International Journal of Management Science
Vol 12, No 1, May 2006

Assessment of Effects of Predictors on the Corporate Bankruptcy Using Hierarchical Bayesian Dynamic Model

Minje Sung*
School of Business Administration, Ajou University
San 5 Wonchun-dong, Suwon, Korea

Sungbin Cho**
School of Business, Sogang University
1 Sjinsu-dong, Seoul, Korea

(Received Dec. 2005; Revised Feb. 2006; Accepted Mar. 2006)

ABSTRACT

This study proposes a Bayesian dynamic model in a hierarchical way to assess the time-varying effect of risk factors on the likelihood of corporate bankruptcy. For the longitudinal data, we aim to describe dynamically evolving effects of covariates more articulately compared to the Generalized Estimating Equation approach. In the analysis, it is shown that the proposed model outperforms in terms of sensitivity and specificity. Besides, the usefulness of this study can be found from the flexibility in describing the dependence structure among time specific parameters and suitability for assessing the time effect of risk factors.

Keywords: Hierarchical Model, Bayesian Dynamic Model, Bankruptcy Data

1. INTRODUCTION

A number of studies have been conducted in the literature to estimate the risk of corporate bankruptcy and identify the explanatory factors. The modeling strategies in the literature have been either cross-sectional with one time point data [1, 2, 7, 12] or static estimation of the effects of independent variables on the binary bankruptcy indicator variable with longitudinal repeated measured data [6]. When the longitudinal data are available where measurements were collected

* Corresponding author, Email: sungmj@ajou.ac.kr
** Email: sungbincho@sogang.ac.kr

over a span of several years, a static analysis technique such as Generalized Estimating Equation (GEE) [13] approach for binary outcome variables has been most widely used to address the issues of interest. Although not completely wrong this static approach is irrelevant in the sense that it ignores the dynamically evolving effects of predictors on the risk of bankruptcy.

One of the main interests was to find out how the risk of bankruptcy changes over time as a function of various predictors, thus many researchers have conducted a longitudinal study up to a certain point in time. Here, the challenging problem is that the bankruptcy is once-a-lifetime event that can only be observed at the time of death of a company, while the predictors are time dependent and available up to the time of bankruptcy or the time of censoring, which occurs when the data collection stops. A GEE model is appropriate for the case where both response and predictor variables are time dependent because the goal of GEE model is to evaluate the marginal effects of time-varying predictors on the response variable. Basically, GEE approach does not allow researchers to assess how the risk of bankruptcy changes over time, in a dynamic sense, because the effect of an independent variable on the bankruptcy indicator variable may be different from one time point to the other. Such dynamic effect of predictors on the risk of bankruptcy at one time point may be related in some way to the effect at the other time point. The purpose of current study is two folds: First, we employ Bayesian hierarchical modeling approach to assess how the effect of corporate financial variables on the risk of bankruptcy changes over time. Second, we introduce a novel dynamic statistical modeling strategy using a Bayesian framework, which has been popularly used in area of engineering and biostatistics but not much in management science society.

2. STATISTICAL MODEL

Corporate finance data (N=1000) from year 1999 to 2004 were collected for data analysis. Three consecutive years of financial variables were used in the analysis. For instance, the data consist of the variables in years 2001, 2002, and 2003 if the corresponding company went into bankruptcy in 2003. The data were obtained from the Korea Credit Guarantee Fund. Among 54 corporate finance-related variables collected, 27 variables were selected by the t-test with the alpha of 0.05 as a preliminary screening step. As a next step, stepwise logistic regression finally identified nine variables where the entry significance level of 0.05, the staying significance level of 0.10, the logit link function, and the Schwartz Bayesian Criterion are used. The nine variables include gross value added to total assets, increase in assets, net return on assets, net working capital to total assets, current liabilities to total assets, depreciation expense, interest expense to total expense, net interest expenses, and wage and salary expenses.

For repeated measures like our data where the response variable is binary, the measurements within a particular company are, in general, correlated with each other. In order to draw inferences correctly about the effects of independent variables on the binary response variable, the within-company correlation needs to be accounted for. Among various statistical techniques, GEE [13] provides the most popular set of methods for estimated average or marginal effects of independent variables. One of the limited capability of the GEE approach is that the GEE does not maximize the advantage of having longitudinal data. In other words, the most benefit of having the repeated measures collected over a span of several years with all costs is that longitudinal data enables us to investigate how the outcome variable evolves over time and how the time-varying independent variables dynamically affect the outcome variable. For the sake of simplicity in modeling and solvability, the GEE approach assumes that the effect of independent variable on the outcome variable would be constant, and thereby it fails to show the time-evolving structure of genuine effect of independent variables.

Thus, a more appropriate strategy is needed to maximize the benefit of longitudinal study as well as to assess the dynamic nature of relationships among variables. To meet this challenge, we consider Bayesian Dynamic Generalized Linear Models (DGLM) [11]. The key idea of dynamic models is to relate the model parameters to changes in the developing risk of bankruptcy, due to the passage of time. In the Bayesian paradigm, all uncertainties about the unknown quantities are described by probabilities and inferences about them are made with posterior probabilities, which will be assessed using Markov Chain Monte Carlo (MCMC) simulation approach due to the computational difficulties in the present study. A good reference of MCMC application would be the study by Sung et al. [10].

DGLM structure employs the logistic regression to assess the effect of corporate financial variables on the risk of bankruptcy at time T for the ith company. Let us define Y_{iT} as a binary variable representing bankruptcy of company i at time T. Thus Y_{iT} is a Bernoulli random variable with probability π_{iT} and logit transformation is made on π_{iT}. The Bayesian hierarchical modeling framework is as follows.

First Level:

$$(Y_{iT} \mid \pi_{iT}) \sim Bernoulli(\pi_{iT}), \tag{1}$$

for $i = 1,\dots, M$.

Second Level:

$$\log(\frac{\pi_{iT}}{1-\pi_{iT}}) = \log\text{it}(\pi_{iT}) = \beta_{0t} + \sum_{q=1}^{Q} \beta_{qt} X_{qit}, \tag{2}$$

where π_{it} represents the probability of bankruptcy at time t for the ith company, X_{qit} represents independent variables for $i = 1,\dots, M$, $t = 1,\dots,T$, $q = 1,\dots,Q$.
Here, the system equation that describes the parameters evolving over time, according to a first order Markov dependence process, is given by

$$\beta_{qt} = \beta_{q,t-1} + w_{qt}, w_{qt} \sim N(0, W_{qt}). \tag{3}$$

An independent normal distribution is assumed as the prior distribution and β_{qt}'s as

$$(\beta_{q,t} \mid \beta_{q,t-1}, \tau_{\beta qt}) \sim \begin{cases} Normal(\beta_{q,t-1}, \tau_{\beta q}) & if \ \ t > 1 \\ Normal(0, \tau_{\beta q}) & if \ \ t = 1 \end{cases},$$

for $q = 0,\dots,Q$ with unknown precisions $\tau_{\alpha}, \tau_{\beta j}$ which are inverse of variances, i.e., $1/\sigma^2_{\beta_q}$.

Third Level:

The unknown precisions (i.e., inverse of variance) are assumed to follow Gamma distribution as,

$$(\tau_{\beta q} \mid a_{\beta_q}, b_{\beta_q}) \sim Gamma(a_{\beta_q}, b_{\beta_q}), \text{ for } \ q = 0,\dots,Q.$$

where a and b are specified.

Figure 1 is the graphical representation of the hierarchical Bayesian model with the first order Markov structure on a time-varying parameter. This illustrates the model structure, and how parameters at a given time point are related to those at other time points. In this figure, a plate represents repeated components for the range, for example (i in $1{:}M$). The arrow shows the specific relationships between two nodes. The descriptions of nodes and arrows are as follows [9]:

- Constant node $\boxed{c}$ describes a quantity fixed by the study design.
- Stochastic node y describes a variable that is given a distribution.
- Deterministic node describes all logical functions of other nodes.
- An arrow "$\longrightarrow$" represents relationship between parent node and descendants.
- An arrow "$\Longrightarrow$" connects two nodes by logical functions.

Figure 1 illustrates that the probability of bankruptcy for a company i at time T depends on time-varying covariates. The covariate effects on the probability of bankruptcy are represented by regression parameters β's at time t, which are dynamic and evolve over time. Y [i,T] indicates bankruptcy for company i at time T, π [i,T] is the probability of bankruptcy, X [i,t] represents covariates, β [t] represents regression parameters at time t, and σ^2 [t] is the variance of β [t]

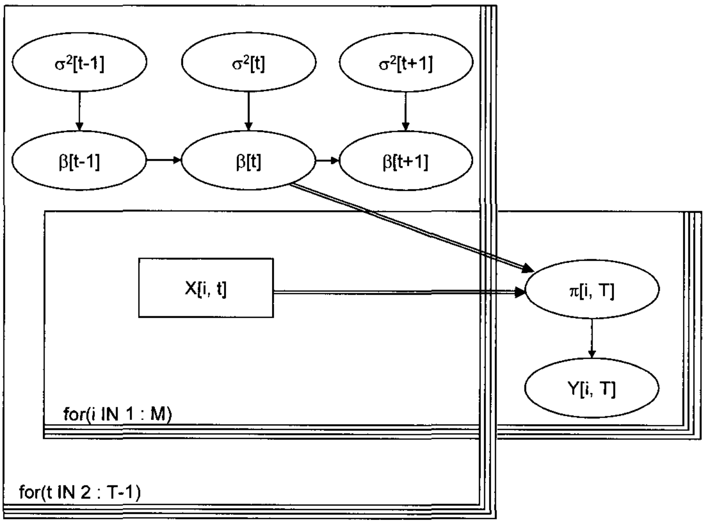

Figure 1. Graphical representation of modeling framework

The conditional independence assumptions are illustrated in Figure 1 such that:

$$Y_{iT} \perp (\tilde{\beta}_{t-1}, \tilde{\beta}_t, \tilde{\beta}_{t+1}, \tilde{\sigma}_t^2, \tilde{\sigma}_{t-1}^2, \tilde{\sigma}_{t+1}^2) \mid \pi_{iT},$$

$$\pi_{iT} \perp (\tilde{\beta}_{t-1}, \tilde{\beta}_{t+1}, \tilde{\sigma}_{t-1}^2, \tilde{\sigma}_t^2, \tilde{\sigma}_{t+1}^2) \mid \tilde{\beta}_t$$

where $\tilde{\beta}_{t-1} = (\beta_{1t-1},...,\beta_{Qt-1})'$, $\tilde{\beta}_t = (\beta_{1t},...,\beta_{Qt})'$, $\tilde{\beta}_{t+1} = (\beta_{1t+1},...,\beta_{Qt+1})'$,

$\tilde{\sigma}^2_{t-1} = (\sigma^2_{1t-1},...,\sigma^2_{Qt-1})$, $\tilde{\sigma}^2_t = (\sigma^2_{1t},...,\sigma^2_{Qt})$, $\tilde{\sigma}^2_{t+1} = (\sigma^2_{1t+1},...,\sigma^2_{Qt+1})$.

That is, the indicator variable representing bankruptcy is independent of all other parameters in the model given the probability of bankruptcy. The probability of bankruptcy is independent of all other parameters in the past and future given current logistic regression parameters, $\tilde{\beta}_t$. However, the current logistic regression parameters at time t will learn from parameters estimated from the previous time point t-1 by the relationship expressed in (3).

The Bayesian analysis requires the full joint posterior distribution which can be written as

$$p(\pi_{1T},...,\pi_{MT},\tilde{\beta}_1,...,\tilde{\beta}_T,\tilde{\sigma}^2_1,...,\tilde{\sigma}^2_T \mid Y_{1T},...,Y_{MT})$$
$$\propto \prod_{i=1}^{M}\prod_{t=1}^{T} p(Y_{iT} \mid \pi_{iT})p(\pi_{iT} \mid \tilde{\beta}_t)p(\tilde{\beta}_t \mid \tilde{\sigma}^2_t).$$

The above posterior distribution can not be obtained in an analytically tractable form. Thus, the evaluation of parameters requires the use of Markov Chain Monte Carlo simulation methods such as the Gibbs sampler [4, 5]. The model will be implemented in the WinBUGS programming environment [8].

3. DATA ANALYSIS

3.1 Inference about the Parameters

The data set consists of N = 1000 for t = 1, ..., 3. We considered the financial data of most recent three years prior to the event of bankruptcy. As discussed in the previous section, nine predictors were considered and the data values were standardized to avoid the scaling issue of variables. Those were gross value added to total assets (X_1), increase in assets (X_2), net return on assets (X_3), net working capital to total assets (X_4), current liabilities to total assets (X_5), depreciation expense (X_6), interest expense to total expense (X_7), net interest expenses (X_8), and wage and salary expenses (X_9). Bayesian analysis requires the specification of prior distribution for each unknown quantity of the model. We used non-informative but proper prior. As discussed in the previous section, β_{qt} is assumed to follow $Normal(\beta_{q,t-1},\tau_{\beta q})$ for $t > 0$ and $Normal(0,\tau_{\beta q})$ for t = 0. We could specify time-dependent priors for the precision, but to simplify parameters to be estimated, we assume that $\tau_{qt} = \tau_q$ for all t. The precision τ_q is assumed to follow Gamma(0.1, 0.1).

The DGLM model discussed in the previous section was implemented using WinBUGS software, which is freely available. After the burn-in sample of 100,000 iterations, we simulated additional 20,000 iterations and obtained a posterior sample of 2,000 realizations from the posterior distributions with thinning at 10th iterations of this sample. Even though this was not necessary in all cases, this thinning approach was taken to ensure the convergence of the Gibbs sampler. It took about 5 minutes to run 10,000 iterations running on a Pentium IV 3 GHz machine, and no convergence problems were suspected. GEE model was also implemented using PROC GENMOD in SAS for comparison purpose.

Table 1. Parameter Estimates (p–value) by GEE and Posterior mean (95% credible interval) of parameters by DGLM

	GEE	DGLM$^+$ (95% CI)*	Odds Ratio from DGLM (95% CI)
Intercept	-3.5663 (<0.0001)	-2.423 (-2.71, -2.15)	0.089 (0.07, 0.12)
X_1	-0.6233 (0.08)	0.0828 (-0.51, 0.62)	1.086 (0.60, 1.86)
X_2	0.5734 (0.04)	0.4734 (-0.04, 0.96)	1.605 (0.96, 2.61)
X_3	-0.4925 (0.004)	-1.1000 (-1.77, -0.44)	0.333 (0.17, 0.64)
X_4	0.2774 (0.20)	0.3553 (-0.03, 0.74)	1.427 (0.97, 2.10)
X_5	0.3349 (0.13)	0.4863 (0.14, 0.86)	1.626 (1.15, 2.36)
X_6	-0.6175 (0.001)	-0.564 (-0.87, -0.27)	0.569 (0.42, 0.76)
X_7	0.2503 (0.08)	0.3117 (0.07, 0.55)	1.366 (1.07, 1.73)
X_8	0.0521 (0.71)	0.3804 (0.13, 0.64)	1.463 (1.14, 1.90)
X_9	0.2201 (0.24)	-0.2129 (-0.56, 0.09)	0.808 (0.57, 1.09)

+ Posterior mean for t=3, which is upon the event of bankruptcy or censored.
*95% credible interval: (β_L, β_U) such that $(1-\alpha) = \Pr(\beta_L < \beta < \beta_U \mid \text{Data})$, see for details [3].

Table 1 shows the results obtained from the posterior samples of DGLM model in comparison to those of GEE. The posterior means of parameters from DGLM presented in Table 1 are for t=3 while the parameter estimates from GEE approach are the averaged marginal effects of predictors on the risk of bankruptcy. Using a 95% credible interval, which is Bayesian analogous to the classical 95% confidence interval, those predictors retaining statistical significance are X_3, X_5, X_6, X_7, and X_8. On the contrary, those significant variables from GEE using

$\alpha = 0.05$ were X_2, X_3, and X_6. The third column of Table 1 shows odds ratios for each predictor, which was obtained by the exponentiating parameter estimates.

3.2 Classification Accuracy

In a logistic regression, the choice of a cutoff value arises when the interest of the researcher lies in making judgment about the event of bankruptcy. The goodness of the model fit in terms of making correct classifications for the event or non-event depends on the cutoff value of the estimated probability of bankruptcy. One way to find the optimal cut point is to find the sensitivity (True Positive Rate: TPR) and specificity (1-False Positive Rate: FPR) pair that maximizes the function {Sensitivity-False Positive Rate}[15]. Table 2 shows how well the DGLM approach classified the bankrupt companies (sensitivity) and how well it classified the non-bankrupt companies (specificity) based on the optimal cutoff point. Both sensitivity and specificity could be maximized (sensitivity = 67.56%, specificity = 65.42%) when the cutoff value of the estimated probability of bankruptcy was 0.095. As evidenced in Table 2, DGLM outperformed GEE approach in terms of correct classification.

Table 2. Classification accuracy: Sensitivity (True Positive Rate) and Specificity (1−False Positive Rate) by the optimal cutoff value

Model	Optimal Cut value	Sensitivity (TPR)	Specificity(1-FPR)
GEE	0.035	0.662	0.622
DGLM	0.095	0.676	0.654

We also used the models to forecast the risk of bankruptcy for the holdout data (N=216) which was not used in model building. As shown in Table 3, the superiority in predictability for the new data is even more obvious with holdout data.

Table 3. Classification accuracy for holdout data: Sensitivity (True Positive Rate) and Specificity (1−False Positive Rate) by the optimal cutoff value

Model	Optimal Cut value	Sensitivity (TPR)	Specificity(1-FPR)
GEE	0.035	0.56	0.52
DGLM	0.080	0.63	0.58

Figure 2 shows the Receiver Operating Characteristic (ROC) curves for GEE and DGLM approaches. An ROC curve, a plot of a model's TPR versus FPR, illustrates how different cut values for classification of bankruptcy produce different values for the judgment's TPR and FPR. The accuracy of the judgment from a model can be represented as the area under the ROC curve, which can take values between 0.0 and 1.0. If the classification is perfectly accurate, the area under the curve would be 1.0. The closer the ROC curve to the upper left hand corner, the more accurately the model can classify the events. In general, the ROC curves assume the bi-normal distribution of scores and the computation of the area under the curve is possible with the bi-normal assumption. However, we empirically obtained the ROC curves, and the area under the curve was estimated by summing the area of trapezoids formed by connecting the points of the ROC curve. A formula for the nonparametric estimate of the area is [14]

$$\frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \Psi(T_{1i}, T_{0j}),$$

where n_0, n_1 represent the numbers of non-bankrupt and bankrupt companies respectively, T_{0j}, T_{1i} represent the estimated probabilities of bankruptcy for non-bankrupt companies and bankrupt companies respectively. Ψ is a function of two variables: $\Psi(X,Y) = 0$ if $Y > X$, 1/2 if $Y = X$, and 1 otherwise. The estimated area under ROC curves were 0.723 for GEE model and 0.747 for DGLM model, i.e., representing better classification of bankruptcy by different cutoff values.

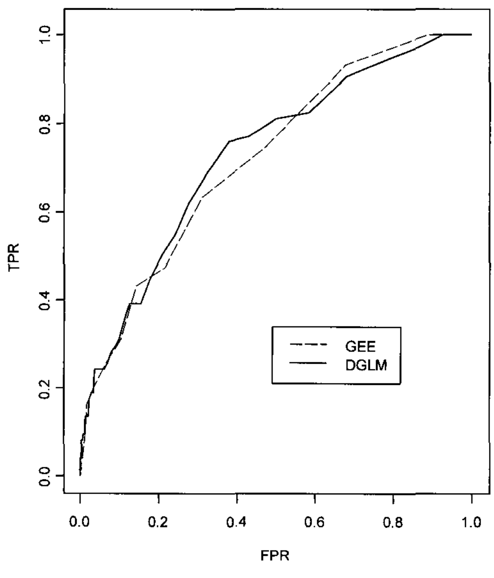

Figure 2. Empirical ROC curves for GEE and DGLM

3.3 Dynamic Parameters

DGLM enables us to examine how the effects of predictors change over time. Figure 3 shows the posterior distributions of the parameters retaining statistical significance in (2). The parameters at time t have been evolved dynamically from t-1 by the first order Markov process as in (3). The effects of X_3(net return on assets) and X_6(depreciation expense) were getting stronger with time progress in a negative direction whereas the effect of X_5(current liabilities to total assets) moved the opposite direction. We note that the effects of X_7(interest expense to total expense) and X_8(net interest expenses) remained constant over time. In other words, X_7 and X_8 are the significant predictors even three years prior to the bankruptcy. For X_3, X_5, and X_6, only concurrent effect of predictors can be found and the past information of those predictors was not very useful in the prediction of bankruptcy.

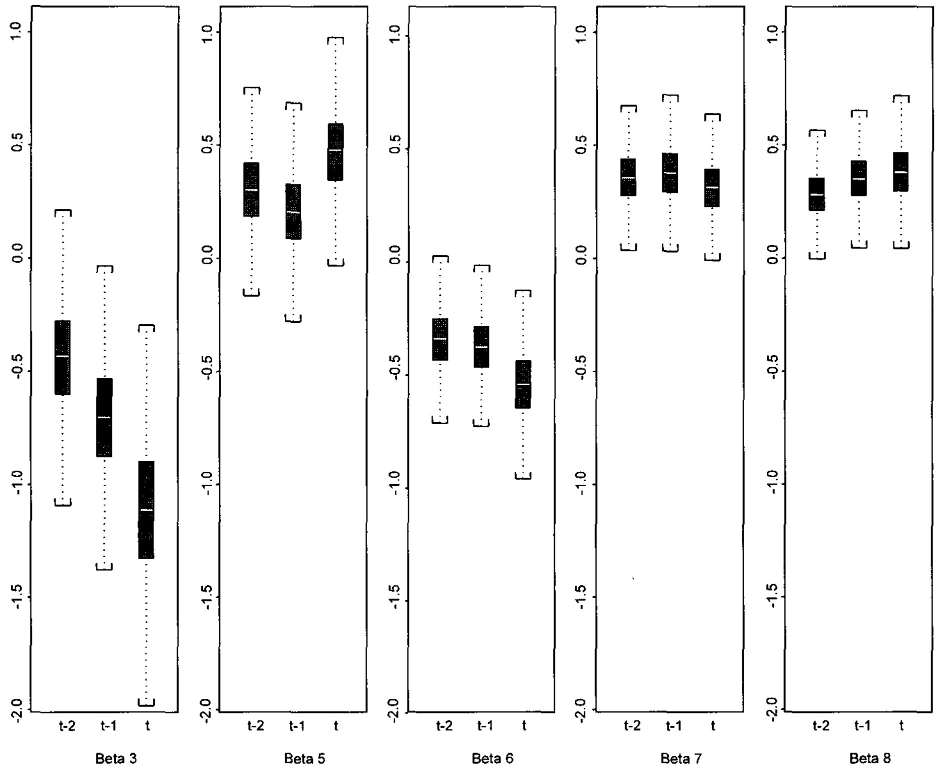

Figure 3. Posterior distributions of the parameters of DGLM: parameters retaining statistical significance with type I error rate = 0.05

4. CONCLUSIONS

We considered a Bayesian dynamic modeling strategy to assess the time-varying effect of risk factors on the likelihood of corporate bankruptcy. Having longitudinal data to estimate the likelihood of corporate bankruptcy, our primary concern was to develop a better modeling strategy to describe dynamically evolving effects of time-varying factors on the risk of bankruptcy than the GEE approach. Note that despite the popularity for analyzing longitudinal data, the GEE approach is static in that it provides the averaged marginal effects of factors.

In comparison with a static modeling approach such as the GEE, Bayesian dynamic model showed better classification accuracy in terms of sensitivity and specificity. Other static modeling approaches such as logistic regression model and discriminant analysis approach were not considered in the present study because these models are suitable for cross sectional data rather than for the longitudinal data. For the same reason, some popular machine learning approaches such as neural networks were not considered. Even though a few machine learning techniques might provide a better predictive power depending on the data, they are not compatible to the parametric statistical modeling approach that is used in the present study. In general, the forecasting accuracy using statistical approaches tends to be lower than that using data mining approaches and this seems to be true for this case, too. One of the reasons would be due to the oversampling of the samples with the case to match the proportion of the case with that of non-case. This will increase the power to detect statistical significance of the effects. Also, non-parametric approaches such as machine learning would raise over fitting issues.

Besides the accuracy in predicting the risk of bankruptcy, the benefits of utilizing Bayesian dynamic approach can be pointed out as the flexibility of describing the dependence structure among time specific model parameters and the ability of assessing how effects of factors change over time. As for the time dependence structure among time specific parameters, we assumed that parameters evolve over time via a random walk type of first order Markov process as in (3). Random walk process was used instead of the first order auto regressive process (AR(1) process) such as $\beta_{qt} = a\beta_{q,t-1} + w_{qt}, w_{qt} \sim N(0, W_{qt})$. The difference between random walk process and AR(1) process would be whether β_{qt}'s are smoothed via the coefficient 'a' in AR(1). We considered random walk process for the dependence structure in order to capture the unsmoothed time specific parameters among different points in time.

Figure 3 could not be obtained in the static approach scheme. The time specific effects of model parameters in Figure 3 show how the effects of factors on the risk of bankruptcy change over time. Among the six statistically significant effects (using type I error rate of 0.05), the effects of X_3 were getting stronger in a negative direction with time, while the effects of the other five variables did not change much from one time point to another. Thus, as the effects of some variables did not show much variation by time, one can raise a question about the usefulness of longitudinal study to predict the risk of corporate bankruptcy. If modeled correctly using a dynamic approach as in the present study, a more sophisticated approach would provide better predictability than the static approach and one can monitor how those effects change over time.

The limitations of the present study would be the time intervals when the data were collected. In Figure 3, some parameters showed stable effects over three time points. This may be due to the fact that the data were collected yearly. Quarterly or monthly collected financial data would be more useful to predict the risk of bankruptcy. Also, the usefulness of Bayesian dynamic modeling approach in estimating the risk of bankruptcy may need to be supported with more replicated studies with other data sets in the future study.

REFERENCES

[1] Altman, E., Financial Ratios, "Discriminant Analysis and the Prediction of Corporate Bankruptcy," *Journal of Finance* 23, 3 (1968), 589-609.

[2] Beaver, W., "Financial Ratios as predictors of failure," *Journal of Accounting Research* 4, 1 (1966), 71-127.

[3] Bernardo, J. M. and A. F. M. Smith, *Bayesian Theory*, John Wiley and Sons, New York, 1994.

[4] Gelfand, A. E. and A. F. M. Smith, "Sampling-based Approaches to Calculating Marginal Densities," *Journal of American Statistical Association* 85 (1990), 972-985.

[5] Gilks, W. and P. Wild, "Adaptive Rejection Sampling for Gibbs Sampling," *Journal of the Royal Statistical Society B* 41 (1992), 337-348.

[6] Lee, G. and Y. Chee, "Bankruptcy Prediction Model with AR process," *Korean Operations Research and Management Science Society* 26, 1 (2001), 109-116.

[7] Ohlson, J., "Financial Ratios and the Probabilistic Prediction of Bankruptcy," *Journal of Accounting Research* 18, 1 (1980), 109-131.

[8] Spiegelhalter, D., A. Thomas, N. Best, and W. Gilks, *Bayesian Inference Using Gibbs Sampling Manual (version ii).* MRC Biostatistics Unit, Cambridge University, 1996.

[9] Sung, M., *Bayesian analyses of Markov Chains: Applications to Longitudinal Data from Psychiatric Treatment Programs*, Ph.D. Dissertation, George Washington University, 2001.

[10] Sung, M., A. Erkanli, A. Angold, and E. J. Costello, "Effects of Age at First Substance Use and Psychiatric Comorbidity on the Development of Substance Use Disorders," *Drug and Alcohol Dependence* 75 (2004), 287-299.

[11] West, M., J. Harrison, and H. Migon, "Dynamic Generalized Linear Models and Bayesian Forecasting," *Journal of American Statistical Association* 80 (1985), 73-83.

[12] Zavgren, C. V., M. T. Dugan, and J. M. Reeve, "The Association between Probabilities of Bankruptcy and Market Responses: A Test of Market Anticipation," *Journal of Business Finance and Accounting* 15, 1 (1988), 19-45.

[13] Zeger, S. L., K. Y. Liang, and P. S. Albert, "Models for Longitudinal Data: A Generalized Estimating Equation Approach," *Biometrics* 44 (1988), 1049-1060.

[14] Zhou, X., N. A. Obuchowski, and D. K. McClish, *Statistical methods in diagnostic medicine*, John Wiley and Sons, New York, 2002.

[15] Zweig, M. H. and G. Campbell, "Receiver-operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine," *Clinical Chemistry* 39 (1993), 561-577.