Robust Logistic and Probit Methods for Binary and Multinomial Regression

Tabatabai MA¹, Li H², Eby WM³, Kengwounou-Keumo JJ⁴, Manne U⁵, Bae S¹, Fouad M⁵ and Singh KP**

¹School of Graduate Studies and Research, Meharry Medical College, Nashville, TN 37208, USA
²Department of Mathematical Sciences, Cameron University, Lawton, OK 73505, USA
³Department of Mathematics, New Jersey City University, Jersey City, NJ 07305, USA
⁴Department of Pathology and Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, AL 35294, USA
⁵Department 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 systemic 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 $y_i = \pi(x_i\beta) + \epsilon_i$, where $\epsilon_1, \ldots, \epsilon_n$ are independent random variables with $E(\epsilon) = 0$ and $\text{Var}(\epsilon) = \pi(x_i\beta)(1-\pi(x_i\beta))$, and $y_1, \ldots, y_n$ are independent Bernoulli random variables with $E(y_i) = \pi(x_i\beta)$ and $\text{Var}(y_i) = \pi(x_i\beta)(1-\pi(x_i\beta))$ such that the conditional success probability is given by $P(y_i = 1 | x_i) = \pi(x_i\beta)$ and $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$. $1 \leq i \leq n$ 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 $\beta$. The most commonly used method is the logistic
regression which is used to analyze the effects of explanatory variables on the binary response \( y \). In the logistic regression the link function \( \pi_l(x; \beta) \) is assumed to have the following functional form

\[
\pi_l(x; \beta) = \frac{1}{1 + \exp(-\sum_{j=0}^{p} \beta_j x_{ij})}
\]

The logistic transformation of \( \pi_l(x; \beta) \) is called the logit function and is given by

\[
\text{logit} (\pi_l(x; \beta)) = \log \left( \frac{\pi_l(x; \beta)}{1-\pi_l(x; \beta)} \right) = \sum_{j=0}^{p} \beta_j x_{ij}
\]

The probit function is a link function of the form

\[
\pi_p(x; \beta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\pi(x; \beta)} e^{-\frac{t^2}{2}} dt
\]

with the probit transformation function

\[
\text{probit}(\pi_p(x; \beta)) = \sqrt{2\pi} \exp \left( -\frac{x^2}{2} \right) + 1
\]

Where \( \text{Erf} \) function is defined as

\[
\text{Erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt
\]

which has the probit transformation function

\[
\text{probit}(\pi_p(x; \beta)) = \sqrt{2\pi} \exp \left( -\frac{x^2}{2} \right) + 1
\]

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

\[
\pi_T(x; \beta) = \frac{1}{1 + \text{arcsinh} \left[ \exp \left( -\sum_{j=0}^{p} \beta_j x_{ij} \right) \right]}
\]

Where \( \text{arcsinh} () \) represents the inverse hyperbolic sine function.

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

\[
\text{tabit} (\pi_T(x; \beta)) = \log \left( \frac{1 - \pi_T(x; \beta)}{\pi_T(x; \beta)} \right) = \sum_{j=0}^{p} \beta_j x_{ij}
\]

Where \( \text{Csch} () \) denotes the hyperbolic cosecant function and the complementary log-log model link function has the form

\[
\pi_{CLL}(x; \beta) = 1 - \exp \left( -\sum_{j=0}^{p} \beta_j x_{ij} \right)
\]

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

\[
\text{cllogit}(\pi_{CLL}(x; \beta)) = \log \left[ \log \left( 1 - \pi_{CLL}(x; \beta) \right) \right] = \sum_{j=0}^{p} \beta_j x_{ij}
\]

Figure 1 shows the graph of \( \pi_l, \pi_{CLL}, \pi_T, \) and \( \pi_e \) where of \( \pi_l, \pi_{CLL}, \pi_T \) and \( \pi_e \) take values between zero and one. The solid curve is the graph of \( \pi_l \) function, the dotted curve is the graph of \( \pi_T \) function, the dot-dashed curve is the graph of \( \logit(\pi_l) \) function, the graph of \( \text{probit}(\pi_p) \) function, the graph of \( \text{cllogit}(\pi_{CLL}) \) function, and the graph of \( \text{tabit}(\pi_T) \) function, respectively. The principle of maximum likelihood is ordinarily used to estimate the model parameters by maximizing the log-likelihood function of the form

\[
LL(\beta) = \sum_{i=1}^{n} \left[ y_i \log(\pi_l(x_i; \beta)) + (1 - y_i) \log(1 - \pi_l(x_i; \beta)) \right]
\]

In other words, the estimate \( \hat{\beta} \) of \( \beta \) is

\[
\hat{\beta} = \arg \max_{\beta \in \mathbb{R}} (LL(\beta))
\]

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 \( \rho_\omega(x) = 1 - \text{Sech} (\omega x) \), where the positive real number \( \omega \) is called the tuning constant.

The bounded function \( \rho_\omega : \mathbb{R} \to \mathbb{R} \) is a differentiable function satisfying the following properties:

\[
\rho_\omega(0) = 0
\]

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

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

\[
(\forall x) (x \in \mathbb{R} \Rightarrow \rho_\omega(x) = \rho_\omega(-x))
\]

\[
(\forall a, b \in \mathbb{R} \land |b| > |a| \Rightarrow \rho_\omega(a) = \rho_\omega(b))
\]

\[
(\forall x > 0) \lim_{x \to 0} \rho_\omega(x) = 1
\]

\[
\lim_{x \to 0} \frac{d\rho_\omega(x)}{dx} = 0
\]

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

\[
Aeff = \left[ \frac{\text{E}[\psi_\omega(t)]^2}{\text{E}[\psi_\omega^2(t)]} \right]
\]

Where \( \psi_\omega \) is the derivative of \( \rho_\omega \) and is equal to

\[
\psi_\omega(x) = \omega \text{Sech}(\omega x) \text{Tanh}(\omega x)
\]

Where \( \text{Sech} \) and \( \text{Tanh} \) represent the hyperbolic secant and hyperbolic tangent, respectively.

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

With

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

The tuning constant \( \omega \) can be calculated by solving the following equation (1) for \( \omega \). The numerical values for \( \omega \) 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 \( \omega \) is left for the investigator to decide, we do recommend an efficiency of approximately 90 percent which corresponds to \( \omega = 1/2 \). We now consider the hat matrix of the form
For $j = 1, 2, \ldots, k$, define

$$M_j = \text{Median}(\{|x_{j1}|, |x_{j2}|, \ldots, |x_{jn_j}|\})$$

$$X_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

where $X$ is the design matrix defined as

$$X = \begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1k} \\ x_{20} & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{nk} \end{pmatrix} = (X_0 X_1 \cdots X_K).$$

If the model has intercept, then the column vector $X_i$ has the form

$$H = X(X'X)^{-1}X'$$

$X_0$.
and for \( i = 1, 2, \ldots, n \), define
\[
L_i = \max_{j=1}^{n} \{M_{ij}, x_{ij}\}
\]
and for \( \omega > 0 \) define the function \( G_{\omega}(u) \) as
\[
G_{\omega}(u) = \int_{0}^{u} \psi(-\ln(t))dt.
\]
(2)

The estimator
\[
\hat{\beta} = \arg\min_{\beta \in \omega} \sum_{i=1}^{n} \left[ -h_i \psi(\beta^T x_i) + \psi(\beta^T x_i) \right]
\]
(3)

Where \( h_i \) is the \( i \)-th diagonal element in the hat matrix. For the logistic regression, \( H_{\omega}(u) \) is the Gauss hypergeometric function \( 2F1 \) with \( \omega \).

To test the following hypothesis
\[
H_0 : \beta \in \Omega_0
\]
against the alternative
\[
H_1 : \beta \in \Omega_1
\]
(4)

and one can use the Wald type test statistic which is defined as
\[
W^2 = n(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)^T \Omega_0^{-1}(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)
\]

The null distribution of the statistic \( W^2 \) 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
\[
y_{ij} = \begin{cases}
1 & \text{if } y_i = j \\
0 & \text{if } y_i \neq j
\end{cases}
\]
This means that \( y_i = 1 \) whenever the \( i \)-th response is in category \( j \).

For \( i = 1, 2, \ldots, n \), \( \sum_{j=1}^{k} y_{ij} = 1 \) and \( x_i = (x_{i1}, x_{i2}, \ldots, x_{ik}) \). The parameter vector
\[
\beta = (\beta_1, \beta_2, \ldots, \beta_k)^T
\]
and the multinomial log-likelihood function is
\[
ML(\beta) = \sum_{i=1}^{n} \prod_{j=1}^{k} \left[ \pi_{ij}(x_i; \beta) \right]^{y_{ij}}
\]
and the multinomial log-likelihood function is
\[
MLLL(\beta) = \sum_{i=1}^{n} \prod_{j=1}^{k} \left[ \pi_{ij}(x_i; \beta) + y_{ij} \ln(\pi_{ij}(x_i; \beta)) + \cdots + y_{ik} \ln(1 - \sum_{j=1}^{k} \pi_{ij}(x_i; \beta)) \right].
\]

For the generalized logit when the first category is the designated reference category and the intercepts are \( \beta_{i1}, \beta_{i2}, \ldots, \beta_{ik} \), we get
\[
\pi_j(x_i; \beta) = P(y_i = j | x_i; \beta) = \frac{1}{1 + \sum_{s=2}^{k} \exp[-\eta_j(x_i; \beta)]}
\]
and for \( j = 2, \ldots, k \),
\[
\eta_j(x_i; \beta) = \ln \frac{\pi_j(x_i; \beta)}{\pi_1(x_i; \beta)} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{i,p} x_{ip}
\]
and for \( j = 1, \ldots, k-1 \) 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
\[
\eta_j(x_i; \beta) = \ln \frac{\pi_j(x_i; \beta)}{\pi_1(x_i; \beta)} = \beta_{0j} + \beta_{1j} x_{i1} + \beta_{2j} x_{i2} + \cdots + \beta_{ipj} x_{ip}.
\]

The principle of maximum likelihood can be used to estimate model parameters. The maximum likelihood estimate of the parameters vector \( \hat{\beta} \) is
\[
\hat{\beta} = \arg\max_{\beta \in \Omega} \sum_{i=1}^{n} d_i^T
\]
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
\[ \hat{\beta} = \arg \min \sum_{i=1}^{n} \frac{(1 - h_i)[\rho_u(d_i^*) - \sum_{j=1}^{k} G_u(\pi_j(x_i; \beta)) - G_u(1)]}{L(i)} \]  

(5)

Define the Hessian matrix \( H_p \) as
\[ H_p = \begin{pmatrix} \frac{\partial^2 \text{MLL}(\beta)}{\partial \beta_0^2} & \ldots & \frac{\partial^2 \text{MLL}(\beta)}{\partial \beta_p \partial \beta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \text{MLL}(\beta)}{\partial \beta_p \partial \beta_{p-1}} & \ldots & \frac{\partial^2 \text{MLL}(\beta)}{\partial \beta_p^2} \end{pmatrix} \]

Then, for the multinomial response, an estimate of the variance-covariance matrix for vector \( \hat{\beta} \) is \( \text{Var}(\hat{\beta}) = (H_p)^{-1} \) with an estimated variance \( \sigma^2 \) given by
\[ \sigma^2 = \frac{1}{n - (p + 1)(k - 1)} \sum_{j=1}^{k} \frac{y_{ij}^2 - \pi_j(x_i; \hat{\beta})^2}{\pi_j(x_i; \hat{\beta})} \]

For the cumulative logit model for ordinal \( k \)-category response, the cumulative probability for the \( j \)th response belongs to the response category less than or equal to \( j \) is
\[ F_j(x_i) = P(y \leq j | x_i) \]

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
\[ O_j(x_i; \beta) = \ln \left( \frac{F_j(x_i)}{1 - F_j(x_i)} \right) = \rho_{0j} + \rho_1 x_{i1} + \rho_2 x_{i2} + \ldots + \rho_p x_{ip}, \]

where \( \rho_{0j} < \rho_{02} < \ldots < \rho_{0k-1} \). The ordered logit model is sometimes called the proportional odds model.

Let \( \pi_j^0(x_i; \beta) = P(y = j | x_i; \beta) \). Then we have
\[ \pi_j^0(x_i; \beta) = \begin{cases} \frac{1}{1 + \exp[-O_j(x_i; \beta)]} & \text{if } j = 1 \\ \frac{1}{1 + \exp[-O_j(x_i; \beta)]} - \frac{1}{1 + \exp[-O_{j-1}(x_i; \beta)]} & \text{if } 2 \leq j \leq k - 1 \\ \frac{1}{1 + \exp[-O_k(x_i; \beta)]} & \text{if } j = k \end{cases} \]

and for the ordinal probit we have
\[ \pi_j^0(x_i; \beta) = \begin{cases} \frac{1}{2}(1 + \text{erf} \frac{O_j(x_i; \beta)}{\sqrt{2}}) & \text{if } j = 1 \\ \frac{1}{2}(1 + \text{erf} \frac{O_j(x_i; \beta)}{\sqrt{2}}) - \frac{1}{2}(1 + \text{erf} \frac{O_{j-1}(x_i; \beta)}{\sqrt{2}}) & \text{if } 2 \leq j \leq k - 1 \\ \frac{1}{2}(1 + \text{erf} \frac{O_k(x_i; \beta)}{\sqrt{2}}) & \text{if } j = k \end{cases} \]

The robust estimate of ordinal multinomial parameter vector is given by
\[ \hat{\beta}^0 = \arg \min \sum_{i=1}^{n} \frac{(1 - h_i)[\rho_u(d_i^0) - \sum_{j=1}^{k} G_u(\pi_j^0(x_i; \beta)) - G_u(1)]}{L(i)} \]

(6)

where \( d_i^0 = \sum_{j=1}^{k} y_{ij} \ln(\pi_j^0(x_i; \beta)) \)

**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 dilatation. 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 \( y = \{1, 0\} \) 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 \( X_{arc} \) is the lowest for the new robust method. The \( X_{arc} \) is defined by Kordzakhia et al. [14] as

\[ X_{arc} = \sum_{i=1}^{n} \left( \arcsin \sqrt{y_i} - \arcsin \sqrt{\pi_i(x_i; \beta)} \right)^2 \]

**Plasma example**

The erythrocyte sedimentation rate (ESR) is very important. It is a common hematolgy 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.
One important factor that can slow the rate of sedimentation is fibrinogen in blood in conditions such as those mentioned above, in osteomyelitis, AIDS, pelvic inflammatory disease, prostate cancer, and the uses of ESR. Saadeh [23] mentions some potential new applications be used in evaluating this process. In addition to the well-established applications, the elevated ESR and heart failure, suggesting both that inflammation is 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 atherothrombotic disease has been accompanied by a recent study of increased 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 $\chi^2_{arc}$.

### 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

<table>
<thead>
<tr>
<th>Methods</th>
<th>Coefficients</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$b_0$</td>
<td>$b_1$</td>
</tr>
<tr>
<td>ML</td>
<td>-6.8451</td>
<td>1.8271</td>
</tr>
<tr>
<td>(Influential observations removed)</td>
<td>($\chi^2_{arc} = 38.9405$)</td>
<td>($\chi^2_{arc} = 29.1415$)</td>
</tr>
<tr>
<td>BY</td>
<td>-59.62</td>
<td>17.46</td>
</tr>
<tr>
<td>(Influential observations removed)</td>
<td>($\chi^2_{arc} = 29.1415$)</td>
<td>($\chi^2_{arc} = 37.0421$)</td>
</tr>
<tr>
<td>New method</td>
<td>-8.3774</td>
<td>2.2870</td>
</tr>
<tr>
<td>(Influential observations removed)</td>
<td>($\chi^2_{arc} = 29.2719$)</td>
<td>($\chi^2_{arc} = 29.2719$)</td>
</tr>
</tbody>
</table>

Table 2: Parameter Estimates for Plasma Data Using ML, BY and New Robust Method.
Table 5: Simulation Results for Logistic Regression (b0=1, b1=0.5, b2=2, N=100, m=1000).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimated</th>
<th>Standard Error</th>
<th>Robust estimate</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept 1</td>
<td>-0.2819</td>
<td>0.6423</td>
<td>-2.374</td>
<td>0.7265</td>
</tr>
<tr>
<td>Intercept 2</td>
<td>1.2128</td>
<td>0.6607</td>
<td>1.1923</td>
<td>0.7507</td>
</tr>
<tr>
<td>Intercept 3</td>
<td>2.2094</td>
<td>0.7210</td>
<td>2.2981</td>
<td>0.8364</td>
</tr>
<tr>
<td>Life</td>
<td>-0.3189</td>
<td>0.1210</td>
<td>-0.3181</td>
<td>0.1618</td>
</tr>
<tr>
<td>Socioeconomic status</td>
<td>1.1112</td>
<td>0.8109</td>
<td>1.1423</td>
<td>0.7789</td>
</tr>
</tbody>
</table>

Table 4: Simulation Results for Logistic Regression (b0=1, b1=3, N=100, m=1000).

<table>
<thead>
<tr>
<th>Method</th>
<th>Bias</th>
<th>MSE</th>
<th>Bias (5%)</th>
<th>MSE (5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>b_0</td>
<td>0.0430</td>
<td>0.0876</td>
<td>0.08065</td>
</tr>
<tr>
<td></td>
<td>b_1</td>
<td>0.1784</td>
<td>0.6310</td>
<td>1.00777</td>
</tr>
<tr>
<td>BY</td>
<td>b_0</td>
<td>0.0463</td>
<td>0.0923</td>
<td>0.0027</td>
</tr>
<tr>
<td></td>
<td>b_1</td>
<td>0.1857</td>
<td>0.6662</td>
<td>0.2336</td>
</tr>
<tr>
<td>New Robust</td>
<td>b_0</td>
<td>0.0342</td>
<td>0.0897</td>
<td>0.0125</td>
</tr>
<tr>
<td></td>
<td>b_1</td>
<td>0.1125</td>
<td>0.5321</td>
<td>0.2642</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Bias</th>
<th>MSE</th>
<th>Bias (5%)</th>
<th>MSE (5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>b_0</td>
<td>0.04283</td>
<td>0.08162</td>
<td>0.00312</td>
</tr>
<tr>
<td></td>
<td>b_1</td>
<td>0.01071</td>
<td>0.25287</td>
<td>0.19777</td>
</tr>
<tr>
<td>BY</td>
<td>b_0</td>
<td>0.10448</td>
<td>0.38209</td>
<td>0.36612</td>
</tr>
<tr>
<td></td>
<td>b_1</td>
<td>0.0485</td>
<td>0.0856</td>
<td>0.0232</td>
</tr>
<tr>
<td>New Robust</td>
<td>b_0</td>
<td>0.0128</td>
<td>0.2650</td>
<td>0.1711</td>
</tr>
<tr>
<td></td>
<td>b_1</td>
<td>0.1143</td>
<td>0.4028</td>
<td>0.0909</td>
</tr>
</tbody>
</table>

Table 5: Simulation Results for Logistic Regression (b0=1, b1=3, N=100, m=1000).

<table>
<thead>
<tr>
<th>Method</th>
<th>Bias</th>
<th>MSE</th>
<th>Bias (5%)</th>
<th>MSE (5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>b_0</td>
<td>0.0410</td>
<td>0.0876</td>
<td>0.0270</td>
</tr>
<tr>
<td></td>
<td>b_1</td>
<td>0.0255</td>
<td>0.2799</td>
<td>0.1045</td>
</tr>
<tr>
<td>BY</td>
<td>b_0</td>
<td>0.1034</td>
<td>0.4037</td>
<td>0.0553</td>
</tr>
</tbody>
</table>

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 $X_1$ 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

$$y_i = \begin{cases} 1 & \text{if} & 1 + 3x_1 + \epsilon_i > 0 \\ 0 & \text{if} & 1 + 3x_1 + \epsilon_i \leq 0 \end{cases}$$

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

$$bias = \frac{\sum_{i=1}^{m} (\hat{\theta}_i - \theta)^2}{m}$$

$$MSE = \frac{\sum_{i=1}^{m} (\hat{\theta}_i - \theta)^2}{m}$$

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 $\epsilon_i$ 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

$$y_i = \begin{cases} 1 & \text{if} & 1 + 5x_{1i} + 2x_{2i} + \epsilon_i > 0 \\ 0 & \text{if} & 1 + 5x_{1i} + 2x_{2i} + \epsilon_i \leq 0 \end{cases}$$

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