ROBUST ESTIMATION OF QUALITATIVE RESPONSE REGRESSION MODELS

Alexander A. Kalinin, Daniil V. Lisitsin
Novosibirsk State Technical University
Novosibirsk, Russian Federation
e-mail: kalinin.a.letters@gmail.com

Abstract

Qualitative response regression models such as logistic regression are typically estimated by the maximum likelihood method. To improve its robustness, two special cases of the M-estimation based approach for quantitative continuous random variables were extended to the variant of qualitative and mixed variables modeling. Expressions of the score functions for polytomous regression models were derived. In according to results of the research some conclusions and practical recommendations were given.

Keywords: qualitative response, Bayesian dot contamination, polytomous regression, robust estimation, influence function

Introduction

The classical statistic procedures are based on a number of assumptions which can't be fulfilled in practice. Under such conditions a lot of widespread statistic procedures lose their positive qualities. For instance, the procedures, which rest on the maximum likelihood method. But this problem can be solved by using robust estimators. The general robust theory is developed in Huber [7] and Hampel, Ronchetti, Rousseeuw, and Stahel [6]. Recent work describing robust statistics in detail is Maronna et al. [10]. Generally robustness theory has been developed for the quantitative continuous random variables modeling. Qualitative and mixed variables modeling are paid much less attention. Several authors have studied the logistic regression model in terms of the robustness properties of the maximum likelihood estimation (MLE) and it's modifications. The maximum likelihood estimator attains the minimum asymptotic variance under the model and then it is optimal, but it is very sensitive to atypical data. Observations with extreme covariates, in particular, have a large influence on the estimator, and if they are accompanied by misclassified responses, the resulting estimates can be seriously biased. Pregibon (see [11]) made the earliest systematic attempts to fix this problem; he proposed methods to unmask influential observations and robust estimators for the logistic model. Later robust proposals in this area include Carroll and Pederson [2], Bianco and Yohai [1], Croux and Haesbroeck [4], and Gervini [5]. Typically, in these works binary regression models are considered. Many approaches for binary choice estimators development were introduced as alternatives to the maximum likelihood estimators, but they often are of semi-heuristic nature. Also note, that the numeric character of a binary variable is assumed in many papers. Recent examples include Victoria-Feser [13], Čížek [3], and Kotlyarova and Zinde-Walsh [8]. All these estimators differ greatly in terms of outlier resistance and efficiency under the model.

The one of the most perspective approaches was suggested by Shurygin in [7] (see also [9]). Shurygin's approach based on Bayesian dot contamination of model distribution allows to get the estimators possessing a high robustness and efficiency. Originally the estimators within Shurygin's approach were formed only for continuous random variables models. However the theory developed in [12, 9] can be easily extended to the cases of scalar qualitative or count and vector mixed response models, where the latter consists of qualitative polytomous and quantitative responses. Qualitative polytomous (multinomial) response can be nominal or ordinal. In the latter case, one uses cumulative link model, continuation ratio model, stereotype model, and others. So, the purpose of this study is to develop a general theory of robust estimation for regression models with polytomous response and its application to the case of the nominal response.

1 Model Specification

Assume that discrete random variable Z has a fixed number of acceptable values $\{1, 2, ..., J\}$. Distribution of Z_t under observation t is set of model probabilities

$$P\left\{Z_t = j|x_t, \alpha\right\} = \pi_j\left(x_t, \alpha\right), t = 1, ..., N,$$

where x_t is a vector of covariates, α is a vector of parameters.

M-estimation $\hat{\alpha}$ of vector of parameters α is obtained by solving equations system

$$\sum_{t=1}^{N} \Psi(Z_t, x_t, \hat{\alpha}) = 0, \tag{1}$$

where $\Psi(Z_t, x_t, \hat{\alpha})$ is a vector score function satisfying further condition for all t

$$\sum_{j=1}^{J} \pi_j(x_t, \alpha)\Psi(j, x_t, \alpha) = 0. \tag{2}$$

2 Robust Estimation

One of the major indicators of estimator's robustness is an influence function which in the case under some regularity conditions takes the form

$$IF(Z, x, \alpha) = M^{-1}\Psi(Z, x, \alpha), \tag{3}$$

where

$$M = \sum_{t=1}^{N}\sum_{j=1}^{J} \Psi(j, x_t, \alpha)\frac{\partial}{\partial \alpha^T}\pi_j(x_t, \alpha).$$

In the Bayesian dot contamination model the distribution of Z_t is defined by the set of probabilities

$$P\left\{Z_t = j|x_t, \alpha, Z_t^*\right\} = (1 - \varepsilon)\pi_j\left(x_t, \alpha\right) + \varepsilon\delta_{jZ_t^*},$$

where Z_t^* is discrete random variable with fixed number of acceptable values $\{1, 2, ..., J\}$ and distribution $P\{Z_t^* = j|x_t, \alpha, Z_t^*\} = s_j(x_t, \alpha)$, ε is contamination level ($0 < \varepsilon < 0.5$), δ is Kronecker delta.

Indicator of estimation badness in Bayesian dot contamination model can be written as functional

$$U_t(\Psi) = \sum_{t=1}^{N} \sum_{j=1}^{J} IF(j, x_t, \alpha) IF^T(j, x_t, \alpha) s_j(x_t, \alpha). \tag{4}$$

Corresponding optimum score function in Bayesian dot contamination model is of the form represented

$$\Psi(Z, x, \alpha) = C \left[\frac{\partial}{\partial \alpha} \ln \pi_Z(x, \alpha) + \beta \right] \frac{\pi_Z(x, \alpha)}{s_Z(x, \alpha)} =$$

$$= C \sum_{j=1}^{J} \frac{\partial}{\partial \alpha} \ln \pi_j(x, \alpha) \left[\delta_{jZ} - \frac{\pi_j^2(x, \alpha)/s_j(x, \alpha)}{\sum_{l=1}^{J} \pi_l^2(x, \alpha)/s_l(x, \alpha)} \right] \frac{\pi_Z(x, \alpha)}{s_Z(x, \alpha)}, \tag{5}$$

where C is nonsingular matrix, vector $\beta = \beta(x, \alpha)$ provides fulfillment of the condition (2).

2.1 Generalized Radical Estimator

Generalized radical estimation (GRE) corresponds to the case:

$$s_j(x, \alpha) = [\pi_j(x, \alpha)]^{1-\lambda} / \Delta(x, \alpha, \lambda),$$

where λ is estimator parameter ($\lambda \geq 0$), value of $\Delta(x_t, \alpha, \lambda)$ either equals $\sum\limits_{l=1}^{J} [\pi_l(x_t, \alpha)]^{1-\lambda}$ (used for satisfying probabilities normalizing condition) or is identity, if that condition is not used. Note that the case of $\lambda = 0$ matches maximum likelihood estimation.

For modeling dependence of nominal response from covariates polytomous logistic regression is often used. Corresponding probabilities are of the form

$$\pi_j(x_t, \alpha) = exp\left[\Phi(x_t)\alpha_j)\right] \left\{ 1 + \sum_{k=1}^{J-1} exp\left[\Phi(x_t)\alpha_k\right] \right\}^{-1}, \tag{6}$$

where $\Phi(x_t)$ is a vector of regressors, α_j is a subvector of α (subvectors α_j, $j = 1, 2, ..., J-1$, are not intersected), α_J is a null vector.

Generalized radical estimation of subvector α_j in polytomous logistic regression model is defined by the score function

$$\Psi_j(Z_t, x_t, \alpha) = \left\{ \delta_{jZ_t} - \frac{[\pi_j(x_t, \alpha)]^{1+\lambda}}{\sum\limits_{l=1}^{J} [\pi_j(x_t, \alpha)]^{1+\lambda}} \right\} [\pi_{Z_t}(x_t, \alpha)]^{\lambda} \Delta(x_t, \alpha, \lambda) \Phi^T(x_t). \tag{7}$$

2.2 Conditionally Optimal Estimator

In the set of robust estimators also can be used estimation with the optimum score function in Bayesian dot contamination model given by

$$\Psi(Z,x,\alpha)=C\left[\frac{\partial}{\partial\alpha}\ln\pi_Z(x,\alpha)+\beta\right]\frac{1}{1+\dfrac{k^2}{\pi_Z(x,\alpha)}},$$

where k^2 is estimator parameter and C,β are the same as in (5).

To obtain conditionally optimal estimator, assume that distribution of Z_t^* is given by

$$s_j(x_t,\alpha)=\pi_j(x_t,\alpha)+k^2.$$

Hence, taking into account (6) the score function for conditional optimal estimation in polytomous logistic regression model is of the folowing form

$$\Psi_j(Z_t,x_t,\alpha)=\left\{\delta_{jZ_t}-\frac{\dfrac{\pi_j^2(x_t,\alpha)}{\pi_j(x_t,\alpha)+k^2}}{\sum\limits_{l=1}^{J}\dfrac{\pi_l^2(x_t,\alpha)}{\pi_l(x_t,\alpha)+k^2}}\right\}\frac{1}{1+\dfrac{k^2}{\pi_{Z_t}(x_t,\alpha)}}\Phi^T(x_t). \tag{8}$$

3 Experimental Research

In practice, there may be several solutions of equation system (1). Thus some methods of selection solutions are should to be used. Also it is necessary during solving to distinguish between consistent and inconsistent solutions and leave the latter out.

As a check of working capacity of proposed approaches was performed experimental research of generalized radical estimation of polytomous logistic regression model with nominal response having three levels. Maximum likelihood and robust estimators were compared under following values of estimators' parameters and model's parameters α. Vector of regressors is of the form $[1,x,x^2]$. True values of model's parameters are $\alpha_1=[-8,2,1]$ and $\alpha_2=[-5,4,1]$. The number of observations is 1000, the values of x are uniformly distributed on $[-10,10]$. The response has contaminated distribution with level $\varepsilon=0.05$. Contamination also has uniform distribution. And the parameter of the generalized radical estimator has the value $\lambda=1$ (this case is equivalent to conditionally optimal estimator with parameter $k^2=\infty$) The results of MLE are $\hat{\alpha}_1=[-3.45433123701182,0.149008965849041,0.214074248534943]$ and $\hat{\alpha}_2=[-1.3719459314693,0.907693561128312,0.149007714343864]$. And corresponding results of GRE are $\hat{\alpha}_1=[-11.8998482654318,-1.09199066567912,0.432142103840774]$ and $\hat{\alpha}_2=[-7.95057570361495,7.29031267939047,1.15240358346135]$.

Figure 1 provide us MLE-estimated probabilities dependence on the covariates. True probabilities are presented by black lines and estimated probabilities by grey. Solid lines correspond to the value of the response $j=1$, dashed lines to the value $j=3$ and dash-dot lines to the value $j=2$.

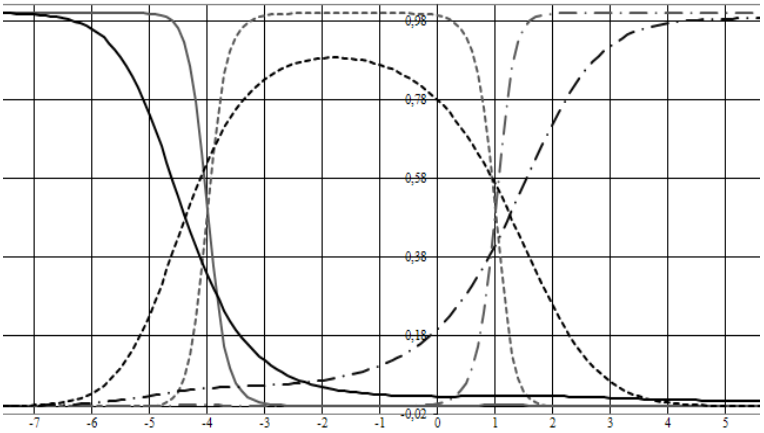

Figure 1: Probabilities estimated by MLE

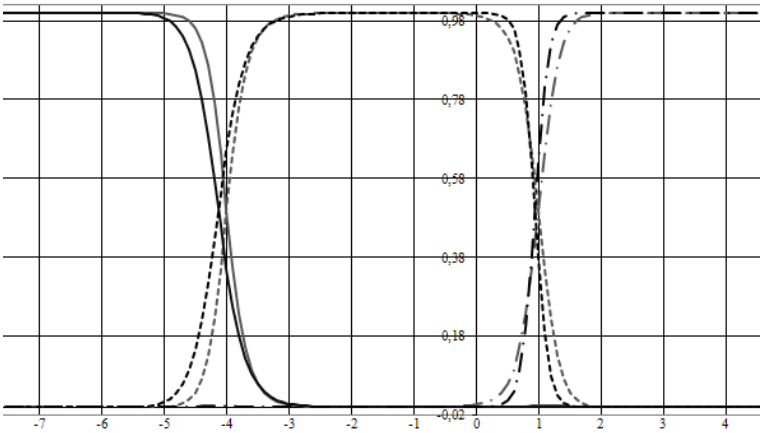

Figure 2: Probabilities estimated by GRE

Figure 2 provide us GRE-estimated probabilities dependence on the covariates. Designations for this figure are the same as for Figure 1.

As the results of the study, robust estimate is less affected by contamination than the MLE estimate. Although robust estimation of parameters quite substantially differ from the true values, dependences of the estimated probabilities are close enough to the true. Hence it is obvious that generalized radical estimator shows more accurate results of probabilities estimation than maximum likelihood estimator.

Conclusions

Due to the results of the research we conclude that:

- the proposed robust method is effective when level of contamination is not too high;
- it is often necessary to use more robust estimator for obtaining good results;
- high quality of estimation requires a great number of observations;
- whereas methods of estimation are sensitive to initial point it is essential to develop special techniques for obtaining good initial approximation.

Acknowledgements

Research work has been done with a partial support of President of Russian Federation grant (№ МД-2690.2008.9).

References

[1] Bianco A.M., Yohai V.J. (1996). Robust estimation in the logistic regression model. *Robust Statistics, Data Analysis, and Computer Intensive Methods, Lecture Notes in Statistics*, Vol. **109**, pp. 17-34.

[2] Carroll R.J., Pederson S. (1993). On robustness in the logistic regression model. *Journal of the Royal Statistical Society*, Vol. **B 55**, pp. 693-706.

[3] Čížek P. (2006). Trimmed likelihood-based estimation in binary regression models, *Austrian Journal of Statistics*, Vol. **35**, pp. 223-232

[4] Croux C., Haesbroeck G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics and Data Analysis*, Vol. **44**, pp. 273-295.

[5] Gervini D. (2005). Robust adaptive estimators for binary regression models. *Journal of Statistical Planning and Inference*, Vol. **131**, pp. 297-311.

[6] Hampel F.R., Rouchetti E.M., Rousseeuw P.J., Stahel W.A. (1986). *Robust Statistics: The approach based on influence functions*. Wiley, New York.

[7] Huber P.J. (1981). *Robust Statistics*. Wiley, New York.

[8] Kotlyarova Y., Zinde-Walsh V. (2010). Robust estimation in binary choice models. *Communications in Statistics - Theory and Methods*, Vol. **39:2**, pp. 266-279.

[9] Lisitsin D.V. (2009).On estimation of model parameters in presence of Bayesian dot contamination. *Reports of Russian higher education academy of sciences*. № **1(12)**, pp. 41-55 (in Russian).

[10] Maronna R.A., Martin R.D., Yohai V.J. (2006). *Robust Statistics: Theory and Methods*. Wiley, England.

[11] Pregibon D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics*, Vol. **38**, pp. 485-498.

[12] Shurygin A.M. (2009). *Mathematical methods of prediction*. Goryachaya Liniya - Telecom, Moscow (in Russian).

[13] Victoria-Feser M.-P. (2002). Robust inference with binary data. *Psychometrika*, Vol. **67**, pp. 21-32.