ISSN 0971-0388

Aligarh Journal of Statistics
Vol. 38 (2018), 33-42

A Generalized Logit Model –An Alternative to Probit and Logit Models

C. Satheesh Kumar* and L. Manju

[Received on January, 2015. Accepted on December, 2018]

ABSTRACT

Probit and logit models are two commonly used techniques for regressing certain independent variables on a dichotomous dependent variable. Through this paper we propose an alternative regression model to the probit and logit models, based on the generalized logistic distribution of Balakrishnan and Leung (Communications in Statistics - Computation and Simulation, 1988). We discuss the maximum likelihood estimation of the parameters of the model and illustrate its usefulness with the help of a real life data set.

1. Introduction

Probit and logit models are two commonly used techniques for regressing certain independent variables on a dichotomous dependent variable when the categories are assumed to reflect an underlying normal/logistic distribution of the dependent variable. The probability distribution of both these models is symmetric. But, in practice, there are several situations where asymmetry arises in the distributional pattern of the dependent variable. In such cases, the assumption of symmetry of the probit and logit models violates and there by these procedures become inappropriate. So through this paper, we propose a generalized logit model based on type II generalized logistic distribution of Balakrishnan and Leung (1988). They defined the distribution as follows:

*Corresponding author**: C. Satheesh Kumar, Department of Statistics, University of Kerala, Trivandrum. E-mail: drcsatheeshkumar@gmail.com and L.Manju , Department of Community Medicine, Sree Gokulam Medical College & Research Foundation, Trivandrum.

A continuous random variable Z is said to follow "the type II generalized logistic distribution (GLD_{II})" if its probability density function (p.d.f) is of the following form, for any $z \in R = (-\infty, +\infty)$ and $c > 0$.

$$f(z;c) = ce^{-cz}\left(1+e^{-z}\right)^{-(c+1)} \qquad (1.1)$$

The cumulative distribution function (c.d.f) of the GLD_{II} with p.d.f (1.1) is the following, for any $z \in R = (-\infty, +\infty)$.

$$F(z) = 1 - e^{-cz}\left(1+e^{-z}\right)^{-c} \qquad (1.2)$$

Note that the GLD_{II} is an asymmetric distribution, and a regression model based on this distribution will be capable for tackling the asymmetric distributional pattern of the dependent variable. So through this paper we propose a generalized logit model based on the GLD_{II}.

The rest of the paper is organized as follows. In section 2 we describe the probit and logit models and in section 3 we present the definition and some important properties of the GLD_{II}. In section 4 we consider the estimation of the parameters of the proposed generalized logit model and in section 5 a real life data application is considered for illustrating the usefulness of the model compared to the existing models. A generalized likelihood ratio test procedure is also suggested for testing the significance of the additional parameter c of the generalized model and a brief simulation study is conducted in section 5.

2. Probit and Logit Models

The probit models were introduced in the mid 1930s in toxicology studies and the idea of probit analysis was originally due to Chester Ittner Bliss in 1934 (Agresti, 2007). In some situations we need a regression model which will predict the response probabilities p_i, say $P(Y_i = 1)$ of the dependent variable Y_i. The dependent random variable Y_i is assumed to be binary taking values, say 0 and 1. That is, $Y_i \in \{0,1\}$, for each $i = 1,2,\ldots,n$. The outcomes on Y are assumed to be mutually exclusive and exhaustive, and assumed to depend on k observable variables $X_1, X_2, \ldots, X_k$. We can indicate this relationship by writing $p = P(Y = 1 \mid X_1, X_2, \ldots, X_k)$, or simply $p = P(Y \mid X)$ where X denote a set of k independent variables. It is assumed that no exact or near linear dependencies exist among these k independent variables. The link

function for the model, called the probit link function, transforms probabilities to z-scores from the standard normal distribution. The probit link function transforms p so that the regressions curve for p or for $1-p$ has the appearance of the normal c.d.f. The probit link function applied to p gives the standard normal z-score at which the left tail probability equals p. The model can be represented as

$$\Phi^{-1}(p) = Z = a + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k, \tag{2.1}$$

where p is the proportion and Φ is the c.d.f of the standard normal distribution. $p = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du.$ (2.2)

The logistic regression models were not in much use until 1970s but they are now more popular than the probit model in several application studies (Agresti, 2007). The shape of the logistic distribution, similar to that of the normal distribution but with slightly thicker tails, makes it simpler and also more appropriate in certain occasions. In such cases, the model can be represented as in the following, in which $z \in R = (-\infty, \infty)$.

$$p = (1 + e^{-z})^{-1}, \tag{2.3}$$

which shows a smooth S-shaped curve symmetric about the point $z = 0$ and such a characteristic of the function make it an attractive alternative to the linear probability model for dichotomous dependent variables.

3. The Generalized Logit Model

An important drawback of standard logit model is that it is not suitable for asymmetric distributional patterns. In most of the practical situations, the data may not be symmetric. Since the GLD_{II} with c.d.f (1.2) is a skewed one, based on this distribution, we propose a generalized logit regression model through the following representation, in which $z \in R$ and $c > 0$.

$$p = 1 - e^{-cz}(1 + e^{-z})^{-c} \tag{3.1}$$

From Balakrishnan and Hossain (2007), it can be noted that the generalized logistic distribution with p.d.f (1.1) is negatively skewed when $c < 1$ and positively skewed when $c > 1$. Also, its skeweness measure can be viewed as a

decreasing function of c. If the value of c tends to infinity, the GLD_{II} has 'heavier tails' than the normal distribution.

Analogous to Aldrich and Nelson (1984), we describe the estimation of the parameters of the generalized logit regression model by method of maximum likelihood as follows: For $i = 1,2,\ldots,n$, let

$$p_i = P(Y_i = 1 \mid X_i) = 1 - e^{-cz}(1 + e^{-z})^{-c}, \tag{3.2}$$

where

$$z = a + \sum_{j=1}^{k} b_j X_{ij} \tag{3.3}$$

Thus $P(Y_i = 0 \mid X_i) = 1 - p_i$ and the probability of observing outcome Y_i, whether it be 0 or 1 is given by

$$P(Y_i \mid X_i) = p_i^{Y_i}(1 - p_i)^{1-Y_i}.$$

The probability of observing a particular sample of n values of Y, say $\boldsymbol{Y}$, given all n sets of values of X_i, say $\boldsymbol{X}$, is given by the product of the 'n' probability expressions as given below, since the observations are independent.

$$P(Y \mid X) = \prod_{i=1}^{n} p_i^{Y_i}(1 - p_i)^{1-Y_i} \tag{3.4}$$

Let $\theta = (a, b_1, b_2, \ldots, b_k, c)$ be the vector of parameters of the generalized logit regression model and let $\hat{\theta} = (\hat{a}, \hat{b}_1, \hat{b}_2, \ldots, \hat{b}_k, \hat{c})$ be the maximum likelihood estimator (MLE) of θ. Since $P(Y \mid X)$ depends on θ, we can write the log-likelihood function $\log L\,(y \mid z; \theta)$ as given below, in which z is as defined in (3.3).

$$\log L\,(y \mid z;\, \theta) = \sum_{i=1}^{n} Y_i \log \left(1 - \frac{e^{-cz}}{\left(1+e^{-z}\right)^{c}}\right) + \sum_{i=1}^{n} (1 - Y_i) \log \left(\frac{e^{-cz}}{\left(1+e^{-z}\right)^{c}}\right) \tag{3.5}$$

Now, the MLE of the parameters are obtained by solving the following set of likelihood equations, in which

$$\delta_c = (1 + e^{-z})^{-c} \text{ and } \delta_{c+1} = (1 + e^{-z})^{-(c+1)}$$

$$\frac{\partial \log L\,(y \mid z; \theta)}{\partial c} = 0$$

or equivalently

$$\left[\sum_{i=1}^{n} Y_i \frac{1}{1-\delta_c e^{cz}} - \sum_{i=1}^{n}(1-Y_i)\frac{e^{cz}}{\delta_c}\right] * \left[e^{-cz}\delta_c \log\left(1+e^{-z}\right) + ze^{-cz}\delta_c\right] = 0, \tag{3.6}$$

$$\frac{\partial \log L\,(y \mid z;\theta)}{\partial a} = 0$$

or equivalently

$$\left[\sum_{i=1}^{n} Y_i \frac{1}{1-\delta_c e^{-cz}} - \sum_{i=1}^{n}(1-Y_i)\frac{e^{cz}}{\delta_c}\right] * \left[c\, e^{-cz}\, \delta_c - c\, e^{-(c+1)z}\, \delta_{c+1}\right] = 0, \tag{3.7}$$

$$\frac{\partial \log L\,(y \mid z;\theta)}{\partial b_j} = 0$$

or equivalently

$$\left[\sum_{i=1}^{n} Y_i \frac{1}{1-\delta_c e^{-cz}} - \sum_{i=1}^{n}(1-Y_i)\frac{e^{cz}}{\delta_c}\right] * \left[ce^{-cz}\delta_c - ce^{-(c+1)z}\delta_{c+1}\right] X_{ij} = 0, \tag{3.8}$$

for each $j = 1, 2, \ldots, k$. Since the likelihood equations as given in (3.6) to (3.8) do not have a solution, the maximum of the log-likelihood equation is attained at the border of the domain of the parameters. So we obtained second order partial derivatives of (3.5) with respect to the parameters and we observed, with the help of MATHEMATICA software that the equation give negative values for all $a \in R$, $b \in R$ and $c > 0$. Hence the MLE of the parameters are unique under these parametric restrictions (Puig, 2003). Thus, one can obtain the value of $\hat{\theta}$ by solving the likelihood equations (3.6) to (3.8) with the help of the mathematical software *MATHEMATICA.*

4. An Application

For numerical illustration of the procedures discussed in the above section, we consider the "Prostrate cancer data set" available in https://www.umass.edu/

statdata/statdata/. This is also used by Hosmer and Lemeshow (2000). Among the 380 patients 153 had tumor that penetrated the prostatic capsule. The variable ***capsule*** denotes the status of the tumor, whether it has penetrated or not, which is considered as the dichotomous outcome variable and ***Prostatic Specimen Antigen Value (PSA)*** in mg/ml as the exogenous variable. Here we consider the simplest case involving one exogenous variable so that our model (3.3) reduces to $z = a + bx$. We obtained the MLEs of the parameters a, b and c of the generalized logit regression model by using the maxLik package in R software which involves the Newton–Raphson optimization procedure (Henningsen and Toomet, 2011). MaxLik package is a function with the same name **maxLik**. This function has two mandatory arguments, logLik and start. The first argument (logLik) is the function that calculates the log-likelihood value as a function of the parameter (usually parameter vector). The second argument (start) is a vector of starting values. This function returns the log-likelihood value, estimated values of the parameters, standard errors and p-values.

For comparisons, we consider the existing tools - probit and logit models, and estimated the parameters a and b of both these models by maximum likelihood estimation method. The computation results obtained in case of each of the models along with the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) as a measure of the goodness of fit are presented in Table1. Also, we have plotted the empirical cumulative distribution of the data set along with the respective models in Figure 1. From Table 1 and Figure 1 it can be observe that the generalized logit model gives better fit to the data set compared to the standard models in the literature- logit and probit models.

Table 1: Estimated values of the parameters and the corresponding value of the AIC and BIC for the different models.

Model	$\hat{a}$	$\hat{b}$	$\hat{c}$	AIC	BIC
Probit	-0.677	0.029	--	467.350	468.516
Logit	-1.114	0.050	--	467.161	468.320
Generalized logit	43.990	7.518	0.004	464.071	465.810

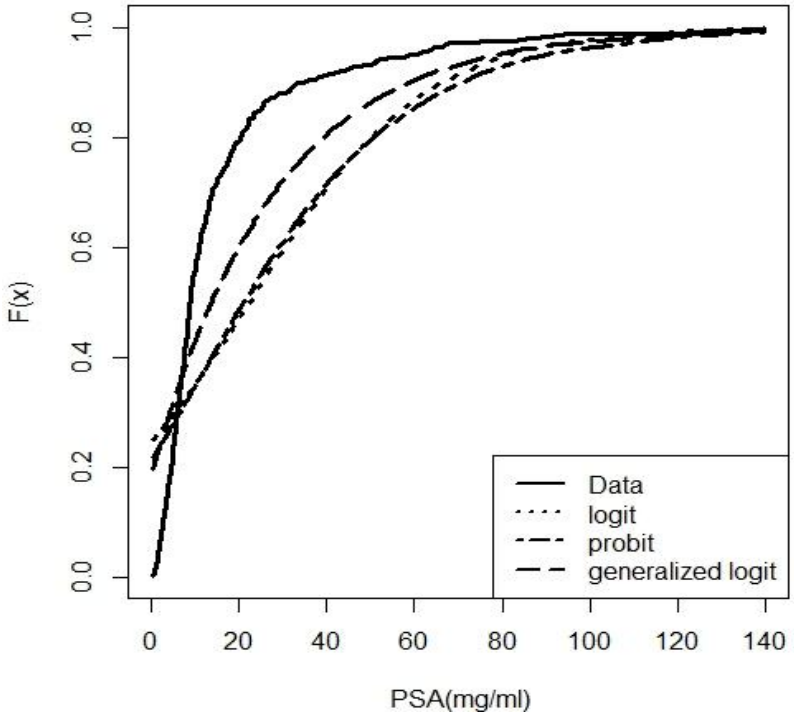

Figure 1**:** Empirical distribution of the data set along with fitted regression models- probit, logit, generalized logit models

5. Testing of Hypothesis and Simulation

In this section we discuss the generalized likelihood ratio test procedure for testing the hypothesis $H_0 : c = 1$ against the alternative hypothesis $H_1 : c \neq 1$ and attempted a brief simulation study. Here the test statistic is,

$$-2\log\Lambda = 2[\log L(\hat{\Omega}; y \mid x) - \log L(\hat{\Omega}^*; y \mid x)] \tag{5.1}$$

where $\hat{\Omega}$ is the maximum likelihood estimator of $\Omega = (a,b,c)$ with no restriction, and is $\hat{\Omega}^*$ the maximum likelihood estimator of Ω when $c = 1$. The test statistic $-2\log\Lambda$ given in (5.1) is asymptotically distributed as χ^2 with one

degree of freedom (Rao, 1973). The computed values of $\log L(\hat{\Omega}; y \mid x)$ is -255.462, $\log L(\hat{\Omega}^*; y \mid x)$ is -260.503 and the value of test statistic is 10.082. Since the critical value at the significance level 0.05 and degree of freedom one is 3.84, the null hypothesis is rejected. Hence it can be concluded that the generalized logit model is more appropriate to the asymmetric data set considered in this paper. In order to assess the performance of the MLEs of the parameters of the generalized logistic regression model, we have conducted a brief simulation study based on values of the following set of parameters, $a = 43.99, b = 7.518, c = 0.004$. Here we utilized the inverse transform method of Ross (1997) for generating random numbers. The computed values of bias and mean square error (MSE) corresponding to sample sizes 100, 200 and 300 respectively are given in Table 2.

Table 2: Bias and Mean square error of each of the parameters of the simulated data sets

Sample size	Bias			MSE		
	a	b	c	a	b	c
100	-0.05714	-0.00344	0.06530	0.01681	0.00019	0.03246
200	0.02100	-0.00160	0.04371	0.01515	0.00013	0.02235
300	0.00831	-0.00045	0.00921	0.01075	0.00007	0.00800

From Table 2 it can be seen that both the absolute bias and MSEs in respect of each parameters of the generalized logit model are in decreasing order as the sample size increases.

Concluding Remarks

The paper introduces the concept of a generalized logit model based on the generalized logistic distribution of Balakrishnan and Leung (1988). We discussed the maximum likelihood estimation of the parameters of the generalized

regression model and demonstrated the procedures by using a real life data set on Prostrate cancer. Generalized likelihood ratio test is considered for testing the significance of the parameters of the model and a brief simulation study is attempted for establishing the better performance of the maximum likelihood estimators of the model.

Acknowledgements

The authors would like to express their sincere thanks to the Editor and the anonymous Referees for carefully reading the paper and for their valuable comments.

References

Agresti, A. (2007): *An Introduction to Categorical Data Analysis,* Second Edition, Wiley, New York.

Aldrich, J.H. and Nelson, F.D. (1984): *Linear Probability, Logit and Probit models,* Sage University Paper series on Quantitative Applications in the Social Sciences, Sage Publications, London.

Balakrishnan, N. and Leung, M.Y. (1988*)*: *Means, variances and covariances of order statistics, BLUEs for the type I generalized logistic distribution, and some applications, "Communications in Statistics - Computation and Simulation",* **17, (1),** 51–84.

Balakrishnan, N. and Hossain, A. (2007): *Inference for the Type II generalized logistic distribution under progressive type II censoring, "Journal of Statistical Computation and Simulation",* **77, (12),** 1013-1031.

Henningsen, A. and Toomet, O. (2011): *MaxLik: A package for maximum likelihood estimation in R, "Computational Statistics",* **26**, 443-458.

Hosmer, D.W. and Lemeshow, S. (2000): *Applied Logistic Regression,* Wiley, New York.

Puig, P. (2003*)*: Characterizing additively closed discrete models by a property of their MLEs, with an application to generalized Hermite distributions, *"Journal of American Statistical Association",* **98**, 687 – 692.

Ross, S.M. (1997): *Simulation,* 2nd edition, Academic Press, Inc: USA.

Rao, C.R. (1973): *Linear Statistical Inference and Its Applications,* Wiley, NewYork.