Invited Paper

Modelling Clustered Survival Data with Cured Fraction

Angela D. Nalica
School of Statistics
University of the Philippines Diliman

Iris Ivy M. Gauran
School of Statistics
University of the Philippines Diliman
and University of Maryland Baltimore County

In modelling lifetime data, standard parametric theory assumes that all observations will eventually experience the event of interest if they are monitored for a very long period. While every unit starts as susceptible to the event of interest, a fraction of observations may switch into a nonsusceptible group. A mixture cured fraction model with covariates is modified to incorporate random clustering effect to characterize the switch mechanism. Simulation studies and telecommunications data show that cured fraction models with random clustering effect perform better than their parametric counterpart in terms of predictive ability. Moreover, results show that the nonparametric method is superior than modified parametric Cox PH model.

Keywords: mixture cured fraction models, random clustering effect, right-censored lifetime data

1. Introduction

Customer care units of service companies aim to cultivate satisfied customers who will eventually develop loyalty and subsequent retention as a customer. Companies are becoming more and more aware of the crucial importance to fully exploit their existing consumer database to serve as inputs for marketing planning and strategies. A lot of information can be derived from these data such as predicting customer behaviour, customer loyalty and customer satisfaction among others.

Customer relationship management (CRM) plays a critical role in establishing mutually beneficial customer-company relationship. Chalmeta (2005) defines CRM as a set of business, marketing and communication strategies and technological infrastructures designed with the aim of building lasting relationship with the customers, which involves identifying, understanding and meeting their needs. Coltman (2007) illustrates that a superior CRM capability can create positional advantage and subsequent improved performance. Further, it is shown that to be most successful, CRM programs should focus on latent or unarticulated customer needs that emphasize a proactive rather than a reactive market orientation.

There is now a shift in focus from building a large client base to keep the existing ones since markets are becoming saturated and competitive pressure is becoming intense. Extant literature shows that it is much cheaper to retain old clients than to gain new ones. Thus, attention has been directed on detecting customers that are becoming less loyal, also called churners.

The churn model presented in this study is based on the theory of survival analysis. It is assumed that if complete follow-up were possible for all elements of the study population, then each element would eventually experience the event of interest. However, the focus is shifted towards modelling survival data wherein a substantial proportion of the individuals does not experience the event at the end of the observation period. These individuals are classified as long term survivors and are viewed as "cured" in the sense that even after an extended follow-up period, the event of interest may not be observed. Failing to account for such cured subjects would lead to incorrect inferences (Corbiere and Joly, 2007).

When right-censoring is a possibility, the survival function is usually modelled by a mixture model (Boag, (1949) and Farewell, (1982)). This approach allows the simultaneous estimation of the occurrence of the event of interest (incidence) and time of occurrence, given that it can occur (latency). Let $U = 1$ indicate that an individual is susceptible while $U = 0$ denote that the individual is nonsusceptible to the event of interest. Define T to be a non-negative random variable denoting the failure time of interest, defined only when $U = 1$. The mixture model is given by

$$S(t|x, z) = \pi(z)\, S(t|U = 1, x) + 1 - \pi(z) \qquad (1)$$

where $S(t|x, z)$ is the unconditional survival function of T for the entire population, $S(t|U = 1, x) = P(T > t \,|\, U = 1, x)$ is the survival (latency) function for susceptible individuals given a covariate vector x and $\pi(z) = P(U = 1 \mid z)$ is the probability of being susceptible given a covariate vector z, which may include the same covariates as x. The survival function of cured individuals can be set to one for all finite values of t because they will never experience the event of interest. It should be noted that $S(t|x, z) \to 1 - \pi(z)$ as $t \to \infty$. If all the individuals are susceptible

to the event of interest, $\pi(z_i) = 1$ for all z_i. This means that when there is no cured fraction then the mixture cure model reduces to the standard survival model.

The conditional latency distribution $S(t|U=1)$ can take the form of parametric distributions. Among the parametric models, Weibull distribution is commonly used to model survival data. After reparametrization (Gamel et. al., 2000), this distribution can be expressed as:

$$S(t|U=1) = \exp\left\{-\exp\left(\frac{\log(t-\mu)}{\sigma}\right)\right\} \tag{2}$$

In proportional hazards models, $S(t|U=1, x) = S_0(t|U=1)^{\exp(\beta' x)}$ where $S_0(t|U=1)$ is the baseline hazard function. If $S_0(t|U=1)$ is left arbitrary, the model is defined as the Cox's Proportional Hazards (PH) Mixture Cure Model (Cox, 1972).

The Cox Proportional Hazards Mixture Cure Model was transformed into an additive model so that the components can be estimated separately with the appropriate estimation method. There are three components of the model namely, the baseline hazard function, the cluster component and the covariate. These three components of the model are estimated one at a time, i.e., the most important term comes first. Backfitting algorithm is known to provide good estimates among the model terms estimated early on (Santos and Barrios, 2012). The estimation procedure is summarized as follows:

1. Since the richest information is provided by the covariate, it is believed that the covariate dominates the postulated model. Initially, the baseline hazard function and the cluster component are ignored first and the covariate is estimated nonparametrically using smoothing splines. Spline smoothing is used since there is only one function to be estimated.
2. The partial residual is computed. At this point, the partial residual contains information on the baseline hazard function and cluster component.
3. Since the cluster component is deemed as the next important part of the model, the baseline hazard function is ignored further and the partial residual in (2) is used to estimate the cluster component. The residual is computed again, now a function of the baseline hazard function alone.
4. Once more, the residual will be used to estimate baseline hazard function. This procedure is iterated until convergence is observed.

The data used in this study are represented as clustered survival information to take into account the homogeneity of the individuals belonging in the same cluster. However, in most of the survival data models described in the literature,

heterogeneities between individuals have been taken into account only in the form of the observable covariates. Thus, to be able to model clustered survival data, the Cox PH Model was modified to include a cluster-specific random component.

To establish a direct comparison, different predictive churn models are considered in this study: the modified Cox PH model and the Weibull model. However, for the modified Cox PH model, two methods are employed in the estimation procedure: the parametric and the nonparametric approaches. The predictive abilities of these models are compared using simulation and a telecommunications data. Comparison is done through the relative difference in median absolute percentage error (MAPE). In the context of the telecommunications data set, the definition of prepaid churn is based on a number of successive months with zero top-up. Due to the constraints in data availability, definition of churn is restricted to having zero top-up in 6 consecutive months.

2. Results and Discussion

Simulation study

To compare the parametric model and the proposed model, a simulation study was performed. The simulation boundaries include (1) the percentage of observations classified as right-censored and (2) presence of misspecification in the model. We considered 5%, 10%, and 20% censoring in assessing the predictive ability of the model. On the other hand, presence or absence of misspecification was considered.

Table 1. Relative Difference in Mean of MAPE based on the Simulation Scenarios

Simulation Scenarios	% Censoring	Relative Difference in Mean of MAPE		
		Parametric vs. Modified Cox PH (Nonparametric)	Modified Cox PH (Parametric) vs. Modified Cox PH (Nonparametric)	Parametric vs. Modified Cox PH (Parametric)
With Misspecification	5%	10.6401	6.4433	4.4859
	10%	9.3905	4.7524	4.8696
	20%	9.9885	5.3974	4.8531
Without Misspecification	5%	61.1479	27.6651	46.2885
	10%	61.4868	26.3581	47.7020
	20%	61.5044	27.3299	47.0269

In Table 1, modified Cox PH model performs better than the parametric model whether in the presence or absence of misspecification in the model. When the nonparametric modified Cox PH model is pitted against the parametric model,

it can also be noted that as the percentage of the right-censored observations increases, the corresponding relative difference in MAPE also increases. This indicates that if a sizable amount of observations have been subject to an intervening variable or a curing situation, then the predictive ability of the model will be affected.

Moreover, in the ideal scenario where contamination is not present, the model produces relatively lower MAPE compared to the scenario wherein misspecification error is introduced into the model. This explains why data sets with limited information on covariates and with some degree of contamination would greatly affect the predictive ability of the two models.

Application to telecommunications data

The models are applied to a telecommunications data set which contains monthly transactional details of prepaid customers observed in January 2010 to February 2012. The data consists of about 38,000 subscribers and is limited only on billing information (amount of and frequency of top-up).

Based on the amount of top-up, ten clusters were identified. This number is based on the result of the study of Arceneaux and Nickerson (2009) which highlights that an increase in the number of clusters leads to an increase in efficiency. The first cluster includes all subscribers whose total amount of top-up is less than twenty pesos per month. This implies that the subscribers belonging in this cluster are the ones that bought the Subscriber Identity Module (SIM) card, used all the freebies and opted to recharge on a short duration only. Almost all subscribers belonging in this cluster were the earliest churners. Meanwhile, the tenth cluster is characterized by subscribers whose total amount of top-up is more than PhP 250 per month. Upon consideration of the wide range of promos and freebies that this network offers, this amount is already considered as high.

Estimation of hazard probabilities

The sample estimates of hazard probabilities, also known as the marginal hazard probability estimates, are calculated from the event history data. The estimated hazard probability for time j is the number of events that are observed to occur in time period j divided by the total number of subjects at risk in time period j. "Subjects at risk" in this context refers to all those observations in period j that are not censored during period j.

The 20 sample hazard probabilities are plotted by month as shown in Figure 1, suggesting that the marginal hazard function is decreasing from July 2010 to April 2011 but there is a slightly increasing pattern from April 2011 to February 2012 wherein the fluctuations range from 0.03 to 0.06.

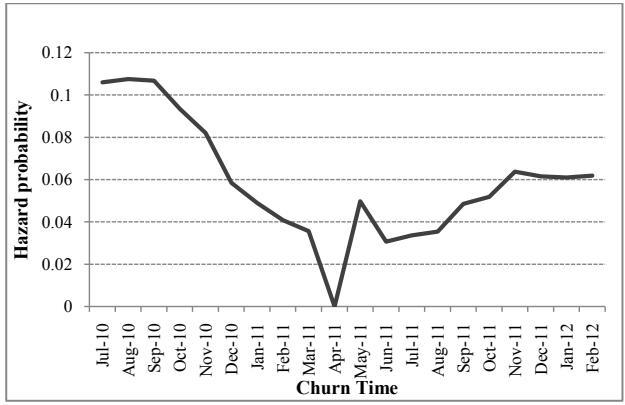

Figure 1. Sample-estimated Hazard Probabilities of Churn

The proportions of subscribers surviving through each month (i.e., the survival probabilities) can be estimated directly from the estimated hazard probabilities. Figure 2 displays the plot of the estimated survival probabilities by month. There is an increase in the proportion of the total subscribers churned over time with almost 70% churned by the end of the 20th month.

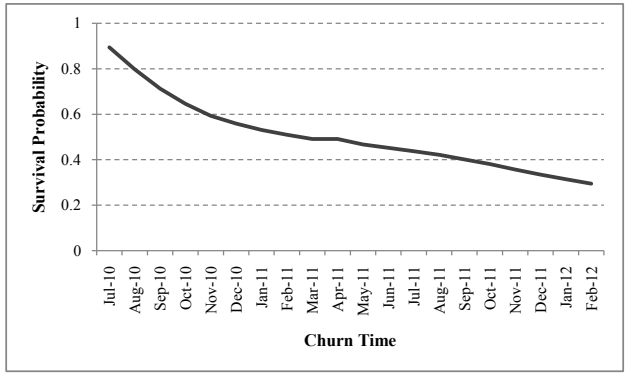

Figure 2. Sample-estimated survival probabilities of churn

Comparison of the parametric model and the modified mixture cured fraction model

The parametric method with covariate and the modified mixture cured model with clustering variable using both the parametric and nonparametric approaches are applied on a telecommunications data set. Customers are grouped according to the mean amount of top-up incurred in the study period. Although the models

include only one covariate, four different indicators were used to determine if there will be differences on the performance of the models.

Table 2. Relative Difference in MAPE of the Parametric and Modified Cured Fraction Model

Covariate	Relative Difference in MAPE		
	Parametric vs. Modified Cox PH (Nonparametric)	Modified Cox PH (Parametric) vs. Modified Cox PH (Nonparametric)	Parametric vs. Modified Cox PH (Parametric)
Average frequency of top-up	26.6409	15.2264	13.4647
Maximum frequency of top-up	31.2911	18.5712	15.6208
Total frequency of top-up	26.6409	15.2264	13.4647
Median frequency of top-up	31.0212	15.6468	18.2262

Regardless of the covariate used, the modified mixture cured models have smaller MAPEs compared to the parametric method. Between the two modified Cox PH models, the nonparametric approach yields smaller MAPEs. It can also be noticed that the MAPEs when the models are applied on the data are even larger than the simulation results when there is misspecification in the model. It may be explained by lack of access to other possible covariates or the existence of random shock on certain periods. Since we are limited by the number as well as the kind of explanatory variables available, data on number of inbound/outbound calls as well as the minutes of use among other variables may help better explain the churning behaviour.

3. Conclusions

The simulation study confirms that clustered survival data can be better characterized by the proposed model. In a perfect scenario (i.e. there is no misspecification error), the modified cured model is superior than its parametric counterpart, relatively robust to varying censoring percentages. With misspecification error, while predictive ability declines, the proposed model still outperforms the parametric model. Results are even better if model is fitted nonparametrically.

The cured model with random clustering effect also performed better (predictive ability) than the parametric counterpart in the application to telecommunication data. Regardless of the covariate used, the proposed model still exhibits lower MAPE and predictive ability is improved if we use the nonparametric approach.

REFERENCES

ARCENEAUX, K. and NICKERSON, D., 2009, Modeling certainty with clustered data: A comparison of methods, *Political Analysis* 17: 177-190.

BOAG, J. W., 1949, Maximum likelihood estimates of the proportion of patients cured by cancer therapy, *Journal of the Royal Statistical Society* 11: 15-44.

CARONI C. and ECONOMOU P., 2012, A hidden competing risk model for censored observations, *Brazilian Journal of Probability and Statistics*.

CHALMETA, R., 2006, Methodology for Customer Relation Management, *The Journal of Systems and Software* 79: 1015-1024.

COLTMAN, 2007, Why build a customer relationship management capability?, *Journal of Strategic Information Systems* 16: 301-320.

CORBIERE, F. and JOLY, P., 2007, A SAS macro for parametric and semiparametric mixture cure models, *Computer Methods and Programs in Biomedicine* 85(2): 173-180.

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

FAREWELL, V. T., 1982, Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics* 14: 257-262.

GAMEL, J. W., WELLER, E. A., WESLEY, M. N., FEUER, E. J., 2000, Parametric cure models of relative and cause-specific survival for group survival times, *Comput. Methods Programs Biomedicine* 61: 99-110.

GLADY, N., BAESENS, B., CROUX, C., 2009, Modeling churn using customer lifetime value, *European Journal of Operational Research* 197: 402-411.

LAMBERT, P. C., 2007, Modeling of the cure fraction in survival studies, *The Stata Journal* 7(3): 351-375.

SANTOS, E. and BARRIOS, E., 2012, Decomposition of Multicollinear Data and Time Series using Backfitting and Additive Models 41:1693-1710.