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

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

$$\times \left[1 - F\left(\boldsymbol{x}_i'\boldsymbol{\beta} + \delta z_i\right)\right]^{1-y_i} g(z_i)\mathrm{d}z_i,$$

where $D = (n, \boldsymbol{y}, \boldsymbol{x})$ represents the observed data.

Using the Bayesian approach, the posterior distribution of $(\boldsymbol{\beta}, \delta)$ based on the observed data D is represented by

$$\begin{aligned} p(\boldsymbol{\beta}, \delta \mid D) &\propto l(\boldsymbol{\beta}, \delta \mid D)\,\pi(\boldsymbol{\beta}, \delta) \\ &= \left\{\prod_{i=1}^{n} \int_{-\infty}^{\infty} \left[F\left(\boldsymbol{x}_i'\boldsymbol{\beta} + \delta z_i\right)\right]^{y_i} \right. \\ &\quad \left. \times \left[1 - F\left(\boldsymbol{x}_i'\boldsymbol{\beta} + \delta z_i\right)\right]^{1-y_i} g(z_i)\mathrm{d}z_i \right\} \pi(\boldsymbol{\beta}, \delta), \end{aligned} \tag{2}$$

where $\pi(\boldsymbol{\beta}, \delta)$ is the prior distribution of $(\boldsymbol{\beta}, \delta)$.

Finally, we can sample $(\boldsymbol{\beta}, \delta)$ from its posterior distribution given in (2) by using Markov Chain Monte Carlo (MCMC) for Bayesian Inference. To sample from the posterior distribution $p(\boldsymbol{\beta}, \delta \mid D)$, we introduce the new latent variables $\boldsymbol{z} = (z_1, \ldots, z_n)'$. Then the joint posterior distribution for $(\boldsymbol{\beta}, \delta, \boldsymbol{z})$ is represented as

$$\begin{aligned} p(\boldsymbol{\beta}, \delta, \boldsymbol{z} \mid D) &\propto \left\{\prod_{i=1}^{n} \int_{-\infty}^{\infty} \left[F\left(\boldsymbol{x}_i'\boldsymbol{\beta} + \delta z_i\right)\right]^{y_i} \right. \\ &\quad \left. \times \left[1 - F\left(\boldsymbol{x}_i'\boldsymbol{\beta} + \delta z_i\right)\right]^{1-y_i} g(z_i)\mathrm{d}z_i \right\} \pi(\boldsymbol{\beta}, \delta, \boldsymbol{z}) \end{aligned}$$

where $\pi(\boldsymbol{\beta}, \delta, z)$ is the prior distribution of $(\boldsymbol{\beta}, \delta, z)$.

Finally, to run the Gibbs sampler, we need to sample from the following conditional distributions: $[z_i \mid \boldsymbol{\beta}, \delta, D]$ for $i = 1, 2 \ldots, n$; $[\boldsymbol{\beta} \mid \delta, z, D]$, and $[\delta \mid \boldsymbol{\beta}, z, D]$. More details about the algorithm and implementation used here can be found in Section 5.2.

3. Prior specification

We now examine the problem of eliciting the prior distribution for the regression parameters. The objective is to choose an appropriate joint prior distribution for the random variables $\boldsymbol{\beta}$ and δ. One possibility is to use conditional probabilities assuming dependence between them. On the other hand, Chen et al. (1999) propose two alternative solutions: the use of non-informative or informative prior distributions for $\pi(\boldsymbol{\beta}, \delta)$. In the first case it is required that the matrix $\mathbf{X}_{n\times k}$ has a full rank (see Theorem 1, in Chen et al. (1999)). This condition does not appear in our database, because we have a great many repeated observations, and therefore our research is guided through an informative prior distribution for $(\boldsymbol{\beta}, \delta)$.

Since the coefficients of the logit model do not have a natural interpretation, we have accepted, in a simple way, the following assumptions: (1) $\boldsymbol{\beta}$ and δ are independent, and (2) the odds ratios represented by $\text{odd} = p_i/(1 - p_i) = \exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}$ can be easily obtained. The odds ratios reflect the probability that the analyzed event will happen divided by the probability that it will not happen. Obviously, if the odds are greater than one, then the event is more likely to happen than not. These odds values can be used to elicit the prior normal distribution in the following way. First, we take into account the prior beliefs about the mean and variance of the odd ratio for each covariate. In this sense, we can consider the behavior for these covariates observed in other studies, or take into account the claim adjuster's prior beliefs about them. Assuming that the prior distribution of the coefficients vector $\boldsymbol{\beta}$ is normal, the prior distribution of the odds ratio is log-normal. Second, we can elicit the prior mean and variance by using the following relationship:

$$\begin{aligned} \beta_k \sim \mathcal{N}\left(\beta_k^0, \sigma_k^{-1}\right) \Longleftrightarrow \text{odd}_k &= \exp\{\beta_k\} \\ \sim \quad \mathcal{LN}\left(\text{odd}_k^0, \Sigma_{\text{odd}_k}^{-1}\right), & \end{aligned}$$

where $\mathcal{LN}$ denotes the log-normal distribution, and the two first moments are:

$$\begin{aligned} E[\text{odd}_k] &= \text{odd}_k^0 = \exp(\beta_k^0 + \frac{1}{2}\sigma_k^{-1}), \\ \text{Var}[\text{odd}_k] &= \Sigma_{\text{odd}_k}^{-1} = \exp(2 \cdot \beta_k^0 + \sigma_k^{-1}) \cdot (\exp(\sigma_k^{-1}) - 1). \end{aligned} \tag{3}$$

The claim adjusters or previous studies can help us to obtain some information about the mean and the variance of the odds ratios. Then, we can obtain prior information about the coefficients β by solving the above equations.

Using the odds provides $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta^0}, \sigma_1\mathbf{I})$, where $\mathbf{I}$ represents the identity matrix and $\boldsymbol{\beta^0} = (\beta_1^0, \ldots, \beta_k^0)$. Now, to force an asymmetric skewed model, δ must be estimated to take values not equal to 0. With this objective, we take a symmetric normal distribution with large standard deviation, i.e. $\delta \sim \mathcal{N}(0, \sigma_2)$, with large σ_2. At this point, it is worth mentioning that a large variance will be less informative and will have less impact on the estimation.

4. The database

We now illustrate the application of the specified asymmetric skewed model to a Spanish insurance fraud database. The data correspond to a random sample of 10 000 automobile claims filed in Spain in the year 2000. All claims were investigated by the insurance company and were classified as honest (coded using zeros, 9899 cases), or fraudulent (coded using ones, 101 cases). All claims coded "one" were truly fraudulent because the insured admitted they had cheated (in Spain, legal prosecution of insurance fraud is rare).

The explanatory variables included in the model are defined in Table 1. These fraud indicators have been used previously in other studies (Belhadji et al., 2000; Artís et al., 2002; Caudill et al., 2005; Pinquet et al., 2007) and all of them reflect information obtained from claim reports (only the variable *coverage* is directly related to the policy). The variable *delay* indicates that the claim was not reported to the insurance company within the legal period fixed by Spanish Insurance Law (one week). The claim is more likely to be fraudulent if there is a long delay in contacting the company (Weisberg and Derrig, 1998). The variable *sameco* refers to the fact that all people involved in an accident were insured by the same company. This situation may encourage falsification of the claim and may mean that clients are more familiar with the

Table 1
Explanatory variables used in the model

Fraud (y_i)	Type of claim (fraudulent, 1; legitimate 0).
Delay (x_1)	Claim not reported to the company within the legally established period, 1; otherwise 0.
Sameco (x_2)	Insured and the other driver in the same insurance company, 1; otherwise 0.
Proxim (x_3)	Accident occurred between the policy issue date and the effective starting date, 1; otherwise 0.
Coverage (x_4)	Full coverage, 1; otherwise 0.
Blame (x_5)	The other driver accepts blame for the accident, 1; otherwise 0.

Table 2
Descriptive measures

	Total sample ($N = 10\,000$)		Observed fraudulent ($N = 101$)		Observed honest ($N = 9899$)	
Variable	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Delay	0.300	0.458	0.376	0.487	0.299	0.458
Sameco[a]	0.026	0.159	0.148	0.357	0.025	0.155
Proxim	0.003	0.059	0.020	0.140	0.003	0.058
Coverage[a]	0.498	0.500	0.129	0.336	0.502	0.500
Blame[a]	0.467	0.499	0.802	0.400	0.463	0.499

[a] Indicates variables which have a significant difference of means at 95% confidence level.

claim investigation process. The variable *proxim* may indicate that there is some evidence of planned fraud (Weisberg and Derrig, 1998), for instance, when the policy is contracted after the accident. The variable *coverage* indicates that the claimant has full coverage and may not need to commit fraud in order to recover compensation for damages. Finally, the variable *blame* indicates which driver is responsible for the accident.

Some descriptive measures for the overall sample and for the two subsamples of fraudulent and legitimate claims are shown in Table 2. In this table we indicate with an asterisk which variables have a significant difference of means in the two samples (observed honest and observed fraudulent), at the 95% confidence level.

5. Results

Using the Spanish insurance database described in the previous section, we aim to compare the estimation and fit of various models that try to explain fraud behavior in the automobile insurance market. Firstly, we compare the standard logit model with the corresponding Bayesian logit model. Once we conclude that both models provide similar results in terms of parameter estimates and fit, we compare them (specifically, the classical Bayesian logit model) with the asymmetric or skewed model presented here. We show that the new approximation notably improves the overall fit.

5.1. Standard logit model versus the Bayesian logit model

The results of estimating the standard logit model using maximum likelihood from expression (1) are given in Table 3. The estimations are very similar to those obtained by Artís et al. (2002), and Caudill et al. (2005). All the parameters are significant at least at the 5% level (only the parameter for the variable *delay* is not significant), and the Chi-square statistic shows the general significance for the specified model. The model with all the explanatory variables is significantly better than the restricted model, which has only the constant term.

Table 3
Standard logit estimation results

Variables	Coefficients	Standard deviation	P-value
Constant	−5.051	0.293	0.000[a]
Delay	0.324	0.209	0.121
Sameco	1.717	0.295	0.000[a]
Proxim	1.541	0.749	0.040[b]
Coverage	−1.271	0.340	0.000[a]
Blame	0.903	0.287	0.002[a]

Dependent variable: Fraud; $N = 10\,000$; Chi-square = 99.895 (0.000).
[a] Indicates 1% significance level.
[b] Indicates 5% significance level.

As expected, several characteristics of the claim are positively related to a higher probability of fraud. For instance, cases in which both drivers involved in the accident are insured by the same company, the accident took place between the policy issue date and the policy effective date, and the other driver involved in the claim accepts blame for the accident. Only the parameter for the variable *coverage* has a negative sign but it is also significant. When the insured has full coverage he will probably recover compensation for all claim damages without any difficulty, and so, the probability of committing fraud will be lower.

The logistic regression analysis from the Bayesian point of view, with the assumption of a symmetric link function (see Albert and Chib (1993)) is summarized in Table 4. Following the Bayesian methodology, the posterior distributions of the models have been obtained by using the WinBUGS package, part of the BUGS (Bayesian inference Using Gibbs Sampling) project that provides flexible software for the Bayesian analysis of complex statistical models using Markov Chain Monte Carlo (MCMC) methods.

The Bayesian logistic regression summarized in Table 4 corresponds to the Bayesian skewed logit model presented in Section 2 with $\delta = 0$, by using a non-informative prior distribution for parameters β. Due to this prior distribution and the high number of observations, we see that the coefficients

Table 4
Bayesian logistic regression analysis

Variables	Coefficients	Standard deviation	MC error	Confidence interval (95%)
Constant	−5.097	0.302	0.010	(−5.743, −4.533)
Delay	0.315	0.203	0.003	(−0.099, 0.694)
Sameco	1.681	0.319	0.004	(1.024, 2.268)
Proxim	1.298	0.833	0.008	(−0.518, 2.734)
Coverage	−1.278	0.346	0.007	(−2.038,−0.632)
Blame	0.939	0.301	0.009	(0.386, 1.605)

Dependent variable: Fraud; $N = 10\,000$. MC error column gives an estimate of Monte Carlo standard error (it should be less than 5% to ensure MCMC simulation convergence).

and their significance are very similar to those shown in Table 3, with non-Bayesian inference. Therefore, the same conclusions for the influence of some characteristics of the claim or the policy on the probability of fraud can be reached by the use of this model. There is only one difference, the parameter for the variable *proxim* is no longer significant at 95% level. This means that the Bayesian methodology is unable to signal its significance. This could be the result of the very small number of claims, only thirty five claims, that occurred before the effective starting date.

To assess the quality of fit for the standard logit model and the Bayesian model analyzed we propose two different measures: (1) a confusion matrix containing both information about observed and estimated claims classification in terms of fraud; and (2) a statistical fit measure such as the Akaike information criterion (AIC), or the deviance information criterion (DIC). The latter is a hierarchical modelling generalization of the AIC and is particularly useful in Bayesian model selection problems where the Markov Chain Monte Carlo (MCMC) simulation is used.

The driving idea behind the AIC and DIC is to examine the complexity of the model and the adequacy of its fit to the sample data (obtaining a measure which balances complexity and goodness of fit). For our insurance database, we obtained an AIC of 1041.32 for the standard logit model, and the same DIC value for the Bayesian logit model. This confirms that the two models are also similar with regard to fit.

We obtained the same confusion matrix for the two models analyzed[1] (Table 5). From this matrix we can conclude that the accuracy, i.e. the proportion of fraudulent and non-fraudulent claims correctly classified by the models, is around 0.607. However, while the true positive rate (proportion of observed fraud cases correctly classified as fraudulent by the models) is about 0.852, the true negative rate (proportion of observed legitimate cases correctly classified as honest) is only 0.604, i.e. notably lower. The use of another kind of modelling that improves the fit quality is absolutely justified, taking into account the auditing cost for the percentage of legitimate claims incorrectly classified as fraud by the model. Auditing costs are those related to all the steps of the claim examination, including the first screening of the vehicle carried out by the adjuster, and the special investigation carried out by the SIU's (Special Investigation Units) looking for fraud. The mean auditing cost for this kind of claims (observed honest, predicted fraud) is around 72.26 euros (Viaene et al., 2007, pp. 575, Table 7). So the total cost related to this classification error could be around 283 042 euros in our sample. In conclusion, we found that the lack of fit is due to the incorrect classification of legitimate cases.[2]

Table 5
Confusion matrix for standard logit and Bayesian logit models

		Estimated		
		Legitimate	Fraud	Total
Observed	Legitimate	5982	3917	9 899
	Fraud	15	86	101
	Total	5997	4003	10 000

Given the similarities observed between models, it seems appropriate to apply the Bayesian methodology to the database we have selected. Moreover, as we show in the next section, the Bayesian skewed logit model presented in this paper allows us to improve the classification performance.

5.2. Non-skewed logit models versus the Bayesian skewed logit model

As we have shown, the commonly used links (such as the classical logit link) do not always provide the best fit for a given database. Specifically, in our data the probability of one binary response approaches 0 or 1 at different rates. Therefore, the use of a symmetric link could be inappropriate. One way to solve this problem is to substitute the symmetric link by a skewed or asymmetric link, as Chen et al. (1999) proposed for the Bayesian model that is presented here.

According to the analysis presented in Sections 2 and 3, for a fully Bayesian analysis, we must specify prior distributions for the parameters of interest $(\boldsymbol{\beta}, \delta, z)$. As we discussed in Section 3, we must use an informative prior distribution for $\boldsymbol{\beta}$.

Before the Bayesian study was carried out, the claims adjusters expected a high probability of identifying fraudulent

[1] We have selected a threshold of $c = 0.0101$, according the percentage of fraudulent claims in the total sample. This proportion reflects the observed fraud rate for all claims submitted to the insurance company which make up our database (Pinquet et al., 2007).

[2] Perhaps the inclusion of additional regressors related to the policy (such as the type of vehicle insured), the claim (such as the existence of a limited coverage or not), or the accident (such as the occurrence place) could help us to improve the classification performance. However, some previous studies that have included a higher number of variables have shown that the classification error is still relatively very high (Viaene et al., 2007, Table 9).

Table 6
Bayesian skewed logistic regression analysis

Variables	Coefficients	Standard deviation	MC error	Confidence interval (95%)
Constant	−3.608	1.208	0.023	(−7.039,−2.173)
Delay	0.346	0.224	0.001	(−0.094,0.782)
Sameco	1.853	0.380	0.003	(1.136, 2.638)
Proxim	1.337	0.952	0.005	(−0.685, 3.086)
Coverage	−1.323	0.351	0.002	(−2.031, −0.659)
Blame	0.953	0.304	0.002	(0.379, 1.561)
Delta	−4.652	3.034	0.047	(−12.061, 0.680)

Dependent variable: Fraud; $N = 10\,000$. MC error column gives an estimate of Monte Carlo standard error (it should be less than 5% to ensure MCMC simulation convergence).

claims among those not reported to the insurance company within the legally established period (odd ratio of 1.8). On the other hand, they expected a lower probability of fraudulent claims among those related with full coverage, and in cases when another driver accepted responsibility for the accident (odds ratios of 0.7 and 0.8 respectively). There was no prior information about the rest of variables, taking $\beta^0 = 0$ and $\sigma_1 = 100$ in that case. A small value of 0.01 was assigned to the prior variance for all the odds ratios. Then, after solving the equations in (3), the prior elicitation was implemented by using the following parameter assignments:

$$\boldsymbol{\beta^0} = (0, 0.5862, 0, 0, -0.3667, -0.2308),$$
$$\sigma_1 \mathbf{I} = \text{diag}(100, 0.003, 100, 100, 0.0202, 0.155).$$

A normal distribution with large standard deviation was established for δ, i.e. $\delta \sim \mathcal{N}(0, \sigma_2)$, with $\sigma_2 = 100$. We selected a large value for the δ variance because of the lack of previous information about this skewness parameter (in order to ensure the minimum influence of the prior distribution on the estimation). Finally, the model proposed by Chen et al. (1999) considers the latent variable z distributed as a half-standard normal distribution with mean 0 and variance 1.

With the specified prior distributions and using the WinBUGS package, the posterior distribution of $(\boldsymbol{\beta}, \delta, z)$ was obtained. Table 6 summarizes the main results for the specified skewed logit model. As we see, there are no substantial differences between skewed and non-skewed models regarding the estimation of coefficients $\boldsymbol{\beta}$, except for the constant term, which receives the effect of asymmetry together with the skewness parameter δ.

Note that δ is negative. This result confirms the applicability of a skewed link to our database, and gives evidence that the probability of p_i does not approach 0 at the same rate as it approaches 1. Specifically, as $\delta < 0$, the probability of p_i approaches 0 at a faster rate than it approaches 1. In other words, with $\delta < 0$ the skewed link model increases the probability of $y_i = 0$. Nevertheless, the 95% probability interval includes both positive and negative values, and we cannot claim that it is statistically relevant. However, from the posterior marginal distribution, we find a probability of 91.7% that the parameter δ is negative.

To compare the skewed and the non-skewed Bayesian logit models, we use the same measures as in Section 5.1 to assess the model fit. The WinBUGS package allows us to calculate the DIC measure, which is equal to 850.27 for the skewed link model, i.e, notably lower than that obtained by the non-skewed model (1041.32). This major reduction in the DIC measure indicates a significant increase in the level of fit.

Table 7
Confusion matrix for Bayesian skewed logit model

		Estimated		
		Legitimate	Fraud	Total
Observed	Legitimate	9867	32	9 899
	Fraud	15	86	101
	Total	9882	118	10 000

The confusion matrix presented in Table 7 shows an improvement on the classification results we have obtained using the new skewed model. We see that the accuracy is now about 0.995, which means that the fit with this model is much better than with the symmetric models. The proportion of honest claims correctly classified is around 0.997, in contrast to the 0.604 obtained by the previous models. The proportion of fraud cases that were correctly identified remained unchanged. Obviously, the increase of the probability of $y_i = 0$ induced by the skewed model, since δ was negative, explains these results.

Finally, another way to assess the model's fit, as shown by Guillén (2004), is by using the receiver operating characteristic (ROC) curve, and the area under the ROC curve, called AUROC. In order to visualize the ROC curve for the skewed model, we selected a grid of different thresholds $c \in [0, 0.02]$, and then computed the confusion matrix for each c. In Fig. 1 we plotted the sensitivity in the y-axis versus (1-specificity) in the x-axis, for each different threshold c. Finally, all the points in the plot were linked together starting at (0, 0) up to (1, 1) with a continuous line. The same procedure was repeated for the non-skewed model obtaining the ROC curve shown in Fig. 1 (dotted line).

The AUROC is the zone under the ROC curve and above the bisectrix in Fig. 1. The larger the AUROC, the better the fit of the model; if a model produces a good fit, sensitivity and specificity should be equal to 1, which corresponds to the point (0, 1) where the maximum area is obtained. Therefore, the skewed model produces a better fit than the non-skewed models.

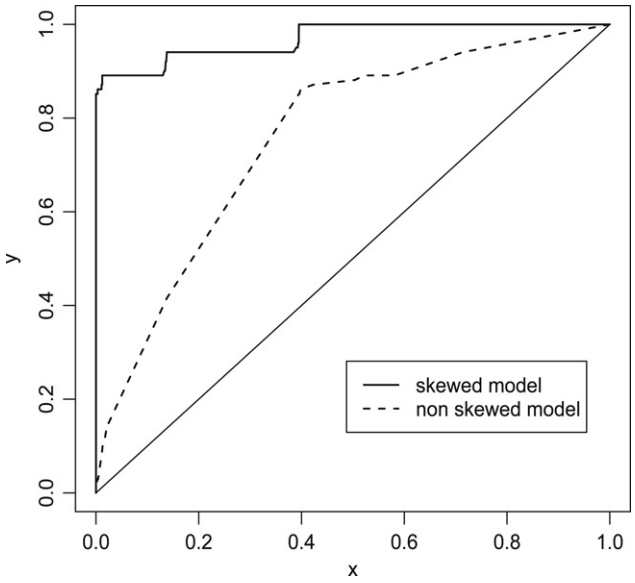

Fig. 1. ROC curve for skewed and non-skewed models.

6. Conclusions

In this paper, we introduced a simulation-based approach by applying a Monte Carlo Bayesian Gibbs sampling for fitting an insurance fraud database using a dichotomous model. Our approach identifies the likelihood of the data by using an asymmetric logit model, and then assuming a proper prior distribution of the parameters of the model. These considerations, combined with the Gibbs sampler, allow us to simulate on the basis of the exact posterior distribution of these parameters.

Comparing the standard and the Bayesian logit (with symmetric links) estimation results, we see that the Bayes logistic model gives posterior estimations for the coefficients that are quite similar to the classical ones. A logistic model with classical inference gives almost the same estimations as a Bayesian inference with non-informative normal priors. However, the Bayesian methodology is unable to signal the significance of the parameter associated to the variable *proxim*, as we have commented in Section 5.1. Moreover, both models present a lack of fit due to the incorrect classification of zero cases (see Table 5).

A large difference can be observed in the confusion matrix (see Table 7) when we consider a Bayesian skewed logit model, which is more suitable for fitting data than a classical logistic model if the zero response is observed more often than the one response. Still a non-significant parameter for *proxim* has been found. Finally, the Bayesian procedure presents standard errors for the parameters slightly higher than those obtained with the classical logit.

These conclusions are analyzed in the context of insurance fraud, where poor classification of honest and fraudulent claims implies high auditing costs for the insurance company. The auditing cost related to honest claims incorrectly classified as fraud will be notably reduced with the application of a skewed logit model. Since the Bayesian skewed logit model presented here is only used for fitting purposes, it is necessary to search for an asymmetric link function which would model the insurance fraud database so that the best predictive model would be obtained. In this way, we have already used the most common asymmetric link functions (log–log and complementary log–log), but we have not found a relevant improvement in terms of predicting or classification performance. Therefore, a natural extension of this paper is looking for skewed link functions which help us to obtain better predictions.

Another interesting extension of the study is the application of our results to a multinomial logit model, where multiple responses are possible (Holmes and Held, 2006; Bazán et al., 2006). The multinomial logit model has been previously used in modelling automobile insurance fraud behavior (Artís et al., 1999), but not from a Bayesian point of view.

Acknowledgements

We thank the anonymous referee for his/her useful comments. The authors are grateful for the valuable suggestions from the participants in the 10th International Congress on Insurance: Mathematics and Economics in Leuven during July 18–20, 2006. This study has received support from the Spanish Ministry of Education and Science and FEDER grants.

References

Albert, H., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. Journal of American Statistical Association 88 (422), 669–679.

Amemiya, T., 1981. Qualitative response models: A survey. Journal of Economic Literature 19, 1483–1536.

Artís, M., Ayuso, M., Guillén, M., 1999. Modelling different types of automobile insurance fraud behavior in the Spanish market. Insurance: Mathematics & Economics 24, 67–81.

Artís, M., Ayuso, M., Guillén, M., 2002. Detection of automobile insurance fraud with discrete choice models and misclassified claims. Journal of Risk and Insurance 69 (3), 325–340.

Basu, S., Mukhopadhayay, S., 2000. Bayesian analysis of binary regression using symmetric and asymmetric links. Sankhyā 3, 372–387.

Bazán, J., Branco, M., Bolfarine, H., 2006. A skew item response model. Bayesian Analysis 1 (4), 861–892.

Belhadji, E.B., Dionne, G., 1997. Development of an expert system for the automatic detection of automobile insurance fraud. Working Paper 97-06. École des Hautes Études Commerciales. Université de Montréal.

Belhadji, E.B., Dionne, G., Tarkani, F., 2000. A model for the detection of fraud. Geneva Papers on Risk and Insurance — Issues and Practice 25 (5), 517–538.

Carlin, B., Polson, N., 1992. Monte Carlo Bayesian methods for discrete regression models and categorical time series. Bayesian Statistics 4, 577–586.

Caudill, S., Ayuso, M., Guillén, M., 2005. Fraud detection using a multinomial logit model with missing information. Journal of Risk and Insurance 72 (4), 539–550.

Chen, M., Dey, D., Shao, Q., 1999. A new skewed link model for dichotomous quantal response data. Journal of the American Statistical Association 94 (448), 1172–1186.

Guillén, M., 2004. Fraud in Insurance. In: Encyclopedia of Actuarial Science, vol. 2. John Wiley and Sons, Chichester, pp. 729–739.

Hausman, J., McFadden, D., 1984. Specification tests for the multinomial logit model. Econometrica 52 (5), 1219–1240.

Holmes, L., Held, L., 2006. Bayesian auxiliary variables models for binary and multinomial regression. Bayesian Analysis 1 (1), 145–168.

Koop, G., Poirier, D., 1993. Bayesian analysis of logit models using natural conjugate priors. Journal of Econometrics 56 (3), 323–340.

McCulloch, R., Polson, N., Rossi, P., 1999. A Bayesian analysis of the multinomial probit model with fully identified parameters. Journal of Econometrics 99 (1), 173–193.

McFadden, D., 1974. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York, pp. 105–142.

McFadden, D., 1981. Econometric models of probabilistic choice. In: Manski, C., McFadden, D. (Eds.), Structural Analysis of Discrete Data. MIT Press, Cambridge, pp. 198–272.

Pinquet, J., Ayuso, M., Guillén, M., 2007. Selection bias and auditing policies for insurance claims. Journal of Risk and Insurance 74 (2), 425–440.

Stukel, T., 1988. Generalized logistic model. Journal of the American Statistical Association 83, 426–431.

Viaene, S., Derrig, R., Baesens, B., Dedene, G., 2002. A comparison of State-of-the-Art classification techniques for expert automobile insurance claim fraud detection. Journal of Risk and Insurance 69 (3), 373–421.

Viaene, S., Ayuso, M., Guillén, M., Van Gheel, D., Dedene, G., 2007. Strategies for detecting fraudulent claims in the automobile insurance industry. European Journal of Operational Research 176, 565–583.

Weisberg, H.I., Derrig, R.A., 1998. Detection de la fraude: Methodes quantitatives. Risques 35, 75–99 (in English translation).