Robustness against separation and outliers in binary regression

P. Rousseeuw[1] and A. Christmann

[1] Universitaire Instelling Antwerpen, Universiteitsplein 1, B-2610 Antwerpen, Belgium

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 act as a hidden layer in a neural net. We propose the *maximum estimated likelihood* method in this model, which is robust against separation unlike all existing methods for logistic regression. We then construct an outlier-robust modification of this estimator, called the *weighted maximum estimated likelihood* (WEMEL) method, which is robust against both problems.

Motivation

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).

It is possible 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 (2002). However, when we know that there is no overlap we still have to solve the identifiability problem.

Here we will 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 such situations 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 response plays the role of a hidden layer in a stochastic neural network, which is why we call this the hidden logistic regression model.
In this model we propose the maximum estimated likelihood (MEL) technique, and show that it is immune to the identification problem described above. Furthermore, we construct an outlier-robust estimator in this setting, the WEMEL method. We then study the behavior of the MEL and WEMEL estimators on real and simulated data.

References

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

A. Christmann, P. Fischer, and T. Joachims (2002).
Comparison between the regression depth method and the support vector machine to approximate the minimum number of misclassifications.
To appear in: *Computational Statistics*, 2.

A. Christmann, P.J. Rousseeuw (2001).
Measuring overlap in logistic regression.
Computational Statistics and Data Analysis, 37, 65–75.

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.

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

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.

Please fill in this form and mail it together with your abstract.

My abstract fits best to topic number 3 or 20.

List of Topics:

1. Algorithms
2. Applications
3. Biostatistics $\Longleftarrow$
4. Computing and graphics
5. Data analysis
6. Data mining
7. Economics, finance
8. Efficiency and robustness
9. Functionals and bias
10. Fuzzy statistics
11. Geostatistics
12. Inference for robust methods, model testing
13. Location depth and regression depth
14. Multivariate methods
15. Neural networks
16. Rank-based methods
17. Regression quantiles, trimming
18. Robust covariance
19. Robust designs
20. Robust regression $\Longleftarrow$
21. Time series analysis
22. Wavelets
23. Other (please specify)