A CROSS-SECTIONAL SURVIVAL ANALYSIS REGRESSION MODEL WITH APPLICATIONS TO CONSUMER CREDIT RISK

Mercy Marimo [1]
University of the Witwatersrand, South Africa
e-mail: *mercy.marimo@yahoo.com*

Musa Clive Malwandla
University of Cape Town, South Africa
e-mail: *malwandla@live.co.za*

Douw Gerbrand Breed
North West University, South Africa
e-mail: *gerbrand.breed@gmail.com*

Key words: Competing risks, Logistic regression, Loss given default, Proportional hazards.

Abstract: When performing long-range survival estimations, longitudinal survival analysis methods such as Cox Proportional Hazards (PH) and accelerated lifetime models may produce estimates that are outdated. This paper introduces a cross-sectional survival analysis regression model for discrete-time survival analysis. The paper describes a number of variations to the model, including how the model can be used to model competing risks. The model is applied to a portfolio of defaulted loans to estimate the probability of loss. The model's performance is benchmarked against the Cox PH model. Results show that cross-sectional survival analysis performs better than the conventional methods of survival. This is attributable to the fact that the cross-sectional survival method is able to use only the most recent survival information to inform predictions.

1. Introduction

Survival analysis is a general class of techniques used to determine the distribution of time until a target event occurs. The applications of survival analysis techniques is widely varied. In actuarial science, survival analysis is applied to measure lifetime distribution of insured lives (Wetterstrand, 1981). In finance, survival analysis was used to derive the Gaussian copula function for dependent risk (Li, 2000). In operations research, survival analysis is commonly applied to measure the reliability of appliances and systems (Peña and Hollander, 2004). In medicine, survival analysis is used

[1] Corresponding author.

AMS: 62P05, 62N02, 91G40

to measure the success of drug trials (Knaus et al., 1993). In consumer credit risk survival analysis is used to model default rates and loss given default (Bellotti and Crook, 2009).

There are a number of different types of survival analysis methods in common application. Parametric survival analysis is one of the most tamed of these. This requires a parametric form for the waiting-time distribution to be assumed, so that the analysis reduces to the estimation of the parameters of this assumed distribution. This form of survival analysis is relatively inflexible to data, only being reliable when the nature of the distribution is well-understood. Otherwise, the analysis suffers from the risk of the assumed parametric form being wrong.

An accelerated life (or accelerated failure time) survival model is a special class of parametric survival methods where occurrence of the event of interested is modelled as being accelerated or decelerated by the presence of certain risk factors (Wei, 1992). In this type of model, the waiting-time distribution function of all members of the population is found as a horizontal shift of some baseline distribution function. Due to this strict restriction, this form of survival analysis is only applicable to a selected number of cases.

One of the most common type of survival analysis is proportional hazard survival analysis. Here it is assumed that the rate of occurrence (or hazard rate) between any two members of the population is proportional at all survival times, e.g., the j^{th} subject of the study is always r times riskier than the i^{th} subject at all survival times. Cox regression is the most popular type of proportional hazard regression (Cox, 1972). It owes its popularity to the fact that is allows the hazard function to be specified semi-parametrically.

Cox regression is widely applied in almost all applications of survival analysis. Bellotti and Crook (2009) apply survival analysis to the analysis of risk in consumer loans. Petrie et al. (2002) apply Cox regression in studying the effect of intervention on recovery following myocardial infarctions. Chollet et al. (2002) apply Cox regression to study induction chemotherapy in operable breast cancer patients.

Although widely applied, Cox regression has some weaknesses. Firstly, the technique is unwieldy when the covariates are allowed to vary with survival time. The second concern with Cox regression is that it is fitted with the presumption that the baseline hazard remains unchanged over time. Although this is usually not a problem in typical applications, it stands to compromise the validity of the model when the application is over a longer term. The study conducted by Strauss, Shavelle and Ashwal (1999) provides a good example of the effects of this issue. Their study was on the mortality of humans in permanent vegetative state. They found marked increases in infant (2 years old and younger) life expectancy between 1981 and 1996. Since no similar effect was found in older age groups, this study would invalidate the assumption of a constant baseline hazard.

A closely-related issue to this is the fact that Cox regression estimated the baseline hazard on a longitudinal basis. Certain changes in survival rates are observed easier through cross-sectional analysis than longitudinal analysis. The study of Strauss et al. (1999) is an example of these. Furthermore, longitudinal analysis is less suitable for left-censored data, while cross-sectional analysis is usable under both left and right censoring.

Having accounted for these weaknesses, Cox regression is still one of the most practical forms of survival analysis. This is mainly due to its intuitiveness and flexibility to data. Strauss, Shavelle, DeVivo and Day (2000) attempt to overcome some of the weaknesses of Cox regression through logistic regression. In this paper we continue in this theme and offer a different form of logistic regression, as an alternative to Cox regression.

The paper describes how logistic regression can be used to model survival probability using time-varying covariates. By introducing an offset variable reflecting the baseline survival probability, we show how logistic regression is able to model survival probabilities in a population exposed to multiple decrements. Furthermore, by estimating the baseline survival probability on a cross-sectional basis, we show that the resultant probability is more reactive to changes in the nature of the survival distribution.

The remainder of the paper is organised as follows. Section 2 discusses the theoretical merits of a complementary-log-log survival model over standard survival analysis, and offers ways in which such a model can be extended under different situations. Sections 3 and 4 present a practical comparison of the model to Cox regression and Section 5 concludes with a review of the findings.

2. Complementary Log Log Survival Model

2.1. Logistic Regression

Logistic regression is one of the most popular modelling techniques applied for modelling binary outcomes. It produces a model that relates the transformed probability of the target outcome to a linear function of a set of covariates.

Let Y_k be a Bernoulli random variable representing the outcome of the k^{th} trial out of a set of n trials. Let $X_k = \{X_{k,1}, X_{k,2},, X_{k,p}\}$ be a vector of covariates associated with the outcome Y_k. Logistic regression is interested in modelling $\rho_k = P[Y_k = 1]$ as a function of X_k. The canonical link function of the Bernoulli distribution is the logit function, so that the natural form of the logistic regression model is as follows:

$$\ln\left(\frac{\rho_k}{1-\rho_k}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{k,j},$$

where $\beta = \{\beta_1, \beta_2, \ldots, \beta_p\}$ is the vector of parameters of the model. Other common link functions for logistic regression models include the probit function and the complementary log-log (CLL) function.

For a general link function $g^{-1}(x)$, we have the following model for ρ_k:

$$\rho_k = g\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{k,j}\right).$$

Therefore, we can estimate $\boldsymbol{\beta}$ by maximising the following log-likelihood function with respect to $\boldsymbol{\beta}$:

$$l(\boldsymbol{\beta}) = \sum_{k=1}^{n}\left[Y_k g\left(\beta_0 + \sum_{j=1}^{p}\beta_j X_{k,j}\right) + (1-Y_k)\left(1 - g\left(\beta_0 + \sum_{j=1}^{p}\beta_j X_{k,j}\right)\right)\right].$$

The maximisation can be accomplished numerically. For a general discussion on the application and interpretation of logistic regression, see Lottes, DeMaris and Adler (1996).

Complementary Log Log Link Function with an Offset

Logistic regression with the CLL link function yields the following model:

$$\ln\left(-\ln\left(1-\rho_k\right)\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{k,j}.$$

Therefore, the model yields the following expression for ρ_k:

$$\rho_k = 1 - e^{-e^{\beta_0+\Sigma_{j=1}^{p}\beta_j X_{k,j}}}.$$

Suppose that, in addition to X_k, we have prior estimates relating to the value of ρ_k within the population. Let $\pi_k \in (0,1)$ be a baseline estimate of ρ_k, satisfying the following property:

$$\rho_k = 1 - \left[1-\pi_k\right]^{e^{\beta_0+\Sigma_{j=1}^{p}\beta_j X_{k,j}}}.$$

The baseline estimate π_k can be treated as another covariate of the model. However, in order to force the relationship above, we include π_k into the model as $\ln\left(-\ln\left(1-\pi_k\right)\right)$ and force a parameter estimate of one, as follows:

$$\ln\left(-\ln\left(1-\rho_k\right)\right) = \ln\left(-\ln\left(1-\pi_k\right)\right) + \beta_0 + \sum_{j=1}^{p} \beta_j X_{k,j},$$

i.e., we fit a CLL logistic regression model, with $\ln\left(-\ln\left(1-\pi_k\right)\right)$ included as an offset variable. This simplifies to produce the following model for ρ_k:

$$\rho_k = 1 - \left[1-\pi_k\right]^{e^{\beta_0+\Sigma_{j=1}^{p}\beta_j X_{k,j}}}.$$

The use of the offset variable to produce survival probabilities is the core idea that is contributed by this paper. The following section develops how this idea is used under different situations where survival analysis is required.

2.2. Survival Models

The Basic Model

The form of the logistic regression model described above is ideal for modelling survival probabilities. Suppose we are given a set of n subjects of a survival study. For each subject k, we are interested in modelling the following binary outcome:

$$Y_k\left(t,t+s\right) = \begin{cases} 0, & \text{if subject survives to time } t+s, \text{ given the subject survived to time } t \\ 1, & \text{if subject has experienced the target event by time } t+s, \text{given survival to time } t \end{cases}$$

Define $\rho_k\left(t,t+s\right)$ to be $P\left[Y_k\left(t,t+s\right)=1\right]$ and $\pi\left(t,t+s\right)$ to be the prior estimate for $\rho_k\left(t,t+s\right)$. Note here that the estimate $\pi\left(t,t+s\right)$ is the same for all k, i.e.., it is a baseline estimate. Using the CLL model, $\rho_k\left(t,t+s\right)$ is as follows:

$$\rho_k\left(t,t+s\right) = 1 - \left[1-\pi\left(t,t+s\right)\right]^{e^{\beta_0+\Sigma_{j=1}^{p}\beta_j X_{k,j}(t)}},$$

where $X_k(t) = \{X_{k,1}(t), X_{k,2}(t), \ldots, X_{k,p}(t)\}$ is a vector of categorical covariates of subject k. For simplicity, we assume all variables are categorical. This allows that $\rho_k(t,t+s) = \pi(t,t+s)$ when $\beta_0 + \sum_{j=1}^{p} \beta_j X_{k,j} = 0$, i.e., $\pi(t,t+s)$ is set to equal $\rho_k(t,t+s)$ for a chosen reference sub-population.

In order to fit the model, $\pi(t,t+s)$ needs to be estimated from the observed experience of the baseline population. One approach for doing this uses the Kaplan-Meier estimator for the survival function (Kaplan and Meier, 1958). Let $S_0(t)$ be the probability that a subject from the baseline population survives to time t. The Kaplan-Meier estimator for $S_0(t)$ is as follows:

$$\widehat{S}_0(t) = \prod_{j=1}^{t} \left(1 - \frac{D_0(t)}{E_0(t)}\right),$$

where $D_0(t)$ is the number of subjects in the observed experience that experienced the event at time t and $E_0(t)$ is the number of subjects yet to experience the event at time t, excluding any subjects that were censored before time t. The hazard function associated with $\widehat{S}(t)$ is $\widehat{h}(t) = \frac{D_0(t)}{E_0(t)}$. The baseline probability can then be estimated as:

$$\pi(t,t+s) = \frac{\widehat{S}_0(t+s)}{\widehat{S}_0(t)}.$$

Therefore, the probability that a subject with covariate vector x at time t survives to time $t+s$ is modelled as:

$$\rho(t,t+s,x) = 1 - [1 - \pi(t,t+s)]^{e^{\beta_0 + \Sigma_{j=1}^{p} \beta_j x_j}}.$$

Notice that the current survival time t and the length of the event horizon s are only featured in the model formula through $\pi(t,t{+}s)$. In other words, covariate vector x only adjusts for non-temporal risk effects. Therefore, for the model to be valid, the following must hold:

$$e^{\beta_0 + \Sigma_{j=1}^{p} \beta_j x_j} = \frac{\ln[1 - \rho(t,t+s,x)]}{\ln[1 - \pi(t,t+s)]},$$

i.e., the ratio $\frac{\ln[1-\rho(t,t+s,x)]}{\ln[1-\pi(t,t+s)]}$ must be constant for all t and s, for a given x. We can test this empirically by plotting the following ratio for different values of t for a given x:

$$r(t,s,x) = \frac{\ln\left[1 - \frac{\widehat{S}(t+s,x)}{\widehat{S}(t,x)}\right]}{\ln\left[1 - \frac{\widehat{S}_0(t+s)}{\widehat{S}_0(t)}\right]},$$

where $\widehat{S}(t,x)$ is the Kaplan-Meier estimate of the survival function in the population with covariate vector x. The plot of $r(t,s,x)$ against either t or s should produce a flat line in order for the model to be valid. This test is analogous to the test for the proportional hazards assumption under Cox regression.

The Competing-Risk Model

Competing risk survival analysis is concerned with estimating the waiting-time distribution in the case where the subjects of the study are exposed to more than one decrement. A decrement in this case is defined as the predetermined event of interest. A common approach to competing-risk survival analysis is the Cumulative Incidence Curve (CIC) approach (see Lin, 1997). This involves developing separate survival models for each decrement (in each case treating the other decrement as a type of censorship) and combining the models using the CIC.

Let $h_{[i]}(t)$ be the hazard function for the i^{th} decrement of the study. Define $S(t)$ to be the overall survival function of the study, taking into account all the decrements, and $I_{[i]}(t)$ to be the incident function for the i^{th} decrement. The survival function $S(t)$ can be computed recursively from:

$$S(t) = S(t-1) - \sum_i I_{[i]}(t-1),$$

with the initial condition $S(0) = 1$. Values for $I_i(t)$ are computed from $S(t)$ as follows:

$$I_{[i]}(t) = h_{[i]}(t)\,S(t)\,.$$

The probability of surviving decrement i until $t+s$, given survival to t, is then computed as follows:

$$\rho_{[i]}(t,t+s) = \sum_{u=t+1}^{t+s} \frac{I_{[i]}(u)}{S(t)}.$$

We call $p_{[i]}(t,t+s)$ the *forward-looking* survival probability for decrement i. Therefore, the overall survival probability $S(t)$ can also be written as

$$S(t) = 1 - \sum_i \rho_{[i]}(t,t+s).$$

The CLL model can be adapted for competing risk survival analysis by using the CIC approach. Suppose, as before, we are given a set of n subjects of a survival study. For each subject k, we are interested in modelling the following binary outcome:

$$Y_{k,[i]}(t,t+s) = \begin{cases} 0 & \text{if subject survives decrement } i \text{ to time } t+s, \text{ given the subject survived to time } t \\ 1 & \text{if subject has experienced decrement } i \text{ by time } t+s, \text{ given survival to time } t \end{cases},$$

for each decrement i in the study.

Define $\rho_{k,[i]}(t,t+s)$ to be $P\left[Y_{k,[i]}(t,t+s) = 1\right]$ and $\pi_{[i]}(t,t+s)$ to be the baseline estimate for $\rho_{k,[i]}(t,t+s)$. Therefore, the probability that a subject with covariate vector x at time t survives decrement i to time $t+s$ is modelled as:

$$\rho_{[i]}(t,t+s,x) = 1 - \left[1 - \pi_{[i]}(t,t+s)\right]^{e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_j}}.$$

As before, $\pi_{[i]}(t,t+s)$ corresponds to the baseline population (i.e., $x = 0$), so that $\pi_{[i]}(t,t+s) = \rho_{[i]}(t,t+s,0)$. Therefore, we can estimate $\pi_{[i]}(t,t+s)$ through the CIC approach, as $p_{[i]}(t,t+s)$ for the population $x = 0$.

The probability of not surviving to time $t+s$ is the sum of the probability of exiting due to each decrement, expressed as $\sum_i \rho_{[i]}(t,t+s)$. Therefore, building a separate logistic regression model for each decrement may lead to negative survival probabilities, since $\sum_i \rho_{[i]}(t,t+s)$ may exceed one. Polytomous logistic regression can be used to overcome this potential difficulty (Engel, 1988).

The Long-Term Competing Risk Model

So far we have discussed how to build a simple CLL survival model and seen how this model can be generalised to competing risk survival analysis through the CIC approach. We now consider the case when the model is used to make long-term predictions.

The estimation of the survival function can be done longitudinally or through cross-sectional approaches. The Kaplan-Meier estimator is an example of a longitudinal estimator for $S(t)$. Cross-sectional estimation can be done by considering the number of subjects that have been exposed to each decrement for t periods at the beginning of period T, $n_{t,T}$, and the number of subjects experiencing decrement i between period T and period $T+1$, $d_{t,[i],T}$, for each decrement i *(note that T corresponds to a calendar period, such as January 2015 for instance, while t corresponds to survival period, such as t months since birth).* We then estimate the hazard function for period T and decrement i as follows:

$$h_{[i],T} = \frac{d_{t,[i],T}}{n_{t,T}}.$$

Therefore, we can compute an estimate of the decrement i survival function from period T cross-sectional data as follows:

$$\widehat{S}_{[i],T}(t) = \prod_{j=1}^{t} \left(1 - \frac{d_{t,[i],T}}{n_{t,T}}\right).$$

Furthermore, in order to make survival predictions during period T, we could use $\widehat{S}_{[i],T-1}(t)$ which would be based on the latest cross-sectional information available. In certain instances, a cross-sectional estimator would be expected to perform better than a longitudinal estimator. Notably, in cases where the nature of survival risk is affected by period effects (such as calendar month effects) rather than cohort effects (and said effect transitions smoothly from one period to the next), $\widehat{S}_{[i],T-1}(t)$ should be the most relevant estimator of the survival rates of the subjects under observation during period $T+1$.

As a final extension to the CLL model, consider a set of n_T subjects of a survival study, for periods $T=1$ to m. For each subject k, we are interested in modelling the following binary outcome:

$$Y_{k,[i],T}(t,t+s) = \begin{cases} 0, & \text{if subject survives decrement } i \text{ to time } t+s, \text{ given survival} \\ & \text{to time } t \text{ during period } T \\ 1, & \text{if subject experienced decrement } i \text{ by time } t+s, \text{ given survival} \\ & \text{to time } t \text{ at period } T. \end{cases}$$

Define $p_{[i],T}(t,t+s)$ to be the forward-looking probability based on the cross-sectional estimator $\widehat{S}_{[i],T}(t)$. Therefore, the probability that a subject with covariate vector x at time t during period T survives decrement i to time $t+s$ is modelled as:

$$\rho_{[i],T}(t,t+s,x) = 1 - \left[1 - \pi_{[i],T}(t,t+s)\right]^{e^{\beta_0 + \Sigma_{j=1}^{p} \beta_j x_j}},$$

where $\pi_{[i],T}(t,t+s)$ is estimated as $p_{[i],T}(t,t+s)$ for the baseline population $x=0$.

Note, the essential difference between the model $\rho_{[i],T}(t,t+s,x)$ and $\rho_{[i]}(t,t+s,x)$ is that the former allows the baseline to change over time and should produce estimates that are more reactive to changes in survival behaviour.

2.3. The Model in Practice

The CLL model applied to cross-sectional data is ideal for modelling when the prediction horizons, represented by s in the above discussion, are long. The process involved in fitting the model can be summarised as follows:

1. For each subject in the study, create target variables $Y_{k,[i],T}(t,t+s)$ for each decrement i.
2. Select a baseline population (where $X_{k,T}=0$, through dummy coding). This may be selected to be the largest population in the state space of $X_{k,T}$, the covariate vector of the k^{th} subject during period T, as this would lead to greater credibility in the estimation of the baseline.
3. For each period T and decrement i, calculate the cross-sectional survival function estimator $\widehat{S}_{[i],T}(t)$. Use this to compute the forward looking probabilities $p_{[i],T}(t,t+s)$.
4. Compute the ratio $r(t,s,x)=\ln\left[1-\frac{\widehat{S}_{[i],T}(t+s,x)}{\widehat{S}_{[i],T}(t,x)}\right]/\ln\left[1-p_{[i],T}(t,t+s)\right]$ for each x in the state space of $X_{k,T}$, where $\widehat{S}_{[i],T}(t,x)$ is the period T cross-sectional (decrement i) survival function estimator for the population with covariate vector x. Ensure that $r(t,s,x)$ is a constant function of t and s by modifying and removing variables where appropriate.
5. Set $\pi_{[i],T}(t,t+s)=p_{[i],T}(t,t+s)$ and apply the CLL function to create an additional covariate $b_{k,[i],T}=\ln\left(-\ln\left(1-\pi_{[i],T}(t,t+s)\right)\right)$.
6. Apply polytomous logistic regression with the CLL link function with $b_{k,[i],T}$ set as an offset variable and X_k as normal covariates.

The CLL model, when used to make survival prediction over long horizons, has the following advantages over Cox regression:

1. The model retains tractability when covariates are allowed to vary with time.
2. The model is more reactive to gradual (non-cohort specific) changes in the nature of the baseline risk, since $\widehat{S}_{[i],T}(t)$ is based on the latest cross-sectional data.
3. The model easily allows for time-varying covariates, e.g., in a mortality study, we may include factors that vary with time, such as income, which is not always easy to do via Cox regression.

Additionally, the model is arguably as simple as Cox regression and is able to handle competing risks in a similar way. The main disadvantage of the CLL model in general is that its assumptions are more difficult to test, compared with the proportional hazard assumption in Cox regression. Unlike Cox regression, the CLL model is a discrete-time survival model. It is thus unable to produce estimates in continuous time. Where estimates are required in continuous time, this may be seen as a drawback to the model.

In practice, there are alternatives to the CLL model form. For instance, the CLL link function can be replaced with a probit link function to produce:

$$\Phi^{-1}\left(\rho_{[i],T}(t,t+s,x)\right)=\beta_0+\Phi^{-1}\left(\pi_{[i],T}(t,t+s)\right)+\sum_{j=1}^{p}\beta_j x_j.$$

3. Application: Credit Loss Given Default Modelling

In this section, the CLL model is applied to model loss and cure probabilities in defaulted mortgage loans. Defaulted accounts from this portfolio were observed between October 2006 and November 2014. The aim of the study was to estimate the probability that an account is written off or *cures* within h months of default.

The life of a loan is modelled as a four-state stochastic process, with the following states:

1. Performing: consists of accounts that are deemed to still be performing, broadly, within the parameters of the loan contract.
2. Default: consists of accounts that are performing outside of the parameters of the loan contract.
3. Write-Off: consists of accounts where the issuer of the loan deems it unlikely to receive any further repayments on the loan.
4. Closed: consists of accounts where the loan has been fully repaid.

3.1. Definition of Decrements or Events

An account that is in the default state can either proceed to get written off or return to performing, depending on the behaviour of the borrower. *Write off* or *loss* decrement occurs when there are no prospects of further repayments on the loan as defined by the issuer. A *cure* decrement is defined as the transition from default to performing. This study is interested in modelling the probability of write off and the probability of cure within the first $h = 60$ months of default. The study thus aims to produce an estimate for the probability that an account k observed in calendar month T with covariate vector $X_{k,T}$ in its t^{th} month of default will cure within $h-t$ months, $\rho_{[c],T}\left(t,h,X_{k,T}\right)$, and the probability that the account will be written off within $h-t$ months, $\rho_{[w],T}\left(t,h,X_{k,T}\right)$.

In credit risk analysis, the probability of write-off and the probability of cure are inputs into the calculation of loss-given-default (LGD), as follows:

$$LGD = \rho_{[c],T}\left(t,h,X_{k,T}\right) \times E\left[\text{Loss}|\,\text{Cure}\right] + \rho_{[w],T}\left(t,h,X_{k,T}\right) \times E\left[\text{Loss}|\,\text{Write-off}\right],$$

where $E\left[\text{Loss}|\,\text{Cure}\right]$ is the expected loss from re-default following a cure and $E\left[\text{Loss}|\,\text{Write-off}\right]$ expected write-off amount given write-off.

3.2. Model Fitting

A CLL survival model was fitted for both the cure decrement and the write-off (loss) decrement. The model development process summarised in Section 2.3 was followed to construct the CLL model. Figure 1 illustrates the way in which the assumption of the model was assessed, as per Section 2.3., for two variables in the cure model. We require each variable to produce a constant ratio across the time dimension — this requirement is only approximately met.

The variables selected for the model, along with the parameter estimates, are contained in Table 1.

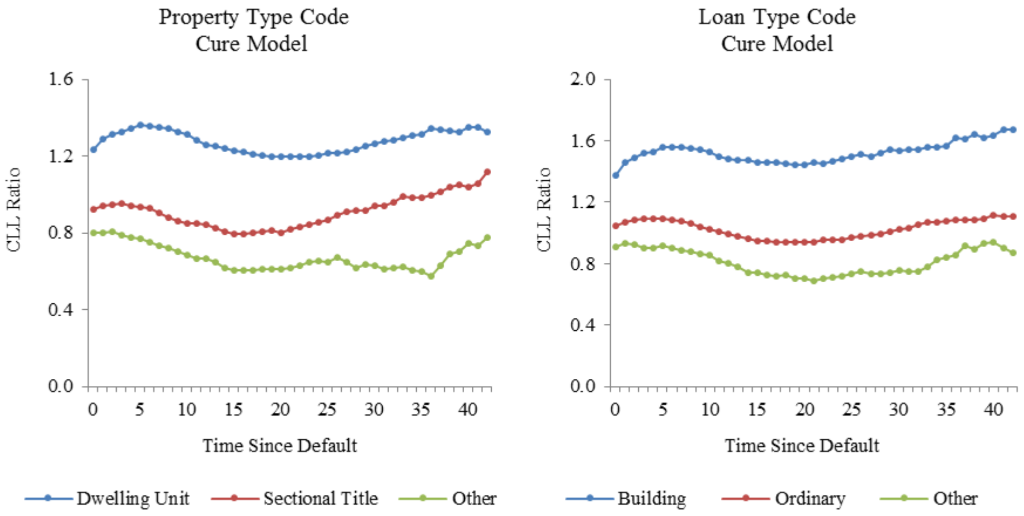

Figure 1: Checking the assumptions of the CLL model.

Table 1: CLL survival model parameter estimates.

CLL Survival Model					
Cure Estimates			**Loss Estimates**		
Variable	**Class**	**Estimate**	**Variable**	**Class**	**Estimate**
Legal Type Code	Collections	1.0323	Legal Type Code	Legal Action	1.4599
	Other	0.0000		Other	0.0000
LTV	<70%	0.0000	Balance	≤R60,000	-0.7774
	≥ 70%	-0.2559		>R60,000	0.0000
Property Type Code	Dwelling Unit	0.0000	Loan Type Code	Ordinary, Building	-0.3404
	Sectional Title	-0.2336		Other	0.0000
	Other	-0.5119			
Loan Type Code	Building	0.2948	Collateral Code	No Collateral	0.0000
	Ordinary	0.0000		Collateral, Surety	0.3495
	Other	-0.2688			

A Cox survival model was also fitted for both the cure decrement and the loss decrement, using the CIC approach. Table 2 contains the variables included in the model, along with parameter estimates. Although the two models are developed on the same data set, we see that the Cox regression model ends up with less covariates.

Table 2: Cox model estimates.

Cox Survival Model					
Cure Estimates			**Loss Estimates**		
Variable	**Class**	**Estimate**	**Variable**	**Class**	**Estimate**
Loan-to-Value (LTV) Ratio	<70%	0.0000	Balance	$\leq$ R60,000	0.0000
	[70% - 80%)	-0.3115		>R60,000	0.8329
	[80% - 90)	-0.5132			
	90+	-0.7447			
Loan Type Code	Further Advanced	0.1514			
	Other	0.0000			
Property Type Code	Dwelling Unit	0.0000			
	Sectional Title	-0.0413			
	Other	-0.2917			

3.3. Model Comparisons

Several metrics were used to compare the performance of the two survival model. These include the Receiver Operating Characteristic (ROC) curves, Area Under the Curves (AUC), accuracy across range and accuracy over time (Mair, Reise and Bentler, 2008).

The ROC Curves: The ROC test plots the sensitivity against 1-specificity of the models at various cut-off values of risk. For the cure event, sensitivity refers to a fraction of cured accounts that the model correctly identifies as cured. The same goes for the loss event. Specificity refers to a fraction of non-cured accounts that the model correctly identifies as non-cured. The ROC curves for the CLL and the Cox regression model are provided in Figure 2 for the cure and loss models. For both decrements, the CLL model performs better than the Cox model.

AUC and Model Gini Statistics: Another metric that can be used to compare the performance of different models is the AUC. It quantifies the overall ability of the model to discriminate between those accounts that eventually cure and those that never cure. A completely random model (one no better at identifying true positives than flipping a coin) has a theoretical AUC of 50%. A perfect model has an AUC of 100%. In addition, the overall model Gini statistics were calculated for each

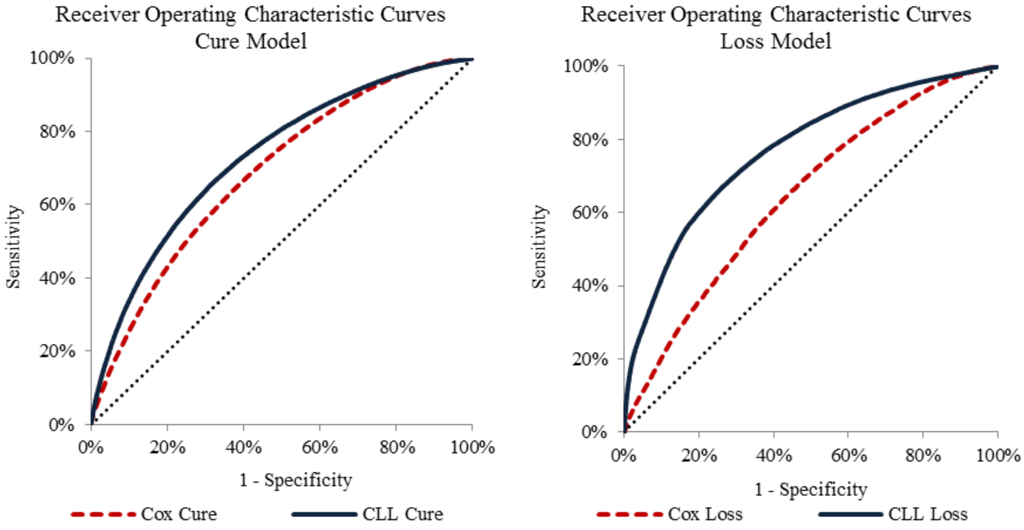

Figure 2: Receiver Operating Characteristic curves for loss and cure models.

model as generalised metrics to measure the ability of the models to differentiate risk. Results are given in Table 3.

Table 3: AUC and model Gini statistics.

Area Under the ROC Curves and Model Gini Statistics			
Model	**Variant**	**AUC**	**Gini Statistic**
Cure	Cox	70.28%	40.56%
	CLL	71.75%	43.49%
Loss	Cox	68.83%	37.67%
	CLL	77.09%	54.19%

The CLL survival model has improved discriminatory power over the Cox survival model. This is particularly the case for the loss decrement, where the Gini statistic for the CLL model is almost 17% greater than that of the Cox model.

Model Accuracy: To test across range, accounts were ranked separately for each model into 10 risk groups, based on their respective cure and loss probabilities. For each risk group the actual empirical probabilities and the corresponding expected values produced by the models were determined. A scatter-plot of these are provided in Figure 3. For the models to be accurate, scatter points should not deviate significantly from the 45 degree diagonal. In both models, the CLL model scatter points lie closer to the 45 degree than the Cox model scatter points.

The model accuracy was also assessed by prediction horizon, referred to as the workout period,

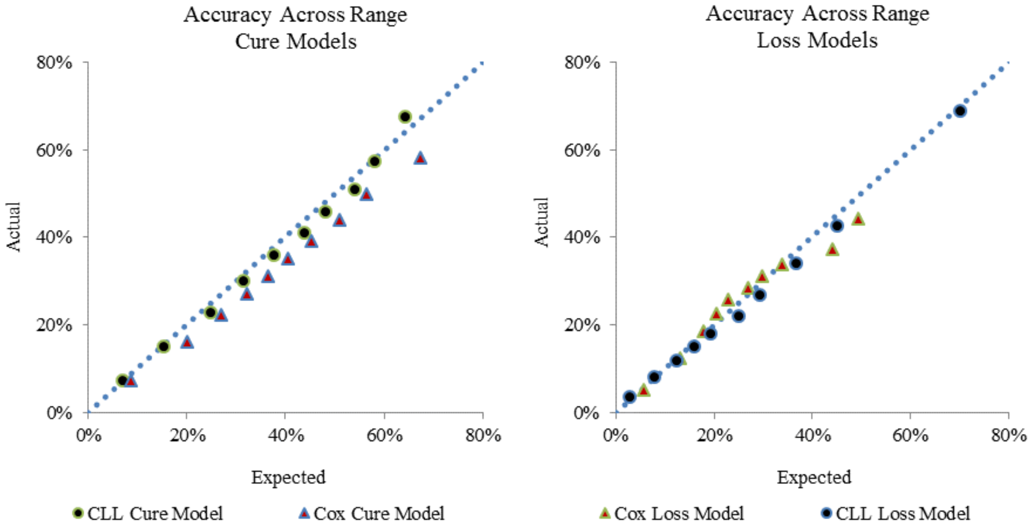

Figure 3: Accuracy across range.

which is given by $h-t$ in $\rho_{[i],T}\left(t,h,X_{k,T}\right)$. This assesses how well the model performs at different prediction horizons. Plots of this assessment are shown in Figure 4, where the CLL model tends to perform better than the Cox model.

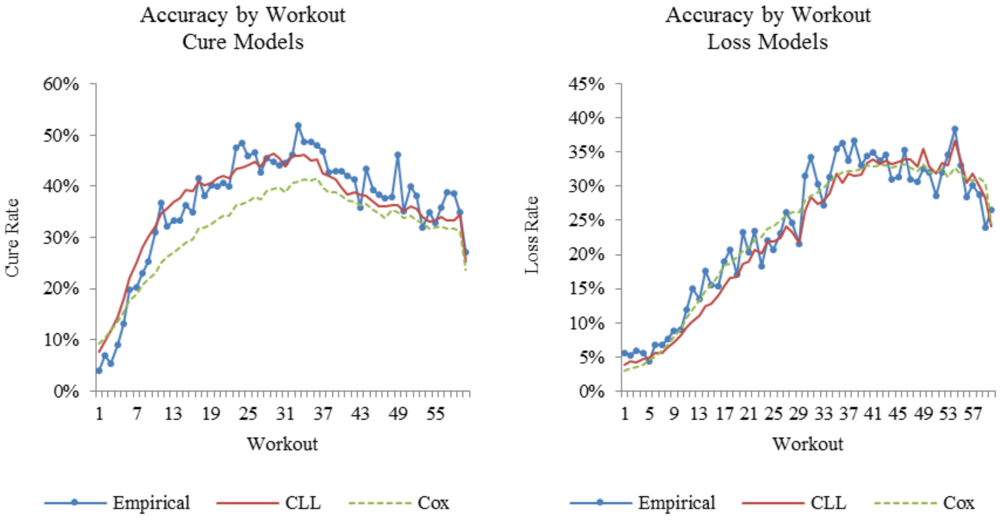

Figure 4: Accuracy by workout.

4. Application under Cyclical Conditions

4.1. Stock Market Analysis

In this section we consider how the CLL may be used for technical and fundamental analysis. For this, we reduce the task of stock selection to the task of modelling the probability that the price of a given share exceeds its current value by more than 10% within some chosen horizon T (an up-probability) in comparison to the probability that a share price falls below 90% of its current value within the same horizon (a down-probability). The *up* and *down* events are treated as competing risks.

If S_t is the price of the share at time t, the up-probability can be determined by modelling the waiting time until the share price exceeds a value of $S_0 \times 1.10$. Similarly, the down-probability is determined from the distribution of time until the share price falls below a value of $S_0 \times 0.90$. If it is assumed that S_t follows a geometric Brownian motion both these waiting time distributions are known to be inverse Gaussian (Lee and Whitmore, 2006). However, it is also understood that securities may exhibit more extreme movements than those suggested by the Gaussian distribution, i.e., the inverse Gaussian distribution may not always be appropriate.

Consider now a derivative contract placed on the share under which a payment is made on the first instance of the two possible events: the price rising above $S_0 \times 1.10$ (an up-movement) and the price falling below $S_0 \times 0.90$ (a down-movement).

Both the situations described above can be modelled using the CLL model: the former being a single decrement survival analysis and the latter requiring the competing risk version of the CLL.

A competing risk CLL model and a competing risk Cox regression model were fitted to estimate the probability of an up-movement in competition with a down-movement, on a set of selected daily time series of share prices from the Johannesburg Stock Exchange, over the period from January 2000 to December 2014. The two models were assessed by using them to estimate the competing probability of an up-movement and that of a down-movement over a 30 day horizon. Figure 5 summarises the model performance.

The graphs show that the CLL model performs significantly better than the Cox regression model. This is unsurprising, since the CLL model includes the most recent price performance as inputs into the estimation of the waiting time distributions. In modelling something that is as cyclical as share price, the CLL clearly outperforms the conventional Cox regression model, with an R-Squared (coefficient of determination) differential of 85.57% and 76.58% in the up and down models respectively. This is significantly better than the R-squared values obtained by the Cox model, as can be seen in Table 4.

Table 4: R-squared values of Cox vs. CLL.

	R-square Cox	R-square CLL
Ups	0.41%	85.98%
Downs	3.31%	79.89%

Although the CLL model performs better than the Cox model, we note that CLL model has a

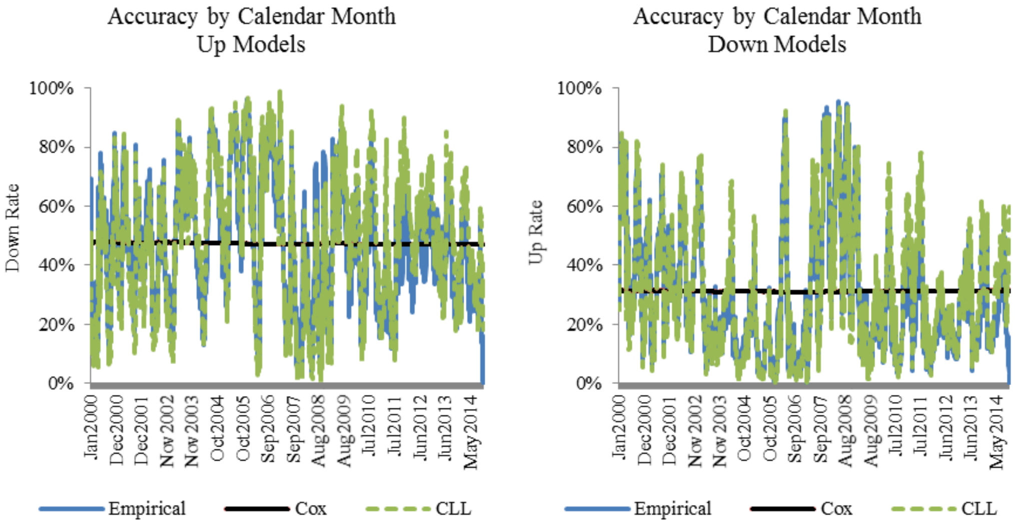

Figure 5: Accuracy by month.

lagged reaction to the actual price performance. This is to be expected since the model does not include any forward-looking information. Such information may be included into the model by adjusting the baseline for expected changes in price performance, or including forward-looking indicators as covariates, e.g. including forecasts of Gross Domestic Product and sectorial performance forecasts. We also note the implications of the Efficient Markets Hypothesis, which (in its semi-strong form) predicts that it is impossible to statistically outperform the market through technical analysis (Fama, 1970). If this assumption holds, then the CLL model is only economically useful when forward-looking information or adjustments to the baseline performance are included.

4.2. Extended Stock Market Analysis

As an extension to the analysis carried out on stock analysis, we consider a single-decrement simulation study of the model. We are concerned with investing how the CLL model performs against the Cox regression model under different conditions of cyclicality.

A sample of 20,000 subjects were simulated as entrants into the population at each time from time 1 to time 180. The purpose of the study was to model the probability that the subject survive to time 12. It was assumed that subject k aged t during time T will be exposed to the following hazard rate:

$$h(t,X,T) = \left(\frac{1}{4}t^{\frac{1}{4}}\right)^{a_0+a_1X+a_2\sin\left(\frac{T}{k}+e_T\right)},$$

where X and e_T are normal random variables. As k influences the cyclicality of the experience, the simulation was carried out for different values of k. The study involved using the cohort of subjects that entered the population between time $t-12-x$ and time $t-12$ to test the model of the experience of the cohort of subjects entering between time t and time $t+12$, i.e., we are assuming that the shelf-life of the model is 12 periods, after which a new model is developed on the latest 12 cohorts for which performance experience is available for x periods. Figure 6 shows the results of the simulation study.

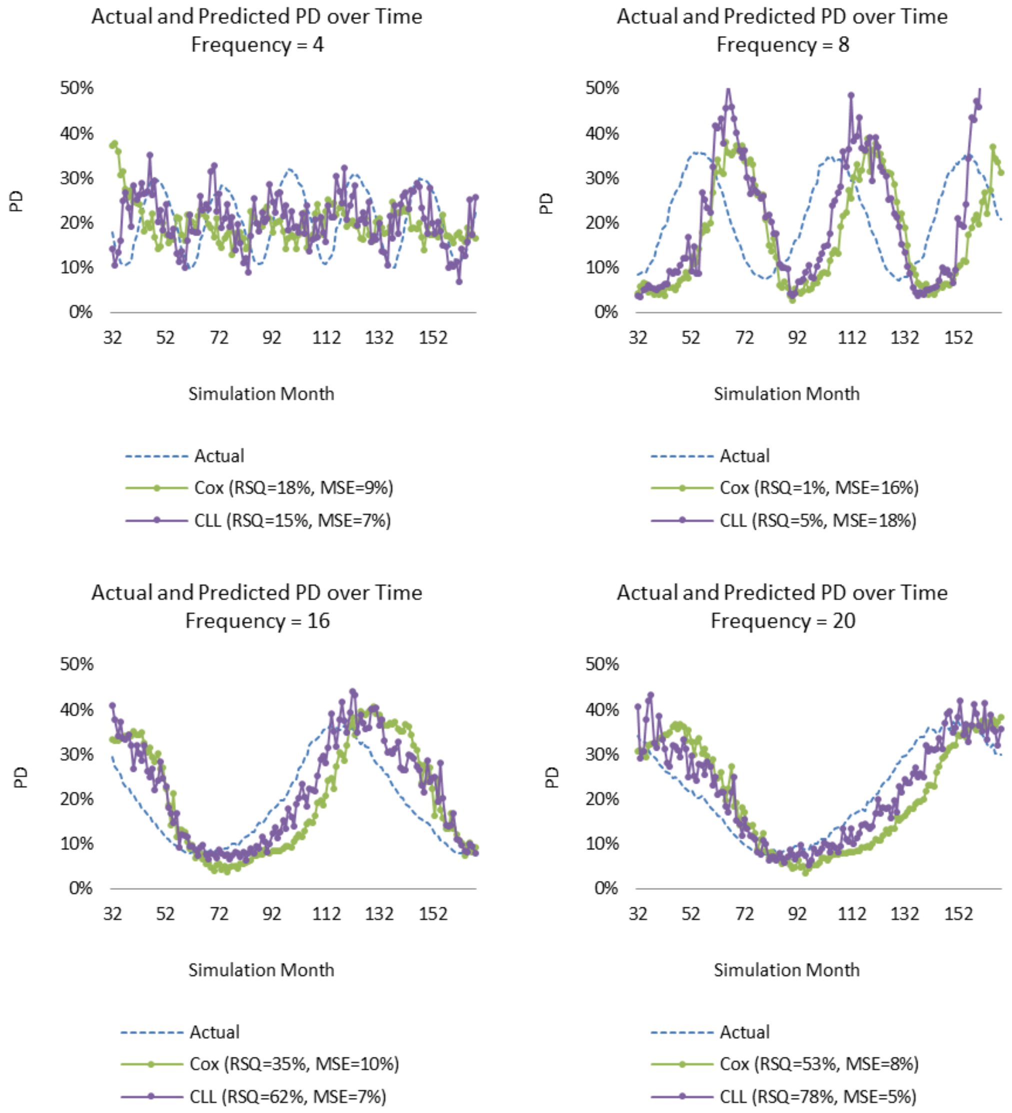

Figure 6: Accuracy by month.

From the graphs, it is evident that the CLL has faster reactivity time than the standard Cox regression model at different frequencies. This is seen by looking at both the coefficient of determination and the mean-squared-error of the predicted rate.

5. Discussion

The paper described a logistic regression approach to survival analysis, where three variations were described:

1. The approach can be conducted with either a single or multiple decrement, the latter of which is done via polytomous logistic regression.
2. The baseline can be estimated either longitudinally or via cross-sectional analysis, which could improve the reactiveness of the model.
3. The link function can be specified in a number of ways: the model discussed here uses the CLL link function, but the probit model was also given as an example..

The model's workings were illustrated on a loan portfolio, where models for probability of loss and cure from default were developed. It was found that the CLL model tends to perform better than the Cox model on the portfolio. A limited attempt was made to discuss the theoretical merits of the CLL over other forms of survival analysis. Further research could look into applications of this approach in different domains of study, such as medical research, social psychology, engineering, agriculture and other fields where pre-determined events can be observed and tracked over discrete time periods.

References

BELLOTTI, T. AND CROOK, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, **60** (12), 1699–1707.

CHOLLET, P., AMAT, S., CURE, H., DE LATOUR, M., LE BOUEDEC, G., MOURET-REYNIER, M., FERRIERE, J., ACHARD, J., DAUPLAT, J., AND PENAULT-LLORCA, F. (2002). Prognostic significance of a complete pathological response after induction chemotherapy in operable breast cancer. *British Journal of Cancer*, **86** (7), 1041–1046.

COX, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*, **34** (2), 187–220.

ENGEL, J. (1988). Polytomous logistic regression. *Statistica Neerlandica*, **42** (4), 233–252.

FAMA, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, **25** (2), 383–417.

KAPLAN, E. L. AND MEIER, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53** (282), 457–481.

KNAUS, W. A., HARRELL, F. E., FISHER, C. J., WAGNER, D. P., OPAL, S. M., SADOFF, J. C., DRAPER, E. A., WALAWANDER, C. A., CONBOY, K., AND GRASELA, T. H. (1993). The clinical evaluation of new drugs for sepsis: A prospective study design based on survival analysis. *JAMA*, **270** (10), 1233–1241.

LEE, M.-L. T. AND WHITMORE, G. (2006). Threshold regression for survival analysis: Modeling event times by a stochastic process reaching a boundary. *Statistical Science*, 501–513.

LI, D. X. (2000). On default correlation: A copula function approach. Working Paper 99-07, The Riskmetrics Group.

LIN, D. (1997). Non-parametric inference for cumulative incidence functions in competing risks studies. *Statistics in Medicine*, **16** (8), 901–910.

LOTTES, I. L., DEMARIS, A., AND ADLER, M. A. (1996). Using and interpreting logistic regression: A guide for teachers and students. *Teaching Sociology*, **24** (3), 284–298.

MAIR, P., REISE, S. P., AND BENTLER, P. M. (2008). IRT goodness-of-fit using approaches from logistic regression. Technical report, Department of Statistics, UCLA. UCLA: Department of Statistics, UCLA. Retrieved from: `http://escholarship.org/uc/item/1m46j62q`.

PEÑA, E. A. AND HOLLANDER, M. (2004). Models for recurrent events in reliability and survival analysis. *In Mathematical Reliability: An Expository Perspective*. Springer: New York, pp. 105–123.

PETRIE, K. J., CAMERON, L. D., ELLIS, C. J., BUICK, D., AND WEINMAN, J. (2002). Changing illness perceptions after myocardial infarction: An early intervention randomized controlled trial. *Psychosomatic Medicine*, **64** (4), 580–586.

STRAUSS, D., SHAVELLE, R. M., DEVIVO, M. J., AND DAY, S. (2000). An analytical method for mortality studies. *Journal of Insurance Medicine*, **32**, 217–225.

STRAUSS, D. J., SHAVELLE, R. M., AND ASHWAL, S. (1999). Life expectancy and median survival time in the permanent vegetative state. *Pediatric Neurology*, **21** (3), 626–631.

WEI, L. (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, **11** (14-15), 1871–1879.

WETTERSTRAND, W. H. (1981). Parametric models for life insurance mortality data: Gompertz's law over time. *Transactions of the Society of Actuaries*, **33**, 159–175.

Manuscript received, 2015-12-07, revised, 2016-09-20, accepted, 2016-11-13.