Sakate and Kashid *Environ Syst Res (2016) 5:9*
DOI 10.1186/s40068-016-0060-7

Environmental Systems Research
a SpringerOpen Journal

RESEARCH Open Access

A new robust model selection method in GLM with application to ecological data

D. M. Sakate* and D. N. Kashid

Abstract

Background: Generalized linear models (GLM) are widely used to model social, medical and ecological data. Choosing predictors for building a good GLM is a widely studied problem. Likelihood based procedures like Akaike Information criterion and Bayes Information Criterion are usually used for model selection in GLM. The non-robustness property of likelihood based procedures in the presence of outliers or deviation from assumed distribution of response is widely studied in the literature.

Results: The deviance based criterion (DBC) is modified to define a robust and consistent model selection criterion called robust deviance based criterion (RDBC). Further, bootstrap version of RDBC is also proposed. A simulation study is performed to compare proposed model selection criterion with the existing one. It indicates that the performance of proposed criteria is compatible with the existing one. A key advantage of the proposed criterion is that it is very simple to compute.

Conclusions: The proposed model selection criterion is applied to *arboreal marsupials* data and model selection is carried out. The proposed criterion can be applied to data from any discipline mitigating the effect of outliers or deviation from the assumption of distribution of response. It can be implemented in any statistical software. In this article, R software is used for the computations.

Keywords: Arboreal marsupials, Bootstrap, DBC, Robust estimation

Background

In the last two decades, generalized linear models (GLM) have emerged as a useful tool to develop models from ecological data to explain the nature of ecological phenomena. GLM encompass a wide range of nature of response variable like 'presence-absence' and 'count'. It can also be used to estimate the survivorship as can be seen in the conservation literature. GLM builds a predictive model for a response variable based on the predictors. Given a data on response and predictors, the model is fitted using maximum likelihood estimates (MLE) of the unknown regression coefficients. Under certain regularity conditions, the MLE is consistent asymptotic normal estimator of regression coefficients in GLM (McCullagh and Nelder 1989). In the presence of over dispersion, maximum quasi-likelihood estimation (MQLE) (Wedderburn 1974; McCullagh and Nelder 1989; Heyde 1997) is a popular estimation method. In the process of model building, the researcher may be confronted to a pool of predictors of which some might be redundant in nature. If such predictors are included in the model, the response will be predicted with less accuracy. The fitted GLM may contain some predictors which are redundant in nature and are required to be eliminated from the model based on the observed data.

In the linear regression set up, Murtaugh (2009) evaluated the prediction power of various variable selection methods for ecological and environmental data sets. GLM is a wider class of models with linear regression as a particular case when distribution of response is normal. In GLM, there are many methods available in the literature for variable selection. When the likelihood is known, Akaike information criterion (AIC) (Akaike 1974), Bayes information criterion (BIC) (Akaike 1978) and distribution function criterion (DFC) (Sakate and

*Correspondence: dms.stats@gmail.com
Department of Statistics, Shivaji University, Kolhapur, MS, India

Kashid 2013) find applications. Sakate and Kashid (2014) proposed a deviance based criterion (DBC) for model selection in GLM which uses MLE of parameters. BIC and DBC are consistent model selection criteria while AIC is not. Sakate and Kashid (2014) empirically established the superiority of DBC over BIC. They also showed that DBC performs better than $\bar{R}^2$ proposed by Hu and Shao (2008).

In practice, the data available for fitting a GLM may be contaminated and the MLE fit of the GLM may not be appropriate. In fact, both MLE and MQLE share the same non robustness property against contamination. Non robustness of MLE in the GLM is extensively studied in the literature (Pregibon 1982; Stefanski et al. 1986; Künsch et al. 1989; Morgenthaler 1992; Ruckstuhl and Welsh 2001). Hence, the use of MLE or MQLE in the presence of contaminated data may give misleading results. The non-robustness of MLE to contamination results in non-robustness of AIC, BIC, DFC and DBC. Hence, using MLE based model selection criterion in presence of contaminated data may be erroneous.

To overcome the problem of contamination in GLM, Cantoni and Ronchetti (2001) introduced robust estimation of regression coefficients. Müller and Welsh (2009) proposed a robust consistent model selection criterion by extending the method in Müller and Welsh (2005) to GLM. It is based on a penalized measure of predictive ability of GLM that is estimated using m-out-of-n bootstrap method. It is flexible as it can be used with any estimator. Further, Müller and Welsh (2009) empirically established that its performance is best with the robust estimator due to Cantoni and Ronchetti (2001). However, this method is computationally intensive.

In this article, we propose a new robust model selection criterion in GLM. We show that it is a consistent model selection criterion in the sense that as sample size tends to infinity, the model selected coincides with the true model with probability approaching to one. A simulation study is presented to compare its performance with its competitors. The proposed model selection criterion along with the other criteria is applied to a data on diversity of *arboreal marsupials* (possums) in montane ash forest (Australia) for model selection.

Results and discussion

An important assumption in the GLM is that the distribution of response is a member of exponential family with the general form of the density given by (McCullagh and Nelder 1989)

$$P(y_i;\theta_i,\varphi) = e^{\frac{y_i\theta_i - b(\theta_i)}{a(\varphi)} + h(y_i,\varphi)}, \quad i = 1,2,\ldots,n$$

where, θ_i is the natural location parameter and ϕ is a scale parameter and y_i is the ith observation on the response variable Y.

A GLM is defined via a link function

$$g(\mu_i) = \boldsymbol{X}_i^T\boldsymbol{\beta}, \quad i = 1,2,\ldots,n \tag{1}$$

where, $\mu_i = E(Y_i) = \frac{db(\theta_i)}{d\theta_i}$, $\boldsymbol{X}_i^T$ is the ith row of $n \times k$ matrix X whose first column is of ones and the remaining columns contain observations on the predictors $X_1, X_2, \ldots, X_{k-1}$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{k-1})^T$ is the vector of regression coefficients.

The log likelihood function is

$$l(\boldsymbol{\beta};\boldsymbol{y}) = \sum_{i=1}^{n}\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\varphi)} + h(y_i,\varphi)\right\}.$$

The maximum likelihood score equations in matrix notations can be written as

$$X^T(\boldsymbol{y} - \boldsymbol{\mu}) = 0,$$

where, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$. The MLE of regression parameter $\boldsymbol{\beta}$ using iteratively reweighted least squares (IRLS) at convergence is (McCullagh and Nelder 1989)

$$\widehat{\boldsymbol{\beta}} = \left(X'V^{-1}X\right)^{-1}X'V^{-1}\boldsymbol{z},$$

where, V is an $n \times n$ diagonal matrix whose diagonal elements are $v_i = \frac{d\theta_i}{d\mu_i}a(\phi)$ and ith component of $n \times 1$ vector $\boldsymbol{z}$ is $z_i = g(\hat{\mu}_i) + (y_i - \hat{\mu}_i)\frac{dg(\mu_i)}{d\mu_i}$.

Robust estimation

The quasi-likelihood estimator is the solution of the system of estimating equations

$$\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\beta}}Q(y_i,\mu_i) = \sum_{i=1}^{n}\frac{(y_i - \mu_i)}{V(Y_i)}\mu_i' = 0, \tag{2}$$

where, $\mu_i' = \frac{\partial\mu_i}{\partial\boldsymbol{\beta}}$ and $Q(y_i, \mu_i)$ is the quasi-likelihood function. The solution to Eq. (2) can be viewed as an M-estimator (Huber 1981; Hampel et al. 1986) with score function $\tilde{\psi}(y_i,\mu_i) = \frac{(y_i-\mu_i)}{V(Y_i)}\mu_i'$. Its influence function (Hampel 1974; Hampel et al. 1986) is proportional to $\tilde{\psi}$ and is unbounded. Therefore, large deviations of the response from its mean or outlying points in the explanatory variables can have a large influence on the estimator and hence is non-robust (Cantoni and Ronchetti 2001).

Cantoni and Ronchetti (2001) proposed a robust estimation procedure based on quasi-likelihood. It is the solution of the estimating equations,

$$\sum_{i=1}^{n}\boldsymbol{\psi}(y_i,\mu_i) = 0, \tag{3}$$

where, $\boldsymbol{\psi}(y_i,\mu_i) = v(y_i,\mu_i)w(\boldsymbol{X}_i)\mu_i' - a(\boldsymbol{\beta})$, $a(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} E\left[v(y_i,\mu_i)\right]w(\boldsymbol{X}_i)\mu_i'$ with expectation taken with respect to the conditional distribution of $Y|\boldsymbol{x}$, $v(y_i,\mu_i)$ and $w(\boldsymbol{X}_i)$ are weight functions. The constant $a(\boldsymbol{\beta})$ ensures Fisher consistency of the estimator. Equation (3) corresponds to the minimization of the quantity,

$$Q_M(\boldsymbol{y},\boldsymbol{\mu}) = \sum_{i=1}^{n} Q_M(y_i,\mu_i), \tag{4}$$

with respect to $\boldsymbol{\beta}$ where,

$$\begin{aligned} Q_M(y_i,\mu_i) = &\int_{\tilde{s}}^{\mu_i} v(y_i,t)w(\boldsymbol{X}_i)dt \\ &- \frac{1}{n}\sum_{j=1}^{n}\int_{\tilde{t}}^{\mu_j} E\left[v(y_j,t)w(\boldsymbol{X}_j)\right]dt \end{aligned} \tag{5}$$

with $\tilde{s}$ such that $v(y_i,\tilde{s}) = 0$ and $\tilde{t}$ such that $E\left[v(y_j,\tilde{t})\right] = 0$.

Let $r_i = \frac{(y_i-\mu_i)}{\sqrt{V(Y_i)}}$ be the Pearson residual and ψ_c be the Huber function defined by

$$\psi_c(r) = \begin{cases} r, & |r| < c, \\ c\,\text{sign}(r), & |r| \geq c, \end{cases} \tag{6}$$

where, c is tuning constant. The simple choices for the weight functions $v(\cdot,\cdot)$ and $w(\cdot)$ could be $v(y_i,\mu_i) = \psi_c(r)\frac{1}{\sqrt{V(Y_i)}}$ and $w(\boldsymbol{X}_i) = \sqrt{1-h_{ii}}$, where, h_{ii} is the ith diagonal element of the hat matrix. The estimator defined in such a way is called the Mallows' quasi-likelihood estimator. When $w(\boldsymbol{X}_i) = 1$, this estimator is called as the Huber quasi-likelihood estimator. The details on the properties and computational aspect of this estimator are given in Cantoni and Ronchetti (2001).

Robust quasi-deviance

Cantoni and Ronchetti (2001) introduced the concept of robust quasi-deviance based on the notion of robust quasi-likelihood function to evaluate the adequacy of a model. The robust goodness of fit measure called robust quasi-deviance is defined as

$$D_{QM}(\boldsymbol{y},\boldsymbol{\mu}) = -2Q_M(\boldsymbol{y},\boldsymbol{\mu}) = -2\sum_{i=1}^{n} Q_M(y_i,\mu_i). \tag{7}$$

We call the model given in Eq. (1) as full model and denote its sub model as M_α, where $\alpha = \alpha_0 \cup \alpha_l$, $\alpha_0 = \{0\}$ denotes intercept and α_l denotes a non empty subset of $\{1,2,\ldots,k-1\}$. Hence, M_α is an individual model containing the predictors whose suffices are present in the set α. The model M_α, is defined as

$$g(\mu_{i,\alpha}) = X_{i,\alpha}^T\beta_\alpha \tag{8}$$

where, $X_{i,\alpha}$ denotes the sub-vector of X_i containing components indexed by α, β_α is a p_α-vector, and p_α denotes cardinality of α. Suppose α_N denotes all necessary predictors. Following Shao (1993) and using the notations similar to Hu and Shao (2008), Sakate and Kashid (2013, 2014), we define two exclusive classes of models. If the model M_α contains all the necessary predictors then it is a correct model. Collection of all such correct models is the class of correct models and is denoted by $\mathcal{M}_c$. Therefore,

$$\mathcal{M}_c = \{M_\alpha : \text{all the necessary predictors are present}\},$$
$$\text{i.e.}\mathcal{M}_c = \{M_\alpha : \alpha_N \subseteq \alpha\}$$

Similarly, if the model M_α doesn't contain at least one necessary predictor then it is a wrong model. Collection of all such wrong models is the class of wrong models and is denoted by $\mathcal{M}_w$. Therefore,
$\mathcal{M}_w = \{M_\alpha : \text{at least one necessary predictor is missing}\}$,
i.e. $\mathcal{M}_w = \{M_\alpha : \alpha_N \subsetneq \alpha\}$. The model M_α is called the optimal model if it contains only all the necessary predictors. It is denoted by M_{α_N}. In the following, we discuss robust model selection.

Müller and Welsh (MW) model selection criterion

Müller and Welsh (2009) combined a robust penalized measure of fit to the sample with a robust measure of out of sample predictive ability that is estimated using post-stratified m-out-of-n bootstrap to define a robust model selection criterion $A(M_\alpha)$ in GLM. Let $\boldsymbol{\beta}_\alpha$ be the vector of regression coefficients in the model M_α, g be the link function, ρ be a non negative loss function, δ be a specified function of sample size n, $V(Y_i) = \sigma^2 var(\boldsymbol{X}_{i,\alpha}^T\boldsymbol{\beta}_\alpha)$ where, σ^2 is a scale parameter and $var(\boldsymbol{X}_{i,\alpha}^T\boldsymbol{\beta}_\alpha)$ is the variance function and $\tilde{\boldsymbol{y}}$ be a vector of future observations at X that are independent of $\boldsymbol{y}$. Then the model M_α is selected for which the criterion function (Müller and Welsh 2009)

$$\begin{aligned} A(M_\alpha) = \frac{\sigma^2}{n}\Bigg\{ &E\sum_{i=1}^{n} w(\boldsymbol{X}_{i,\alpha})\rho\left[\frac{y_i - g^{-1}(\boldsymbol{X}_{i,\alpha}^T\boldsymbol{\beta}_\alpha)}{\sigma V(\boldsymbol{X}_{i,\alpha}^T\boldsymbol{\beta}_\alpha)}\right] + \delta(n)p_\alpha \\ &+ E\left(\sum_{i=1}^{n} w(\boldsymbol{X}_{i,\alpha})\rho\left[\frac{\tilde{y}_i - g^{-1}(\boldsymbol{X}_{i,\alpha}^T\boldsymbol{\beta}_\alpha)}{\sigma V(\boldsymbol{X}_{i,\alpha}^T\boldsymbol{\beta}_\alpha)}\right]\Bigg|\boldsymbol{y}, X\right)\Bigg\} \end{aligned} \tag{9}$$

is small. A common choice of $\delta(n)$ is $2\log n$ (Schwarz 1978; Müller and Welsh 2005). σ^2 is usually known. If it is unknown, it is estimated based on the full model by Pearson Chi square divided by its degrees of freedom. Let $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\mu}}$ be the estimate of $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ based on the full model. The in-sample term in the criterion function in Eq. (9) is estimated by $\widehat{\sigma}^2\left\{A_1(M_\alpha) + \frac{1}{n}\delta(n)p_\alpha\right\}$, where

$$A_1(M_\alpha) = \frac{1}{n}\sum_{i=1}^{n} w(\boldsymbol{X}_{i,\alpha})\rho\left[\frac{y_i - g^{-1}\left(\boldsymbol{X}_{i,\alpha}^T\widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}}\right)}{\widehat{\sigma}V\left(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}\right)}\right] \tag{10}$$

To compute the second term, a proportionally allocated, stratified m-out-of-n bootstrap is implemented. It can be summarized in the following steps (Müller and Welsh 2009).

Step I Compute and order Pearson residuals from the full model.

Step II Set the number of strata K between 3 and 8 (Cochran 1977, pp. 132–134) depending on the sample size.

Step III Set stratum boundaries at the $K^{-1}, 2K^{-1}, \ldots, (K-1)K^{-1}$ quantiles of the Pearson residuals.

Step IV Allocate observations to the strata in which the Pearson residuals lie.

Step V Sample $\frac{(\text{number of observations in stratum } K)m}{n}$ (rounded if necessary) rows of $(\boldsymbol{y}, X)$ independently with replacement from stratum K so that the total sample size is m.

Step VI Use these data to construct the estimator $\widehat{\boldsymbol{\beta}}^*_{\alpha,m}$, repeat steps V and VI, B independent times and then estimate the conditional expected prediction loss by $\widehat{\sigma}^2 A_2(M_\alpha)$, where

$$A_2(M_\alpha) = \frac{1}{n}\sum_{i=1}^{n} w(\boldsymbol{X}_{i,\alpha})\rho\left[\frac{y_i - g^{-1}\left(\boldsymbol{X}_{i,\alpha}^T\left\{\widehat{\boldsymbol{\beta}}^*_{\alpha,m} - E_*\left(\widehat{\boldsymbol{\beta}}^*_{\alpha,m} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}}\right)\right\}\right)}{\widehat{\sigma}V\left(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}\right)}\right] \tag{11}$$

and E_* denotes expectation with respect to the bootstrap distribution. Combining Eqs. (10) and (11) we get an estimate of the criterion function given in Eq. (9) as

$$\hat{A}(M_\alpha) = \widehat{\sigma}^2\left\{A_1(M_\alpha) + \frac{1}{n}\delta(n)p_\alpha + A_2(M_\alpha)\right\} \tag{12}$$

Müller and Welsh (2009) suggest using $\frac{n}{4} \le m \le \frac{n}{2}$ for moderate sample size n ($50 \le n \le 200$) and if n is large, m can be smaller than $\frac{n}{4}$.

The robust model selection criterion $A(M_\alpha)$ due to Müller and Welsh (2009) requires computations of the quantities given in Eqs. (10) and (11). Also, a computer intensive proportionally allocated, stratified m-out-of-n bootstrap is required to compute the quantity in Eq. (11). This makes its implementation by a researcher quite difficult. There is a need of a robust criterion which is easy to implement. We propose a robust version of deviance based criterion (DBC) called robust DBC (RDBC).

Proposed robust model selection criterion

The DBC proposed by Sakate and Kashid (2014) is defined as

$$\text{DBC}(M_\alpha) = \frac{D\left(y, \widehat{\boldsymbol{\beta}}_\alpha\right) - D\left(y, \widehat{\boldsymbol{\beta}}\right)}{\varphi} - (k - p_\alpha) + C(n, p_\alpha),$$

where, $C(n, p_\alpha)$ is a penalty term which measures the complexity of the model.

Instead of deviance, we use robust quasi deviance (Cantoni 2004) as a measure of discrepancy of the fitted GLM. Using the notion behind the DBC, we combine the robust discrepancy measure between a nested model M_α and the full model with the measure of complexity of the model M_α to define a robust model selection criterion in GLM. The robust measure of discrepancy between a nested model M_α and the full model (Cantoni and Ronchetti 2001) is

$$\Lambda_{QM} = D_{QM}\left(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}\right) - D_{QM}\left(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}\right), \tag{13}$$

where, $\widehat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\mu}}$ are robust estimators of $\boldsymbol{\mu}_\alpha$ and $\boldsymbol{\mu}$ respectively.

We define the robust version of DBC (RDBC) as

$$\text{RDBC}(M_\alpha) = \Lambda_{QM} + C(n, p_\alpha). \tag{14}$$

RDBC selects the model M_α if RDBC(M_α) is minimum in the class of all possible sub models. In the following, we establish the consistency property of the criterion given in Eq. (14). Note that, Λ_{QM} is always positive and vanishes when M_α is a full model. We require the following condition to establish the consistency property.

Condition 1 For $M_\alpha \in \mathcal{M}_w$ and $M_{\alpha_*} \in \mathcal{M}_c$,

$$\lim_{n\to\infty} \inf \left(D_{QM}\left(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}\right) - D_{QM}\left(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{\alpha_*}}\right) + C(n, p_\alpha) - C\left(n, p_{\alpha_*}\right)\right) > 0.$$

The following theorem ensures that the model selected using RDBC falls in the class of the correct models as n tends to infinity.

Theorem 1 *Under the Condition 1, for any correct model $M_{\alpha_*} \in \mathcal{M}_C$ and any wrong model M_α we have,*

$$\lim_{n\to\infty} \inf \Pr\left(\text{RDBC}(M_\alpha) > RDBC\left(M_{\alpha_*}\right)\right) = 1.$$

The proof of the Theorem is deferred to the "Appendix".

The next Theorem establishes the consistency property of

RDBC. Let M_n be the model selected using RDBC when sample size is n.

Condition 2 $C(n,p_\alpha) = o(n)$ and $C(n,p_\alpha) \uparrow \infty$ as $n \to \infty$.

Theorem 2 *Under the Condition 2, with probability approaching to one, as n tends to infinity, RDBC selects the optimal model* (M_{α_N}) *in the class of all correct models* $(\mathcal{M}_c)$ i.e.

$$\lim_{n\to\infty} \Pr\left(M_n = M_{\alpha_N}\right) = 1.$$

The proof of the Theorem is deferred to the "Appendix".
RDBC criterion is based on robust estimator which is biased. Therefore, to improve the performance of RDBC, we make use of bias reduction technique like bootstrap to modify RDBC.

Bootstrap RDBC

It is well known fact that the robust estimators of the unknown regression coefficients in the GLM are not unbiased and their bias is non negligible for small to moderate sample sizes. RDBC is based on the robust estimator due to Cantoni and Ronchetti (2001) which is a biased estimator. Therefore, we propose a modified version of RDBC using proportionally allocated, stratified m-out-of-*n* bootstrap.

The simulation study in Shao (1996), Wisnowski et al. (2003) and Simpson and Montgomery (1998) indicate that a straightforward implementation of the bootstrap fails in general in the presence of outliers. Müller and Welsh (2005) attributed this partly to non-robust loss function and hence non robust selection criterion and partly to the fact that some of the bootstrap samples may consist almost entirely of outliers. Proportionally allocated, stratified m-out-of-*n* bootstrap was used by Müller and Welsh (2005) for the first time in robust model selection. We propose a modified version of RDBC by replacing $\widehat{\boldsymbol{\beta}}$ by a proportionally allocated, stratified m-out-of-*n* bootstrap estimate of $\boldsymbol{\beta}$. We call this modified version of RDBC as bootstrap RDBC (B-RDBC). It is given by

$$\text{B-RDBC}(M_\alpha) = \Lambda^*_{QM} + C(n,p_\alpha), \tag{15}$$

where, $\Lambda^*_{QM} = \frac{1}{B}\sum_{j=1}^{B}\left(D_{QM}\left(\boldsymbol{y},\widehat{\boldsymbol{\mu}}^*_{\alpha,m,j}\right) - D_{QM}\left(\boldsymbol{y},\widehat{\boldsymbol{\mu}}^*_{m,j}\right)\right)$, $\widehat{\boldsymbol{\mu}}^*_{\alpha,m,j} = g^{-1}\left(X_{\alpha,m,j}\widehat{\boldsymbol{\beta}}^*_{\alpha,m,j}\right), \widehat{\boldsymbol{\mu}}^*_{m,j} = g^{-1}\left(X_{m,j}\widehat{\boldsymbol{\beta}}^*_{m,j}\right)$ and $X_{m,j}$ is the jth m-out-of-n bootstrap sample of size m from the matrix X. To compute the first term in B-RDBC defined in Eq. (15), we use the same algorithm given in section "Results and discussion" with Step VI replaced by Step VI*.

Step VI* Use these data to construct the estimator $\widehat{\boldsymbol{\beta}}^*_{\alpha,m,j}$, repeat steps V and VI*, B independent times and then compute Λ^*_{QM}.

Simulation results

Simulated data is used to compare the performance of the RDBC and B-RDBC with the MW criterion. The simulation design used to generate the data is described in detail in the "Methods" section. Table 1 gives the percentage of optimal model selection by RDBC, B-RDBC and MW criterion. From the Table 1, it is quite evident that B-RDBC outperforms RDBC for 5 % as well as 10 % contamination for all the sample sizes considered. Use of bootstrapping elevated the optimal model identification ability of RDBC when sample size is small. The performance of B-RDBC with penalty P_2 is compatible with MW criterion for small sample sizes. For large sample sizes, the performance of RDBC and B-RDBC with penalty P_2 is compatible with MW criterion. As sample size increases, the performance of all the criteria considered in this paper, becomes more or less same. This is because negligible bias is introduced in robust estimates of the regression coefficients when sample size is large. As such, RDBC which is easy to understand and implement can be used in place of MW criterion for model selection in GLM when sample size is large.

Real data application

We illustrate the proposed criterion using data on diversity of *arboreal marsupials* (possums) in montane ash forest (Australia). This data is described by Lindenmayer et al. (1990, 1991) and is a part of the 'robustbase' package in R (possumDiv.rda). For details on the study under consideration and the data collection method employed, we refer to Lindenmayer et al. (1990, 1991). The response is the count of different species (diversity) observed on $n = 151$ sites. Hence, a Poisson regression model is considered. The explanatory variables are shrubs, stumps, stags, bark, acacia, habitat, Eucalyptus and aspect. Cantoni and Ronchetti (2001) found observation number 59, 110, 133 and 139 as potentially influential data points. In the presence of these influential points in the data, Cantoni and Ronchetti (2001) advocated the use of robust estimator over MLE. We apply the proposed criterion, MW method, AIC and BIC for model selection on *arboreal marsupials* data. The results are reported in Table 2. B-RDBC and Müller and Welsh method based on Mallows' quasi-likelihood estimator select the same model based on the minimum number of variables. RDBC, AIC and BIC tend to select a model with larger number of variables.

The GLM are providing a satisfactory answer to many practical problems in the emerging quantitative analysis in the fields like environmental science and ecology. The data produced in the studies of air pollution, ozone

Table 1 Percentage of optimal model selection

ϵ	γ	n	RDBC		B-RDBC		$A(M_\alpha)$
			P_1	P_2	P_1	P_2	
0.05	2	64	89	93	98	99	98
		128	92	95	99	100	99
		192	94	96	100	100	100
	5	64	87	93	93	96	99
		128	89	93	99	99	100
		192	91	93	99	99	99
	10	64	86	91	91	96	98
		128	91	94	98	99	99
		192	92	95	99	99	100
0.10	2	64	88	91	96	98	97
		128	90	93	98	99	99
		192	92	95	99	100	99
	5	64	78	85	78	87	99
		128	85	89	94	97	100
		192	86	90	96	98	100
	10	64	78	83	70	80	98
		128	84	88	92	96	99
		192	86	91	97	99	100

Table 2 Selected models

Selection criterion	Selected variables in the best model
RDBC	Stags, bark, acacia, habitat, aspect
B-RDBC	Stags, habitat
MW method based on Mallows' quasi-likelihood estimator	Stags, habitat
MW method based on bias corrected Mallows' quasi-likelihood estimator (stratified bootstrap)	Stags, habitat
AIC	Stags, bark, acacia, habitat, aspect
BIC	Stags, bark, acacia, aspect

exceedance, ground water contamination, avian population monitoring, boreal treeline dynamics, aquatic bacterial abundance, conservation biology, marine and fresh water fish populations, etc. can be analyzed using GLM. GLM will provide a satisfactory solution to model based inference if only relevant variables are included in the model and there is no deviation from the assumed distribution of response. Identification and safe removal of the redundant predictors from the model in the presence of slight deviation from the assumed distribution of the response can be effectively done by the proposed criterion. Our criterion is robust to outliers which are common in any real data. It is also shown to be a consistent model selection criterion. Hence, our criterion is a good addition to easy implement and consistent model selection toolbox of researchers.

Methods

Simulation design

The empirical comparison of the proposed and existing model selection criteria is done using simulation study. The simulated data was generated according to a Poisson regression model with canonical link (log) and three predictors with intercept i.e. $\log\mu_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$. The predictors were generated from the standard uniform distribution i.e., $X_{ij} \sim U(0, 1)$, $j = 1, 2, 3$. The observations on the responses Y_i's were generated from Poisson distribution $P(\mu_i)$ and a perturbed distribution of the form $(1-\epsilon)P(\mu_i) + \epsilon P(\gamma\mu_i)$, where, $\epsilon = 0.05, 0.10$ and $\gamma = 2, 5, 10$.

To simulate the data, the regression parameters were set to $\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = 2$ and $\beta_3 = 0$. The choice of these parameters is not intentional but only for the purpose of illustration. We considered three different sample sizes, $n = 64$, 128 and 192. To compute B-RDBC and $A(M_\alpha)$, we divided the entire sample into eight equal-sized strata based on the Pearson residuals from the full model. In case of sample size $n = 64$, we draw 3 observations from each strata with replacement so that the sample size becomes 24. Similarly, for $n = 128$ and 192, we draw 5 and 7 observations and sample size becomes 40 and 56 respectively. This is in the accordance with the algorithm mentioned in section "Results and discussion".

In such a way, we obtain B = 50 bootstrap samples for each sample size. To implement RDBC and B-RDBC, we used the penalty functions $P_1 = p_\alpha log(n)$ and $P_2 = p_\alpha \big(log(n) + 1\big)$ for $C(n, p_\alpha)$. The Huber score function with tuning constant $c = 2$ was used to compute the robust estimator due to Cantoni and Ronchetti (2001). It can be easily computed using the `robustbase` (Rousseeuw et al. 2014) package in R software. This experiment was repeated 1000 times and the percentage of optimal model selection using these three criteria was obtained.

Conclusions

We proposed a robust model selection criterion in GLM called as RDBC. RDBC takes into account goodness of fit as well as complexity of the model. The consistency property of RDBC is also established. Performance evaluation and comparison with MW method is done using simulation study. These methods are also applied to the real ecological data. We also defined a bootstrap version of RDBC and called it as B-RDBC. Any suitable penalty function can be used without changing the form of RDBC and B-RDBC.

In case of quantitative analysis of environmental and ecological data using GLM, the distribution of response may deviate from the assumed distribution in the model and there might be some redundant predictors present in the model which are to be identified and safely removed from the model. The proposed criterion can be used effectively to perform model selection in GLM. It is robust to slight deviations from the assumed response distribution and the presence of outliers in the data. Overall, the proposed model selection criterion is robust, consistent and easy to implement model selection criterion as compared to its competitors.

Authors' contributions

DS has defined the proposed method, stated and proved the theorems, performed the simulation study and illustrated model selection for *arboreal marsupials* data. DK formulated the concept behind the method and contributed in writing, drafting the manuscript and revising it critically for intellectual content. Both authors read and approved the final manuscript.

Acknowledgements

The authors wish to thank the Editor and anonymous referees for their suggestions which led to the improvement in the paper.

Competing interests

The authors declare that they have no competing interests.

Appendix

Proof of Theorem 1 Consider,

$$\begin{aligned}\Pr\big(\text{RDBC}(M_\alpha) > RDBC\big(M_{\alpha_*}\big)\big) &= \Pr\big(D_{QM}\big(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}\big) - D_{QM}\big(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{\alpha_*}}\big) \\ &\quad + C(n, p_\alpha) - C\big(n,\ p_{\alpha_*}\big) > 0\big).\end{aligned}$$

Therefore using Condition 1 we have,

$$\begin{aligned}&\lim_{\text{n}\to\infty} \inf\ \Pr\big(\text{RDBC}(M_\alpha) > RDBC\big(M_{\alpha_*}\big)\big) \\ &= \lim_{\text{n}\to\infty} \inf\ \big\{\Pr\big(D_{QM}\big(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}\big) - D_{QM}\big(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{\alpha_*}}\big) + C(n, p_\alpha) - C\big(n, p_{\alpha_*}\big) > 0\big)\big\} \\ &> Pr\Big\{\lim_{\text{n}\to\infty} \inf\ \big(D_{QM}\big(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}\big) - D_{QM}\big(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{\alpha_*}}\big) + C(n, p_\alpha) - C\big(n, p_{\alpha_*}\big) > 0\big)\Big\} = 1.\end{aligned}$$

Proof of Theorem 2 In the light of Theorem 1, to prove Theorem 2 it is enough to prove that the value of RDBC for any correct model is larger than that for the optimal model as n tends to infinity. For this consider,

$$\begin{aligned}\Pr\big(\text{RDBC}\big(M_{\alpha_*}\big) > RDBC\big(M_{\alpha_N}\big)\big) &= \Pr\big(D_{QM}\big(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}_{\alpha_*}\big) - D_{QM}\big(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}_{\alpha_N}\big) \\ &\quad + C\big(n, p_{\alpha_*}\big) - C\big(n, p_{\alpha_N}\big) > 0\big) \\ &= \Pr\big(D_{QM}\big(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}_{\alpha_N}\big) - D_{QM}\big(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}_{\alpha_*}\big) \\ &\quad < C\big(n, p_{\alpha_*}\big) - C\big(n, p_{\alpha_N}\big)\big)\end{aligned}$$

According to Proposition 1 in Cantoni and Ronchetti (2001), $D_{QM}\big(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}_{\alpha_N}\big) - D_{QM}\big(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}_{\alpha_*}\big)$ is distributed as T, where, T is a linear combination of independent Chi square random variables with positive coefficients. Thus, T is a positive valued random variable. Therefore,

$$\begin{aligned}&\Pr\big(\text{RDBC}\big(M_{\alpha_*}\big) > RDBC\big(M_{\alpha_N}\big)\big) \\ &= \Pr\left(T < C\big(n, p_{\alpha_*}\big) - C\big(n, p_{\alpha_N}\big)\right)\end{aligned}$$

Under the Condition 1 and $p_{\alpha_*} > p_{\alpha_N}$, we have, $C\big(n, p_{\alpha_*}\big) - C\big(n, p_{\alpha_N}\big) > 0$ and increases to infinity as n tends to infinity, we have

$$\lim_{n\to\infty} \Pr\big(\text{RDBC}\big(M_{\alpha_*}\big) > RDBC\big(M_{\alpha_N}\big)\big) = \Pr\left(T < \infty\right) = 1$$

This indicates that, with probability approaching to one, asymptotically value of RDBC for the optimal model is the smallest in the class of all correct models. Moreover, RDBC selects that model for which its value is minimum among all possible models. Therefore,

$$\lim_{\text{n}\to\infty} \Pr\big(M_n = M_{\alpha_N}\big) = 1.$$

□

Received: 23 December 2015 Accepted: 8 February 2016
Published online: 24 February 2016

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19:716–723
Akaike H (1978) A Bayesian analysis of the minimum AIC procedure. Ann Inst Stat Math 30:9–14
Cantoni E (2004) Analysis of robust quasi-deviances for generalized linear models. J Stat Softw 10:i04
Cantoni E, Ronchetti E (2001) Robust inference for generalized linear models. J Am Stat Assoc 96:1022–1030
Cochran WG (1977) Sampling techniques, 3rd edn. Wiley, New York
Hampel FR (1974) The influence curve and its role in robust estimation. J Am Stat Assoc 69:383–393
Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986) Robust statistics: the approach based on influence functions. Wiley, New York
Heyde CC (1997) Quasi-likelihood and its application. Springer, New York
Hu B, Shao J (2008) Generalized linear model selection using R^2. J Stat Plan Inference 138:3705–3712
Huber PJ (1981) Robust statistics. Wiley, New York
Künsch HR, Stefanski LA, Carroll RJ (1989) Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. J Am Stat Assoc 84:460–466
Lindenmayer DB, Cunningham RB, Tanton MT, Smith AP, Nix HA (1990) The conservation of arboreal marsupials in the montane ash forests of the Victoria, South-East Australia: I. Factors influencing the occupancy of trees with hollows. Biol Conserv 54:111–131
Lindenmayer DB, Cunningham RB, Tanton MT, Nix HA, Smith AP (1991) The conservation of arboreal marsupials in the montane ash forests of the Central Highlands of Victoria, South-East Australia: III. The habitat requirements of Leadbeater's Possum Gymnobelideus leadbeateri and models of the diversity and abundance of arboreal marsupials. Biol Conserv 56:295–315
McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman & Hall, London
Morgenthaler S (1992) Least-absolute-deviations fits for generalized linear models. Biometrika 79:747–754
Müller S, Welsh AH (2005) Outlier robust model selection in linear regression. J Am Stat Assoc 100:1297–1310
Müller S, Welsh AH (2009) Robust model selection in generalized linear model selection. Stat Sin 19:1155–1170
Murtaugh PA (2009) Performance of several variable-selection methods applied to real ecological data. Ecol Lett 12:1061–1068
Pregibon D (1982) Resistant fits for some commonly used logistic models with medical applications. Biometrics 38:485–498
Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Koller M and Maechler M (2014) Robustbase: basic robust statistics. R package version 0.92-2. http://CRAN.R-project.org/package=robustbase
Ruckstuhl AF, Welsh AH (2001) Robust fitting of the binomial model. Ann Stat 29:1117–1136
Sakate DM, Kashid DN (2013) Model selection in GLM based on the distribution function criterion. Model Assist Stat Appl 8:321–332
Sakate DM, Kashid DN (2014) A deviance-based criterion for model selection in GLM. Statistics 48:34–48
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464
Shao J (1993) Linear model selection by cross-validation. J Am Stat Assoc 88:486–494
Shao J (1996) Bootstrap model selection. J Am Stat Assoc 91:655–665
Simpson JR, Montgomery DC (1998) The development and evaluation of alternative generalized m-estimation techniques. Commun Stat Simul Comput 27:1031–1049
Stefanski LA, Carroll RJ, Ruppert D (1986) Optimally bounded score functions for generalized linear models with applications to logistic regression. Biometrika 73:413–424
Wedderburn RWM (1974) Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. Biometrika 61:439–447
Wisnowski JW, Simpson JR, Montgomery DC, Runger GC (2003) Resampling methods for variable selection in robust regression. Comput Stat Data Anal 43:341–355