Available online at www.sciencedirect.com
ScienceDirect
Insurance: Mathematics and Economics 42 (2008) 779–786
INSURANCE MATHEMATICS & ECONOMICS
www.elsevier.com/locate/ime

A Bayesian dichotomous model with asymmetric link for fraud in insurance

Ll. Bermúdez[a,*], J.M. Pérez[b], M. Ayuso[c], E. Gómez[d], F.J. Vázquez[d]

[a] *Department of Economic, Financial and Actuarial Mathematics, University of Barcelona, 08034-Barcelona, Spain*
[b] *Department of Quantitative Methods in Economics, University of Granada, 18011-Granada, Spain*
[c] *Department of Econometrics, University of Barcelona, 08034-Barcelona, Spain*
[d] *Department of Quantitative Methods, University of Las Palmas de Gran Canaria, 35017-Las Palmas de G.C., Spain*

Received March 2007; received in revised form August 2007; accepted 21 August 2007

Abstract

Standard binary models have been developed to describe the behavior of consumers when they are faced with two choices. The classical logit model presents the feature of the symmetric link function. However, symmetric links do not provide good fits for data where one response is much more frequent than the other (as it happens in the insurance fraud context). In this paper, we use an asymmetric or skewed logit link, proposed by Chen et al. [Chen, M., Dey, D., Shao, Q., 1999. A new skewed link model for dichotomous quantal response data. J. Amer. Statist. Assoc. 94 (448), 1172–1186], to fit a fraud database from the Spanish insurance market. Bayesian analysis of this model is developed by using data augmentation and Gibbs sampling. The results show that the use of an asymmetric link notably improves the percentage of cases that are correctly classified after the model estimation.

IB classification: IB40

IM classification: IM20

JEL classification: C11

Keywords: Bayesian statistics; Logit model; Gibbs sampling; Automobile insurance; Fraud

1. Introduction

Classically, standard binary models have been used to assess the behavior of consumers faced with binary choices (McFadden, 1974, 1981). Popular binary models use symmetric links as the logit or the probit link for analyzing variables which are related to the probability of choosing between category zero or one. In the context of generalized linear modelling for binary response, the link function is defined as a transformation of the expected value of the response variable (i.e. the probability that the dependent variable takes value zero or one) so that fitted values must be inside the range [0, 1]. A symmetric link function $F(\cdot)$ satisfies the property $F(k-x) = F(k+x)$ for a given constant k and all x's. Sometimes, however, the individual choice is clearly related to one category more than to the other. This happens in the context of insurance fraud, where databases are not normally balanced: they contain a higher number of non-fraudulent than fraudulent cases. In this situation, the use of an asymmetric or skewed logit link (like the one proposed by Chen et al. (1999)) can help us to improve the fitted logit model quality, providing very good results for the success percentages at the matrix confusion. In this paper we describe the Bayesian analysis of this model, using data augmentation and Gibbs sampling.

In many fields of application, dichotomous qualitative models have been studied using non-Bayesian techniques. Amemiya (1981), Hausman and McFadden (1984) and McFadden (1981) are excellent references for a review. However, recently there has been great interest in Bayesian analysis of binary and polychotomous response models.

* Corresponding address: Departament de Matemàtica Econòmica, Financera i Actuarial, Universitat de Barcelona, Diagonal 690, 08034-Barcelona, Spain. Tel.: +34 93 4034853; fax: +34 93 4034892.
E-mail address: lbermudez@ub.edu (Ll. Bermúdez).

0167-6687/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.insmatheco.2007.08.002

McCulloch et al. (1999), Albert and Chib (1993), Koop and Poirier (1993), Stukel (1988), Basu and Mukhopadhayay (2000) and Bazán et al. (2006), among others, provide good examples of this approach. In the area of insurance, Artís et al. (1999), Artís et al. (2002), Belhadji and Dionne (1997), Belhadji et al. (2000) and Caudill et al. (2005) estimated discrete choice models for fraud behavior from a non-Bayesian point of view. However, very little has been said in the literature (Viaene et al., 2002; Viaene et al., 2007) about the Bayesian analysis of fraud behavior in the automobile insurance market.

An insurance portfolio is a collection of N individuals or contracts where a binary random variable y_i is observed for the policyholder i, $i = 1, \ldots, N$. In this case, y_i equals one if the ith individual admits a fraud claim, and zero otherwise. It is obvious that y_i follows a Bernoulli distribution where $y_i = 1$ with probability p_i, and $y_i = 0$ with probability $1 - p_i$. Thus, $E(y_i) = p_i$, and $\mathrm{Var}(y_i) = p_i(1 - p_i)$. Let $\mathbf{y} = (y_1, \ldots, y_n)'$ be a sample of $n \leq N$ observations and $l(\mathbf{y}) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}$, be the the likelihood function. In the dichotomous response models, $p_i = F(\boldsymbol{x_i'\beta})$, where $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ik})'$ is a $k \times 1$ vector of covariates, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ is a $k \times 1$ vector of regression coefficients. The likelihood function can be written as:

$$l(\mathbf{y}) = \prod_{i=1}^{n} \left[F(\boldsymbol{x_i'\beta})\right]^{y_i} \left[1 - F(\boldsymbol{x_i'\beta})\right]^{1-y_i}. \tag{1}$$

Two special, well-known cases assume that the link function $F^{-1}(\cdot)$ is the inverse of the standard normal cumulative distribution function (probit model), or the inverse of the standard logistic cumulative distribution function (logit model). The main advantage of the logit model over the probit one is that it makes the interpretation of coefficients much easier. Although probit models are clearly appealing due to the relative ease of their computation and the modelling of the covariance structure, they present some problems with the parameter interpretation.

In this paper we focus on the logit model but using an asymmetric link, from a Bayesian point of view. In previous works on insurance fraud (Artís et al., 1999, 2002; Belhadji and Dionne, 1997), based on the use of symmetric links, the discrete model was estimated by maximizing a weighted likelihood function. Weights were included in the estimation procedure in order to account for the effect of the fraud over-representation in the samples used. In other studies such as Belhadji et al. (2000) or Pinquet et al. (2007) the authors set the probability threshold for the fraud claim classification according to the observed fraud detection rate in the portfolios used (which are representative of the total population).

In this study, we use a Bayesian skewed logit model (Chen et al., 1999) for fitting an insurance fraud database. This model incorporates the possibility of using asymmetric links in order to measure the probability of $y_i = 0$ and $y_i = 1$ in non-balanced samples (with a high proportion of zeros or ones). Firstly, we quantify the prior distribution for each of the parameters and for the skewness parameter. Secondly, we use the Bayes' theorem to calculate posterior model probabilities. In the empirical section of the paper, we observe a notable improvement for the regression fit results when the asymmetric Bayesian approach is used compared with those obtained with the classical logit model or the symmetric Bayesian model (which gives results similar to those of the classical model).

The rest of the paper is structured as follows. In Section 2, we present the Bayesian procedures for analyzing the new skewed logit model. Section 3 is devoted to the prior elicitation. In Sections 4 and 5, we present the data and results from an application of the proposed model to the Spanish automobile insurance fraud database. Finally in Section 6, we show the main conclusions and suggestions for future research related to this study.

2. Bayesian skewed logit model

The use of a logit skewed model can produce significantly better fits than the symmetric link model (Stukel, 1988). Recently, some Bayesian models proposing asymmetric links have been presented in the literature, as we noted in the introduction (Chen et al., 1999; Basu and Mukhopadhayay, 2000; Bazán et al., 2006). Although these models may complicate the computation of the required posterior distribution, we use the Markov Chain Monte Carlo (MCMC) and the Gibbs sampling procedures to obtain it (see Carlin and Polson (1992)).

Following Chen et al. (1999) we assume that the underlying latent variable has a skewed distribution. Thus, the model uses a vector of latent variables $\boldsymbol{w} = (w_1, \ldots, w_n)'$ in the following form:

$$y_i = \begin{cases} 0, & w_i < 0, \\ 1, & w_i \geq 0, \end{cases}$$

where

$$w_i = \boldsymbol{x_i'\beta} + \delta z_i + \varepsilon_i, \quad z_i \sim G, \varepsilon_i \sim F.$$

Here, z_i and ε_i are independent; G is the cdf (cumulative distribution function) of a skewed distribution, and F is the cdf of a symmetric distribution. Following Chen et al. (1999), we assume G to be the cdf of the half-standard normal distribution, and F the standard normal cdf. As the authors show, these functions allow us to ensure the identifiability of the model and produce several attractive properties.

The novelty of this model is the incorporation of the term δz_i in which $\delta \in (-\infty, \infty)$ is a skewness parameter. If $\delta > 0$, it means that the model increases the probability of $y_i = 1$, in our case, the probability of committing fraud. On the other hand, if $\delta < 0$, then the probability of $y_i = 0$ increases. In this way, the skewed model allows us to modify the commonly used symmetric link model by increasing or decreasing the probabilities that y_i equals zero or one. Obviously, if $\delta = 0$, then the skewed link model is reduced to a standard symmetric link model, because in this case the link function corresponds to a symmetric distribution.

Under this new model, the likelihood function in (1) can be rewritten as

$$l(\boldsymbol{\beta}, \delta \mid D) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \left[F\left(\boldsymbol{x_i'\beta} + \delta z_i\right)\right]^{y_i}$$