Discussiones Mathematicae
Probability and Statistics 36 (2016) 43–51
doi:10.7151/dmps.1180

A NOTE ON ROBUST ESTIMATION IN LOGISTIC REGRESSION MODEL

Tadeusz Bednarski

Wroclaw University

e-mail: t.bednarski@prawo.uni.wroc.pl

Abstract

Computationally attractive Fisher consistent robust estimation methods based on adaptive explanatory variables trimming are proposed for the logistic regression model. Results of a Monte Carlo experiment and a real data analysis show its good behavior for moderate sample sizes. The method is applicable when some distributional information about explanatory variables is available.

Keywords: logistic model, robust estimation.

2010 Mathematics Subject Classification: 62F35, 62J12.

1. Introduction

The logistic model plays important role in predictive inference in many fields: diagnostics in medicine, behavioral prediction in social and economic processes. A more recent area of its interesting applications concern the uplift modeling in marketing. The binary response variable Y of the model satisfies the logistic equation

$$P_\beta(Y = 1|X = x) = \frac{e^{[1,x']\beta}}{1 + e^{[1,x']\beta}},$$

where X is a k-vector of explanatory variables and $\beta = [\beta_0, \beta_1, \ldots, \beta_k]$ is a vector of regression parameters. The logistic function, as well as the normal distribution in the probit model, is supposed to measure the relationship between the probability of the event of interest and the value of its linear predictor. This mathematical relationship does not seem to follow precisely from any real phenomena. It is rather a mixture of mathematical convenience and some intuitive

knowledge about a monotone regularity between probability of 'success' and explanatory variables.

As for any statistical model, inference based on maximum likelihood may be very sensitive to 'erroneous' observations and therefore supplementary robust data analysis is highly recommended. There is a vast literature on robust inference for the generalized linear models and in particular for the logistic one [1, 2, 3, 5, 8, 10, 11, 12]. The methods are either based on downweighting the influential observations or on modification of the likelihood function or both approaches are used. Some gain in inferential efficiency for the logistic (or probit) model can be achieved by adjustment of robust methodology to a particular experimental situation. In contrast to a standard linear regression model, where outliers in dependent variable alone can do much harm, this is unlikely to happen for the logistic model, where 'true' oulyingness is possible only in independent variable.

The estimation method proposed here for the logistic model is related to [9, 12] with weighting based exclusively on the design or explanatory variable. It assumes some prior knowledge about the distribution of the explanatory variables. In repeated experiments such knowledge is sometimes available and it either has a distributional character or is implied by, say, physical limitations of studied objects. Further on we shall assume that this prior knowledge would be an approximately normal distribution for all or some of the explanatory variables.

2. The method

If $(X_1, Y_1), \ldots, (X_n, Y_n)$ denote a random sample of explanatory and dependent variables then the likelihood function for the logistic model is equal to

$$L_{\beta(X,Y)} = \prod_{i=1}^{n} \left(\frac{e^{[1,X_i']\beta}}{1 + e^{[1,X_i']\beta}} \right)^{Y_i} \left(\frac{1}{1 + e^{[1,X_i']\beta}} \right)^{1-Y_i}$$

with the corresponding score function

$$S_\beta = \sum_{i=1}^{n} \begin{bmatrix} 1 \\ X_i \end{bmatrix} \left(Y_i - \frac{e^{[1,X_i']\beta}}{1 + e^{[1,X_i']\beta}} \right).$$

There are two basic modifications of the score function proposed here. In both of them it is initially assumed that the constant term in the linear predictor is equal to zero. The first modification has the form

$$S_\beta^w = \sum_{i=1}^{n} w(X_i) \left(Y_i - \frac{e^{w(X_i)'\beta}}{1 + e^{w(X_i)'\beta}} \right), \tag{1}$$

where $w(x) = xI_B(x)$ for B a Borel set. The proposed method is equivalent to the maximum likelihood estimation for the logistic regression for possibly modified observations in explanatory variables and under the assumption of zero shift in the linear predictor.

The second modification assumes the following form of the score function

$$S_\beta^w = \sum_{i=1}^{n} h(w(X_i)) \left(Y_i - \frac{e^{w(X_i)'\beta}}{1 + e^{w(X_i)'\beta}} \right). \tag{2}$$

Theorem 1. *Let us consider the estimation methods given by score functions* (1) *and* (2).

(a) *The data modifying function* $w(x) = xI_B(x)$ *for* B *a Borel set such that* $P(X \in B) > 0$ *gives the conditional Fisher consistency of method* 1.

(b) *Assume the distribution of* X *is symmetric around the origin,* w *is an odd function,* h *is even and for some* $\beta \neq \beta_0$ *we have* $E_{\beta_0}\left[h(w(X_i))(Y_i - \frac{e^{w(X_i)'\beta}}{1+e^{w(X_i)'\beta}})\right] \neq 0$. *Then method* 2 *is Fisher consistent.*

Proof. Justification of the two facts is elementary. In case (a) we have the following immediate equalities:

$$\begin{aligned}
&\int w(x) \left(y - \frac{e^{w(x)'\beta}}{1 + e^{w(x)'\beta}} \right) dP_{\beta_0}(y|x) dG(x) \\
&= \int w(x) \left(1 - \frac{e^{w(x)'\beta}}{1 + e^{w(x)'\beta}} \right) P_{\beta_0}(y = 1|x) \\
&\quad + w(x) \left(0 - \frac{e^{w(x)'\beta}}{1 + e^{w(x)'\beta}} \right) P_{\beta_0}(y = 0|x)] dG(x) \\
&= \int w(x) \frac{1}{1 + e^{w(x)'\beta}} \frac{e^{x'\beta_0}}{1 + e^{x'\beta_0}} - w(x) \frac{e^{w(x)'\beta}}{1 + e^{w(x)'\beta}} \frac{1}{1 + e^{x'\beta_0}} dG(x) \\
&= \int w(x) \frac{1}{(1 + e^{w(x)'\beta})(1 + e^{x'\beta_0})} \left(e^{x'\beta_0} - e^{w(x)'\beta} \right) dG(x),
\end{aligned}$$

where $dP_{\beta_0}(y|x)dG(x)$ denotes the true model distribution. By the definition of w we get the zero value of the expression under the integral if $\beta = \beta_0$.

The second fact follows then from the symmetry assumption and oddness of the pseudo-score function. Notice that

$$h(w(-x))\frac{1}{(1+e^{w(-x)'\beta})(1+e^{-x'\beta_0})}\left(e^{-x'\beta_0}-e^{w(-x)'\beta}\right)$$

$$=h(w(x))\frac{e^{w(x)'\beta}e^{x'\beta}}{(1+e^{w(x)'\beta})(1+e^{x'\beta_0})}\left(e^{-x'\beta_0}-e^{w(-x)'\beta}\right)$$

$$=-h(w(x))\frac{1}{(1+e^{w(x)'\beta})(1+e^{x'\beta_0})}\left(e^{x'\beta_0}-e^{w(x)'\beta}\right).$$

■

A drawback of the first method is a non-smooth structure of the data modifying function w. The second method allows smooth $w's$ which is a merit in estimation since the variance assessment becomes more reliable. Further on we shall discuss properties of the first method only. The method is meant to be an intermediate step in any practical implementation of a standard logistic model (with arbitrary shift) when the distribution of explanatory variables is approximately normal. In the first step the original data are robustly standardized with a consequent reparametrization of the model. Then the maximum likelihood method with standardized data modified by $w(x) = xI_B(x)$ is applied. The final step consists in returning to the original parametrization.

3. Monte Carlo and real data case study

A simulation study was conducted to compare efficiency of the above described simplified robust estimation procedure (1) with mle and some standard robust estimates. A logistic model with two dimensional explanatory vector of independent standard normal variables $X = (X_1, X_2)$ was assumed so that

$$P_\beta(Y = 1|X = x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}$$

and $\beta = (\beta_0, \beta_1, \beta_2) = (0, 1, 1)$. The sample size was taken $n = 300$. In contaminated samples 5% of the original standard normal vector observations were replaced by the vectors $(0, -k)$ where for each k taking values $1, 2, \ldots, 7$ the estimation process was simulated 500 times for the original and contaminated data. This sort of contamination was chosen because it strongly exhibit efficiency differences. The mle was used for the original data and for the contaminated data to give a reference point. The contaminated data were also estimated with standard robust methods for the logistic model and the proposed robust method.

The data simulation and estimation process was conducted with R software. The maximum likelihood estimation was done with glm function while standard robust estimation with glmRob function of the 'robust' package. Two robust

methods were applied: 'mallows' for Mallow's leverage downweighting estimator and 'cubif' for the conditionally unbiased bounded influence estimator. For the simplified method the mle was applied to transformed explanatory variables: in the first step robust location and covariance matrix were computed using covRob procedure with constrained M estimator. Then for standardized data the mle was applied under zero value of the shift and $B = [-3, 3]^{\times 3}$. Finally using the obtained robust estimates of shift and covariance matrix final estimates were obtained.

Figure 1 gives two graphs summarizing the estimation effects. The one on the left-hand side shows plots of the mean Euclidian distance between estimates and the vector $\beta = (0, 1, 1)$ at different contaminating values k. The one on the right-hand side is the mean angle formed by the vector of estimates $(\hat{\beta}_1, \hat{\beta}_2)$ and $(1, 1)$. In each case there are four plotted curves: the mle for non-contaminated, the mle for contaminated samples, robust 'mallows' estimate for contaminated data and finally simplified estimation for contaminated samples. The mean distance and angle of estimators with respect to the true parameter value were employed as stability characteristics since they may behave differently depending on the data distribution.

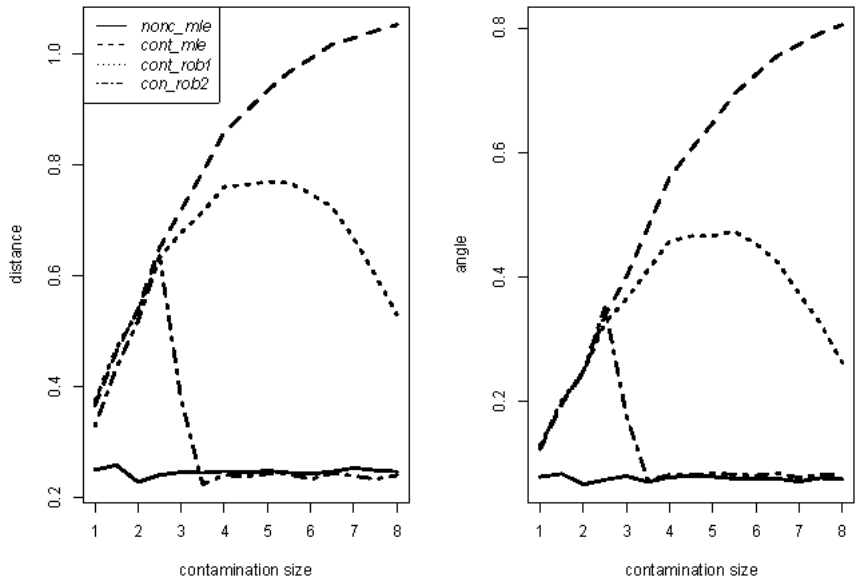

Figure 1. Mean distance (left graph) and angle (right graph) for the mle, standard robust (rob1) and simplified robust method (rob2).

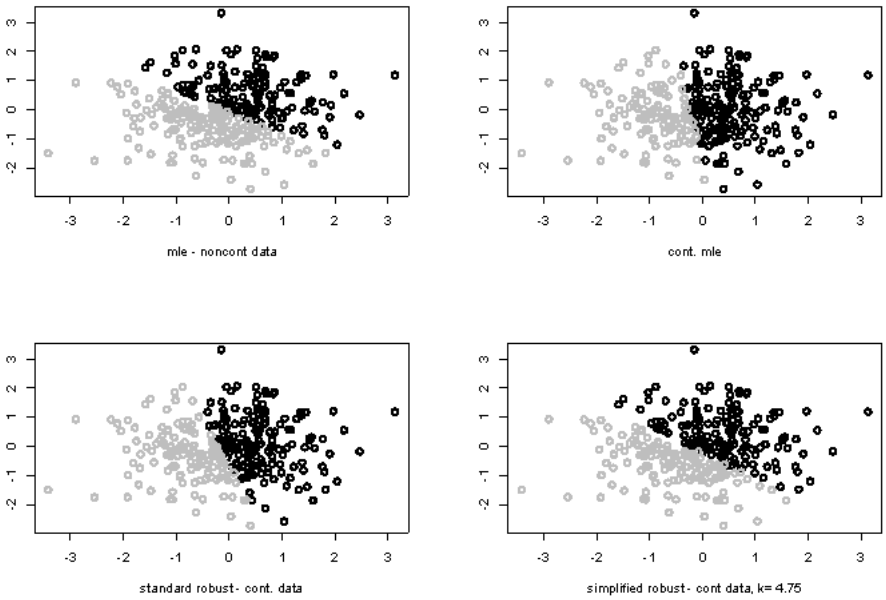

Figure 2. Exemplary result of classification.

In spite of the fact that the simplified method 'completely ignores' the information from outlying observations, we can see its good behavior compared to other estimation methods. The 'cubif' procedure showed similar results. Also changes in sample size from 100 to 500 did not essentially changed the situation. We can see that as the contamination values increase, the standard robust methods start to work better. However, they seem to achieve the efficiency of the simple method at extremely large outliers values, detectable probably by any method. The procedures were used in their default form, therefore there might be a potential for their better behavior.

The following graph depicts assignment of sample units to two classes depending on the estimated success probability value, smaller or greater than 0.5. To make the visual comparison possible only the non-contaminated data are plotted. The black dots correspond to sample elements for which estimated success probability was greater than 0.5. We can see that even for relatively small contamination size ($k = 4.75$) the angular effect differs very much for the presented methods. The simplified method is very stable and it gives results similar to mle for non-contaminated data. These classification results would be rather typical for contamination size k between 4 and 7.

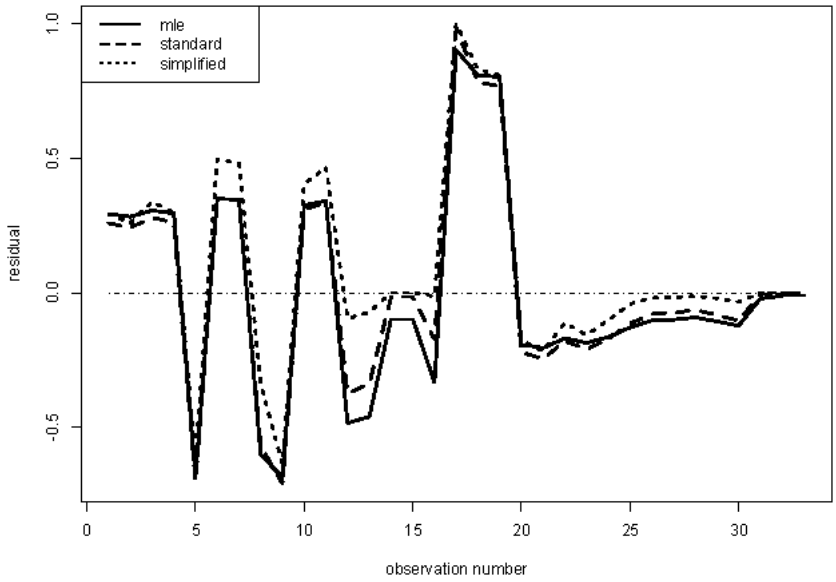

Figure 3. Differences between dependent variable and estimated probabilities for the Feigl and Zelen (1965) data.

The real data case analyzed below originates from [6] and it was discussed in [4]. The data are available in R as BGPhazard package. There are four variables of 33 leukemia patients: time – survival times in weeks from diagnosis, AG – indicator of positive result of a test related to the white blood cell characteristics, wbc – white blood cells counts in thousands and delta – a status variable indicating censoring. As in [4] the dependent variable indicating survival longer than 52 weeks was introduced and logistic model was applied with explanatory variables AG and wbc. The mle, standard robust and simplified robust methods were employed. The inferential problem for this data set is that five patients extreme values of wbc = 100 lead to contradictory predictions about survival time and AG. The trimming value for the explanatory variable in the simplified method was 69.7 – the mean plus 3 standard deviations computed without the outlying values of wbc = 100. The three methods indicate significance of AG variable. The simplified method gave the smallest p-value to wbc $p = 0.079$ (0.082 for standard robust and 0.088 for mle) and it led to smallest null and residual deviance. The graph below shows the residual curves for the three methods.

The simplified method behaves very well for unimodal elliptically symmetric distributions close to the normal one. Simulations show that the variability of

the method is comparable to other robust methods for the logistic model except for the case when outliers of robustly standardized data are around the points of discontinuity of the function w, where the simplified method shows higher variability. The method also looses its advantage when the distribution of the vector of explanatory variables is for instance uniformly distributed or it comes from an asymmetric distribution.

The asymptotic distributions of the proposed estimators are relatively straightforward under the assumption that explanatory variables come indeed from the standard normal population. Otherwise they become more difficult to compute due to transformations of observed variables by robust estimates of shift and covariance. A description of the asymptotic distributions along with their validity for moderate sample sizes will be given in a subsequent paper.

REFERENCES

[1] A.M. Bianco and E. Martinez, *Robust testing in the logistic regression model*, Computational Statistics and Data Analysis **53** (2009) 4095–4105. doi:10.1016/j.csda.2009.04.015

[2] A.M. Bianco and V.J. Yohai, *Robust estimation in the logistic regression model*, Lecture Notes in Statistics, Springer Verlag, New York **109** (1996) 17–34.

[3] E. Cantoni and E. Ronchetti, *Robust inference for generalized linear models*, Journal of the American Statistical Association **96** (2001) 1022–1030. doi:10.1198/016214501753209004

[4] R.D. Cook and S. Weisberg, Residuals and Influence in Regression (Chapman and Hall, London, 1982).

[5] C. Croux, G. Haesbroeck and K. Joossens, *Logistic discrimination using robust estimators: An influence function approach*, Canadian J. Statist. **36** (2008) 157–174. doi:10.1002/cjs.5550360114

[6] P. Feigl and M. Zelen, *Estimation of exponential probabilities with concomitant information*, Biometrics **21** (1965) 826–38. doi:10.2307/2528247

[7] D.J. Finney, *The estimation from individual records of the relationship between dose and quantal response*, Biometrika **34** (1947) 320–334. doi:10.1093/biomet/34.3-4.320, doi:10.2307/2332443

[8] H.R. Kunsch, L.A. Stefanski and R.J. Carroll, *Conditionally Unbiased Bounded Influence Estimation in General Regression Models, with Applications to Generalized Linear Models*, J. Amer. Statist. Assoc. **84** (1989) 460-466. doi:10.2307/2289930

[9] C.L. Mallows, On some topics in robustness (Tech. Report, Bell Laboratories, Murray Hill, NY, 1975).

[10] S. Morgenthaler, *Least-absolute-deviations fits for generalized linear model*, Biometrika **79** (1992) 747–754. doi:10.1093/biomet/79.4.747

[11] D. Pregibon, *Resistant Fits for some commonly used Logistic Models with Medical Applications*, Biometrics **38** (1982) 485–498. doi:10.2307/2530463

[12] L. Stefanski, R. Carroll and D. Ruppert, *Optimally bounded score functions for generalized linear models with applications to logistic regression*, Biometrika **73** (1986) 413–424. doi:10.2307/2336218

Received 21 December 2015