RESEARCH ARTICLE

A semiparametric complementary log-log (C-loglog) model with applications in binary rare event identification

Kai Peng[a], Cheng Peng[b],

[a]School of Science, Ningbo University of Technology, Ningbo, Zhejiang 315048, China.
[b]Department of Mathematics, West Chest University of Pennsylvania, West Chester, PA 19383, USA.

This article is under peer review with a statistics journal.

Contents

Address of correspondence: C. Peng. Email: cpeng@wcupa.edu

ABSTRACT
We consider fitting the complementary log-log (C-loglog) model to a stratified random sample using the empirical profile likelihood approach. The link function in the C-loglog model uses an asymmetric link function that allows capturing the asymmetry of the predicted scores and, hence, improves the performance of binary prediction. We developed an explicit algorithm for the semiparametric maximum likelihood estimation for the model parameters. The simulation results show that the C-loglog model is a valid binary regression model. Finally, we implement this model using real-world credit card fraud prediction.

KEYWORDS
Machine learning; Logit model; Complementary log-log; Imbalanced data; Rare event mining; Anomaly detection

1. Introduction

The binomial regression models are in a sub-family of Nelder and Wedderburn's [16] Generalized Linear Models (GLM).

Let Y be the Bernoulli random variable that takes value 0 and 1 and $X = (X_1, X_2, \cdots, X_p)$ be a vector of k independent predictor variables. Consider binary regression model

$$g(\pi(x)) = \beta_0 + x\beta \quad \text{or equivalently,} \quad \pi(x) = g^{-1}(\beta_0 + x\beta)$$

where $\pi(x) = P(Y = 1|x)$ and $g(\cdot)$ is the *link function*.

Different link functions define different binary regression models. When the inverse of the link function takes the cumulative distribution function of the standard logistic, standard normal, and standard Gumbel distributions respectively, we define three popular binary regression models: logistic, probit, and complementary log-log (C-loglog) models.

Piegorsch [18] studied the complementary log regression model with a log-link function and also developed goodness-of-link procedure for practical applications. This model is different from the C-loglog model.

The link functions of logistic and probit models are the cumulative distribution function (CDF) of the standard logistic and normal distributions, respectively. However, the link function of C-loglog model is the inverse of the CDF of the standard extreme value (Type-I) distribution. Unlike the logit and probit links which are symmetric, the c-loglog link is asymmetric.

Numerous studies have been conducted to compare the performance of the three models under prospective data. Logistic and probit models usually yield similar results. The readers are referred to Hosmer and Lemeshow [12], Hardin and Hilbe [11], Agresti [2], Cox and Snell [5], and Greene [10] for more details.

Some studies such as Long [15] indicate that the choice between the logistic and probit models is largely dependent on convenience and convention. Since logit and probit CDF are symmetric link functions, their CDF curves approach 0 at the same rate as they approach 1. However, the skewed C-loglog link curve approaches 1 faster than 0. Intuitively, it should be able to result in a fit better to the imbalanced data in which one category is infrequently observed (see Agostino et. al. [1]).

The C-loglog model is frequently used in applications in different fields based on simple random samples. To name a few, see the recent works of Calabrese [4], Fujisawa et al. [9], Kitali [14], Penman [17], and Shim et al. [21].

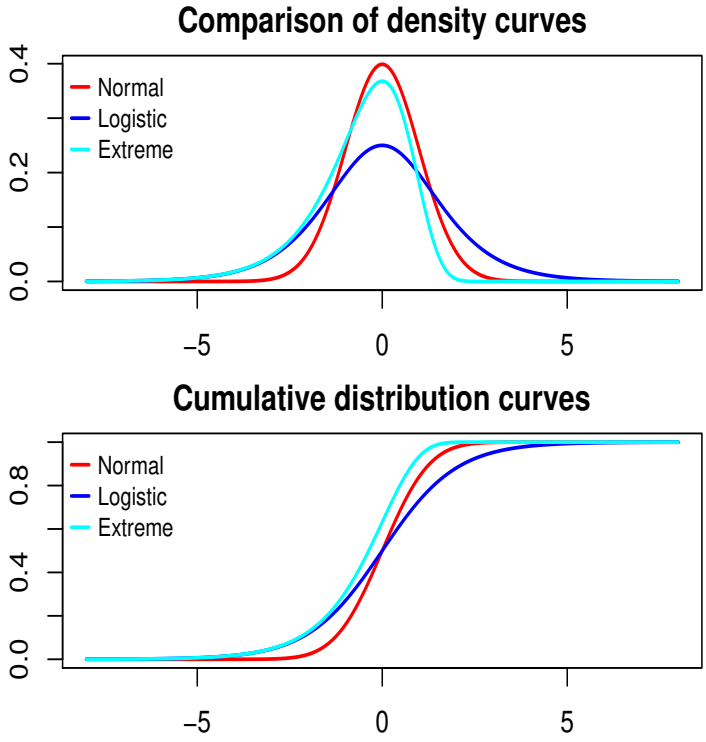

Figure 1. Three distributions are used in the definition of the link functions of logistic, probit, and C-loglog models.

Using the C-loglog model for rare event prediction is rarely discussed in the literature. Although some work was explored in this direction, the limited work was based on a simple random sample and from an applications perspective. See, for example, the recent work of Alves et al. [3]. However, C-loglog models based on stratified samples have not been studied yet.

The goal of this paper is to Qin's [19] framework of density ratio model to develop a C-loglog model based on stratified random samples via a semiparametric empirical likelihood theorem. The resulting model could incorporate the prior information in the estimation and a naturally suited for rare event prediction.

The rest of the paper is organized in the following. In Section 2, we briefly the formulation of the likelihood function of the general binary regression model based on prospective data. In Section 3, we formulate the semiparametric C-loglog regression model using Qin's empirical likelihood approach based on stratified case-control data. The semiparametric empirical likelihood estimation of the model parameters will be discussed in Section 4. The random number generation process of the two involved stratified population distributions is discussed in Section 5. In Sections 6 and 7, we present a case study and a small simulation study to assess the performance of the semiparametric maximum likelihood estimator. A real-world challenging problem - credit card fraud detection in Section 8. Some discussions and future work are presented in Section 9.

2. Likelihood Function of General Binary Regression Models

Let $\{y_i, x_i\}_{i=1}^n$ be an i.i.d. simple random sample taken from a population. y_i is from a Bernoulli population. The likelihood of a single observation is given by

$$L(\beta : y_i, x_i) = \left[g^{-1}(\beta_0 + x_i\beta)\right]^{y_i} \left[1 - g^{-1}(\beta_0 + x_i\beta)\right]^{1-y_i}$$

Let $G_0(x) = P(x|Y = 0)$ be probability distribution function of population corresponding $Y = 0$ and $G_1(x) = P(x|Y = 1)$ be probability distribution function of population corresponding $Y = 1$. Using the Bayes rule, we have

$$P(x|y = 0) = \frac{P(x)P(y = 0|x)}{P(y = 0)} \quad \text{and} \quad P(x|y = 1) = \frac{P(x)P(y = 1|x)}{P(y = 1)} \tag{1}$$

The ratio of the above two probabilities gives the following model

$$\frac{P(x|y = 1)}{P(x|y = 0)} = \frac{1 - \pi}{\pi} \frac{P(y = 1|x)}{P(y = 0|x)} \tag{2}$$

where

$$\pi = P(y = 1) = 1 - P(y = 0)$$

and

$$\text{odds}(x) = \frac{P(y = 1|x)}{P(y = 0|x)} = \frac{P(y = 1|x)}{1 - P(y = 1|x)} \tag{3}$$

is the odds of $Y = 1$ for given x. Under the logit link function

$$Q[P(y = 1|x)] = \log[\frac{P(y = 1|x)}{1 - P(y = 0|x)}] = \alpha + x\beta^T \tag{4}$$

or equivalently,

$$P(y = 1|x) = \frac{\exp(\alpha + x\beta^T)}{1 + \exp(\alpha + x\beta^T)} \tag{5}$$

Let $g_0(x)$ be the density function of $G_0(x)$ and $g_1(x)$ be the density function of $G_1(x)$. The semiparametric model is re-expressed by

$$g_1(x) = \exp(\alpha^* + x\beta^T)g_0(x) \tag{6}$$

where $\alpha^* = \alpha + \log[(1-\pi)/\pi]$ is a scale parameter and β is a $p \times 1$ vector of regression parameters.

Note that the logit link function (2) is a symmetric function with respect to (0,0.5). It is restrictive. So is the probit link function.

The following figure depicts the features of the three commonly used link functions in the GLM.

The C-log-log link function relaxes the restriction of symmetry, we will use the C-log-log model based on the retrospective study design in this project.

The rest of the paper is organized as follows. In section 2, we define the new density ratio model and give the likelihood function based on the empirical likelihood method. In section 3, we introduce the empirical likelihood estimators of the regression coefficients based on the Lagrange multiplier method. A simulation experiment to verify the correctness of the parameter estimation program is given in section 4. In the last section, we use real-world data to illustrate the implementation of the model and compare the results based on the regular logistic regression model.

3. Model Specification

Instead of using the logit link function, we use the complementary Log-log link and obtain

$$Q[P(y=1|x)] = \log[-\log(1-p(y=1|x))] = \alpha + \beta^T x \tag{7}$$

or equivalently,

$$P(y=1|x) = 1 - \exp[-\exp(\alpha + \beta^T x)] \equiv \psi(x)$$

and

$$P(y=0|x) = \exp[-\exp(\alpha + \beta^T x)] \equiv 1 - \psi(x)$$

where α) is a scale parameter and β is a $p \times 1$ vector of regression coefficients.

Let $P(x|y=0)$ be the probability distribution for the population labeled with $Y=0$ and $P(x|y=1)$ be the probability distribution for the population labeled with $Y=1$. The two corresponding density functions are given respectively by $h(x)$ and $g(x)$. Using the Bayes rule, we have the following retrospective C-loglog model

$$\frac{h(x)}{g(x)} = \frac{\theta\psi(x)}{1-\psi(x)} = \theta\{\exp[\exp(\alpha + \beta^T x)] - 1\} \equiv \theta w(x) \tag{8}$$

where $w(x) = \exp\left[\exp(\alpha + \beta^T x)\right] - 1$ is dependent on the unknown parameters α and β. $\pi = P(Y=1)$ is the population proportion of observing $Y=1$ and $\theta = (1-\pi)/\pi$ is the odds of of observing $Y=0$. For the retrospective case-control study, the two stratified samples in (9) do not have the information about θ. In other words, θ is inestimable and will be. In practical application, we need to use the prior information of θ from the population in an algorithm for estimating α and β to be developed later. Hereafter, we will consider θ as a scalar.

Next, we establish the semiparametric estimation of parameters in the model by assuming that the two independent samples are taken from $Y = 1$ and $Y = 0$ respectively as follows

$$\begin{cases} x_1, x_2, \cdots, x_{n_0} \to g(x) \\ z_1, z_2, \cdots, z_{n_1} \to h(x) = \theta w(x) g(x) \end{cases} \tag{9}$$

Let $G(x)$ and $H(x)$ be the cumulative distribution associated with density functions $g(x)$ and $h(x)$. The semiparametric empirical likelihood function under model (8) with constraints is given by

$$\mathcal{L}(\alpha, \beta, G) = \prod_{i=1}^{n_0} dG(x_i) \prod_{j=1}^{n_1} w(z_j) dG(z_j) = \prod_{i=1}^{n} p_i \prod_{j=1}^{n_1} \theta w(z_j) \tag{10}$$

Subject to, for $i = 1, 2, \cdots, n$

$$0 \le p_i \le 1, \quad \sum_{i=1}^{n} p_i - 1 = 0, \quad \sum_{i=1}^{n} p_i [\theta w(t_i) - 1] = 0 \tag{11}$$

where $p_i = dG(t_i) \quad (i = 1, 2, \cdots, n)$ are non-negative jumps with total mass unity. The optimization of the problem of the above likelihood function with constraints is equivalent to the following constrained optimization problem.

$$l(\alpha, \beta, G) = \sum_{i=1}^{n} \log(p_i) + \sum_{j=1}^{n_1} \log[\theta w(z_j)] \tag{12}$$

where

$$0 \le p_i \le 1, \quad \sum_{i=1}^{n} p_i - 1 = 0, \quad \sum_{i=1}^{n} p_i [\theta w(t_i) - 1] = 0. \tag{13}$$

4. Parameter Estimation

In this section, we estimate the parameters of the proposed C-log-log model by solving the constrained optimization problem with the following objective function (14).

$$F(\alpha, \beta, p_i, \gamma, \lambda) = \sum_{i=1}^{n} \log(p_i) + \sum_{j=1}^{n_1} \log[\theta w(z_j)] - \gamma\{\sum_{i=1}^{n} p_i - 1\} - \lambda\{\sum_{i=1}^{n} p_i [\theta w(t_i) - 1]\}. \tag{14}$$

From $\partial F / \partial p_i = 0$, we have

$$\begin{cases} \gamma = n \\ p_i = \frac{1}{n+\lambda[\theta w(t_i)-1]}. \end{cases} \tag{15}$$

To use the information of the third constraint in (13), we need to solve for λ from the following equation

$$1 = \sum_{i=1}^{n} p_i = \sum_{k=1}^{n} \frac{1}{n + \lambda\left[\theta w(x) - 1\right]} \tag{16}$$

where $w(x) = \exp(\exp(\alpha+\beta x))-1$. We now re-write the semiparametric log-likelihood function (12) in the following

$$l(\alpha, \beta, \lambda) = \sum_{j=1}^{n_1} \log[\theta w(z_j)] - \sum_{i=1}^{n} \log\left\{n + \lambda\left[\theta w(t_i) - 1\right]\right\} \tag{17}$$

The semiparametric maximum likelihood estimator denoted by $(\widetilde{\alpha}, \widetilde{\beta})$ is the solution to the following maximization problem.

$$\underset{\alpha,\beta,\lambda}{\operatorname{argmax}}\, l(\alpha, \beta, \lambda) \tag{18}$$

Before deriving score equations, we introduce the following notations.

$$\begin{aligned} u(x) &= \exp[\exp(\alpha + \beta^T x)] \\ v(x) &= \exp(\alpha + \beta^T x) \\ r(x) &= u(x)v(x) \end{aligned} \tag{19}$$

Note that $w(x) = u(x) - 1$. With the above notations, we have the following results regarding the partial derivatives about α and β. As an example, we use u'_α to denote the partial derivative of $u(x) = u(\alpha.\beta, : x)$ with respective to α. That is, $u'_\alpha(x) = \partial u(x)/\partial\alpha$

$$\begin{aligned} u'_\alpha(x) &= r(x) \\ u'_\beta(x) &= xr(x) \\ r'_\alpha(x) &= r(x)\left[v(x) + 1\right] \\ r'_\beta(x) &= xr(x)\left[v(x) + 1\right] \end{aligned} \tag{20}$$

The kernel of the log-likelihood function can be re-expressed explicitly in α, β and θ in the following

$$l_k(\alpha, \beta, \lambda) = \sum_{j=1}^{n_1} \log\left[w(z_j)\right]\} - \sum_{i=1}^{n} \log\left\{n + \lambda\left[\theta w(t_i) - 1\right]\right\}, \tag{21}$$

α and β in $w(\cdot)$ are unknown parameters.

Taking the partial derivative with respect to α and β yields the following system of score equations.

$$f_1 = \frac{\partial l(\alpha, \beta, \lambda)}{\partial \alpha} = \sum_{j=1}^{n_1} \frac{r(z_j)}{w(z_j)} - \sum_{i=1}^{n} \frac{\lambda \theta r(t_i)}{n + \lambda \left[\theta w(t_i) - 1\right]} = 0 \tag{22a}$$

$$f_2 = \frac{\partial l(\alpha, \beta, \lambda)}{\partial \beta} = \sum_{j=1}^{n_1} \frac{z_j r(z_j)}{w(z_j)} - \sum_{i=1}^{n} \frac{\lambda \theta t_i r(t_i)}{n + \lambda \left[\theta w(t_i) - 1\right]} = 0 \tag{22b}$$

$$f_3 = \frac{\partial l(\alpha, \beta, \lambda)}{\partial \lambda} = -\sum_{i=1}^{n} \frac{\theta w(t_i) - 1}{n + \lambda \left[\theta w(t_i) - 1\right]} = 0 \tag{22c}$$

The solution to the above system of nonlinear equation, denoted by $(\widetilde{\alpha}, \widetilde{\beta})$, is the semiparametric estimate of (α, β). To solve the above nonlinear system, we need the following Jacobian matrix

$$J(\alpha, \beta) = \begin{pmatrix} \partial f_1/\partial \alpha & \partial f_1/\partial \beta \\ \partial f_2/\partial \alpha & \partial f_2/\partial \beta \end{pmatrix}$$

where the cell elements are explicitly given in the following with some new notations

$$\frac{\partial f_1}{\partial \alpha} = \frac{\partial^2 l(\alpha, \beta)}{\partial \alpha^2} = \sum_{j=1}^{n_1} \frac{k(z_j)}{w^2(z_j)} - \sum_{i=1}^{n} \frac{k(t_i) + m(t_i)}{\left[\Omega + w(t_i)\right]^2} \tag{23a}$$

$$\frac{\partial f_1}{\partial \beta^T} = \frac{\partial^2 l(\alpha, \beta)}{\partial \alpha \partial \beta^T} = \sum_{j=1}^{n_1} \frac{z_j k(z_j)}{w^2(z_j)} - \sum_{i=1}^{n} \frac{t_i \left[k(t_i) + m(t_i)\right]}{\left[\Omega + w(t_i)\right]^2} \tag{23b}$$

$$\frac{\partial f_2}{\partial \beta^T} = \frac{\partial^2 l(\alpha, \beta)}{\partial \beta \partial \beta^T} = \sum_{j=1}^{n_1} \frac{z_j^2 k(z_j)}{w^2(z_j)} - \sum_{i=1}^{n} \frac{t_i^2 \left[k(t_i) + m(t_i)\right]}{\left[\Omega + w(t_i)\right]^2} \tag{23c}$$

We denote $f = (f_1, f_2)$ to be the vector of the three functions specified in (17) and $v = (\alpha, \beta)$ to be the vector of the true value of parameters. We use the Newton method to find the MLE of the regression coefficients (α, β) with given prior information of θ_0

Newton Algorithm:

(1) Choose a vector of initial values $v_0 = (\alpha_0, \beta_0)$
(2) Solve for λ_0 from

$$1 = \sum_{k=1}^{n} \frac{1}{n + \lambda_0 \left\{\theta[\exp(\exp(\alpha_0 + \beta_0 t_i))] - 1)\right\}} \tag{24}$$

(3) Evaluate the Jacobian matrix

$$J(\alpha_0, \beta_0, \lambda_0) = \begin{pmatrix} \frac{\partial f_1}{\partial \alpha} & \frac{\partial f_1}{\partial \beta} \\ \frac{\partial f_2}{\partial \alpha} & \frac{\partial f_2}{\partial \beta} \end{pmatrix} \Bigg|_{v_0=(\alpha_0,\beta_0,\lambda_0)} \tag{25}$$

(4) Assume that $J(\alpha_0, \beta_0, \lambda_0)$ is invertible. Let $J_{v_0}^{-1}$ be the inverse of $J(\alpha_0, \beta_0, \lambda_0)$. The next updated vector of the parameter is given by

$$v_1 = v_0 - J_{v_0}^{-1} f(\alpha_0, \beta_0, \lambda_0)$$

(5) Let $v_i = (\alpha_i, \beta_i)$ be the i-th updated vector. and λ_i be the solution to the following equation

$$1 = \sum_{k=1}^{n} \frac{1}{n + \lambda_i \{\theta[\exp(\exp(\alpha_i + \beta_i t_i))] - 1)\}}. \tag{26}$$

The (i+1)-th updated vector is

$$v_{i+1} = v_i - J_{v_i}^{-1} f(\alpha_i, \beta_i, \lambda_i)$$

where $J_{v_i}^{-1}$ is the inverse of

$$J(\alpha_i, \beta_i, \lambda_i) = \begin{pmatrix} \frac{\partial f_1}{\partial \alpha} & \frac{\partial f_1}{\partial \beta} \\ \frac{\partial f_2}{\partial \alpha} & \frac{\partial f_2}{\partial \beta} \end{pmatrix} \Bigg|_{v_0=(\alpha_i,\beta_i,\lambda_i)} \tag{27}$$

(6) Repeat step 5 until $|v_{n+1} - v_n| < \varepsilon_0$. ε_0 is a pre-selected small error bound.

Remark: Since there is a nested iterative process to find λ within each iteration of the Newton method, some optimization routines in statistical programs cannot be used directly to find the MLE of the regression coefficients.

We will use the above algorithm in the simulation study and numerical examples.

5. Random Number Generation

In this section, we introduce steps for generating random, numbers from the derived density function $g(x)$ for any given baseline distribution $h(x)$ specified in (9).

We choose the baseline population to be a uniform population: $G(x) \sim Unif(a, b)$. Then for given a, b, β and θ, the density function of $H(x)$ based on the C-log-log model is given by

$$h(x) = \frac{\theta\{\exp[\exp(\alpha + \beta x)] - 1\}}{b - a}.$$

We need to find the value of α such that $\int_a^b h(x)\, dx = 1$. This means

$$\frac{(b-a)(1+\theta)}{\theta} = \int_a^b \exp[\exp(\alpha+\beta x)]\ dx \tag{28}$$

Because a, b, β, and θ are given, we can use a numerical method to approximate the value of the integration in the above denominator to find the value of α.

Since there is no existing program that can be used to generate random numbers with density $h(x)$, we use the following rejection-acceptance algorithm to generate random samples from $h(x)$.

For convenience, we summarize the well-known accept-rejection method of the random variable generation before introducing the steps for generating the derived distribution with a specific baseline distribution. The detail and proof of this algorithm can be found in Chapter 11 of the classic textbook of Ross [20]

Acceptance-Rejection algorithm for generating random numbers from $h(x)$.

(1) Choose a density function $k(x)$ that has an efficient algorithm for generating random numbers from it and $h(x)/k(x)$ bounded by some positive constant M that is close to 1.
(2) Generate a random number x from a distribution that has a density $k(x)$.
(3) Generate a [0,1)-uniform random number u.

(a). If $u > \frac{h(x)}{Mk(x)}$, reject the sample and return to Step 2.

(b). Otherwise, accepting x as a random number from $h(x)$.

6. Case Study: Uniform Baseline

In our simulation, we choose $a = 1, b = 2, \theta = 0.5, \beta = -1$ and solve α from (28). To be more specific,

$$3 = \int_1^2 \exp[\exp(\alpha - x)]\ dx$$

Solving the normalizing parameter from (19), we have $\alpha \approx 1.509441$. The density function $h(x)$ is explicitly given by

$$h(x) = 0.5\{\exp[\exp(1.509441 - x)] - 1\}$$

Before we find the reference distribution for generating random numbers from $h(x)$, we find the mean and variance of $h(x)$ and use it to check with the simulated sample mean and variance.

$$E_h(X) = \int_1^2 0.5x\{\exp[\exp(1.509441 - x)] - 1\}dx \approx 1.368528 \tag{29}$$

and

$$V_h(X) = \int_1^2 0.5(x-E)^2\{\exp[\exp(1.509441-x)]-1\}dx \approx 0.07493256 \qquad (30)$$

Finding the reference random variable in the above *acceptance-rejection* algorithm is not straightforward, we use the following density plot to assist in the selection of the reference distribution.

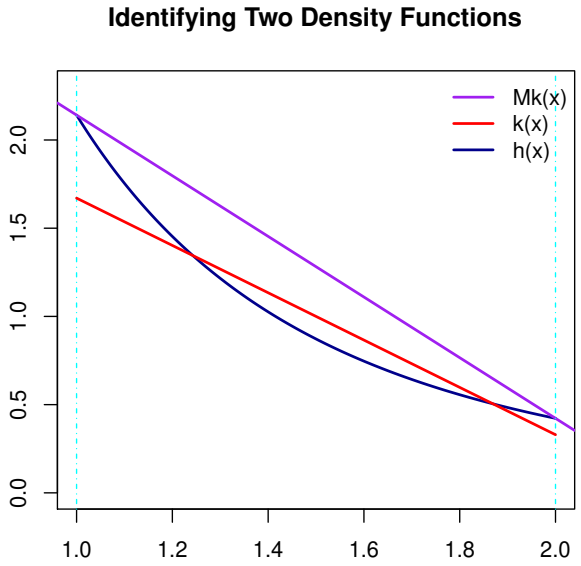

Figure 2. *Identifying the reference density to simulate the random numbers from the derived density* $h(x)$

We first find $Mk(x)$ such that

$$\frac{h(x)}{Mk(x)} \leq 1$$

for some M density function $k(x)$ that has an algorithm to generate random numbers.

From the density curse of $h(x)$, we select a straight line that is above the density curve as $Mk(x): \ y = 3.859978 - 1.718829x$. Note that

$$\int_1^2 (3.859978 - 1.718829x)\ dx = 1.281735$$

This means that $M = 1.281735$ and $k(x) = 3.011526 - 1.341018x$. The corresponding CDF is given by

$$K(t) = \int_1^t (3.011526 - 1.341018x)\ dx = -0.670509t^2 + 3.011526t - 2.341017$$

where $1 \leq t \leq 2$. To generate random numbers from $k(x)$, we need to find the inverse of the CDF given by

$$K^{-1}(U) = \frac{3.011526 - \sqrt{2.790597 - 2.682036U}}{1.341018} \tag{31}$$

Remark: we choose "-" in the numerator since the mean of the reference distribution. $k(x)$, is approximately equal to 1.388247.

Algorithm: Steps for generating random numbers from $h(x)$:

(1) Choose reference density $k(x) = 3.256821 - 1.450246x$ and $M = 1.185198$ Then

$$\frac{h(x)}{Mk(x)} \leq 1$$

(2) Generate a uniform random number u from Unif$(0, 1)$. Let

$$x = \frac{3.011526 - \sqrt{2.790597 - 2.682036U}}{1.341018} \tag{32}$$

(3) Evaluate the following ratio

$$R = \frac{0.5\{\exp[\exp(1.509441 - x)] - 1\}}{3.859978 - 1.718829x} \tag{33}$$

(4) Generate another random number u from Unif$(0, 1)$.
   (a). If $u > R$, rejecting x in (32) as a random number of $h(x)$;
   (b). If $u < R$, accepting x in (32) as a random number of $h(x)$.

(5) Repeat steps 2 and 3 until reaching the desired sample size.

Here we use uniform distribution as the baseline. One can use other baselines to derive the density of $h(x)$ under the C-loglog model. We perform a small simulation study to assess the performance of the semiparametric estimator of the regression coefficients under the C-loglog model.

7. Simulation

In this section, we conduct a numerical simulation experiment to assess the finite sample performance of the semiparametric estimators of the regression coefficients.

We take n_0 and n_1 respectively from $(30, 60, 90, 120, 150)$. Two independent random samples from $U(1, 2)$ and $h(x)$ based on the procedures introduced above with various sample sizes. From the above section, the true values of the two parameters are $\alpha = 1.51$ and $\beta = -1$.

Using the algorithms introduced in Section 3 to find the semiparametric empirical likelihood estimates. For each scenario of the combination of (n_0, n_1), we simulated 1000 samples. The following tables report the performance of the parameter estimation with the mean square error (MSE) and the coverage probability.

Table 1. MSE of semiparametric MLE of parameters

Regression		n_2				
Coefficients	n_1	30	60	90	120	150 height
30	0.472	0.352	0.319	0.281	0.267	
	60	0.341	0.223	0.190	0.176	0.161
α	90	0.292	0.177	0.147	0.123	0.114
	120	0.295	0.165	0.130	0.117	0.092
	150	0.277	0.154	0.124	0.099	0.094
	30	0.113	0.085	0.078	0.069	0.065
	60	0.081	0.053	0.046	0.043	0.039
β	90	0.069	0.042	0.035	0.030	0.028
	120	0.070	0.039	0.031	0.028	0.022
	150	0.065	0.037	0.030	0.024	0.023

We can see from table 1, that the MSE decreases as one or both sample sizes increase. Use the same configuration of the sample sizes to calculate the coverage probabilities of both estimated parameters and summarize the results in the following table. The nominal confidence level used in the simulation is 95%.

Table 2. The coverage probabilities of the semi-parametric MLE of the parameters

Regression		n_2				
Coefficients	N	30	60	90	120	150 height
30	0.940	0.936	0.936	0.942	0.942	
	60	0.942	0.926	0.938	0.938	0.944
α	90	0.934	0.926	0.938	0.938	0.944
	120	0.930	0.924	0.928	0.936	0.940
	150	0.936	0.944	0.934	0.936	0.938
	30	0.938	0.936	0.936	0.944	0.942
	60	0.942	0.926	0.938	0.940	0.944
β	90	0.934	0.928	0.938	0.938	0.944
	120	0.932	0.922	0.926	0.934	0.940
	150	0.936	0.944	0.934	0.934	0.938

As expected, the coverage probabilities (table 2) increase as sample sizes increase although the magnitude is not significant. One pattern worth mentioning is that all of these coverage probabilities are less than the nominal level of 95%. This could be because the uniform baseline is too special. The other possible reason could be the rounding error in finding the value of α (β is exact in the simulation). Since this is not the primary interest of C-loglog regression, we will not dive deep in this direction. However, the simulation results confirmed the validity of the semiparametric estimation. We will present a numerical example in the next section.

8. Real-world Application - Credit Card Fraud Detection

We will devote an entire section to a real-world application of the C-loglog regression for credit card fraud detection. The data used in the section was taken from a financial company. We need to process the data by extracting fraud information from the raw data before fitting the C-loglog model to the analytic data. Due to the complexity of the information extraction process, we will illustrate this process in several subsections.

8.1. Raw Data Processing

The initial raw data had 25575 genuine cards and 7762 compromised cards. For each selected card, 41 historical transactions were extracted from the database. The only variable used in this study is the transaction dollar amount. The most recent transactions associated with 7662 compromised cards are fraudulent transactions the rest of the 40 historical dollar amounts are genuine. For all 25575 genuine cards, all transactions are not fraudulent. We define the binary response variable under the name *fraud status* based on the most recent transaction. The processed data will stored in a 33337×42 rectangular matrix with the last column being fraud status. In the following section, we use the method of [22] to define the fraud index based on the first 41 columns in the data matrix.

8.2. Fraud Index

The fraud index is defined based on the process capability index that is customarily used to assess the capability of a process. We consider transaction dollar amounts of a credit card to be the outcomes of the customer's "spending process". We use a data-driven approach to determine the upper specification limit (USL) and the lower specification limits (LSL), the mean ($\bar{X}$), and the standard deviation (s) in order to define the index. To be more specific, we use the earliest 36 transaction dollar amounts to determine LSL, USL, $\bar{X}$, and s. The most 5 recent transaction dollar amounts to define the fraud index. This index has been developed by the second author and implemented in a production system for real-time fraud detection. The details of how to extract this type of fraud index and how to better extract the index will be discussed elsewhere. Here focus on building the framework. There should be a more accurate definition of fraud index.

The fraud index defined based on the 41 historical transaction dollar amounts and the corresponding fraud status will form the final simple analytic data set with only two columns. We sort the data set using the fraud indices and the corresponding fraud status to draw the cumulative probability - empirical *S-Curve*.

The empirical S-curve in the above Figure 3, is not not symmetric. This implies that the C-loglog model is more appropriate than the regular binary logit and probit models.

8.3. Fitting C-loglog Model

We fit the C-loglog model to the data set using the procedure in Section 4. For ease of demonstration, we choose the prior parameter $\theta = 0$ (this is also called the tuning parameter whose value can be tuned to gain the optimal predictive power). The SMLE (semiparametric maximum likelihood estimation) of the regression coefficient and their

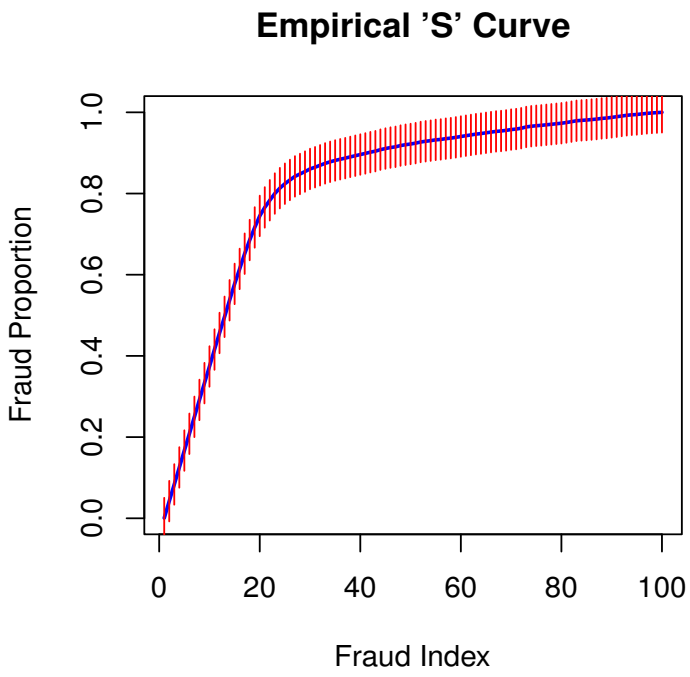

Figure 3. Empirical distribution of cumulative fraud probability based on fraud indices.

standard errors are summarized in the following table.

Parameter	SMLE	Standard Error	Z	P-value
$\hat{\alpha}$	-4.40021e-02	2.272079e-04	-193	0
$\hat{\beta}$	6.07741e-02	1.008228e-7	607700	0

Therefore, the fraud probability function is given explicitly by

$$P(Y = \text{fraud}|\text{fraud.index}) = \exp[\exp(-0.044 + 0.067 \times \text{fraud.index}) - 1]$$

The fraud index itself has some capability to separate fraudulent transactions from genuine transactions. The following figure shows the C-loglog model-based fraud probability significantly high separability of the fraud.

8.4. ROC Analysis

The performance of a classification model at all classification thresholds is assessed using the ROC curve. This curve plots two parameters: true positive rate (i.e., sensitivity, TPR) and false positive rate (i.e., specificity, FPR) of the model. A ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows the ROC curve of the C-loglog model based on the fraud data.

ROC curve is a descriptive and visual tool for assessing the global performance of all binary classification models. The area under the curve (AUC) is a numeric measure

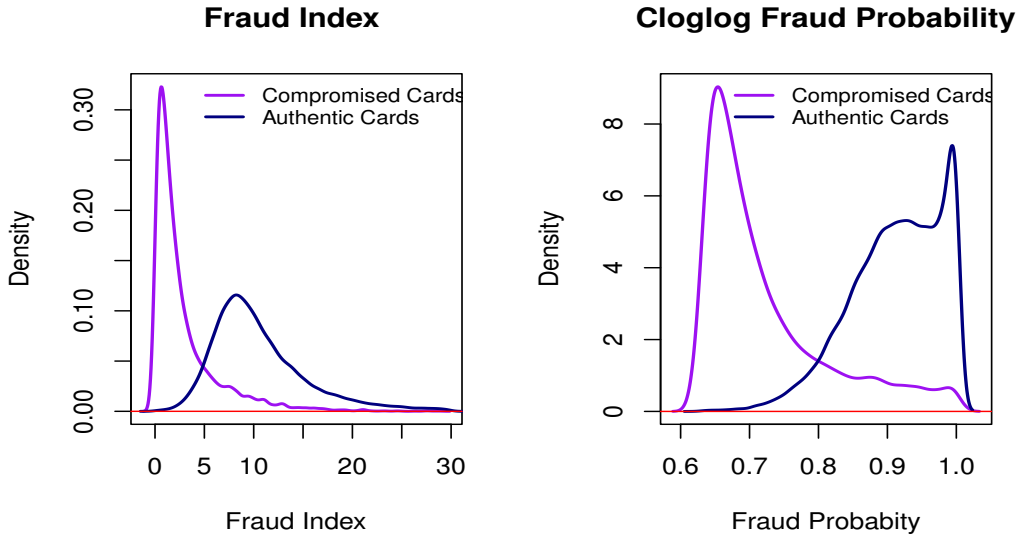

Figure 4. The separability of fraud index and model-based fraud probability.

of global performance. They are in-variant in scale and classification boundary. ROC and AUC are used for relative comparisons between different models and algorithms.

We report both ROC and AUC only for the purpose of illustration. When comparing two models and algorithms, the larger the AUC, the better the model (algorithm).

9. Discussions and Future Work

This paper developed a semiparametric C-loglog regression model using empirical likelihood estimation. Unlike the binary logit models that are widely used in practice, C-loglog model was occasionally used in some disciplines for field-specific applications. The semi-parametric C-loglog regression model developed in this paper is the first kind of model that applies to stratified samples. The event odds (θ) of the population are considered as prior information. We either provide it beforehand or use it as a hyperparameter and tune it for better predictive performance.

First of all, both parameters are solely based on the stratified samples. However, the information contained in the stratified samples is dependent on the sampling ratio (the ratio of the two sample sizes). The sampling ratio is, in general, determined by the practitioners. Therefore, the prior information should be used in order to find the appropriate estimates of α and β for prediction and association between response and predictors. This is also very different from the semiparametric logistic regression model.

Because the semiparametric C-loglog model allows stratified random sample and adjusting the sampling bias in SMEL by incorporating the prior information of the event odds (θ) of the population. This gives practitioners an approach to rare event learning (also called imbalanced learning) that is challenging in both the academic research community and industry applications as well. This method is different from Gary and Zeng's [13] approach of adjusting the intercept of logit models when fit to stratified random samples.

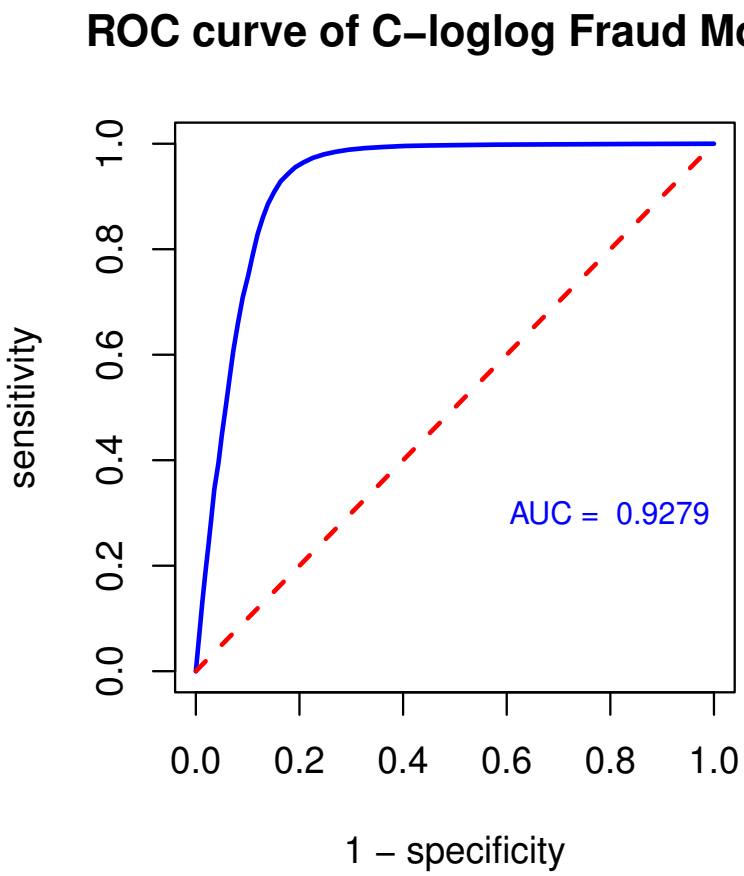

Figure 5. ROC analysis for the predictive power of the semiparametric C-loglog model.

In this semiparametric C-loglog model, the prior information (θ) is implicitly incorporated in the SMLE of the regression coefficients, The resulting SMLE of the estimated probability function can be used directly for prediction including rare event prediction.

This work is essentially an algorithm for both association and prediction analysis. Simulation results validate the algorithm. The asymptotic results of the model (structural) parameters and formal goodness-of-link test will be discussed elsewhere. More comprehensive simulations of the small sample performance will also be discussed as well.

References

[1] Agostino, M., Errico, L., Rondinella, S., & Trivieri, F. (2023). Enduring lending relationships and european firms default. *Research in Economics*, **77(4)**, 459-477.

[2] Agresti, A. (2007), *Introduction to Categorical Data Analysis*, John Wiley & Sons, 2nd edition.

[3] Alves, J. S., Bazán, J. L., and Arellano-Valle, R. B. (2023). Flexible cloglog links for

binomial regression models as an alternative for imbalanced medical data. Biometrical Journal, 65(3), 2100325.

[4] Calabrese, R., and Osmetti, S. A. (2011). Generalized extreme value regression for binary rare events data: an application to credit defaults. Bulletin of the International Statistical Institute LXII, 58th Session of the International Statistical Institute, 5631-5634.

[5] Cox, D. R. and Snell, E. J. (1989), *Analysis of Binary Data*. Chapman & Hall, NY, 2nd edition.

[6] Finney, D. J. (1971), *Probit Analysis*, Cambridge University Press, Cambridge, UK, 3rd edition.

[7] Greenberg, B. G. (1980), Chester I. Bliss, 1899-1979. *International Statistical Review*, **8(1)**, 135–136.

[8] Cramer, J. S. (1971), *Logit Models from Economics and Other Fields, Chapter 9: Origin and Development of the Probit and Logit Models*. Cambridge University Press, Cambridge, UK.

[9] Fujisawa, K., Mitomi, K., and Tahata, K. (2021). Extension of Marginal Complementary Log–Log Model and Separations of Marginal Homogeneity for Ordinal Categorical Data. *Journal of Statistical Theory and Practice*, **15(3)**, 62.

[10] Greene, W. H. (2018), *Econometric Analysis*. Pearson, New York, NY. 8th edition.

[11] Hardin, J. W. and Hilbe, J. M. (2007), *Generalized Linear Models and Extensions*. Stat Press Publication, College Station, TX, 2nd edition.

[12] Hosmer, D. W. and Lemeshow, S. (2000) *Applied Logistic Regression*. John Wiley & Sons, NY, 2nd edition.

[13] King, G. and Zeng, L. (2001). *Logistic Regression in Rare Events Data." Political Analysis*, **9**, 137–163.

[14] Kitali, A. E., Alluri, P., Sando, T., Haule, H., Kidando, E., and Lentz, R. (2018). Likelihood estimation of secondary crashes using Bayesian complementary log-log model. Accident Analysis & Prevention, 119, 58-67.

[15] Long, S. J. (1997), *Regression Models for Categorical and Limited Dependent Variables* (Advanced Quantitative Techniques in the Social Sciences). Sage Publications, Thousand Oaks, CA.

[16] Nelder, J. and Wedderburn, R. (1972), Generalized Linear Models, *Journal of the Royal Statistical Society. Series A (General)*. **135 (3)**, 370–384.

[17] Penman, A. D., and Johnson, W. D. (2009). Complementary log–log regression for the estimation of covariate-adjusted prevalence ratios in the analysis of data from cross-sectional studies. Biometrical Journal: Journal of Mathematical Methods in Biosciences, 51(3), 433-442.

[18] Piegorsch, W. W. (1992), Chester I. Bliss, 1899-1979. *International Statistical Review*, **8(1)**, 135–136.

[19] Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, **85(3)**, 619-630.

[20] Ross, S. M. (2019), *Introduction to Probability Models, 12th Edition*, Academic Press, San Diego, CA.

[21] Shim, H., Bonifay, W., and Wiedermann, W. (2023). Parsimonious asymmetric item response theory modeling with the complementary log-log link. Behavior Research Methods, 55(1), 200-219.

[22] Xu, J. and Peng, C. (2019) Parametric bootstrap process capability index control charts for both mean and dispersion, *Communications in Statistics - Simulation and Computation,* **48(10)**, *2936-2954, DOI: 10.1080/03610918.2018.1471505.*