2013

On the predictive analytics of the probit and logit link functions

Necla Gunduz

Ernest Fokoue

On the Predictive Analytics of the Probit and Logit Link Functions

Necla Gündüz
Department of Statistics
Gazi Üniversitesi, Fen Fakültesi, Istatistik Bölümü
06500 Teknikokullar, Ankara, Turkey
ngunduz@gazi.edu.tr

Ernest Fokoué*
Center for Quality and Applied Statistics
Rochester Institute of Technology
98 Lomb Memorial Drive, Rochester, NY 14623, USA
ernest.fokoue@rit.edu

Abstract

Researchers and practitioners have long held that the probit and logit link functions generally yield the same performance in binary classification. Despite this widespread recognition of the strong similarities between these two link functions, very few (if any) researchers have dedicated time to carry out a formal study aimed at establishing and characterizing firmly all the aspects of the similarities and differences. This paper proposes a definition of both structural and predictive equivalence link functions-based binary regression models, and provides both a theoretical and computational justification of the long held claim that probit and logit are indeed very similar. From a predictive analytics perspective, it turns out that not only are probit and logit perfectly predictively equivalent, but the other link functions like cauchit and complementary log log enjoy very high percentage of predictive equivalence. Throughout this paper, simulated and real life examples demonstrate clearly all the equivalence results that we prove theoretically.

I. Introduction

Given $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)$, where $\mathbf{x}_i^\top \equiv (x_{i1}, x_{i2}, \cdots, x_{ip})$ denotes the p-dimensional vector of characteristics and $y_i \in \{0, 1\}$ denotes the binary response variable, binary regression seeks to model the relationship between $\mathbf{x}$ and y using

$$\pi(\mathbf{x}_i) = \Pr[Y_i = 1 | \mathbf{x}_i] = F(\eta(\mathbf{x}_i)) \tag{1}$$

where

$$\eta(\mathbf{x}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = \tilde{\mathbf{x}}_i^\top \boldsymbol{\beta} \qquad i = 1, \cdots, n \tag{2}$$

for a $(p+1)$-dimensional vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \cdots, \beta_p)^\top$ of regression coefficients and $F(\cdot)$ is the cdf corresponding to the link functions under consideration. Specifically, the cdf $F(\cdot)$ is the inverse of the link function $g(\cdot)$, such that $\eta(\mathbf{x}_i) = F^{-1}(\pi(\mathbf{x}_i)) = g(\pi(\mathbf{x}_i)) = g(\mathbb{E}(Y_i|\mathbf{x}_i))$. Table

*Corresponding Author

(1) provides specific definitions of the link functions considered in this paper, along with their corresponding cdfs.

Model	Link function	cdf
Probit	$\Phi^{-1}(v)$	$\Phi(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$
Compit	$\log[-\log(1-v)]$	$1 - e^{-e^u}$
Cauchit	$\tan\left[\pi v - \frac{\pi}{2}\right]$	$\frac{1}{\pi}\left[\tan^{-1}(u) + \frac{\pi}{2}\right]$
Logit	$\log\left[\frac{v}{1-v}\right]$	$\Lambda(u) = \frac{1}{1+e^{-u}}$

Table 1: Link functions along with corresponding cdfs

The above link functions have been used extensively in a wide variety of applications in fields as diverse as medicine, engineering, economics, psychology, education just to name a few. The logit link function for which

$$\pi(\mathbf{x}_i) = \Pr[Y_i = 1|\mathbf{x}_i] = F(\eta(\mathbf{x}_i)) = \Lambda(\eta(\mathbf{x}_i)) = \frac{1}{1 + e^{-\eta(\mathbf{x}_i)}} \tag{3}$$

is arguably the most commonly used of all of them, probably because it provides a nice interpretation of the regression coefficients in terms of the ratio of the odds. In fact, the literature on both the theory and applications based on the logistic distribution is so vast it would be unthinkable to reference even a fraction of it. Some recent authors like Zelterman (1989), Schumacher et al. (1996), Nadarajah (2004), Lin and Hu (2008) and Nassar and Elmasry (2012) provide extensive studies on the characteristics of generalized logistic distributions, somehow answering the ever increasing interest in the logistic family of distributions. Indeed, applications abound that make use of both the standard logistic regression model and the so-called generalized logistic regression model, as can be seen in Hout et al. (2007) and Tamura and Giampaoli (2013). The probit link, for which

$$\pi(\mathbf{x}_i) = \Pr[Y_i = 1|\mathbf{x}_i] = F(\eta(\mathbf{x}_i)) = \Phi(\eta(\mathbf{x}_i)) = \int_{-\infty}^{\eta(\mathbf{x}_i)} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \tag{4}$$

is clearly the second most commonly used of all the link functions, with Bayesian researchers seemingly topping the charts in its use. See Basu and Mukhopadhyay (2000), Csató et al. (2000), Chakraborty (2009) for a few examples of probit use in binary classification in the Bayesian setting. Armagan and Zaretzki (2011) is just another one of the references pointing to the use of the probit link function in the statistical data mining and machine learning communities.

In the presence of some many possible choices of link functions, the natural question to ask is: how does one go about choosing the right/suitable/appropriate link function for the problem at hand? To say that the logit link function is popular is an understatement. Most experts and non-experts alike who deal with binary classification tend to almost automatically choose the logit link, to the point that it - the logit link - has almost been attributed a transcendental place. From experience, experimentation and mathematical proof, it is our view, a view shared by Feller (1971) and Feller (1940), that all these link function are equivalent, both structurally and predictively. Indeed, our conjectured equivalence of binary regression link functions is strongly supported by William Feller in his vehement criticism of the overuse of the logit link function and

a tendency to give it a place above the rest of existing link functions. In Feller (1971)'s own words: *An unbelievably huge literature tried to establish a transcendental "law of logistic growth"; measured in appropriate units, practically all growth processes were supposed to be represented by a function of the form* (3) *with t representing time. Lengthy tables, complete with chi-square tests, supported this thesis for human populations, for bacterial colonies, development of railroads, etc. Both height and weight of plants and animals were found to follow the logistic law even though it is theoretically clear that these two variables cannot be subject to the same distribution. Laboratory experiments on bacteria showed that not even systematic disturbances can produce other results. Population theory relied on logistic extrapolations (even though they were demonstrably unreliable). The only trouble with the theory is that not only the logistic distribution but also the normal, the Cauchy, and other distributions can be fitted to the same material with the same or better goodness of fit. In this competition the logistic distribution plays no distinguished role whatever; most contradictory theoretical models can be supported by the same observational material.*

As a matter of fact, it's obvious from the plot of their densities for instance that the probit and logit are virtually identical, almost superposed one on top of the other. It is therefore not surprising that one would empirically notice virtually no difference when the two are compared on the same binary regression task. Despite this apparent indistinguishability due to many of their similarities, it is fair to recognize that the two functions different, at least by definition and by their very algebra. Chambers and Cox (1967) argue in their paper that probit and logit will yield different results in the multivariate context. Their work is a rarety in a context where most researchers seem to have settled comfortably with the acceptance of the fact that the two links are essentially the same from a utility perspective. For such researchers, using one over the other is determined solely by mathematical convenience and a matter of taste. We demonstrate both theoretically and computationally that they all predictively equivalent in the univariate case, but we also provide a characterization of the conditions under which they tend to differ in the multivariate context.

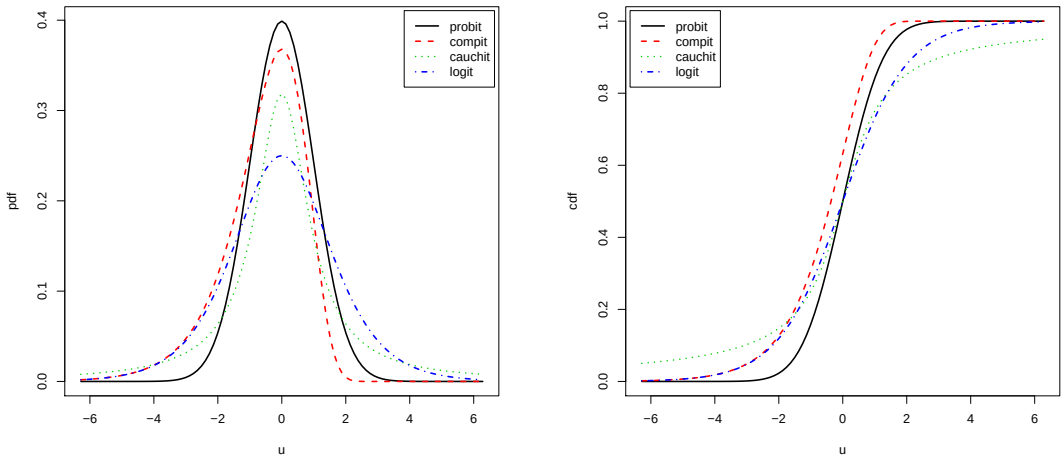

Figure 1: (`Left`) *Densities corresponding to the link functions* (`Right`) *cdfs corresponding to the link functions. The similarities are clear around the center of the distributions, but also some differences can be seen at the tails*

Throughout this work, we perform model comparison and model selection using both Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Taking the view that the

ability of an estimator to generalize well over the whole population, provides the best measure of it ultimate utility, we provide extensive comparisons of the performances of each link functions based on their corresponding test error. In the present work, we perform a large number of simulations in various dimensions using both artificial and real life data. Our results persistently reveal the performance indistinguishability of the links in univariate settings, but some sharp differences begin to appear as the dimension of the input space (number of variables measured) increased.

The rest of this paper is organized as follows: section 2 presents some general definitions, namely our meaning of the terms predictive equivalence and structural equivalence, along with some computational demonstrations on simulated and real life data. This section also clearly describes our approach to demonstrating/verifying our claimed results. We show in this section, that for low to moderate dimensional spaces, goodness of fit and predictive performance measures reveal the equivalence between probit and logit, and even shows equivalence among the other link functions. Section 3 provides our formal proof of the equivalence of probit and logit. Section 4 reveals that there might be some differences in performance when the input space becomes very large. Our demonstration in this section in based on the famous AT&T 57-dimensional Email Spam Data set. Section 5 provides a conclusion and a discussion, along with insights into extensions of the present work.

II. Definitions, Methodology and Verification

Throughout this work, we consider comparing models both on the merits of goodness of fit, and predictive performance. With that in mind, we can then define equivalence both from a goodness of fit perspective and also from a predictive optimality perspective.From a predictive analytics perspective for instance, an important question to ask is: given a randomly selected vector $\mathbf{x}$, what is the probability that the prediction made by probit will differ from the one made by logit? In other words, how often do the probit and logit link functions yield difference predictions? This is particularly important in predictive analytics in the data mining and machine learning where the nonparametric nature of most models forces the experimenter to focus on the utility of the estimator rather than its form. We respond to this need by defining what we call the $100(1-\alpha)\%$ predictive equivalence.

II.1 Basic definitions and results

Definition 1. *Given a input space $\mathcal{X}$ and a binary response space $\mathcal{Y} = \{0, 1\}$ along with a link function having corresponding cdf $F(\cdot)$, a classifier h will take $\mathbf{x} \in \mathcal{X}$ and yield the prediction*

$$h(\mathbf{x}) = \frac{1}{2}\left\{1 + \mathsf{sign}\left(\pi(\mathbf{x}) - \frac{1}{2}\right)\right\},$$

where $\pi(\mathbf{x}) = \Pr[Y = 1|\mathbf{x}] = F(\eta(\mathbf{x}))$.

Definition 2. *Let h_1 and h_2 be two classifiers defined on the same p-dimensional input space $\mathcal{X}$. We shall say that h_1 and h_2 are $100(1-\alpha)\%$ predictively equivalent if $\forall X \in \mathcal{X}$ drawn according to the density $p_X(\mathbf{x})$,*

$$\Pr\Big[h_1(X) \neq h_2(X)\Big] = \alpha.$$

In other words, the probability of disagreement between the two classifiers is α. When $\alpha = 0$, we say that h_1 and h_2 are perfectly predictively equivalent.

Lemma 1. *Let* $\lambda = \sqrt{\frac{\pi}{8}}$. *If* $X \sim \texttt{Logistic}(0,1)$, *and* $Y = \sqrt{\frac{\pi}{8}}X$, *then* $Y \overset{approx}{\sim} N(0,1)$.

Proof. Figure (2) below shows that the scaled version of the logistic cdf lines up almost perfectly with the standard normal. □

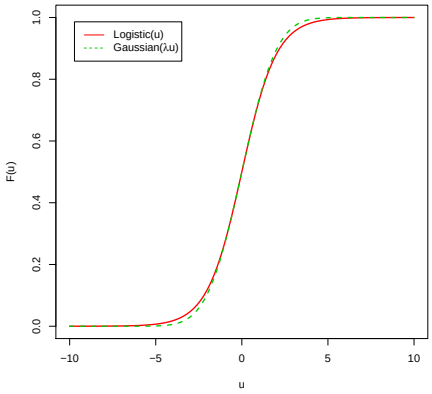

Figure 2: *Scaled version of the logistic CDF superposed on the standard normal CDF.*

Definition 3. *Let* M_1 *and* M_2 *be two binary regression models based on two different link functions defined on the same p-dimensional input space. We shall say that* M_1 *and* M_2 *are structurally equivalent if there exists a nonzero real constant* $\lambda \in \mathbb{R}^*$ *such that* $\beta_j^{(\texttt{M}_1)} \approx \lambda \beta_j^{(\texttt{M}_2)}$ *for all* $j = 1, \cdots, p$. *In other words, the parameters of* M_1 *are just a scaled version of the parameters of* M_2, *so that knowing the parameters of* M_1 *is sufficient to completely determine the parameters of* M_2, *and vice-versa.*

Theorem 1. *The logit and probit models are structurally equivalent.*

Proof. Thanks to Lemma (1), we can write

$$\Lambda(\mathbf{x}^\top \boldsymbol{\beta}^{(\texttt{logit})}) \approx \Phi(\lambda \mathbf{x}^\top \boldsymbol{\beta}^{(\texttt{logit})}) = \Phi(\mathbf{x}^\top \lambda \boldsymbol{\beta}^{(\texttt{logit})}) = \Phi(\mathbf{x}^\top \boldsymbol{\beta}^{(\texttt{probit})}),$$

where

$$\boldsymbol{\beta}^{(\texttt{probit})} \approx \lambda \boldsymbol{\beta}^{(\texttt{logit})}.$$

□

Lemma 2. *Let* $\Phi(\cdot)$ *denote the standard normal cdf. Let* $\lambda = \sqrt{\frac{\pi}{8}}$. *Then*

$$\text{sign}\left(\Phi(z) - \frac{1}{2}\right) = \text{sign}\left(\Phi(\lambda z) - \frac{1}{2}\right).$$

Theorem 2. *The probit and logit link functions are perfectly predictively equivalent. Specifically, given an input space* $\mathcal{X}$ *and a density* $p_X(\mathbf{x})$ *on* $\mathcal{X}$,

$$\Pr\left[h_{\texttt{logit}}(X) \neq h_{\texttt{probit}}(X)\right] = 0,$$

for all $X \in \mathcal{X}$ *drawn according to* $p_X(\mathbf{x})$.

Proof. For a given X and parameter sets $\boldsymbol{\beta}^{(\texttt{probit})}$ and $\boldsymbol{\beta}^{(\texttt{logit})}$, we must show that

$$\delta = \Pr\left[\text{sign}\left(\Lambda(X^\top \boldsymbol{\beta}^{(\texttt{logit})}) - \frac{1}{2}\right) \neq \text{sign}\left(\Phi(X^\top \boldsymbol{\beta}^{(\texttt{probit})}) - \frac{1}{2}\right)\right] = 0,$$

Based on Lemma (1), we can write

$$\delta \approx \Pr\left[\text{sign}\left(\Phi(\lambda X^\top \boldsymbol{\beta}^{(\texttt{probit})}) - \frac{1}{2}\right) \neq \text{sign}\left(\Phi(X^\top \boldsymbol{\beta}^{(\texttt{probit})}) - \frac{1}{2}\right)\right]$$

Thanks to Lemma (2), it is straightforward to see that $\delta = 0$. □

II.2 Computational Verification via Simulation

To get deeper into how strongly related the probit and logit models are, we now seek to estimate via simulation, the constant coefficient that relates their parameter estimates. Indeed, we conjecture that $\hat{\boldsymbol{\beta}}^{(\texttt{logit})}$ and $\hat{\boldsymbol{\beta}}^{(\texttt{probit})}$ are linearly related via the regression equation

$$\hat{\boldsymbol{\beta}}^{(\texttt{probit})} = \tau + \theta \hat{\boldsymbol{\beta}}^{(\texttt{logit})} + \nu,$$

where τ is the intercept and ν is the noise term. To estimate one instance of θ, we generate M random replications of the dataset, and for each replication we estimate a copy of $\hat{\boldsymbol{\beta}}$, and with it we also compute an estimate of $\rho = \texttt{cor}(\hat{\boldsymbol{\beta}}^{(\texttt{probit})}, \hat{\boldsymbol{\beta}}^{(\texttt{logit})})$ the correlation coefficient between $\hat{\boldsymbol{\beta}}^{(probit)}$ and $\hat{\boldsymbol{\beta}}^{(logit)}$. By repeating the estimation R times, we gather data to determine the central tendency of θ and the corresponding correlation.

```
For r = 1 to R

    For s = 1 to S
      * Generate a replicate of the random sample of {(x_i, y_i), i = 1, ..., n}
      * Estimate the logit model coefficient β̂_s^(logit)
      * Estimate the probit model coefficient β̂_s^(probit)
    End
  - Store the simulated data D^(r) = {(β̂_s^(logit), β̂_s^(probit)), s = 1, ..., S}
  - Fit M^(r), the regression model β̂_s^(probit) = τ + θβ̂_s^(logit) + ν_s using D^(r)
  - Extract the coefficient θ̂^(r) from M^(r)
  - Compute ρ̂^(r) estimate of correlation between β̂^(probit) and β̂^(logit)

End

Collect {θ̂^(r) and ρ̂^(r), r = 1, ..., R}, then compute relevant statistics.
```

`Example 1:` We consider a random sample of $n = 199$ observations $\{(\mathbf{x}_i, y_i), i = 1, \cdots, n\}$ where the $\mathbf{x}_i$ are equally spaced points in an interval $[a, b]$, that is, $\mathbf{x}_i = a + \left(\frac{b-a}{n-1}\right)(i-1)$, and y_i are drawn from one of the binary regression models. For instance, we set the domain of $\mathbf{x}_i$ to

$[a, b] = [0, 1]$ and generate the Y_i's from a Cauchit model with slope $1/2$ and intercept 0, i.e., $Y_i \overset{iid}{\sim} \texttt{Bernoulli}(\pi(\mathbf{x}_i))$, with

$$\Pr[Y_i = 1|\mathbf{x}_i] = \pi(\mathbf{x}_i) = \frac{1}{\pi}\left[\tan^{-1}\left(\frac{1}{2}\mathbf{x}_i\right) + \frac{\pi}{2}\right],$$

Using $R = 99$ replications each running $S = 199$ random samples, we obtain the following results, see Fig (3). The most striking finding here is that the estimated coefficient of determination is roughly equal to 1, indicating that the knowledge of logit coefficient almost entirely helps determine the value of the probit coefficient. Hence our claim of structural equivalence between probit and logit. The value of the slope θ appears to be in the neighborhood of 0.6.

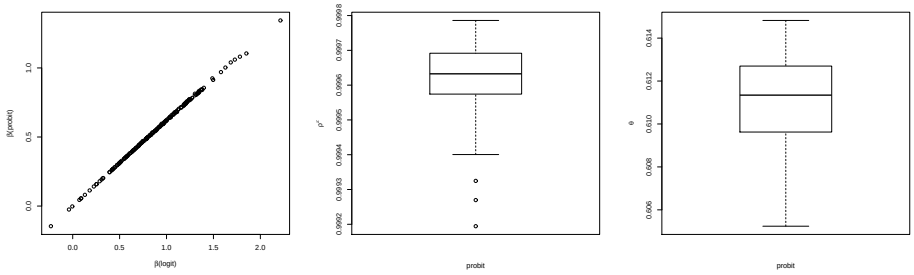

Figure 3: *(Left) Scatterplot of $\hat{\boldsymbol{\beta}}^{(probit)}$ against $\hat{\boldsymbol{\beta}}^{(logit)}$ based on the R replications generated; (Center) Boxplot of the R replications of the estimate of the coefficient of determination between $\hat{\boldsymbol{\beta}}^{(probit)}$ and $\hat{\boldsymbol{\beta}}^{(logit)}$; (Right) Boxplot of the R replications of the estimate of the slope θ*

`Example 2:` We now consider the famous Pima Indian Diabetes dataset, and obtain parameter estimates under both the logit and the probit models. The dataset is 7-dimensional, with $\mathtt{x}_1 =$ `npreg`, $\mathtt{x}_2 =$ `glu`, $\mathtt{x}_3 =$ `bp`, $\mathtt{x}_4 =$ `skin`, $\mathtt{x}_5 =$ `bmi`, $\mathtt{x}_6 =$ `ped` and $\mathtt{x}_7 =$ `age`. Under the logit model, the probability that patient i has diabetes given her characteristics $\mathbf{x}_i$ is given by

$$\Pr[\texttt{Diabetes}_i = 1|\mathbf{x}_i] = \pi(\mathbf{x}_i) = \frac{1}{1 + e^{-\eta(\mathbf{x}_i)}},$$

where

$$\eta(\mathbf{x}_i) = \beta_0 + \beta_1\texttt{npreg} + \beta_2\texttt{glu} + \beta_3\texttt{bp} + \beta_4\texttt{skin} + \beta_5\texttt{bmi} + \beta_6\texttt{ped} + \beta_7\texttt{age}.$$

We obtain the parameter estimates using `R`, and we display in the following table their values.

Model	npreg	glu	bp	skin	bmi	ped	age
Probit	0.0592	0.0192	−0.0024	−0.0017	0.0505	1.0682	0.0249
Logit	0.1031	0.0321	−0.0047	−0.0019	0.0836	1.8204	0.0411
Probit/Logit	0.57434	0.5987	0.5181	0.9073	0.6044	0.5868	0.6064

Table 2: Parameter estimates under probit and logit for the Pima Indian Diabetes Data Set

As can be seen in the above Table (2), the ratio of the probit coefficient over the logit coefficient is still a number around 0.6 for almost all the parameter. Indeed, the relationship

$$\hat{\beta}_j^{(probit)} \simeq \tau + 0.6\hat{\beta}_j^{(logit)} + \nu$$

appears to still hold true. The deviation from that pattern observed in variable `skin` is probably due to the extreme outlier in its distribution. It is important to note that although our theoretical justification was built under the simplified setting of a univariate model with no intercept, the relationship uncovered still holds true in a complete multivariate setting, with each predictor variable obeying the same relationship.

`Example 3:` We also consider the benchmark Crabs Leptograpsus dataset, and obtain parameter estimates under both the logit and the probit models. The dataset is 5-dimensional, with $\mathtt{x}_1 = \mathtt{FL}$, $\mathtt{x}_2 = \mathtt{RW}$, $\mathtt{x}_3 = \mathtt{CL}$, $\mathtt{x}_4 = \mathtt{CW}$ and $\mathtt{x}_5 = \mathtt{BD}$. Under the logit model, the probability that the sex of crab i is `male` given her characteristics $\mathbf{x}_i$ is given by

$$\Pr[\mathtt{sex}_i = 1|\mathbf{x}_i] = \pi(\mathbf{x}_i) = \frac{1}{1 + e^{-\eta(\mathbf{x}_i)}},$$

where

$$\eta(\mathbf{x}_i) = \beta_0 + \beta_1\mathtt{FL} + \beta_2\mathtt{RW} + \beta_3\mathtt{CL} + \beta_4\mathtt{CW} + \beta_5\mathtt{BD}.$$

We obtain the parameter estimates using `R`, and we display in the following table their values.

Model	FL	RW	CL	CW	BD
Probit	−3.5572	−11.4801	5.5364	−0.3101	1.6651
Logit	−6.1769	−19.9569	9.6643	−0.5746	2.8927
Probit/Logit	0.5758	0.5752	0.5728	0.5396	0.5756

Table 3: Parameter estimates under probit and logit for the Crabs Leptograpsus Data Set

As can be seen in the above Table (3), the estimate $\hat{\theta}$ of the ratio θ of the probit coefficient over the logit coefficient is still a number around 0.6 for almots all the parameter. Indeed, the relationship

$$\hat{\beta}_j^{(probit)} \simeq \tau + 0.6\hat{\beta}_j^{(logit)} + \nu$$

appears to still hold true. It is important to note that although our theoretical justification was built under the simplified setting of a univariate model with no intercept, the relationship uncovered still holds true in a complete multivariate setting, with each predictor variable obeying the same relationship.

Fact 1. *As can be seen from the examples above, the value of $\hat{\theta}$ lies in the neighborhood of* 0.6*, regardless of the task under consideration. This supports and confirms our conjecture that there is a fixed linear relationship between probit coefficients and logit coefficients to the point that knowing one implies knowing the other. Hence, the two models are structurally equivalent. In a sense, wherever logistic regression has been used successfully, probit regression will do just as a job. This result confirms what was already noticed and strongly expressed by Feller (1971) (pp 52-53).*

II.3 Likelihood-based verification of structural equivalence

In the proofs presented earlier, we focused on the parameters and never mentioned their estimates. We now provide a likelihood based verification of the structural equivalence of probit and logit. Without loss of generality, we shall focus on the univariate case where the underlying linear model does not have the intercept β_0, so that $\eta(\mathbf{x}_i) = \boldsymbol{\beta}\mathbf{x}_i$. With $\mathbf{x}_i$ denoting the predictor variable for the ith observation, we have the probability model $\Pr[Y_i = 1|\mathbf{x}_i] = \pi(\mathbf{x}_i) = F(\eta(\mathbf{x}_i)) = F(\boldsymbol{\beta}\mathbf{x}_i)$. Let $\hat{\boldsymbol{\beta}}^{(\texttt{logit})}$ and $\hat{\boldsymbol{\beta}}^{(\texttt{probit})}$ denote the estimates of $\boldsymbol{\beta}$ for the logit and the probit link functions respectively. Our first verification of the equivalence of the above link functions consists of showing that $\hat{\boldsymbol{\beta}}^{(\texttt{logit})}$ and $\hat{\boldsymbol{\beta}}^{(\texttt{probit})}$ are linearly related through $\hat{\boldsymbol{\beta}}^{(\texttt{probit})} = \tau + \theta\hat{\boldsymbol{\beta}}^{(\texttt{logit})} + \nu$, with a coefficient of determination very close to 1 and a slope θ that remains fixed regardless of the task at hand. We derive the approximate estimates of θ theoretically using Taylor series expansion, but we also confirm their values computationally by simulation.

Theorem 3. *Consider an i.i.d sample $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)$ where $\mathbf{x}_i \in \mathbb{R}$ is a real-valued predictor variable, and $y_i \in \{0, 1\}$ is the corresponding binary response. First consider fitting the probit model $\Pr[Y_i = 1|\mathbf{x}_i] = \pi(\mathbf{x}_i) = \Phi(\boldsymbol{\beta}\mathbf{x}_i)$ to the data, and let $\hat{\boldsymbol{\beta}}^{(\texttt{probit})}$ denote the corresponding estimate of $\boldsymbol{\beta}$. Then consider fitting the logit model and $\Pr[Y_i = 1|\mathbf{x}_i] = \pi(\mathbf{x}_i) = 1/(1 + \exp(-\boldsymbol{\beta}\mathbf{x}_i))$ to the data, and let $\hat{\boldsymbol{\beta}}^{(\texttt{logit})}$ denote the corresponding estimate of $\boldsymbol{\beta}$. Then,*

$$\hat{\boldsymbol{\beta}}^{(\texttt{probit})} \simeq 0.625\hat{\boldsymbol{\beta}}^{(\texttt{logit})}.$$

Proof. Given an i.i.d sample $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)$ and the model $\Pr[Y_i = 1|\mathbf{x}_i] = \pi(\mathbf{x}_i)$, the loglikelihood for $\boldsymbol{\beta}$ is given by

$$\ell(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Big\{ y_i \log \pi(\mathbf{x}_i) + (1 - y_i) \log(1 - \pi(\mathbf{x}_i)) \Big\}. \tag{5}$$

Under the logit link function, we have $\pi(\mathbf{x}_i) = 1/(1 + e^{-\boldsymbol{\beta}\mathbf{x}_i})$. Now, using a Taylor series expansion around zero for the two most important parts of the loglikelihood function, we get

$$\frac{\partial \log(\pi(\mathbf{x}_i))}{\partial \boldsymbol{\beta}} = \frac{\mathbf{x}_i}{2} - \frac{\mathbf{x}_i^2}{4}\boldsymbol{\beta} + \frac{\mathbf{x}_i^4}{48}\boldsymbol{\beta}^3 - \frac{\mathbf{x}_i^6}{480}\boldsymbol{\beta}^5,$$

and

$$\frac{\partial \log(1 - \pi(\mathbf{x}_i))}{\partial \boldsymbol{\beta}} = -\frac{\mathbf{x}_i}{2} - \frac{\mathbf{x}_i^2}{4}\boldsymbol{\beta} + \frac{\mathbf{x}_i^4}{48}\boldsymbol{\beta}^3 - \frac{\mathbf{x}_i^6}{480}\boldsymbol{\beta}^5.$$

The derivative of the approximate log-likelihood function for the logit model is then given by

$$\ell'(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i \left(\frac{\mathbf{x}_i}{2} - \frac{\mathbf{x}_i^2}{4}\boldsymbol{\beta} + \frac{\mathbf{x}_i^4}{48}\boldsymbol{\beta}^3 - \frac{\mathbf{x}_i^6}{480}\boldsymbol{\beta}^5 \right) + (1 - y_i) \left(-\frac{\mathbf{x}_i}{2} - \frac{\mathbf{x}_i^2}{4}\boldsymbol{\beta} + \frac{\mathbf{x}_i^4}{48}\boldsymbol{\beta}^3 - \frac{\mathbf{x}_i^6}{480}\boldsymbol{\beta}^5 \right) \right\},$$

which, upon ignoring the higher degree terms in the expansion becomes

$$\ell'(\boldsymbol{\beta}) \simeq \sum_{i=1}^{n} \Big\{ 4y_i\mathbf{x}_i - 2\mathbf{x}_i - \mathbf{x}_i^2\boldsymbol{\beta} \Big\}.$$

It is straightforward to see that solving $\ell'(\boldsymbol{\beta}) = 0$ for $\boldsymbol{\beta}$ yields

$$\hat{\boldsymbol{\beta}}^{(\texttt{logit})} \simeq 2 \left[\frac{2\sum_{i=1}^{n} \mathbf{x}_i y_i - \sum_{i=1}^{n} \mathbf{x}_i}{\sum_{i=1}^{n} \mathbf{x}_i^2} \right].$$

If we now consider the probit link function, we have $\pi(\mathbf{x}_i) = \Phi(\boldsymbol{\beta}\mathbf{x}_i) = \int_{-\infty}^{\boldsymbol{\beta}\mathbf{x}_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$. Using a derivation similar to the one performed earlier, and ignoring higher order terms, we get

$$\begin{aligned} \ell'(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &\simeq \sum_{i=1}^{n} \left\{ y_i \left(c_1 \mathbf{x}_i - 2c_2 \boldsymbol{\beta} \mathbf{x}_i^2 \right) + (1 - y_i) \left(-c_1 \mathbf{x}_i - 2c_2 \boldsymbol{\beta} \mathbf{x}_i^2 \right) \right\} \\ &= \sum_{i=1}^{n} \left\{ 2c_1 \mathbf{x}_i y_i - c_1 \mathbf{x}_i - 2c_2 \boldsymbol{\beta} \mathbf{x}_i^2 \right\} \end{aligned}$$

where $c_1 = 0.797885$ and $c_2 = 0.31831$. This leads to

$$\hat{\boldsymbol{\beta}}^{(\texttt{probit})} \simeq \frac{c_1}{2c_2} \left[\frac{2\sum_{i=1}^{n} \mathbf{x}_i y_i - \sum_{i=1}^{n} \mathbf{x}_i}{\sum_{i=1}^{n} \mathbf{x}_i^2} \right].$$

It is then straightforward to see that

$$\frac{\hat{\boldsymbol{\beta}}^{(\texttt{probit})}}{\hat{\boldsymbol{\beta}}^{(\texttt{logit})}} \simeq \frac{c_1}{4c_2} = 0.625, \quad \texttt{or equivalently} \quad \hat{\boldsymbol{\beta}}^{(\texttt{probit})} \simeq 0.625 \hat{\boldsymbol{\beta}}^{(\texttt{logit})}.$$

□

It must be emphasized that the above likelihood-based theoretical verifications are dependent on Taylor series approximations of the likelihood and therefore the factor of proportionality are bound to be inexact. It's re-assuring however to see that our computational verification does confirm the results found by theoretical derivation.

III. Similarities and Differences beyond Logit and Probit

Other aspects of our work reveal that the similarities proved and demonstrated above between the probit and the logit link functions extend predictively to the other link functions mentioned above. As far as structural equivalence or the lack thereof is concerned, Appendix A contains similar derivations for the relationship between cauchit and logit, and the relationship between compit and logit. As far as, predictive equivalence is concerned, we now present a verification based on the computation of many replications of the test error.

III.1 Computational Verification of Predictive Equivalence

We now computationally compare the predictive merits of each of the four link functions considered so far. To this end, we compare the estimated average test error yielded by the four link

functions. We do so by running $R = 10000$ replications of the split of the data set into training and test set, and at each iteration we compute the corresponding test error for the classifier corresponding to each link functions. Specifically, given $\{(\mathbf{x}_i, y_i), i = 1, \cdots, n\}$, we randomly form a training set $\{(\mathbf{x}_i^{(\mathtt{tr})}, y_i^{(\mathtt{tr})}), i = 1, \cdots, n_{\mathtt{tr}}\}$ and a test set $\{(\mathbf{x}_i^{(\mathtt{te})}, y_i^{(\mathtt{te})}), i = 1, \cdots, n_{\mathtt{te}}\}$. We typically run $R = 10000$ replications of this split, with $2/3$ of the data allocated to the training set and $1/3$ to the test set. We define the test error here under the symmetric zero-one loss as

$$\hat{R}_{\mathtt{test}}(\hat{f}) = \frac{1}{n_{\mathtt{te}}} \sum_{i=1}^{n_{\mathtt{te}}} 1_{\{y_i^{(\mathtt{te})} \neq \hat{f}(\mathbf{x}_i^{(\mathtt{te})})\}} = \frac{\#\{y_i^{(\mathtt{te})} \neq \hat{f}(\mathbf{x}_i^{(\mathtt{te})})\}}{n_{\mathtt{te}}}.$$

For one iteration/replication for instance, $\hat{R}_{\mathtt{test}}(\hat{f}^{(\mathtt{probit})})$, $\hat{R}_{\mathtt{test}}(\hat{f}^{(\mathtt{compit})})$, $\hat{R}_{\mathtt{test}}(\hat{f}^{(\mathtt{cauchit})})$ and $\hat{R}_{\mathtt{test}}(\hat{f}^{(\mathtt{logit})})$ are the values of the test error generated by probit, compit, cauchit and logit respectively. After R replications, we have R random realizations of each of those four test errors. We then perform various statistical calculations on the R replications, namely *median, mean, standard deviation, kurtosis, skewness, IQR etc...*, to assess the similarity and the differences among the link functions. We perform the similar R replications for model comparison using both AIC and BIC.

`Example 4:` Verification of Predictive Equivalence on Artificial Data: $\{(\mathbf{x}_i, y_i), i = 1, \cdots, n\}$ where $\mathbf{x}_i \sim \mathtt{Normal}(0, 2^2)$ and $y_i \in \{0, 1\}$ are drawn for a cauchy binary regression model with $\beta_0 = 1$ and $\beta_1 = 2$, namely $Y_i \sim \mathtt{Bernoulli}(\pi(\mathbf{x}_i))$ where

$$\pi(\mathbf{x}_i) = \Pr[Y_i = 1 | \mathbf{x}_i] = \frac{1}{\pi}\left[\tan^{-1}(1 + 2\mathbf{x}_i) + \frac{\pi}{2}\right].$$

Table(4) shows some statistics on $R = 10000$ replications of the test error. It goes without saying that all the four link functions are clearly indistinguishable as the estimated statistics are almost all perfectly equally across the board.

	probit	compit	cauchit	logit
median	0.16	0.16	0.16	0.16
mean	0.16	0.16	0.16	0.16
sd	0.04	0.04	0.03	0.04
skewness	0.21	0.26	0.26	0.24
kurtosis	3.18	3.51	3.20	3.20
cv	22.56	22.46	22.25	22.57
IQR	0.05	0.04	0.05	0.05
min	0.06	0.04	0.06	0.06
max	0.30	0.32	0.31	0.30

Table 4: *Statistics based on $R = 10000$ replicates of the test error on the artificial data set described above. It's clear that the values are indistinguishable across the four link functions.*

`Example 5:` Verification of Predictive Equivalence on the Pima Indian Diabetes Dataset: We once again consider the famous Pima Indian Diabetes dataset. The Pima Indian Diabetes Dataset is arguably one the most used benchmark data sets in the statistics and pattern recognition community. As can be see in Table (5), there is virtually no difference between the models. In other words, on the Pima Indian Diabetes data set, the four link functions are predictive equivalent.
It's also noteworthy to point out that all the four models also yield similar goodness of fit measures when scored using AIC and BIC. Indeed, Figure (4) reveals that over the $R = 10000$ replica-

	probit	compit	cauchit	logit
median	0.25	0.24	0.25	0.25
mean	0.25	0.25	0.26	0.25
sd	0.04	0.04	0.05	0.04
skewness	0.06	0.07	0.06	0.06
kurtosis	2.92	2.95	2.95	2.92
cv	17.84	18.33	17.62	17.85
IQR	0.06	0.06	0.07	0.06
min	0.09	0.07	0.10	0.09
max	0.43	0.40	0.45	0.42

Table 5: *Statistics based on $R = 10000$ replicates of the test error on the Pima Indian Diabetes data set. It's quite obvious that the values are indistinguishable across the four link functions.*

tions of the split of the data into training and test set, both the AIC and BIC are distributionally similar across all the four link functions. Despite the slight difference shown by the Cauchit

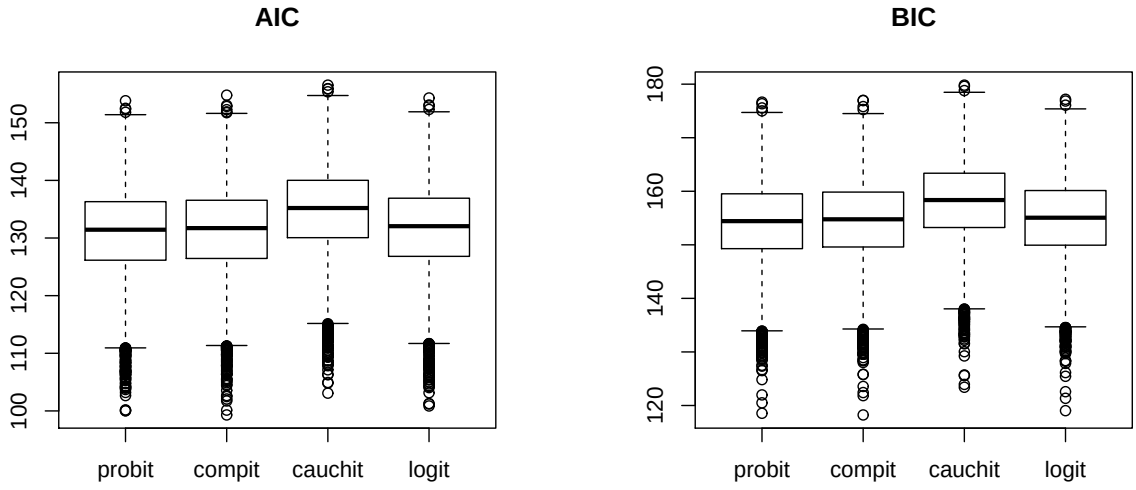

Figure 4: *Comparative boxplots of both AIC and BIC across the four link functions based on $R = 10000$ replications of model fittings.*

model, it is fair to say that all the link functions are equivalent in terms of goodness of fit. Once again, this is yet another evidence to support and somewhat reinforce/confirm Feller (1971)'s strong claim that all these link functions are equivalent in terms of goodness of fit, and that the over-glorification of the logit model is at best misguided if not unfounded.

III.2 Evidence of Differences in High Dimensional Spaces

Simulated evidence: We generate $s = 10000$ observations in the interval $[-15, 15]$. For each link function, we compute the sign of $F(\mathbf{x}_i) - 1/2$ for $i = 1, \cdots, s$. We then generate a table containing the percentage of times the signs differ.

	probit	compit	cauchit	logit
probit	0.000	0.004	0.000	0.000
compit	0.004	0.000	0.004	0.004
cauchit	0.000	0.004	0.000	0.000
logit	0.000	0.000	0.000	0.000

Table 6: *All the pairs reveal a disagreement of* 0% *except the pairs involving the compit.*

Computational Demonstrations on the Email Spam Data: Unlike all the other data sets encountered thus far, the email spam data set is a fairly high dimensional data set. It has a total of $p = 57$ variables and a remarkable $n = 4601$ observations.

	probit	compit	cauchit	logit
median	0.08	0.13	0.07	0.07
mean	0.10	0.13	0.07	0.08
sd	0.03	0.03	0.04	0.01
skewness	1.75	0.88	9.16	4.86
kurtosis	9.29	8.56	103.95	40.58
cv	34.41	20.61	51.15	14.32
IQR	0.04	0.04	0.01	0.01
min	0.06	0.07	0.05	0.06
max	0.41	0.38	0.62	0.18

Table 7: *Email Spam Data Set Results*

Clearly, the results depicted in Table (7) reveal some drastic differences in performance among the four link functions on this rather high dimensional data. The boxplots below reinforce these findings as they show that in terms of goodness of fit measured through AIC and BIC, the compit model deviates substantially from the other models.

IV. Conclusion and discussion

Throughout this paper, we have explored both conceptually/methodologically and computationally the similarities among four of the most commonly used link functions in binary regression. We have theoretically shed some light on some of the structural reasons that explain the indistinguishability in performance in the univariate settings among the four link functions considered. Although section 2 concentrated mainly on the equivalence of the logit and probit, the Appendix provides a similar derivation for both the cauchit and the complementary log log link functions. We have also demonstrated by computational simulations that the four link functions are essentially equivalent both structurally and predictively in the univariate setting and in low dimensional spaces.

Our last example showed computationally that the four link functions might differ quite substantially when the dimensional of the input space becomes extremely large. We notice specifically that the performance in high dimensional spaces tends to defend on the internal structure of the input: completely orthogonal designs tending to bode well with all the perfectly symmetric link functions while the non orthogonal designs deliver best performances under the complementary log log. Finally, the sparseness of the input space tends to dictate the choice of the most appropriate link function, Cauchit tending to be the model of choice under high level of sparseness.

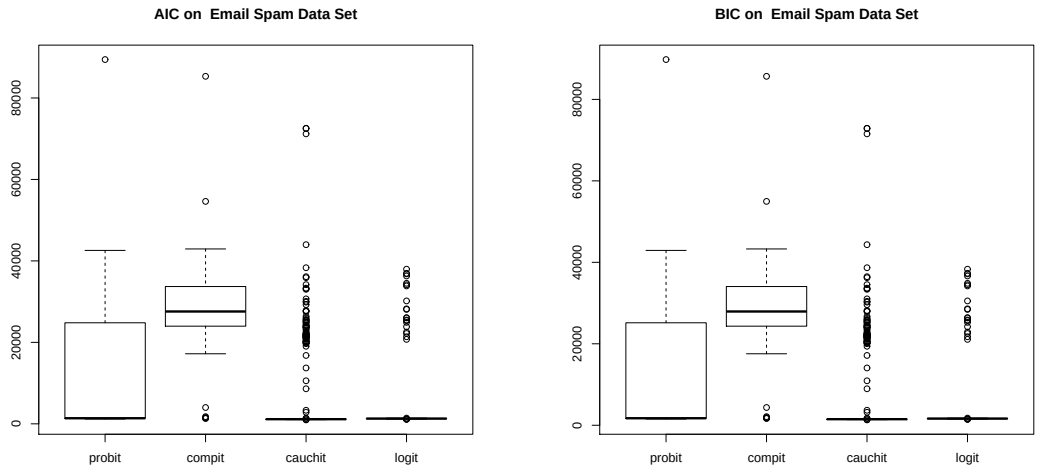

Figure 5: *Comparative Boxplots assessing the goodness of fit of the four link functions using AIC and BIC over* $R = 10000$ *replications of model fitting under each of the link functions.*

In our future work, we intend to provide as complete a theoretical characterization as possible in extremely high dimensional spaces, namely providing the conditions under which each of the link function will yield the best fit for the data.

References

Armagan, A. and R. Zaretzki (2011). A note on mean-field variational approximations in bayesian probit models. *Computational Statistics and Data Analysis 55*, 641–643.

Basu, S. and S. Mukhopadhyay (2000). Bayesian analysis of binary regression using symmetric and asymmetric links. *Sankhya: The Indian Journal of Statistics 62*(3), 372–387.

Chakraborty, S. (2009). Bayesian binary kernel probit model for microarray based cancer classification and gene selection. *Computational Statistics and Data Analysis 53*, 4198–4209.

Chambers, E. and D. Cox (1967). Discrimination between alternative binary response models. *Biometrika 54*(3/4), 573–578.

Csató, L., E. Fokoué, M. Opper, B. Schottky, and O. Winther (2000). Efficient approaches to gaussian process classification. In S. A. Solla, T. K. Leen, and e. K.-R. Müller (Eds.), *Advances in Neural Information Processing Systems*, Number 12. MIT Press.

Feller, W. (1940). On the logistic law of growth and its empirical verification in biology. *Acta Biotheoretica 5*, 51–66.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications* (Second ed.), Volume II. New York: John Wiley and Sons.

Hout, A., P. Heijden, and R. Gilchrist (2007). The logistic regression model with response variables subject to randomized response. *Computational Statistics and Data Analysis 51*, 6060–6069.

Lin, G. D. and C. Y. Hu (2008). On characterizations of the logistic distribution. *Journal of Statistical Planning and Inference 138*, 1147–1156.

Nadarajah, S. (2004). Information matrix for logistic distributions. *Mathematical and Computer Modelling 40*, 953–958.

Nassar, M. M. and A. Elmasry (2012). A study of generalized logistic distributions. *Journal of the Egyptian Mathematical Society 20*, 126–133.

Schumacher, M., R. Robner, and W. Vach (1996). Neural networks and logistic regression: Part i. *Computational Statistics and Data Analysis 21*, 661–682.

Tamura, K. A. and V. Giampaoli (2013). New prediction method for the mixed logistic model applied in a marketing problem. *Computational Statistics and Data Analysis 66*, 202–216.

Zelterman, D. (1989). Order statistics for the generalized logistic distribution. *Computational Statistics and Data Analysis 7*, 69–77.

V. Appendix A

Theorem 4. *Consider an i.i.d sample* $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)$ *where* $\mathbf{x}_i \in \mathbb{R}$ *is a real-valued predictor variable, and* $y_i \in \{0, 1\}$ *is the corresponding binary response. First consider fitting the cauchit model* $\Pr[Y_i = 1|\mathbf{x}_i] = \pi(\mathbf{x}_i) = \frac{1}{\pi}\left[\tan^{-1}(\boldsymbol{\beta}\mathbf{x}_i) + \frac{\pi}{2}\right]$ *to the data, and let* $\hat{\boldsymbol{\beta}}^{(\texttt{cauchit})}$ *denote the corresponding estimate of* $\boldsymbol{\beta}$*. Then consider fitting the logit model and* $\Pr[Y_i = 1|\mathbf{x}_i] = \pi(\mathbf{x}_i) = 1/(1 + \exp(-\boldsymbol{\beta}\mathbf{x}_i))$ *to the data, and let* $\hat{\boldsymbol{\beta}}^{(\texttt{logit})}$ *denote the corresponding estimate of* $\boldsymbol{\beta}$*. Then,*

$$\hat{\boldsymbol{\beta}}^{(\texttt{cauchit})} \simeq \frac{\pi}{4}\hat{\boldsymbol{\beta}}^{(\texttt{logit})}.$$

Proof. Given an i.i.d sample $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)$ and the model $\Pr[Y_i = 1|\mathbf{x}_i] = \pi(\mathbf{x}_i)$, the loglikelihood for $\boldsymbol{\beta}$ is given by

$$\ell(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Big\{ y_i \log \pi(\mathbf{x}_i) + (1 - y_i) \log(1 - \pi(\mathbf{x}_i)) \Big\}. \tag{6}$$

For the Cauchit for instance, $\pi(\mathbf{x}_i) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(\boldsymbol{\beta}\mathbf{x}_i)$. We use the Taylor series expansion around zero for both $\log(\pi(\mathbf{x}_i))$ and $\log(1 - \pi(\mathbf{x}_i))$.

$$\log \pi(\mathbf{x}_i) = -\log 2 + \frac{2\boldsymbol{\beta}\mathbf{x}_i}{\pi} - \frac{2\boldsymbol{\beta}^2\mathbf{x}_i^2}{\pi^2} - \frac{2(\pi^2 - 4)\boldsymbol{\beta}^3\mathbf{x}_i^3}{3\pi^3} + O(\mathbf{x}_i^4)$$

and

$$\log(1 - \pi(\mathbf{x}_i)) = -\log 2 - \frac{2\boldsymbol{\beta}\mathbf{x}_i}{\pi} - \frac{2\boldsymbol{\beta}^2\mathbf{x}_i^2}{\pi^2} + \frac{2(\pi^2 - 4)\boldsymbol{\beta}^3\mathbf{x}_i^3}{3\pi^3} + O(\mathbf{x}_i^4)$$

A first order approximation of the derivative of the log-likelihood with respect to $\boldsymbol{\beta}$ is

$$\begin{aligned} \ell'(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \left\{ y_i \left(\frac{2\mathbf{x}_i}{\pi} - \frac{4\boldsymbol{\beta}\mathbf{x}_i^2}{\pi^2} \right) + (1 - y_i) \left(-\frac{2\mathbf{x}_i}{\pi} - \frac{4\boldsymbol{\beta}\mathbf{x}_i^2}{\pi^2} \right) \right\} \\ &= \sum_{i=1}^{n} \left\{ \frac{4}{\pi}\mathbf{x}_i y_i - \frac{2}{\pi}\mathbf{x}_i - \frac{4}{\pi^2}\boldsymbol{\beta}\mathbf{x}_i^2 \right\} \end{aligned}$$

Solving $\ell'(\boldsymbol{\beta}) = 0$ yields

$$\hat{\boldsymbol{\beta}} = \frac{\frac{4}{\pi}\sum_{i=1}^{n}\mathbf{x}_i y_i - \frac{2}{\pi}\sum_{i=1}^{n}\mathbf{x}_i}{\frac{4}{\pi^2}\sum_{i=1}^{n}\mathbf{x}_i^2}$$

which simplifies to

$$\hat{\boldsymbol{\beta}}^{(\texttt{cauchit})} = \frac{\pi}{2}\left[\frac{2\displaystyle\sum_{i=1}^{n}\mathbf{x}_i y_i - \sum_{i=1}^{n}\mathbf{x}_i}{\displaystyle\sum_{i=1}^{n}\mathbf{x}_i^2}\right]$$

□

Finally, if we now consider the complementary log-log link function, we have $\pi(\mathbf{x}_i) = 1 - e^{-e^{\boldsymbol{\beta}\mathbf{x}_i}}$. Using aderivation similar to the ones performed earlier, we get

$$\hat{\boldsymbol{\beta}}^{(\texttt{cloglog})} = \left[\frac{(1+c_1)\displaystyle\sum_{i=1}^{n}\mathbf{x}_i y_i - \sum_{i=1}^{n}\mathbf{x}_i}{\displaystyle\sum_{i=1}^{n}\mathbf{x}_i^2 - (1+2c_2)\sum_{i=1}^{n}\mathbf{x}_i^2 y_i}\right]$$

where $c_1 = 1/(e-1)$ and $c_2 = 1/2(e-1)^2$.

It is easy to see that

$$\frac{1}{\hat{\boldsymbol{\beta}}^{(\texttt{cloglog})}} = A + \frac{1}{\left[\dfrac{(1+c_1)\displaystyle\sum_{i=1}^{n}\mathbf{x}_i y_i - \sum_{i=1}^{n}\mathbf{x}_i}{\displaystyle\sum_{i=1}^{n}\mathbf{x}_i^2}\right]} \simeq A + \frac{2}{\hat{\boldsymbol{\beta}}^{(\texttt{logit})}}$$