Detecting outliers using weights in logistic regression*

JH Venter & T de la Rey†

1 Introduction

Logistic regression (LR) is concerned with explaining the probability of a specific response in terms of a number of regressors using a sample of relevant data. Pregibon (1981) states that the estimated LR relationship may be severely affected by outliers; this motivates the need for robust logistic regression procedures. Studies in this direction have been reported by Pregibon (1981), Copas (1988), Rousseeuw & Christmann (2003), Huber (1973), Rousseeuw & Leroy (1987) and Yohai (1987). Trimming is a broad approach towards robustifying of statistical procedures. It allows one to identify outliers and remove them from the data used in the estimation process. Trimming has been developed extensively by a number of authors in least squares regression, multivariate analysis and other fields (see for example Rousseeuw (1984), Rousseeuw & Van Driessen (1999a,b), where further references can be found). At first thought it seems attractive to use trimming also in LR to identify outliers and to limit their effects. When trimming, a subset of the data that is highly likely to be free from outliers is needed and a method is required to select such a subset. One possibility is to use maximum likelihood considerations, but this approach tends to run into the separation problem. The problem is that those observations that are considered as outliers are usually the same observations that will provide some overlap in the data. Therefore, trimming these observations removes the overlap and may lead to non-existence (indeterminacy) of the maximum likelihood estimator (MLE) applied to the remaining data as pointed out by Christmann and Rousseeuw (2001). They produced methodology to measure this overlap, enabling the user to judge the closeness to indeterminacy. In a further contribution Rousseeuw and Christmann (2003) overcame the non-existence problem by introducing the hidden logistic regression model with an associated estimator referred to as the maximum estimated likelihood (MEL) estimator which always exists even when there is no overlap in the data. They also proposed a robustified form of the MEL estimator, called the weighted maximum estimated likelihood (WEMEL) estimator. WEMEL does not trim, but downweights leverage points, where the choice of leverage points is based on the robust distances in the regressor space. Using a simulation study they show that WEMEL performs very well as a robust procedure compared to its competitors. However, WEMEL does not take outliers in the response direction into account and is not really an outlier detection procedure in the sense that it produces a subset of the observations that may be labelled as outliers. In this paper we use a different form of downweighting to introduce a procedure that may be thought of as both a robust LR estimation procedure and an outlier detection method.

Our procedure may be described as a method that "Detects Outliers Using Weights" and is referred to below as the DOUW method. DOUW begins by selecting two sets of weights, namely high and low weights and then splits the data optimally into two subsets to which the high and the low weights are attached, the subset with the high weights including the observations that are more likely not to be outliers. A corresponding weighted ML estimator of the regression coefficients is computed. This is used to estimate the response probabilities of the individual observations. Observations with success response but low estimated response probability and observations with failure response but high estimated response probability can then be

*This is a shortened version of the original article. Full version of this article can be emailed on request. Contact Tanja.delaRey@nwu.ac.za

†JH Venter and T de la Rey are at the Centre for Business Mathematics and Informatics (BMI), North-West University, Potchefstroom, South Africa. Correspondence email address Tanja.delaRey@nwu.ac.za. The research was supported by the National Research Foundation (NRF and THRIP) and by industry (ABSA and SAS Institute). This work was completed as part of the second author's Ph.D at the North-West University. The original version of this article was accepted for publication and should appear in SASJ 2007.

classified as outliers. A final weighted MLE can then be computed by redistributing the high and low weights according to the outlyingness of the relevant observations. The method depends on the specification of some items, such as the size of the initial high weight subset and the levels of the high and low weights as well as the threshold according to which outlyingness is decided. We study the effects of the choices of these items below and also compare DOUW with ML, MEL and WEMEL.

The layout of this paper is as follows. Section 2 introduces the notation and terminology used here and reviews briefly some notions regarding outliers relevant to LR. Section 3 formulates the basic DOUW procedure and lists a number of more elaborate versions that can also be used. Section 4 reports the result of a simulation study that evaluates the cost/benefit balance that has to be taken into account when specifying the tuning parameters of the procedure. Section 5 discusses the application of the DOUW procedure to a number of standard datasets in the literature as well as a new large dataset relating to success probabilities in sales promotion campaigns. Section 6 concludes while technical details regarding the C-step algorithm used here are provided in the appendix (appendix not shown in this shortened version).

2 Notation and terminology

In linear logistic regression we have a dichotomous response variable Y that can take the values 1 ("success") or 0 ("failure"), and we have K regressors $x_1, ..., x_K$. Let $\mathbf{x}^T = (1, x_1, ..., x_K)$ with T denoting transpose. We fit the LR model

$$P(Y = 1) = p(\mathbf{x}, \boldsymbol{\beta}) = 1/\left(1 + \exp\left(-\boldsymbol{\beta}^T \mathbf{x}\right)\right) \tag{2.1}$$

where $\boldsymbol{\beta}^T = (\beta_0, \beta_1, ..., \beta_K)$ is the vector of LR coefficients (see Hosmer & Lemeshow, 1989 and Kleinbaum, 1994). Assume that we have N observations, where the n^{th} observation is $(y_n, \mathbf{x}_n^T)$, with y_n the observed value of Y and $\mathbf{x}_n^T = (1, x_{n,1}, ..., x_{n,K})$ the vector of observed values of the K regressors. $l(\mathbf{x}) = \boldsymbol{\beta}^T \mathbf{x}$ is often referred to as the logit value of $\mathbf{x}$ and $p(\mathbf{x}, \boldsymbol{\beta})$ considered as a function of $l(\mathbf{x})$ as the success probability function or curve of the model we are fitting to the data.

Under the independence assumption the log likelihood of the N observations is given by

$$\sum_{n=1}^{N} D_n(\boldsymbol{\beta}) \text{ with } D_n(\boldsymbol{\beta}) = y_n \log p(\mathbf{x}_n, \boldsymbol{\beta}) + (1 - y_n) \log(1 - p(\mathbf{x}_n, \boldsymbol{\beta})) \tag{2.2}$$

and the MLE of $\boldsymbol{\beta}$ is obtained by maximizing this expression over $\boldsymbol{\beta}$**.** For datasets in which there is no overlap between the 0 and 1 responses (i.e. the $\mathbf{x}_n$'s corresponding to $y_n = 0$ can be separated by a hyperplane from the $\mathbf{x}_n$'s corresponding to $y_n = 1$), the MLE does not exist. Rousseeuw and Christmann (2003) introduced the MEL estimator to overcome this difficulty. MEL may be summarized as follows. Let $\delta = 0.01$, $\overline{\pi} = \frac{1}{N}\sum_{n=1}^{N} y_n$, $\widehat{\pi} = \max(\delta, \min(1 - \delta, \overline{\pi}))$, $\delta_0 = \widehat{\pi}\delta/(1 + \delta)$, $\delta_1 = (1 + \widehat{\pi}\delta)/(1 + \delta)$, and transform the y_n's to $\widetilde{y}_n = (1 - y_n)\delta_0 + y_n\delta_1$. Then MEL chooses $\boldsymbol{\beta}$ to maximize the "estimated" log likelihood

$$\sum_{n=1}^{N} \widetilde{D}_n(\boldsymbol{\beta}) \text{ with } \widetilde{D}_n(\boldsymbol{\beta}) = \widetilde{y}_n \log p(\mathbf{x}_n, \boldsymbol{\beta}) + (1 - \widetilde{y}_n) \log(1 - p(\mathbf{x}_n, \boldsymbol{\beta})) \tag{2.3}$$

Unlike the classical MLE, the MEL estimator always exists. The related robust estimator WEMEL is defined as the maximizer over $\boldsymbol{\beta}$ of the weighted estimated log likelihood $\sum_{n=1}^{N} w_n \widetilde{D}_n(\boldsymbol{\beta})$. Here the weights are determined by the position of the data in x-space according to $w_n = M/\max\left\{RD^2(\mathbf{x}_n^*), M\right\}$, where $\mathbf{x}_n^* = (x_{n,1}, ..., x_{n,K})^T$ and $RD^2(\mathbf{x}_n^*)$ is the robust distance (RD) and M is the 75^{th} percentile of all $RD^2(\mathbf{x}_n^*)$, $n = 1, ..., N$. For more details on the robust distances see Rousseeuw and Christmann (2003) who also provides other properties and results on the performance of these estimators.

As we have mentioned in the introduction, outliers may severely affect the fitted model (2.1). This motivates the need for robust LR procedures (of which WEMEL is an example). One can distinguish between outliers in the x-space and in the y-direction. Many methods have been developed to deal with outliers in the x-space. Perhaps the most prominent of these is

the fast minimum covariance determinant (FAST-MCD) methodology due to Rousseeuw & Van Driessen (1999b) which is used in WEMEL. In this paper our emphasis is more on outliers in the y-direction. Figure 2.1 illustrates the situation in the case of two regressors.

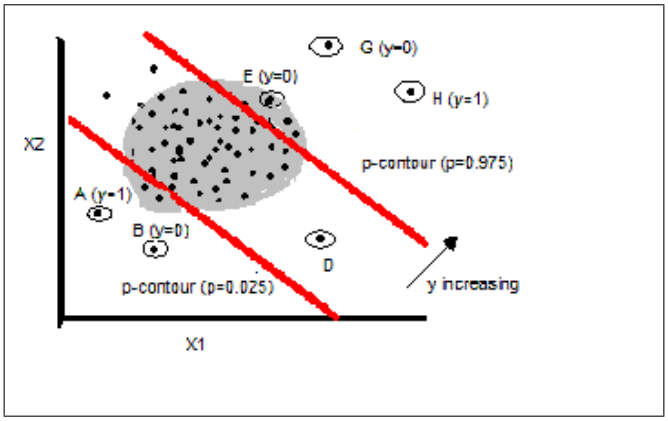

Figure 2.1: x- and y-outliers

The grey area which includes most of the (x_1, x_2) pairs may be thought of as the x-inliers while those outside this area are x-outliers, for example A, B, D, G and H. For illustrative purposes we have also added two contours $p(\mathbf{x}, \boldsymbol{\beta}) = d$ and $p(\mathbf{x}, \boldsymbol{\beta}) = 1 - d$ (with d small). Observations to the outside of these contours with inappropriate y-values may be thought of as possible y-outliers, especially those that are far from the contours, for example A and G. Copas (1988) calls an observation with $y = 1$ and p close to 0 an "uplier" (for example A) and an observation with $y = 0$ but p close to 1 a "downlier" (for example G and E). Upliers and downliers may also be thought of as bad leverage points, in the sense that they are likely to influence the estimated regression coefficients seriously in directions different from that implied by the other observations. By contrast observations such as those corresponding to B and H may be called good leverage points in the sense that although they are outlying in the x-space, there y-values are consistent with what is to be expected from the x-region in which they are.

3 Detecting outliers using weights

The methodology introduced here has some parallels with the least trimmed squares (LTS) methodology of Rousseeuw and Van Driessen (1999a,b) in ordinary regression; hence we briefly review LTS. LTS starts with a lower bound, g_1, on the number of good observations (inliers). We then look for a subset $\mathbf{G}$ of the observations with size $\#(\mathbf{G}) = g_1$, which has smallest residual sum of squares among all subsets of size g_1. The optimal $\mathbf{G_1}$ found in this way is thought to be most likely to be free of outliers and should therefore result in an estimate of the regression coefficients that is least influenced by possible outliers. The estimated regression coefficients obtained in this way are the LTS estimates based on g_1 observations. It may be that the choice of g_1 is conservative in the sense that there could be observations outside of $\mathbf{G}_1$ which are also good. One could use the LTS estimates to calculate residuals for the observations outside of $\mathbf{G}_1$ and use these residuals to decide which observations to add to $\mathbf{G}_1$ to obtain a larger subset $\mathbf{G}_2$ containing $g_2 \geq g_1$ good observations and then base the final trimmed regression estimate on $\mathbf{G_2}$. The LTS optimization is computationally difficult and if done exactly, requires a number of steps that grows combinatorially with the size of the number of observations and is practically infeasible especially in large problems. Rousseeuw and Van Driessen (1999a, b) handles the computational issue by starting with an initial random choice of $\mathbf{G}$, where $\#(\mathbf{G}) = g_1$, and then improves this choice of $\mathbf{G}$ iteratively until convergence, using a so-called C-step procedure. This is repeated many times and the best solution is kept and taken to represent the optimal $\mathbf{G}_1$ required. More detail can be found in Rousseeuw and Van Driessen (1999b).

The LTS procedure is quite attractive and successful in ordinary regression and our initial aim was to formulate an analogue for the LR case, using trimmed likelihood instead of trimmed

least squares. This would mean that we want to select a subset $\mathbf{G}$ containing g_1 observations (together with an associated $\boldsymbol{\beta}$) which maximizes $\sum_{n\in\mathbf{G}} D_n(\beta)$. Again this is computationally difficult but the equivalent to the C-step is already available from the work of Neykov & Muller (2002). However, when implementing the procedure we found that the "optimal" $\mathbf{G}_1$ usually tends to get bogged down among subsets with no overlap for which the corresponding MLE does not exist. This especially happens when one starts with a conservative g_1 which is much smaller than N. Replacing MLE with MEL avoids the non-existence issue but does not eliminate the possibility that the "optimal" $\mathbf{G}_1$ may be chosen poorly . To circumvent these problems we decided to follow a downweighting approach rather than a trimming approach. For this purpose let $0 < \epsilon < 1$ and for a given subset $\mathbf{G}$ define a corresponding weighted log likelihood by

$$l_\epsilon(\boldsymbol{\beta}, \mathbf{G}) = \sum_{n\in\mathbf{G}} D_n(\boldsymbol{\beta}) + \epsilon \sum_{n\notin\mathbf{G}} D_n(\boldsymbol{\beta}) \tag{3.1}$$

This expression is a weighted log likelihood function with $w_n = 1$ for $n \in \mathbf{G}$ and $w_n = \epsilon$ for $n \notin \mathbf{G}$. Thus the observations in $\mathbf{G}$ are associated with the higher weight 1 and the observations outside of $\mathbf{G}$ with the lower weight ϵ. Good observations in the sense of making a large contribution $D_n(\boldsymbol{\beta})$ to the log likelihood function should be in $\mathbf{G}$ and bad ones in the sense of making low contribution should be outside of $\mathbf{G}$ in order to make (3.1) large. As in LTS to capture the good observations in $\mathbf{G}$ we now look for that subset $\mathbf{G}$ with size $\#(\mathbf{G}) = g_1$ and associated $\boldsymbol{\beta}$ which maximize $l_\epsilon(\boldsymbol{\beta}, \mathbf{G})$ among all subsets of given size g_1. For any given set $\mathbf{G}$ it is relatively easy to calculate the corresponding optimizer, $\boldsymbol{\beta}^*(\mathbf{G})$ of $l_\epsilon(\boldsymbol{\beta}, \mathbf{G})$ over $\boldsymbol{\beta}$ using for example a Newton-Raphson method (see Hastie et al. (2001)). However, optimizing $l_\epsilon(\boldsymbol{\beta}, \mathbf{G})$ over both $\boldsymbol{\beta}$ and $\mathbf{G}$ is computationally difficult but it can be handled by a more general form of the C-step procedure of Neykov & Muller (2002), presented in the Appendix (not shown in this shortened version). Once this is done we have an optimal $\mathbf{G_1}$ with an associated estimator $\boldsymbol{\beta}^*(\mathbf{G}_1)$. Next we must set up a criterion which can be used to identify outliers. To do this, we use a small cut-off level c with $0 < c < 1$ and then declare observation n as an outlier if $y_n = 1$ but $p(\mathbf{x}_n, \boldsymbol{\beta}^*(\mathbf{G})) \leq c$ or if $y_n = 0$ but $p(\mathbf{x}_n, \boldsymbol{\beta}^*(\mathbf{G})) > 1 - c$. The reasoning here is that if $y_n = 1$, and $p(\mathbf{x}_n, \boldsymbol{\beta}^*(\mathbf{G}))$ is small it is probable that observation n is an uplier and therefore should be downweighted (given the weight ϵ). Similarly if $y_n = 0$, and $p(\mathbf{x}_n, \boldsymbol{\beta}^*(\mathbf{G})))$ is large it is probable that observation n is an downlier and should therefore be downweighted. The remaining observations are given high weights 1. We could now let $\mathbf{G_2}$ be the set of observations that are given the high weights in this classification step and then compute a final weighted estimate for the regression coefficients as $\boldsymbol{\beta}^*(\dot{\mathbf{G}_2})$ which by definition maximizes $l_\epsilon(\boldsymbol{\beta}, \mathbf{G_2})$. This is the DOUW procedure.

There are further issues that have to be dealt with to complete the specification of the DOUW method. Among these are the initial choice of g_1, the number of iterations in the C-steps, the choice of ϵ and the choice of cut-off c. The choices of ϵ and c will be dealt with after we have reported the results of a simulation study in Section 4. The choice of g_1 and the number of iterations are spelled out in the form of the following **pseudocode for DOUW**.

1. Select $g_1 = \max\{[(N + K + 1)/2], K + 1\}$ where $[x]$ is the integer part of x. This is in line with the suggestion of Rousseeuw & van Driessen (1999a) for the FAST-LTS method.
2. Repeat 50 times:
   - Select a starting subset $\mathbf{H} \subset \{1, ..., N\}$ at random with $\#(\mathbf{H}) = K + 1$.
   - Calculate $\boldsymbol{\beta}^*(\mathbf{H})$, the $D_n(\boldsymbol{\beta}^*(\mathbf{H}))$'s and find the π_i's so that $D_{\pi_1}(\boldsymbol{\beta}^*(\mathbf{H})) \geq ... \geq D_{\pi_N}(\boldsymbol{\beta}^*(\mathbf{H}))$. Put $\mathbf{G} = \{\pi_1, ..., \pi_{g_1}\}$.
   - Carry out 2 C-steps starting with this $\mathbf{G}$ and ending with $\mathbf{G}''$ say.
   - Store the 5 best results, in terms of the highest values of $l_\epsilon(\boldsymbol{\beta}^*(\mathbf{G}''), \mathbf{G}'')$.
   - For each of these 5 best results, repeat the C-step iteration until convergence.
   - Retain the overall best subset $\mathbf{G_1}$ which is the best of the converged 5 iterations.
3. Put $\mathbf{G_2} = \{n : (y_n = 1$ and $p(\mathbf{x}_n, \boldsymbol{\beta}^*(\mathbf{G_1})) \geq c\ )$ or $(y_n = 0$ and $p(\mathbf{x}_n, \boldsymbol{\beta}^*(\mathbf{G_1})) \leq 1 - c)\}$ and then calculate the final weighted estimate $\boldsymbol{\beta}^*(\dot{\mathbf{G}_2})$ which maximizes $l_\epsilon(\boldsymbol{\beta}, \mathbf{G_2})$. The outliers are the observations outside of $\mathbf{G}_2$.

This basic DOUW procedure can be varied in a number of ways and some remarks are in the full version of the article.

4 Simulation studies

4.1 Design

To study the properties of DOUW and to get some guidance on the effects of the choices of the tuning parameters, we report the results of simulation studies in this section. As Rousseeuw & Christmann (2003) we use $K = 2$ regressors x_1 and x_2 and we take $\boldsymbol{\beta}^{\mathbf{T}} = (1, 1, 2)$. We will study the following **cases**:

1. $N = 100$ with x_1 and x_2 independently $N(0, 1)$ distributed
2. $N = 100$ with x_1 and x_2 independently $N(0, 4)$ distributed
3. $N = 100$ with x_1 and x_2 independently t_3 distributed
4. $N = 20, 50, 100, 200$ with x_1 and x_2 independently $N(0, 1)$ distributed

These specifications give the true values of the model $p(\mathbf{x}, \boldsymbol{\beta})$ of (2.1) to be fitted to the data (the "**approximating family**" in the terms of Linhart & Zucchini (1986)). Of course data generated under this model may contain apparent up- and downliers since there is always positive probability to get $y_n = 1$ even though $p(\mathbf{x}, \boldsymbol{\beta})$ may be small and vice versa. One way to study the sensitivity of procedures to outliers is to manually add a number of additional up- and downliers into the data. Instead of following this approach we choose to select a data generating model $q(\mathbf{x})$ (the "**operating model**" in the terms of Linhart & Zucchini (1986)) with a flatter success probability curve than $p(\mathbf{x}, \boldsymbol{\beta})$. Data generated from such a model will then contain more up- and downliers than expected under $p(\mathbf{x}, \boldsymbol{\beta})$. An example of such a $q(\mathbf{x})$ is

$$q(\mathbf{x}) = q(\mathbf{x}, \boldsymbol{\beta}, \alpha) = \begin{cases} \alpha & \text{if } p(\mathbf{x}, \boldsymbol{\beta}) < \alpha \\ p(\mathbf{x}, \boldsymbol{\beta}) & \text{if } \alpha \leq p(\mathbf{x}, \boldsymbol{\beta}) \leq 1 - \alpha \\ 1 - \alpha & \text{if } p(\mathbf{x}, \boldsymbol{\beta}) > 1 - \alpha \end{cases} \tag{4.1}$$

which will be used extensively in this section. Here α is a parameter ranging between 0 and 1/2. Figure 4.1 panel (a) illustrates the corresponding success probability curves for the choice $\alpha = 0.2$. For $p(\mathbf{x}, \boldsymbol{\beta})$ between α and $1 - \alpha$ the two curves coincide. For $p(\mathbf{x}, \boldsymbol{\beta}) < \alpha$ the probability of getting $y = 1$ (an uplier) is larger under the data generating model $q(\mathbf{x})$ so that relatively more upliers will tend to be present in the actual data than expected under the fitting model $p(\mathbf{x}, \boldsymbol{\beta})$. For $p(\mathbf{x}, \boldsymbol{\beta}) > 1-\alpha$ the probability of getting $y = 0$ (a downlier) is larger under the data generating model $q(\mathbf{x})$ so that relatively more downliers will tend to be present in the actual data than expected under $p(\mathbf{x}, \boldsymbol{\beta})$. If $\alpha = 0$ the two models do not differ but as α increases the number of outliers produced under the data generating model increases. Many other examples of $q(\mathbf{x})$ are possible, for example the hidden logistic regression (HLR) model used in Copas (1988) and Rousseeuw & Christmann (2003). For our purposes we may take this HLR model to given by $q(\mathbf{x}, \boldsymbol{\beta}, \alpha) = \alpha + (1 - 2\alpha)p(\mathbf{x}, \boldsymbol{\beta})$ and illustrate it in panel (b) of Figure 4.1. Again the severity of the number of outliers increases as α increases. The results for both these models are largely similar and to save space we will report only in terms of the data generating model of (4.1).

4.2 Performance criteria

The first performance criterion to be used may be described as the average mean squared error of estimation of the true $\boldsymbol{\beta}$'s, i.e. if $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_o, \widehat{\beta}_1, ..., \widehat{\beta}_K)$ is the estimator of $\boldsymbol{\beta} = (\beta_o, \beta_1, ..., \beta_K)$ then $CB = (1/(K + 1))\, E \sum_{k=0}^{K} (\widehat{\beta}_k - \beta_k)^2$. It is often also important that the success probabilities be estimated accurately. A criterion to judge this accuracy may be described as the average mean weighted squared error of estimation of the success probabilities at the actual x-values i.e. if $\widehat{p}_n = p(\mathbf{x}_n, \widehat{\boldsymbol{\beta}})$ then $CWP = (1/N) E \sum_{n=1}^{N} \lambda_n \left(\widehat{p}_n - p(\mathbf{x}_n, \boldsymbol{\beta})\right)^2$, where λ_n is a weight associated with the n^{th} observation. Often the observations with small or large response probabilities are particularly important and we can choose the weights to emphasize this. One possibility is to take $\lambda_n = 1/p(\mathbf{x}_n, \boldsymbol{\beta})\left(1 - p(\mathbf{x}_n, \boldsymbol{\beta})\right)$, but then λ_n tends to infinity when $p(\mathbf{x}_n, \boldsymbol{\beta})$ tends to 0 or 1 causing these extreme cases to dominate the others. Therefore to lessen the influence of the extreme cases on the weights we take $\lambda_n = 1/\left(a + p(\mathbf{x}_n, \boldsymbol{\beta})\left(1 - p(\mathbf{x}_n, \boldsymbol{\beta})\right)\right)$. As long as a is

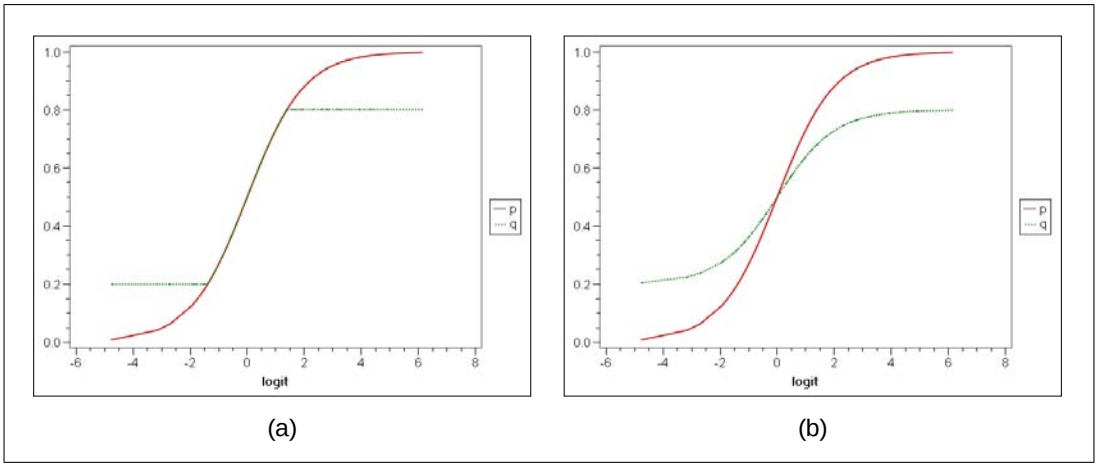

Figure 4.1: Examples of success probability curves of $p(\mathbf{x}, \boldsymbol{\beta})$ and $q(\mathbf{x}, \boldsymbol{\beta}, \alpha)$ with $\alpha = 0.2$. Panel (a) $q(\mathbf{x}, \boldsymbol{\beta}, \alpha)$ given by (4.1) and panel (b) $q(\mathbf{x}, \boldsymbol{\beta}, \alpha)$ of HLR

positive the weights no longer tends to infinity when $p(\mathbf{x}_n, \boldsymbol{\beta})$ tends to 0 or 1. The choice of a is somewhat arbitrary but does not seem to matter much for our purposes and we report mostly for the choice $a = 0.01$. We could also considered an unweighted form where we take $\lambda_n = 1$, so that the criterion becomes $CP = (1/N)E\sum_{n=1}^{N}\left(\widehat{p}_n - p(\mathbf{x}_n, \boldsymbol{\beta})\right)^2$. Many other criteria could be considered (for example judging bias and variance separately or using absolute deviation, etc.). There are however, other features that need to be varied (for example the sample size N, the number of regressors K, and the distribution of the $\mathbf{x}_n$'s as well as the DOUW tuning constants, etc.). It is not feasible to report on a large number of criteria in conjunction with all these other features; hence attention will be restricted to these criteria.

4.3 Choice of ϵ and c

We first study the effect of the choice of ϵ on the performance of DOUW. Provisionally we choose the cut-off $c = 0.05$ and will comment on its effect later on. We show that choosing ϵ very small causes the procedure to perform very poorly while choosing a moderate value for ϵ leads to good performance in the presence of outliers, at limited cost when no outliers are present.

To begin with consider **case 1**. The top two panels of Figure 4.2 show CB for the choices $\epsilon = \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5\}$ as functions of α with the data generating model (4.1) based on 1000 simulation runs. In the left hand panel it is evident that for $\epsilon = 0.01$, CB is uniformly very large. If ϵ is made smaller than 0.01 CB rapidly becomes even worse. To understand the reasons for this phenomenon we studied the best subset $\mathbf{G_1}$ for the very small choices of ϵ. It turns out that most often the observations in this subset can be separated. Strictly speaking in such cases the ML estimators of the $\boldsymbol{\beta}$'s do not exist; however the optimizer that we used in this study did converge but to values rather different from the true beta's and this discrepancy causes the large values for CB shown in Figure 4.2. Therefore ϵ should not be chosen very small. The right hand top panel of Figure 4.2 shows the same curves as the left hand panel with the curve corresponding to $\epsilon = 0.01$ omitted in order to increase the vertical scale and display more details of the remaining curves. For comparison purposes we also included the curves of MEL and WEMEL. Considering first the CB curve corresponding to $\epsilon = 0.1$, DOUW outperforms both MEL and WEMEL if there are a substantial number of outliers ($\alpha > 0.15$) but this advantage comes at a cost when there are few or no outliers in that CB for DOUW is larger than for CB for MEL and WEMEL when α is small. If we increase ϵ to 0.2 DOUW outperforms MEL and WEMEL over a larger range of α values but not to the same extent as for $\epsilon = 0.1$. At the same time a smaller cost when few outliers are present is involved for $\epsilon = 0.2$ as compared to $\epsilon = 0.1$. With further increases in ϵ the extent of the improvement when many outliers are present diminishes and the required cost when few outliers are present also diminishes so that the CB curve resembles those of MEL and WEMEL progressively more closely. Examining the

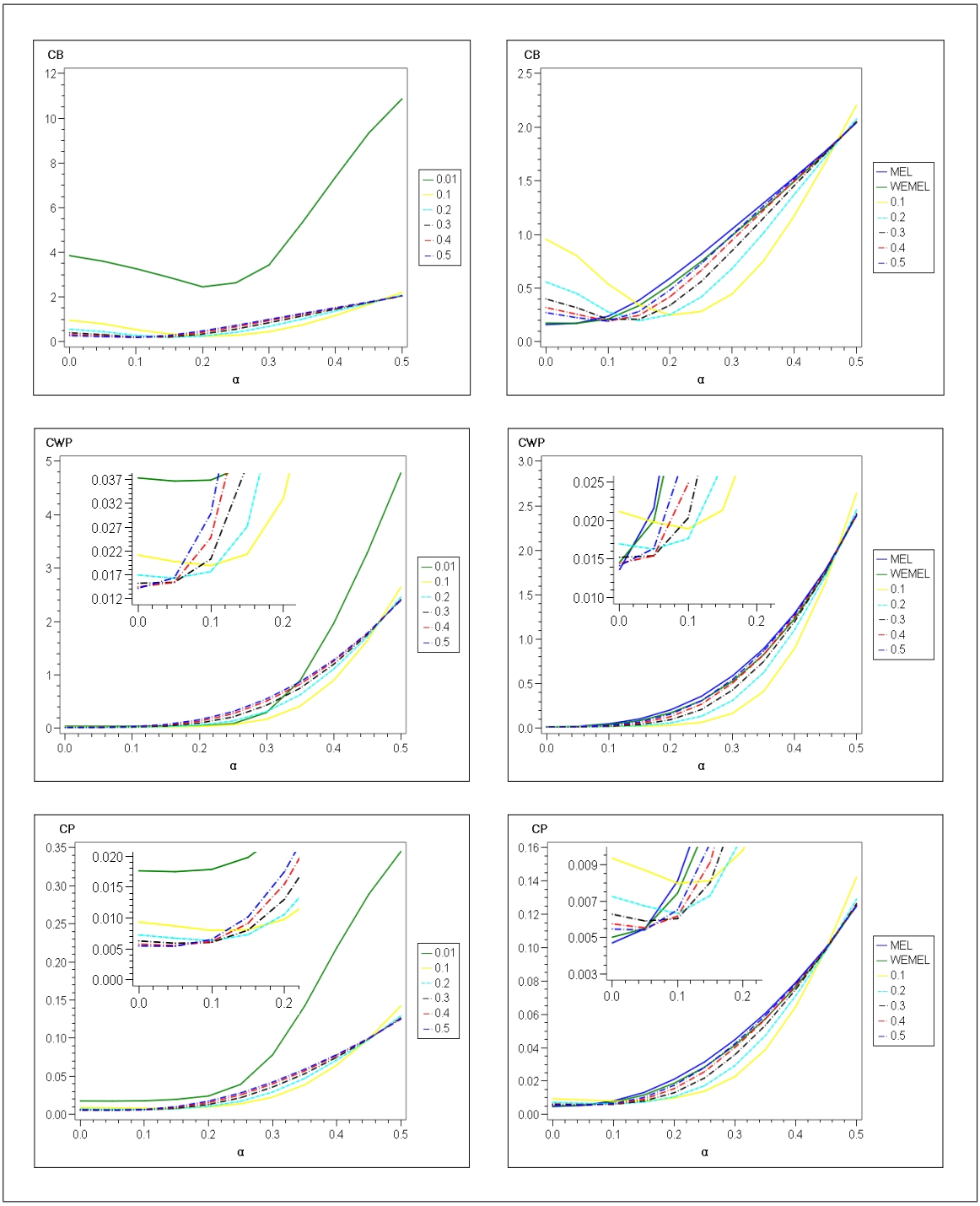

Figure 4.2: CB, CWP and CP values for $\epsilon = \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5\}$ and $c = 0.05$ for case 1

curves of $\epsilon = 0.4$ and $\epsilon = 0.5$ it would seem that choosing ϵ greater than 0.3 leaves little scope for different performance of DOUW as compared to MEL and WEMEL so that the interesting range for ϵ is from 0.1 to 0.3 and our impression at this time is that $\epsilon = 0.2$ is a reasonable compromise for the scenarios modeled.

The middle and the bottom four panels of Figure 4.2 show CWP and CP for the cases corresponding to the top panels. We have also inserted an enlargement on each of these four graphs to show details of the curves at small values of α. In the left of these four panels it is again evident that choosing $\epsilon = 0.01$ leads to very poor performance in terms of the CWP and CP criteria and we have confirmed that smaller choices of ϵ lead to even worse performance. The right hand middle and bottom panels show that DOUW with $\epsilon = 0.1$ improves substantially on MEL and WEMEL when many outliers are present (for $\alpha > 0.09$) but again this comes at a cost when few outliers are present as is shown in the enlargements. Making ϵ larger decreases the cost but also decreases the benefit when many outliers are present. We can clearly see that the results for CWP and CP are similar. This similarity was true for all the different cases and from now on we will only report on CB and CP. Thus the effects of varying ϵ are quite similar for all performance criteria. The choice of ϵ must take into account the cost/benefit balance associated with ϵ. This feature is a typical dilemma in the choice of tuning constants for ϵ in statistical procedures; for example selecting a small size for a test is desirable to make the type I error rate small, but also decreases the power of the test to reject the null hypothesis when it is not true and vice versa.

In Figure 4.3 we repeat the analysis but this time for a different cut-off value c. We used $c = 0.01$ at the two left panels and $c = 0.1$ for the two right panels. The results are qualitatively similar to Figure 4.2 where we used $c = 0.05$. However with the choice of $c = 0.01$ we have to make ϵ smaller (for example 0.1) for the cost/benefit balance not to disappear, since otherwise DOUW performs similar to MEL and WEMEL. By contrast, for the choice $c = 0.1$. the benefit increases but so does the cost and to keep these in balance we need to make a larger choice of ϵ (for example 0.3). It appears that the combination of choices, $(\epsilon, c) = (0.1, 0.01), (0.2, 0.05), (0.3, 0.1)$ are reasonable but to some extent this combination is a judgement call based on limited experience and the issue of making sensible choices in practice is still open at this stage. Perhaps the best practice is to vary these tuning constants and to judge the answers accordingly.

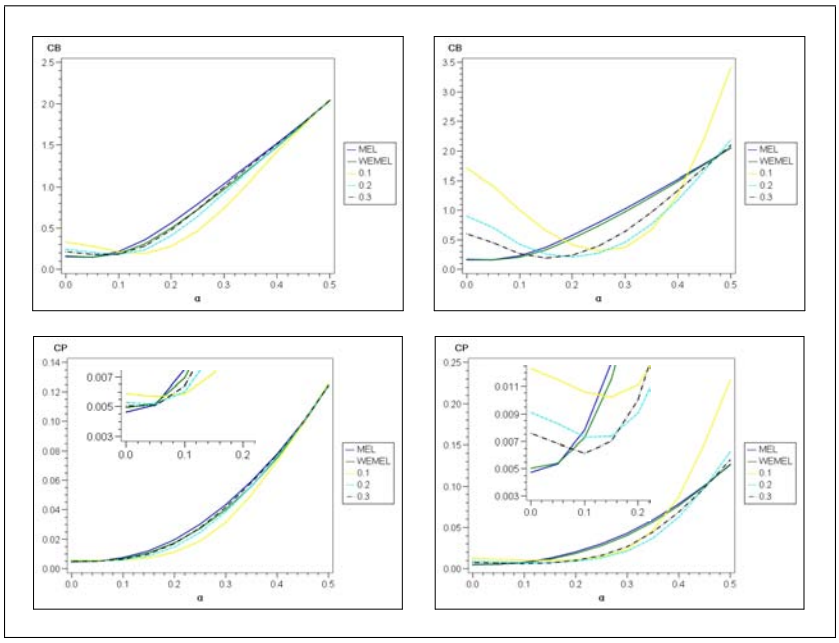

Figure 4.3: CB and CP values for $\epsilon = \{0.1, 0.2, 0.3\}$ (with $c = 0.01$ left and $c = 0.10$ right) for case 1

In Figure 4.4 we compared the DOUW procedure with ML replaced by MEL throughout (last variation in Section 3). Here we see that DOUW-MEL performs better than DOUW-ML when α

is small (few outliers), but otherwise the two versions of DOUW perform quite similarly.

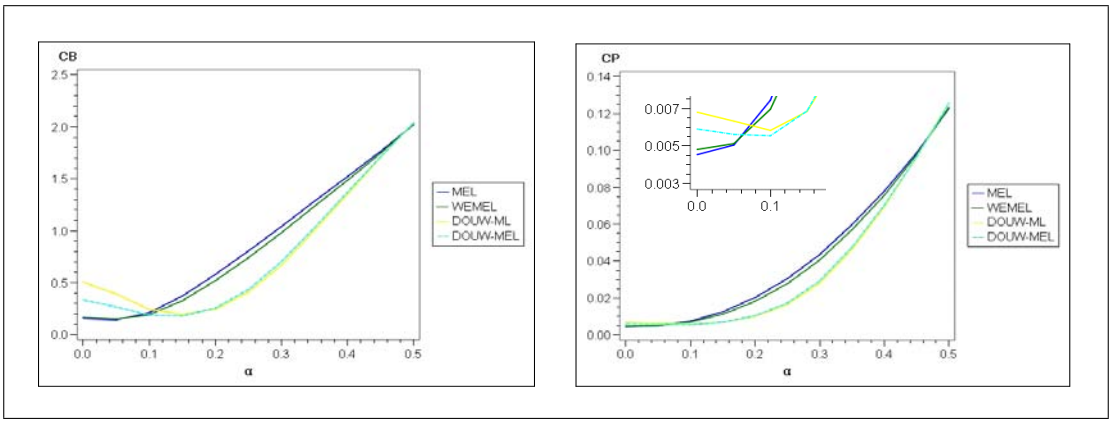

Figure 4.4: CB and CP values for DOUW when using ML and MEL for case 1 ($\epsilon = 0.2$, $c = 0.05$)

Proceeding to **cases 2** and **3**, Figures 4.5 and 4.6 show the corresponding CB and CP curves respectively. A similar cost/benefit balance effect is again evident as in case 1. In both these figures, we used $c = 0.05$ on the left and $c = 0.10$ on the right while ϵ varies over the values 0.1 to 0.3. We see that the choice of the cut-off again influences the extent of the cost/benefit balance. For illustrative purposes, we will from now on use $\epsilon = 0.2$ and $c = 0.05$. We also did the same analysis with data generated under the HLR model and found similar results.

For **case 4** we have plotted the CB and CP values for $n = 50$ and $n = 200$ (not shown in this shortened version). These may be compared to the case $n = 100$ in the top right and bottom right panels of Figure 4.2. We can see in all these cases that we have similar behavior with regards to the cost/benefit balance of DOUW vs MEL and WEMEL. However DOUW does even better than MEL and WEMEL as n increases. The benefit is very large and the cost is minimal if we consider $n = 200$, while the benefit of DOUW above WEMEL and MEL is less when $n = 50$ and the cost of DOUW is higher when $n = 50$.

We have also run simulations on 3 and more regressors and found similar results namely increasing ϵ diminishes both the cost and the benefit associated with DOUW and so does decreasing c. The effects are more pronounced in large samples than in small samples. More research is required to establish concrete guidelines for choices of these tuning constants in practice.

5 Examples

In this section we will apply the DOUW procedure to a number of benchmark data sets as well as a large new dataset from so-called RFM analysis. The benchmark datasets are:

- banknotes (Rousseeuw & Christmann, 2003)
- toxoplasmosis (Efron, 1986)
- vaso constriction (Finney, 1947; Pregibon, 1981)
- food stamp (Kunsch et al. 1989)

These benchmark datasets were also used for illustration purposes by Rousseeuw & Christmann (2003). The RFM dataset will be explained later on.

Tables 5.1 - 5.4 present the results for the benchmark datasets and the column labelled #O indicates the number of outliers found. In these examples we used MEL in DOUW. The parameter estimates for ML, MEL and WEMEL are the same as those of Rousseeuw & Christmann (2003), barring the slight difference in the case of WEMEL which may be due to our using different software. The **banknotes** dataset in Table 5.1 has no overlap and therefore the ML does not exist. The DOUW procedure found no outliers and therefore has the same estimated β's

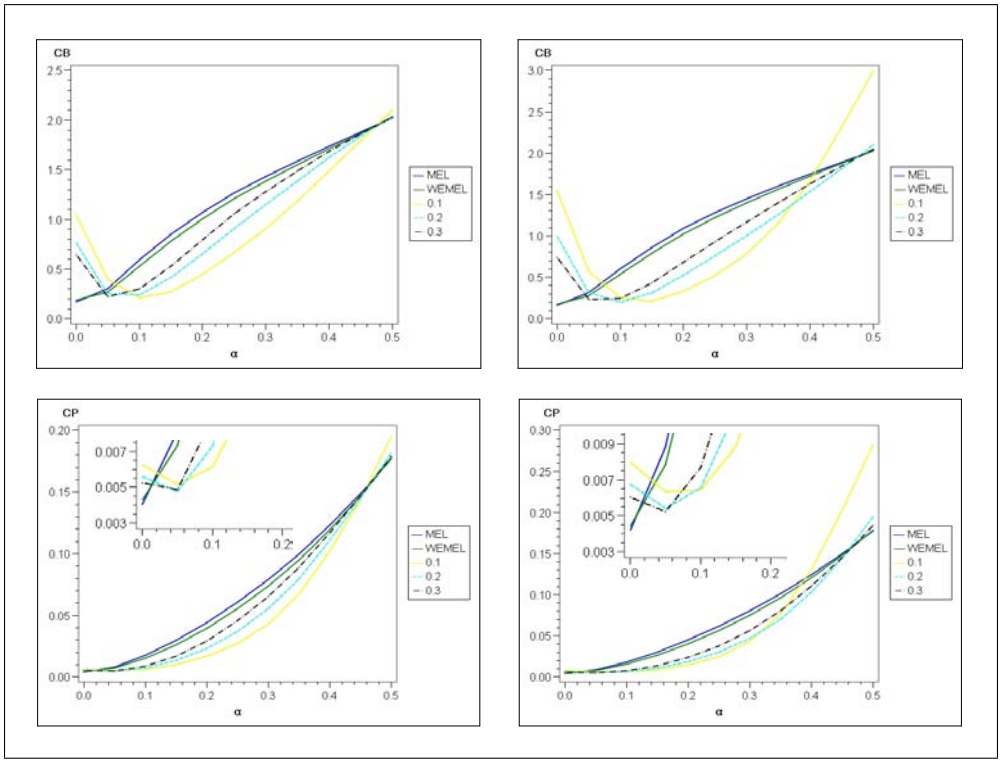

Figure 4.5: CB and CP values for $\epsilon = \{0.1, 0.2, 0.3\}$ and $c = 0.05$(left), $c = 0.1$(right) for case 2

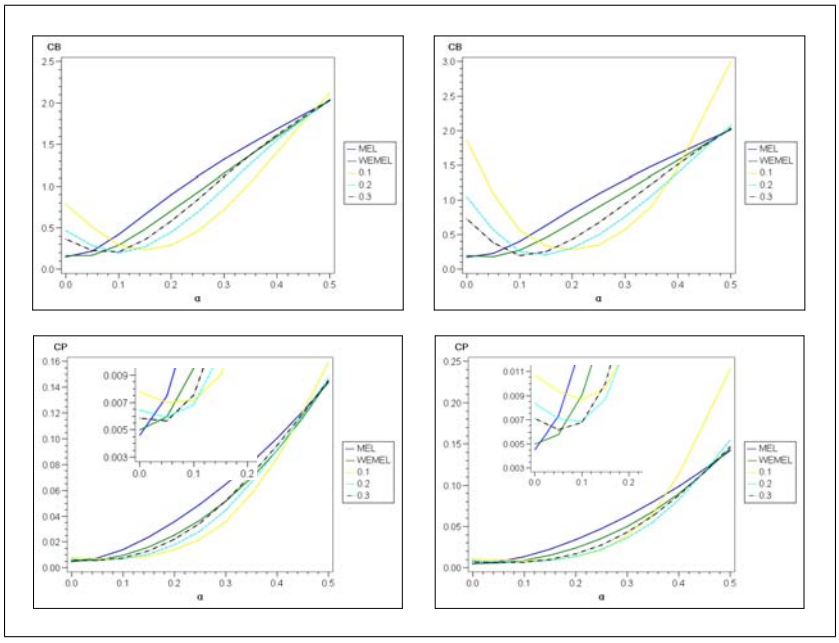

Figure 4.6: CB and CP values for $\epsilon = \{0.1, 0.2, 0.3\}$ and $c = 0.05$(left), $c = 0.1$(right) for case 3

as the MEL procedure, differing from WEMEL due to the radically different weighting scheme used by the latter. In Table 5.2 we show the results for the **toxoplasmosis** data set. Again no observations are identified as outliers when using DOUW and the estimated β's are quite similar for all the procedures. In Table 5.3 we have the results for the **vaso constriction** dataset which was extensively used in the literature often reporting observations 4 and 18 as outliers. With the small cut-off $c = 0.01$ DOUW reports no outliers but with $c >= 0.05$ observations 4 and 18 are flagged as outliers here also. Note that the estimates of β's for these choices are substantially different from the estimates found by ML, MEL and WEMEL and DOUW with $c = 0.01$. Again this is a reflection of the substantial influence that outliers have on the estimated parameters.

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	**#O**
ML	- does not exist -							
MEL	147.09	0.4649	-1.0204	1.3316	2.2049	2.3218	-2.3703	
WEMEL	252.55	-0.2541	-1.5791	2.03376	2.1012	2.3706	-2.1496	
DOUW$_{c=0.01,\epsilon=0.1}$	147.09	0.4649	-1.0204	1.3316	2.2049	2.3218	-2.3703	0
DOUW$_{c=0.05,\epsilon=0.2}$	147.09	0.4649	-1.0204	1.3316	2.2049	2.3218	-2.3703	0
DOUW$_{c=0.10,\epsilon=0.3}$	147.09	0.4649	-1.0204	1.3316	2.2049	2.3218	-2.3703	0

Table 5.1: Banknotes (N=200, K=7)

Method	β_0	β_1	β_2	β_3	**#O**
ML	0.09939	-0.44846	-0.18727	0.21342	
MEL	0.09882	-0.44395	-0.18536	0.21126	
WEMEL	0.09932	-0.40999	-0.16756	0.20259	
DOUW$_{c=0.01,\epsilon=0.1}$	0.1346	-0.46283	-0.27584	0.25642	0
DOUW$_{c=0.05,\epsilon=0.2}$	0.13002	-0.46045	-0.26424	0.25064	0
DOUW$_{c=0.10,\epsilon=0.3}$	0.12562	-0.45814	-0.25308	0.24508	0

Table 5.2: Toxoplasmosis (N=694, K=4)

Method	β_0	β_1	β_2	**#O**
ML	-2.92382	5.2205	4.6312	
MEL	-2.76789	4.9844	4.4064	
WEMEL	-2.73954	4.9487	4.3641	
DOUW$_{c=0.01,\epsilon=0.1}$	-2.76789	4.9844	4.4064	0
DOUW$_{c=0.05,\epsilon=0.2}$	-4.12743	6.8738	6.0565	2
DOUW$_{c=0.10,\epsilon=0.3}$	-6.11277	9.6801	8.5351	2

Table 5.3: Vaso constriction (N=39, K=3)

In Table 5.4 we show the results of the **foodstamp** data. Again the small choice of $c = 0.01$ identifies no outliers but either 3 (observations 66, 137 and 147) or 6 outliers (the former plus observations 22, 103 and 120) are declared when $c = 0.05$ and 0.10 respectively. Pregibon (1981) developed logistic regression diagnostic plots as a tool to identify outliers. One of these namely the deviance residual plot can be stated as follows: Define $Dev_n = -\sqrt{-D_n(\widehat{\boldsymbol{\beta}})}$ if $y_n = 0$ and $Dev_n = \sqrt{-D_n(\widehat{\boldsymbol{\beta}})}$ if $y_n = 1$. We plot Dev_n against the observation number n (not shown in this shortened version). Then the downliers are represented by the extreme lower observations and the upliers by the extreme upper observations in this plot. Again observations 66, 137 and 147 show up as downliers when the cut-off level is is about -2.2 and in addition also observations 22, 103 and 120 when the cut-off level is about -1.9. It is noteworthy that the estimates of the β_i's produced by the estimators are substantially different for this dataset.

The next example relates to what is called RFM (Recency, Frequency, Monetary) analysis in CRM (customer relations management). This example is not shown in this shortened version, only in the complete version available from Tanja.delaRey@nwu.ac.za.

Method	β_0	β_1	β_2	β_3	**#O**
ML	0.92638	-1.85021	0.89606	-0.33275	
MEL	0.8936	-1.82665	0.88498	-0.32772	
WEMEL	5.37607	-1.75504	0.61952	-1.06607	
DOUW$_{c=0.01,\epsilon=0.1}$	1.21335	-2.14949	1.06178	-0.39777	0
DOUW$_{c=0.05,\epsilon=0.2}$	0.93637	-2.314	1.13623	-0.35559	3
DOUW$_{c=0.10,\epsilon=0.3}$	0.51745	-3.00769	0.75962	-0.25222	6

Table 5.4: Food stamp (N=150, K=4)

6 Conclusion

Outliers in logistic regression data can be flagged by so-called deviance diagnostic (or similar) analysis. Once we have the ML estimate $\widehat{\beta}$ of the regression coefficients, we calculate the deviances of the observations and classify those with the most extreme negative deviances as the downliers and those with the most extreme positive deviances as the upliers, using some cut-off level to express the extent of outlyingness. The problem with this approach is that the outliers (if any) were included in the observations on which the ML estimate $\widehat{\beta}$ was based to begin with and this inclusion may seriously effect the results. Among other consequences, it leaves the procedure vulnerable to the well-known dangers of masking and swamping in outlier identification. To guard against these risks one must use a procedure that does parameter estimation and outlier detection simultaneously. Trimming approaches have been used successfully for this purpose in a number of areas in statistics but in logistic regression trimming runs into the separation problem which makes it difficult to apply. In this paper we presented an approach based on associating pre-assigned large and small (but positive) weights with the observations in an optimal way as part of the likelihood maximization. This device enables the identification of the outliers as those that are assigned the small weights. The required maximization is handled by a search method based on repeated random starting subsets to which the high weights are assigned, followed by C-step improvements similar to that used in ordinary LTS regression. We refer to the method as DOUW and its properties depend on two tuning constants, namely the ratio of the small to the large weights and the probability cut-off level used to measure outlyingness. We present a simulation study to show the effects of these constants on the performance of DOUW and illustrate the results in terms of four benchmark data sets as well as a large new data set from the application area of retail marketing campaign analysis. On-going research is aimed at improving specification of the procedure for practical use.

References

CHRISTMANN, A. & ROUSSEEUW P.J. 2001. Measuring overlap in binary regression. *Computational Statistics & Data Analysis*, 37(1):65-75, July 2001.

COPAS, J.B. 1988. Binary regression models for contaminated data. With discussion. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(2):225-265.

EFRON, B. 1986. Double exponential families and their use in generalized linear regression. *Journal of American Statistical Association*. 81 (395), 709-721, Sep 1986.

FINNEY, D.J. 1947. The estimation from individual records of the relationship between dose and quantal response. *Biometrika*. 32(1): 320-334.

HASTIE T, TIBSHIRANI R. & FRIEDMAN J. 2001. The Elements of Statistical Learning: Data Mining, Inference and Prediction. New York: Springer. 533 p.

HOSMER, D.W. & LEMESHOW, S. 1989. Applied logistic regression. New York: Wiley. 307 p.

HUBER, P.J. 1973. Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics*, 1(5):799-821, Sep. 1973.

KLEINBAUM, D.G. 1994. Logistic regression: a self-learning text. New York: Springer. 282 p.

KUNSCH, 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(406):460-466, Jun. 1989.

LINHART, H. & ZUCCHINI, W. 1986. Model Selection. New York: John Wiley & Sons. 301 p.

NEYKOV, N.M. & MULLER, C.H. 2002. Breakdown Point and Computation of Trimmed Likelihood Estimators in GLMs. In: R. Dutter et al., editors, *Developments in robust statistics*, Physica Verlag, Heidelberg. 286 pp.

NOVA, J. 2000. Drilling Down: Turning Customer Data into Profits with a Spreadsheet. Florida: Deep South Publishing Company. 196 p.

PREGIBON, D. 1981. Logistic Regression Diagnostics. *The Annals of Statistics*, 9(4):705-724, Jul.

ROUSSEEUW, P.J. 1984. Least Median of Squares Regression. *Journal of the American Statistical Association*, 79(388):871 -880, Dec. 1984.

ROUSSEEUW, P.J. & CHRISTMANN, A. 2003. Robustness against separation and outliers in logistic regression. *Computational Statistics & Data Analysis*, 43(3):315-332, July 2003.

ROUSSEEUW, P.J. & LEROY, A.M. 1987. Robust regression and outlier detection. New York: John Wiley & Sons. 329p.

ROUSSEEUW, P.J. & VAN DRIESSEN, K. 1999a. Computing LTS Regression for Large Data Sets. Technical report, University of Antwerp. 21 p.

ROUSSEEUW, P.J. & VAN DRIESSEN, K. 1999b. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, 41(3):212 -223, August 1999..

YOHAI, V.J. 1987. High Break-Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics*, 15(2): 642-656, Jun. 1987.