Figure 6: β_1 **estimates distribution with leverages in** x_2 ($n = 100$, $\beta = (-2, 1, 3)$)

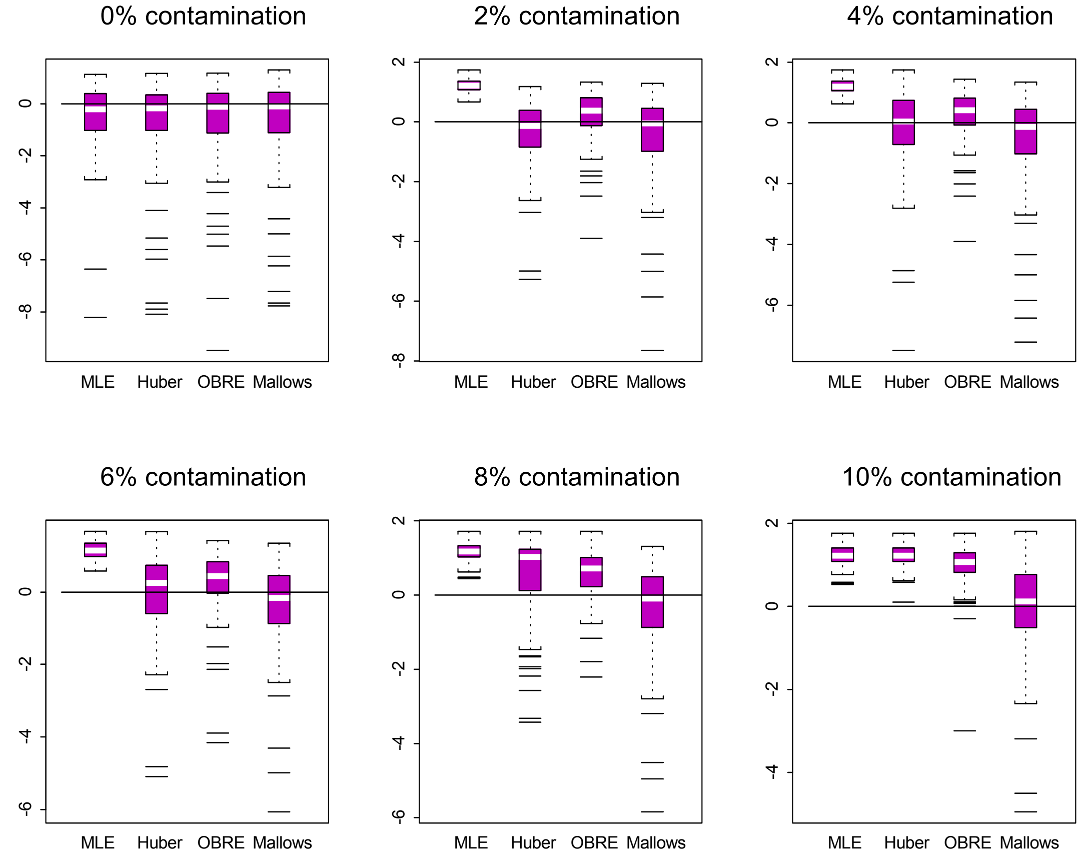

Figure 7: β_0 **estimates distribution with leverages in** x_1 **and/or** x_2 **(**$n = 50$**,** $\beta = (-2, 1, 3)$**)**

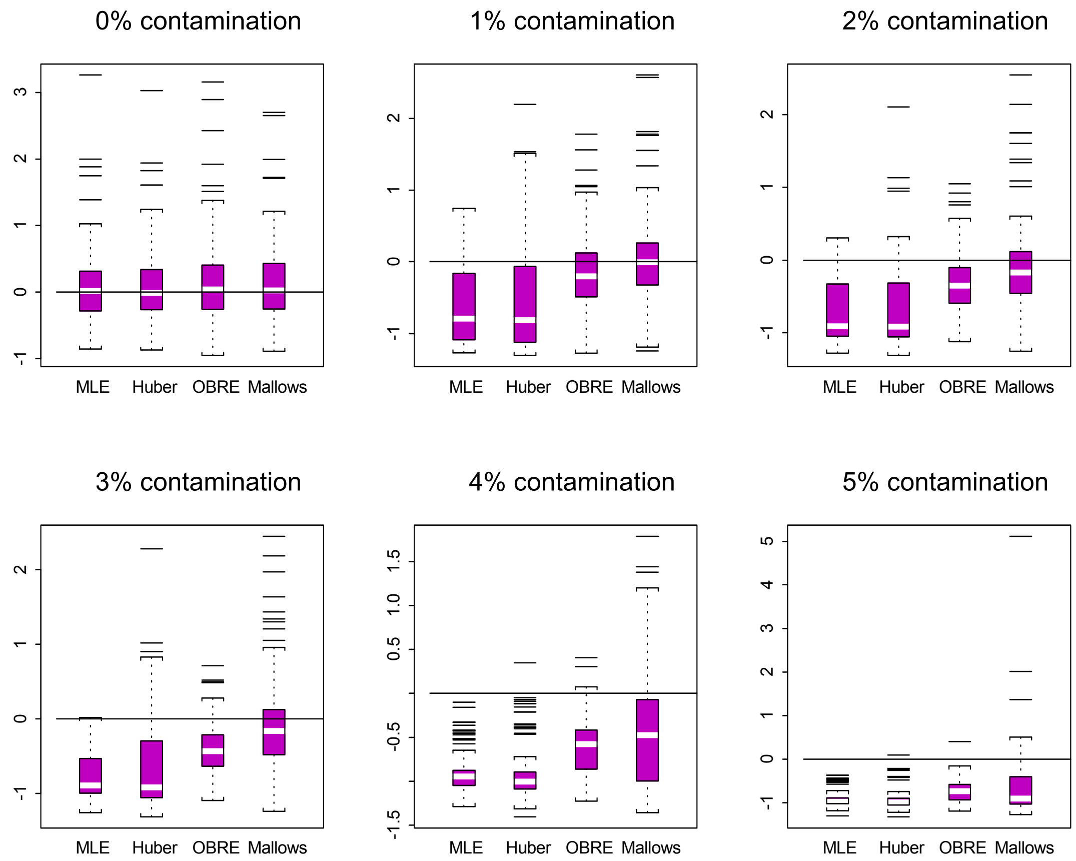

Figure 8: β_2 **estimates distribution with misclassification errors and leverages in** x_1 **and/or** x_2 **(**$n = 100$**,** $\beta = (2, 3, 1)$**)**

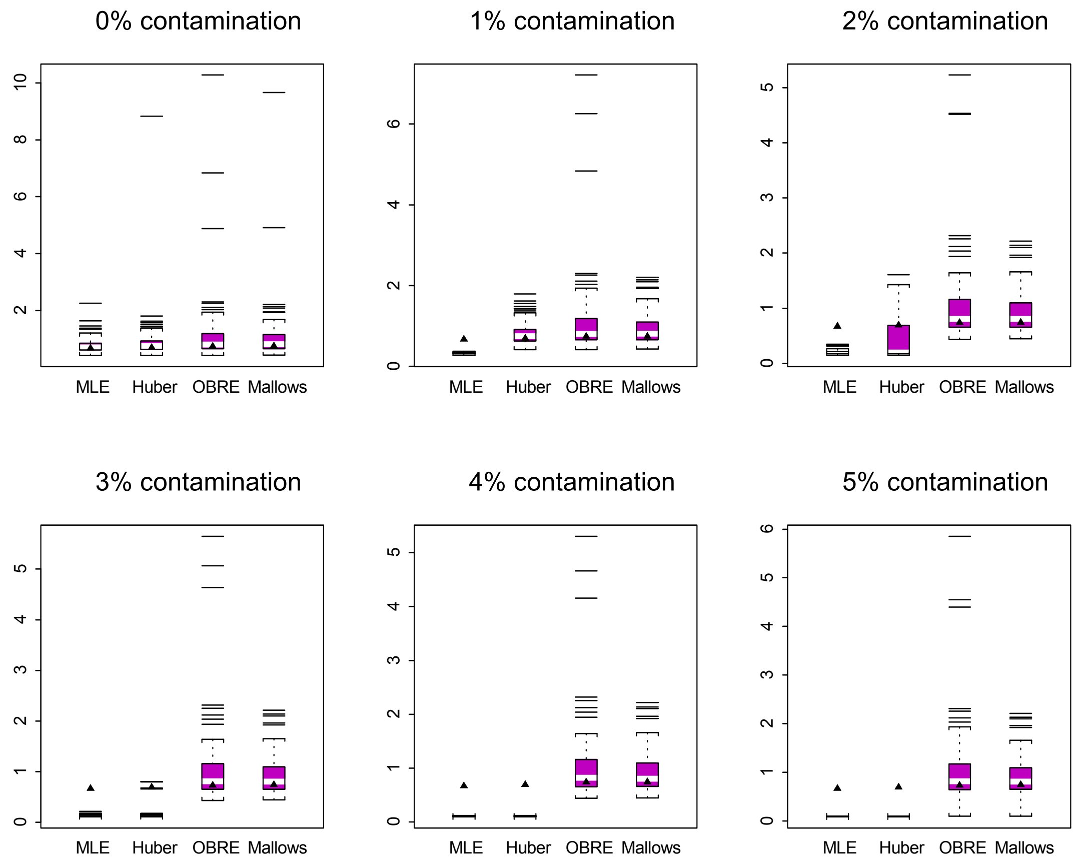

Figure 9: **Estimated standard errors distributions for β_2 with leverages in x_2 ($n = 100$, $\beta = (-2, 1, 3)$)**

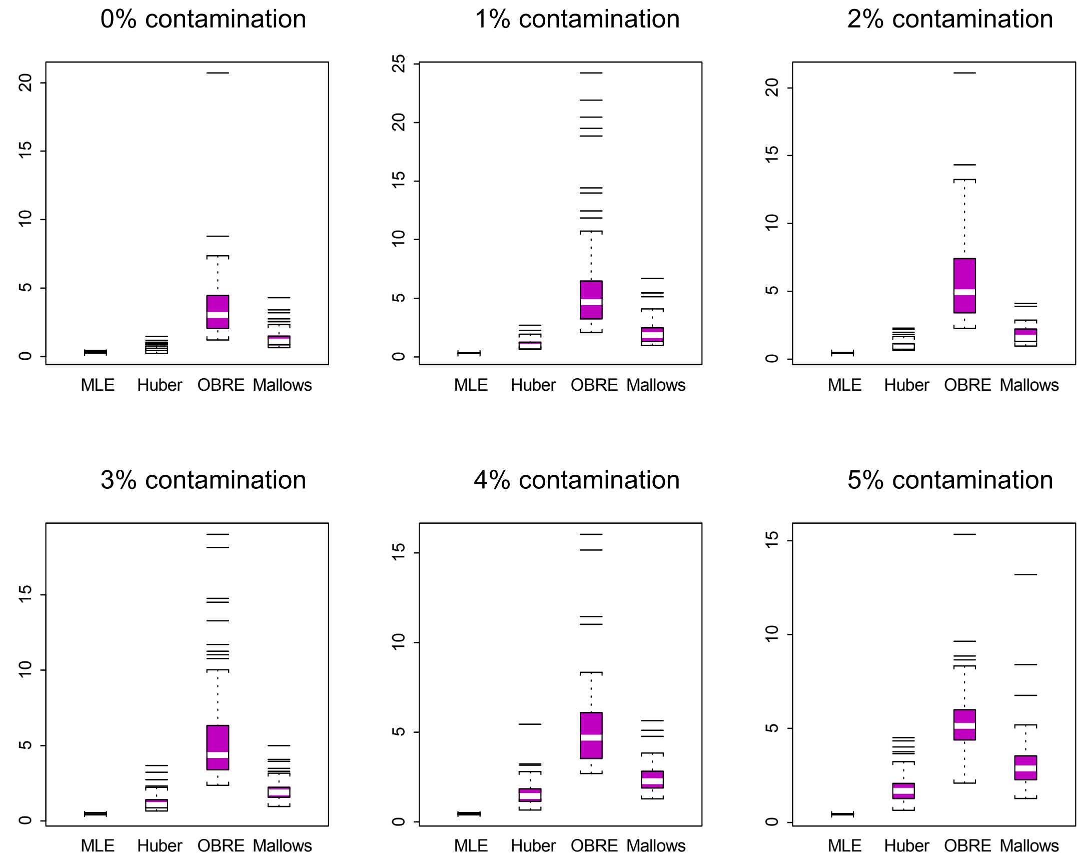

Figure 10: **Computational times for the different estimators with misclassification errors** ($n = 100$, $\beta = (-2, 1, 3)$)

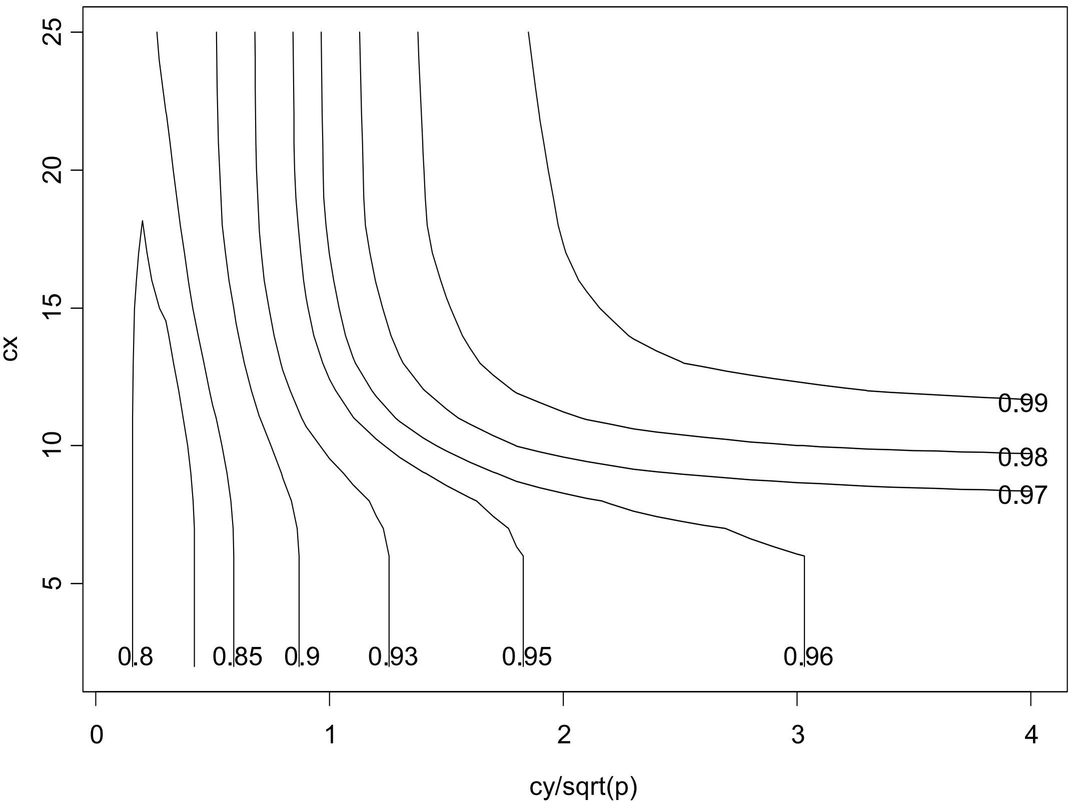

Figure 11: **Contour plot for the efficiency of Mallows estimator on the breastfeeding data**

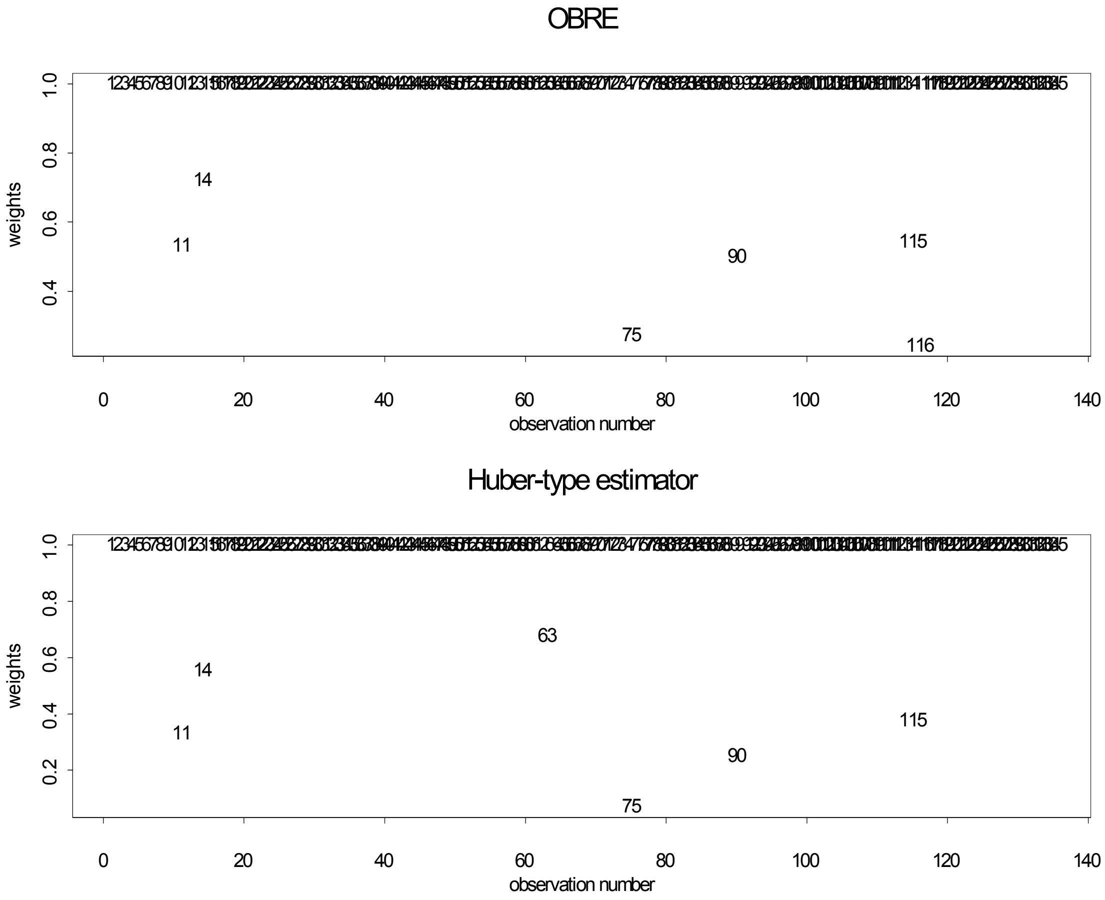

Figure 12: **Weights of robust estimators for the breastfeeding data**

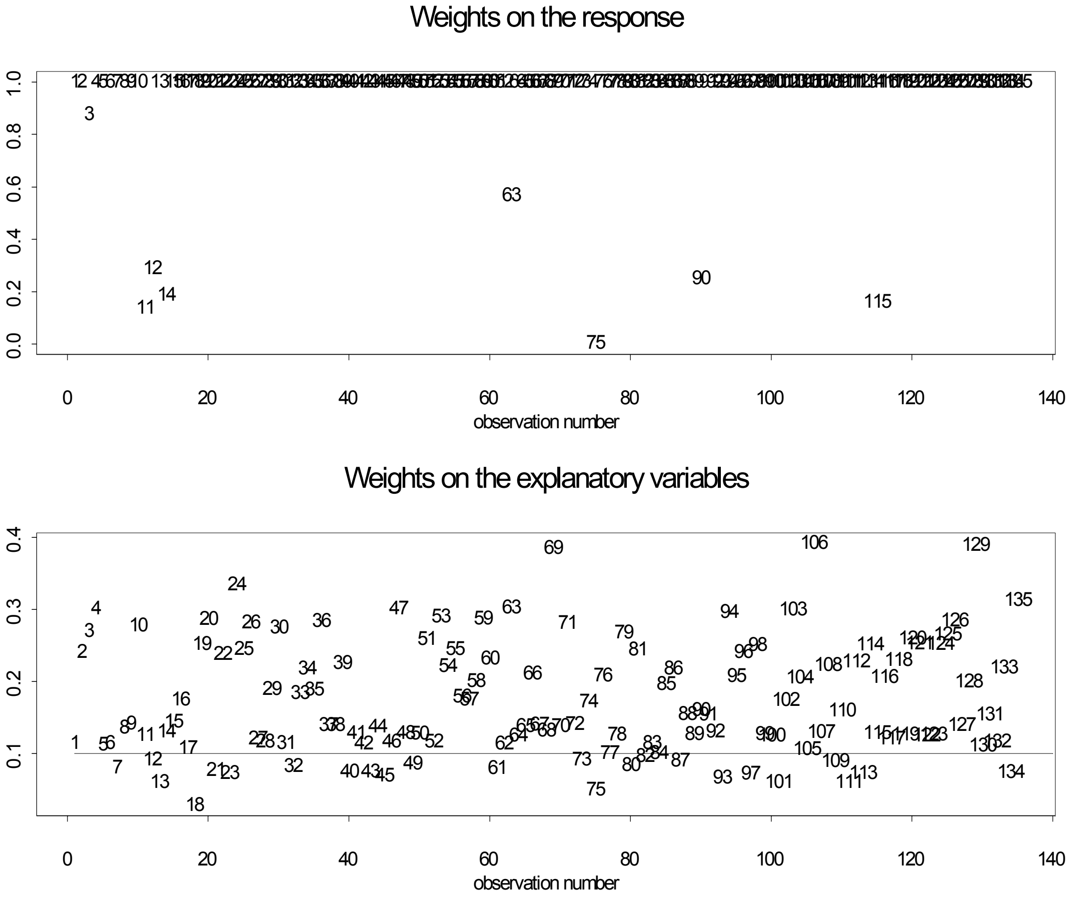

Figure 13: **Weights of Mallows-type estimator for the breastfeeding data**

	Int	X_1	X_2	X_3	X_4	X_5	X_6	X_7	$X_{8/1}$	$X_{8/2}$
MLE	-4.12	-0.98	0.31	**1.50**	1.08	0.027	0.17	**-1.96**	**1.57**	**3.31**
(SE)	2.39	0.58	0.59	0.59	0.70	0.05	0.13	0.76	0.59	1.01
Huber	**-7.12**	-0.90	0.51	**1.45**	0.85	0.03	**0.38**	**-2.64**	**1.91**	**3.51**
(SE)	3.24	0.68	0.69	0.68	0.80	0.06	0.18	1.05	0.67	1.11
OBRE	-5.06	-0.80	0.31	1.35	0.91	0.023	0.25	-2.21	**1.71**	**3.22**
(SE)	6.93	0.89	0.88	0.85	1.00	0.074	0.41	1.37	0.80	1.40
Mallows	**-14.31**	-0.68	0.85	1.51	0.66	0.04	**0.83**	**-3.09**	**1.85**	**3.93**
(SE)	6.19	0.82	0.82	0.81	0.91	0.07	0.37	1.46	0.77	1.43

Table 1: **Classical and robust estimates for the breastfeeding data**

	Classical test (p-value)	Robust test (p-value)
$H_0 : \beta_2 = \beta_3 = 0$	0.017	0.094
$H_0 : \beta_4 = \beta_6 = 0$	0.086	0.061
$H_0 : \beta_{8/1} = \beta_{8/2} = 0$	0.0002	0.0004

Table 2: **Classical and robust scores test for the breastfeeding data**