Journal of Statistical Planning and Inference 149 (2014) 116–124

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Infinitely imbalanced binomial regression and deformed exponential families

Tomonari Sei

Department of Mathematics, Keio University, 3-14-1, Hiyoshi, Kohoku-ku, Yokohama, Kanagawa, Japan

ARTICLE INFO

Article history:
Received 13 July 2013
Received in revised form
6 January 2014
Accepted 9 January 2014
Available online 24 January 2014

Keywords:
Binomial regression
Extreme value theory
Imbalanced data
Poisson point process
q-Exponential family

ABSTRACT

The logistic regression model is known to converge to a Poisson point process model if the binary response tends to infinity imbalanced. In this paper, it is shown that this phenomenon is universal in a wide class of link functions on binomial regression. The proof relies on the extreme value theory. For the logit, probit and complementary log–log link functions, the intensity measure of the point process becomes an exponential family. For some other link functions, deformed exponential families appear. A penalized maximum likelihood estimator for the Poisson point process model is suggested.

1. Introduction

Let $\{(X_i, Y_i)\}_{i=1}^m$ be m independently and identically distributed observable data on $\mathbb{R}^p \times \{0,1\}$. The conditional distribution of Y_i given X_i is assumed to be

$$P(Y_i = 1|X_i, a, b) = G(a + b^{\mathrm{T}} X_i), \quad a \in \mathbb{R}, \quad b \in \mathbb{R}^p, \tag{1}$$

where $G(\cdot)$ is a one-dimensional cumulative distribution function. The inverse function $G^{-1}(p) = \sup\{z : G(z) \le p\}$ is the link function in terms of generalized linear models. Denote the marginal distribution of X_i by $F(dX_i)$. The distribution function G is typically the logistic, standard normal or Gumbel distributions. The corresponding link functions are the logit, probit and complementary log–log functions, respectively. For the three examples, the log-likelihood function of (1) is concave; see Wedderburn (1976).

Our interest is the situation that the data is highly imbalanced. In other words, the probability of success is almost zero. Examples of such cases are fraud detection, medical diagnosis, political analysis and so forth; see e.g. Bolton and Hand (2002), Chawla et al. (2004), Jin et al. (2005), and King and Zeng (2001). For the data without covariates, Poisson's law of rare events is well known: if $P(Y_i = 1) = \lambda/m + o(m^{-1})$, then the probability distribution of $\sum_{i=1}^m Y_i$ converges to the Poisson distribution with the mean parameter λ. From this observation, for highly imbalanced data, it is natural to consider that the true parameter (a,b) in (1) depends on m, say (a_m, b_m), and $G(a_m) \to 0$ as $m \to \infty$.

Owen (2007) showed that the maximum likelihood estimator of the logistic regression model converges to that of an exponential family if $\sum_{i=1}^m Y_i$ is fixed and m goes to infinity. This result is roughly derived as follows. Consider the model (1)

E-mail address: sei@math.keio.ac.jp

0378-3758/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2014.01.002

with the logistic distribution $G(z)=e^z/(1+e^z)$. Take $a_m(\alpha)=-\log m+\alpha$ and $b_m(\beta)=\beta$ for any fixed α and β. Then we obtain

$$P(Y_i=1|X_i,a_m(\alpha),b_m(\beta))=\frac{e^{-\log m+\alpha+\beta^{\mathrm{T}}X_i}}{1+e^{-\log m+\alpha+\beta^{\mathrm{T}}X_i}}=\frac{e^{\alpha+\beta^{\mathrm{T}}X_i}}{m}+o(m^{-1}) \tag{2}$$

as $m\to\infty$. By Bayes' theorem, the conditional density of X_i given $Y_i=1$ with respect to the distribution $F(dX_i)$ is, at least formally,

$$\frac{e^{\beta^{\mathrm{T}}X_i}}{\int e^{\beta^{\mathrm{T}}x}F(dx)}+o(1). \tag{3}$$

This is an exponential family with the sufficient statistic X_i, and Owen's result follows.

Remark 1. To be precise, Owen (2007) proved the convergence result under a different setting from here. He assumed that the true conditional distribution of X_i given $Y_i=j$, $j\in\{0,1\}$, is any distribution F_j. In our setting, F_0 is asymptotically equal to F, and the density of F_1 with respect to F should satisfy (3). In other words, our setting becomes misspecified unless this equality is satisfied. We discuss this point again in Section 5.

Warton and Shepherd (2010) pointed out that the likelihood of logistic regression converges to a Poisson point process model with a specific form of intensity. Indeed, by (2), the probability $P(Y_i=1,X_i\in A)$ is approximately $m^{-1}\int_A e^{\alpha+\beta^{\mathrm{T}}x}F(dx)$ for any compact subset A of $\mathbb{R}^p$. Therefore, by Poisson's law of rare events, the number of observations X_i for which $X_i\in A$ and $Y_i=1$ is approximately distributed according to the Poisson distribution with mean $\int_A e^{\alpha+\beta^{\mathrm{T}}x}F(dx)$. This is the Poisson point process with the intensity measure $e^{\alpha+\beta^{\mathrm{T}}x}F(dx)$.

Baddeley et al. (2010) investigated approximation of the spatial Poisson point process by the pixel-based logistic regression. They showed that the complementary log–log link function is consistent under change of pixel sizes and that a split-pixel strategy works well when the covariate data is irregular.

In this paper, we consider the limit of various binomial regression models other than the logistic model. As expected from the result on logistic regression, the limit becomes a Poisson point process. A remarkable fact we prove is that the intensity measure of the point process should be a q-exponential family for some real number q. The q-exponential family, also called the deformed exponential family or α-family, is recently much investigated in the literature of statistical physics and information geometry; see e.g. Amari (1985), Amari and Nagaoka (2000), Amari and Ohara (2011), Naudts (2002), Naudts (2010), and Tsallis (1988). The precise definition is given in Section 2. The proof relies on the theory of extreme values. For example, for the probit or complementary log–log link functions, the limit of binomial regression is the usual exponential family as with the logit link. On the other hand, if G is the Cauchy distribution, then the limit becomes a q-exponential family with $q=2$. If the uniform distribution is used, $q=0$.

As a related work, Ding et al. (2011) introduced the t-logistic regression, that uses the q-exponential family for binary response, where $q=t$. In Section 3, we show that the t-logistic regression converges to the q-exponential family if $q\geq 0$.

In Section 4, we study a penalized maximum likelihood estimator on the q-exponential family of intensity measures. For some special cases, the estimator is reduced to a known admissible estimator for the Poisson mean parameter; see Ghosh and Yang (1988).

Some related problems are discussed in Section 5.

2. Imbalanced asymptotics of binomial regression

For each real number q, define the q-exponential function by

$$\exp_q(z)=\begin{cases} e^z & \text{if } q=1,\\ [1+(1-q)z]_+^{1/(1-q)} & \text{if } q\neq 1,\end{cases} \tag{4}$$

where $[z]_+=\max(z,0)$ and $[0]_+^{-1}=\infty$. This is inverse of the Box–Cox transformation with parameter $\lambda=1-q$. Note that $\exp_q(z)=\infty$ for $z\geq -1/(1-q)$ if $q>1$. The function $\exp_q(z)$ is convex if and only if $q\geq 0$.

Consider the binomial regression model (1) and put the following assumption on the distribution function G.

Assumption 1. There exist $q>0$, $c_m\in\mathbb{R}$ and $d_m>0$ such that

$$G(c_m+d_m z)=\frac{1}{m}\exp_q(z)+o(m^{-1}) \tag{5}$$

as $m\to\infty$ for each $z\in\mathbb{R}$.

In the extreme value theory, it is known that there is no other asymptotic form than (5) as long as it exists; see e.g. de Haan and Ferreira (2006, Theorems 1.1.2 and 1.1.3). The number q controls the lower tail structure of G. For example, the logistic distribution satisfies Assumption 1 with $q=1$, $c_m=-\log m$ and $d_m=1$. Other examples including the normal and Cauchy distributions are considered in Section 3.

We define

$$a_m(\alpha)=c_m+d_m\alpha \quad \text{and} \quad b_m(\beta)=d_m\beta \tag{6}$$