A SURVIVAL ANALYSIS INCORPORATING AUXILIARY INFORMATION BY A BAYESIAN GENERALIZED METHOD OF MOMENTS: APPLICATION TO PURCHASE DURATION MODELING

Ryosuke Igari∗ and Takahiro Hoshino†

ABSTRACT

In this study, we propose a new estimation procedure for incomplete survival data caused by nonignorable nonresponses or missing censoring indicators. It is widely known that if there is any nonignorable missingness or censoring indicators cannot be fully observed, the results from survival analysis such as the Kaplan-Meier estimator or the Cox proportional hazard model may be biased. However, it sometimes occurs that nonignorable missingness cannot be specified and that the censoring indicators are never or partially observed. We propose a Bayesian generalized method of moments (GMM) approach that utilizes population-level information to identify true survival time and estimates parameters. We apply the proposed model to analyze purchase duration in marketing using purchase history data.

1. Introduction

1.1. Incomplete Survival Data Caused by Missingness or Missing Censoring Indicator Problems

Missing data problems, in which complete data cannot always be obtained, are widely known to researchers in many fields such as social science or nature science. Missing or incomplete data problems are caused by non-response, censoring, truncation, or dropout. In this study, we focus on survival analysis for incomplete data where there is some non-ignorable missingness and censoring indicators cannot be fully observed. Survival analysis is widely used in fields such as biostatistics, medical statistics, economics, and marketing. In the context of survival analysis or any usual statistical modeling, like linear regression or generalized linear models, it is known that simple estimators may yield biased results in the case of nonignorable missingness (e.g., Little and Rubin, 2002). Various methods have been proposed to deal with nonignorable missingness, but these methods require the correct specification of missing-mechanisms, which is quite difficult in real data analyses. In this study, we propose a new estimation method incorporating auxiliary information without specifying the missing-mechanisms.

Furthermore, the censored data problem is important in the specific case of survival analysis. In survival analysis, right censoring is common, and censoring indicators δ are assumed to be fully available. The subject i’s observed time-to-event is usually set to

∗Graduate School of Economics, Keio University, 2-15-45, Mita, Minato-ku, Tokyo 108-8345, Japan
E-mail: igariryosuke@yahoo.co.jp
†Faculty of Economics, Keio University, 2-15-45, Mita, Minato-ku, Tokyo 108-8345, Japan
E-mail: hoshino@econ.keio.ac.jp

Key words: Proportional Hazard Model; Incomplete Data; Nonignorable Missingness; Missing Censoring Indicator; Quasi-Bayesian Inference; MCMC
$T_i = \min(T_i^*, C_i)$, where $T_i^*$ is the survival time and $C_i$ is the censoring time. It is widely known in survival analysis that ignoring censoring may cause biased estimates (Gaynor et al., 1993; Klein and Moeschberger, 2003). If the censoring indicator is fully obtained by researchers, the Kaplan-Meier estimator (Kaplan and Meier, 1958) or the Cox proportional hazard model (Cox, 1972) can be applied. However, if the censoring time $C_i$ is different for each subject $i$ and $C_i$ cannot be known to researchers, then the censoring indicator cannot be fully observed in some cases. These incomplete censoring problems are also known as unknown censoring times or missing censoring indicators (Jonker, 2003; Subramanian, 2006, 2011; Qiu et al., 2015). For example, van der Laan and McKeague (1998) noted in studies that relevant death certificate information can be missing or autopsy results and hospital case notes can be epidemiologically inconclusive, that is, the censoring indicators are sometimes missing. In this situation, survival analysis such as the Kaplan-Meier estimator or the Cox proportional hazard model does not work appropriately. In application studies, researchers have often imputed the last observed event as the unobserved censoring time. However, it is inevitable that the last observed event precedes the true censoring, which underestimates the risk of events and leads to large biases (Jonker, 2003). Subramanian (2011) and Qiu et al. (2015) propose a model considering the missing censoring indicator by imputation methods. Concretely, the authors assume the missing at random (MAR) for missing censoring indicator, that is, the authors model the missing censoring indicators, which are explained by observed time-to-event and covariates using the non-parametric imputation method or the parametric multiple-imputation method. However, if the missing-mechanism cannot be specified, their models that assume MAR cannot be applied. We propose a method incorporating auxiliary information without specification of the missing-mechanism.

This situation is common in other fields such as the social sciences. We apply our model to purchase duration analysis in marketing. There are many studies that deal with purchase duration using survival data analysis (e.g., Allenby et al. 1999; Seetharaman and Chintagunta, 2003; Bijwaard et al., 2006; Igari and Hoshino, 2016). The role of duration analysis in marketing is to estimate the effect of marketing promotions, such as price coupons, and predict the time when consumers are highly likely to purchase products. In purchase duration modeling, researchers analyze the time to event for the next purchase event using purchase history data. Then, researchers usually delete some observations for consumers who have not purchased products for a long time, because researchers are interested in active consumers who may purchase products at the stores in the future. However, the customers whose purchase data are deleted may purchase the products in the future. This case is considered the same as unknown censoring or the missing censoring indicator problem.

As a workaround, for example, Sinha and Chandrashekaran (1992) proposed a split hazard model that divides customers into two groups: (1) customers who will eventually purchase a new product and (2) customers who will never purchase a new product. In the split hazard model, the indicators, which show that customers will eventually purchase a product, are partially unobserved. The two customer groups are divided using a logistic regression model, and the basic idea of Sinha and Chandrashekaran (1992) is the same as that of Subramanian (2011) or Qiu et al. (2015), in which there is a need to specify the missing-mechanisms. Additionally, the same situations are found in marketing fields such as customer base analysis (e.g., Schmittlein and Peterson 1994). In marketing, the number of customers who are active in a company’s database is called the customer base, and understanding the customer base is important because most of the profit originates
from high frequency customers. In customer base analysis, researchers sometimes assume that customer defection occurs randomly and active customers purchase products at regular intervals (e.g., Ehrenberg, 1972; Schmittlein and Peterson, 1994). These assumptions are common in practical business marketing but are the same as the unknown censoring and missing censoring indicator problems.

We address incomplete data problems caused by nonignorable missingness, in which specifying the missing-mechanism is difficult, and we also deal with incomplete data problems caused by unknown censoring or missing censoring indicators, using auxiliary information obtained from other data sources. Then, we consider the case where the parameters are estimated using only the observed data, that is, we ignore the durations with missing values. Furthermore, in the case of missing censoring indicators, we do not consider whether or not each observation is censored, and we estimate the parameters from the only obtained times to events using auxiliary information. For the use of auxiliary information, we propose the Bayesian generalized method of moments (GMM), which is one of the quasi-Bayesian inferences, and we estimate parameters using the Markov chain Monte Carlo (MCMC) method. It is known that survival models are generally hard to fit, especially in the presence of complex censoring schemes, and Bayesian inference using MCMC is straightforward in survival analysis (Ibrahim et al., 2005). Even when the optimization of the objective function is difficult, the Bayesian GMM using MCMC may still be available. Additionally, Bayesian inference, including the Bayesian GMM, can incorporate prior information into a prior distribution.

1.2. The Purpose of This Study: Incorporating Auxiliary Information

For real data, the observed individual-level data are likely to be biased for various reasons, including selection bias and nonignorable missingness, causing the resulting estimates to be biased. Therefore, complete data, without any biases, are occasionally unavailable, and analysis from biased data would lead to biased estimators. On the other hand, researchers can sometimes obtain auxiliary information from other sources such as government statistics or research institutions. The auxiliary information given by other sources are limited to summary statistics such as averages or proportions of variables, and the parameters that show the relationships between dependent and independent variables cannot usually be obtained by researchers. This information cannot be incorporated into the prior distribution of Bayesian modeling except for simple probability distribution models such as normal or binomial distributions, in which averages or proportions imply the parameters of models directly. In this setup, some studies use auxiliary information to strengthen the accuracy of individual-level data modeling. Imbens and Lancaster (1994) and Hellerstein and Imbens (1999) proposed a method of incorporating the auxiliary information into individual-level models using the GMM. Similarly, Qin (2000) and Chaudhuri et al. (2008) propose empirical likelihood approaches that include auxiliary information in individual-level modeling. However, there is no study that addresses auxiliary information in individual-level modeling using the Bayesian GMM. From the viewpoint of missing data analysis, Nevo (2003), Qin (2000), and Qin and Zhang (2007) use auxiliary information to strengthen the analysis of data with missing responses using the GMM or empirical likelihood. However, their models require the specification of missing-mechanisms. As we mentioned, censoring indicators sometimes cannot be fully obtained by researchers. In this situation, the proposed model enables us to estimate parameters appropriately without fully observed censoring indicators by incorporating auxiliary information. Additionally, the proposed model can manage incomplete data without the specification of missing-mechanisms, that is, our model can be
applied to data analysis for which missing values are not missing at random.

Thus, there is no study that addresses survival data analysis with nonignorable missingness or missing censoring indicators using auxiliary information. We use auxiliary information to manage biased data that have nonignorable missingness or missing indicators, that is, unknown or missing censoring indicators. Additionally, our approach is not limited to only nonignorable missingness or missing censoring indicator problems in survival analysis but rather is applicable to various statistical models, such as linear regression models, generalized linear models, and longitudinal analysis using incomplete data. The essence of our approach is not to substitute auxiliary information for incomplete data such as missing durations or missing censoring information in survival analysis, but to estimate proper parameters from biased individual-level data by incorporating auxiliary information using the Bayesian GMM.

2. Model

2.1. Proportional Hazard Model

First, we consider the general hazard function in the proportional hazard model,

\[
\lambda(t) = \lim_{\Delta t \to 0} \frac{p(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \lambda_0(t) \exp(\mathbf{x}^T \beta),
\]

where \( \lambda_0(t) \) is the baseline hazard function, \( \mathbf{x} \) is the covariate vector, and \( \beta \) is its coefficients vector. Equation (1) covers the Cox proportional hazard model (Cox, 1972).

The survival function \( S(t) \) is

\[
S(t) = \exp\left\{ - \int_0^t \lambda(u) du \right\}.
\]

In the general formulation, the probability density function (pdf) of the proportional hazard model is

\[
f(t) = \lambda(t)^\delta S(t),
\]

where there is a need for the censoring indicator \( \delta \) to be fully observed. If \( \delta \) is missing partially or fully, the proportional hazard model cannot work appropriately (van der Laan and McKeague, 1998; Jonker, 2003).

For the baseline hazard function \( \lambda_0(t) \), we can use either parametric or nonparametric models. In the empirical analysis, we focus on the parametric hazard models. In the parametric hazard models, exponential, Weibull, log-logistic and log-normal distributions are typically employed (e.g., Klein and Moeschberger, 2003; Ibrahim et al., 2005). In the real data analysis, we assume the Weibull distribution in the baseline hazard function \( \lambda_0(t) \), because the Weibull distribution is recommended in the marketing purchase duration model (Sinha and Chandrashekaran, 1992; Grewal et al., 2004; Hoshino, 2013; Igari and Hoshino, 2016). Although we employ the Weibull hazard model from this point on, the proposed model can be applied to other parametric or nonparametric hazard models. When we use a nonparametric model such as Cox’s proportional hazard model, we can define the score function of the partial likelihood as a moment restriction.

The baseline hazard function of the Weibull model is

\[
\lambda_0(t) = \alpha t^{\alpha-1},
\]
where $\alpha(>0)$ is the shape parameter. We do not contain scale parameter in the baseline hazard function, because we include the intercept in coefficients $\beta$, that is, $x^T \beta = \beta_0 + x^T \beta_1$, in which $\beta_0$ plays a role on the scale parameter.

Then, the survival function $S(t)$ of the Weibull hazard model is

$$S(t) = \exp\left\{-\exp(\beta_0 + x^T \beta_1)t^\alpha\right\}. \quad (5)$$

In the general formulation, the pdf of the Weibull hazard model is

$$f(t) = \left[\alpha t^{\alpha - 1}\exp(\beta_0 + x^T \beta_1)\right]^{\delta} \left\{\exp\left(-\exp(\beta_0 + x^T \beta_1)t^\alpha\right)\right\}. \quad (6)$$

For the estimation of the hazard models, maximum likelihood estimation or Bayesian estimation via MCMC are widely used (e.g., Klein and Moeschberger, 2003; Ibrahim et al., 2005). The general MCMC procedure of Bayesian estimation in Weibull hazard models is shown in Ibrahim et al. (2005) or Igari and Hoshino (2016). In this study, we assume that the censoring indicator $\delta$ is not available completely and do not consider $\delta$ in the pdf thereafter. In other words, we delete the data that has missing censoring indicators and use only the complete data in which the event occurs during the observed period. The auxiliary information plays a role in the censoring indicator. In this study, we use the Bayesian GMM to estimate the parameters, because we can incorporate moment restrictions from the auxiliary information into the objective function in the GMM.

### 2.2. Bayesian Generalized Method of Moments

The Bayesian GMM is considered a quasi-Bayesian inference (Chernozhukov and Hong, 2003). Let $T = (T_1, ..., T_n)$ be an independent vector, and $\theta$ be a parameter with $r$-dimensional vector. The quasi-Bayesian posterior is

$$q(\theta|T) = \frac{\exp\{L_n(\theta)\}p(\theta)}{\int_\Theta \exp\{L_n(\theta)\}p(\theta)d\theta} \propto \exp\{L_n(\theta)\}p(\theta), \quad (7)$$

where $p(\theta)$ is a prior distribution for $\theta$, $\Theta$ is the parameter space of $\theta$, and $L_n(\theta)$ is an objective function such as GMM, M-estimators, or empirical likelihoods instead of log-likelihood functions (Chernozhukov and Hong, 2003; Hoshino, 2008; Yin, 2009; Yang and He, 2012).

The quasi-Bayesian posterior means are represented as follows:

$$\hat{\theta} = \int_\Theta \theta q(\theta|T)d\theta = \int_\Theta \theta \left(\frac{\exp\{L_n(\theta)\}p(\theta)}{\int_\Theta \exp\{L_n(\theta)\}p(\theta)d\theta}\right) d\theta. \quad (8)$$

In is shown that under mild regularity conditions, the quasi-Bayesian posterior means are consistent and asymptotically normally distributed (Kim, 2002; Chernozhukov and Hong, 2003; Yin, 2009; Yang and He, 2012).

The GMM-type objective function is defined as follows,

$$L_n(\theta) = -\frac{n}{2} \left(\frac{1}{n} \sum_{i=1}^{n} m(T_i|\theta)^T \Omega_n^{-1}(\theta) \left(\frac{1}{n} \sum_{i=1}^{n} m(T_i|\theta)\right)\right), \quad (9)$$

where $m(T_i|\theta)$ is a moment restriction that is $E[m(T|\theta)] = 0$, and $\Omega_n(\theta)$ is the optimal weight matrix,

$$\Omega_n(\theta) = E\left[m(T|\theta)m(T|\theta)^T\right]. \quad (10)$$
As we mentioned, even when the optimization of the objective function is difficult, the Bayesian GMM can be available with the incorporation of population-level information into \( m(T_i|\theta) \).

### 2.3. Moment Restriction

Next, we define the moment restriction \( m(T_i|\theta) \). We divide the moment restriction into two components: (1) restriction for estimating each parameter, \( m_E(T_i|\theta) \), and (2) restriction from population-level information, \( m_R(T_i|\theta) \),

\[
m(T_i|\theta) = \begin{bmatrix} m_E(T_i|\theta) \\ m_R(T_i|\theta) \end{bmatrix},
\]

where \( m_E(T_i|\theta) \) is an \( r \)-dimensional vector that is the same number of dimensions as parameter \( \theta \), and \( m_R(T_i|\theta) \) is an \( S \)-dimensional vector, which means the number of dimensions in the population-level information. That is, \( m(T_i|\theta) \) has \( (r+S) \) dimensions. In the GMM, we can impose more moment restrictions than the number of parameters due to additional external theory constraints (Hansen, 1982).

**Restriction for Estimating Each Parameter**

For the moment restriction for parameter estimation in the Weibull hazard model, we employ the generalized residual (Lancaster, 1985), which is one of the estimating equation methods,

\[
m_E(T_i|\theta) = \begin{bmatrix} \log(T_i)^2 - \alpha^{-2}\left\{ -\left(\beta_0 + x_i^T\beta_1 + \psi(1)\right)\right\}^2 + \psi'(1) \\ \log(T_i) - \alpha^{-1}\left\{ -\left(\beta_0 + x_i^T\beta_1 + \psi(1)\right)\right\} \\ x_i \left[ \log(T_i) - \alpha^{-1}\left\{ -\left(\beta_0 + x_i^T\beta_1 + \psi(1)\right)\right\} \right] \end{bmatrix},
\]

where \( \theta = (\alpha, \beta_0, \beta_1^T)^T \), \( \psi(1) \) is Euler’s constant, that is, \( \psi(1) = 0.57772\cdots \), and \( \psi'(1) = \pi^2/6 = 1.6449\cdots \).

**Restriction from Population-level Information**

The \( S \)-dimensional population-level information \( t^* = (t_1^*, ..., t_S^*)^T \) is set to identify the parameters of the true time-to-event distribution. The moment restrictions from population-level information \( m_R(T_i|\theta) \) are obtained by letting

\[
m_R(T_i|\theta) = \begin{bmatrix} I_1^i [t_1^* - E[T_i|x_i,\theta]] \\ \vdots \\ I_S^i [t_S^* - E[T_i|x_i,\theta]] \end{bmatrix},
\]

where \( I_s^i = 1 \) when individual \( i \) belongs to group \( s \) (e.g., gender or range of age) . The expectation of the Weibull hazard model is

\[
E[T_i|x_i,\theta] = \Gamma(1 + \alpha^{-1}) \exp\left\{ -\frac{\beta_0 + x_i^T\beta_1}{\alpha} \right\},
\]

where \( \Gamma() \) means the gamma function (Klein and Moeschberger, 2003).

We consider the case that has population-level information, and the proposed method is easily generalized to manage the case using statistics from unbiased external surveys.
Survival Analysis Incorporating Auxiliary Information

(Imbens and Lancaster, 1994). For such cases, we can combine the stochastic information obtained in the external surveys by adding the statistics to moment restriction \( m_R(T_i|\theta) \) and its variance matrix to the relevant part of \( \Omega_n(\theta) \).

### 2.4. Estimation via MCMC

For the estimation of each parameter, we use the Markov chain Monte Carlo (MCMC) method (e.g., Gelman et al., 2013; Koop et al., 2007). We draw \( \theta \) from the quasi-Bayesian posterior in equation (7) using the Metropolis-Hastings (MH) algorithms.

For example, the quasi-Bayesian posterior of \( \beta \) in the Weibull hazard model is

\[
q(\beta|T,\alpha) \propto \exp\{L_n(\beta,\alpha)\}p(\beta)
\]

\[
\propto \exp\left\{-\frac{n}{2}\left(\frac{1}{n}\sum_{i=1}^{n}m(T_i|\beta,\alpha)\right)^T\Omega^{-1}(\beta,\alpha)\left(\frac{1}{n}\sum_{i=1}^{n}m(T_i|\beta,\alpha)\right)\right\}
\]

\[
\times \exp\left\{-\frac{1}{2}(\beta - b_0)^TB_0^{-1}(\beta - b_0)\right\},
\]

where \( b_0 \) and \( B_0 \) are the mean vector and variance-covariance matrix of prior distribution \( p(\beta) \).

We draw the new candidate \( \beta \) using random-walk MH algorithms,

\[
\beta^{\text{cand}} \sim N(\beta^{\text{old}}, \Psi),
\]

where \( \beta^{\text{old}} \) means the parameter of the previous iteration in MCMC, and \( \Psi \) is composed of the variance parameters of random-walk MH algorithms. Then, we calculate the probability of accepting the new candidate sample,

\[
p(\beta^{\text{old}} \rightarrow \beta^{\text{cand}}) = \min\left\{1, \frac{q(\beta^{\text{cand}}|T,\alpha)}{q(\beta^{\text{old}}|T,\alpha)} \right\}.
\]

### 3. Application

#### 3.1. Simulation Study

Here, we show the performance of the proposed model using two simulation studies: Simulation 1. nonignorable missing data, and Simulation 2. a missing censoring indicator problem, with simple exponential hazard models. The exponential hazard model is a special case of the Weibull hazard model when the shape parameter is fixed to one (\( \alpha = 1 \)). The hazard function of the exponential hazard model is

\[
\lambda(t) = \exp(\beta_0 + x_i\beta_1).
\]

**Simulation 1. Nonignorable Missing Data**

In simulation 1, we show how the proposed method performs when analyzing data with non ignorable missingness. Now, we explain the method of generating data with nonignorable missingness. First, we generate artificial datasets for the usual hazard model using true parameters. Second, we create a missing indicator using a logistic regression model as in Qiu et al. (2015). To be more concrete, we set the probability that the time to event is missing to be a decreasing function of the time-to-event value \( T_i \). Third, we delete the durations in which the times to events are missing. In parameter estimation, we do not use the missing-mechanism (or selection) model and missing durations but the auxiliary information for incomplete data. We set the proportion of missing values to about 20 \( \sim 30\%\),
which are deleted in the estimation of parameters. We compose the number of moment restrictions ($NMR = (0, 3, 6)$) from auxiliary information on the true time to event among the range of covariate $x$. We set the sample size $n = (200, 500, 1000, 5000)$, and we generate 1,000 datasets for each sample size. For model comparison, we estimate four models: (1) normal Bayes models without auxiliary information, which are estimated by the usual Bayesian inference using the likelihood function, (2) proposed models using the Bayesian GMM without auxiliary information ($NMR = 0$), (3) proposed models using the Bayesian GMM with three pieces of auxiliary information ($NMR = 3$), and (4) proposed models with six pieces of auxiliary information ($NMR = 6$).

We show the MSE ($\times 10^2$) and coverage from a 95% Bayesian credible interval in Table.1. From Table.1, normal Bayes models and $NMR = 0$ without auxiliary information perform the worst in terms of MSEs or coverage. The table shows that as the sample sizes increases, the coverage of the normal Bayes models decreases. On the other hand, the Bayesian GMM models have adequate coverage in both $NMR = 3$ and $NMR = 6$. Comparing the results of $NMR = 3$ and $NMR = 6$, the MSEs of $NMR = 6$ are better than the MSEs of $NMR = 3$, but the large difference in the coverages is not seen between the two proposed models. Next, we show the boxplot of each parameter in the cases $n = 500$ and 5,000 in Figure.1 and the traceplot of MCMC in Figure.2. From Figure.1 and 2, we understand that the Bayesian GMM models can reproduce the true parameters appropriately, but those from the normal Bayes models have large biases. Particularly, the results of the Bayesian GMM when $NMR = 6$ are less biased than the results when $NMR = 3$.

Table 1: Results of Simulation 1 (MSEs and Coverages)

|       | $n=200$ |       |       |       | $n=500$ |       |       |       |       | $n=1000$ |       |       |       |       | $n=5000$ |       |       |       |       |
|-------|---------|-------|-------|-------|---------|-------|-------|-------|-------|----------|-------|-------|-------|-------|----------|-------|-------|-------|-------|
|       | MSE($10^2$) | Proposed Model | Coverage(95%) | Proposed Model | Coverage(95%) | Proposed Model | Coverage(95%) | Proposed Model | Coverage(95%) |
|       | Normal Bayes | NMR=0 | NMR=3 | NMR=6 | Normal Bayes | NMR=0 | NMR=3 | NMR=6 | Normal Bayes | NMR=0 | NMR=3 | NMR=6 | Normal Bayes | NMR=0 | NMR=3 | NMR=6 |
| $n=200$ | $\beta_0$ | 24.336 | 38.649 | 1.158 | 0.148 | 0.111 | 0.109 | 0.953 | 0.890 |
|        | $\beta_1$ | 2.878 | 6.290 | 0.451 | 0.102 | 0.617 | 0.537 | 0.971 | 0.922 |
| $n=500$ | $\beta_0$ | 21.222 | 34.394 | 0.539 | 0.043 | 0.000 | 0.001 | 0.967 | 0.919 |
|        | $\beta_1$ | 2.136 | 5.285 | 0.210 | 0.027 | 0.316 | 0.161 | 0.978 | 0.961 |
| $n=1000$ | $\beta_0$ | 20.241 | 32.734 | 0.285 | 0.017 | 0.000 | 0.000 | 0.950 | 0.946 |
|        | $\beta_1$ | 1.916 | 4.841 | 0.107 | 0.012 | 0.111 | 0.017 | 0.964 | 0.979 |
| $n=5000$ | $\beta_0$ | 19.566 | 31.748 | 0.064 | 0.003 | 0.000 | 0.000 | 0.940 | 0.962 |
|        | $\beta_1$ | 1.723 | 4.533 | 0.027 | 0.004 | 0.000 | 0.000 | 0.954 | 0.941 |

Simulation 2. Missing Censoring Indicator Problem

In simulation 2, we show the performance of the proposed method for the data with missing censoring indicators. Now, we explain the method of generating the data with missing censoring indicators. First, we generate censoring time $C_{i}$, which differs for each subject.
i, and true time to event $T_i^*$ using true parameters. Second, we set $T_i = \min(T_i^*, C_i)$, and we assume that the censoring indicator $\delta_i$ and the censoring time $C_i$ are not available to researchers. That is to say, it is difficult for researchers to identify whether the observed survival times are censored or not, and equation (3), which is a general formation of the proportional hazard model, cannot be applied. In parameter estimation, we do not use the missing-mechanism model and missing censoring indicator but the auxiliary information for incomplete data, as in simulation 1. We set the proportion of missing censoring indicators to about 40%. We compose the number of moment restrictions ($NMR = (0, 3, 6)$) from auxiliary information on the true time to event among the range of covariate $x$. We
set the sample size $n = (200, 500, 1000)$ and generate 1,000 datasets for each sample size. We estimate and compare five models: (1) normal Bayes models without censoring in the likelihood function (without $\delta$), (2) normal Bayes models with censoring in the likelihood function (with $\delta$), (3) proposed models using the Bayesian GMM without auxiliary information ($NMR = 0$), (4) proposed models using the Bayesian GMM with three pieces of auxiliary information ($NMR = 3$), and (5) proposed models with six pieces of auxiliary information ($NMR = 6$). Although, as we mentioned, the censoring time cannot be observed by researchers, we assume that average censoring time is known to researchers and calculate the censoring indicator $\delta_i$ empirically in (2) normal Bayes models (with $\delta$). That is to say, we use the average censoring time virtually (in fact, it is unobserved in censoring missing indicator problems).

We show the MSE ($\times 10^2$) and coverage in Table.2. From Table.2, normal Bayes models without $\delta$ perform the worst out of all the models. Similarly, the results of normal Bayes models with $\delta$ are also severely biased in comparison to the proposed methods ($NMR = 3$ or $NMR = 6$). The table shows that the Bayesian GMM models have adequate coverage.
Survival Analysis Incorporating Auxiliary Information

for both $NMR = 3$ and $NMR = 6$. Comparing the results of the two Bayesian GMM models, the MSEs when $NMR = 6$ are better than those when $NMR = 3$. Next, we show the boxplot of each parameter in the cases $n = 500$ and 1,000 in Figure.3. From Figure.3, we understand that the Bayesian GMM models with $NMR = 3$ or $NMR = 6$ can reproduce the true parameters appropriately, but those from the two normal Bayes models or when $NMR = 0$ have large biases. Additionally, the results of the Bayesian GMM when $NMR = 6$ are less biased than the results when $NMR = 0$ or $NMR = 3$, which shows that the more auxiliary information is used, the less biased the estimated results are. Thus, it is shown that if researchers could obtain the average censoring time, the estimated results would be biased and may lead to an incorrect interpretation under missing censoring indicators. However, the proposed model with auxiliary information can estimate the parameters appropriately.

Table 2: Results of Simulation 2 (MSEs and Coverages)

|        | MSE($\times 10^2$) | Coverage(95%) |
|--------|--------------------|---------------|
|        | Normal Bayes       | Proposed Model| Normal Bayes       | Proposed Model|
|        | without $\delta$   | with $\delta$ | NMR=0 NMR=3 NMR=6| NMR=0 NMR=3 NMR=6|
| n=200  |                    |               |                  |
| $\beta_0$ | 46.583            | 18.928        | 22.695 0.468 0.077| 0.00 0.00 0.00 0.990 0.986|
| $\beta_1$ | 30.851            | 12.063        | 18.463 4.292 0.275| 0.00 0.035 0.006 0.997 0.961|
| n=500  |                    |               |                  |
| $\beta_0$ | 46.255            | 18.580        | 22.050 0.114 0.006| 0.00 0.00 0.00 0.986 0.992|
| $\beta_1$ | 30.955            | 11.960        | 18.336 0.592 0.029| 0.00 0.00 0.00 0.980 0.961|
| n=1000 |                    |               |                  |
| $\beta_0$ | 46.023            | 18.306        | 21.842 0.074 0.003| 0.00 0.00 0.00 0.972 0.991|
| $\beta_1$ | 30.935            | 11.781        | 18.230 0.315 0.016| 0.00 0.00 0.00 0.966 0.952|

3.2. Application to Purchase Duration Model in Marketing

We apply our model to purchase duration analysis in marketing. In marketing practice, researchers sometimes assume that when a customer does not purchase products for over 30 days from the last purchase, the data are deleted in the analysis, which is considered to be the same as missingness or missing censoring indicators. These uses of data are common in business practice in marketing and are called recency and frequency (RF) analysis (see Figure.4). In RF analysis, R means the last purchase incidence. However, customers whose purchase data are deleted may still purchase the products in the future. Here, we show that the results from this analysis cause biased estimates and wrong interpretations.

In the empirical analysis, we use the Syndicated Consumer Index (SCI) data provided by Intage Inc. The SCI data are the de facto standard for purchase panel data in the Japanese marketing field. The SCI records the purchase incidence, purchased products, number of products purchased by consumers, amounts and prices of products, and stores in which the purchase occurred, with dates and times. In the analysis, the observations are deleted when a customer does not purchase products for over 30 days from the last purchase incidence. This dataset is regarded to be incomplete data, which can yield severely biased
results. In the analysis, we use tissue/toilet paper as the product category. We analyze data from January to June 2015 (n = 11579). We show the summary statistics and histograms of observed duration in Table 3 and Figure 5. The tissue/toilet paper products, which we use in the empirical analysis, are purchased every 20 days on average. If researchers can use panel data that capture purchase incidences for a long time (e.g., two or three years), the bias from unobserved censoring indicators might be minimal. However, if we analyze the records of products with a long interval between purchases, like durable goods, the bias from unobserved censoring indicators might be too large to ignore despite the use of long-term panel data. Thus, it is valid for our approach to deal with purchase data with an unknown censoring time.
We define the day from the last purchase as the dependent variable $T$ and four types of covariates $x$, *price*, *gender* (male=1), and *family size*. Then, *price* is scaled and equals 1 when the price in the purchase incidence is equal to the regular price. The coefficients of price should be negative because consumers are likely to purchase products when price discounts are available. The coefficients of gender (male) should be negative because female customers purchase products frequently, such as commodity goods in many cases. On the other hand, the coefficients of family size should be positive because the quantities of consumption in large families are typically more than those of a one or two-person household. Finally, we define the auxiliary information $t^*$. We set the true durations as the auxiliary information, which is calculated from the whole SCI data. We use eight total groups for $x$; the three price ranges (*under 0.9, from 0.9 to 1*, and *over 1*); the *male* or *female* genders; and the *presence* or *absence* of a child (See Table.4). Although we calculate the auxiliary information from the whole SCI data, we can calculate this information using government statistics or external syndicated surveys, which are usually provided by marketing research companies.

| Summary                         |
|---------------------------------|
| Number of Observations         | 11579 |
| Duration                       | 18.24 |
| Price                          | 0.99  |
| Gender (Male=1)                | 0.22  |
| Family Size                    | 3.21  |
For the real data analysis, we estimate the parameters from two models, the normal Bayes model and the proposed model. In the two models, we draw 20,000 MCMC samples after the convergence, and we confirm the convergence of each model by the Geweke (1992) method. The estimated results are shown in Table 5, which represents the posterior mean and 95% credible intervals. The results show that the coefficients of the two models are different, especially for price and gender. The normal Bayes model underestimates the effect of price. Additionally, the coefficient on gender in the normal Bayes model is significantly positive, but the coefficient on gender in the proposed model is significantly negative. The results from the normal Bayes model contradict the previous studies and empirical knowledge for marketing practice, but the results from the proposed model have adapted to that knowledge and practice. Thus, the survival analysis from the incomplete data leads to biased estimates, and researchers should consider the incomplete observations. Then, the proposed method can improve the results effectively.
4. Conclusion

In this study, we proposed a new estimation procedure for survival data where times to events are partially missing or censoring indicators are fully or partially unobserved. We incorporated population-level information into moment restrictions and proposed the Bayesian GMM approach using MCMC. In the empirical analysis, the proposed model is applied to purchase duration modeling in the marketing field. We show that analysis from incomplete data may cause biased estimates, and the proposed model can improve the accuracy of estimates by simulation study and real data analysis. In data analysis, when a customer does not purchase products for over 30 days since the last purchase, the data are deleted. However, this timeframe can be defined differently, for instance, 20 or 40 days since the last purchase. Additionally, the ideal timeframe considered may vary across consumers, that is, it may be defined by the degree of detachment from the previous purchase durations for each consumer. Moreover, the timeframe also depends on the average interval between purchases for each product category. For example, detergents and eggs have different intervals between purchases, and it is difficult to generalize the timeframe in practice. The effect of changing this timeframe on the estimated results and the auxiliary information work should be investigated in future research.

For another future research project, we will incorporate latent variables into quasi-Bayesian estimation. If latent variables are available in quasi-Bayesian inference, we can use complex models such as frailty models, random effect models, factor models, and multi-level models with constraining moment restrictions. Additionally, we use SCI data, which capture the purchase behaviors in all stores for the same person in empirical data analysis. These data have been used by many in academic marketing science and this study contributes to research in marketing science. However, store-centric data, such as ID-POS data, which captures only the purchase behaviors in a specific store, have been used in marketing practice. In ID-POS data, if there are any purchase events in other stores, the observed purchase intervals are considered incomplete. For future research, we will propose a duration model with repeated events that has unobserved intermittent missingness using a new quasi-Bayesian method that can incorporate latent variables and auxiliary information regarding intermittent missingness. For such models, it is not possible to obtain valid estimates to make use of auxiliary information and employ latent variable modeling.

Finally, we refer to the accessibility and accuracy of macro information. Not only government statistics, such as the census, but also questionnaire surveys with finite sample

|                   | Normal Bayes Mean        | Proposed Mean        |
|-------------------|--------------------------|----------------------|
| $\alpha$ shape parameter | 1.144 (1.129, 1.159) | 1.017 (1.0142, 1.020) |
| $\beta_0$ intercept  | -3.455 (-3.609, -3.304) | -3.107 (-3.117, -3.097) |
| $\beta_1$ price     | -0.140 (-0.274, -0.001) | -0.313 (-0.323, -0.302) |
| $\beta_1$ gender(male=1) | 0.084 (0.040, 0.128) | -0.098 (-0.100, -0.096) |
| $\beta_1$ family size | 0.060 (0.045, 0.075) | 0.064 (0.062, 0.065) |
numbers can be used as auxiliary information. If these data have a finite sample size, the stochastic information can be used for the auxiliary information in the same manner, as we mentioned. However, the proposed model is based on the premise that macro information is accurate or is free of errors (Chaudhuri et al., 2008). If the accuracy of macro information is low, the estimated results may be biased. Thus, the macro information used should be chosen carefully in practice. On the other hand, recently, government statistics, such as the Family Income and Expenditure Survey or the Consumer Confidence Survey have been made available to the public, and custom-made aggregation services are available in Japan. In the future, auxiliary information will be open and available not only to researchers but also to private companies, and then we think our approach will be helpful and powerful. Additionally, the databases of companies that provide consumer products record purchase histories of only the companies’ brands, not of competing brands. In this situation, our model assists marketing managers through the use of auxiliary information such as the market share or large-scale syndicated surveys, to develop subsequent marketing strategies.

Acknowledgement

We would like to thank the referees for very insightful and constructive comments that substantially improved the paper. We also appreciate that Intage Inc. provided us the dataset we used for the illustrative analysis. This work was supported by JPSP KAKENHI (15J03947 and 26285151) and Yoshida Hideo Memorial Foundation.

REFERENCES

Allenby, G. M., Leone, R. P. and Jen, L. (1999). A dynamic model of purchase timing with application to direct marketing. Journal of the American Statistical Association, 94, 365–374.

Bijwaard, G. E., Franses, P. H. and Paap, R. (2006). Modeling purchases as repeated events. Journal of Business & Economic Statistics, 24, 487–502.

Chaudhuri, S., Handcock, M. S. and Rendall, M. S. (2008). Generalized linear models incorporating population level information: An empirical-likelihood-based approach. Journal of the Royal Statistical Society: Series B, 70, 311–328.

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. Journal of Econometrics, 115, 293–346.

Cox, D. R. (1972). Regression models and life-tables (with discussion). Journal of the Royal Statistical Society: Series B, 34, 187–220.

Ehrenberg, A. S. C. (1972). Repeat-buying: Theory and applications. North-Holland Pub. Co.

Gaynor, J.J., Feuer, E.J., Tan, C.C., Wu, D.H., Little, C.R., Straus, D.J., Clarkson, B. D. and Brennan, M.F. (1993). On the use of cause-specific failure and conditional failure probabilities: Examples from clinical oncology data, Journal of the American Statistical Association, 88, 400–409.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013). Bayesian data analysis 3rd. Chapman & Hall/CRC.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In, Bayesian Statistics 4, J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith eds., 169–193. Oxford University Press.
Grewal, R., Mehta, R. and Kardes, F. R. (2004). The timing of repeat purchases of consumer durable goods: The role of functional bases of consumer attitudes. *Journal of Marketing Research, 41*, 101–115.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica, 50*, 1029–1054.

Hellerstein, J. K. and Imbens, G. W. (1999). Imposing moment restrictions from auxiliary data by weighting. *The Review of Economics and Statistics, 81*, 1–14.

Hoshino, T. (2008). A Bayesian propensity score adjustment for latent variable modeling and MCMC algorithm. *Computational Statistics & Data Analysis, 52*, 1413–1429.

Hoshino, T. (2013). Joint random effect modeling for repeated durations and discrete choices with selection bias correction: Application to promotion policy planning for potential clients using web access-log data. *Journal of the Japan Statistical Society, 43*, 41–58. (in Japanese).

Ibrahim, J. G., Chen, M. H. and Sinha, D. (2005). *Bayesian survival analysis*. John Wiley & Sons, Ltd.

Igari, R. and Hoshino, T. (2016). A joint model of consumer’s interpurchase-timing and purchase amounts in Online-Offline channels. *Communications of the Operations Research Society of Japan, 61*, 589–599 (in Japanese).

Imbens, G. W. and Lancaster, T. (1994). Combining micro and macro Data in microeconometric models. *The Review of Economic Studies, 61*, 655–680.

Jonker, M. A. (2003). Estimation of life expectancy in the middle ages. *Journal of the Royal Statistical Society: Series A, 166*, 105–117.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*, 457–481.

Kim, J. Y. (2002). Limited information likelihood and bayesian analysis. *Journal of Econometrics, 107*, 175–193.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Koop, G., Poirier, D. J. and Tobias, J. L. (2007). *Bayesian econometric methods*. Cambridge University Press.

Lancaster, T. (1985). Generalised residuals and heterogeneous duration models: With applications to the Weibull model. *Journal of Econometrics, 28*, 155–169.

Little, R. J. and Rubin, D. B. (2002). *Statistical analysis with missing data*. John Wiley & Sons.

Nevo, A. (2003). Using weights to adjust for sample selection when auxiliary information is available. *Journal of Business & Economic Statistics, 21*, 43–52.

Qin, J. (2000). Combining parametric and empirical likelihoods. *Biometrika, 87*, 484–490.

Qin, J. and Zhang, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of the Royal Statistical Society: Series B, 69*, 101–122.

Qiu, Z., Chen, X. and Zhou, Y. (2015). A kernel-assisted imputation estimating method for the additive hazards model with missing censoring indicator. *Statistics & Probability Letters, 98*, 89–97.

Schmittlein, D. C. and Peterson, R. A. (1994). Customer base analysis: An industrial purchase process application. *Marketing Science, 13*, 41–67.

Seetharaman, P. B. and Chintagunta, P. K. (2003). The proportional hazard model for purchase timing: A comparison of alternative specifications. *Journal of Business & Eco-
Sinha, R. K. and Chandrashekaran, M. (1992). A split hazard model for analyzing the diffusion of innovations. *Journal of Marketing Research, 29*, 116–127.

Subramanian, S. (2006). Survival analysis for the missing censoring indicator model using kernel density estimation techniques. *Statistical Methodology, 3*, 125–136.

Subramanian, S. (2011). Multiple imputations and the missing censoring indicator model. *Journal of Multivariate Analysis, 102*, 105–117.

van Der Laan, M. J. and McKeague, I. W. (1998). Efficient estimation from right-censored data when failure indicators are missing at random. *Annals of Statistics, 26*, 164–182.

Yang, Y. and He, X. (2012). Bayesian empirical likelihood for quantile regression. *Annals of Statistics, 40*, 1102–1131.

Yin, G. (2009). Bayesian generalized method of moments. *Bayesian Analysis, 4*, 191–207.

(Received: May 11, 2017, Accepted: September 18, 2017)