On the Comparison of Some Link Functions In Binary Response Analysis

S. A. Damisa[1]*; F. Musa[2]; S. Sani[1]

[1]Department of Statistics,
Ahmadu Bello University,
Zaria-Nigeria.
email: asdamisa@abu.edu.ng*

[2]Department of Mathematics and Statistics,
Kaduna Polytechnic,
Kaduna, Nigeria.
email: famusamusa08@gmail.com[2]

***Abstract*—Binary response analysis is modeled when the response variable is nominal and as such violates the use of the ordinary linear regression models. This paper utilizes the classical approach to fit a categorical response regression model using the logit, probit and the complementary log log (Cloglog) link functions. It is captured in past studies that we can only make comparisons between these link functions only when n is large say (n>1000),In this study we fit these models on the participation in cybercrime among youths using the Akaike information criterion (AIC) and the Bayesian information Criterion (BIC) as a basis of comparison for the three different model fitting techniques on responses on perception of cybercrime at Federal College of Education Abeokuta. We adopted the use of questionnaire as a tool for gathering data from the respondents using a sample size of 50 as opposed to the (n>1000) to see if there are differences between the link functions. The R package was initiated in running the inferential statistics. The result of the simulated data of sample size 50 revealed that there are differences between the three link functions with differing values of AICs and BICs with the cloglog outperforming the logit and probit analysis. While the link functions on the data on participation in cybercrime had the same (AICs) and (BICs) due to the zero inflation of the response variable and not the small size of n. This implies that any of them would fit well in the choice of modeling the binary responses of participation in cybercrime in this research and related researches that portray similar characteristics otherwise the Cloglog should be adopted based on the nature of the data.**

Keywords-*Binary Responses analysis, Cybercrime, CLogLog, Probit and Logit links.*

I. INTRODUCTION

A vast literature in statistics, certain aspects of biometrics and econometrics researches is concerned with the analysis of binary response data. Binary responses can be described by generalized linear models McCullagh and Nelder [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.

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, $\ln(\mu/(1-\mu)) = \eta$. Gill [5] puts it especially plainly in discussing link functions including the Cloglog, he indicates that they "provide identical substantive conclusions" Gill[5]. Elsewhere, similar advice appears regularly when the topic is discussed e.g.,Maddala, [6]; Davidson and MacKinnon, [7]; Long, [1]; Powers and Xie, [8]; Fahrmeir and Tutz, [9]; Hardin and Hilbe, [10].

Empirical support for the recommendations regarding both the similarities and differences between the probit and logit models can be traced back to results obtained by Chambers and Cox [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. We discussed their work and extend it to the cloglog analysis in comparison with the logit and probit analysis.

Many researchers, especially epidemiologists, prefer to fit logit models than probit models because of the odds-ratio interpretation of the logit coefficients. The odds ratio are the probability(p)of an event occurring to the probability(q) of the event not occurring that is (P/q).Also the logitlink is considered the default link. You may want to ask if the logit is considered the default link, then why do we still use probit and Complementary log loglinks. These are the few reasons.

- Theoretical Considerations
- Influences by disciplinary traditions
  - Economists favourprobit models
  - Toxicologist favour logit models
- Underlying characteristics of the data
  - Complementary log log works best with extremely skewed distributions.

Long [1] says the choice between the logit and probit models is largely one of convenience and convention, since the substantive results are generally indistinguishable.

Albert and Chib [2] examined the choice of link function in binary response models from the Bayesian perspective. As mentioned above, Chambers and Cox [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 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 were to take on the values 1, 2, 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., greater than 99%)

II. Research Methodology

Traditional Bayesian model comparison is performed using Bayes factors Kass and Raftery [12]. More recently, Spiegelhalter *et al.* [13] 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 (Akaike, [14]). Which is the Akaike Information Criterion (AIC) and the Bayesian Information Criterion Which would be used for the comparison in this paper. The three links transform probabilities are;

Cloglog link function: $\eta(p) = \log(-\log(1 - P))$ (1)

Logit link function $\eta(p) = \log\left(\frac{p}{1-p}\right)$ (2)

probit link function $\eta(p) = \Phi^{-1}(p)$ (3)

We applied the logit, probit and Clog log analysis on dummy variables of perception of cybercrime among tertiary education students with 3 independent variables which are the Knowledge of cybercrime (x), ever being a victim of cybercrime(y) and perception about those who partake in it (z) and where all dummy coded with yes = 1 for participation and No = 0 for non-participation, yes = 1 for Knowledge about cybercrime and No = 0 for no knowledge about cybercrime, yes = 1 for ever being a victim of cybercrime and No = 0 for never being a victim of cybercrime and good = 1 & bad = 0 for perception of cybercrime respectively.

TABLE 1: CYBERCRIME DATA

Sn	Pt	Kw	Vt	Pc	Sn	Pt	Kw	Vt	Pc
1	0	1	0	0	26	0	1	0	0
2	0	0	0	0	27	0	1	0	0
3	0	1	0	1	28	0	1	0	0
4	0	1	0	1	29	0	1	0	0
5	0	1	0	1	30	0	1	0	0
6	0	1	0	0	31	0	1	1	0
7	0	1	0	1	32	0	1	0	0
8	0	1	0	1	33	0	1	0	0
9	0	1	0	1	34	0	1	0	0
10	0	1	0	0	35	0	1	0	0
11	0	1	0	1	36	0	1	0	0
12	0	1	0	0	37	0	1	0	0
13	0	0	0	0	38	0	1	0	0
14	0	0	0	0	39	0	1	0	0
15	0	1	1	0	40	0	1	0	0
16	0	1	0	0	41	0	1	1	0
17	0	1	0	0	42	0	1	0	0
18	0	1	0	0	43	0	1	1	0
19	0	1	0	1	44	0	1	1	0
20	0	1	0	0	45	0	1	0	0
21	0	1	0	0	46	0	1	0	0
22	0	1	0	0	47	0	1	0	0
23	0	1	0	0	48	1	1	1	1
24	0	1	0	0	49	0	1	0	0
25	0	0	0	0	50	1	1	1	1

Sn – Serial Number
Pt – participation
Kw- knowledge about cybercrime
Vt– victim of cybercrime
Pc – Perception about cybercrime

We targeted a population of size 60 in a class at the Federal College of Education, Abeokuta during a reading

session of the students from various departments. Questionnaires were structured based on the above configuration to elicit information from these students. Fifty-two questionnaires were administered based on the students that were present at that period out of which fifty completed questionnaires were returned, an indication of about 96% response rate which is quite reasonable [11]. So, the sample size used for this study is 50 and the final data collected, after coding, are presented in Table 1.

In a Monte-Carlo study, data were simulated from Binomial distribution based on the above four variables' definitions (a, x, y and z) and configurations using sample of size 50 using the R package. These four variables were simulated from a binomial distribution with probability of success (having a yes outcome) set at 0.5 for all the variables. The three variables x. y and z were regressed on response variable (a) that represents whether a responded has participated in cybercrime before or not. This process was repeated three different times in order to be sure about the consistency of the results to enable valid comparisons.

III. Analysis

Binary Response Analysis Using the R Package.

Simulated data Analysis: R codes

```
> 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)

> logit =
glm(a~x+y+z,family=binomial(link="logit"),dataplus)
>probit =
glm(a~x+y+z,family=binomial(link="probit"),dataplus)
>cloglog =
glm(a~x+y+z,family=binomial(link="cloglog"),dataplus)
```

Analysis on perception of cybercrime (Dummy coded variables)

```
> logit =
glm(pct~knwdg+victim+percptn,family=binomial(link="log
it"),data=cyber)
>probit =
glm(pct~knwdg+victim+percptn,family=binomial(link="pro
bit"),data=cyber)
>cloglog =
glm(pct~knwdg+victim+percptn,family=binomial(link="clo
glog"),data=cyber).
```

IV. Results

TABLE 2: 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 3: 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 4: RESULT OF SIMULATED DATA FOR THE CLOGLOG LINK

	cloglog			
	estimate	*Std error*	*Z value*	*Pr(>|z|)*
(Intercept)	0.07478	0.38283	0.195	0.8451
x	0.61731	0.38629	1.598	0.1100
y	-0.72296	0.38017	-1.896	0.0577
z	0.22558	0.37948	0.594	0.5522

TABLE 5: INFORMATION CRITERION TABLE OF SIMULATED DATA

	LINK FUNCTIONS		
	LOGIT	*PROBIT*	*CLOGLOG*
AIC	66.59990	66.22168	65.51923
BIC	73.86978	73.70808	73.16732

Results For Real life data (Dummy coded variables)

TABLE 6: RESULT OF PERCEPTION DATA FOR THE LOGIT LINK

	logit			
	estimate	*Std error*	*Z value*	*Pr(>|z|)*
(Intercept)	-26.25	178062.0	0.000	1.000
knowledge	-45.19	194875.6	0.000	1.000
Victim	47.72	65162.8	0.001	0.999
Perception	47.28	65710.8	0.001	0.999

TABLE 7: RESULT OF PERCEPTION DATA FOR THE PROBIT LINK

	probit			
	estimate	*Std error*	*Z value*	*Pr(>|z|)*
(Intercept)	-6.991	36990.34	0.000	1.000
knowledge	-12.688	39850.95	0.000	1.000
Victim	13.102	12228.94	0.001	0.999
Perception	13.032	12285.69	0.001	0.999

TABLE 8: RESULT OF PERCEPTION DATA FOR THE CLOGLOG LINK

	cloglog			
	estimate	*Std error*	*Z value*	*Pr(>|z|)*
(Intercept)	-26.48	170241.0	0.000	1
knowledge	-25.33	182050.0	0.000	1
Victim	27.73	45151.00	0.001	1
Perception	27.31	46325.78	0.001	1

TABLE 9: INFORMATION CRITERION TABLE OF PERCEPTION DATA

	LINK FUNCTIONS		
	LOGIT	*PROBIT*	*CLOGLOG*
AIC	8.0	8.0	8.0
BIC	15.64809	15.64809	15.64809

V. DISCUSSIONS

Some literatures, as mentioned earlier, have established that we can only discriminate between the link functions when the sample size is large say ($n \geq 1000$), but from our investigations in this paper using a simulated data of sample size 50 we were able to establish that there are differences between the link functions with the Cloglog having the least AIC and BIC in comparison with the other two link functions (logit and probit). Though, the interpretations in the three links were the same for the simulated and real life data based on the significance of the models' parameters.

Furthermore there was no distinction in the link functions of the Cloglog, logit and probit analysis based on their AICs and BICs this could be as a result of the inflation of the zeros on the binary response and not the due to the small sample size of 50. The choice of the link function as a result of this study should be based on the nature of the data.

The results from real life data simply indicated that knowledge about cybercrime, ever been a victim of cybercrime and perception of those involved in it does not significantly contribute to youth participation in cybercrime at 0.05 level of significance in all the three models of youth participation in cybercrime.

VI. CONCLUSION

It is concluded that any of the link functions of the binary response analysis is as suitable as its counterparts in fitting a model for the participation in cybercrime among youth that portrays a zero inflated response data. While the simulated data suggests that the Cloglog performs better than the other two link functions which is as a result of the nature of the data.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewer for their useful comments and observations. The authors equally thank the Chairman, editorial board for effecting necessary corrections on the original manuscript to improve the work.

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] McCullagh, 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. NewYork: 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.