PSYCHOMETRIKA—VOL. 67, NO. 1, 21–32
MARCH 2002

ROBUST INFERENCE WITH BINARY DATA

MARIA-PIA VICTORIA-FESER

FACULTY OF PSYCHOLOGY AND EDUCATIONAL SCIENCES
UNIVERSITY OF GENEVA

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. 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's 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. They are finally applied to the analysis of binary data from a study on breastfeeding.

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

1. Introduction

The focus in this paper is on a robust approach to logistic regression. One issue addressed is the influence of data misclassification (e.g., a yes mistaken from a no) and/or of a singular subject on the value of the parameter's estimates as well as on the results of significance tests. A general theory of robustness is developed in Huber (1981) and Hampel, Ronchetti, Rousseeuw, and Stahel (1986) and 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. In the case of logistic regression several authors (see Carroll & Pederson, 1993; Copas, 1988; Kuensch, Stefanski, & Carroll, 1989; Markatou, Basu, & Lindsay, 1997; Pregibon, 1982) have made different proposals. In this paper, a general framework for robust inference is stated and then robust estimators and testing procedures are proposed and compared to previous results.

The paper is organized as follows. In section 2, the theoretical framework is set in which first the logistic model and its MLE is presented, then a general framework for robust estimation is given and links are made with other results for the logistic model, and 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 to real data from a study on breastfeeding.

2. Theoretical Results for Robust Inference

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 with parameter $\mu = \mathrm{E}[Y] = P(Y = 1)$ (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

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.

Requests for reprints should be sent to Maria-Pia Victoria-Feser, HEC, University of Geneva, 40, bd du Pont d' Arve, CH-1211 Geneva 4, SWITZERLAND.

0033-3123/2002-1/1999-0706-A $00.75/0

© 2002 The Psychometric Society

expectation of Y by means of a link function $g(\mu) = \mathbf{X}\beta$ such that $g^{-1}(\mathbf{X}\beta)$ takes 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 & Nelder, 1989), that is,

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

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

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

How is the MLE of β influenced by model misspecification? It can be shown analytically (see Victoria-Feser, 2000) that the MLE can be influenced by extreme values in the design space. The case of misclassification errors on the other hand has been studied by, for example, Copas (1988) and Pregibon (1982). Their results and mine (see below) show that misclassification errors can also lead to a biased MLE.

2.2. *A General Framework for Robust Estimation*

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 general formulation for a consistent robust estimator for the logistic regression model is given by the solution in β of

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

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). Equation (1) can be compared to Equation 2.1 of Carroll and Pederson (1993). If $w_i = 1$ and $b(\mathbf{x}_i, \beta) = 0\ \forall i$, then (1) yields the MLE. Equation (1) actually 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

$$V(\beta) = M^{-1}Q(M')^{-1} \tag{2}$$

where

$$M = \frac{1}{n}\sum \mu_i(1 - \mu_i)\mathbf{x}_i'\mathbf{x}_i\left(\mu_i w_i|_{y=0} + (1 - \mu_i)w_i|_{y=1}\right) \tag{3}$$

and

$$Q = \frac{1}{n}\sum \mu_i(1 - \mu_i)\mathbf{x}_i'\mathbf{x}_i\left(\mu_i w_i|_{y=0} + (1 - \mu_i)w_i|_{y=1}\right)^2. \tag{4}$$

We consider here consistency as defined in Kuensch et al. (1989), namely conditional Fisher consistency. This means that given $\mathbf{x}_i$, $E[(y - \mu(\mathbf{x}_i\beta))w_i\mathbf{x}_i'] - b(\mathbf{x}_i, \beta) = 0, \forall i$ so that $b(\mathbf{x}_i, \beta) = E[(y - \mu(\mathbf{x}_i\beta))w_i\mathbf{x}_i']$. 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 standardized residuals or Pearson residuals $\frac{y_i-\mu_i}{[\mu_i(1-\mu_i)]^{1/2}}$ (see McCullagh & Nelder, 1989) as proposed in Cantoni (1999) for generalized linear models; namely,

$$w_i = w_{y_i} = \min\left\{1; c\left|\frac{y_i - \mu_i}{[\mu_i(1-\mu_i)]^{1/2}}\right|^{-1}\right\} \tag{5}$$

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

$$w_{y_i}^0 = \min\left\{1; c\left|\frac{\mu_i}{[\mu_i(1-\mu_i)]^{1/2}}\right|^{-1}\right\} \quad \text{and} \quad w_{y_i}^1 = \min\left\{1; c\left|\frac{1-\mu_i}{[\mu_i(1-\mu_i)]^{1/2}}\right|^{-1}\right\}.$$

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 a weight function of the type $w_i = w_{y_i} \cdot w_{x_i}$ which separates the weights on extreme residuals (w_{y_i}) for the misclassification errors and the weights on extreme data in the design space (w_{x_i}). One possibility would be to base w_{x_i} on the diagonal elements of the hat matrix $H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ (see Victoria-Feser, 2000). I prefer, however, to choose a weighting function for the $\mathbf{x}$'s based on the influence function (*IF*) (see Hampel et al., 1986). The *IF* carries most of the information about the robustness properties of an estimator (or a test statistic) since it measures a first-order approximation of its (asymptotic) bias due to an infinitesimal deviation (of any type) from the assumed model. (For an illustration see Hampel et al., 1986, Figure 1, p. 42). A controlled bound on the *IF* therefore ensures robustness of the resulting estimator.

Equation (1) can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\left[w_{y_i}(y_i - \mu_i) - a_i\right]\mathbf{x}_i' w_{x_i} = 0 \tag{6}$$

where the constants $a_i = E[w_{y_i}(y-\mu_i)] = \mu_i(1-\mu_i)(w_{y_i}^1 - w_{y_i}^0)$ ensure conditional Fisher consistency. The *IF* is equal to $M^{-1}[w_y(y-\mu) - a]\mathbf{x}' w_x$, M given in (3) with $w_i|_{y=0} = w_{x_i} \cdot w_{y_i}^0$ and $w_i|_{y=1} = w_{x_i} \cdot w_{y_i}^1$. Actually there exists several ways to bound the *IF* (see Hampel et al., 1986), one of them being the standardized version which leads in our case to the condition $|w_y(y-\mu) - a|w_{x_i}[\mathbf{x}Q^{-1}\mathbf{x}']^{1/2} \leq c$, with Q given in (4). We therefore propose here the following weighting system: Huber weights on the response's standardized residuals given by (5) and Huber weights on $[\mathbf{x}Q^{-1}\mathbf{x}']^{1/2}$, that is,

$$w_{x_i} = \min\left\{1; \frac{c_x}{\left[\mathbf{x}_i Q^{-1}\mathbf{x}_i'\right]^{1/2}}\right\}. \tag{7}$$

A referee noted that the weights w_{x_i} are a variation of Mahalanobis distances with the exception that the matrix Q is not the covariance matrix of $\mathbf{x}$. The estimator we propose belongs to the so-called Mallows class of estimators (Mallows, 1975). To compute it one needs an iterative algorithm, whereby given a current value for the estimates one computes the weights and then a Newton-Raphson step for (6)[1]. Alternatively, by using a scoring method, these estimators can be seen as reweighted least squares estimators defined by (see Victoria-Feser, 2000)

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

[1] In order to simplify the estimation, the weights in (4) are taken to be only the weights on the response, that is, $w_i|_{y=0} = w_{yi}^0$ and $w_i|_{y=1} = w_{yi}^1$.

where $\mathbf{z} = \mathrm{vec}(\mathbf{x}_i\beta^{(k)} + v_i)$, $v_i = \{w_{y_i}(y_i - \mu_i) - a_i\}/[\mu_i w^0_{y_i} + (1 - \mu_i)w^1_{y_i}]\mu_i(1 - \mu_i)$ and $\mathbf{W} = \mathrm{diag}([\mu_i w^0_{y_i} + (1 - \mu_i)w^1_{y_i}]\mu_i(1 - \mu_i)w_{x_i})$.

We note here that Markatou, Basu, and Lindsay (1997) proposed also weighted likelihood estimating equations for the logistic regression model. It is however not applicable in the setting we consider here but in settings where the responses are made of the results from several trials (like the number of dead eggs in vials as considered in Markatou et al., 1997). It may be possible to extend the weighted likelihood methodology in this context. However, we do not pursue this goal here. 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

$$w_i = \min\left\{1; \frac{c}{|y_i - \mu_i - d_i|\left[\mathbf{x}_i\mathbf{A}^{-1}\mathbf{x}'_i\right]^{1/2}}\right\} \tag{8}$$

where $\mathbf{A}$ is defined implicitly by $Q = (\mathbf{A}^{-1})(\mathbf{A}^{-1})'$, c is a tuning constant and d_i is given in Kuensch et al. (1989). 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* with a bound controlled by a unique tuning constant c, but also 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 (1) 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 (nonconsistent) 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 & Ronchetti, 1994). In order to test hypotheses on linear combinations of regressors we choose Rao's score test which has a robust analogue. 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 $\beta_{(2)} = 0$. Victoria-Feser (2000) shows in the particular case of logistic regression by using the results of Heritier and Ronchetti (1994) that the level of Rao's score test can become arbitrarily biased either because of misclassification in the response 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.

Using the results of Heritier and Ronchetti (1994), Victoria-Feser (2000) shows that a suitable robust version of Rao's test statistic for the logistic regression model can be based on (6) and is given by

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

where $Z_M = \frac{1}{n}\sum_{i=1}^{n}[w_{y_i}(y_i - \mu(\mathbf{x}_i\hat{\beta})) - a_i]\mathbf{x}'_{(2)i}w_{x_i}$, with the weights given in (5) and (7), $\hat{\beta}_{(2)} = 0$ and $\hat{\beta}_{(1)}$ defined implicitly by $\frac{1}{n}\sum_{i=1}^{n}[w_{y_i}(y_i - \mu(\mathbf{x}_i\hat{\beta})) - a_i]\ \mathbf{x}'_{(1)i}w_{x_i} = 0$. The standardization matrix is $C = M_{(22.1)}V_{(22)}M'_{(22.1)}$ where $M_{(22.1)} = M_{(22)} - M_{(21)}M^{-1}_{(11)}M_{(12)}$ and $V_{(22)}$ are obtained by computing respectively (2) and (3) at $\hat{\beta}_{(2)} = 0$ and $\hat{\beta}_{(1)}$. nR^2_M is then compared to a χ^2_q with $q = \dim(\beta_{(2)})$. 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, taking the OBRE as a benchmark, we performed a simulation study. The design matrix consists 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. We also considered two arbitrary parameter sets, $\beta = (2, 3, 1)$ and $(-2, 1, 3)$. The values of the mean and standard deviation of the resulting true means $\mu_i = x_i'\beta$ are respectively $(0.72, 0.35)$ 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

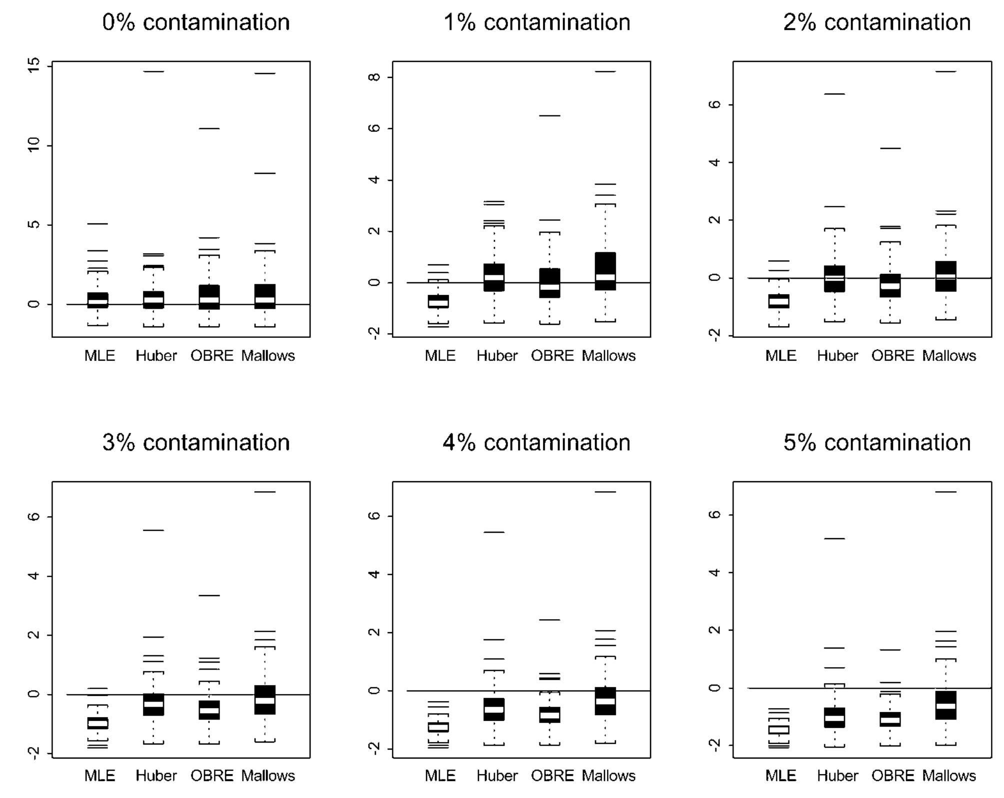

FIGURE 1.
β_2 estimators' bias distribution with missclassification errors ($n = 100$, $\beta = (-2, 1, 3)$).

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 also considered simultaneously.

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 estimators achieves the same degree of efficiency at the model compared to the MLE. 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. On how to compute the tuning constants and their values for the simulation exercise, see Victoria-Feser (2000). The latter also noticed that the parameters's value does not seem to change the efficiency ratio, whereas the sample size and the design matrix do.

3.1. 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 we found that 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%. This behavior is however not always observed. With $n = 50$ and/or with $\beta = (-2, 1, 3)$, the MLE becomes biased with only 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 1 that for β_2, the Mallows estimator performs well even with 4% of contaminated data. The graphics actually represent the distribution of the bias, namely the estimates minus the value of the corresponding

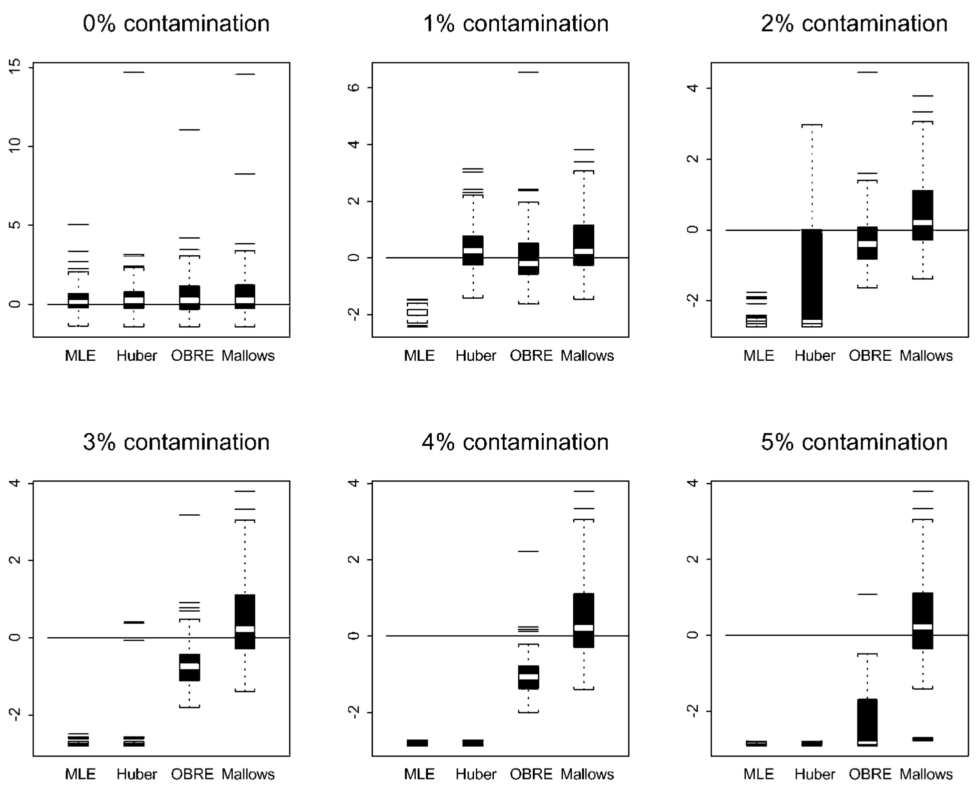

FIGURE 2.
β_2 estimators' bias distribution with leverages in x_2 ($n = 100$, $\beta = (-2, 1, 3)$)

parameter. All simulations results cannot be presented here but the following conclusions can be drawn: with misclassification errors, the MLE can become biased with just 1% misclassified response, whereas the robust estimators can withstand at least 3% of contaminated data.

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, the Mallows' estimator is very resistant with contaminated data at least up to 3%. For example, in Figure 2 are presented the β_2 estimates distributions for $n = 100$. Only the Mallows estimator remains unbiased. Moreover, a bias can also be present for the other parameters as presented in Figure 3 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 do not 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 $\beta = (-2, 1, 3)$. 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 the 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 the Mallows estimator having the best performance overall. For example, in Figure 4 are presented the distributions of β_2 estimates and one can see that only Mallows's estimator is

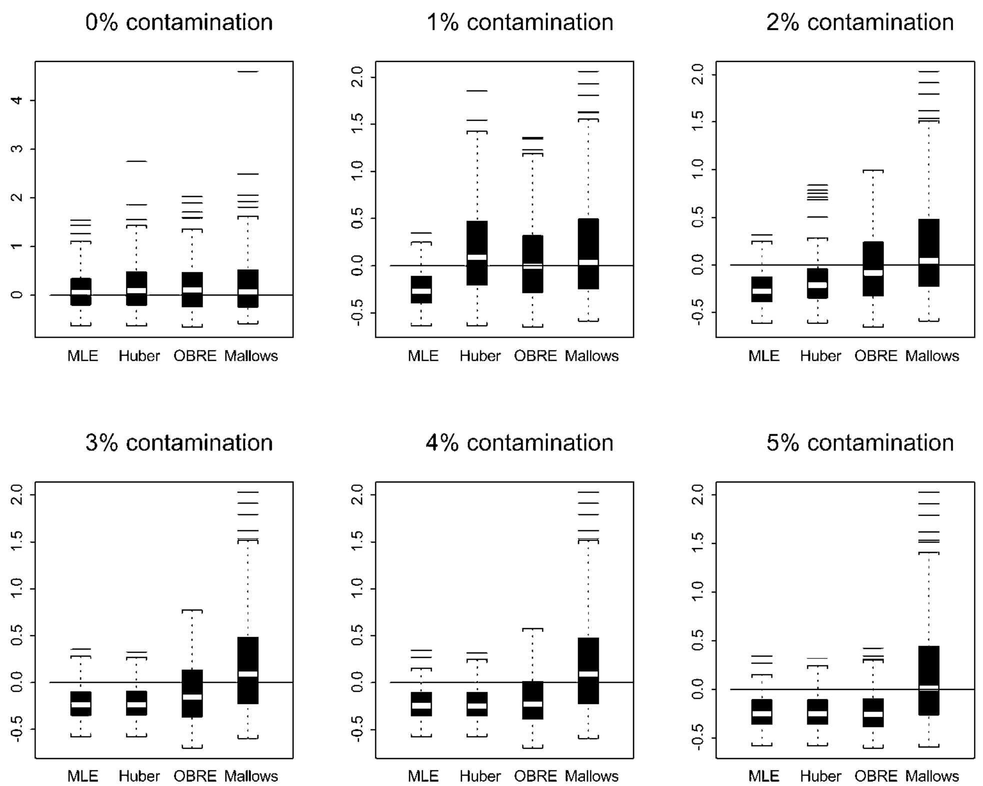

FIGURE 3.
β_1 estimators' bias distribution with leverages in x_2 ($n = 100$, $\beta = (-2, 1, 3)$).

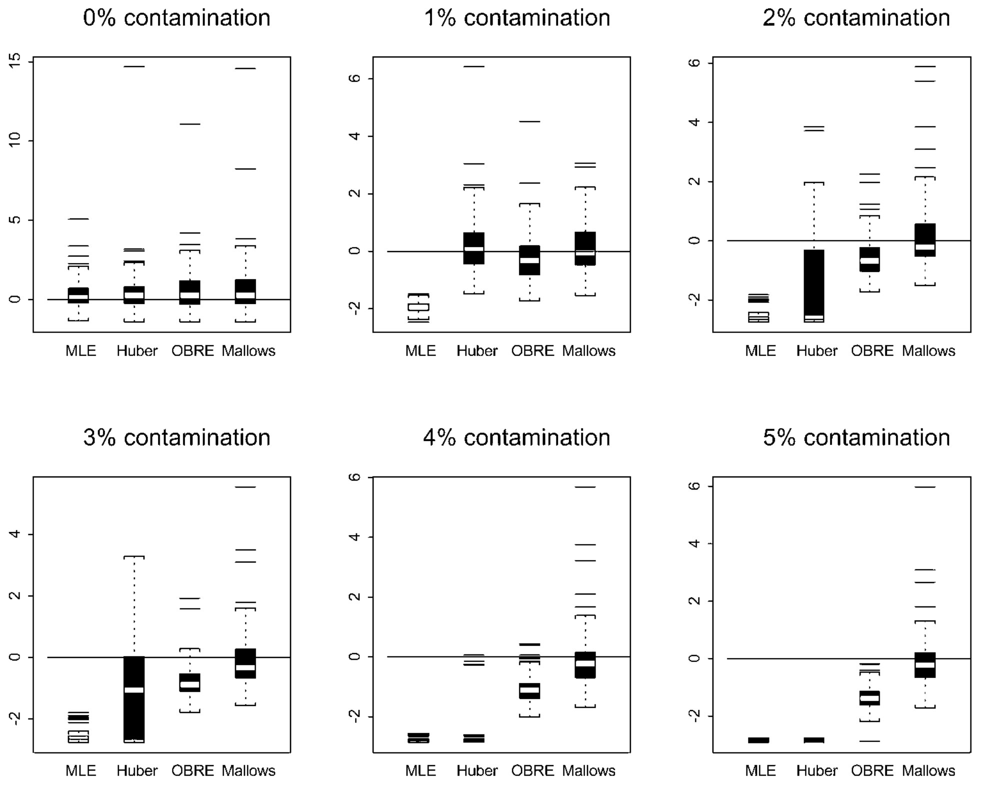

FIGURE 4.
β_2 estimators' bias distribution with missclassification errors and leverages in x_1 and/or x_2 ($n = 100$, $\beta = (-2, 1, 3)$).

resistant to up to 5% of both types of errors. We found that the smallest amount of contamination Mallows's estimator can withstand is of 3% of misclassification errors and 3% of leverages, that is, 6% of contaminated data all together.

It should be stressed that when analyzing 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 the Mallows estimator which has overall the best performance.

3.2. *Estimating Standard Errors*

Standard errors are important for judging the significance of the parameters through a t-test. This subsection considers the effect of contaminated data on the estimated standard errors. In the same simulation study, we also computed standard errors for the estimators using the diagonal elements of (2) with parameters replaced by their estimates. 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 (8) 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 5. The triangle

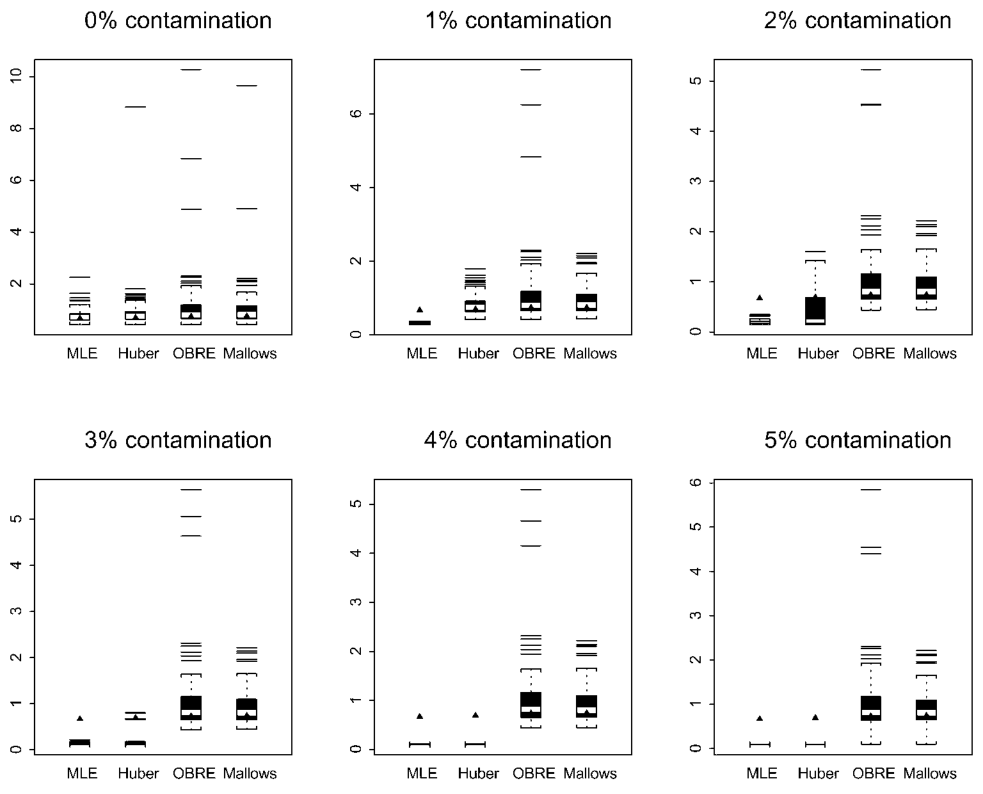

FIGURE 5.
Estimated standard errors distributions for β_2 with leverages in x_2 ($n = 100$, $\beta = (-2, 1, 3)$).

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.

3.3. *Computational Time*

Finally, a referee asked that the computational times of the different estimators be compared. As expected, the computational times (as measured 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 6 the different computational times for $n = 100$ and $\beta = (-2, 1, 3)$. It is clear that time is gained by using the 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 U.K. hospital on the decision of pregnant women to breastfeed their babies or not. 135 expecting mothers were asked what kind of

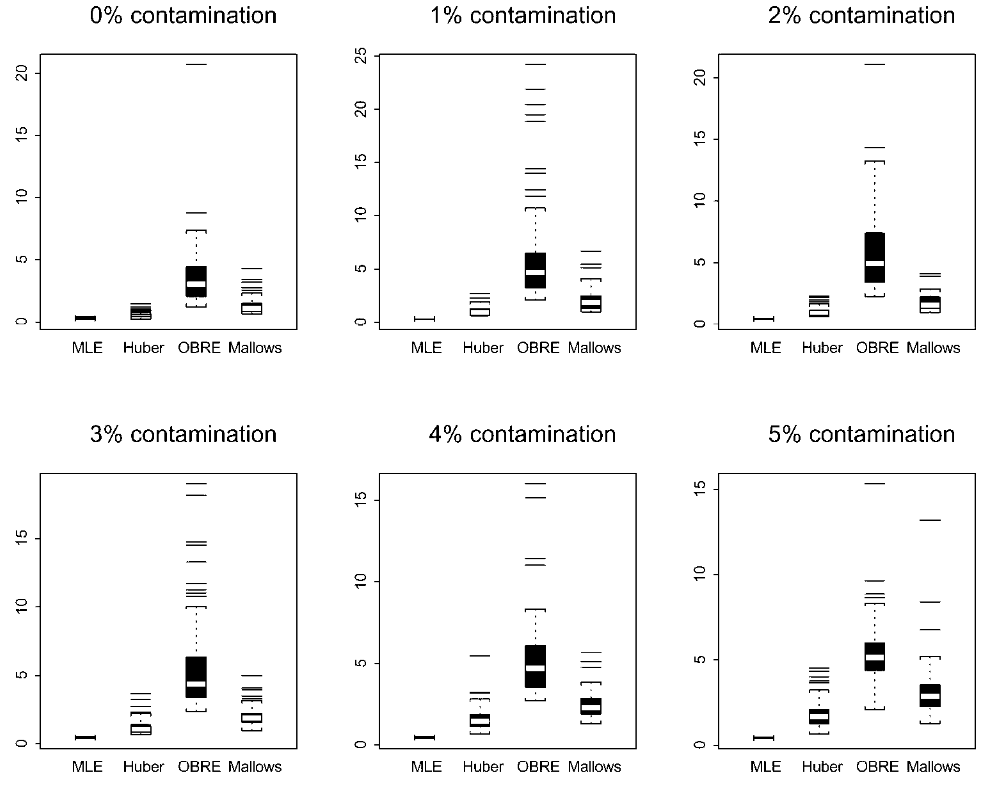

FIGURE 6.
Computational times for the different estimators with misclassification errors ($n = 100$, $\beta = (-2, 1, 3)$).

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 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 among 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. Robust Estimation

For the robust estimators, we chose their tuning constant so that they all achieve 85% efficiency compared to the MLE. The different estimates and their standard errors are given in Table 1[2]. On the whole, the estimates are of a similar order across methods. However, 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 non 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

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

TABLE 1.
Classical and robust estimates 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

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.

These estimated differences mean 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 my opinion no definite answer, but by construction of the estimators and from the simulations results, it is my 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 non 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 expecting mother.

4.2. Testing

With this data set, a few hypotheses 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

TABLE 2.
Classical and robust scores test 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

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 fail to reject the null hypothesis at the 5% level, but the robust one is nearly significant. 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 analyzing 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.

References

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

Cantoni, E. (1999). Resistant Techniques for Non Parametric Regression, Generalized Linear and Additive Models. Unpublished doctoral dissertation, University of Geneva, Switzerland.

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

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

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., & Stahel, W.A. (1986). Robust statistics: The approach based on influence functions. New York, NY: John Wiley.

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

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

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

Kuensch, H.R., Stefanski, L.A., & 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.

Mallows, C.L. (1975). *On some topics in robustness* (Tech. Rep.). Murray Hill, NJ: Bell Telephone Laboratories.

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

McCullagh, P., & Nelder, J.A. (1989). *Generalized linear models* (2nd ed.). London, U.K.: Chapman and Hall.

Moustaki, I., Victoria-Feser, M.-P, & Hyams, H. (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* (Statistics Research Report LSERR44). London, U.K.: London School of Economics.

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

Victoria-Feser, M.-P. (2000). *Robust logistic regression for binomial responses* (Tech. Rep. 2000). Geneva, Switzerland: University of Geneva.

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

Manuscript received 25 JAN 1999
Final version received 18 MAY 2001