sciendo

ECONOMETRICS. EKONOMETRIA
Advances in Applied Data Analysis
Year 2019, Vol. 23, No. 2
ISSN 1507-3866; e-ISSN 2449-9994

COMPETING RISK MODELS OF DEFAULT IN THE PRESENCE OF EARLY REPAYMENTS

Ewa Wycinka
University of Gdańsk, Gdańsk, Poland
e-mail: ewa.wycinka@ug.edu.pl
ORCID: 0000-0002-5237-3488

DOI: 10.15611/eada.2019.2.07
JEL Classification: C14, C34, H81

Abstract: One of the central tasks of credit institutions is credit risk assessment, in which the estimation of the probability of default is an important element. The size of an institution's credit portfolio can decrease as a result of early repayments, which changes the probability of default over time. Prognosis of the probability of default should therefore also take into consideration the prognosis of early repayments. In this paper, methods of evaluating the probability of default over time, using competing risks regression models, are considered. Methods of evaluation for models of default over time are proposed. A sample of retail credits, provided by a Polish financial institution, was empirically examined.

Keywords: Cox model, Fine-Gray model, pseudo-observations, mixture models, vertical modelling.

1. Introduction

Survival analysis was introduced to credit scoring by Narain in 1992. Two advantages of survival analysis are its ability to model the probability of default over time and the ability to deal with censored observations. Sources of censored observations include the end of, and early repayments during, the follow-up period, and in some credit portfolios the number of early repayments is many times greater than the number of defaults.

Such heavy censoring during follow-up can cause biased estimates of the models' parameters in classic survival analysis, in which there is also the unverifiable assumption of the independence of the events of interest and censoring, which would be debatable in the case of risk of default and risk of early repayment. Another problem is time-scale; time is assumed to be a continuous random variable, whereas cred-

it repayment is measured in a finite number of instalments. As a result, the same default time can be recorded for more than one credit, creating tied observations. Competing risk analysis, which is an extension of classic survival analysis for more than one event, allows these problems to be overcome.

In this paper, the use of competing risks models in the prediction of defaults over credit life, in the presence of early repayments, is considered. A sample of 5,000 consumer credit accounts from a 24-month personal loan portfolio of a Polish financial institution is investigated, with the cohort of credits observed for 15 months. The characteristics of both credit and creditor were used as covariates in the regression models, with five such models for competing risks developed and compared: cause--specific hazard regression, subdistribution hazard regression, mixture models, vertical modelling, and regression based on pseudo-observations. The final discussion focuses on the usefulness of these models in predicting the probability of default.

2. Competing risks

In survival analysis, time to event is the object of study; in the competing risks scenario we assume there is a single lifetime for each individual, but events may be of different types or have different causes [Lawless 2003]. Every event can be assigned one, and only one, cause from a given set of causes [Crowder 2001]. If only one type of event is of particular interest, all other events can be summarised into a single category of competing risks, creating two types of events: events of interest and competing risks.

There are two approaches to competing risks: the first analyses a bivariate variable of time and type of event, the second a multivariate latent variable of unobserved times to different types of events.

2.1. First approach

Let (T, C) be a bivariate random variable in which T is a continuous variable representing time of the first event and $C = j$ $(j = 0, 1, \dots, p)$ is a discrete variable denoting the type of the first event $(j = 1, \dots, p)$ [Pintilie 2006; Crowder 2001]. It is assumed that one, and only one, event type is assigned to every event from the given set of p event types. If the time of observation for some units is earlier than the time of the first event, we have encountered right censoring.

In such a situation, $C = 0$ and T_c is the time at which the observation was censored, and the only thing we know is that $T > T_c$. Due to the right censoring, the variable (T, C) is only partially observable. We observe a pair $(\min\{T, T_c\}, C)$; as a result, the joint distribution of (T, C) is difficult to identify and can be estimated only by making assumptions, often unverifiable.

Marginal and conditional distributions of the bivariate random variable can be expressed in relation to the joint distribution as

$$P(T = t|C = j) = \frac{P(T=t,C=j)}{P(C=j)} \tag{1}$$

and

$$P(C = j|T = t) = \frac{P(T=t,C=j)}{P(T=t)}, \tag{2}$$

where $P(C = j)$ specifies the marginal distribution of the type of the first event and $P(T = t)$ is the marginal distribution of the time of the first event. The subdistribution function of event type j (cumulative incidence function (*CIF*)) is the probability that event type j will occur at or before time t

$$F_j(t) = P(T \leq t, C = j). \tag{3}$$

Subdistribution is not a proper distribution because

$$lim_{t\to\infty} F_j(t) = P(C = j) \leq 1. \tag{4}$$

The equality $P(C = j) = 1$ holds if there is only one type of event, i.e. that there are no competing risks.

The subsurvivor function is given by

$$S_j(t) = P(T > t, C = j). \tag{5}$$

The subdistribution and subsurvival functions are related by

$$F_j(t) + S_j(t) = P(C = j). \tag{6}$$

The sum of the subdistribution functions for all types of events is a marginal distribution function of variable T

$$F(t) = P(T \leq t) = \sum_{j=1}^{p} F_j(t) \tag{7}$$

and the sum of the subsurvival functions is a marginal survival function [Lindqvist 2008

$$S(t) = P(T > t) = \sum_{j=1}^{p} S_j(t). \tag{8}$$

The subdensity function $f_j(t)$ and marginal density $f(t)$ can be calculated as

$$f_j(t) = \frac{\partial F_j(t)}{\partial t} \tag{9}$$

and

$$f(t) = \sum_{j=1}^{p} f_j(t). \tag{10}$$

The subhazard function defined as

$$\tilde{h}_j(t) = \lim_{\partial t\to 0} \frac{P(t<T\leq t+\partial t, C=j|T>t)}{\partial t} = \frac{f_j(t)}{S(t)} \tag{11}$$

is the hazard of the event type j under the condition that the entity survived until time t, being the risk of all types of events $j = 1, \ldots, p$.

The hazard function in the marginal distribution of T, also called the overall hazard rate, is defined as

$$h(t) = \lim_{\partial t \to 0} \frac{P(t<T\leq t+\partial t|T>t)}{\partial t} = \frac{f(t)}{S(t)} \tag{12}$$

and the sum of the subhazards is

$$h(t) = \sum_{j=1}^{p} \tilde{h}_j(t). \tag{13}$$

The subdistribution function (3) for event type j can be expressed by the subhazard function as

$$F_j(t) = \int_0^t \tilde{h}_j(u)S(u)du, \tag{14}$$

here $S(t)$ can be expressed as

$$S(t) = \exp\Big[-\Big(\int_0^t h(u)du\Big)\Big]. \tag{15}$$

Gray [1988] proposed another type of hazard function – the hazard of subdistribution:

$$h_j^*(t) = lim_{\partial t \to 0} \frac{P((t<T\leq t+\partial t, C=j|T>t) \vee (T\leq t \wedge C\neq j))}{\partial t} = \frac{f_j(t)}{1-F_j(t)}, \tag{16}$$

which is the probability of the occurrence of event type j during the time interval $(t, t+\partial t)$, under the condition that the entity has not experienced any such event before time t nor has experienced any other type of event before time t [Pintilie 2006]. Individuals failing before time t from any event not of type j remain in the risk set for all future event times. The subdistribution function (3) can be directly derived from the hazard of subdistribution

$$F_j(t) = 1 - \exp(-\int_0^t h_j^*(u)du). \tag{17}$$

2.2. Second approach

Another approach to competing risks assumes that each event type j is assigned an event of time T_j. Only the first event is observed. $\boldsymbol{T}$ is a multivariate latent random variable $\boldsymbol{T} = (T_1, T_2, \ldots T_p)$ of p unobserved event times. Variables T_j $(j = 1, \ldots, p)$ are continuous and $P(T_j = T_k) = 0$ for all $j \neq k$ [Crowder 2001]. Two situations should be considered: the scenario with independent T_j and the scenario with possible dependency between T_j.

The joint distribution function of $\boldsymbol{T}$

$$G(\boldsymbol{t}) = P(\boldsymbol{T} \leq \mathbf{t}) \tag{18}$$

is the probability that each of the T_j variables is lower or equal to $\boldsymbol{t} = (t_1, t_2, \ldots, t_p)$ and

$$\bar{G}(\boldsymbol{t}) = P(\boldsymbol{T} > \mathbf{t}) \tag{19}$$

is the joint survival function. Marginal distributions specified by marginal distribution functions

$$G_j(t_j) = P(T_j \leq t_j) \tag{20}$$

and marginal survival functions

$$\bar{G}_j(t_j) = P(T_j > t_j) \tag{21}$$

do not define the joint distribution unless T_j is independent. Tsiatis [1975] demonstrated that, given any joint distribution with arbitrary dependence between component variables, there also exists a different joint distribution with independent component variables which has exactly the same marginal distributions as the first [Crowder 2001]. An assumption of independence cannot be verified because only the first event $T = \min\{T_1, T_2, \ldots T_p\}$ can be observed.

Another observed variable is the discrete variable $C = j$ $(j = 0,1, \ldots, p)$, where $j = 0$ is a censored observation and $j = 1, \ldots, p$ is an event type. Marginal distributions infer how risks would act in isolation (net risk). Marginal distributions do not describe events that actually occur, but rather describe events from isolated causes in situations in which all other types of events have been removed. However, one should note that after the isolation of competing risks, circumstances can change and, as a result, a distribution of T_j observed in isolation could be different to a marginal distribution derived from the joint distribution [Crowder 2001]. Peterson [1976] proved the relation $S_j(t) \leq \overline{G_j}(t_j)$, where $S_j(t)$ is the subsurvival function (5) and $\overline{G_j}(t_j)$ is the marginal survival function (21).

The hazard of the marginal distribution (cause-specific hazard) for event type j is defined as

$$h_j(t) = \lim_{\partial t \to 0} \frac{P(t < T_j \leq t + \partial t | T_j > t)}{\partial t} = -\frac{\partial ln \bar{G}_j(t)}{\partial t}. \tag{22}$$

As we assume that T_j $(j = 1, \ldots, p)$ are independent, then cause-specific hazards are equal to subhazards, for all t and j [Pintilie 2006]. Consequently,

$$\bar{G}(\boldsymbol{t}) = \prod_{j=1}^{p} \bar{G}_j(t_j), \tag{23}$$

where $\bar{G}(\boldsymbol{t})$ is the joint survival function (19) and $\overline{\mathrm{G}}_j(t_j)$ are marginal survival functions of T_j (21) [Cox, Oakes 1984; Lindqvist 2008].

3. Methods

Since the first applications of survival analysis to credit scoring, a range of competing risks models of default and early repayment have been developed. Banasik et al. [1999] and Stepanova and Thomas [2002] used the Cox Proportional Hazards models separately for both default and early repayment. Deng et al. [2000], Pavlov [2001], and Ciochetti et al. [2002] examined the joint risks of default and early repayment, using the Cox model for overall hazard (12). This function was calculated as the sum of the subhazards for default and early repayment.

Steinbuks (2015) used extensions of the Cox PH model to investigate the effect of prepayment regulations on the termination of subprime mortgages. With the popularity of the Cox models, other regression models for competing risks received very little attention. In this article, regression models for competing risks in credit risk assessment are investigated, using models which were originally developed in biostatistics. Of the methods presented in the literature, the five most popular regression methods reviewed by Haller et al. [2013] were chosen. These are: cause-specific hazard regression, subdistribution hazard regression, mixture models, vertical modelling, and regression based on pseudo-observations.

3.1. Cause-specific hazard regression

Prentice et al. [1978] proposed modelling cause-specific hazards (22) using Cox-type regression [Cox 1972], assuming proportional cause-specific hazards for all types of events

$$h_j(t|X) = h_{j0}(t)\exp(\sum_{k=1}^{m}\beta_k X_k). \tag{24}$$

Here, $h_{j0}(t)$ is a cause-specific baseline hazard for event type j, $X = (X_1, X_2, \ldots, X_m)$ is a vector of covariates, and $\beta = (\beta_1, \ldots, \beta_m)$ is a vector of regression coefficients for event type j. Maximum partial likelihood estimation of the regression coefficients can be conducted by the iterative Newton-Raphson algorithm. It is assumed that, at each particular time point, only one event type j occurs. When time is not a strictly continuous variable, i.e. when the time of event j is equal for two or more units in a dataset, then the problem of tied event times occurs and a modification of the approximation algorithm is needed [Therneau, Grambsch 2000].

The model assumes proportionality of cause-specific hazards, which means that, for two individuals with vectors of covariates X and X^*, the ratio of their hazard rates is

$$\frac{h_j(t|X)}{h_j(t|X^*)} = \frac{h_{j0}(t)\exp(\sum_{k=1}^{m}\beta_k X_k)}{h_{j0}(t)\exp(\sum_{k=1}^{m}\beta_k X_k^*)} = exp(\sum_{k=1}^{m}\beta_k(X_k - X_k^*)), \tag{25}$$

which is a constant over time. A number of graphical methods and tests have been proposed in the literature to check this assumption (e.g. [Pintilie 2006; Li et al. 2015]). In this paper, omnibus tests proposed by Li et al. (2015) are used. The ad-

vantage of this method is that the same type of test can also be used to check the proportionality of subdistribution hazards in the Fine-Gray model (described later).

In the CoxPH model only the event type j is analysed, with all competing risks assumed to be censored observations. The cause-specific hazard for event type j is modelled as if this event were the only possible one. The effect of covariates on the cause-specific hazard cannot be translated directly on the CIF (see (3)). To estimate the cumulative incidence function, the Cox PH models for cause-specific hazards have to be estimated for each type of event. Beyersmann and Schumacher [2007], in the case of two competing risks (event of interest ($j = 1$) and all competing risks combined ($j = 2$)), expressed the cumulative incidence function for the event of interest in terms of cause-specific hazards of two risks. The cumulative incidence function for the event of interest ($C = 1$) can be expressed as

$$F_1(t) = \int_0^t h_1(u) \exp\left\{-\left[\int_0^t h_1(u) + \int_0^t h_2(u)\right]\right\} du. \tag{26}$$

3.2. Subdistribution hazard regression

The hazard of subdistribution (16) can be modelled by regression, as developed by Fine and Gray [1999]. This is a Cox-type regression

$$h_j{}^*(t|X) = h_{j0}{}^*(t)\exp(\textstyle\sum_{k=1}^{m} \beta_k X_k). \tag{27}$$

Here, $h_{j0}{}^*(t)$ is the baseline hazard of subdistribution. The difference in estimation of parameters between the Cox PH model (24) and the Fine-Gray model (27) is in the definition of the risk set needed for the partial likelihood. In the Cox model, the risk set is the set of individuals still at risk at t, whereas in the Fine-Gray model, the risk set comprises those units who did not experience the event of interest by time t and those who experienced a competing event before time t.

Additionally, for the partial likelihood, in the Fine-Gray model, weights are added such that units who experience the competing risk at time t have wages after time t which decrease over time from one to zero. Thus the share of the units who experienced competing risks decreases in the likelihood function (c.f. [Pintilie 2006, pp. 87-92]). The Fine-Gray model assumes the proportionality of hazards of subdistributions (c.f. (25)), but Grambauer et al. [2010] showed that a subdistribution hazard regression model has a proper interpretation, even when subdistribution hazards are not proportional. On the basis of equation (17), the cumulative incidence function (3) for the Fine-Gray model can be directly estimated as

$$F_j(t|X) = 1 - exp(-\int_0^t h_j^*(u|X)du). \tag{28}$$

3.3. Mixture models

Larson and Dinse [1985] based their method on the conditional distributions presented in equation (1). The transformation of equation (1) gives the joint distribution of event types

$$P(T = t, C = j) = P(C = j)P(T = t|C = j) \tag{29}$$

as a mixture of the marginal distribution of the types of event and the conditional distribution of the times of the accordant type of event, given the type of event. The disadvantage of this method is that an estimated probability of type of event $P(C)$ depends on the length of follow-up [Nicolaie et al. 2010].

A number of different distributions can be used for each of the two components of a mixture model. In this paper, logistic regression will be used for the distribution of types of events and the Cox PH model for the conditional distribution of times to the given type of event. The sets of covariates for both component models can be different. In the case of two types of risks, with $C = 1$ as the risk of interest and $C = 2$ for all competing risks, the probability of the event of interest, given a set of covariates Y, can be expressed by a binary logistic model

$$P(C = 1|Y) = \frac{\exp(\alpha_0 + \sum_{k=1}^{m} \alpha_k Y_k)}{1 + \exp(\alpha_0 + \sum_{k=1}^{m} \alpha_k Y_k)}, \tag{30}$$

where α_0 is an intercept, $\alpha = (\alpha_1, \dots, \alpha_m)$ is a vector of parameters, and $Y = (Y_1, \dots, Y_m)$ is a vector of covariates.

The conditional distribution of the survival times for a given type of event and given set of covariates can be modelled by a Cox PH model. Survival function for a risk of interest $C = 1$ and a given set of covariates X can be denoted as

$$\mathrm{S}_1(\mathrm{t}|\mathrm{C} = 1, \mathrm{X}) = \mathrm{P}(\mathrm{T} > \mathrm{t}|\mathrm{C} = 1, \mathrm{X}) = \exp(-\int_0^t h_{01}(u) exp(\sum_{k=1}^{r} \beta_k X_k) du). \tag{31}$$

Here, $h_{01}(t)$ is the null cause-specific hazard function for event type $C = 1$ for an individual with all covariates set to zero, $X = (X_1, X_2, \dots, X_r)$ is a vector of covariates, and $\beta = (\beta_1, \beta_2, \dots, \beta_r)$ is a vector of regression coefficients for the event type $C = 1$.

The $\mathrm{S}_1(\mathrm{t}|\mathrm{C} = 1, \mathrm{X})$ function is a proper survival function, i.e. $\lim_{t\to\infty} \mathrm{S}_1(\mathrm{t}|\mathrm{C} = 1, \mathrm{X}) = 0$. In the Cox model, hazard rates are assumed to be proportional and the survival function for the competing risk $C = 2$ can be estimated analogously.

Finally, the CIF (see (3)) for each event of type j can be expressed as

$$\mathrm{F}_\mathrm{j}(\mathrm{t}) = (1 - \mathrm{S}_\mathrm{j}(\mathrm{t}|\mathrm{C} = \mathrm{j}, \mathrm{X})) \cdot P(C = j|Y) \tag{32}$$

(c.f. [Lau et al. 2008]). In the mixture-model approach, no assumption about the independence of competing risks is necessary, which is an appealing feature of this model [Ng, McLachlan 2003].

In the literature, a few algorithms have been proposed to estimate regression coefficients of mixture models for competing risks (e.g. [Ng, McLachlan 2003; Chang et al. 2007]). For the case of a semiparametric mixture model with proportional hazards for failure time, conditional on type of cause and with a marginal multinomial model for type of cause, Chang et al. [2007] provided algorithms for a non--parametric maximum-likelihood estimate (NPMLE) of the parameters of covariates. This method assumes the same set of covariates as in the Cox PH and logit models. The above algorithm will be used in the empirical part of this study. The method proposed by Ng and McLachlan [2003] is not analysed in this paper.

3.4. Vertical modelling

Nicolaie et al. [2010] proposed another method to model joint probability. The transformation of equation (2) is

$$P(T = t, C = j) = P(T = t) \cdot P(C = j|T = t). \tag{33}$$

The joint distribution is therefore a mixture of the marginal distribution of time to all types of events and the conditional distribution for events of type j, given the event time. If the covariate effect has to be included, the marginal distribution of time to event can be estimated by the Cox PH model, while the conditional distribution of types of events, given the time of the event, can be estimated by a multinomial logit model. The whole dataset is used to evaluate the marginal distribution of survival times, whereas the conditional distribution of types of events is evaluated only using the set of complete observations. Each of the above models is estimated separately [Nicolaie et al. 2010].

Let the relative subhazard for event type j be

$$\pi_j(t) = P(C = j|T = t) = \frac{\tilde{h}_j(t)}{h(t)}. \tag{34}$$

Here $\tilde{h}_j(t)$ is the subhazard function (11) and $h(t)$ is the overall hazard (12). The relative subhazard is the conditional probability that, at time t, the event is of type j, given that an event occurs at time t. For each t, a relationship exists such that $\sum_{j=1}^{p} \pi_j(t) = 1$ [Cox, Oakes 1984; Nicolaie et al. 2010].

Taking equation (14), the reversal of equation (34), and adding to the model two sets of covariates, X and Y, to model the distribution of time to event and type of event, respectively, the cumulative incidence function can be expressed as

$$F_j(t|X, Y) = \int_0^t \pi_j(u|Y)h(u|X)S(u|X)du. \tag{35}$$

Relative subhazards $\pi_j(t|Y)$ can be estimated by a multinomial logit model. In the case of only two types of events, this model reduces to a binary logit model

$$\pi_j(t|Y) = \frac{\exp(\alpha_0 + \sum_{k=1}^{l} \gamma_k B(t) + \sum_{k=1}^{r} \alpha_k Y_k)}{1 + \exp(\alpha_0 + \sum_{k=1}^{l} \gamma_k B(t) + \sum_{k=1}^{r} \alpha_k Y_k)}, \tag{36}$$

with time T and covariates Y as independent variables. Here, $B(t)$ are spline functions of time. Using functions of time is justified by the need to smooth the changes of the relative subhazard over time. The use of raw ratios could lead to inexplicable variations; Nicolaie et al. [2010] suggested including the interactions of covariates with time functions in the model, if data is sufficient.

The marginal distribution of time to event in (33) can be estimated by a Cox PH model for all types of events considered as the event of interest. Thus we do not observe competing risks here. The overall hazard $h(t|X)$ can be modelled by

$$h(t|X) = h_0(t)\exp(\sum_{k=1}^{m} \beta_k X_k). \tag{37}$$

Here, $h_0(t)$ is an overall baseline hazard, $X = (X_1, X_2, \dots, X_r)$ is a vector of co--variates, and $\beta = (\beta_1, \beta_2, \dots, \beta_m)$ is a vector of regression coefficients. The marginal survival function is then

$$S(t|X) = [S_0(t)]^{\exp(\sum_{k=1}^{m} \beta_k X_k)}. \tag{38}$$

Here, $S_0(t)$ is the baseline marginal survival function (8) that corresponds to the baseline hazard function [Kleinbaum, Klein 2012]. The assumption of proportionality of the hazards in this model is required. Finally, the cumulative incidence function given by (35) is calculated with formulas (36)-(38) as components.

3.5. Competing risks regression based on pseudo-observations

The regression based on pseudo-observations directly models a cumulative incidence function – a methodology first proposed by Andersen et al. [2003] for multi-state models. The main idea of this approach is to replace each censored observation by the appropriate proxy, which consequently allows the use of regression methods for completed data. In the case of competing risks, such a proxy is a cumulative incidence function for event j (3).

Let n be the sample size. At an arbitrary predefined series of time points $t \in \{t_1, \dots, t_H\}$, pseudo-observations for unit i $(i = 1, \dots, n)$ and event j are evaluated as

$$\hat{\theta}_{ji}(t) = n\hat{F}_j(t) - (n-1)\hat{F}_j^{(-i)}(t). \tag{39}$$

Here, $\hat{F}_j(t)$ is the estimated cumulative incidence function for event type j at time t, using all units, and $\hat{F}_j^{(-i)}(t)$ is the estimated cumulative incidence function derived from all but the i-th unit. For each unit, both completed and censored, H pseudo-observations are calculated, and the final augmented data set consists of an $n \times H$ matrix of pseudo-observations.

Andersen et al. [2003] proposed the use of pseudo-observations as a dependent variable in a regression model. However, multiple pseudo-observations for each unit can be a source of correlation in a dataset, and therefore a regression model for corre-

lated data should be chosen. Generalised estimation equations (GEE), introduced by Liang and Zeger [1986], are generalisations of generalised linear models for correlated data. The GEE model for pseudo-observations is

$$g\left(\hat{\theta}_j(t)\middle|X^*\right) = \alpha_o + \sum_{k=1}^{m+H} \beta_k X_k. \tag{40}$$

Here, $g(.)$ is a link function, and vector X^* includes covariates X_k $(k = 1, \dots, m)$ and indicators of time points (as dummy variables) X_k $(k = m + 1, \dots, m + H)$. Estimated regression coefficients for time points can be expressed as time-dependent intercepts. As a result, the model can be presented as

$$g\left(\hat{\theta}_j(t)\middle|X\right) = \alpha_o + \alpha_o(t) + \sum_{k=1}^{m} \beta_k X_k. \tag{41}$$

When a complementary log-log link function on $(1 - x)$ is used, $g(x) = -\log(-\log(1 - x))$, then the regression model has the form

$$-\log\left[-\log\left(1 - F_j(t)\right)\right] = \alpha_o + \alpha_o(t) + \sum_{k=1}^{m} \beta_k X_k \tag{42}$$

which can be expressed in a form analogous to a Fine-Gray model (28) as

$$F_j(t) = 1 - \exp\{-\exp[-(\alpha_o + \alpha_o(t) + \sum_{k=1}^{m} \beta_k X_k)]\}. \tag{43}$$

Estimates of the coefficients are based on the estimating equations with a prespecified type of working covariance matrix [Andersen, Perme 2010]. Klein and Andersen (2005), in the Monte Carlo study, showed that the choice of an independence-model working covariance matrix for pseudo-observations gives estimations of GEE that do not significantly differ from the results for other, more complex working covariance matrices.

4. Model evaluation

There are two objectives of credit risk assessments. First, the lender wants to know the number or percentage of credits that default in each month of the credit's life; second, the lender wants to know which credits are more susceptible to the risk of default during these months. A ranking of credits according to the risk of default for each time point is therefore needed. To assess which of the evaluated models best realises these objectives, two kinds of measures were used in this study.

Firstly, empirical and theoretical cumulative incidence functions for the whole sample were compared at all time points. To measure the mean error of classification, the following measure was used

$$(MSE)^{1/2} = \sqrt{\sum_{i=1}^{H} (y_{t_i} - \hat{y}_{t_i})^2 / H}\ . \tag{44}$$

Here, y_t is an empirical number of defaults at time t in the sample and $\hat{y}_t$ is a theoretical number of defaults at time t according to the given model.

Secondly, evaluation of the discriminant power of the estimated models was assessed by the receiving operating curve (ROC) and the area under it (AUC). For any binary outcome ($D = \{0,1\}$), where $D = 1$ are cases and $D = 0$ are controls, and a continuous predictor Z, the ROC curve is created by plotting, for various thresholds k ($k \in Z$), true positive rates ($TPR(k) = P(Z > k|D = 1)$) against false positive rates ($FPR(k)) = P(Z > k|D = 0)$.

In survival analysis, ROC curves can be estimated for different time points (t). Moreover, the outcome at time t ($D(t) = 1$) can be considered as the presence of the event of interest at time t (*incident case*) or before time t (*cumulative case*). Additionally, due to the presence of competing risks and censored observations, controls ($D(t) = 0$) are those units with $T_i > \tau$, for a large time τ, called *static controls*, or those units with $T_i > t$ (*dynamic controls*). Finally, predictor Z can be a *fixed predictor* (i.e., measured once at time $t = 0$) or measured at each time point t for which $AUC(t)$ is evaluated (*longitudinal predictor*). Due to these particularities, in survival analysis, time-dependent ROC curves should be used [Blanche et al. 2013]. $ROC(t)$ plots $TPR(k,t)$ against $FPR(k,t)$ for varying k. The *AUC*(*t*) is defined as

$$AUC(t) = \int_{-\infty}^{\infty} TPR(k,t) \left| \frac{\partial FPR(k,t)}{\partial k} \right| dk, \tag{45}$$

where $TPR(k,t) = P(Z > k|D(t) = 1)$ and $FPR(k,t) = P(Z > k|D(t) = 0)$.

Some modifications of $AUC(t)$ in comparison to AUC are necessary. Blanche et al. (2013) reviewed the estimators for time-dependent $AUC(t)$, with different definitions given for the cases, controls, and predictors. In this paper, the cumulative/dynamic approach to compute time-dependent $ROC(t)$ was applied [c.f. Blanche at al. 2013]. The outcomes at time t ($D(t) = 1$) are considered defaults that have occurred before time ($T \leq t$), whereas controls ($D(t) = 0$) are units that are free of any event before time t ($T > t$). The *CIF* (see (3)) at time $t = 3$ was used as a predictor. Estimators of the *AUC(t)* for different approaches to cases, controls, and predictors are presented in Blanche et al. [2013].

6. Data

In this paper, we investigate the use of competing risks models for a sample of 5,000 consumer credits granted for 24 months by a Polish financial services organisation. All of these credits were granted within a period of 12 calendar months. A cohort of credits was observed for 15 months following the origination of the first credit, and thus the earliest granted credits were observed for 15 months, while credits granted in the 12th calendar month were observed for only 3 months.

Default was considered as 90 days overdue in payment and is the event of interest. Early repayment was repayment before the indicated end date of the loan and is considered a competing risk. Censored observations are those credits for which neither default nor early repayment occurred before the end of follow-up. There

were 446 creditors who defaulted during the follow-up, 3,454 creditors who repaid the credit early, and 1,100 censored observations. The distribution of events and censoring over time is presented in Table 1.

Table 1. The distribution of defaults, early repayment, and censoring in the cohort of credits through to follow-up

	Month of credits' life														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
No. of defaults	0	0	79	50	56	45	39	35	21	27	32	19	22	15	6
No. of early repayments	129	174	396	368	340	323	409	266	238	248	175	166	115	84	23
No. of censored observations	0	0	125	110	87	93	120	73	82	88	109	91	61	52	9

Source: own study.

Table 2. Variables and their attributes, and the inclusion of covariates in the regression models

Covariates	Number of			% of total	Attribute not included in the model				
	censored	defaulted	early rep.		Cox (def)	Cox (e.r.)	F-G	Mixt.	Vert.
X1_0	132	91	405	13					
X1_1	819	270	2537	73		N		N	N
X1_2	149	85	512	15					
X2_0	84	54	388	11					
X2_1	116	77	523	14					
X2_2	574	228	1837	53					
X2_3	326	87	706	22					
X3_0	106	89	281	10					
X3_1	176	85	497	15					
X3_2	543	164	1750	49					
X3_3	275	108	926	26					N
X4_0	74	60	350	10					
X4_1	133	74	518	15		N		N	
X4_2	562	247	1786	52					
X4_3	331	65	800	24		N		N	N
X5_0	527	212	1753	50					
X5_1	270	92	897	25					
X5_2	303	142	804	25					
X6_0	47	75	197	6					
X6_1	371	244	1232	37		N	N	N	N
X6_2	682	127	2025	57		N	N	N	N
X7_0	579	302	1777	53					
X7_1	521	144	1677	47	N	N		N	N
X8_0	913	402	2932	85					
X8_1	69	21	210	6					
X8_2	118	23	312	9					
X9_0	673	202	1970	57					
X9_1	148	104	464	14	N	N		N	N
X9_2	279	140	1020	29					

Source: own study.

The dataset, in addition to information about payment behaviour, also contains typical application characteristics such as amount of credit, amount of an instalment, purpose of the loan, age of the applicant, property, and educational level. These variables are used in the regression models as covariates.

To anonymise the data all the variables are denoted by the letter *X* appended with consecutive numbers. Covariates were categorised and replaced by dummy variables created for each attribute of the variable. Numbers preceded by an underline show the number of the attribute. Attributes denoted by _0 are reference groups. To avoid collinearity, the reference group for each variable is excluded from the models. Percentages of the units with particular attributes of each variable are given in Table 2.

7. Results

All analyses were performed using the statistical software R [R Development Core Team 2017] and its libraries (*survival* [Thernau 2017], *cmprsk* [Gray 2014], *NPMLEcmprsk* [Chen, Chang, Hsiung 2015], *goftte* [Sfumato, Boher 2017], *timeROC* [Blanche 2015], *splines* (part of R), *pseudo* [Pohar Perme et al. 2017], and *geepack* [Højsgaard et al. 2016]).

The models compared in this paper are based on different principles and require the fulfilment of different assumptions. As a result, it is difficult to find one common set of criteria for variable selection that could be applied to all. Backward elimination is ambiguous in mixture models and can lead to undesirable models with poor performance, while criteria based on likelihood functions, such as the Akaike Information Criterion, cannot be used in the GEE models. Due to these difficulties, and in order to compare the performance of the estimated models, no variable selection methods were implemented. All the variables listed in Table 2 were included in the models, except those that did not meet the assumption of proportionality that is required by most of the models.

In the first step of the analysis, the proportionality of the hazards for all covariates in the Cox-type models and in the Fine-Gray model were checked by the omnibus test proposed by Lin et al. [1993] and Li et al. [2015] (*goftte* package). In the Cox PH model, variables were separately excluded from the model for default (Cox def.) and from the/ model for early repayment (Cox e.r.). In the Fine-Gray model (F-G) only the variable X6 failed to meet the assumption of proportionality of subhazards. The mixture model, estimated with the use of the NPMLE algorithm (see Section 3.3), contains the same set of variables as all of the components: time until default, time until early repayment, and probability for type of event. Therefore, all variables that did not meet the proportionality assumption for the Cox PH models for default or for early repayment were also excluded from mixture model (c.f. Table 2).

In the vertical approach the Cox model was estimated for all types of events. The proportionality assumption was verified for the Cox model with combined events. None of the variables were removed from the GEE model for pseudo-observations

because, of all the compared models, this was the only one that did not require the assumption of proportionality. The variables that were excluded from models due to the lack of proportionality are marked in Table 2. Extant variables were applied in the models as covariates.

Table 3. Coefficients for covariates in competing risks models for default

Cov.	Cox PH (def.)		Cox PH (e.r.)		Fine and Gray		Mixture model					
	B	*p*-value	β	*p*-value	β	*p*-value	α	*p*-value	β_1	*p*-value	β_2	*p*-value
Int.	.	.	.	.	.	.	1.12	0.000	.	.	.	.
X1_1	-0.35	0.011	.	.	-0.16	0.250	.	.	.	.	.	.
X1_2	0.05	0.759	-0.02	0.643	0.19	0.240	0.29	0.004	0.06	0.343	-0.07	0.088
X2_1	0.02	0.932	0.06	0.339	0.02	0.920	-0.9	0.000	3.08	0.000	-0.42	0.000
X2_2	-0.2	0.226	-0.1	0.069	-0.23	0.160	-0.91	0.000	2.69	0.000	-0.53	0.000
X2_3	-0.35	0.094	-0.27	0.000	-0.36	0.081	-1.21	0.000	2.75	0.000	-0.67	0.000
X3_1	-0.15	0.355	0.13	0.090	-0.45	0.008	-1.23	0.000	1.00	0.000	-0.37	0.000
X3_2	-0.79	0.000	0.14	0.030	-0.92	0.000	-1.94	0.000	0.75	0.000	-0.46	0.000
X3_3	-0.63	0.001	0.16	0.025	-0.61	0.003	-1.89	0.000	0.85	0.000	-0.42	0.000
X4_1	-0.28	0.114	.	.	-0.08	0.650	.	.	.	.	.	.
X4_2	-0.14	0.412	-0.02	0.515	0.07	0.670	0.32	0.000	0.28	0.022	-0.04	0.133
X4_3	-0.49	0.036	.	.	-0.18	0.450	.	.	.	.	.	.
X5_1	-0.08	0.564	-0.04	0.326	-0.19	0.170	-0.5	0.000	0.76	0	-0.14	0.001
X5_2	0.22	0.073	-0.05	0.249	0.39	0.002	0.47	0.000	0.25	0.044	0	0.495
X6_1	-0.57	0	.	.	.	.	.	.	.	.	.	.
X6_2	-1.66	0	.	.	.	.	.	.	.	.	.	.
X7_1	.	.	.	.	-0.5	0.000	.	.	.	.	.	.
X8_1	-0.36	0.110	0.02	0.784	-0.37	0.100	-0.67	0.000	-0.05	0.427	-0.07	0.184
X8_2	0.50	0.024	0.03	0.605	-0.63	0.004	-0.93	0.000	-0.04	0.454	-0.02	0.359
X9_1	.	.	.	.	0.26	0.072	.	.	.	.	.	.
X9_2	0.35	0.003	0.05	0.232	0.34	0.004	0.44	0	0.16	0.17	0.1	0.009

Source: own study.

The cause-specific hazards for default and for early repayment were modelled by the Cox PH regression (21) with the *survival* package. The coefficients of these models are presented in Table 3. Both survival functions and baseline hazards for these models were used to estimate cumulative incidence functions (22). The coefficients of the Fine-Gray model (27) were estimated with the *cmprsk* package. The mixture model (32) was estimated with the use of the *NPMLEcmprsk* package. The results of the estimations are presented in Table 3.

Table 3. (continuation). Coefficients for covariates in competing risks models for default

Vertical model								GEE model					
Logit model for default						Cox PH for time to any event							
Cov.	γ	*p*-value	Cov.	*α*	*p*-value	*β*	*p*-value	Cov.	*α*	*p*-value	Cov.	*β*	*p*-value
Int.	0.87	0.1	X1_1	-0.25	0.123	.	.	α0	-18.55	0.000	X1_1	-0.17	0.285
B1(t)	-1.94	0.072	X1_2	0.17	0.389	0.02	0.58	α0 (2)	-6.14	0.000	X1_2	0.27	0.146
B2(t)	-0.76	0.081	X2_1	-0.1	0.64	0.06	0.352	α0 (3)	15.76	0.000	X2_1	0.17	0.414
B3(t)	-0.43	0.501	X2_2	-0.32	0.087	-0.12	0.025	α0 (4)	16.31	0.000	X2_2	-0.09	0.646
			X2_3	-0.53	0.029	-0.28	0.000	α0 (5)	16.69	0.000	X2_3	-0.17	0.472
			X3_1	-0.28	0.159	0.02	0.676	α0 (6)	16.92	0.000	X3_1	-0.27	0.15
			X3_2	-0.85	0.000	-0.05	0.186	α0 (7)	17.08	0.000	X3_2	-0.58	0.003
			X3_3	-0.52	0.033	.	.	α0 (8)	17.21	0.000	X3_3	-0.33	0.163
			X4_1	-0.07	0.739	0.14	0.01	α0 (9)	17.28	0.000	X4_1	-0.11	0.601
			X4_2	0.27	0.167	0.04	0.332	α0 (10)	17.38	0.000	X4_2	0.22	0.246
			X4_3	0.23	0.383	.	.	α0 (11)	17.49	0.000	X4_3	-0.05	0.855
			X5_1	-0.27	0.094	-0.04	0.34	α0 (12)	17.56	0.000	X5_1	-0.04	0.793
			X5_2	0.51	0.000	-0.01	0.895	α0 (13)	17.67	0.000	X5_2	0.41	0.003
			X6_1	-0.7	0.000	.	.	α0 (14)	17.75	0.000	X6_1	-0.53	0.000
			X6_2	-1.77	0.000	.	.	α0 (15)	17.82	0.000	X6_2	-1.56	0.000
			X7_1	-0.5	0.000	.	.				X7_1	-0.45	0.002
			X8_1	-0.3	0.233	-0.02	0.777				X8_1	-0.37	0.169
			X8_2	-0.58	0.016	-0.03	0.669				X8_2	-0.48	0.06
			X9_1	0.27	0.103	.	.				X9_1	0.22	0.171
			X9_2	0.3	0.026	0.09	0.017				X9_2	0.24	0.083

$\alpha_0(t)$ – time-dependent intercept at time t

Source: own study.

In the vertical modelling approach, the Cox PH model for any type of event was estimated. Logit models were then estimated only for complete observations (see Section 3.4). For variations of relative subhazard (34) over time, cubic b-splines (t) with $k = 3$ knots were applied to smooth the variations (Figure 1). The interactions of time and covariates were not considered in the model due to the large number of covariates used, in comparison with the number of complete observations. Estimates of both parts of the vertical model are shown in Table 3 (continuation).

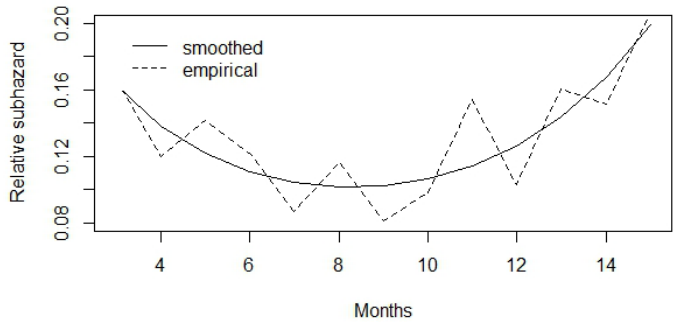

Fig. 1. Empirical relative subhazard, and smoothed by *B*-splines

Source: own study.

In the pseudo-values approach (see Section 3.5), pseudo-observations for all units and for 15 time points were calculated (*pseudo* package). Pseudo-observations were used as values of the dependent variable in a GEE model. Estimations were made with the *geepack* package and estimates of the model are given in Table 3 (continuation). The use of dummy variables for time points resulted in time--dependent intercepts ($\alpha_0(2) - \alpha_0(15)$) in the model.

To evaluate the performance of the above models, CIF (see (3)) were estimated and compared. In Figure 2, the solid line represents the empirical cumulative incidence function for defaults in the sample and the dotted lines represent theoretical CIFs for estimated models. The Cox PH model best mimics jumps in, and the level of cumulative incidences for, all time points. Only the vertical model predicts the presence of events after time=1 rather than time=3. The whole theoretical CIF curve for this model lies above the empirical curve. The mixture model gives overestimated values of CIF for most time points.

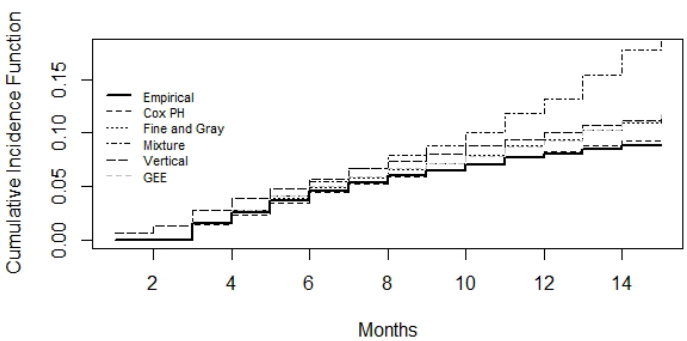

Fig. 2. Empirical and theoretical cumulative incidence functions for default in the sample of 5,000 credits

Source: own study.

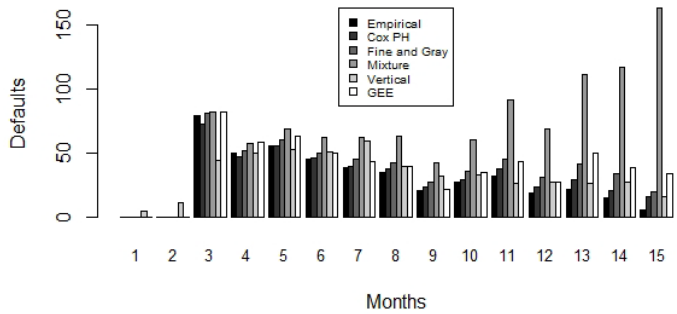

Fig. 3. Empirical and theoretical (by model) distributions of defaults through the months of credit life

Source: own study.

The differences between the actual and theoretical number of defaults in consecutive months of credit life are shown in Figure 3. For months 3 through 5, the Cox PH model underestimates the real number of defaults while in later months it overestimates. However, in comparison with the other models, the Cox PH model gives the best estimation of defaults at most of the time points.

Table 4. Time-dependent *AUC* (and $(MSE)^{1/2}$ in the last row) for competing estimated risks models

Month	Cox PH (95% CI)	Fine and Gray (95%CI)	Mixture (95%CI)	Vertical (95%CI)	GEE (95%CI)
4	0.748 (0.700-0.801)	0.664 (0.607-0.736)	0.599 (0.552-0.680)	0.819 (0.749-0.879)	0.767 (0.704-0.821)
5	0.785 (0.750-0.821)	0.709 (0.669-0.756)	0.610 (0.577-0.714)	0.829 (0.778-0.878)	0.792 (0.751-0.832)
6	0.779 (0.750-0.813)	0.707 (0.671-0.750)	0.601 (0.571-0.725)	0.798 (0.749-0.848)	0.782 (0.746-0.816)
7	0.768 (0.742-0.801)	0.696 (0.667-0.735)	0.609 (0.574-0.72)	0.764 (0.712-0.821)	0.77 (0.738-0.803)
8	0.771 (0.747-0.803)	0.692 (0.665-0.730)	0.624 (0.591-0.708)	0.749 (0.696-0.808)	0.774 (0.743-0.803)
9	0.762 (0.739-0.796)	0.682 (0.654-0.719)	0.615 (0.581-0.699)	0.721 (0.667-0.785)	0.764 (0.73-0.791)
10	0.757 (0.734-0.79)	0.672 (0.648-0.711)	0.602 (0.567-0.689)	0.692 (0.634-0.765)	0.753 (0.72-0.784)
11	0.751 (0.727-0.787)	0.670 (0.649-0.708)	0.587 (0.551-0.678)	0.664 (0.594-0.745)	0.744 (0.709-0.777)
12	0.724 (0.695-0.766)	0.671 (0.65-0.707)	0.579 (0.544-0.666)	0.627 (0.54-0.73)	0.719 (0.684-0.759)
13	0.72 (0.683-0.774)	0.674 (0.653-0.710)	0.578 (0.547-0.687)	0.611 (0.498-0.733)	0.712 (0.667-0.766)
14	0.774 (0.706-0.844)	0.678 (0.657-0.711)	0.634 (0.55-0.752)	0.644 (0.428-0.838)	0.761 (0.681-0.835)
$(MSE)^{1/2}$	4.48	10.10	59.23	12.67	13.14

Source: own study.

The values of the $\sqrt{MSE}$ measure (44) for all the models are given at the bottom of Table 4. The lowest error (4.48 defaults per month) was generated by the Cox PH model, which confirms the observation from Figure 3 that the Cox PH model gives estimates closest to the real distribution of defaults.

To measure the applicability of the CIF as a score function, time-dependent ROC curves and the area above them, $AUC(t)$ were calculated for months 4 to 14. Due to the construction of the estimators of the $AUC(t)$, estimates *of* $AUC(t)$ for months 3 and 15 could not be calculated. The results are presented in Table 4; confidence intervals were estimated as percentiles 2.5 and 97.5 from 1000 bootstraps.

The *CIFs* created by the models give satisfactory results, with all AUCs significantly greater than 0.5. For months 4 to 6, the vertical model gives the best predictive accuracy of the compared models, whereas during the subsequent months it gives one of the worst. It is also worth noting that, even though the *CIFs* for the whole sample are very close (c.f. Figure 2), as given by the model for pseudo-observations (GEE model) and by the Fine-Gray model, the $AUC(t)$ are higher for the first model for all t.

8. Discussion

In this paper, the application of competing risks regression models was proposed to evaluate the probability of default over time. The empirical study showed that competing risks models can be effectively used in credit risk assessment, but the results of the study do not prejudge which of the models would be the most efficient scoring tool.

The differences in the model prerequisites result in different choices of sets of covariates. The advantage of the pseudo-observations approach is that, in contrast with other models, it does not require the assumption of proportionality to be met, which allows the inclusion of variables which are excluded from other types of model. The predictive accuracy of this model, measured by $AUC(t)$, was also one of the highest. However, the weakness of this model is that it overestimated the *CIF* at all time points.

The CIF calculated on the basis of the Cox PH models for default and for early repayment best resembled the distribution of defaults over time. A worthwhile feature of this model is the option to use different sets of covariates in survival functions for time to default and time to early repayment.

Despite these positive results, one should keep in mind that this model requires an unverifiable assumption of independence of default and early repayment. Borrowers who decide to repay credit before the predefined end date are often those who have their own assets for repayment and prefer not to pay interest to the bank. However, the decision to repay early could also be part of a refinancing process; borrowers who are in a bad financial condition and are unable to make instalment payments could take additional credit, for a longer period of time and with lower instalment

payments, to pay back the first credit. This would allow avoidance, or at least postponement, of default.

Some banks, especially those offering services to creditors excluded from the broader financial market, may have such customers. A bank is not usually informed of the reason for early repayment, and heterogeneity in the group of borrowers who repay credit early could influence the estimation and performance of models for probability of default. It would therefore be valuable to compare the performance of models for competing risks in different portfolios of retail credit.

Bibliography

Andersen P.K., Klein J.P., Rosthøj S., 2003, *Generalised linear models for correlated pseudo-observations, with applications to multi-state models*, Biometrika, 90, pp. 15-27.

Andersen P.K., Perme M., 2010, *Pseudo-observations in survival analysis*, Statistical Methods in Medical Research, 19(1), pp. 71-99.

Banasik J., Crook J.N., Thomas L.C., 1999, *Not if but when borrowers default*, The Journal of the Operational Research Society, 50(12), pp. 1185-1190.

Beyersmann J., Schumacher M., 2007, *Misspecified regression model for the subdistribution hazard of a competing risk by Latouche A, Boisson V, Chevret S and Porcher R.*, Statistics in Medicine 2006; Statistics in Medicine, 26, pp. 1649-1651.

Blanche P., 2015, *timeROC: Time-Dependent ROC Curve and AUC for Censored Survival Data*. URL https://CRAN.R-project.org/package=timeROC, R package version 0.3.

Blanche P., Dartigues J.F., Jacqmin-Gadda, H. 2013, *Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks*, Statistics in Medicine, 32(30), pp. 5381-5397.

Chang I.-S., Hsiung C.A., Wen C.-C., Wu Y.-J., Yang C.-C., 2007, *Non-parametric maximum-likelihood estimation in a semiparametric mixture model for competing-risks data*, Scandinavian Journal of Statistics, 34, pp. 870-895.

Chen Ch.-H., Chang I-S., Hsiung C.A. 2015, *NPMLEcmprsk: Type-Specific Failure Rate and Hazard Rate on Competing Risks Data*, URL https://cran.r-project.org/web/packages/NPML Ecmprsk, R package version 2.1.

Ciochetti D., Deng Y., Gao B., Yao R. 2002, *The termination of commercial mortgage contracts through prepayment and default: a proportional hazards approach with competing risks,* Real Estate Economics, 30(4), pp. 595-633.

Cox D., 1972, *Regression models and life-tables*, Journal of the Royal Statistical Society. Series B (Methodological), 34(2), pp. 187-220.

Cox D., Oakes D., 1984, *Analysis of Survival Data*, Chapman and Hall.

Crowder M., 2001, *Classical Competing Risks*, CRC Press, London.

Deng Y., Quigley J., Van Order R., 2000, *Mortgage termination, heterogeneity, and the exercise of mortgage options*, Econometrica, 68(2), pp. 275-307.

Fine J.P., Gray R.J., 1999, *A proportional hazards model for the subdistribution of a competing risk,* Journal of the American Statistical Association, 94, pp. 496-509.

Grambauer N., Schumacher M., Beyersmann J., 2010, *Proportional subdistribution hazards modeling offers a summary analysis, even if misspecified*, Statistics in Medicine, 29, pp. 875-884.

Gray B., 2014, *cmprsk: Subdistribution Analysis of Competing Risks*. URL http://CRAN.R-project.org/package=cmprsk, R package version 2.2-7.

Gray R., 1988, *A class of k-sample tests for comparing the cumulative incidence function in the presence of a competing risk,* The Annals of Statistics, 16, pp. 1141-1154.

Haller B., Schmidt G., Ulm K., 2013, *Applying competing risks regression models: an overview*, Lifetime Data Analysis, 19, pp. 33-58.

Højsgaard S., Halekoh U., Yan J., 2016, *geepack: Generalized Estimating Equation Package*, URL https://CRAN.R-project.org/package=geepack, R package version 1.2_1.

Klein J.P., Andersen P.K., 2005, *Regression modelling of competing risks data based on pseudo-values of the cumulative incidence function*, Biometrics, 61(1), pp. 223-229.

Kleinbaum D., Klein M., 2012, *Survival Analysis. A Self-Learning Text*, Third Edition, Springer--Verlag, New York.

Larson M.G., Dinse G.E., 1985, *A mixture model for the regression analysis of competing risks data*. Journal of the Royal Statistical Society. Series C, 34, pp. 201-211.

Lau B., Cole S.R., Moore R.D., Gange S.J., 2008, *Evaluating competing adverse and beneficial outcomes using a mixture model*, Statistics in Medicine, 27(21), pp. 4313-4327.

Lawless J.F., 2003*, Statistical Models and Methods for Lifetime Data*, John Wiley and Sons, Inc., New York.

Li J., Scheike T.H., Zhang M.J., 2015, *Checking Fine & Gray subdistribution hazards model with cumulative sums of residuals*, Lifetime Data Analysis, 21(2), pp. 197-217.

Liang K., Zeger S., 1986, *Longitudinal data analysis using generalized linear models*, Biometrika, 73(1), pp. 13-22.

Lin D.Y., Wei J.L., Ying Z., 1993, *Checking the Cox model with cumulative sums of Martingale based residuals*, Biometrika, 80(3), pp. 557-572.

Lindqvist H., 2008, *Competing Risks. Encyclopedia of Statistics in Quality and Reliability*, Wiley, Chichester.

Narain B., 1992, *Survival Analysis and the Credit-Granting Decision,* [in:] Thomas L., Crook J., Edelman D. (ed.), *Credit Scoring and Credit Control*, Oxford University Press, pp. 109-122.

Ng S.K., McLachlan G.J., 2003, *An EM-based semi-parametric mixture model approach to the regression analysis of competing-risks data*, Statistics in Medicine, 22, pp. 1097-1111.

Nicolaie M.A., van Houwelingen H.C., Putter H., 2010, *Vertical modeling: a pattern mixture approach for competing risks modeling*, Statistics in Medicine, 29, pp. 1190-1205.

Pavlov A.D., 2001, *Competing risks of mortgage termination: who refinances, who moves, and who defaults?*, The Journal of Real Estate Finance and Economics, 23, pp. 185-211.

Peterson A., 1976, *Bounds for a joint distribution function with fixed sub-distribution functions: application to competing risks*, Proceedings of the National Academy of Sciences of the United States of America, 73(1), pp. 11-13.

Pintilie M., 2006, *Competing Risks: a Practical Perspective*, Wiley, Chichester.

Pohar Perme M., Gerster M., Rodrigues K., 2017, *Package pseudo*, https://CRAN.R-project.org/package=pseudo, R package version 1.4.3.

Prentice R., Kalbfleisch J., Peterson A., Flournoy N., Farewell V., Breslow N., 1978, *The analysis of failure times in the presence of competing risks*, Biometrics 34, pp. 541-554.

R Development Core Team, 2017, *R*: *A Language and Environment for Statistical Computing*, the R Foundation for Statistical Computing, Vienna, Austria.

Sfumato P., Boher J.M., 2017, *goftte: Goodness-of-Fit for Time-to-Event Data*, https://CRAN.R-project.org/package=goftte, R package version 1.0.5.

Steinbuks J., 2015, *Effects of prepayment regulations on termination of subprime mortgages*, Journal of Banking and Finance, 59C, pp. 445-456.

Stepanova M., Thomas L., 2002, *Survival analysis methods for personal loan data*, Operations Research, 50(2), pp. 277-289.

Therneau T.M., 2017, *Survival: Survival analysis*, http://CRAN.R-project.org/package=survival, R package version 2.41-3.

Therneau T.M., Grambsch P.M., 2000, *Modeling Survival Data: Extending the Cox Model*, Springer, New York.

Tsiatis A., 1975, *A non-identifiability aspect of the problem of competing risks*, Proceedings of the National Academy of Sciences of the United States of America, 72(1), pp. 20-22.

ZASTOSOWANIE MODELI ZDARZEŃ KONKURUJĄCYCH DO OCENY RYZYKA KREDYTOWEGO

Streszczenie: Jednym z podstawowych zadań instytucji kredytowych jest ocena ryzyka kredytowego, którego podstawowym elementem jest ocena niewypłacalności kredytobiorcy. Wielkość portfela kredytowego może zmniejszać się w czasie z powodu nie tylko wystąpienia niewypłacalności, ale również wcześniejszych spłat kredytów. Zmienia to prawdopodobieństwo niewypłacalności w kolejnych okresach. Szacując prawdopodobieństwo niewypłacalności, należy więc uwzględnić prawdopodobieństwo wcześniejszych spłat w kolejnych okresach, co można osiągnąć za pomocą modeli zdarzeń konkurujących. W badaniu do oceny ryzyka niewypłacalności zaproponowano wybrane modele regresji dla zdarzeń konkurujących. Rozważane są modele: hazardu według przyczyny, hazardu subrozkładu, mieszanki modeli (podejście horyzontalne i wertykalne) oraz modele uogólnionych równań estymacyjnych GEE dla pseudoobserwacji. Badanie empiryczne przeprowadzono na próbie portfela kredytów udzielonych przez jedną z instytucji finansowych w Polsce.

Słowa kluczowe: model Coxa, model Fine'a-Graya, pseudoobserwacje, modele mieszane, modele wertykalne.