A New Generalized Class of the Complementary LogLog Link Function

Brian Bader

Department of Statistics

University of Connecticut

Email: brian.bader@uconn.edu

December 13, 2014

Abstract

The complementary loglog (cloglog) link is one of the most widely used links to model asymmetric binomial response data. While symmetric response functions have the same rate as $F(x)$ approaches 0 or 1, asymmetric response functions may have different rates as $F(x) \to 1$ and $F(x) \to 0$. This new generalized class of the cloglog link allows for greater flexibility of the rates for each tail individually. This provides better model fitting, with the known cloglog being a special case. Properties of this new link function class are proven and discussed. Two real data applications are provided to demonstrate its use and compare with other existing links.

KEY WORDS: complementary loglog, link function, generalized cloglog, general class of asymmetric link functions

1. INTRODUCTION

The generalized cloglog link function provides a new alternative to modeling binary responses when the assumption of symmetry is not satisfied. It allows for greater flexibility in modeling the shape/rate of the tails of the response curve. The addition of two parameters to the (simplified) cloglog link function provide this flexibility without over-complicating the model. The known, simplified cloglog link can be shown to be a special case of the generalized cloglog link. Although the logit link is arguably the most popular and has a nice interpretation, the generalized cloglog can actually provide better overall fit to the data due to its flexibility.

A link function is used to model the probability of a 'success' via a Generalized Linear Model (GLM) applied to binary or binomial response data. The three most popular and well-known link functions (logit, probit, and simplified cloglog) are discussed in Agresti (2014). Recently, there have been new additions on how to model both positive and negative skewness via flexible link functions. Stukel (1988) proposed a class of generalized logistic models that depend only upon two parameters, dictated by the data. Stukel showed that the model can approximate many known link functions, such as the three discussed above. Aranda-Ordaz (1981) proposed using two separate, one-parameter models and Guerrero and Johnson (1982) used the Box-Cox transformation applied to the odds ratio to achieve a more flexible link. (Chen et al. 1999) propose a latent variable random effects model to introduce skewness into the link function. Later, (Kim et al. 2008) develop a class of generalized skewed t-link models under a latent variable model.

The most similar approaches to the generalized cloglog link function are proposed by Nagler (1994), Gupta and Gupta (2008), and Wang and Dey (2010). All introduce power parameters into the cumulative distribution function to allow the skewness flexibility. The Scobit model (Nagler 1994) introduces a power parameter to the logistic model. Gupta and Gupta (2008) propose using the power normal distribution and discuss the differences from the skew normal distribution. Wang and Dey (2010) use a symmetric baselink link function

and introduce a power parameter to model skewness in both positive and negative directions.

This paper is organized as follows. In the next section, several mathematical properties are proven and discussed for the generalized cloglog link. Section 3 shows two data applications for the generalized cloglog and compares its results to several other known link functions. Section 4 provides a discussion of the advantages and disadvantages of the generalized cloglog link function in practice and in theory.

2. PROPERTIES

The generalized cloglog link function is taken from the standard Kumaraswamy Gumbel distribution, which has CDF given by $F(x) = 1 - \{1 - \exp(-au)\}^b$ for $a > 0$, $b > 0$, $u = \exp(-x)$, and $x \in (-\infty, +\infty)$. Cordeiro et al. (2012) derived this distribution, based on the work of Kumaraswamy (1980). See Eljabri (2013) for a detailed discussion of its properties. Thus, define the response curve:

$$\pi_i = F(\mathbf{x_i}\boldsymbol{\beta}) = \{1 - \exp(-a\exp(\mathbf{x_i}\boldsymbol{\beta}))\}^b$$

and link function

$$F^{-1}(\pi_i) = \mathbf{x_i}\boldsymbol{\beta} = \log\Big(\frac{-\log(1-{\pi_i}^{1/b})}{a}\Big)$$

for $a > 0$ and $b > 0$. It is clear that when $a = b = 1$, this reduces to the simplified cloglog link function.

Theorem 1: The generalized cloglog is asymmetric and positive skewed.

Proof of Theorem 1: Let

$$r = \lim_{u\to\infty} \frac{1 - F(u)}{F(-u)}.$$

If r=0, then it can be said that the link F^{-1} is positive skewed or skewed to the left. Here,

$$F(u) = \{1 - \exp(-a\exp(u))\}^b$$

so

$$r = \lim_{u \to \infty} \frac{1 - F(u)}{F(-u)} = \frac{1 - \{1 - \exp(-a \exp(u))\}^b}{\{1 - \exp(-a \exp(-u))\}^b} = \frac{0}{0} \tag{1}$$

where $a > 0$ and $b > 0$ are fixed. But, it can be shown that $r = 0$. This is due to the fact that

$$\lim_{u \to \infty} \frac{\exp(-a \exp(u))}{1 - \exp(-a \exp(-u))} = 0$$

which implies that the numerator in (1) approaches zero faster than the denominator. Thus, $r = 0$ and the generalized cloglog is asymmetric and positive skewed. An alternate proof may use L'Hopital's rule. ■

To explore the link properties, changes in the parameters a and b are explored. Denote

$$F_{(a,b)}(u) = \{1 - \exp(-a \exp(u))\}^b.$$

For $b_1 > b_2$, with a fixed:

$$\frac{F_{(a,b_1)}(u)}{F_{(a,b_2)}(u)} = \{1 - \exp(-a \exp(u))\}^{b_1/b_2} < 1$$

for every $u < \infty$.

For $a_1 > a_2$, with b fixed:

$$\frac{F_{(a_1,b)}(u)}{F_{(a_2,b)}(u)} = \left\{\frac{1 - \exp(-a_1 \exp(u))}{1 - \exp(-a_2 \exp(u))}\right\}^b > 1$$

for every $u < \infty$.

These two arguments show that if a is fixed and $b_1 > b_2$, $F_{(a,b_1)}(u) < F_{(a,b_2)}(u)$ for all u. This can be seen in Figure 4. Similarly, if b is fixed and $a_1 > a_2$, $F_{(a_1,b)}(u) > F_{(a_2,b)}(u)$ for all u. This can be seen in Figure 5. For changes in both parameters, there is no clear boundary between the different link functions - i.e. the two link functions may cross at some point, specifically when $F_{(a_1,b_1)}(u) = F_{(a_2,b_2)}(u)$. This can be seen in Figure 3.

3. APPLICATION TO TWO DATASETS

3.1 Age of Menarche in Warsaw

The data (Table 4) examines the proportions of female children at various ages during adolescence who have reached menarche. The sample was taken from Warsaw in 1965. The data has three columns, with each row representing an age group. Variable **Total** refers to the total number of observations in the age group, **Menarche** refers to the number of children that reached menarche, and **Proportion** is defined as **Menarche**/**Total**.

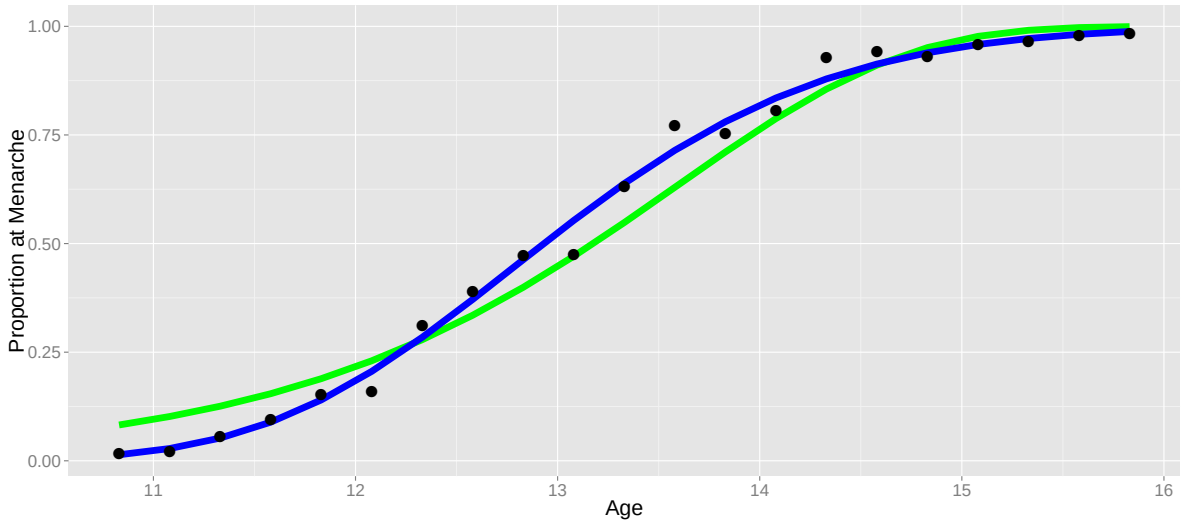

Figure 1: Warsaw menarche proportions by age. Blue curve is the fitted Generalized CLogLog link ($\hat{a}$=3.000, $\hat{b}$=197.335). Green curve is the CLogLog link (a=b=1). The actual proportions from the data are shown for reference.

The generalized linear model used is $\pi_{age} = F(\beta_0 + \beta_1 age)$ where F refers to the response function. Table 1 shows the fitted coefficient estimates and standard errors, AIC, and residual deviance to the menarche data for four different link functions. The simplified cloglog arguably performs the worst, having the largest AIC and residual deviance. The generalized cloglog has the lowest AIC and residual deviance of the link functions analyzed here.

Here, it is clear that the generalized cloglog performs significantly better than the sim-

plified cloglog, but it can be tested using the likelihood ratio test.

$H_0 : a = b = 1$

H_1 : At least one of a or b are not equal to 1.

$W = -2\ln[L(H_0)/L(H_1)] = 104.487$ follows χ^2_2 under H_0, so the null hypothesis is clearly rejected and the generalized cloglog model is necessary.

Table 1: Warsaw Menarche Data Output. For Generalized CLogLog, ($\hat{a}$=3.000, $\hat{b}$=197.335).

	β_0			β_1				Residual
Model	Est.	S.E.	P-Val.	Est.	S.E.	P-Val.	AIC	Deviance
Probit	-11.819	0.387	<0.001	0.908	0.030	<0.001	110.940	22.887
Logit	-21.226	0.771	<0.001	1.632	0.059	<0.001	114.760	26.703
CLogLog	-12.985	0.426	<0.001	0.953	0.031	<0.001	206.870	118.820
Gen. CLogLog	-1.791	0.079	<0.001	0.187	0.006	<0.001	102.390	14.334

3.2 Bacteria Growth Data

The bacteria growth data (Table 3) originates from growing the bacteria *V. natriegens* in a flask for 160 minutes. The population density was measured every 16 minutes, and the time index variable t measures the number of these 16 minute intervals since the beginning of the experiment. The 'population density' units are defined as absorbance measured by a spectrophotometer. The maximum density possible is one and the minimum is zero. A purely exponential growth curve is not appropriate here, as the population growth appears to slow over time.

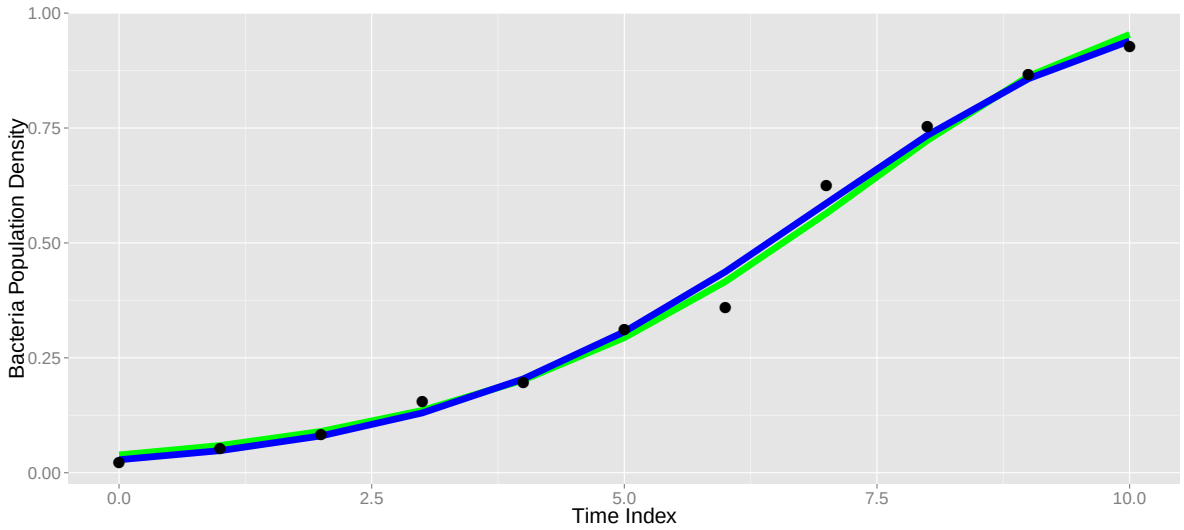

Figure 2: Bacteria population density by indexed time. Blue curve is the fitted Generalized CLogLog link ($\hat{a}$=3.212, $\hat{b}$=2.076). Green curve is the CLogLog link (a=b=1). The actual densities from the data are shown for reference.

The generalized linear model used is $\pi_{time} = F(\beta_0+\beta_1 time)$ where F refers to the response function. Again, four link functions are fit to the data: logit, probit, simplified cloglog, and generalized cloglog. From the AIC and residual deviance output in Table 2, an asymmetric link function is most appropriate here. Both the generalized and simplified cloglog have lower AIC and residual deviances than the two symmetric links. The generalized cloglog outperforms all the links, with the lowest AIC and residual deviance. Hence, the generalized cloglog with $\hat{a}$=3.212 and $\hat{b}$=2.076 has the best overall fit to the data.

As in the previous example, the LRT can be performed to test if the simplified cloglog link should be used or not.

$H_0 : a = b = 1$

H_1 : At least one of a or b are not equal to 1.

$W = -2\ln[L(H_0)/L(H_1)] = 1.057$ follows χ^2_2 under H_0, so the null hypothesis cannot be

rejected. The simplified cloglog link may be appropriate to use here.

Table 2: Bacteria Growth Data Output. For Generalized CLogLog, ($\hat{a}$=3.212, $\hat{b}$=2.076).

	β_0			β_1				Residual
Model	Est.	S.E.	P-Val.	Est.	S.E.	P-Val.	AIC	Deviance
Logit	-3.797	2.237	0.090	0.609	0.351	0.083	9.106	0.070
Probit	-2.169	1.147	0.059	0.350	0.182	0.054	9.280	0.090
CLogLog	-3.233	1.694	0.056	0.435	0.229	0.057	9.085	0.061
Gen. CLogLog	-2.787	1.054	0.008	0.288	0.150	0.056	9.058	0.043

4. DISCUSSION

The generalized cloglog link allows further specification of the tails in order to provide a more appropriate fit to certain data. The logit, probit, and simplified cloglog do not account for this and may not be sufficient in all settings. Other flexible link functions, such as Stukel have been developed but are not as easily interpreted and implementable. Section 2 discussed and proved the role of the two parameters in the generalized cloglog. For the Warsaw menarche dataset, the generalized cloglog has the lowest AIC and residual deviance out of the four models. It was also shown through the likelihood ratio test that the simplified cloglog is not appropriate to use for this dataset. For the bacteria growth dataset, while the generalized cloglog had the lowest AIC and residual deviance of the four models, the likelihood ratio test did not reject the hypothesis of the simplified cloglog link, so it may be most appropriate to use the simpler model there. An alternative to the likelihood ratio test would be to include the two parameters a and b in the AIC calculations (which they are not in this paper). Essentially, the revised AIC_{new} would equal $\text{AIC}_{old} + 4$.

4.1 Further Work

- Comparisons to other flexible link functions, such as the Stukel and Scobit. What is the relationship between the Stukel link and generalized cloglog?
- A Bayesian perspective for analysis under the generalized cloglog link. How to construct priors for the parameters a and b?

5. APPENDIX & CODE

Table 3: Bacteria Growth Dataset

Time	Time Index	Population Density
0	0	0.022
16	1	0.052
32	2	0.082
48	3	0.154
64	4	0.196
80	5	0.312
96	6	0.360
112	7	0.625
128	8	0.753
144	9	0.867
160	10	0.928

Table 4: Warsaw Menarche Dataset

Age	Total	Menarche	Proportion
9.21	376	0	0.000
10.21	200	0	0.000
10.58	93	0	0.000
10.83	120	2	0.017
11.08	90	2	0.022
11.33	88	5	0.057
11.58	105	10	0.095
11.83	111	17	0.153
12.08	100	16	0.160
12.33	93	29	0.312
12.58	100	39	0.390
12.83	108	51	0.472
13.08	99	47	0.475
13.33	106	67	0.632
13.58	105	81	0.771
13.83	117	88	0.752
14.08	98	79	0.806
14.33	97	90	0.928
14.58	120	113	0.942
14.83	102	95	0.931
15.08	122	117	0.959
15.33	111	107	0.964
15.58	94	92	0.979
15.83	114	112	0.982
17.58	1049	1049	1.000

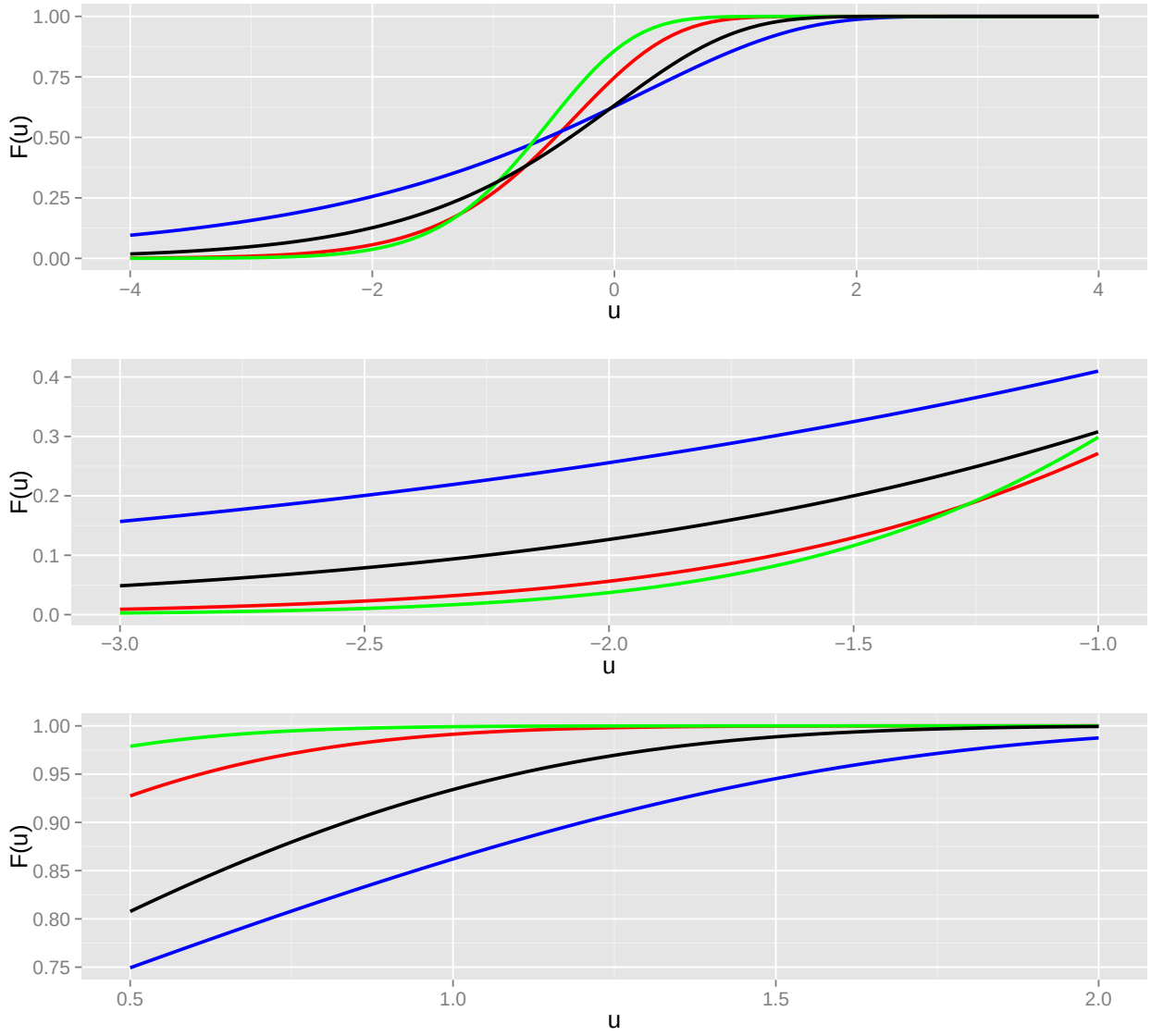

Figure 3: Plots for the Generalized CLogLog link function when $a = b$.

Red:(a,b)=(2,2) Blue:(a,b)=(0.5,0.5) Green:(a,b)=(3,3) Black:(a,b)=(1,1)

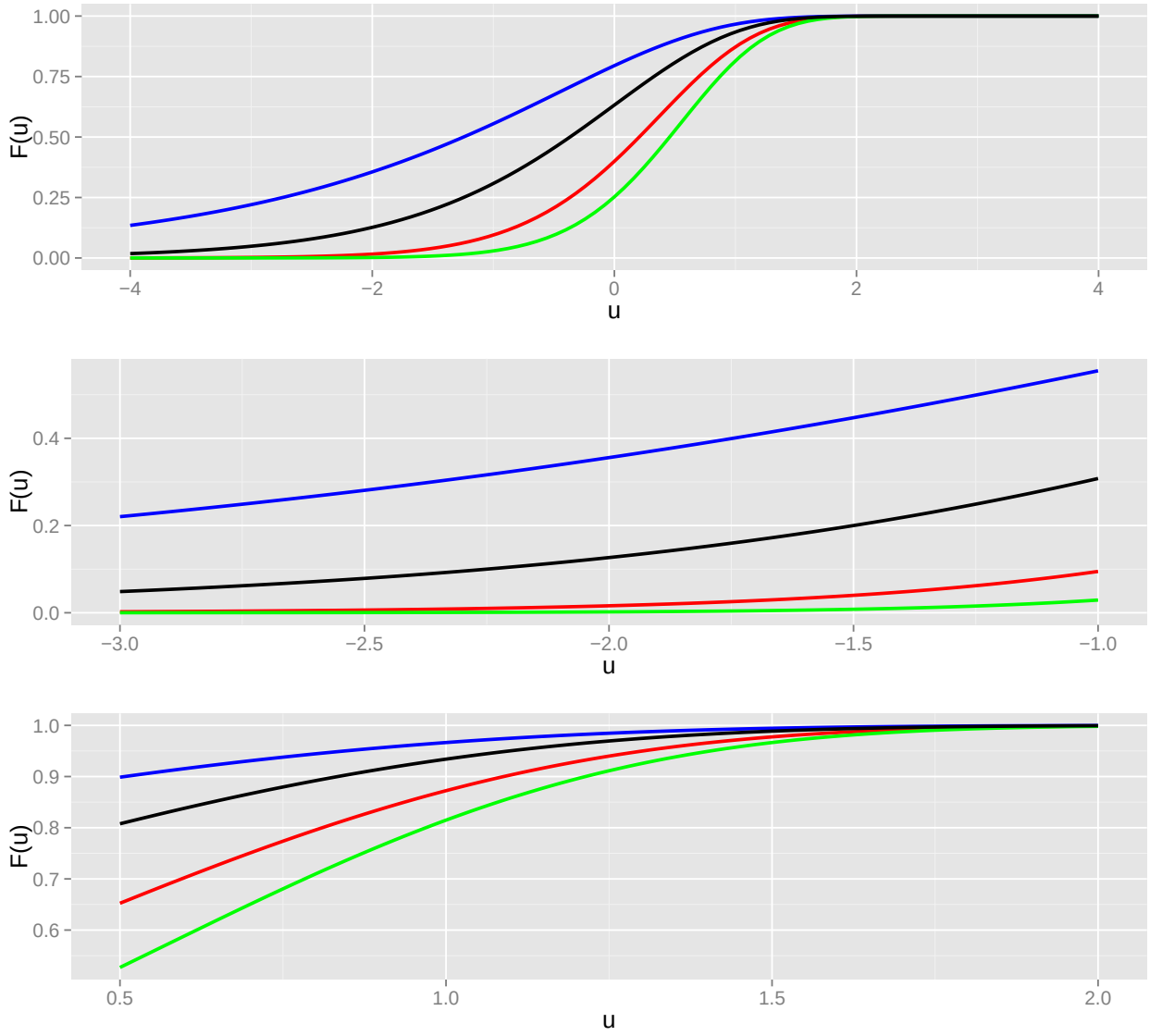

Figure 4: Plots for the Generalized CLogLog link function when $a = 1$ and b varied.

Red:(a,b)=(1,2) Blue:(a,b)=(1,0.5) Green:(a,b)=(1,3) Black:(a,b)=(1,1)

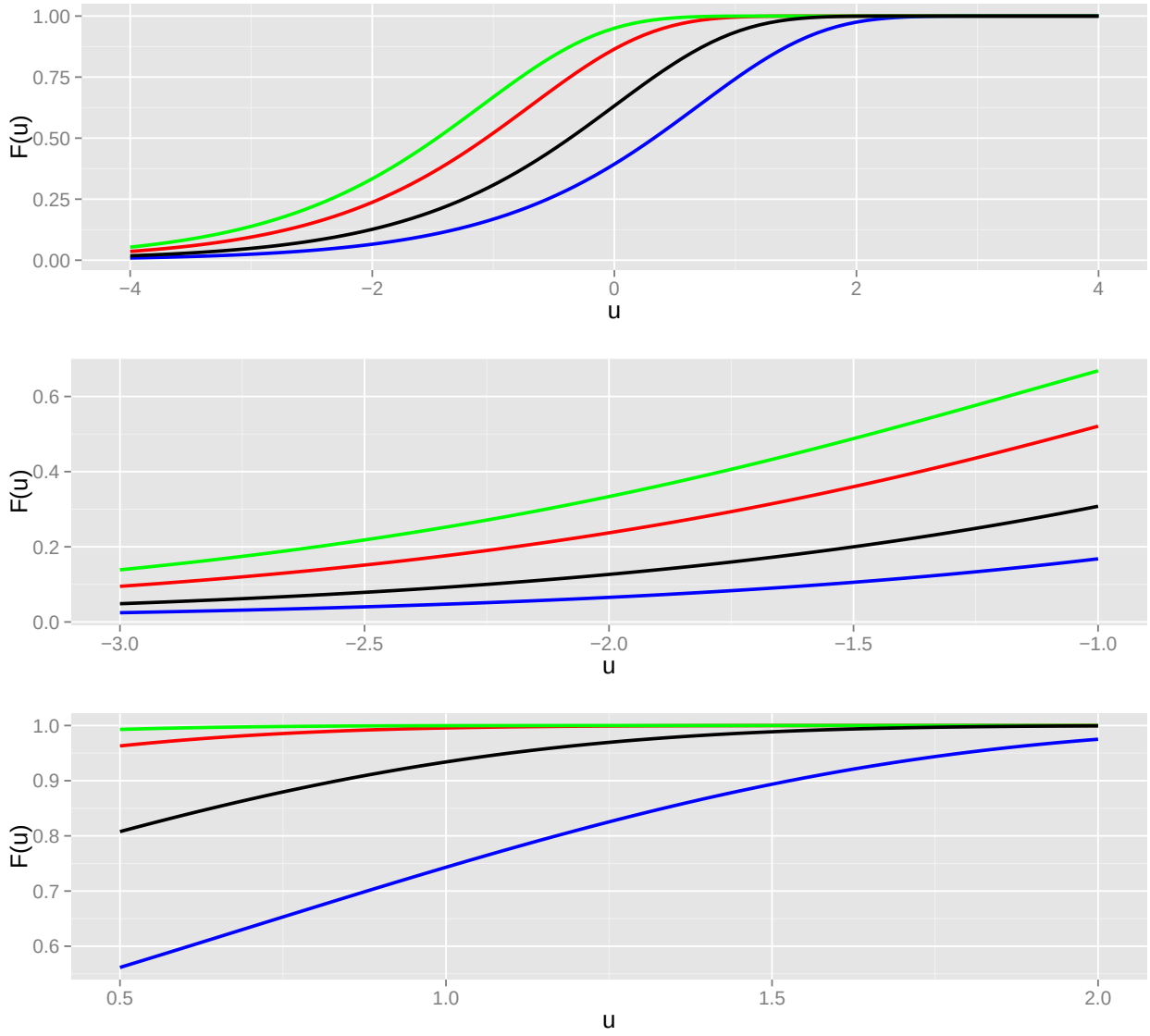

Figure 5: Plots for the Generalized CLogLog link function when $b = 1$ and a varied.

Red:(a,b)=(2,1) Blue:(a,b)=(0.5,1) Green:(a,b)=(3,1) Black:(a,b)=(1,1)

Listing 1: R Code for Link Function

```
## Creating the generalized link function. Specify in the R glm
## function by setting: family=binomial(link=vlog(a,b)))
## Can use function 'optim' to optimize over a and b parameters.

vlog <- function(a , b)
{
    ## link
    linkfun <- function(y) { log(-(1/a)*log(1 - y^(1/b))) }

    ## inverse link
    linkinv <- function(eta)
    {
        pmax(pmin(((1 - exp(-a*exp(eta)))^b), 1 - .Machine$double.eps), .Machine$double.eps)
    }

    ## derivative of invlink wrt eta
    mu.eta <- function(eta)
    {
        eta <- pmin(eta, 700)
        pmax((a*b*exp(eta)*exp(-a*exp(eta))*((1-exp(-a*exp(eta)))^(b-1))), .Machine$double.eps)
    }

    valideta <- function(eta) TRUE
    link <- "log(-(1/a)*log(1 - y^(1/b)))"
    structure(list(linkfun = linkfun, linkinv = linkinv,
                   mu.eta = mu.eta, valideta = valideta,
                   name = link), class = "link-glm")
}
```

REFERENCES

Agresti, A. (2014), *Categorical data analysis*, John Wiley & Sons.

Aranda-Ordaz, F. J. (1981), "On two families of transformations to additivity for binary response data," *Biometrika*, 68, 357–363.

Chen, M.-H., Dey, D. K., and Shao, Q.-M. (1999), "A new skewed link model for dichotomous quantal response data," *Journal of the American Statistical Association*, 94, 1172–1186.

Cordeiro, G. M., Nadarajah, S., and Ortega, E. M. (2012), "The Kumaraswamy Gumbel distribution," *Statistical Methods & Applications*, 21, 139–168.

Eljabri, S. (2013), "New Statistical Models for Extreme Values," .

Guerrero, V. M. and Johnson, R. A. (1982), "Use of the Box-Cox transformation with binary response models," *Biometrika*, 69, 309–314.

Gupta, R. D. and Gupta, R. C. (2008), "Analyzing skewed data by power normal model," *Test*, 17, 197–210.

Jiang, X., Dey, D. K., Prunier, R., Wilson, A. M., Holsinger, K. E., et al. (2013), "A new class of flexible link functions with application to species co-occurrence in cape floristic region," *The Annals of Applied Statistics*, 7, 2180–2204.

Kim, S., Chen, M.-H., and Dey, D. K. (2008), "Flexible generalized t-link models for binary response data," *Biometrika*, 95, 93–106.

Kumaraswamy, P. (1980), "A Generalized Probability Density Function for Double-Bounded Random Processes," *Journal of Hydrology*, 46, 79–88.

MILICER, H. and SZCZOTKA, F. (1966), "Age at menarche in Warsaw girls in 1965," *Human Biology*, 199–203.

Nagler, J. (1994), "Scobit: an alternative estimator to logit and probit," *American Journal of Political Science*, 230–255.

Nykamp DQ, C. J. and RA, A. (2014), "Developing a logistic model to describe bacteria growth." .

Stukel, T. A. (1988), "Generalized logistic models," *Journal of the American Statistical Association*, 83, 426–431.

Wang, X. and Dey, D. K. (2010), "Generalized extreme value regression for binary response data: an application to B2B electronic payments system adoption," *The Annals of Applied Statistics*, 2000–2023.