Positive breakdown point estimation in categorical regression models

Andreas Christmann
University of Dortmund, HRZ and SFB 475, D-44221 Dortmund, GERMANY

1. Introduction

Positive breakdown point estimators for regresssion models with discrete response variables have not received the same attention as robust estimators for linear regression models. It will be shown that there exist strongly consistent positive breakdown point estimators in many regression models, if large strata assumptions are valid. Poisson regression and logistic regression are treated as important special cases. The idea is to transform the original data set by an appropriate transformation to an approximately linear regression model with approximately normal errors, and then use the least median of squares estimator LMS or the least trimmed squares estimator LTS, both proposed by Rousseeuw (1984), for the transformed data set to estimate the unknown parameter vector.

2. Results

Let n and $m_1, \ldots, m_n$ be positive integers, where n is fixed. Let $\mathbf{X} = (x_1, \ldots, x_n)^T \in \mathbb{R}^{n \times p}$ be a design matrix with full column rank $p < n$. Let Y_{i,m_i}, $1 \le i \le n$, be random variables on the measurable space $(\Omega, \mathcal{A})$, where Ω is either a subset of nonnegative integers and $\mathcal{A}$ the set of all subsets of Ω, or $\Omega = \mathbb{R}$ and $\mathcal{A}$ the Borel-$\sigma-$algebra. Define $P_{i,m_i} = Y_{i,m_i}/m_i$, $1 \le i \le n$, and $h_0 = [\![n/2]\!] + [\![(p+1)/2]\!]$, where $[\![r]\!]$ stands for the largest integer less than or equal to r. Let $h \in \{h_0, \ldots, n\}$ be fixed. Of special importance is $h = h_0$ or $h \approx [\![0.75n]\!]$. Let $g : \mathbb{R} \to A \subseteq \mathbb{R}$ be a strictly increasing and continuously differentiable function, such that $g^*(u) = (d/du)g^{-1}(u) \in (0, \infty)$, $u \in A$. Let $v : \mathbb{R} \to (0, \infty)$ be a continuous function. We consider parametric regression models H_0 with the following properties. The random variables $Y_{1,m_1}, \ldots, Y_{n,m_n}$ are independent with cumulative distribution function F_{i,m_i} such that

$$P_{i,m_i} \xrightarrow{\text{wp1}} g(x_i^T\theta)\,, \quad m_i^{1/2}\,\frac{P_{i,m_i} - g(x_i^T\theta)}{v(x_i^T\theta)} \xrightarrow{\mathcal{D}} \mathrm{N}(0,1)\,, \quad \min_{1\le i\le n}(m_i) \to \infty, \tag{1}$$

where $\theta \in \mathbb{R}^p$. Define

$$\tilde{Y}_{i,m_i} = m_i^{1/2}\frac{g^{-1}(P_{i,m_i})}{g^*(P_{i,m_i})v(g^{-1}(P_{i,m_i}))}\,, \quad \tilde{X}_{i,m_i} = m_i^{1/2}\frac{1}{g^*(P_{i,m_i})v(g^{-1}(P_{i,m_i}))}\,x_i \quad 1 \le i \le n. \tag{2}$$

Under H_0 the transformed response vector $(\tilde{Y}_{1,m_1}, \ldots, \tilde{Y}_{n,m_n})^T$ follows approximately a linear regression model with stochastic design matrix $\tilde{\mathbf{X}} = (\tilde{X}_{1,m_1}, \ldots, \tilde{X}_{n,m_n})^T$, which depends on Y_{i,m_i}, $1 \le i \le n$, $\min_i(m_i) \to \infty$. In Table 1 the functions g, g^{-1}, g^*, and v for two important parametric models H_0 are specified: logistic regression and Poisson regression.

Table 1: Examples of parametric models H_0

$P_{Y_{i,m_i}}$	A	$g(u)$	$g^{-1}(u)$	$g^*(u)$	$v(u)$
$\mathrm{Bi}(m_i, \Lambda(x_i^T\theta))$	$(0,1)$	$\Lambda(u)$	$\mathrm{logit}(u)$	$[u(1-u)]^{-1}$	$[\Lambda(u)(1-\Lambda(u))]^{1/2}$
$\mathrm{Poi}(m_i \exp[x_i^T\theta])$	$(0,\infty)$	$\exp(u)$	$\log(u)$	$1/u$	$[\exp(u)]^{1/2}$

The least median of weighted squares estimator (LMWS) and the least trimmed weighted squares estimator (LTWS) are defined as the least median of squares estimator LMS and the least trimmed squares estimator LTS, both proposed by Rousseeuw (1984), for the *transformed* data set to estimate the unknown parameter vector, c.f. Christmann (1994,1998). This approach has several advantages. It can be shown that LMWS and LTWS inherit the positive breakdown point and exact fit property from LMS and LTS. Further, LMWS and LTWS are strongly consistent under a large supermodel of H_0 which even allows some kind of over- or underdispersion, the occurrence of multiple outliers and violations of the assumption that all response variables are independent. There exist software to compute such estimates and corresponding robust residual plots, e.g. S-PLUS. An S-PLUS function for LMWS and LTWS can be downloaded from StatLib (`http://lib.stat.cmu.edu/S/hbdp`). For grouped data the dimensions of the design matrix are often small to moderate such that the least median of weighted squares estimator can sometimes be computed exactly even for very large groups. These estimators can be useful as starting vectors for other robust estimators and can be helpful to find outliers and leverage points. The behaviour of both estimators is compared with the behaviour of the ML-estimator and of M-estimators proposed by Künsch, Stefanski and Carroll (1989) in a simulation study under a logistic regression model and under a logistic regression model with approximately 25% moderate or 25% extreme α_n-outliers in the sense of Davies and Gather (1993). In the situations considered in the simulations LMWS and LTWS have a much smaller bias and mean squared error than the ML-estimator or M-estimators if there are extreme α_n-outliers. The application of both estimators is illustrated using a data set from a multi-disciplinary project, c.f. Jaeger et al. (1997). The aim is to detect risk factors for the capture rate of thrombi of inferior vena cava filters.

REFERENCES

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

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

Jaeger, H.J., Mair, T., Geller, M., Kinne, R.K., Christmann, A., Mathias, K.D. (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.

Rousseeuw, P.J. (1984). Least Median of Squares Regression. J. Amer. Statist. Assoc., 79, 871-880.

RÉSUMÉ

Deux estimateurs convergents avec un point de rupture positive ont eté étudié pour des modéles de régression a variable dépendante catégorielle. La régression logistique et la régression de Poisson sont considérés plus en detail. Le comportement pour des tailles finies est étudié par une étude de simulation. L'application des deux estimateurs est illustée dans deux examples du science médical.