Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Tabatabai et al., J Biomet Biostat 2014, 5:4
DOI: 10.472/2155-6180.1000202

Research Article Open Access

Robust Logistic and Probit Methods for Binary and Multinomial Regression

Tabatabai MA[1], Li H[2], Eby WM[3], Kengwoung-Keumo JJ[2], Manne U[4], Bae S[5], Fouad M[5] and Singh KP[5]*

[1]School of Graduate Studies and Research, Meharry Medical College, Nashville, TN 37208, USA
[2]Department of Mathematical Sciences, Cameron University, Lawton, OK 73505, USA
[3]Department of Mathematics, New Jersey City University, Jersey City, NJ 07305, USA
[4]Department of Pathology and Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, AL 35294, USA
[5]Department of Medicine Division of Preventive Medicine and Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, AL 35294, USA

Abstract

In this paper we introduce new robust estimators for the logistic and probit regressions for binary, multinomial, nominal and ordinal data and apply these models to estimate the parameters when outliers or influential observations are present. Maximum likelihood estimates don't behave well when outliers or influential observations are present. One remedy is to remove influential observations from the data and then apply the maximum likelihood technique on the deleted data. Another approach is to employ a robust technique that can handle outliers and influential observations without removing any observations from the data sets. The robustness of the method is tested using real and simulated data sets.

Introduction

Binary and multinomial regressions are commonly used by medical scientists and researchers for analysis of binary or polytomous outcomes. These methods are routinely used as diagnostic tools in all areas of medicine including oncology and cardiology. Zhou et al. [1] used logistic regression to relate the gene expression with class labels. They also used logistic regression for their microarray-based analysis of cancer classification and prediction. Sator et al. [2] applied a logistic regression model to identify enriched biological groups in gene expression microarray studies. Majid et al. [3] performed logistic regression analysis to predict endoscopic lesions in iron deficiency anemia when there are no gastrointestinal symptoms.

Morris et al. [4] applied multinomial regression technique to analyze the sub-phenotypes by allowing for heterogeneity of genetic effects. Richman et al. [5] investigated the association between European ancestry and renal disease when compared with African Americans, East Asians, and Hispanics. They concluded that European ancestry is protective against the development of renal disease in systematic lupus erythematosus. Their data had some outliers but they were excluded in their final analysis. Timmerman et al. [6] used the logistic regression to distinguish between benign and malignant adnexal mass before surgery. Merritt et al. [7] used the binary and multinomial logistic regressions to investigate the role of dairy food intake and risk of ovarian cancer. The validity of estimation and testing procedures used in the analysis of binary data are heavily dependent on whether or not the model assumptions are satisfied. The maximum likelihood method of estimating binary regression parameters using logistic, probit and many other methods is extremely sensitive to outliers and influential observations.

There is a large literature on the robustness issue of the binary regression. Most of the existing methods attempt to achieve robustness by down weighting observations which are far from the majority of the data, that is, outliers. The reader is referred to papers published by Pregibon [8], Carroll and Pederson [9], and Bianco and Yohai [10]. Bianco and Martinez [11] modified the original score functions of the logistic regression to obtain bounded sensitivity, which is a concept introduced by Morgenthaler [12] using the L^1-norm instead of the L^2-norm in the likelihood, resulting in a weighted score function of the original score function. Cantoni and Ronchetti [13] focused on robustness of inference rather than the model. Pregibon [8] suggested resistant fitting methods which taper the standard likelihood to reduce the influence of extreme observations. Kordzakhia et al. [14] introduced a robust logistic regression by minimizing the mean-squared deviance for the worst case contamination. Bergesio and Yohai [15] introduced projection estimators for generalized linear model. These estimators have the same asymptotic normal distribution as the M-estimators. Hobza et al. [16] introduced a median estimator to estimate the parameters of the logistic regression.

Robust binary and multinomial regression estimators for analysis of biomedical data are proposed. This robust method has a bounded influence and high breakdown point and efficiency under normal distribution and is able to estimate the parameters of logistic and probit regression models. The proposed model is computationally simple and can easily be used by researchers.

Binary Regression Model

Consider the model $y_i=\pi(x_i;\beta)+\varepsilon_i$ where $\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n$ are independent random variables with $E(\varepsilon_i)=0$ and $Var(\varepsilon_i)=\pi(x_i;\beta)(1-\pi(x_i;\beta))$, and $y_1,y_2,\ldots,y_n$ are n independent Bernoulli random variables with $Ey_i=\pi(x_i;\beta)$ and $Var(y_i)=\pi(x_i;\beta)(1-\pi(x_i;\beta))$ such that the conditional success probability is given by $P(y_i=1|\ x_i)=\pi(x_i;\beta)$ and $x_i=(x_{i0},x_{i01},\ldots,x_{ip})^t$; $1\le i\le n$ is a p+1 dimensional vector of predictor variables with $\beta=(\beta_0,\beta_1,\ldots,\beta_p)^t$ as the parameters vector.

There are various estimation methods for the estimation of the parameter vector β. The most commonly used method is the logistic

***Corresponding author:** Karan P Singh, Department of Medicine Division of Preventive Medicine and Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, AL 35294, USA, Tel: 205-934-6887; Fax: 205-934-4262; E-mail: kpsingh@uab.edu

Received July 01, 2014; **Accepted** July 27, 2014; **Published** July 30, 2014

Citation: Tabatabai MA, Li H, Eby WM, Kengwoung-Keumo JJ, Manne U, et al. (2014) Robust Logistic and Probit Methods for Binary and Multinomial Regression. J Biomet Biostat 5: 202. doi:10.4172/2155-6180.1000202

regression which is used to analyze the effects of explanatory variables on the binary response y. In the logistic regression the link function $\pi_L(x_i;\beta)$ is assumed to have the following functional form

$$\pi_L(x_i;\beta)=\frac{1}{1+Exp(-\sum_{j=0}^{p}\beta_j x_{ij})}$$

The logistic transformation of $\pi_L(x_i;\beta)$ is called the logit function and is given by

$$\text{logit}\,(\pi_L(x_i;\beta))=Log\frac{\pi_L(x_i;\beta)}{1-\pi_L(x_i;\beta)}=\sum_{j=0}^{p}\beta_j x_{ij}$$

The probit function is a link function of the form

$$\pi_P(x_i;\beta)=\frac{1}{2}[Erf(\frac{\sum_{j=0}^{p}\beta_j x_{ij}}{\sqrt{2}})+1],$$

Where *Erf* function is defined as

$$Erf\left(x\right)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$$

which has the probit transformation function

$$probit(\pi_P(x_i;\beta))=\sqrt{2}Erf^{-1}\left(-1+2\pi_P(x_i;\beta)\right)=\sum_{j=0}^{p}\beta_j x_{ij}$$

The tabaistic model introduced by Tabatabai and Argyros [17] has the link function:

$$\pi_T(x_i;\beta)=\frac{1}{1+arcsinh[Exp[-\sum_{j=0}^{p}\beta_j x_{ij}]]},$$

Where arcsinh (.) represents the inverse hyperbolic sine function.

The tabaistic transformation function is called the tabit function and is defined as

$$tabit(\pi_T(x_i;\beta))=Log\left(Csch\left(\frac{1-\pi_T(x_i;\beta}{\pi_T(x_i;\beta)}\right)\right)=\sum_{j=0}^{p}\beta_j x_{ij}$$

Where C*sch* (.) denotes the hyperbolic cosecant function and the complementary log-log model link function has the form

$$\pi_{CLL}(x_i;\beta)=1-Exp[-[Exp[\sum_{j=0}^{p}\beta_j x_{ij}]]]$$

with the complementary log-log transformation function (cllogit) defined as

$$\text{cllogit}(\pi_{CLL}(x_i;\beta))=Log[-[Log[1-\pi_{CLL}(x_i;\beta)]]]=\sum_{j=0}^{p}\beta_j x_{ij}$$

Figure 1 shows the graph of π_L, π_{CLL}, π_P and π_T where of π_L, π_{CCL}, π_P and π_T take values between zero and one. The solid curve is the graph of π_L function, the dotted curve is the graph of π_p function, the dot-dashed curve is the graph of π_{CLL} function, and the dashed curve is the graph of π_T function. Figure 2 shows the graph of logit (π_L). cllogit (π_{CLL}), *probit*(π_P), and *tabit*(π_T). The solid curve, the dotted curve, the dot-dashed curve and the dashed curve are the graph of logit (π_L) function, the graph of *probit*(π_p) function, the graph of cllogit (π_{CLL}) function, and the graph of *tabit*(π_T) function, respectively. The principle of maximum likelihood is ordinarily used to estimate the model parameters by maximizing the log-likelihood function of the form

$$LL(\beta)=\sum_{i=1}^{n}[y_i Log(\pi(x_i;\beta))+(1-y_i)Log(1-\pi(x_i;\beta))]$$

In other words, the estimate $\hat{\beta}$ of β is

$$\hat{\beta}=\arg\max_{\beta\in\mathbb{R}^{p+1}}(LL(\beta))$$

Although the maximum likelihood estimator is asymptotically efficient, it is not recommended as a method of choice when outliers are present. The alternative techniques are robust statistical methods.

Tabatabai et al. [18] defined the one parameter family of differentiable functions $\rho_\omega(x)$ of the form $\rho_\omega(x)=1-Sech(\omega x)$, where the positive real number ω is called the tuning constant.

The bounded function $\rho_\omega: R\to R$ is a differentiable function satisfying the following properties:

$$\rho_\omega(0)=0$$

$$\lim_{x\to\infty}\rho_\omega(x)=\lim_{x\to-\infty}\rho_\omega(x)=1$$

$$(\forall x)(x\in\mathbb{R}\Rightarrow\rho_\omega(x)\geq 0)$$

$$(\forall x)(x\in\mathbb{R}\Rightarrow\rho_\omega(x)=\rho_\omega(-x)),$$

$$(\forall a)(\forall b)(a\in\mathbb{R}\wedge b\in\mathbb{R}\wedge|a|>|b|\Rightarrow\rho_\omega(a)\geq\rho_\omega(b))$$

$$(\forall\kappa)(\kappa>0\Rightarrow\lim_{x\to\infty}\frac{\rho_\omega(\kappa x)}{\rho_\omega(x)}=1)$$

$$\lim_{|x|\to\infty}\frac{d\rho_\omega(x)}{dx}=0$$

Under the normality assumption for the error term ε_i, the asymptotic efficiency (*Aeff*) is defined as

$$Aeff=\frac{(E[\psi_\omega^{'}(t)])^2}{E[\psi_\omega^2(t)]} \quad (1)$$

Where ψ_ω is the derivative of ρ_ω and is equal to

$$\psi_\omega(x)=\omega Sech(\omega x)Tanh(\omega x),$$

Where *Sech* and *Tanh* represent the hyperbolic secant and hyperbolic tangent, respectively.

$$E[\psi_\omega^{'}(t)]=\int_{-\infty}^{\infty}\psi_\omega^{'}(t)\frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}dt \text{ with}$$

$$E[\psi_\omega^2(t)]=\int_{-\infty}^{\infty}\psi_\omega^2(t)\frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}dt$$

The tuning constant ω can be calculated by solving the following equation (1) for ω. The numerical values for ω at the efficiency levels 0.80, 0.85, 0.90, and 0.95 are approximately 0.721, 0.628, 0.525 and 0.405, respectively. Although the choice for tuning constant ω is left for the investigator to decide, we do recommend an efficiency of approximately 90 percent which corresponds to ω=1/2. We now consider the hat matrix of the form

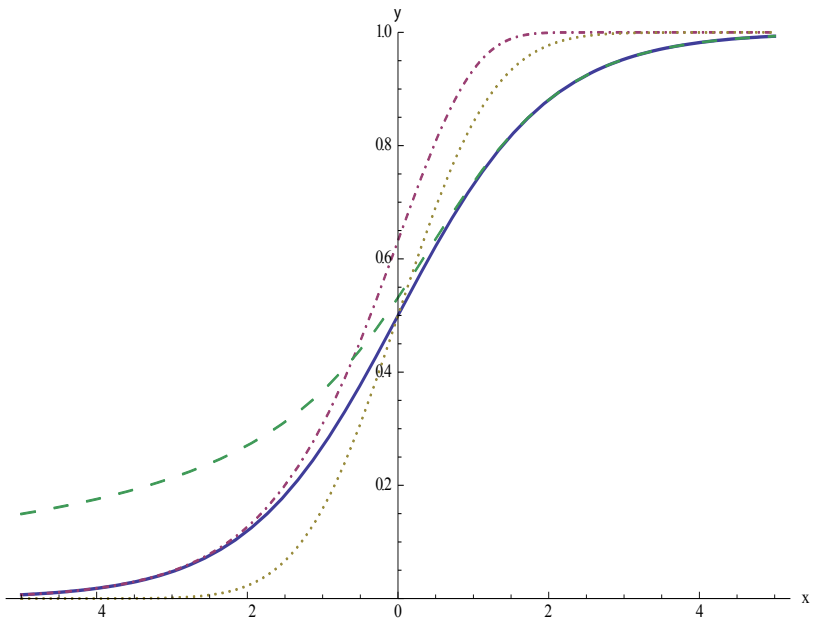

Figure 1: Graphs of π_L, π_{CLL}, π_P and π_T functions.

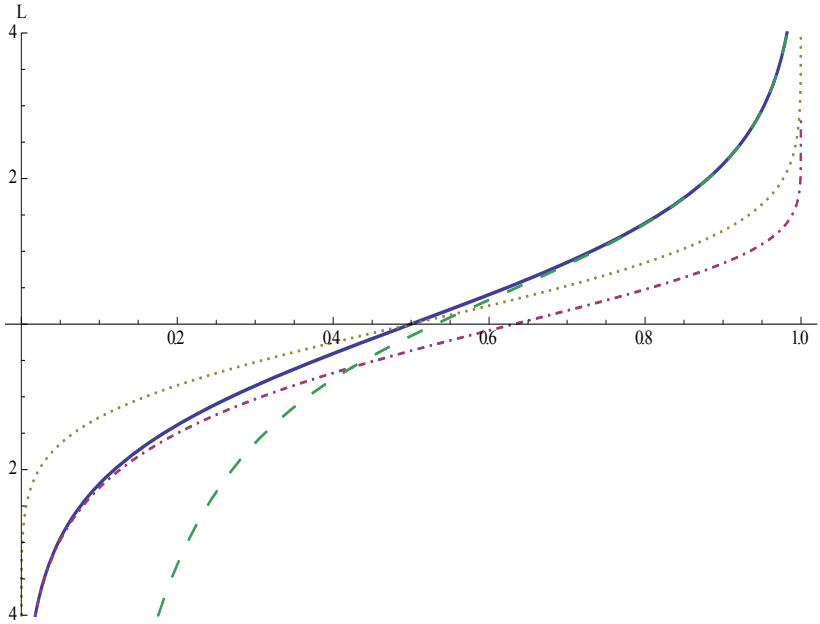

Figure 2: Graphs of logit (π_L). cllogit (π_{CLL}), *probit* (π_P), and *tabit* (π_T) functions.

$$H = X(X^tX)^{-1}X^t,$$

where X is the design matrix defined as

$$X = \begin{pmatrix} x_{10} & x_{11}\ldots & x_{1k} \\ x_{20} & x_{21}\ldots & x_{2k} \\ x_{n0} & x_{n1}\ldots & x_{nk} \end{pmatrix} = (X_0X_1\ldots X_K).$$

If the model has intercept, then the column vector X_0 has the form

$$X_0 = \begin{pmatrix} 1 \\ . \\ . \\ . \\ 1 \end{pmatrix}$$

For j=1,2,…,k, define

$$M_j = Median\{|x_{1j}|, |x_{2j}|, \ldots, x_{nj}|\}$$

and for i=1,2,…,n, define

$$L_i = \sum_{j=1}^{k} Max\{M_j, |x_{ij}|\}$$

and for ω>0 define the function $G_\omega(u)$ as

$$G_\omega(u) = \int_0^u \psi_\omega(-\ln(t))dt \cdot \quad (2)$$

The estimator

$$\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} \frac{(1-h_{ii})[\rho_c(d_i)+s(\beta' x_i)]}{L(i)} \quad (3)$$

Where $d_i = y_i \ln[\pi(x_i;\beta)] + (1-y_i)\ln[1-\pi(x_i;\beta)]$

$$s(t) = G(\pi(t)) + G(1-\pi(t)) - G(1)$$

and h_{ii} is the ith diagonal element in the hat matrix. For the logistic model, we have

$$\pi(t) = \frac{1}{1+Exp(-t)} \cdot$$

So that the above integral (2) can be written as

$$G_\omega(u) = \frac{2u^{1+\omega} H(1, \frac{1+\omega}{2\omega}, \frac{1}{2}(3+\frac{1}{\omega}), -u^{2\omega})}{1+\omega} - uSech(\omega \ln(u)),$$

Where $H(k_1,k_2,k_3,t)$ is the Gauss hypergeometric function 2F1 with parameters k_1,k_2 and k_3. If ω=1, then we have

$$G_1(u) = u^2 H(1,1,2,-u^2) - uSech(\ln(u)) = \frac{2u^2}{1+u^2} - \ln(1+u^2),$$

and if ω=1/2, then we have

$$G_{1/2}(u) = -4\sqrt{u} + 4Arc\tan(\sqrt{u}) + uSech(\frac{\ln(u)}{2}).$$

Define the Hessian matrix H_b for binary data as

$$H_b = \begin{pmatrix} \frac{\partial^2 LL(\beta)}{\partial \beta_0^2} & \cdots & \frac{\partial^2 LL(\beta)}{\partial \beta_0 \beta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 LL(\beta)}{\partial \beta_p^2} & \cdots & \frac{\partial^2 LL(\beta)}{\partial \beta_p \beta_0} \end{pmatrix}.$$

Then an estimate of the variance-covariance matrix for vector $\hat{\beta}$ is $\hat{Var}(\hat{\beta}) = (-H_b)^{-1}$ with an estimated variance σ^2 given by

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}[(y_i - \pi(x_i;\hat{\beta}))^2 / \pi(x_i;\hat{\beta})]}{n-(p+1)}.$$

To perform hypothesis testing, we let $\Omega \subseteq R^p$ be the parameter space and $\{\beta_{j_1}, \beta_{j_2}, ..., \beta_{j_q}\}$ be a subset of $\{\beta_0,\beta_1,\ldots,\beta_p\}$. Define

$$\Omega_0 = \{\beta \in \Omega : \beta_{j_1} = \beta_{j_2} = ... = \beta_{j_q} = 0\}.$$

To test the following hypothesis

$H_0 : \beta \in \Omega_0$ against the alternative $H_1 : \beta \in \Omega_0^c$,

one can use the Wald type test statistic which is defined as

$$W_n^2 = n(\hat{\beta}_{j_1}, \hat{\beta}_{j_2}, ..., \hat{\beta}_{j_q}) V_q^{-1} (\hat{\beta}_{j_1}, \hat{\beta}_{j_2}, ..., \hat{\beta}_{j_q})^t \cdot$$

The null distribution of the statistic W_n^2 is asymptotically a chi-square distribution with q degrees of freedom.

Robust Multinomial Logistic Regression Model

In this section we generalize the robust binary method to multinomial regression where the response y includes k categories. When k=2, this model reduces to the binary regression. Now, consider the response matrix Y given by

$$Y = \begin{pmatrix} y_{11} & \cdots & y_{1k} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nk} \end{pmatrix},$$

where

$$y_{ij} = \begin{cases} 1 & if\ y_i = j \\ 0 & if\ y_i \neq j \end{cases}.$$

This means that y_{ij}=1whenever the ith response is in category j.

For i =1,2,…,n, $\sum_{j=1}^{k} y_{ij} = 1$ and x_i=$(x_{i1},x_{i2},\ldots,x_{ip})^t$. The parameter vector

β_j=$(\beta_{j0},\beta_{j1},\ldots,\beta_{jp})^t$ and the vector β_0 is a zero vector in a p+1 dimensional space with β=$(\beta_0,\beta_1,\ldots,\beta_{k-1})^t$. The multinomial likelihood function of parameter vector β is defined by

$$ML(\beta) = \prod_{i=1}^{n}\prod_{j=1}^{k}[(\pi_j(x_i;\beta))^{y_{ij}}],$$

and the multinomial log- likelihood function is

$$MLL(\beta) = \sum_{i=1}^{n}(y_{i1}\ln[\pi_1(x_i;\beta)] + y_{i2}\ln[\pi_2(x_i;\beta)] + \ldots + y_{ik}\ln[1-\sum_{s=1}^{k-1}\pi_s(x_i;\beta)]).$$

For the generalized logit when the first category is the designated reference category and the intercepts are $\beta_{01},\beta_{02},\ldots,\beta_{0k-1}$, we get

$$\pi_1(x_i;\beta) = P(y_i = 1 | x_i;\beta) = \frac{1}{1+\sum_{s=2}^{k} Exp[-\eta_s(x_i;\beta)]},$$

and for j = 2,…,k,

$$\pi_j(x_i;\beta) = P(y_i = j | x_i;\beta) = \frac{Exp[-\eta_j(x_i;\beta)]}{1+\sum_{s=2}^{k}[Exp[-\eta_s(x_i;\beta)]},$$

and for j = 1,…,k–1 the logit function $\eta_j(x_i;\beta)$ is the log-odds of membership in category j versus the reference category 1 and is equal to

$$\eta_j(x_i;\beta) = \ln\left(\frac{\pi_j(x_i;\beta)}{\pi_1(x_i;\beta)}\right) = \beta_{0j} + \beta_{1j}x_{i1} + \beta_{2j}x_{i2} + \ldots + \beta_{pj}x_{ip}.$$

The principle of maximum likelihood can be used to estimate model parameters. The maximum likelihood estimate of the parameters vector β is

$$\hat{\beta} = \arg\max_{\beta\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} d_i^*, \quad (4)$$

where

$$d_i^* = \sum_{j=1}^{k} y_{ij} \ln(\pi_j(x_i;\beta))$$

The robust estimate of the model parameters is given by

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \frac{(1-h_{ii})[\rho_\omega(d_i^*) + \sum_{j=1}^{k} G_\omega(\pi_j(x_i;\beta)) - G_\omega(1)]}{L(i)} \quad (5)$$

Define the Hessian matrix H_p as

$$H_p = \begin{pmatrix} \frac{\partial^2 MLL(\beta)}{\partial \beta_0^2} & \cdots & \frac{\partial^2 MLL(\beta)}{\partial \beta_0 \beta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 MLL(\beta)}{\partial \beta_p^2} & \cdots & \frac{\partial^2 MLL(\beta)}{\partial \beta_p \beta_0} \end{pmatrix}$$

Then, for the multinomial response, an estimate of the variance-covariance matrix for vector $\hat{\beta}$ is $Var(\hat{\beta}) = (-H_p)^{-1}$ with an estimated variance σ^2 given by

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n}[(y_{ij} - \pi_j(x_i;\hat{\beta}))^2 / \pi_j(x_i;\hat{\beta})]}{(n-(p+1))(k-1)}$$

For the cumulative logit model for ordinal *k*-category response, the cumulative probability for the ith response belongs to the response category less than or equal j is

$$F_j(x_i) = P(y \le j | x_i),$$

and for $j=1,\ldots,k-1$ the ordinal logit $O_j(x_i;\beta)$ is the log-odds of falling into or below category j against falling above it and is given by

$$O_j(x_i;\beta) = \ln\left(\frac{F_j(x_i)}{1-F_j(x_i)}\right) = \beta_{0j} + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip},$$

where $\beta_{01} < \beta_{02} < \ldots < \beta_{0k-1}$. The ordered logit model is sometimes called the proportional odds model.

Let $\pi_j^O(x_i;\beta) = P(y_i = j | x_i;\beta)$. Then we have

$$\pi_j^O(x_i;\beta) = \begin{cases} \frac{1}{1+Exp[-O_1(x_i;\beta)]} & if \quad j=1 \\ \frac{1}{1+Exp[-O_j(x_i;\beta)]} - \frac{1}{1+Exp[-O_{j-1}(x_i;\beta)]} & if \quad 2 \le j \le k-1 \\ 1 - \frac{1}{1+Exp[-O_{k-1}(x_i;\beta)]} & if \quad j=k \end{cases}$$

and for the ordinal probit we have

$$\pi_j^O(x_i;\beta) = \begin{cases} \frac{1}{2}(1+Erf[\frac{O_1(x_i;\beta)}{\sqrt{2}}]) & if \quad j=1 \\ \frac{1}{2}(1+Erf[\frac{O_j(x_i;\beta)}{\sqrt{2}}]) - \frac{1}{2}(1+Erf[\frac{O_{j-1}(x_i;\beta)}{\sqrt{2}}]) & if \quad 2 \le j \le k-1 \\ 1 - \frac{1}{2}(1+Erf[\frac{O_{k-1}(x_i;\beta)}{\sqrt{2}}]) & if \quad j=k \end{cases}$$

The robust estimate of ordinal multinomial parameter vector is giveb

$$\hat{\beta}^O = \arg\min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \frac{(1-h_{ii})[\rho_\omega(d_i^O) + \sum_{j=1}^{k} G_\omega(\pi_j^O(x_i;\beta)) - G_\omega(1)]}{L(i)} \quad (6)$$

where $d_i^O = \sum_{j=1}^{k} y_{ij} \ln(\pi_j^O(x_i;\beta))$

Application

Vasoconstriction example

Vasoconstriction and vasodilation are two important physiological mechanisms used to control the circulation of blood throughout the body. These mechanisms directly affect both the blood pressure and the distribution of the blood in the body. Vasodilation refers to the expansion of blood vessels through relaxation of smooth muscles in the vessel walls. This allows increased flow of blood through these vessels and also decreases the blood pressure. Contraction of the same muscles tightens the blood vessels, which decreases blood flow and increases pressure. Thus, vasodilation and vasoconstriction work in opposition to adjust both blood flow and blood pressure. The usual controls for vasoconstriction and vasodilation are done by smooth muscles and autonomic nervous system, triggered by the medulla. These responses can also be affected by drugs promoting either constriction or dilation. Furthermore, there is a means of control by circulating hormones in the bloodstream, as well as control by intrinsic mechanisms to vasculature, called the myogenic response. The antagonistic operation of vasoconstriction and vasodilation is used by the body for numerous purposes. Primary among these is regulation of the supply of oxygen and nutrients to the cells of the body, to meet their needs. Furthermore, this regulation of blood flow is also needed for thermoregulation within the body. At times of increased metabolic needs or needs for oxygen in certain organs or systems in the body, the blood flow to these regions will also be modulated. Finally, vasoconstriction is also important in restricting blood flow to regions of the body in cases of traumatic injury.

The data set was analyzed originally by Finney [19]. It consists of 39 observations where the binary response variable y=1 or y=0 represent the presence or absence of vasoconstriction of the skin respectively. This experimental data set considers the effect of inhalation of air in a single deep breath on the presence or absence of vasoconstriction in the digits. Presence or absence of vasoconstriction is considered as a categorical variable, and the study considers the effect of two variables, the volume of inhaled air and the rate of inhalation. (Data has hidden outliers so that the robust logistic regression will be useful in its analysis.) In the remainder of this work, we denote by ML and BY, the Maximum Likelihood and Bianco-Yohai methods, respectively. By examining Table 1 we conclude that the new robust estimator has produced the closest parameter estimates to the maximum likelihood estimates when the outliers were removed. In addition, the χ^2_{arc} is the lowest for the new robust method. The χ^2_{arc} is defined by Kordzakhia et al. [14] as

$$\chi^2_{arc} = \sum_{i=1}^{n} 4(\arcsin\sqrt{y_i} - \arcsin\sqrt{\pi(x_i;\hat{\beta})})^2$$

Plasma example

The erythrocyte sedimentation rate (ESR) is very important. It is a common hematology test which is simple and inexpensive but can be used to detect infection or acute phase response, which can alert physicians to a wide variety of conditions. The test is very versatile and can assist physicians in detecting conditions from rheumatoid arthritis

Methods	Coefficients			Standard Error		
	b0	b1	b2	b0	b1	b2
ML	-2.887	5.191	4.578	1.324	1.869	1.843
	(χ^2_{arc} =48.3118)					
ML (Influential observations removed)	-24.590	39.539	31.928	13.974	23.153	17.687
	(χ^2_{arc} =32.0159)					
BY c=1.25	-5.3214	8.4454	7.4801	9.7647	14.1903	12.3199
	(χ^2_{arc} =43.1853)					
New Robust c=0.5	-24.1191	38.4639	30.9608	15.3232	24.9410	19.2365
	(χ^2_{arc} =32.2671)					

Table 1: Parameter estimates for Vaso Data Using ML, BY and New Robust Method.

to systemic lupus erythematosus to multiple myeloma; however it is non-specific and is usually combined with other tests. In practice, ESR is used widely to test for a range of conditions, including inflammation, trauma, and malignant disease. Studies have also suggested the utility of the ESR among the elderly as a general indicator of level of sickness or disease. Recently ESR has also attracted attention for a potential role as a predictor for the development of cardio-vascular disease and heart failure.

The ESR simply measures the rate at which red blood cells precipitate during a period of one hour. Anticoagulated blood is placed in an upright tube, and the rate at which the erythrocytes settle is measured in mm per hour. Although the test is a direct measurement of rate of sedimentation, the balance between factors stimulating sedimentation and factors resisting sedimentation allows for a number of clinically relevant factors to influence this rate. Fibrinogen is the most important factor promoting sedimentation, and the high level of fibrinogen in the blood during the inflammatory process makes this test sensitive to inflammation. High levels of fibrinogen in the blood decrease the repulsive forces experienced between the negatively charged erythrocytes and favor the formation of rouleaux. These stacks of erythrocytes that stick together will settle faster and lead to an increased ESR. Other acute phase reactants, or other large molecules, especially when positively charged, can have a similar effect, although fibrinogen has been observed to have the largest effect.

A recent focus on the inflammatory nature of artherosclerosis has been accompanied by a recent study of increased levels of ESR and elevated risk of coronary heart disease. Erikssen et al. [20] observed that elevated ESR is a strong predictor of mortality from heart failure, suggesting it may serve as a marker for aggressive forms of coronary heart disease. Andresdottir et al. [21] observed an increased risk of coronary heart disease among the top quintile or ESR rates, with a hazard ratio of 1.57 for men and 1.9 for women. The 2005 paper of Inglesson et al. [22] also observes a significant association between elevated ESR and heart failure, suggesting both that inflammation is involved in the processes leading to heart failure and that the ESR may be used in evaluating this process. In addition to the well-established uses of ESR, Saadeh [23] mentions some potential new applications of this test such as bacterial otitis media, acute hematogenous osteomyelitis, AIDS, pelvic inflammatory disease, prostate cancer, and early prediction of stroke severity.

Although the ESR usually detects acute phase response from fibrinogen in blood in conditions such as those mentioned above, in certain cases there are factors which decrease the rate of sedimentation. One important factor that can slow the rate of sedimentation is irregularity in the erythrocytes, either in shape or unusually small size. As a consequence, ESR can detect certain blood diseases (including sickle cell anemia and spherocytosis) which lead to a lower than normal rate of sedimentation, as observed in Bridgen [24]. Other conditions that may also lower ESR include the extreme levels of white blood cells as observed in chronic lymphocytic leukemia. Furthermore the surplus of erythrocytes found in patients with polycythemia makes rouleau formation difficult and decreases the ESR.

In clinical applications the erythrocyte sedimentation rate may in many cases be treated as a categorical variable, with a normal ESR for values less than some given α and an elevated ESR for values greater than α. When representing such a set of data where ESR depends on one or more variables the logistic regression may be used. For instance in the data set from Collett [25], the ESR is considered as a function of two variables, the level of fibrinogen and the level of γ-globulin. The data for 32 individuals represents the levels of fibrinogen and γ-globulin in the blood and whether the ESR level is healthy (< 20 mm/hr) or unhealthy (≥ 20 mm/hr), and the logistic regression is used to describe how both fibrinogen and γ-globulin affect the ESR variable. Since this data set contains (hidden/influential) outliers, both the probit method of regression and the logit method do not give accurate results. However we observed that our new methods for robust logistic regression do represent the data accurately. The logit, when all 32 observations are included in the study, is given by

$$Logit(\hat{\pi}(X_i)) = -6.845 + 1.827 f_i$$

When one removes the influential observations 15, and 23, the logit model becomes

$$Logit(\hat{\pi}(X_i)) = -59.62 + 17.46 f_i$$

The level of γ-globulin was not a statistically significant variable to be included in the model. Thus only the level of fibrinogen f is used in the variable selection.

Again, examining Table 2 reveals that the new robust estimator has produced the closest parameter estimates to the maximum likelihood estimates when the outliers were removed as well as the lowest value for the χ^2_{arc}.

Mental health example

The following example involves the ordinal multinomial regression. The data comes from a mental health study for a random sample of adult

Methods	Coefficients		Standard Error	
	b0	b1	b0	b1
ML	-6.8451	1.8271	2.7703	0.9009
	(χ^2_{arc} =38.9405)			
ML (Influential observations removed)	-59.62	17.46	45.51	13.50
	(χ^2_{arc} =29.1415)			
BY				
c=1.25	-8.3774	2.2870	5.4383	1.6632
	(χ^2_{arc} =37.0421)			
New method c=0.5	-60.5094	17.7654	49.7325	14.7409
	(χ^2_{arc} 29.2719)			

Table 2: Parameter Estimates for Plasma Data Using ML, BY and New Robust Method.

Parameter	Estimated	Standard Error	Robust estimate	Standard Error
Intercept 1	-0.2819	0.6423	-.2374	0.7265
Intercept 2	1.2128	0.6607	1.1923	0.7507
Intercept 3	2.2094	0.7210	2.2981	0.8364
Life	-0.3189	0.1210	-.3181	0.1618
Socioeconomic status	1.1112	0.6109	1.1423	0.7789

Table 3: Parameter estimates for Mental Health data using robust ordinal method.

Method	Bias	MSE	Bias (5% x)	MSE (5%)
ML				
b_0	0.0430	0.0876	0.08065	0.07938
b_1	0.1784	0..6310	1.00777	2.16368
BY				
b_0	0.0463	0.0923	0.0027	0.0842
b_1	0.1857	0.6662	0.2336	0.8885
New Robust				
b_0	0.0342	0.0897	0.0125	0.0779
b_1	0.1125	0.5321	0.2642	0.9771

Table 4: Simulation Results for Logistic Regression (b_0=1, b_1=3, N=100, m=1000).

Method	Bias	MSE	Bias (5%)	MSE (5%)
ML				
b_0	0.04283	0.08152	0.00312	0.07526
b_1	0.01071	0.25287	0.19770	0.23694
b_2	0.10448	0.38209	0.36612	0.75513
BY				
b_0	0.0485	0.0856	0.0232	0.0788
b_1	0.0128	0.2650	0.1711	0.2456
b_2	0.1143	0.4028	0.0909	0.5125
New Robust				
b_0	0.0410	0.0786	0.0270	0.0755
b_1	0.0255	0.2799	0.1045	0.2676
b_2	0.1034	0.4037	0.0553	0.5032

Table 5: Simulation Results for Logistic Regression (b0=1, b1=0.5, b2=2, N=100, m=1000).

residents of Alachua County, Florida. This data was appeared in Agresti [26]. The mental impairment is divided into four categories (well, mild symptom formation, moderate symptom formation, and impaired). The explanatory variables are life events index X_1 and socioeconomic status X_2, where X_2 is binary and takes high and low levels. There is no outlier in this data. We just want to show how the method works even when outliers are not present. The logit is given by

$$logit\left[P\left(Y \le j\right)\right] = \alpha_j + \beta_1 X_1 + \beta_2 X_2$$

Table 3 gives the parameter values using maximum likelihood method as well as the new robust ordinal multinomial method. The R program for this example is provided in the Appendix.

Simulation

To evaluate the performance of the new robust method for logistic regression we conduct a Monte Carlo simulation. In the first round of our simulation, we use one explanatory variable and in the next round, we increase the number of explanatory variables to two. We first generate an independent random sample of size 100 from the standard normal distribution with mean 0 and standard deviation equal to 0.5. We call the variable x. Then we generate a sample of error terms ε_i of size 100 from the logistic distribution with mean zero and standard deviation equal to 1. The dependent variable y is generated using the formula

$$y_i = \begin{cases} 1 & if \quad 1 + 3x_i + \varepsilon_i > 0 \\ 0 & if \quad 1 + 3x_i + \varepsilon_i \le 0 \end{cases}.$$

The model parameters are 1 (intercept) and 3 (coefficient for x). We then select a random sample (5%) from the generated sample x and contaminate the selected sample by multiplying each x value by a factor of 10. Then we repeat the above procedures 1000 times. Finally, we estimate both the bias and mean squared errors using the following equations

$$bias = \left| \frac{\sum_{l=1}^{m} \hat{\theta}_l}{m} - \theta \right|,$$

where m is the number of iterations in the simulation. The mean squared error is estimated by

$$MSE = \frac{\sum_{l=1}^{m} \left(\hat{\theta} - \theta\right)^2}{m}$$

For the two explanatory variables, we generate two independent normal random samples of size 100 from a normal distribution with mean 0 and standard deviation 0.5 and call them x_1 and x_2 respectively. Then we select 5% of this random sample (3% from x_1 and 2% from x_2) and multiply the selected samples by 10. Then we generate a sample of error terms ε_i of size 100 from the logistic distribution with mean zero and standard deviation equal to 1. The dependent variable y is generated using the formula

$$y_i = \begin{cases} 1 & if \quad 1 + .5x_{1i} + 2x_{2i} + \varepsilon_i > 0 \\ 0 & if \quad 1 + .5x_{1i} + 2x_{2i} + \varepsilon \le 0 \end{cases}$$

We calculate the parameter estimates and continue the iteration 1000 times. In addition, we calculate the bias and mean squared errors. Tables 4 and 5 show the results of simulations using ML, BY and the new robust method with one and two explanatory variables, respectively. For binary logistic regression the simulation results indicate that our new robust method is as good as the BY method. The BY method only covers binary logistic regression whereas our method not only covers binary but also covers multinomial regression for both nominal and ordinal responses.

Discussion and Conclusions

In this work we have proposed a new robust method to analyze binary and multinomial regression models. We believe that these new robust methods for binary and multinomial regressions have potential to play a key role in modeling categorical data in medical, biological and engineering sciences. We have shown the lack of robustness of the maximum likelihood technique when outliers are present. In both real examples and simulated ones and when the outliers are present, the new

robust method performed well. In conclusion the motivation was to introduce a new robust loss function of residuals which can attain high breakdown value. The method has high efficiency and high breakdown points with bounded influence function.

Acknowledgement

Research reported in this paper was partially supported by the Center grant of the National Cancer Institute of the National Institutes of Health to the University of Alabama at Birmingham Comprehensive Cancer Center (P30 CA013148), the Cervical SPORE grant (P50CA098252), the Morehouse/Tuskegee University/ UAB Comprehensive Cancer Center Partnership grant (2U54-CA118948), and the Mid-South Transdisciplinary Collaborative Center for Health Disparities Research (U54MD008176). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Zhou X, Liu KY, Wong ST (2004) Cancer classification and prediction using logistic regression with Bayesian gene selection. J Biomed Inform 37: 249-259.
2. Sartor MA, Leikauf GD, Medvedovic M (2009) LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. Bioinformatics 25: 211-217.
3. Majid S, Salih M, Wasaya R, Jafri W (2008) Predictors of gastrointestinal lesions on endoscopy in iron deficiency anemia without gastrointestinal symptoms. BMC Gastroenterol 8: 52.
4. Morris AP, Lindgren CM, Zeggini E, Timpson NJ, Frayling TM, et al. (2010) A powerful approach to sub-phenotype analysis in population-based genetic association studies. Genet Epidemiol 34: 335-343.
5. Richman IB, Taylor KE, Chung SA, Trupin L, Petri M, et al. (2012) European genetic ancestry is associated with a decreased risk of lupus nephritis. Arthritis Rheum 64: 3374-3382.
6. Timmerman D, Testa AC, Bourne T, Ferrazzi E, Ameye L, et al. (2005) Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: A multicenter study by the International Ovarian Tumor Analysis Group. J Clinical Oncology 23: 8794-8801.
7. Merritt MA, Cramer DW, Vitonis AF, Titus LJ, Terry KL (2013) Dairy foods and nutrients in relation to risk of ovarian cancer and major histological subtypes. Int J Cancer 132: 1114-1124.
8. Pregibon D (1982) Resistant fits for some commonly used logistic models with medical application. Biometrics 38: 485-498.
9. Carroll RJ, Pederson S (1993) On Robust Estimation in the Logistic Regression Model. J Roy Statist Soc B 55: 693-706.
10. Bianco AM, Yohai VJ (1996) Robust Estimation in the Logistic Regression Model in Robust Statistics. Data Analysis, and Computer Intensive Methods, 17-34; Lecture Notes in Statistics 109, 1996, Springer Verlag, Ed. H. Rieder, New York.
11. Bianco AM, Martinez E (2009) Robust testing in the logistic regression model. Computational Statistics & Data Analysis 53: 4095-4105.
12. Morgenthaler S (1992) Least-absolute-deviations Øts for generalized linear models. Biometrika 79: 747-754.
13. Cantoni, Ronchetti E (2001b) Robust inference for generalized linear models. Journal of the American Statistical Association 96: 1022-1030.
14. Kordzakhia N, Mishra GD, Reiersolmoen L (2001) Robust Estimation in Logistic regression Model. Journal of Statistical Planning and Inference 98: 211-223.
15. Bergesio A, Yohai VJ (2011) Projection Estimators for Generalized Linear Models. Journal of American Statistical Association 106: 661-671.
16. Hobza T, Pardo L, Vajda I (2008) Robust median estimator in logistic regression. J. Statistical Planning and Inference 138: 3822-3840.
17. Tabatabai MA, Argyros IK (2010) Tabaistic regression and its application to the space shuttle Challenger O-ring data. Jour Appl Math and Com 513-523.
18. Tabatabai MA, Eby WM, Li H, Bae S, Singh KP (2012) TELBS robust linear regression method. Open Access Medical Statistics 2: 65-84.
19. FINNEY DJ (1947) The estimation from individual records of the relationship between dose and quantal response. Biometrika 34: 320-334.
20. Erikssen G, Liestøl K, Bjørnholt JV, Stormorken H, Thaulow E, et al. (2000) Erythrocyte sedimentation rate: a possible marker of atherosclerosis and a strong predictor of coronary heart disease mortality. Eur Heart J 21: 1614-1620.
21. Andresdottir MB, Sigfusson N, Sigvaldason H, Gudnason V (2003) Erythrocyte sedimentation rate, an independent predictor of coronary heart disease in men and women: The Reykjavik Study. Am J Epidemiol 158: 844-851.
22. Ingelsson E, Arnlöv J, Sundström J, Lind L (2005) Inflammation, as measured by the erythrocyte sedimentation rate, is an independent predictor for the development of heart failure. J Am Coll Cardiol 45: 1802-1806.
23. Saadeh C (1998) The erythrocyte sedimentation rate: old and new clinical applications. South Med J 91: 220-225.
24. Brigden ML (1999) Clinical utility of the erythrocyte sedimentation rate. Am Fam Physician 60: 1443-1450.
25. Collett D (1999) Modelling Binary Data. (1stedn), Chapman & Hall 163-169.
26. Agresti A (2007) In introduction to categorical data analysis. (2ndedn), John Wiley: 184-187.

Citation: Tabatabai MA, Li H, Eby WM, Kengwoung-Keumo JJ, Manne U, et al. (2014) Robust Logistic and Probit Methods for Binary and Multinomial Regression. J Biomet Biostat 5: 202. doi:10.4172/2155-6180.1000202