Biomedical Statistics and Informatics
2017; 2(4): 145-149
http://www.sciencepublishinggroup.com/j/bsi
doi: 10.11648/j.bsi.20170204.13

On the Comparison of Some Link Functions of Binary Response Analysis Under Symmetric and Asymmetric Assumptions

Saddam Adams Damisa[1, *], Musa Tasi'u[1], Salamatu Yusuf Bello[1], Farouq Ndamadu Musa[2], Nurudeen Ayobami Ajadi[3], Samson Agboola[1]

[1]Department of Statistics, Ahmadu Bello University, Zaria, Nigeria

[2]Department of Mathematics and Statistics, Kaduna Polytechnic, Kaduna, Nigeria

[3]Department of Statistics, College of Physical Sciences, Federal University of Agriculture, Abeokuta, Nigeria

Email address:

asdamisa@abu.edu.ng (S. A. Damisa)

[*]Corresponding author

To cite this article:

Saddam Adams Damisa, Musa Tasi'u, Salamatu Yusuf Bello, Farouq Ndamadu Musa, Nurudeen Ayobami Ajadi, Samson Agboola. On the Comparison of Some Link Functions of Binary Response Analysis Under Symmetric and Asymmetric Assumptions. *Biomedical Statistics and Informatics.* Vol. 2, No. 4, 2017, pp. 145-149. doi: 10.11648/j.bsi.20170204.13

Received: August 1, 2017; **Accepted**: August 29, 2017; **Published**: September 23, 2017

Abstract: Binary response analysis is modeled when the response variable is nominal and as such violates the use of the ordinary linear regression model. This paper utilizes the classical approach to fit a categorical response regression model using the logit, probit. loglog and the complementary loglog (Cloglog) link functions under symmetric and asymmetric assumptions. It is captured in past studies that we can only make comparisons between these link functions when n is large say (n > 1000), In this study we compared the link functions to investigate this claim with small values of n less than 1000. We fit the Cloglog and loglog models on 600 tuberculosis patients who may be co-infected with hypertension while the R package was initiated in simulating a binary data for fitting the logit and probit models using the Akaike Information Criterion (AIC) as a basis of comparison for the symmetric and asymmetric different model fitting techniques. The result of the simulated data of sample size 50 revealed that there is a difference between the two symmetric link functions with differing values of AIC with the Probit outperforming the logit link having least values of AIC which indicates that the probit link should be preferred under the symmetric assumption. While under the asymmetric link functions the loglog outperformed the cloglog with smaller values of AIC utilized on the life dataset which gives us the notion that the loglog link should be preferred under the asymmetric assumption. Furthermore table 6 also indicates that type of occupation is the only significant factor associated with hypertension in tuberculosis infected patients under study using both the cloglog and loglog link functions. On this note we recommend that patients with diabetes should be given less strenuous jobs and occupations to handle. Finally we were able to show that the link functions can be distinguished even with small values of (n < 1000) under the two assumptions.

Keywords: Binary Responses Analysis, Loglog, Cloglog, Probit and Logit Links

1. Introduction

A vast literature in statistics is concerned with the analysis of binary response data. Binary responses can be described by generalized linear models [4]. The usual link functions in binary regression models are probit, logit, cloglog and loglog, which are based in the CDF of known distributions. Logit and probit are symmetric links while the cloglog and loglog are asymmetric links. Probit and logit models are among the most widely used members of the family of generalized linear models in the case of binary dependent variable. [16] Proposed a simple Bayesian methods in binary quantile regression. They computed the Bayes factor of all candidate models simultaneously based on a single set of MCMC samples from a model that encompasses ass candidate models. In probit models, the

link function relating the linear predictor ($\eta = x\beta$) to the expected value (μ) is the inverse normal cumulative distribution function, $\Phi^{-1}(\mu) = \eta$. In the logit model the link function is the logit transform, $\ln(\mu/(1-\mu)) = \eta$. The conventional wisdom is that in most cases the choice of the link function is largely a matter of taste. For example [4] concludes the discussion of the issue with the summary "in most applications, it seems not to make much difference." [6] Puts it especially plainly in discussing link functions including the Cloglog, the paper indicates that they "provide identical substantive conclusions" [6]. Elsewhere, similar advice appears regularly when the topic is discussed e.g., [1, 7-11]. Empirical support for the recommendations regarding both the similarities and differences between the probit and logit models can be traced back to results obtained [3]. They found that it was only possible to discriminate between the two models when sample sizes were large and certain extreme patterns were observed in the data. This paper discusses their work and extend it to the comparison of cloglog and loglog under the asymmetric assumption and also comparison of the logit and probit Lines under the symmetric assumption. Many researchers, especially epidemiologists, prefer to fit logit models than probit models because of the odds-ratio interpretation of the logit coefficients. The odds are the ratio of a probability (p) to one minus the probability that is (p/1-p) also because it is considered the default link. You may want to ask if the logit is considered the default link, why we still use probit, loglog and complementary log log links. These are the few reasons.

a. Theoretical Considerations
b. Influences by disciplinary traditions
c. Economists favour probit models
d. Toxicologist favour logit models
e. Underlying characteristics of the data
f. Complementary loglog and loglog works best with extremely skewed distributions.

In economics, researchers are often interested in explaining a limited dependent variable, Y, as a function of a set of explanatory variables, X. Because of the bounded nature of the variable of interest, linear specifications often provide an inadequate description of the conditional mean of Y, E(Y/X), since no restriction is imposed on the range of values taken by the predicted outcome. [1] Says the choice between the logit and probit models is largely one of convenience and convention, since the substantive results are generally indistinguishable.

[2] Examined the choice of link function in binary response models from the Bayesian perspective. As mentioned above, [3] established that under certain conditions it was possible to distinguish the results from probit and logit models. In particular, they were able to distinguish between the link functions when sample sizes were large (e.g., $n \geq 1000$) and where there were what can be termed as "extreme independent variable levels". An extreme independent variable level involves the confluence of three events. First, an extreme independent variable level occurs at the upper or lower extreme of an independent variable. For example, say the independent variable x is to take on the values 1, 1.4, 2, 2.1 and 3.2. The extreme independent variable level would involve the values at x = 3.2 (or x = 1). Second, a substantial proportion (e.g., 60%) of the total n must be at this level. Third, the probability of success at this level should itself be extreme (e.g., at least 99%)

Most of the statistical tests we perform are based on a set of assumptions. When these assumptions are violated the results of the analysis can be misleading or completely erroneous.

Typical assumptions are:

a. Normality: Data have a normal distribution (or at least is symmetric)
b. Homogeneity of variances: Data from multiple groups have the same variance
c. Linearity: Data have a linear relationship
d. Independence: Data are independent We can explore in detail what it means for data to be normally distributed in Normal Distribution, but in general it means that the graph of the data has the shape of a bell curve. Such data is symmetric around its mean and has kurtosis equal to zero. In Testing for Normality and Symmetry we provide tests to determine whether data meet this assumption.

Some tests (e.g. ANOVA) require that the groups of data being studied have the same variance. In Homogeneity of Variances we provide some tests for determining whether groups of data have the same variance.

Some tests (e.g. Regression) require that there be a linear correlation between the dependent and independent variables. Generally linearity can be tested graphically using scatter diagrams or via other techniques explored in Correlation, Regression and Multiple Regression.

We touch on the notion of independence in Definition of Basic Probability Concepts. In general, data are independent when there is no correlation between them (Correlation). Many tests require that data be randomly sampled with each data element selected independently of data previously selected. E.g. if we measure the monthly weight of 10 people over the course of 5 months, these 50 observations are not independent since repeated measurements from the same people are not independent. Also the IQ of 20 married couples doesn't constitute 40 independent observations.

Almost all of the most commonly used statistical tests rely of the adherence to some distribution function (such as the normal distribution). Such tests are called parametric tests. Sometimes when one of the key assumptions of such a test is violated, a non-parametric test can be used instead. Such tests don't rely on a specific probability distribution function (Non-parametric Tests).

Another approach for addressing problems with assumptions is by transforming the data. (Data Transformations)

2. Materials and Methods / Research Methodology

Traditional Bayesian model comparison is performed using Bayes factors [13]. More recently, [14] introduced the Deviance Information Criterion (DIC) which combines measures of both model fit and model complexity. Thus, DIC is similar in interpretation and in spirit to other information-theoretic model comparison criterion, AIC [15] The Akaike Information Criterion (AIC) is one of it and would be used for the comparison in this paper. The four links transform probabilities are;

SYMMETRIC ASSUMPTIONS

$$\text{Logit link function } \eta(p) = \log(\frac{p}{1-p}) \quad (1)$$

$$\text{probit link function } \eta(p) = \varphi^{-1}(p) \quad (2)$$

ASYMMETRIC ASSUMPTIONS

$$\text{Cloglog link function } \eta(p) = \log(\log(1-p)) \quad (3)$$

$$\text{Loglog link function, } \log y_i = \beta_0 + \log \beta_i + e_i \quad (4)$$

$$\text{AIC} = -2\text{loglikelihood} + 2k \quad (5)$$

Where k is the number of parameters estimated

This paper applied the logit and probit link functions on a simulated data of sample size 50. While the loglog and cloglog link functions were fitted on a skewed (asymmetric) life data of tuberculosis patients that may be co-infected with hypertension of sample size 600 gotten from the tuberculosis unit of the Ahmadu Bello University Teaching Hospital Zaria, Nigeria.

Tuberculosis is an infectious disease that usually affects the lungs. Compared with other diseases caused by a single infectious agent, tuberculosis is the second biggest killer, globally.

In 2015, 1.8 million people died from the disease, with 10.4 million falling ill.

In the 18th and 19th centuries, a tuberculosis epidemic rampaged throughout Europe and North America, before the German microbiologist Robert Koch discovered the microbial causes of tuberculosis in 1882.

Following Koch's discovery, the development of vaccines and effective drug treatment led to the belief that the disease was almost defeated. Indeed, at one point, the United Nations, predicted that tuberculosis (TB) would be eliminated worldwide by 2025.

However, in the mid-80s, TB cases began to rise worldwide, so much so, that in 1993, the World Health Organization (WHO) declared that TB was a global emergency; the first time that a disease had been labeled as such.

Fortunately, with proper treatment, the vast majority of cases of tuberculosis are curable. Cases of TB have decreased in the United States since 1993, but the disease remains a concern. Without proper treatment, up to two-thirds of people ill with tuberculosis will die.

Hypertension is another name for high blood pressure. It can severely impact quality of life and it increases the risk of heart disease, stroke, and death.

Around 85 million people in the United States (U.S.) have high blood pressure.

Hypertension and heart disease are global problems. The World Health Organization (WHO) suggests that the growth of the processed food industry has impacted the amount of salt consumed, and that this plays a role in hypertension.

Some types of hypertension can be managed through lifestyle and dietary choices, such as engaging in physical activity, reducing alcohol and tobacco use, and avoiding a high-sodium diet.

Here are some key points about hypertension. More detail is in the main article.

a. Normal blood pressure is 120 over 80 mm of mercury (mmHg), but high blood pressure is higher than 140 over 90 mmHg.
b. Acute causes of high blood pressure include stress, but it can happen on its own or it can result from a condition, such as kidney disease.
c. Unmanaged hypertension can lead to a heart attack, stroke, and other problems.
d. Lifestyle factors are the best way to address high blood pressure.

Blood pressure is the force exerted by the blood against the walls of the blood vessels.

How great the pressure is depends on the work being done by the heart and the resistance of the blood vessels.

Medical guidelines define hypertension as a blood pressure higher than 140 over 90 millimeters of mercury (mmHg).

The systolic reading of 140 mmHg refers to the pressure as the heart pumps blood around the body. The diastolic reading of 90 mmHg refers to the pressure as the heart relaxes and refills with blood.

The hypertension status were dummy coded as 0 for (absence of hypertension) 1 for (presence of hypertension) while the covariates were Age, Gender and Occupation which were regressed on the hypertension status of the 600 patients.

3. Analysis

Binary Response Analysis Using the R Package.

3.1. Simulated Data Analysis

```
> a= rbinom(n=50, size=1, p=0.5)
> x= rbinom(n=50, size=1, p=0.5)
> y= rbinom(n=50, size=1, p=0.5)
> z= rbinom(n=50, size=1, p=0.5)
> dataplus = data.frame(a, x, y, z)
hp = read.table("C:/data/hyp.dat", header=T)
> logit = glm(a~x+y+z, family=binomial (link="logit"), dataplus)
> probit = glm(a~x+y+z, family=binomial (link="probit"), dataplus)
```

3.2. Data Analysis on Tuberculosis Hypertensive Patients

```
hp = read.table("C:/data/hyp.dat", header=T)
> clog = glm (hyp~age+gnd+occ, family=binomial (link="cloglog"), data=hp)
> loglog = glm (hypp~age+gnd+occ, family=binomial (link="cloglog"), data=hp)
```

4. Results

Table 1. *Descriptives of the dependents variable (a) for the logit and probit links.*

			Statistic	Std. Error
Data(a)	Mean		.3800	.06934
	95% Confidence Interval for Mean	Lower Bound	.2407	
		Upper Bound	.5193	
	5% Trimmed Mean		.3667	
	Median		.0000	
	Variance		.240	
	Std. Deviation		.49031	
	Minimum		.00	
	Maximum		1.00	
	Range		1.00	
	Interquartile Range		1.00	
	Skewness		.510	.337
	Kurtosis		-1.814	.662

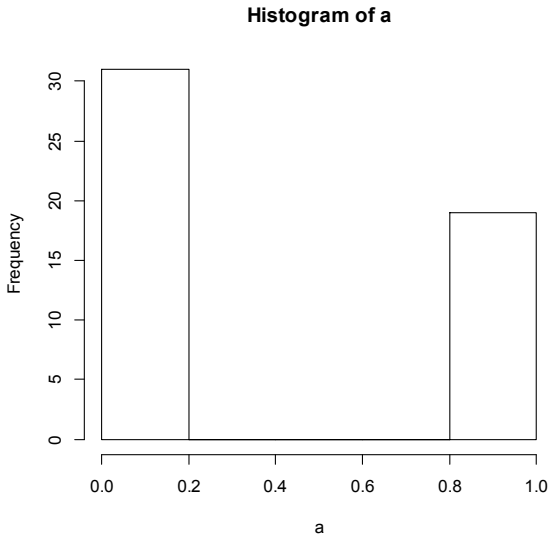

Figure 1. *Histogram of the data (a).*

Table 2. *Descriptive of the life data on tuberculosis hypertensive patients.*

		Statistic	Std. Error
Mean		93	011
95% Confidence Interval for Mean	Lower Bound	91	
	Upper Bound	95	
5% Trimmed Mean		98	
Median		1.00	
Variance		067	
Std. Deviation		258	
Minimum		0	
Maximum		1	
Range		1	
Interquartile Range		0	
Skewness		-3.330	100
Kurtosis		9.116	199

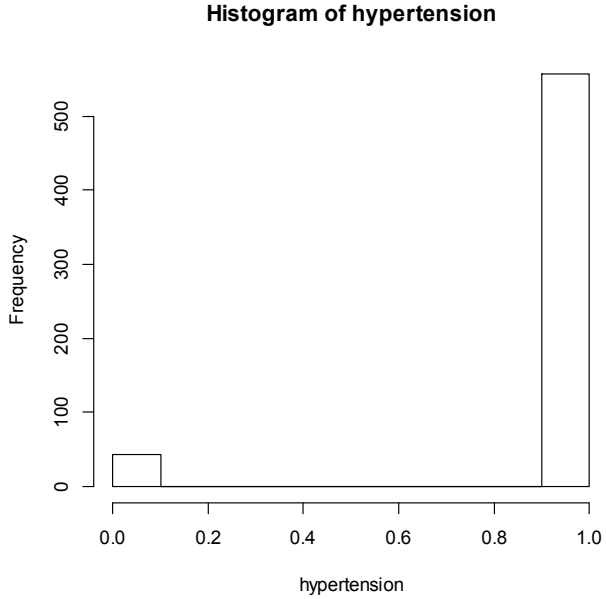

Figure 2. *Histogram of the data on Hypertension.*

Table 3. *Result of simulated data for the logit link.*

	Logit			
	estimate	Std error	Z value	Pr(>|z|)
(Intercept)	0.8427	0.6522	1.292	0.1963
x	0.8461	0.6480	1.306	0.1916
y	-1.1191	0.6651	-1.168	0.0924
z	0.2029	0.6475	0.313	0.7540

Table 4. *Result of simulated data for the probit link.*

	Probit			
	estimate	Std error	Z value	Pr(>|z|)
(Intercept)	0.5046	0.3909	1.291	0.1968
x	0.5356	0.3880	1.380	0.1675
y	-0.6899	0.3931	-1.755	0.0792
z	0.1423	0.3882	0.366	0.7140

Table 5. *Information criterion table of simulated data.*

	Link functions	
	LOGIT	PROBIT
AIC	66.59990	66.22168

Table 6. *Result of perception data for the logit link.*

	Cloglog			
	estimate	Std error	Z value	Pr(>|z|)
(Intercept)	17.3736	3280.2851	0.005	0.996
Age	-14.4397	1288.3601	-0.011	0.991
Gender	-1.0913	2404.2972	0.000	1.000
Occupation	-2.1006	0.3536	-5.940	0.000***

Table 7. *Result of perception data for the cloglog link.*

	Loglog			
	estimate	Std error	Z value	Pr(>|z|)
(Intercept)	-2.66739	242.7068	-0.011	0.991
Age	0.88642	86.13173	0.010	0.992
Gender	0.05639	191.4441	0.000	1.000
Occupation	1.26320	0.22055	5.727	0.000***

Table 8. *Information criterion table on tuberculosis hypertensive patients.*

	Link Function	
	CLOGLOG	**LOGLOG**
AIC	151.5153	149.7694

5. Discussions

Some literatures like [3] have established that comparison can only take place between the link functions when the sample size is large say (n≥1000), but from our investigations in this study using a simulated data of size 50 under the symmetric assumptions we were able to establish that there is a difference between the logit and probit link functions with the probit link having the least AIC in comparison with the logit link. While loglog outperformed the cloglog by having the least AIC value when applied on a life data with sample size 600 of tuberculosis patients who may be co-infected with hypertension also table 6 further indicates that the type of occupation is the only significant factor associated with hypertension in the tuberculosis infected patients in both the cloglog and loglog link functions. The Age and Sex of the tuberculosis infected patients under study were not significantly associated with hypertension at 0.05 level of significance as can be observed in tables 6 and 7 respectively.

6. Conclusion

From our study we found out that it is possible to distinguish between the link functions when the size of n is less than 1000. Also the probit link outperformed the logit link function as such the probit link should be preferred in usage of a binary response data under the symmetric assumption while the loglog link function outperformed the cloglog link which indicates that the loglog link should the preferred in analysis of a binary response data under the asymmetric assumption. Also both the loglog and the cloglog link functions indicated that occupation is the only significant factor associated with tuberculosis-hypertensive patients in Ahmadu Bello University Teaching hospital, Zaria, Nigeria. Therefore further care should be put in place to reduce the work load and probably determine the nature of occupation to be carried out tuberculosis infected patients to reduce the risk of being co-infected with hypertension. The patients already co-infected with hypertension should be probably managed by medical experts to relieve them of the illness.

Acknowledgements

The authors would like to thank the anonymous reviewers for their time.

References

[1] Long, J. S. (1997). Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage.

[2] Albert, J. H. and S. Chib (1993). Bayesian Analysis of Binary and Polychotomous Response Data. Journal of the American Statistical Association 88, 669–679.

[3] Chambers, E. A. and D. R. Cox (1967). Discrimination between alternative Binary Response Models. Biometrika 54, 573–578.

[4] Mc Cullagh, P., and Nelder, J. A. (1989), Generalized Linear Models, 2nd ed, London: Chapman and Hall.

[5] Gill, J. (2001). Generalized Linear Models: A Unified Approach. Thousand Oaks, CA: Sage.

[6] Maddala, G. S. (1983). Limited-Dependent and Qualitative Variables in Econometrics. Cambridge: Cambridge University Press.

[7] Davidson, R. and J. G. MacKinnon (1993). Estimation and Inference in Econometrics. New York: Oxford.

[8] Powers, D. A. and Y. Xie (2000). Statistical Methods for Categorical Data Analysis. San Diego: Academic Press.

[9] Fahrmeir, L. and G. Tutz (2001). Multivariate Statistical Modelling Based on Generalized Linear Models (2nd ed.). New York: Springer.

[10] Hardin, J. and J. Hilbe (2001). Generalized Linear Models and Extensions. College Station, TX: Stata Press.

[11] Krejcie, R. V. and D. W. Morgan (1970) Determining Sample Sizes for Research Activities. Educational and Psychological Measurement 30, 607-610.

[12] Kass, R. E. and A. E. Raftery (1995). Bayes Factors. Journal of the American Statistical Association 90, 773–794.

[13] Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002). Bayesian Measures of Model Complexity and Fit (with discussion). Journal of the Royal Statistical Society: Series B 64, 583–639.

[14] Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In B. N. Petrov and F. Csaki (Eds.), Proceedings of the Second International Symposium on Information Theory, pp. 267–281. Budapest: Akademiai Kiado. Reprinted in breakthroughs in Statistics, vol. 1, pp. 610-624, eds. Kotz, S.

[15] Man-suk, Oh, Eun, Sug Park and Boeng-Soo, So (2016) Bayesian Variable Selection in Binary Quantile Regression. Statistics and Probability Letters.