The hidden logistic regression model

P.J. Rousseeuw[1] and A. Christmann[2]

[1] Universitaire Instelling Antwerpen (UIA), Department of Mathematics and Computer Science, Universiteitsplein 1, B-2610 Wilrijk, Belgium
[2] University of Dortmund, SFB-475, HRZ, 44221 Dortmund, Germany

Keywords: Logistic regression, Hidden layer, Overlap, Robustness.

Abstract

The logistic regression model is commonly used to describe the effect of one or several explanatory variables on a binary response variable. Here we consider an alternative model under which the observed response is strongly related but not equal to the unobservable true response. We call this the *hidden logistic regression (HLR) model* because the unobservable true responses are comparable to a hidden layer in a feedforward neural net. We propose the *maximum estimated likelihood* method in this model, which is robust against separation unlike existing methods for logistic regression. We also consider outlier-robust estimation in this setting.

1 Introduction

The logistic regression model assumes independent Bernoulli distributed response variables with success probabilities $\Lambda(x_i'\theta)$ where Λ is the logistic distribution function, $x_i \in \mathbb{R}^p$ are vectors of explanatory variables, $1 \leq i \leq n$, and $\theta \in \mathbb{R}^p$ is unknown. Under these assumptions, the classical maximum likelihood (ML) estimator has certain asymptotic optimality properties. However, even if the logistic regression assumptions are satisfied there are data sets for which the ML estimate does not exist. This occurs for exactly those data sets in which there is no overlap between successes and failures, cf. Albert and Anderson (1984) and Santner and Duffy (1986). This identification problem is not limited to the ML estimator but is shared by all estimators for logistic regression, such as that of Künsch et al. (1989).

One way to deal with this problem is to measure the amount of overlap. This can be done by exploiting a connection between the notion of overlap and the notion of regression depth proposed by Rousseeuw and Hubert (1999), leading to the algorithm of Christmann and Rousseeuw (2001). A comparison between this approach and the support vector machine is given in Christmann, Fischer and Joachims (2000).

In Section 2 we use an alternative model, which is an extension of the logistic regression model. We assume that due to an additional stochastic mechanism the true response of a logistic regression model is unobservable, but that there exists an observable variable which is strongly related to the true response. E.g., in a medical context there is often no perfect laboratory test procedure to detect whether a specific illness is present or not (i.e., misclassification errors may sometimes occur). In that case, the true response (whether the disease is present) is not observable, but the result of the laboratory test is.

It can be argued that the true unobservable responses are comparable to a hidden layer in a feedforward neural network model, which is why we call this the *hidden logistic regression (HLR) model.* In Section 3 we propose the *maximum estimated likelihood (MEL) technique* in this model, and show that it is immune to the identification problem described above. The MEL estimator is studied by simulations (Section 4) and on real data sets (Section 5). In Section 6 we also consider outlier-robust estimation in this setting, whereas Section 7 provides a discussion and an outlook to further research.

2 The hidden logistic regression model

The classical logistic regression model assumes n observable independent responses Y_i with Bernoulli distributions $\text{Bi}(1, \Lambda(x_i'\theta))$, where $i = 1, \ldots, n$ and $\theta \in \mathbb{R}^p$. Throughout this paper we assume that there is an intercept, so we put $x_{i,1} = 1$ for all i, and thus $p \geq 2$.

The new model assumes that the true responses are unobservable (latent) due to an additional stochastic mechanism. In medical diagnosis there is typically no test procedure (e.g. a blood test) which is completely free of misclassification errors. Another possible cause of misclassifications is the occurrence of clerical errors.

To clarify the model, let us first consider a medical application with only $n = 1$ patient. His/her true status (e.g. presence or absence of the disease) has two possible values, typically denoted as success (s) and failure (f). We assume that the true status T is unobservable. However, we can observe the variable Y which is strongly related to T as in Figure 1. If the true status is $T = s$ we observe $Y = 1$ with probability $\text{P}(Y = 1|T = s) = \delta_1$, hence a misclassification occurs with probability $\text{P}(Y = 0|T = s) = 1 - \delta_1$. Analogously, if the true status is f we observe $Y = 1$ with probability $\text{P}(Y = 1|T = f) = \delta_0$ and we obtain $Y = 0$ with probability $\text{P}(Y = 0|T = f) = 1 - \delta_0$. We of course assume that the probability of observing the true status is higher than 50%, i.e. $0 < \delta_0 < 0.5 < \delta_1 < 1$.

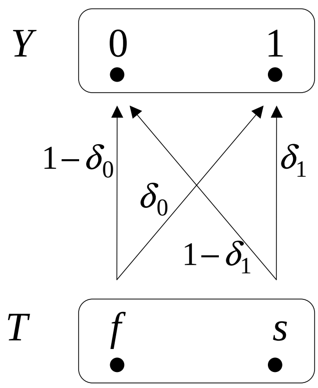

Fig. 1. Unobservable truth T and observable response Y.

Ekholm and Palmgren (1982) considered the general case with n observations. In our notation, there are n unobservable independent random variables T_i resulting from a classical logistic regression model with finite parameter vector $\theta = (\theta_1, \ldots, \theta_p)' = (\alpha, \beta_1, \ldots, \beta_{p-1})'$. Hence T_i has a Bernoulli distribution with success probability $\pi_i = \Lambda(x_i'\theta)$ where $\Lambda(z) = 1/[1 + \exp(-z)]$ and $x_i \in \mathbb{R}^p$. Furthermore, they assume that the observable responses Y_i are related to T_i as in Figure 1. For instance, when $T_i = s$ we obtain $Y_i = 1$ with probability $\text{P}(Y_i = 1|T_i = s) = \delta_1$ whereas $Y_i = 0$ occurs with the complementary probability $\text{P}(Y_i = 0|T_i = s) = 1 - \delta_1$. (The plain logistic model assumes δ_0 and $\delta_1 = 1$.) The entire mechanism in Figure 2 we call the hidden logistic regression model because the true status T_i is hidden by the stochastic structure in the top part of Figure 2. This model can be interpreted as a special kind of neural net, with a single hidden layer that corresponds to the latent variable T.

3 The maximum estimated likelihood method

a. Construction

We now need a way to fit data sets arising from the hidden logistic model. Two approaches already exist, by Ekholm and Palmgren (1982) and by Copas (1988), but here we will proceed in a different way.

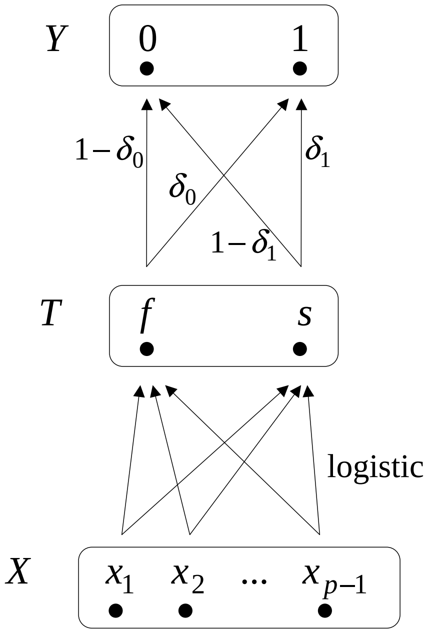

Fig. 2. Hidden logistic regression model.

Let us start by looking only at Figure 1, where Y is observed but T is not. How can we then estimate T? This is actually *the* smallest nontrivial estimation problem, because any such problem needs more than one possible value of the parameter and more than one possible outcome. Here we have exactly two values for both, and the only distributions on two possible outcomes are the Bernoulli distributions. Under f the likelihood of $Y = 0$ exceeds that of $Y = 1$, and under s the opposite holds. Therefore, the maximum likelihood estimator of T given $(Y = y)$ becomes simply

$$\begin{aligned} \hat{T}_{\mathrm{ML}}\ (Y=0) &= f \\ \hat{T}_{\mathrm{ML}}\ (Y=1) &= s \end{aligned} \tag{1}$$

which conforms with intuition.

Let us now consider the conditional probability that $Y = 1$ given $\hat{T}_{\mathrm{ML}}$, yielding

$$\begin{aligned} \mathrm{P}(Y=1|\hat{T}_{\mathrm{ML}}) &= \delta_0 \quad \text{if} \quad y=0 \\ &= \delta_1 \quad \text{if} \quad y=1 \end{aligned} \tag{2}$$

where y is the observed value of Y. Denoting (2) by $\tilde{Y}$, we can rewrite it as

$$\tilde{Y} = \delta_0 + (\delta_1 - \delta_0)Y = (1-Y)\delta_0 + Y\delta_1$$

which is a weighted average of δ_0 and δ_1 with weights $1 - Y$ and Y.

In the model with n observations y_i we obtain analogously

$$\tilde{y}_i = (1-y_i)\delta_0 + y_i\delta_1 \tag{3}$$

which we will call the *pseudo-observations*. In words, the pseudo-observation $\tilde{y}_i$ is the success probability conditional on the most likely estimate of the true status t_i.

We now want to fit a logistic regression to the pseudo-observations $\tilde{y}_i$. (In the classical case, $\tilde{y}_i = y_i$.) There are several estimation methods, but here we will apply the maximum likelihood formula. The goal is thus to maximize

$$L(\theta|(\tilde{y}_1,\ldots,\tilde{y}_n)) = \prod_{i=1}^{n} [\Lambda(x_i'\theta)]^{\tilde{y}_i} [1-\Lambda(x_i'\theta)]^{1-\tilde{y}_i} \tag{4}$$

over $\theta \in \mathbb{R}^p$. We call (4) the *estimated likelihood* because we don't know the true likelihood, which depends on the unobservable $t_1,\ldots,t_n$. (We only know the true likelihood when $\delta_0 = 0$ and $\delta_1 = 1$.) The maximizer $\hat{\theta}$ of (4) can thus be called the *maximum estimated likelihood (MEL)* estimator.

In order to compute the MEL estimator we can take the logarithm of (4), yielding

$$\sum_{i=1}^{n} \tilde{y}_i \ln(\Lambda(x_i'\theta)) + (1-\tilde{y}_i)\ln(1-\Lambda(x_i'\theta)) \tag{5}$$

which always exists since θ is finite. Differentiating with respect to θ yields the ($p-$variate) score function

$$s(\theta|(\tilde{y}_1,\ldots,\tilde{y}_n)) = \sum_{i=1}^{n} (\tilde{y}_i - \Lambda(x_i'\theta))\, x_i \tag{6}$$

for all $\theta \in \mathbb{R}^p$. Setting (6) equal to zero yields the desired estimate.

b. Properties of the MEL estimator

Unlike the classical ML estimator, the MEL estimator always exists.

Property 1. When $0 < \delta_0 < \delta_1 < 1$ and the data set has a design matrix of full column rank, the MEL estimator always exists and is unique.

(Note that when the design matrix is not of full column rank, we can first reduce the dimension of the x_i by means of principal component analysis.)

Proof. The Hessian matrix of (5) equals

$$\frac{\partial}{\partial\theta} s(\theta) = -\sum_{i=1}^{n} \Lambda(x_i'\theta)(1-\Lambda(x_i'\theta))\, x_i x_i' \tag{7}$$

and is thus negative definite because the design matrix has rank p. Therefore the differentiable function (5) is strictly concave. Now let us take any $\theta \neq 0$ and replace θ in (5) by $\lambda\theta$. If we let $\lambda \to +\infty$ then (5) always tends to $-\infty$ because there is at least one x_i in the data set with $x_i'\theta \neq 0$ due to full rank, and neither $\tilde{y}_i$ or $(1-\tilde{y}_i)$ can be zero. Therefore, there must be a finite maximizer $\hat{\theta}_{\text{MEL}}$ of (5), which is unique because the concavity is strict. □

This implies that the MEL estimator exists even when the data set has no overlap. Therefore also the resulting odds ratios $\text{OR}_j = \exp(\hat{\theta}_j)$ always exist, i.e. they are never zero or $+\infty$.

A property shared by all logistic regression estimators is $x-$affine equivariance. This says that when the x_i are replaced by $x_i^* = Ax_i$ where A is a nonsingular $p \times p$ matrix, then the regression coefficients transform accordingly.

Property 2. The MEL estimator is $x-$affine equivariant.

Proof. From (6) it follows that $\hat{\theta}^*_{\text{MEL}} = (A')^{-1}\hat{\theta}_{\text{MEL}}$ hence $(x_i^*)'\hat{\theta}^*_{\text{MEL}} = x_i'A'(A')^{-1}\hat{\theta}_{\text{MEL}} = x_i'\hat{\theta}_{\text{MEL}}$. This also yields the same predicted values. □

In linear regression there exist two other types of equivariance: one about adding a linear function to the response ('regression equivariance') and one about multiplying the response by a constant factor ('$y-$scale equivariance'), but these obviously do no apply to logistic regression.

c. Choice of δ_0 and δ_1

If δ_0 and δ_1 are known from the context (e.g. from the type I and type II error probabilities of a blood test) then we can use these values. But in many cases, δ_0 and δ_1 are not given in advance. Copas (1988, page 241) found that accurate estimation of δ_0 and δ_1 from the data itself is very difficult, if not impossible unless n is extremely large. He essentially considers them as tuning constants that can be chosen, as do we.

The 'symmetric' approach used by Copas is to choose a single constant $\gamma > 0$ and to set

$$\delta_0 = \gamma \quad \text{and} \quad \delta_1 = 1 - \gamma. \tag{8}$$

His computations require that γ be small enough so that terms in γ^2 can be ignored. In his Table 1 the values $\gamma = 0.01$ and $\gamma = 0.02$ occur, whereas he considers $\gamma = 0.05$ to be unreasonably high (page 238). In most of Copas' examples $\gamma = 0.01$ performs well, and this turns out to be true also for our MEL method, so we could use $\gamma = 0.01$ as the default choice. This approach has the advantage of simplicity.

On the other hand, there is something to be said for an 'asymmetric' choice which takes into account how many y_i's are 0 and 1 in the data set. Let us consider the marginal distribution of the y_i (that is, unconditional on the x_i) from which we construct some estimate $\hat{\pi}$ of the marginal success probability $\mathrm{P}(Y = 1)$. It seems reasonable to constrain δ_0 and δ_1 such that the average of the pseudo-observations $\tilde{y}_i$ corresponds to $\hat{\pi}$. This yields

$$\begin{aligned} \hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} \tilde{y}_i &= (1-\hat{\pi})\delta_0 + \hat{\pi}\delta_1 \\ \hat{\pi} - \hat{\pi}\delta_1 &= \delta_0 - \hat{\pi}\delta_0 \\ \frac{1-\delta_1}{\delta_1 - \hat{\pi}} &= \frac{\delta_0}{\hat{\pi} - \delta_0} \,. \end{aligned}$$

Since it is natural to assume that $\delta_0 < \hat{\pi} < \delta_1$ the latter ratios equal a (small) positive number which we will denote by δ. Consequently we can write both δ_0 and δ_1 as functions of δ, as:

$$\delta_0 = \frac{\hat{\pi}\delta}{1+\delta} \quad \text{and} \quad \delta_1 = \frac{1+\hat{\pi}\delta}{1+\delta}\,. \tag{9}$$

However, since we have assumed that $\delta_0 < \hat{\pi} < \delta_1$ we have to construct $\hat{\pi}$ accordingly. We cannot take the standard estimate $\bar{\pi} = \frac{1}{n}\sum_{i=1}^{n} y_i =$ (number of $y_i = 1)/n$ because $\bar{\pi}$ can become 0 or 1. A natural idea is to bound $\bar{\pi}$ away from 0 and 1 by putting

$$\hat{\pi} = \max\left(\delta, \min(1-\delta, \bar{\pi})\right) \tag{10}$$

which means truncation at δ and $1-\delta$. This is sufficient because always

$$\delta_0 = \frac{\hat{\pi}\delta}{1+\delta} < \frac{\hat{\pi}+\hat{\pi}\delta}{1+\delta} = \hat{\pi}$$

and

$$\delta_1 = \frac{1+\hat{\pi}\delta}{1+\delta} > \frac{\hat{\pi}+\hat{\pi}\delta}{1+\delta} = \hat{\pi}$$

hence $\delta_0 < \hat{\pi} < \delta_1$. Note that both misclassification probabilities in Figure 1 are less than δ because

$$\delta_0 = \frac{\hat{\pi}\delta}{1+\delta} < \frac{\delta}{1+\delta} < \delta$$

and

$$1-\delta_1 = \frac{1+\delta-1-\hat{\pi}\delta}{1+\delta} = \frac{(1-\hat{\pi})\delta}{1+\delta} < \frac{\delta}{1+\delta} < \delta\,.$$

Our default choice will be $\delta = 0.01$, which implies smaller classification errors than by putting $\gamma = 0.01$ in formula (8).

When the data are 'balanced' in the sense that there are as many $y_i = 1$ as $y_i = 0$, expression (10) yields $\hat{\pi} = 0.5$ hence $\delta_0 = 1 - \delta_1$ by (9), yielding identical misclassification probabilities, as in the symmetric formulas (8). In all other, 'unbalanced' cases, our asymmetric approach yields less biased predictions. An extreme case is when all $y_i = 1$. (This is a situation where the classical ML estimator does not exist.) The MEL estimator will put all $\tilde{y}_i = \delta_1$ yielding a fit with all slopes $\hat{\beta}_1 = \ldots = \hat{\beta}_{p-1} = 0$ and with intercept $\alpha = \Lambda^{-1}(\delta_1)$. Using the symmetric approach (8) yields $\delta_1 = 0.99$ hence $\hat{\alpha} = \text{logit}(0.99) = \ln(99) \approx 4.595$ so the fitted values are constant and equal to 0.99. On the other hand, the asymmetric approach yields $\hat{\pi} = 0.99$ and $\delta_1 = (1 + (0.99)(0.01))/(1 + 0.01) = 1.0099/1.01 = 0.999901$. This again yields zero slopes but a larger intercept $\hat{\alpha} = \text{logit}(0.999901) = \ln(10099) \approx 9.22$ so the fitted values are 0.9999 which is much closer to 1.

Our recommendation is therefore to compute $\hat{\pi}$, δ_0, and δ_1 as in (9) and (10) with $\delta = 0.01$, to compute the pseudo-observations $\tilde{y}_i$ according to (3) and to carry out the resulting MEL method.

Our S-PLUS code for this method can be downloaded from

`http://win-www.uia.ac.be/u/statis/software/HLR_readme.html`

or

`http://www.statistik.uni-dortmund.de/sfb475/berichte/rouschr2.zip`

The ML estimator has the nice property under the logistic regression model that if $\hat{\theta}$ is the ML estimate for the data set $\{(x_i', y_i),\ 1 \le i \le n\}$, then $-\hat{\theta}$ is the ML estimate for the data set $\{(x_i', 1 - y_i),\ 1 \le i \le n\}$. Hence, recoding all response variables Y_i to $1 - Y_i$ affects the ML estimator only in the way that it changes the signs of the regression coefficients, and the odds ratios become $\exp(-\hat{\theta}_j) = 1/\text{OR}_j$. We call this equivariance with respect to recoding the response variable. The MEL estimator has the same property, whether δ_0 and δ_1 are given by (8) or (9) .

Property 3. The MEL estimator is equivariant with respect to recoding the response variable.

Proof. Writing $y_i^* = 1 - y_i$ and recomputing (10) and (9) [or (8)] yields $\tilde{y}_i^* = 1 - \tilde{y}_i$ by (3). Applying the ML estimator to the $(x_i', \tilde{y}_i^*)$ yields the desired result. □

4 Simulations

In this section we carry out a small simulation to compare the bias and the standard error of the usual ML estimator and the proposed MEL estimator with $\delta = 0.01$ under the assumptions of the logistic regression model. We will estimate $p = 3$ coefficients, including the intercept term. Both explanatory variables are generated from the standard normal distribution. As true parameter vectors we use $\theta_A = (1, 0, 0)'$ and $\theta_B = (1, 1, 2)'$. The number of observations n will be 20, 50, and 100. For each situation 1,000 samples are generated.

We use the depth-based algorithm (Christmann and Rousseeuw 2001) to check whether the data set has overlap, i.e. whether the ML estimate exists. It turned out that there were 12 data sets without overlap for $n = 20$ with θ_A, and 129 data sets without overlap for $n = 20$ with θ_B. This contrasts sharply to the MEL estimate, which existed for all data sets.

Table 1 compares ML and MEL for the data sets with overlap. In situation A, where the true slopes are zero, there is not much difference between the estimators. But in situation B, the MEL estimator has a substantially smaller bias and standard error than the ML estimator. This can be explained by the well-known phenomenon that ML tends to overestimate the magnitude of nonzero coefficients, whereas MEL exhibits a kind of 'shrinkage' behavior.

5 Examples

In this section we consider some benchmark data sets. Both the banknotes data set (Riedwyl 1997) and the hemophilia data set (Hermans and Habbema 1975) have no overlap, hence their ML estimate does not exist. The vaso constriction data (Finney 1947, Pregibon 1981) and the food stamp data (Künsch et al. 1989) are well-known in the literature on outlier detection and robust logistic regression. They both have little overlap: it suffices to delete 3 (resp. 6) observations in these data sets to make the ML estimate nonexistent (see Christmann and Rousseeuw 2001). Some of these observations are considered as outliers in Künsch et al. (1989). The cancer remission data set (Lee 1974) is chosen because $n/p \approx 4$ is small. The toxoplasmosis data set (Efron 1986) and the IVC data set (Jaeger et al. 1997, 1998) have a large n.

Table 1: Bias and Standard Error of the ML estimator and the MEL estimator with $\delta = 0.01$.

n		ML Bias	ML SE	MEL Bias	MEL SE
Case A with $\theta = (1,0,0)'$					
20	α	0.291	0.032	0.272	0.028
	β_1	0.010	0.031	0.009	0.029
	β_2	-0.014	0.035	-0.004	0.030
50	α	0.097	0.012	0.095	0.012
	β_1	-0.015	0.011	-0.015	0.011
	β_2	-0.021	0.012	-0.021	0.012
100	α	0.053	0.008	0.052	0.008
	β_1	0.004	0.008	0.004	0.008
	β_2	-0.004	0.008	-0.004	0.008
Case B with $\theta = (1,1,2)'$					
20	α	0.586	0.067	0.360	0.039
	β_1	0.652	0.083	0.364	0.045
	β_2	1.372	0.159	0.780	0.057
50	α	0.133	0.019	0.097	0.017
	β_1	0.156	0.022	0.104	0.019
	β_2	0.350	0.030	0.247	0.025
100	α	0.061	0.011	0.038	0.010
	β_1	0.085	0.012	0.050	0.011
	β_2	0.154	0.016	0.084	0.015

The IVC data set describes an in vitro experiment to study possible risk factors of the thrombus-capturing efficacy of inferior vena cava (IVC) filters. We focus on the study of a particular conical IVC filter, for which the design consisted of 48 different settings x_i. For each vector x_i there were m_i replications with $m_i \in \{50, 60, 90, 100\}$, yielding a total of $n = 3200$.

Table 2: Comparison between MEL estimates with $\delta = 0.01$ and ML estimates.

Data set (n,p)	Method	$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
Banknotes	ML	– no overlap, ML does not exist –						
$(200,7)$	MEL	147.09	0.46	-1.02	1.33	2.20	2.32	-2.37
Hemophilia	ML	– no overlap, ML does not exist –						
$(52,3)$	MEL	-5.43	-56.59	47.39				
Vaso constriction	ML	-2.92	5.22	4.63				
$(39,3)$	MEL	-2.77	4.98	4.41				
Food stamp	ML	0.93	-1.85	0.90	-0.33			
$(150,4)$	MEL	0.89	-1.83	0.88	-0.33			
Cancer remission	ML	58.04	24.66	19.29	-19.60	3.90	0.15	-87.43
$(27,7)$	MEL	58.51	18.20	12.20	-12.19	3.68	0.14	-81.42
Toxoplasmosis	ML	0.10	-0.45	-0.19	0.21			
$(697,4)$	MEL	0.10	-0.44	-0.19	0.21			
IVC	ML	-1.79	0.67	-1.05	-1.25	1.83		
$(3200,5)$	MEL	-1.73	0.65	-1.03	-1.22	1.79		

Table 2 shows that the MEL estimates with $\delta = 0.01$ were quite similar to the ML estimates for the data sets with overlap. This is even true for the cancer remission data set taking into account the huge standard errors of the ML coefficients, namely 71.23, 47.84, 57.95, 61.68, 2.34, 2.28, and 67.57. The odds ratios $\exp(\hat{\theta}_j)$ based on the ML and MEL estimates were quite similar too (see Table 3).

Figure 3 shows that the choice of δ has relatively little impact on the MEL estimates for the food stamp data set, which has overlap. Figure 4 shows the effect of δ for the banknotes data. Because the latter data set has no overlap we know that $||\hat{\theta}||$ tends to $+\infty$ as δ goes to 0 (since $\delta = 0$ corresponds to the ML estimator). One could therefore use δ like a 'ridge parameter' in Figure 4.

Table 3: Comparison of odds ratios based on ML and MEL.

Data set (n,p)	Method	$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Vaso constriction	ML	0.05	185.03	102.64	
	MEL	0.06	146.13	81.97	
Food stamp	ML	2.53	0.16	2.45	0.72
	MEL	2.44	0.16	2.42	0.72
Toxoplasmosis	ML	1.10	0.64	0.83	1.24
	MEL	1.10	0.64	0.83	1.24

6 Outlier-robust estimation

In the literature on logistic regression, many robust alternatives to the maximum likelihood estimator have been proposed. They can easily be modified for the hidden logistic regression model in the same way that we constructed the MEL estimator, i.e. by applying them to the pseudo-observations (3).

As an example we will consider a modification of the least trimmed weighted squares (LTWS) estimator of Christmann (1994a) which is defined as follows. We assume large strata, i.e. each design point x_i has m_i responses Y_i^j for $j = 1, \ldots, m_i$. One then adds all the Y_i^j corresponding to that x_i yielding

$$Z_i = \sum_{j=1}^{m_i} Y_i^j \in \{0, \ldots, m_i\}$$

and redefines n as the number of the x_i's (which is less than the total number of original responses Y_i^j). The large strata assumption says that n and p are fixed while $\min_{1 \le i \le n} m_i \to \infty$ and $m_i / (\sum_{j=1}^n m_j) \to k_i \in (0,1)$. One then puts $\pi_i = Z_i/m_i$ and $Z_i^* = (m_i \pi_i (1 - \pi_i))^{1/2} \Lambda^{-1}(\pi_i)$ as well as $X_i^* = (m_i \pi_i (1 - \pi_i))^{1/2} x_i$. For large values of m_i the Z_i^* approximately follow a linear regression model in the X_i^*. Christmann (1994a) defined the LTWS estimator of θ as the least trimmed squares estimator (Rousseeuw 1984) applied to the transformed variables Z_i^* and X_i^*, that is

$$\hat{\theta}_{\mathrm{LTWS}} = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{h} r_{i:n}^2$$

where $r_{1:n}^2 \le \ldots \le r_{n:n}^2$ are the ordered squared residuals where $r_i = Z_i^* - \theta' X_i^*$. The robustness aspects and asymptotic behavior of $\hat{\theta}_{\mathrm{LWTS}}$ were investigated in Christmann (1994a, 1998) .

In the hidden logistic model, we apply the LTWS method to the pseudo-observations $\tilde{y}_i$ defined in (3), with δ_0 and δ_1 given by (9) and (10). That is, we put

$$\tilde{Y}_i^j = (1 - Y_i^j)\delta_0 + Y_i^j \delta_1$$

yielding the corresponding variable $\tilde{Z}_i = \sum_{j=1}^{m_i} \tilde{Y}_i^j$. Substituting $\tilde{Z}_i$ for Z_i yields $\tilde{\pi}_i = \tilde{Z}_i / m_i$ and

$$\begin{aligned} \tilde{Z}_i^* &= (m_i \tilde{\pi}_i (1 - \tilde{\pi}_i))^{1/2} \Lambda^{-1}(\tilde{\pi}_i) \\ \tilde{X}_i^* &= (m_i \tilde{\pi}_i (1 - \tilde{\pi}_i))^{1/2} x_i \end{aligned}$$

to which we apply LTS regression. Like the MEL estimator, this modified LTWS estimator exists for all data sets (and it is still $x-$affine equivariant). In addition, it is also robust to outliers in Z_i and x_i. (The latter means that the modified LTWS estimator can resist the effect of leverage points, unlike some other robust approaches.)

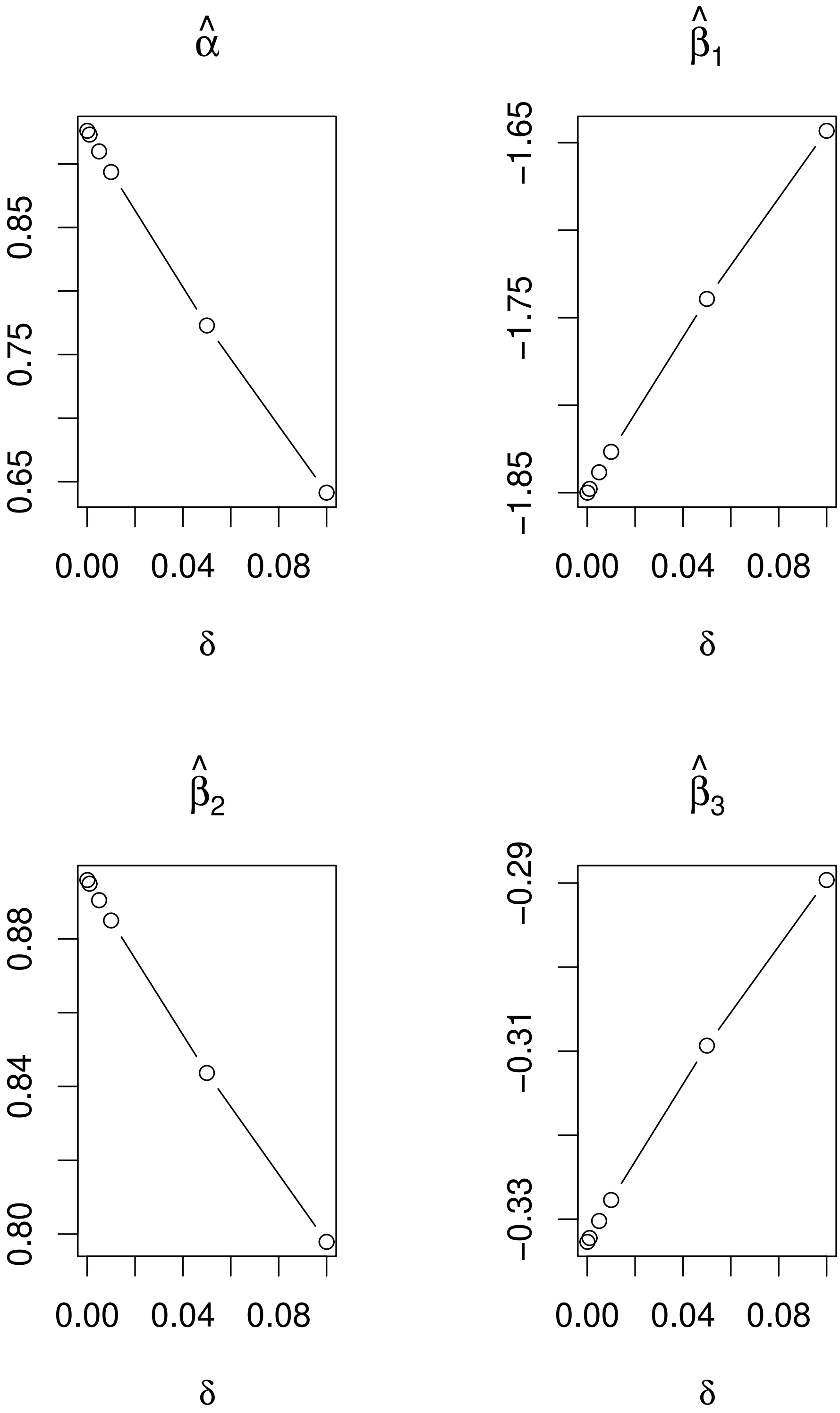

Fig. 3. Graphs of the MEL coefficients versus δ for the food stamp data set, for $\delta = 0.0001, 0.001, 0.005, 0.01, 0.05$, and 0.1.

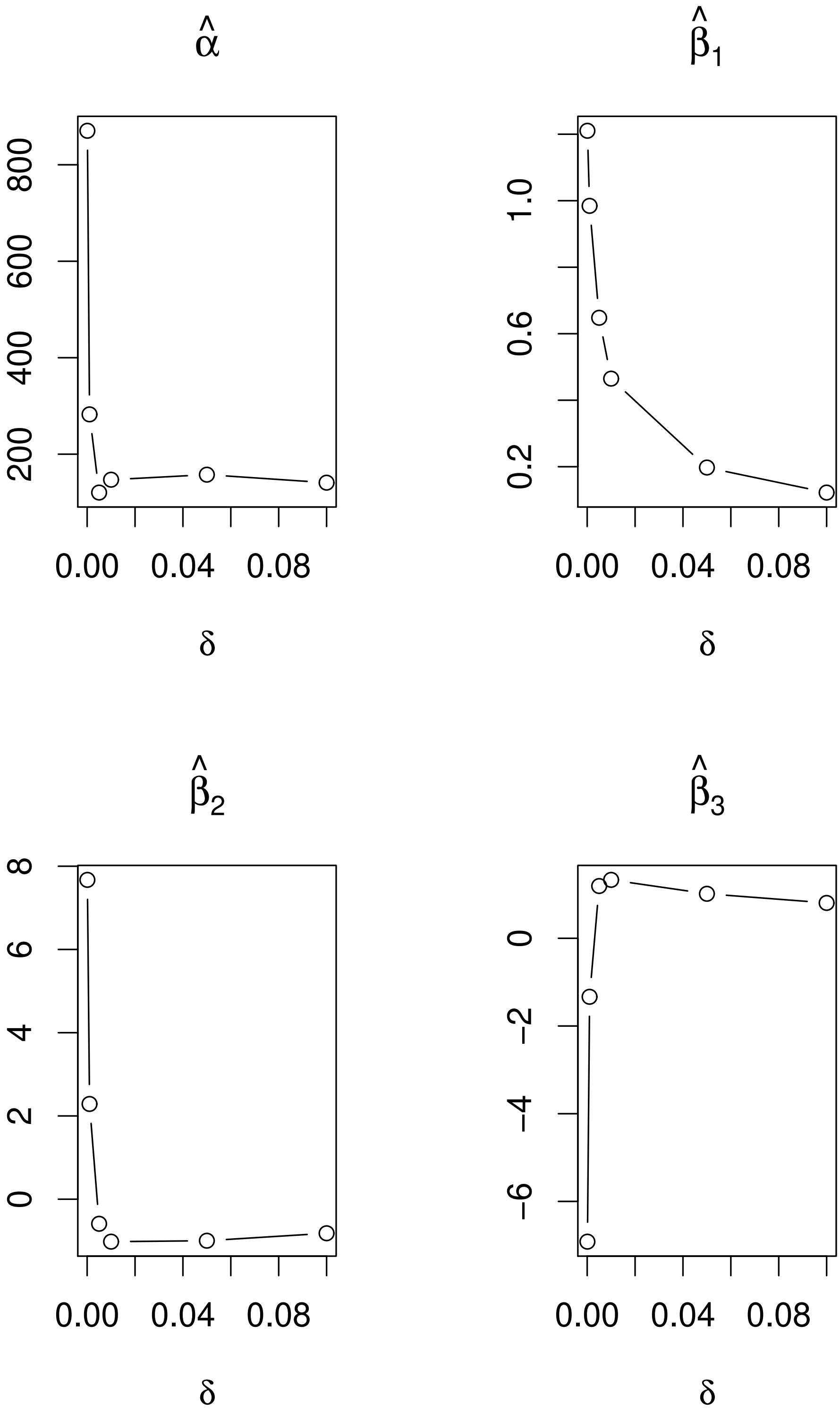

Fig. 4. Graphs of the first four MEL coefficients versus δ for the banknotes data set, for $\delta = 0.0001, 0.001, 0.005, 0.01, 0.05$, and 0.1.

Let us illustrate the modified LTWS estimator on the toxoplasmosis data set (Efron 1986) of Section 5. In aggregated form this data set has $n = 34$ observations, with m_i ranging from 1 to 82 with a mean of 20.5. We ran the modified LTWS method with the default choices $\delta = 0.01$ and $h = [\![0.75n]\!] = 25$, which took only a few seconds because we used the FAST-LTS program (Rousseeuw and Van Driessen 1999b). The resulting coefficients were $(-0.37, -1.26, -0.17, 0.42)'$ which clearly differ from the non-robust coefficients given in Table 2. Of course, the odds ratios 0.69, 0.28, 0.84, and 1.52 based on the outlier-robust approach also differ from the non-robust odds ratios in Table 3. The observations 27, 28, and 30 stick out in the robust residual plot (Figure 5), which agrees with findings based on a robust minimum Hellinger distance approach (Christmann 1994b).

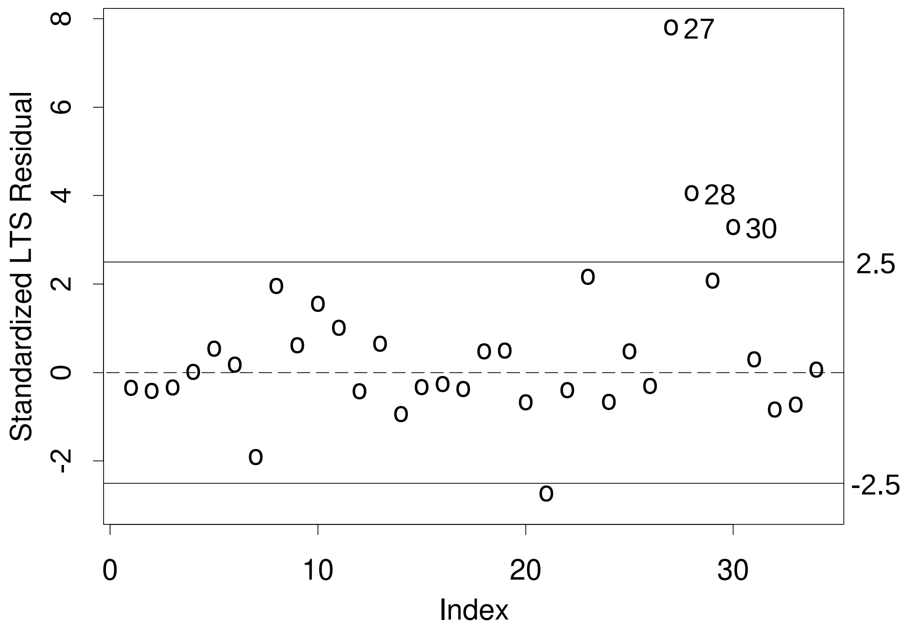

Fig. 5. Toxoplasmosis data: Index plot of residuals based on modified LTWS.

7 Discussion and outlook

The main problem addressed in this paper is that the coefficients of the binary regression model (with logistic or probit link function) cannot be estimated when the x_i's of successes and failures don't overlap. This is a deficiency of the model itself, because the fit can be made perfect by letting $||\theta||$ tend to infinity. Therefore, this problem is shared by all reasonable estimators that operate under the logistic model.

Our approach to resolve this problem is to work with a generalized model, which we call the hidden logistic model. Here we compute the pseudo-observations $\tilde{y}_i$, defined as the probability that $y_i = 1$ conditional on the maximum likelihood estimate of the true status t_i. The resulting MEL estimator always exists and is unique, even though the hypothetical misclassification probabilities (based on our default setting $\delta = 1\%$) are so small that they would not be visible in the observed data.

The hidden logistic model was previously used (under a different name) in an important paper by Copas (1988). However, his approach and ours are almost diametrically opposite. Copas' motivation is to reduce the effect of the outliers that matter, which are the observations (x_i, y_i) where x_i is far away from the bulk of the data and y_i has the value which is very unlikely under the logistic model. In the terminology of Rousseeuw and van Zomeren (1990) these are bad leverage points. In logistic regression their effect is always to flatten the fit, i.e. to bring the estimated slopes close to zero. Copas' approach shrinks the logistic distribution function Λ away from 0 and 1 (by letting it range between γ and $1-\gamma$), so that bad leverage points are no longer that unlikely under his model, which greatly reduces their effect. On the other hand, his approach aggravates the problems that arise when there is little overlap between successes and failures, as in his analysis of the vaso constriction data.

Our approach goes into the other direction: rather than shrinking Λ while leaving the responses y_i unchanged, we leave Λ unchanged and shrink the y_i to the pseudo-observations $\tilde{y}_i$ which are slightly larger than zero or slightly less than 1. This completely eliminates the overlap problem. It does not help at all for the problem of bad leverage points, but for that problem we can use existing techniques from the robustness literature. For instance, for grouped data (i.e. tied x_i's) we saw in Section 6 that the fitting can be done by the LTS regression method, which is robust against leverage points.

In general, also other robust techniques can be applied to the $(x_i, \tilde{y}_i)$. For instance, note that the score function (6) is similar to an M-estimator equation. Since the (pseudo-)residual is always bounded due to

$$|\tilde{y}_i - \Lambda(x_i'\theta)| < 1$$

the main problem comes from the factor x_i which need not be bounded (this corresponds to the leverage point issue). A straightforward remedy is to downweight leverage points, yielding the weighted maximum estimated likelihood (WEMEL) estimator defined as the solution $\hat{\theta}$ of

$$\sum_{i=1}^{n} (\tilde{y}_i - \Lambda(x_i'\theta))\, w_i x_i = 0 \tag{11}$$

where the weights w_i only depend on how far away x_i is from the bulk of the data. For instance, we can put

$$w_i = \frac{M}{\max\{RD^2(x_i^*), M\}} \tag{12}$$

where $x_i^* = (x_{i,2}, \ldots, x_{i,p}) \in \mathbb{R}^{p-1}$, $RD(x_i^*)$ is its robust distance, and M is the 75th percentile of all $RD^2(x_j^*)$, $j = 1, \ldots, n$.

When all regressor variables are continuous and there are not more than (say) 30 of them, we can use the robust distances that come out of the minimum covariance determinant (MCD) estimator of Rousseeuw (1984), for which the fast algorithm of Rousseeuw and Van Driessen (1999a) is available. This algorithm has been incorporated in the packages S-Plus (as the function `cov.mcd`) and SAS/IML (as the routine `MCD`), and both provide the robust distances in their output. In case that not all regressor variables are continuous or there are very many of them (even more than one thousand), we can use the robust distances provided by the robust principal components algorithm of Hubert, Rousseeuw and Verboven (2001).

We have not yet studied the WEMEL estimator in any detail, but we note that it is easy to compute because most GLM algorithms (including the one in S-Plus) allow the user to input prior weights w_i.

We also have not yet addressed the issue of bias correction for either MEL or WEMEL, which is a subject for further research. It may be possible to apply the same type of calculus as for formula (27) of Copas (1988).

Last but not least are the computation of influence functions and breakdown values. It would be interesting to connect our work in the hidden logistic model with the existing body of literature on outlier detection and robust estimation in the classical logistic model, including the work of Pregibon (1982), Stefanski et al. (1986), Künsch et al. (1989), and Müller and Neykov (2000).

Acknowledgement

The second author was supported by the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures").

References

A. Albert, J. A. Anderson (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71, 1–10.

A. Christmann (1994a). Least median of weighted squares in logistic regression with large strata. *Biometrika*, 81, 413–417.

A. Christmann (1994b). *Ausreißeridentifikation im logistischen Regressionsmodell.* In: S.J. Pöppl, H.-G. Lipinski, T. Mansky (Eds.). Medizinische Informatik: Ein integrierender Teil arztunterstützender Technologien. MMV Medizin Verlag, München, pp. 478-481.

A. Christmann (1998). *On positive breakdown point estimators in regression models with discrete response variables.* Habilitation thesis, University of Dortmund, Department of Statistics.

A. Christmann, P. Fischer, and T. Joachims (2000). *Comparison between the regression depth method and the support vector machine to approximate the minimum number of misclassifications.* Technical report 53/00, University of Dortmund, SFB 475, submitted.

A. Christmann, P. J. Rousseeuw (2001). Measuring overlap in logistic regression. To appear in *Computational Statistics and Data Analysis.*

J.B. Copas (1988). Binary regression models for contaminated data. With discussion. *J. R. Statist. Soc.*, B, 50, 225–265.

B. Efron (1986). Double exponential families and their use in generalized linear regression. *J. Amer. Statist. Assoc.*, 81, 709–721.

A. Ekholm, J. Palmgren (1982). A model for binary response with misclassification. In: R. Gilchrist, editor, *GLIM-82: Proc. Int. Conf. Generalized Linear Models*, pp. 128–143, Springer: Heidelberg.

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

J. Hermans, J.D.F. Habbema (1975). Comparison of five methods to estimate posterior probabilities. *EDV in Medizin und Biologie*, 6, 14–19.

M. Hubert, P.J. Rousseeuw, and S. Verboven (2001). A fast method for robust principal components with applications to chemometrics. To appear in *Chemometrics and Intelligent Laboratory Systems.*

H.J. Jaeger, T. Mair, M. Geller, R.K. Kinne, A. Christmann, K.D. Mathias (1997). A physiologic in vitro model of the inferior vena cava with a computer-controlled flow system for testing of inferior vena cava filters. *Investigative Radiology*, 32, 511–522.

H.J. Jaeger, S. Kolb, T. Mair, M. Geller, A. Christmann, R.K. Kinne, K.D. Mathias (1998). In vitro model for the evaluation of inferior vena cava filters: effect of experimental parameters on thrombus-capturing efficacy of the Vena Tech-LGM Filter. *Journal of Vascular and Interventional Radiology*, 9, 295–304.

H. R. Künsch, L. A. Stefanski, and R. J. Carroll (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *J. Amer. Statist. Assoc.*, 84, 460–466.

E.T. Lee (1974). A computer program for linear logistic regression analysis. *Computer Programs in Biomedicine*, 80–92.

C. Müller, N.M. Neykov (2000). Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. Technical report.

D. Pregibon (1981). Logistic regression diagnostics. *Ann. Statist.*, 9, 705–724.

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

H. Riedwyl (1997). *Lineare Regression und Verwandtes.* Birkhäuser, Basel.

P.J. Rousseeuw (1984). Least median of squares regression. *J. Amer. Statist. Assoc.*, 79, 871–880.

P. J. Rousseeuw, M. Hubert (1999). Regression depth. *J. Amer. Statist. Assoc.*, 94, 388–433.

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

P.J. Rousseeuw, K. Van Driessen (1999b). Computing LTS regression for large data sets. Technical Report, University of Antwerp, submitted.

P.J. Rousseeuw, B.C. Van Zomeren (1990). Unmasking multivariate outliers and leverage points. *J. Amer. Statist. Assoc.*, 85, 663–651.

T. J. Santner, D. E. Duffy (1986). A note on A. Albert and J.A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 73, 755–758.

L.A. Stefanski, R.J. Carroll, and D. Ruppert (1986). Optimally bounded score functions for generalized linear models with applications to logistic regression. *Biometrika*, 73, 413–424.