A Bayesian Partition Model for Customer Attrition

C. J. HOGGART and J. E. GRIFFIN
Knowledge Lab, NCR

Abstract: This paper presents a nonlinear Bayesian model for covariates in a survival model with a surviving fraction. The work is a direct extension of the cure rate model of Chen *et al.* (1999). In their model the covariates depend naturally on the cure rate through a generalised linear model. We use a more flexible local model of the covariates utilizing the Bayesian partition model of Holmes *et al.* (1999). We apply the model to a large retail banking data set and compare our results with the generalised linear model used by Chen *et al.* (1999).

Keywords: SURVIVAL ANALYSIS, CURE RATE MODEL, PARTITION MODEL, DATA-MINING.

1. INTRODUCTION

A common problem faced by banks is customer attrition. By this we simply mean customers leaving the bank. We are interested in the case where customer attrition occurs after a particular event, for example, customers may leave a bank after paying off a loan. If we can predict who will leave the bank and when, then action can be taken to prevent customers leaving.

This problem has direct parallels with survival analysis. Rather than predicting the time to death of a patient after observing the patient has a disease we are interested in the time to a customer leaving the bank after an event. There is a vast literature on survival models, however typical models assume that all patients will eventually die from the disease. This is not appropriate for our problem as here we believe that a proportion of the customers will not leave the bank due to the event of interest. In medicine this is equivalent to a proportion of the patients being cured.

A popular cure rate model is the mixture model introduced by Berkson and Gage (1952). Their model assumes a proportion π of the population is cured and model the non-cured fraction with survivor function $S^{\star}(t)$. This leads to the following survivor function for the population

$$S(t) = \pi + (1 - \pi)S^{\star}(t).$$

This model has been extensively discussed in the statistical literature, however it has several drawbacks which are discussed in Chen *et al.* (1999). They have developed an alternative

Bayesian cure rate model which in contrast is computationally attractive, has an intuitive interpretation and has a proportional hazards structure with covariates.

We extend the work of Chen *et al.* (1999) as follows. The model described in their paper models the cure rate through a generalised linear model (GLM) of the covariates. For extra flexibility we model the covariate effect locally using the Bayesian partition model (BPM) of Holmes *et al.* (1999). The extension to local modelling maintains the proportional hazards structure and also results in a computationally faster algorithm.

In section 2 we describe the cure rate mode of Chen *et al.* (1999), state some of its basic properties and outline how covariates can be incorporated using a GLM. In section 3 we outline the BPM and describe how to sample from it. Section 4 describes our extension to local modelling of the covariates and how posterior inference can be made using MCMC. In section 5 we analyse data supplied by a bank using a cure rate model where the covariate effect is modelled using a GLM and a BPM. Section 6 contains a brief discussion of the work described in this paper.

2. THE CURE RATE MODEL

In this section we describe the cure rate model of Chen *et al.* (1999). Their model was applied to a cancer trial and is built around modelling an unknown number of cancerous cells. If a patient has no cancerous cells they are cured and as the number of cancerous cells increases the risk increases. More generally the model can be viewed as a latent variable model which maintains the proportional hazards structure. In the banking context, the latent variables capture heterogeneity in the risk of leaving the bank across the population. We refer to the latent variables as risks.

We now describe the model. The number of unknown risks, denoted by N, is modelled as a Poisson distribution with mean θ. The time to failure due to risk i is denoted by Z_i. The model assumes the random variables $Z_i,\ i = 1, 2, \ldots$ are iid with a common distribution function $F(t) = 1 - S(t)$ and are independent of N. With N risks the probability of survival to time t is $P(Z_1 > t, \ldots, Z_N > t)$. Since the Z_i's are independent $P(Z_1 > t, \ldots, Z_N > t) = S(t)^N$. The survivor function is given by the expectation of $S(t)^N$ with respect to N,

$$\begin{aligned} S_p(t) &= \mathrm{E}_N\left(S(t)^N\right) \\ &= \sum_{k=0}^{\infty} S(t)^k \frac{\theta^k}{k!} \exp\{-\theta\} \\ &= \exp\{-\theta + \theta S(t)\} = \exp\{-\theta F(t)\}. \end{aligned} \tag{1}$$

The cure or remaining fraction of (1) is given by

$$S_p(\infty) = P(N = 0) = \exp\{-\theta\}.$$

It follows that as the mean number of risks θ increases the remaining fraction decreases and tends to 0 as $\theta \to \infty$. The density corresponding to (1) is given by

$$f_p(t) = \theta f(t) \exp\{-\theta F(t)\}.$$

We note that $f_p(t)$ is not a proper density since it does not integrate to 1, similarly $S_p(t)$ is not a proper survivor function. The hazard function is given by

$$h_p(t) = \theta f(t).$$

Chen *et al.* (1999) model θ, the parameter of the Poisson distribution, with a GLM

$$\theta = \exp\{\boldsymbol{X}'\boldsymbol{\beta}\}$$

where $\boldsymbol{\beta}$ is a vector of $p \times 1$ regression coefficients including an intercept and $\boldsymbol{X}$ is a vector of covariates. Chen *et al.* (1999) choose to model $f(t)$ independently of the covariates. With this assumption it is clear that $h_p(t)$ has a proportional hazards structure.

Chen *et al.* (1999) show that a uniform improper prior on $\boldsymbol{\beta}$ gives a proper posterior distribution. The posterior distribution of the unknown parameters can be sampled from in a Gibbs sampler. Sampling from the full conditional distributions of α and $\beta_j,\ j = 1, \ldots, p$ is not standard, however the densities are log-concave and thus the adaptive rejection sampler of Gilks and Wild (1992) can be used.

3. THE BAYESIAN PARTITION MODEL

The Bayesian partition model (BPM) (Holmes *et al.* , 1999) is a generic approach to classification and regression problems. The input space is divided into disjoint regions defined by a tessellation structure T. For example the authors use Voronoi tessellations. Within each region the observations are assumed to come from a "simple", conjugate model, for example, the multinomial model with a Dirichlet prior for classification problems. The parameters are assumed to be independent between each region. The marginal likelihood for the tessellation structure is then the product of the marginal likelihoods within each region and since the model is conjugate these are available analytically. The predictive surface for any given tessellation structure will be disjoint at the tessellation boundaries but is smoothed by mixing over the posterior distribution of tessellation structures.

In data-mining problems we prefer to divide the input space with hyperplanes parallel to the axes which take the form $x_i = \alpha$ for some dimension i of the input space and split point α, and use local models whose parameters do not depend on the covariates. This results in the following computationally attractive properties. Firstly, calculation of the marginal likelihood does not involve expensive operations such as matrix inversion, which would be necessary for a local linear regression model. Secondly, since the local model does not depend on covariates, the hyperplanes will effectively perform variable selection by only splitting on those variables for which there is a significant effect on the output variable. Thirdly, allocation to clusters is much quicker than for the Voronoi tessellation. For k hyperplanes a point's allocation can be calculated using k operations. In comparison the Voronoi tessellation involves pk calculations where p is the number of covariates.

The model is fitted using a Metropolis-Hastings sampler directly on the distribution $p(T \mid D)$, where D is the data. This is possible because the conjugate models allow the marginal likelihood of the data given the partition structure, $p(D \mid T)$, to be calculated analytically. A hybrid Metropolis-Hastings sampler with various dimension-jumping moves is used to explore the posterior distribution $p(T \mid D) \propto p(D \mid T)p(T)$.

In practice the prior for the hyperplanes is independent between dimensions and within a dimension has probability $1/n$ at each observed covariate, where n is the number of

observations. This prior ensures that the hyperplanes fall within the convex hull of the data.

3.1. *Computational strategy*

The hyperplane BPM using the prior described in the previous section can be implemented using a hybrid Metropolis-Hastings sampler with three possible moves:

- A new partition can be added to the model by randomly choosing a dimension and then proposing a split point from the prior distribution, i.e. the empirical distribution of that dimension.
- A partition can be removed at random from the model.
- A partition can be moved by redrawing its split point from the prior distribution, i.e. the empirical distribution of that dimension.

A proposed new tessellation structure T' is accepted with probability

$$\alpha(T', T) = \min\left\{1, \frac{p(D \mid T')p(T')}{p(D \mid T)p(T)}\right\}.$$

The number of partition is assigned a geometric prior to encourage models with fewer partitions.

4. LOCAL EXTENSION TO THE CURE RATE MODEL

In this section we describe the extension to local modelling of the covariate effect using the BPM. The orthogonal hyperplane tessellation T defines m disjoint regions $R_1, \ldots, R_m$, which each have a Poisson model for N with constant parameter values $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$. The conjugate prior for the Poisson model is the gamma distribution. There are $n_1, \ldots, n_m$ observations in each region and n observations in total, $n = n_1 + \cdots + n_m$. Again the covariate effect is modelled through θ alone maintaining the proportional hazards structure.

We now define the other notation required to describe the model. We denote the failure time of the ith customer in R_j by t_{ij}, where t_{ij} may be right censored. We define the indicator variable δ_{ij} which is 1 if t_{ij} is uncensored and 0 if t_{ij} is censored. The number of risks for the ij-th customer is denoted by N_{ij}, $j = 1, \ldots, m$, $i = 1, \ldots, n_j$. For brevity of notation we take the vectors $\boldsymbol{t}$ and $\boldsymbol{\delta}$ to denote the set of all observations and the matrix $\boldsymbol{X}$ to denote all of the covariates.

Each risk is modelled as a Weibull distribution defined as

$$\mathrm{We}(t \mid \alpha, \lambda) = \lambda\alpha t^{\alpha-1} \exp\left\{-\lambda t^{\alpha-1}\right\}, \quad t > 0,\ \alpha > 0,\ \lambda > 0.$$

If f and S are the density and survivor functions respectively of the Weibull distribution and $\mathrm{Pn}(\cdot \mid \theta)$ is the probability mass function of the Poisson distribution with mean θ the

full model can be written as

$$
\begin{aligned}
p(\boldsymbol{t}, \boldsymbol{\delta} \mid \boldsymbol{N}, \alpha, \lambda, T) &= \prod_{j=1}^{m} \prod_{i=1}^{n_j} S(t_{ij} \mid \alpha, \lambda)^{N_{ij} - \delta_{ij}} \left(N_{ij} f(t_{ij} \mid \alpha, \lambda)\right)^{\delta_{ij}} \\
&= \prod_{j=1}^{m} \exp\left\{-\lambda \sum_{i=1}^{n_j} N_{ij} t_{ij}^{\alpha_j - 1}\right\} \prod_{i=1}^{n_j} \left(N_{ij} \lambda \alpha t_{ij}^{\alpha_j - 1}\right)^{\delta_{ij}} \\
p(\boldsymbol{N} \mid \boldsymbol{\theta}, T) &= \prod_{j=1}^{m} \prod_{i=1}^{n_j} \mathrm{Pn}(N_{ij} \mid \theta_j) \\
p(\boldsymbol{\theta} \mid T) &= \prod_{j=1}^{m} \mathrm{Ga}(\theta_j \mid \vartheta_0, \vartheta_1).
\end{aligned}
$$

We assign T a geometric prior distribution $p(T) = \mathrm{Ge}(\psi)$ and use default prior distributions for the parameters of the Weibull distribution: $p(\alpha) = \mathrm{Ga}(\alpha_0, \alpha_1)$ and $p(\lambda) = \mathrm{Ga}(\lambda_0, \lambda_1)$. To summarise the hyperparameters of the model are $\alpha_0, \alpha_1, \lambda_0, \lambda_1, \vartheta_0, \vartheta_1, \psi$ and r which we denote by ϕ.

4.1. *Computational strategy*

The full conditional distributions of α and λ required by the Gibbs sampler to draw from the posterior distribution of the model parameters are identical to those for the GLM and are

$$
\begin{aligned}
p(\lambda \mid \alpha, \boldsymbol{N}, T, \phi, \boldsymbol{t}, \boldsymbol{\delta}) &= \mathrm{Ga}\left(\lambda \,\middle|\, \lambda_0 + \sum_{i=1}^{n} \delta_i, \lambda_1 + \sum_{i=1}^{n} N_i t_i^{\alpha}\right) \\
p(\alpha \mid \lambda, \boldsymbol{N}, T, \phi, \boldsymbol{t}, \boldsymbol{\delta}) &\propto \alpha^{\sum_{i=1}^{n} \delta_i} \left(\prod_{i=1}^{n} t_i^{\delta_i}\right)^{\alpha} \exp\left\{-\lambda \sum_{i=1}^{n} N_i t_i^{\alpha}\right\} p(\alpha).
\end{aligned}
$$

The other full conditionals are

$$
\begin{aligned}
p(N_{ij} \mid\ \alpha, \lambda, \boldsymbol{\theta}, T, \phi, \boldsymbol{t}, \boldsymbol{\delta}) &= \begin{cases} \mathrm{Pn}\left(N_{ij} \mid \theta_j \exp\{-\lambda t_{ij}^{\alpha}\}\right), & \text{if } \delta_{ij} = 0 \\ \mathrm{Pn}\left(N_{ij} - 1 \mid \theta_j \exp\{-\lambda t_{ij}^{\alpha}\}\right), & \text{if } \delta_{ij} = 1 \end{cases} \\
p(\boldsymbol{\theta}, T \mid \alpha, \lambda, T, \phi, \boldsymbol{t}, \boldsymbol{\delta}) &= p(T \mid \boldsymbol{N}, \phi) \prod_{j=1}^{m} p(\theta_j \mid T, \boldsymbol{N}, \phi)
\end{aligned}
$$

where

$$
\begin{aligned}
p(\theta_j \mid T, \boldsymbol{N}, \phi) &= \mathrm{Ga}\left(\theta_j \,\middle|\, \vartheta_0 + n_j, \vartheta_1 + \sum_{i=1}^{n_j} N_{ij}\right) \\
p(T \mid \boldsymbol{N}, \phi, \boldsymbol{X}) &\propto p(N_1, \ldots, N_n \mid T) p(T) \\
&= p(T) \prod_{j=1}^{m} p(N_{1j}, \ldots, N_{n_j j} \mid T).
\end{aligned}
$$

The joint density $p(N_{1j}, \ldots, N_{n_j j} \mid T)$ is the marginal likelihood of latent variables in the jth partition. This is straightforward to evaluate

$$\begin{aligned} p(N_1, \ldots, N_n \mid \vartheta_0, \vartheta_1) &= \int \prod_{i=1}^{n} p(N_i \mid \theta) p(\theta \mid \vartheta_0, \vartheta_1)\, d\theta \\ &= \frac{1}{\prod_{i=1}^{n} N_i!} \frac{\vartheta_1^{\vartheta_0}}{\Gamma(\vartheta_0)} \frac{\Gamma(\vartheta_0 + \sum_{i=1}^{n} N_i)}{(\vartheta_1 + n)^{\vartheta_0 + \sum_{i=1}^{n} N_i}}. \end{aligned}$$

Given this marginal likelihood the tessellation structure can be sampled from using a Metropolis-Hastings algorithm (Hastings, 1970) within the Gibbs sampler.

5. EXAMPLE

One of our banking partners collected data on customers who paid off their mortgage between January 1st and December 31st. The bank had noticed that these customers were more likely to leave the bank and wanted to learn which of these customers left and when.

The results presented in this paper were generated from a subset of the customer covariates collected by the bank. For each customer we had the following information

AGE:	age of customer
GENDER:	gender of customer
QUIT:	indicates whether or not the customer has left the bank
END_DATE:	date of end of the customer relationship
M_NUMB:	number of loans closed in year
CARDS:	number of debit cards.

For each loan closed we had the following information

F_AMOUNT:	how much was paid when the account was closed
TYPE:	type of mortgage
ORIGCLOS:	planned date of closure of the loan
DATE_CLO:	date when loan was actually closed $\leq$ ORIGCLOS
S_AMOUNT:	initial size of mortgage.

Thus the target is the number of days between DATE_CLO and END_DATE, those customers who did not leave the bank were censored on December 31st. The variables ORIGCLOS and DATE_CLO were combined to give the number of days the customer pre-empted the planned date for closing the loan.

We analysed the data using both the GLM and the BPM to model the covariates and compared our results. The priors for α and λ were $\text{Ga}(\alpha_0 = 0.1, \alpha_0 = 0.1)$ and $\text{Ga}(\lambda_0 = 0.1, \lambda_0 = 0.1)$ respectively in both the BPM and GLM. Denison and Holmes (2001) shrink the partition specific parameters θ_i to the empirical mean. However, we found that for our data set crossvalidation scores were robust to a range of priors for θ. Within this range the number of partitions increased as the prior variance of θ decreased. We took a vague $\text{Ga}(\vartheta_0 = 1, \vartheta_1 = 1)$ prior. A geometric distribution with mean 10 was taken for $p(T)$. An improper, flat prior for $\boldsymbol{\beta}$ was taken in the GLM. The data was standardized for model interpretation and to stabilize the posterior computation.

Figs. 1 and 2 show Kaplan-Meier, BPM and GLM survivor functions of customers in two cases: Case 1, with one debit card, one closed account, a final payment of 0 and

mortgage type 0; Case 2, customers with no debit cards, one closed account, a final payment of 0 and mortgage type 1. It is clear that BPM models the data (the Kaplan-Meier estimate) more closely than the GLM.

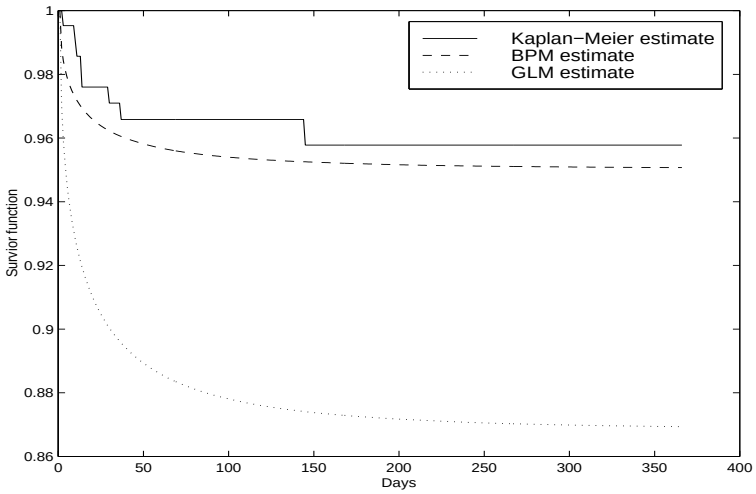

Figure 1. *Survivor functions for Case 1.*

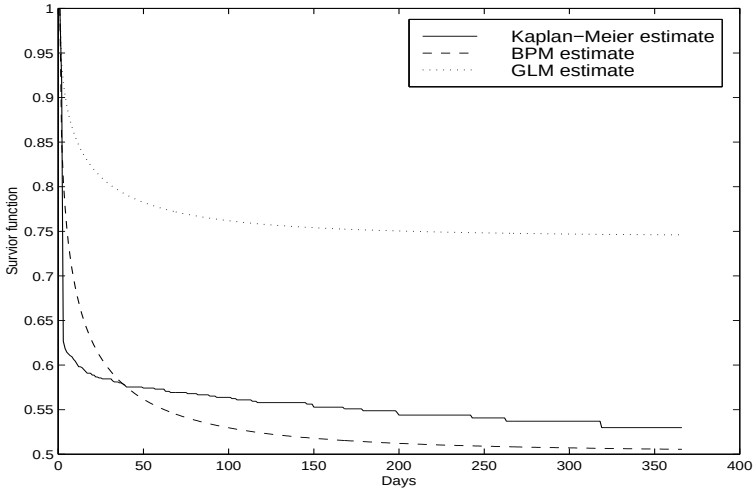

Figure 2. *Survivor functions for Case 2.*

The BPM compared favourably with the GLM in a two way crossvalidation. The data set of 41,979 customers was split into a set of 21,000 customers and a set of 20,979 customers, to perform the crossvalidation. For each crossvalidation set we compared the posterior predictive of one set of observations conditional on the other set under each model. This was evaluated using the following Rao-Blackwellised (Gelfand and Smith, 1990) es-

timate

$$p(\boldsymbol{t}', \boldsymbol{\delta}' \mid \boldsymbol{t}, \boldsymbol{\delta}) = \frac{1}{M} \sum_{j=1}^{M} p(\boldsymbol{t}', \boldsymbol{\delta}' \mid \alpha^{(j)}, \lambda^{(j)}, \boldsymbol{\theta}^{(j)})$$

where

$$p(\boldsymbol{t}', \boldsymbol{\delta}' \mid \alpha, \lambda, \boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{k=0}^{\infty} p(t_i', \delta_i' \mid \alpha, \lambda, N_i = k) p(N_i = k \mid \theta_i)$$

where $(\boldsymbol{t}', \boldsymbol{\delta}')$ is the test data and M is the number of iterations. The results are given in Table 1.

Table 1. *Log posterior predictives under the BPM and GLM.*

	BPM	GLM
Set 1	-4753.27	-5050.10
Set 2	-4582.70	-4869.00

The probability of splitting on the eight customer covariates in the Bayesian orthogonal hyperplane partition model are shown in Table 2. The model rarely splits on AGE, S_AMOUNT, and 'pre-empt' time indicating that these covariates are not informative.

Table 2. *Probability of partitioning on each variable.*

AGE	GENDER	CARDS	M_NUMB
0.00	0.35	1.00	1.00
S_AMOUNT	F_AMOUNT	'pre-empt time'	TYPE
0.00	1.00	0.00	1.00

Table 3 shows the predicted cure rate for 3 different values of customer covariates under the BPM and GLM and the empirical cure rate. The empirical cure rate was naively calculated as the proportion of customers still with the bank. The cure rate is overestimated due to censoring. Three levels of final payment were looked at, \$0, \$1,000 and \$50,000 (the empirical cure rates were calculated by taking observations around \$1,000 and \$50,000). We can see from Table 3 that the cure rate has a nonlinear relationship with F_AMOUNT and depends on whether or not F_AMOUNT is 0. The BPM is well suited to automatically detecting this type of relationship, which the GLM struggles to model.

Table 3. *Estimated cure rates from the GLM and BPM models.*

CARDS	M_NUMB	F_AMOUNT × $1,000	TYPE	GLM cure rate	BPM cure rate	Empirical cure rate
0	1	0	1	0.7445	0.5027	0.5514
0	1	1	1	0.7508	0.9174	0.9288
0	1	50	1	0.9368	0.9174	0.9052

Tables 4 and 5 show posterior estimates of the Weibull parameters of the BPM and the parameters of the GLM respectively. All of the regression coefficients of the GLM except β_{GENDER} have significant posterior mass away from 0. Since the data is normalized we can compare the relative sizes of these coefficients. The effects of CARDS, M_NUMB, F_AMOUNT and TYPE were picked up by the BPM, however the effects of AGE, S_AMOUNT and 'pre-empt time' were not. This was probably because these effects are relatively small. The standard deviations of the parameters of the Weibull distribution are smaller for the BPM than the GLM due to the improved fit of the BPM.

Table 4. *Posterior estimates for the Weibull parameters of the BPM.*

Parameter	Posterior mean	Posterior SD	95% HPD interval
α	0.4968	0.0103	$(0.4774,\ 0.5171)$
λ	0.2628	0.0111	$(0.2442,\ 0.2888)$

6. DISCUSSION

We have shown that a nonparametric extension to the model of Chen *et al.* (1999) can lead to better predictive performance in banking problems whilst retaining the proportional hazards structure. An attractive feature of the orthogonal hyperplane BPM is the natural incorporation of covariate selection. Each hyperplane splits the database on only one covariate and so hyperplanes are only included when the covariate affects the model fit. This results in the orthogonal hyperplane version of the BPM being computationally attractive when only a small proportion of the covariates are useful for prediction. This is often the case in banking problems where databases are very large.

Computationally the orthogonal hyperplane BPM scales as $O(nm)$, where n is the number of customers and m is the number of partitions. In contrast the GLM scales as $O(np)$, where p is the number of covariates. In our experience the number of partitions is relatively small and the orthogonal hyperplane BPM is faster. Sampling for $\boldsymbol{\beta}$ using the adaptive rejection sampler has been replaced by sampling for the tessellation structure and the θ parameters for each partition, both of which are straightforward.

Table 5. *Posterior estimates for the GLM parameters.*

Parameter	Posterior mean	Posterior SD	95% HPD interval
α	0.5160	0.0156	$(0.4859,\ 0.5464)$
λ	0.2455	0.0149	$(0.2184,\ 0.2754)$
β_0	-4.5270	0.0751	$(-4.6743,\ -4.3839)$
β_{AGE}	-0.1148	0.0340	$(-0.1854,\ -0.0515)$
β_{GENDER}	0.0538	0.0380	$(-0.0180,\ 0.1290)$
β_{CARDS}	-1.6706	0.0606	$(-1.7920,\ -1.5503)$
β_{M_NUMB}	-0.1608	0.0527	$(-0.2631,\ -0.0551)$
β_{S_AMOUNT}	-0.2395	0.0380	$(-0.3129,\ -0.1615)$
β_{F_AMOUNT}	-1.2365	0.0748	$(-1.3820,\ -1.0921)$
$\beta_{pre-empt}$	0.0937	0.0149	$(0.0604,\ 0.1191)$
β_{TYPE}	-0.2240	0.0233	$(-0.2688,\ -0.1779)$

REFERENCES

Berkson, J. and Gage, R. P., (1952). Survival curve for cancer patients following treatment, *J. Amer. Statist. Assoc.* **47**, 501–515.

Chen, M. H., Ibrahim, J. G. and Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction, *J. Amer. Statist. Assoc.* **94**, 909–919.

Denison, D. G. T. and Holmes, C. C. (2001). Bayesian partitioning for estimating disease risk, *Biometrics* **57**. To appear.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398–409.

Gilks, W. F. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling, *Appl. Statist.* **41**, 337–348.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**, 97–109.

Holmes, C. C., Denison, D. G. T. and Mallick, B. K. (1999). Bayesian partitioning for classification and regression, *Tech. Rep.*, Imperial College, UK.