STATISTICS IN TRANSITION new series, March 2019
Vol. 20, No. 1, pp. 171–188, DOI 10.21307/stattrans-2019-010

SURVIVAL REGRESSION MODELS FOR SINGLE EVENTS AND COMPETING RISKS BASED ON PSEUDO-OBSERVATIONS

Ewa Wycinka[1], **Tomasz Jurkiewicz**[2]

ABSTRACT

Survival data is a special type of data that measures the time to an event of interest. The most important feature of survival data is the presence of censored observations. An observation is said to be right-censored if the time of the observation is, for some reason, shorter than the time to the event. If no censoring occurs in the data, standard statistical models can be used to analyse the data. Pseudo-observations can replace censored observations and thereby allow standard statistical models to be used.
In this paper, a pseudo-observation approach was applied to single-event and competing-risks analysis, with special attention paid to the properties of the pseudo-observations. In the empirical part of the study, the use of regression models based on pseudo-observations in credit-risk assessment was investigated. Default, defined as a delay in payment, was considered to be the event of interest, while prepayment of credit was treated as a possible competing risk. Credits that neither default nor are prepaid during the follow-up were censored observations. Typical application characteristics of the credit and creditor were the covariates in the regression model. In a sample of retail credits provided by a Polish financial institution, regression models based on pseudo-observations were built for the single-event and competing-risks approaches. Estimates and discriminatory power of these models were compared to the Cox PH and Fine-Gray models.

Key words: generalised estimating equations, cumulative incidence function, probability of default, credit risk, survival analysis.

1. Introduction

In the past few decades, survival analysis methods have become more widely used, not only in biostatistics, where their roots are, but also in many other branches of science, including economics, and the social sciences. Survival analysis is a term that covers a vast collection of different methods that focus on timing and duration prior to an event's occurrence (Mills, 2011). Among these methods are parametric and non-parametric estimation of survival time

[1] University of Gdańsk, Faculty of Management. E-mail: ewa.wycinka@ug.edu.pl. ORCID ID: https://orcid.org/0000-0002-5237-3488.
[2] University of Gdańsk, Faculty of Management. E-mail: tomasz.jurkiewicz@ug.edu.pl. ORCID ID: https://orcid.org/0000-0001-7066-5196.

distributions, and parametric and semiparametric regression models. The common goal of these methods is to handle censored observations that are inevitable in time-to-event analysis. A quite new and innovative approach to the problem of censoring is the idea of pseudo-observations that can replace both complete and censored actual observations. Pseudo-observations can be applied to many different objectives; this paper focuses on the usefulness of pseudo-observations in the development of regression models for survival functions in the case that a single event is analysed and in the competing risk analysis. The first objective was to review the properties of pseudo-observations in these two situations. The second goal was to compare the results and performance of the regression models for pseudo-observations with some of the more classical survival models that are currently most popular – in this case, the Cox Proportional Hazards model for single events (Cox, 1972) and the Fine-Gray model for competing risks (Fine and Gray, 1999).

2. Pseudo-observations for single events and competing risks

The methodology of pseudo-observations was first proposed by Andersen et al. (2003). The main idea of this approach is to replace censored observations by the function of event times $f(T)$, for which an expected value is $E(f(T))$. The condition is that an unbiased estimator $\hat{\theta}$ of $\theta = E(f(T))$ exists. Let n be the sample size $(i = 1, \dots, n)$. A pseudo-observation for $f(T)$ for individual i at a predefined series of time points $t = 1, \dots, H$ is defined as

$$\hat{\theta}_i(t) = n\hat{\theta}(t) - (n-1)\hat{\theta}^{(-i)}(t) \tag{1}$$

and is evaluated by the leave-one-out method. $\hat{\theta}(t)$ is the estimator in the sample of size n at time t, and $\hat{\theta}^{(-i)}(t)$ is the estimator at time t in the sample of size $n-1$, consisting of all units except the i-th individual. The pseudo-observation is then a contribution of the i-th unit to the $E(f(T))$ estimate in the sample of size n. Although the aim of using pseudo-observations is to replace the censored observations, pseudo-observations are calculated for all units in the sample (both completed and censored observations). Therefore, an $n \times H$ matrix of pseudo-observations is obtained. Subsequently, pseudo-observations are used as dependent variables in a generalised regression model with some link function g:

$$g\big(E(f(t)|X)\big) = \beta_0 + \sum \beta_j X_j = \beta^T X. \tag{2}$$

For each unit H pseudo-observations are calculated. Multiple measurement is a source of correlation in the data set; a possible solution to this deficiency would be to use generalised estimating equations (GEE), which are the generalisation of regression models for the case of correlated data (Andersen et al., 2003).

2.1. Single event

Assume that there is only one type of event and T is the time to that event, while T_C is the time to censoring. Due to the right censoring, we can observe $\min(T, T_C)$. The survival function is the probability that the unit does not experience the event until time t

$$S(t) = P(T > t). \tag{3}$$

In the survival analysis to the assumed sole type of event (single event), the survival function $S(t)$ can be estimated with the use of the Kaplan-Meier (KM) estimator

$$\hat{S}(t) = \prod_{t_j \le t}(1 - \frac{D_j}{N_j}), \tag{4}$$

where D_j is the number of events at time t_j, N_j is the number at risk just prior to time t_j, and t_j for $j = 1, \ldots, r$ $(r \le n)$ are distinct event times. The KM estimator is a maximum likelihood estimator (Klein and Moeschberger, 2003).

The i-th pseudo-observation based on the survival function is

$$\hat{\theta}_i(t) = n\hat{S}(t) - (n-1)\hat{S}^{(-i)}(t), \tag{5}$$

where $\hat{S}(t)$ is the estimated survival function at time t in a sample of size n and $\hat{S}^{-i}(t)$ is the estimated survival function derived from the $n-1$ sample (without the i-th observation) (Andersen and Perme 2010). At $t = 0$, the pseudo-observations for survival functions for all units are equal to one.

As t increases, the values of pseudo-observations for units in the cohort increase at each event time observed in the cohort (see Figure 1). Between any two successive event times, the values of pseudo-observations do not change. As a result, the curve of pseudo-observations over time for a particular unit is a step function with a varying length of steps depending on the successive event times. If the event for a unit is observed, the pseudo-observation drops below zero at the event time. At the subsequent time points, the unit that has just been excluded from the cohort has negative and increasing pseudo-values. If the unit is censored, then, beginning at the next event time after censoring, the values of pseudo-observations for that unit start decreasing. They remain, however, positive until the end of the follow-up (see Figure 1).

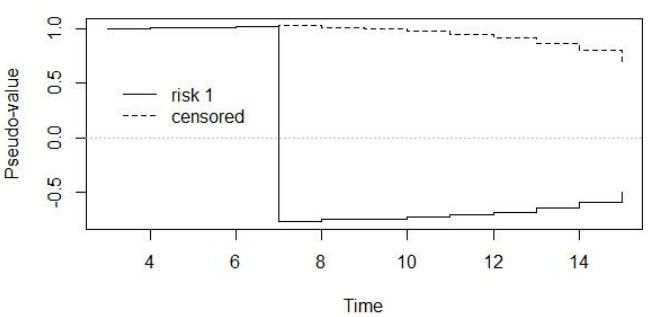

Figure 1. The pseudo-observations for the survival function over time in a censored data set for the individual with event time t=7 (risk 1) and the individual with censored time t_c=7 (censored)

As long as the units are in the cohort, they have similar pseudo-values. The values of pseudo-observations increase at each event time observed in the cohort. Therefore, the value of the pseudo-observation for the unit at its event

time is greater if the event occurred later in time. The later the event occurs, the greater the drop is (see Figure 2). The same pattern is observed if the observation is censored (see Figure 3).

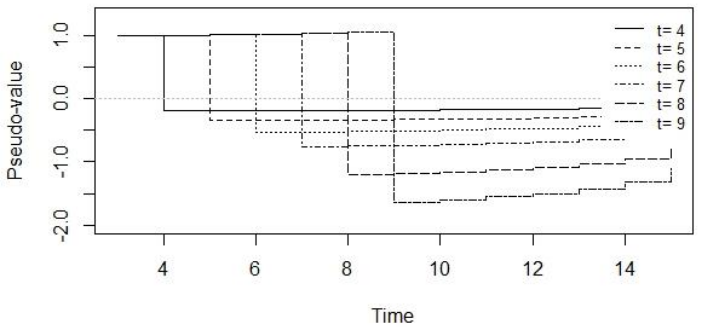

Figure 2. Pseudo-observations for the survival function over time for the units with event times t=4,..9

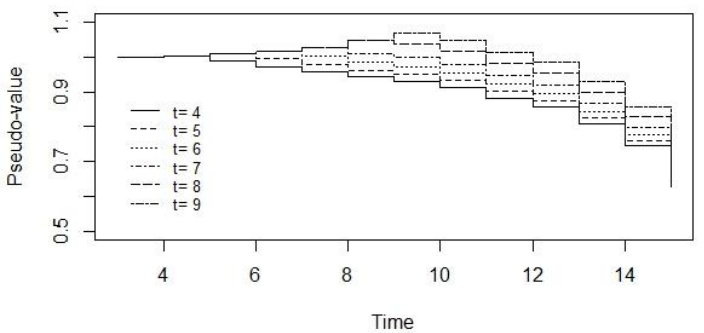

Figure 3. Pseudo-observations for the survival function over time for the units with censoring times C=4,..9

In the absence of censoring, the pseudo-value at time t reduces to the indicator that $T > t$. Therefore, the pseudo-observations are equal as long as the unit is observed in the cohort; after the event, the value of the pseudo-observation falls to zero and is constant until the end of the follow-up (see Figure 4). In this case, pseudo-observations are also independent.

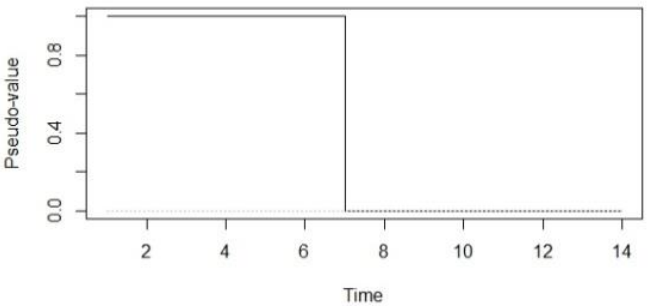

Figure 4. Pseudo-observations for the survival function over time for an individual with a survival time t=7 in a data set with no censoring

2.2. Competing risks

Let (T, C) be a bivariate random variable, such that T is a continuous variable representing the time of the first event, and $C = k$ $(k = 1, \dots, p)$ is a discrete variable denoting the type of event. If the time of the observation for some units is shorter than the time of the first event, we encounter right censoring. In such a situation, $C = 0$ and T_c is the time at which the observation was censored; what we only know is that $T > T_c$. Due to the right censoring, the variable (T, C) is only partially observable, and 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 some unverifiable assumptions (Pintilie, 2006, p. 41).

The subdistribution of event k (cumulative incidence function, CIF) is the probability, until time t, that event k will occur

$$F_k(t) = P(T \leq t, C = k). \tag{6}$$

The subdistribution is not a proper distribution because

$$\lim_{t \to \infty} F_k(t) = P(C = k) \leq 1. \tag{7}$$

The equality $P(C = k) = 1$ holds if there is only one type of event (no competing risks). The sum of the subdistributions for all types of events is a marginal distribution of the variable T

$$F(t) = P(T \leq t) = \sum_{k=1}^{p} F_k(t). \tag{8}$$

The maximum likelihood estimator of the subdistribution is

$$\hat{F}_k(t) = \sum_{t_j \leq t} \hat{h}_{kj} \hat{S}(t_{j-1}), \tag{9}$$

where $\hat{h}_{kj}$ is the cause-specific hazard at time t_j for event k. This can be defined as $\hat{h}_{kj} = \frac{D_{kj}}{N_j}$, where D_{kj} is the number of events of type k at time t_j, N_j is the number at risk just prior to time t_j, and t_j, for $j = 1, \dots, r$ $(r \leq n)$ are distinct event times. $\hat{S}(t_{j-1})$ is the survival function for all types of events just before time t_j.

It is worth noting that the estimate depends not only on the number of individuals who have experienced the k-th type of event, but also on the number of individuals who have not experienced any type of event (Binder et al., 2014). Usually, only one type of event is of interest and other types of event are treated as competing risks to it. In such a situation, it is reasonable to consider only two types of event: the event of interest (risk 1) and every other event combined (risk 2). This approach will be considered later in this paper.

The pseudo-observation for the unit i, at time t, for the event type k, based on the CIF, has the form

$$\hat{\theta}_{ik}(t) = n\hat{F}_k(t) - (n-1)\hat{F}_k^{(-i)}(t). \tag{10}$$

Here, $\hat{F}_k(t)$ is the estimated CIF for the k-th event at time t using all observations, and $\hat{F}_k^{(-i)}(t)$ is the estimated CIF derived from all but the i-th observation. When units are in a cohort, have the same pseudo-observation values for the CIF at subsequent times. At $t = 0$, a pseudo-observation for the CIF equals zero. Then, as time increases, pseudo-observations decrease, taking negative values. Figure 4 shows pseudo-observations over time for a unit that leaves the cohort at time 7. If the unit leaves the cohort due to an event of type 1, the pseudo-observation jumps above one at the time of the event, and then, at subsequent times, gradually decreases towards one. When the unit leaves the cohort due to an event of type 2, the pseudo-values remain negative and decreasing at all subsequent times (Andersen and Perme, 2010). If an individual is censored at time t, the pseudo-observations start increasing as of the next event time recorded in the data set (see Figure 5).

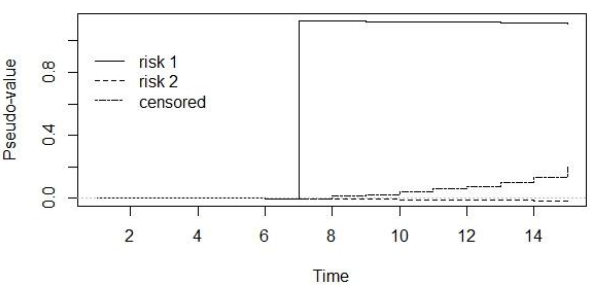

Figure 5. A comparison of the development of pseudo-values over time for three units that leave a cohort at time 7 due to either risk 1, risk 2, or censoring

As we compare the pseudo-observations for units with the same cause of leaving a cohort but at different times, we can see greater changes in the pseudo-values for later departures (see Figure 6). Jumps in the values of pseudo-observations are higher if the event of type 1 happens later, due to the reduction in time of the risk set.

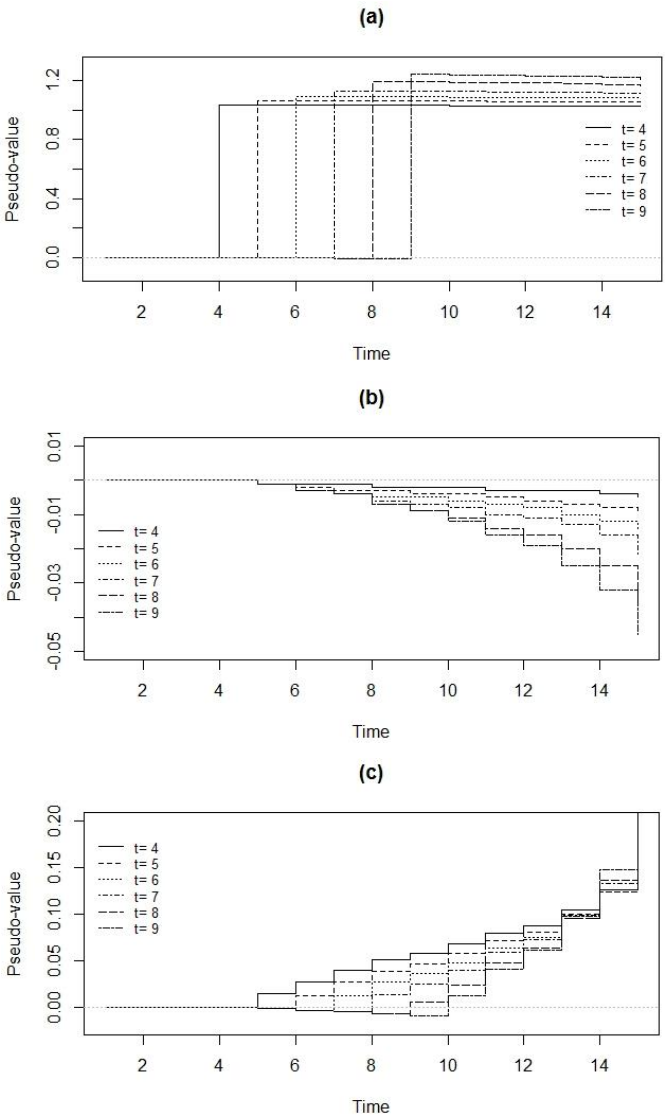

Figure 6. Pseudo-observations for units that experienced (a) risk 1 or (b) risk 2 or were (c) censored at different times (t=4,...9). Different scales are used in each figure

In a special case with no competing risks, the estimated CIF for the type-1 event reduces to the estimation of the distribution function ($\hat{F}_{CIF}$), and the survival function can be estimated as $\hat{S}_{CIF}(t) = 1 - \hat{F}_{CIF}(t)$. If survival functions are estimated directly with a Kaplan-Meier estimator or as $\hat{S}_{CIF}(t)$, the estimations are equal. However, as we show in the empirical part of the study, in the case of pseudo-observations based on these estimators, this equality no longer holds.

If no censoring occurs in the data set, the pseudo-observations for risk 1 reduce to the indicator $F_1(t) = 1[T_1 \leq t]$. They equal zero as long as the unit is in the cohort and rise towards one as the event of type 1 (risk 1) happens. The pseudo-observations for risk 2 equal zero at all time points, even after the occurrence of the event of type 2 (risk 2) (see Figure 7).

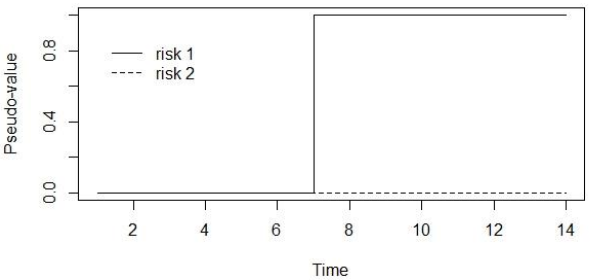

Figure 7. Pseudo-observations for the CIF for risk 1 and risk 2 over time in a data set with no censoring

3. Regression models based on pseudo-observations

For each unit there are H pseudo-observations – one for each predefined point in time. As a result, the data transformed into pseudo-observations is no longer independent, and generalised linear models (GLM) cannot be applied. Generalised estimating equations (GEEs) are the generalisation of GLM models for correlated data, as introduced by Liang and Zeger (1986). This is a method for analysing data collected in clusters where observations within a cluster may be correlated, but observations from different clusters are independent. The variance is a function of the expectation, and a monotone transformation of the expectation is linearly related to the explanatory variable (Højsgaard et al., 2005). The pseudo-observations are dependent variables in GLMs for a given link function $g(.)$. The regression model is

$$g\left(\hat{\theta}_k(t)\middle|X\right) = \beta_0 + \sum_{j=1}^{m+H} \beta_j X_j^* \tag{11}$$

Here, the vector X^* includes indicators of time points $X = (X_{m+1}, \ldots, X_{m+H})$ for $t = 1, \ldots, H$ (as dummy variables), as well as the covariates $X = (X_1, \ldots, X_m)$. When a complementary log-log link function is used, such as $g(x) = \log(-\log(x))$ for a single event, then the regression model has the form

$$\log\left(-\log\left(S(t|X)\right)\right) = \beta_0 + \sum_{j=1}^{m+H} \beta_j X_j^* \tag{12}$$

and can be depicted as

$$S(t|X) = \exp(-\exp(\beta_0 + \sum_{j=1}^{m+H} \beta_j X_j^*)). \tag{13}$$

Estimated coefficients for time points can be put into the model as time-dependent coefficients $\beta_0(t)$:

$$S(t|X) = \exp(-\exp(\beta_0 + \beta_0(t) + \sum_{j=1}^{m} \beta_j X_j)). \tag{14}$$

Finally, the survival function can be expressed as

$$S(\mathrm{t}|\mathrm{X}) = S_0(\mathrm{t})^{\exp\sum_{j=1}^{m}\beta_j X_j}, \tag{15}$$

which is a formula for the Cox PH model. Coefficients can be interpreted as a logarithm of a proportional hazards ratio.

In the case of competing risks, the link function $g(x) = \log(-\log(1-x))$ is used, and the regression model has the form

$$\log\big(-\log\big(1-F_k(t|X)\big)\big) = \beta_0 + \sum_{j=1}^{m+H}\beta_j X_j^*, \tag{16}$$

which can also be expressed in the form

$$F_k(t|X) = 1 - \exp(-\exp(\beta_0 + \beta_0(t) + \sum_{j=1}^{m}\beta_j X_j)). \tag{17}$$

This form is analogous to the proportional hazard model on the subdistribution hazard function in the Fine-Gray model. Coefficients β_j can be interpreted as logarithms of the subdistribution hazard ratios, if all covariates are time independent (Haller et al., 2013, p. 44).

Estimations of the parameters are based on the estimating equations

$$\sum_i \left(\frac{\partial}{\partial\beta} g^{-1}(\beta^T X_i^*)\right)^T V_i^{-1}\left(\hat{\theta}_i - g^{-1}(\beta^T X_i^*)\right) = 0. \tag{18}$$

Here, V_i is a working covariance matrix. The efficiency of the estimators depends on the choice of V_i matrix, which should resemble the true covariance. The GEE method fits marginal mean models and, as a result, only the correct specification of marginal means is required for the parameter estimation to be consistent and asymptotically normal (Højsgaard et al., 2005). The covariance structure does not need to be specified correctly; however, it is necessary to make an assumption about the type of this structure (considered the *working covariance matrix* or *working correlation matrix*). Four different types of working correlation matrix are usually considered.

The simplest – the independent working correlation structure – assumes that $\rho_{t_1,t_2} = corr\left(\hat{\theta}(t_1), \hat{\theta}(t_2)\right) = 0$ for each pair $\left(\hat{\theta}(t_1), \hat{\theta}(t_2)\right)$ and $t_1 \neq t_2$. The compound symmetry (exchangeable) structure treats $\rho_{t_1,t_2} = corr\left(\hat{\theta}(t_1), \hat{\theta}(t_2)\right)$ for all pairs as equal but unknown. The autoregressive structure of order 1 (AR1) has the form $corr\left(\hat{\theta}(t_1), \hat{\theta}(t_2)\right) = \rho^{t_1 - t_2}$, which reflects that observations further apart in time are less correlated. Finally, the unstructured working correlation matrix consists of a set of $corr\left(\hat{\theta}(t_1), \hat{\theta}(t_2)\right)$ that differs for each pair.

Agresti (2007) pointed out that if correlations are small, all working correlation structures yield similar estimates of parameters in GEE models and similar standard errors. In the Monte Carlo study, Klein and Andersen (2005) showed that there are no significant differences in estimations of GEE models for pseudo-observations with different working covariance matrices and recommended the use of the independent working covariance matrix.

The choice of the number of time points has little influence on the model fit. In the Monte Carlo simulations, Klein and Andersen (2005) showed that it is enough to choose five to ten time points, equally spaced on the event scale, to

evaluate pseudo-observations for the fitting model for the entire curve. Parameter estimates are quite insensitive to the number of time points. However, Andersen and Perme (2010) suggested that, nevertheless, all time points should be used if possible.

One of the problems with the implementation of GEE models is that GEE is a non–likelihood-based method. Therefore, information criteria such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) cannot be directly applied, which creates problems with the choice of best model. The GEE models for pseudo-observations with the log-log link function are analogues to the Cox PH and Fine-Gray models; therefore, in the empirical study, a variable selection, and consequently a choice of models, was performed for the last models. The Akaike selection criterion (Akaike 1974) was used to choose the best subset of covariates separately in the Cox PH and Fine-Gray models (Kuk and Varadhan, 2013). Subsequently, these sets of covariates were used in the equivalent GEE models.

4. Empirical study

We considered a cohort of 5,000 retail credits granted during 12 consecutive months by a Polish financial institution. All credits were granted for a fixed term of 24 months. The cohort was followed for 15 months from the moment the first credit was granted. Each credit could terminate in one of two ways: being completely paid back earlier than scheduled (early repayment) or by defaulting. A defaulted credit was considered one that had a delay in instalment payments of at least 90 days. We observed both types of termination in the cohort, as well as censoring. Censored observations were credits for which neither default nor early repayment were observed during the follow-up. That is, for those credits, all instalments were paid on time or with a delay shorter than 90 days. Figure 8 shows the distribution of events and censoring over the months of the credits' life, observed at the end of the follow-up.

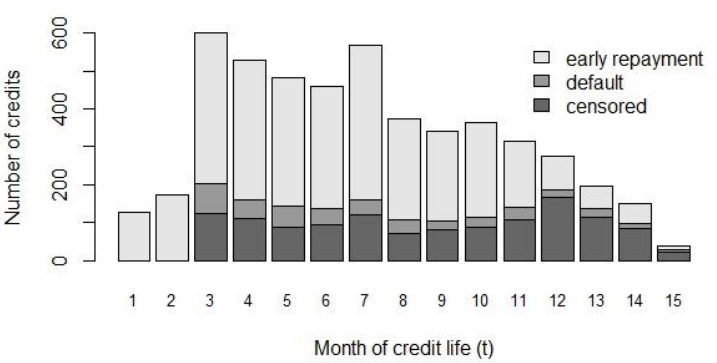

Figure 8. Distribution of the causes of termination during the follow-up of credits

Due to its definition, default cannot be observed for the first three months after credit granting. Early repayments and censoring were not lagged. Through the specificity of the analysed problem in the data set, we observed single events for particular time points with no censoring (t=1,2) and competing events with censoring (t=3,…,15). The objective was to evaluate one model of the probability of default for all time points in the presence of competing risk and heavy censoring.

To evaluate the probability of default, we laid down pseudo-observations and built GEE models for those pseudo-observations. Variables describing a creditor and a credit at the time of credit granting were used as covariates in these models. These variables included information such as age of the applicant, property, educational level, purpose of the credit, amount and instalment payments. To comply with the requirements of the financial institution sharing the data, the names of the variables were anonymised and are denoted in this paper by the letter X and one or more numbers.

All the variables were categorised and included in models as dummy variables. Two approaches were applied that resulted in different methods of assessing the pseudo-observations. The first assumed that the only type of analysed event was default; all other reasons for leaving the cohort of credits were considered to be censoring. In this approach, pseudo-observations were evaluated for the survival function (formula 5).

The second approach considered two causes of events: default and early repayment. Regular payments were handled as censored observations. Pseudo-observations were calculated (formula 10) for all event times due to the small number of analysed time points. To choose the set of covariates for GEE models, variable selection for the Cox PH model for single events, and the Fine-Gray model for competing events, was conducted at the first step with AIC using a stepwise algorithm (Venables and Ripley, 2002). Four different working covariance matrices were then applied.

Parameter estimations for all of the types of matrices were very close. Differences were only observed for estimates of parameters of dummy variables for time points, but these differences had no influence on the models' fit. The independence matrix was a slightly better fit for the model, and the results for this matrix are presented in the latter part of the paper. Table 1 shows the results of estimations of the GEE model for the CIF and estimates of the Fine-Gray model with the same set of covariates. Estimates of the parameters in both models are very similar. The GEE model, apart from covariates, also includes dummy variables for time points for which pseudo-values were calculated. Time point 1 (t1) is not included in the model because it is a reference group.

The CIF changes not only at the time of the considered event, but also at the time of the competing event. This is why time point 2 is included in the model, despite no default having occurred – this is one of the differences between competing- and single-event approaches. Values of the estimates for the subsequent time points increase, which is associated with higher wages for the units leaving a cohort later (compare Figure 6). Standard errors of estimators of covariates in the GEE model are slightly higher than for the Fine-Gray model; this observation is consistent with the findings of Andersen and Perme (2010).

Table 1. Estimates of the GEE model for the CIF and for the Fine-Gray model, both for the risk of default.

	GEE model for CIF								Fine-Gray model			
Time point	β	SE(β)	p-value	Cov.	β	SE(β)	p-value	Cov	β	SE(β)	p-value
Int.	-20.11	0.24	0.000	.	.	.	.	.	.	.	.
t2	-4.67	0.40	0.000	X1_2	-0.34	0.11	0.0027	X1_2	-0.30	0.10	0.0036
t3	17.27	0.40	0.000	X2_2	-0.19	0.13	0.1434	X2_2	-0.17	0.11	0.1300
t4	17.79	0.39	0.000	X2_3	-0.19	0.17	0.2651	X2_3	-0.31	0.15	0.0330
t5	18.18	0.39	0.000	X3_2	-0.30	0.12	0.0098	X3_2	-0.40	0.10	0.0001
t6	18.40	0.39	0.000	X4_2	0.28	0.12	0.0168	X4_2	0.21	0.10	0.0330
t7	18.57	0.39	0.000	X5_2	0.41	0.13	0.0013	X5_2	0.44	0.11	0.0000
t8	18.71	0.39	0.000	X6_2	-0.46	0.15	0.0016	X6_2	-0.61	0.13	0.0000
t9	18.78	0.39	0.000	X6_3	-1.51	0.16	0.0000	X6_3	-1.59	0.15	0.0000
t10	18.87	0.39	0.000	X7_1	-0.49	0.12	0.0000	X7_1	-0.52	0.10	0.0000
t11	18.98	0.39	0.000	X7_2	1.33	0.63	0.0355	X7_2	1.21	0.48	0.0120
t12	19.05	0.39	0.000	X8_1	-0.15	0.28	0.5859	X8_1	-0.31	0.22	0.1700
t13	19.19	0.39	0.000	X8_2	-0.55	0.26	0.0308	X8_2	-0.55	0.22	0.0110
t14	19.33	0.39	0.000	X9_1	0.22	0.15	0.1385	X9_1	0.27	0.13	0.0400
t15	19.58	0.39	0.000	X9_2	0.13	0.13	0.3032	X9_2	0.23	0.11	0.0350

Cov – covariate, Int. – intercept, t – dummy variable for a time point.

The purpose of a credit-risk assessment is not to find the size of the effect of a particular predictor on the risk of default, but to create a model which has the highest discriminatory ability and which allows prediction of the probability of default over the credit's life. To compare the performance of the above models, the following discrimination measures were used: area under the ROC curve (AUC), Kolmogorov-Smirnov test (KS), and Hand measure (H) (Hand, 2009). Additionally, significance tests of the differences between AUCs of both models were calculated (DeLong et al. 1988). Table 2 presents each model's performance at each of the event times. Both models have good and comparable discriminatory power through the whole credit-life. However, the best discrimination was achieved for the first months of credit-life; the slight advantage for the Fine-Gray model according to AUC is significant only for the last six months (see last column of Table 2).

Table 2. Measures of the performance of models for the CIF.

Month	H 95% CI		K-S 95% CI		AUC 95% CI		
	GEE	F-G	GEE	F-G	GEE	F-G	p-value
3	0.215 0.142-0.318	0.221 0.155-0.322	0.422 0.320-0.526	0.406 0.340-0.530	0.751 0.688-0.804	0.754 0.699-0.806	0.457
4	0.236 0.183-0.315	0.236 0.184-0.320	0.458 0.381-0.532	0.456 0.390-0.540	0.775 0.729-0.812	0.778 0.739-0.815	0.357
5	0.219 0.176-0.289	0.216 0.178-0.29	0.419 0.358-0.494	0.424 0.360-0.500	0.764 0.726-0.797	0.765 0.731-0.800	0.750
6	0.206 0.169-0.272	0.205 0.172-0.271	0.399 0.345-0.470	0.407 0.350-0.470	0.754 0.718-0.788	0.756 0.728-0.792	0.316
7	0.205 0.173-0.264	0.207 0.175-0.267	0.400 0.352-0.463	0.407 0.360-0.470	0.756 0.724-0.785	0.759 0.733-0.790	0.128
8	0.200 0.167-0.254	0.201 0.173-0.258	0.389 0.345-0.446	0.394 0.360-0.460	0.753 0.722-0.779	0.757 0.733-0.784	0.073
9	0.189 0.161-0.241	0.191 0.167-0.246	0.375 0.340-0.433	0.384 0.350-0.440	0.747 0.718-0.773	0.75 0.728-0.778	0.069
10	0.182 0.157-0.234	0.183 0.162-0.238	0.366 0.329-0.423	0.372 0.340-0.440	0.741 0.714-0.767	0.745 0.723-0.773	0.046
11	0.176 0.151-0.225	0.177 0.158-0.228	0.355 0.320-0.410	0.359 0.330-0.420	0.735 0.708-0.760	0.739 0.718-0.765	0.019
12	0.171 0.147-0.22	0.173 0.154-0.224	0.350 0.315-0.404	0.357 0.33-0.41	0.731 0.705-0.756	0.736 0.715-0.762	0.014
13	0.174 0.152-0.222	0.176 0.157-0.227	0.355 0.320-0.408	0.362 0.330-0.420	0.733 0.708-0.757	0.737 0.717-0.764	0.023
14	0.174 0.150-0.220	0.175 0.155-0.226	0.35 0.319-0.403	0.359 0.330-0.410	0.73 0.706-0.754	0.734 0.715-0.761	0.013
15	0.174 0.150-0.219	0.174 0.155-0.224	0.351 0.321-0.404	0.36 0.330-0.410	0.73 0.707-0.754	0.733 0.715-0.759	0.037

GEE- generalized estimating equations, F-G – Fine-Gray model, 95% CI – 95% confidence intervals as percentiles form 1000 bootstrapped samples

The application of the competing-risks methodology to credit-risk assessment is quite a recent idea (c.f. Watkins et al., 2014); it is more common to use single-event models (see Dirick et al., 2017). In the single-event approach, only time to default is considered, whereas credits that do not default until data-gathering are censored observations. However, in a credit-risk context, as a loan reaches maturity, default can no longer occur. Moreover, a very large proportion of the population will not go into default; hence, the basic principle in the survival analysis of one event type, that $S(t) \to 0$, does not hold. Therefore, in our study,

we should expect worse performance of single-event models than competing-events models for default. To verify this hypothesis, pseudo-observations for the survival functions were calculated with formula 5. Variable selection for the single-event model was performed using the AIC selection criterion for the Cox PH model. Estimates of the parameters of the Cox PH model and the GEE model for the survival function were calculated (see Table 3).

Table 3. Estimates of the GEE model for the survival function and estimates of the Cox PH model, both for the risk of default.

	GEE model for the survival function							Cox model			
Time point	β	SE(β)	p-value	Cov	β	SE(β)	p-value	Cov	β	SE(β)	p-value
Int	-2.42	0.29	0.000	X1_2	-0.48	0.15	0.001	X1_2	-0.34	0.10	0.001
t4	0.57	0.11	0.000	X2_2	-0.20	0.16	0.188	X2_2	-0.21	0.11	0.058
t5	1.03	0.14	0.000	X2_3	-0.11	0.25	0.653	X2_3	-0.35	0.15	0.020
t6	1.32	0.15	0.000	X3_2	-0.24	0.16	0.138	X3_2	-0.38	0.10	0.000
t7	1.55	0.15	0.000	X4_4	-0.21	0.29	0.460	X4_4	-0.37	0.15	0.010
t8	1.78	0.16	0.000	X5_2	0.28	0.16	0.074	X5_2	0.27	0.11	0.015
t9	1.89	0.16	0.000	X6_2	-0.70	0.23	0.003	X6_2	-0.59	0.13	0.000
t10	2.08	0.17	0.000	X6_3	-1.85	0.28	0.000	X6_3	-1.65	0.15	0.000
t11	2.32	0.17	0.000	X7_1	-0.44	0.16	0.006	X7_1	-0.48	0.11	0.000
t12	2.48	0.18	0.000	X7_2	1.68	0.79	0.033	X7_2	1.83	0.52	0.000
t13	2.82	0.19	0.000	X8_1	-0.04	0.41	0.931	X8_1	-0.35	0.23	0.123
t14	3.12	0.21	0.000	X8_2	-0.64	0.33	0.053	X8_2	-0.53	0.22	0.015
t15	3.68	0.29	0.000	X9_1	0.16	0.21	0.454	X9_1	0.26	0.13	0.055
.	.	.	.	X9_2	0.19	0.17	0.273	X9_2	0.33	0.11	0.004

Cov – covariate, Int. – intercept, t – dummy variable for a time point.

Dummy variables in a single-event approach were evaluated only for time points from 4 to 15. Time point 3 was omitted as the reference group, while time points 1 and 2 were not event times. The Akaike selection criterion applied to the Fine-Gray and Cox PH models gave almost the same set of covariates for both. The only difference was that variable X4_4 was applied to the single-event models instead of X4_2, which was used in the competing-events models. As a result, estimations of the parameters for all covariates in the models can be directly compared. As in the case of competing events, the GEE model for both the survival function and for the Cox PH model gave close estimations of parameters. The fit of the models also does not differ (see Table 4).

For both approaches, an interesting regularity was observed. For most of the covariates, p-values are greater for the GEE models for pseudo-observations than for Cox PH and Fine-Gray models; for some covariates, this resulted in a lack of significance, i.e. X2_3, X9_1, and X9_2 (compare Tables 1 and 3).

Table 4. Measures of the performance of models for the survival function.

Month	H 95% CI		K-S 95% CI		AUC 95% CI		
	GEE	F-G	GEE	F-G	GEE	F-G	p-value
3	0.199 0.130-0.302	0.203 0.148-0.317	0.433 0.314-0.526	0.418 0.329-0.528	0.751 0.680-0.800	0.749 0.694-0.805	0.832
4	0.229 0.163-0.296	0.226 0.175-0.309	0.454 0.365-0.525	0.449 0.384-0.534	0.773 0.720-0.803	0.773 0.734-0.811	0.923
5	0.207 0.157-0.267	0.207 0.171-0.277	0.42 0.346-0.484	0.414 0.359-0.492	0.762 0.718-0.788	0.761 0.728-0.797	0.892
6	0.199 0.148-0.252	0.195 0.163-0.264	0.406 0.332-0.459	0.401 0.351-0.468	0.751 0.708-0.777	0.752 0.722-0.787	0.778
7	0.195 0.152-0.244	0.197 0.166-0.259	0.403 0.339-0.456	0.403 0.360-0.469	0.753 0.714-0.777	0.755 0.729-0.785	0.531
8	0.193 0.152-0.241	0.193 0.166-0.251	0.391 0.333-0.445	0.390 0.352-0.458	0.751 0.714-0.773	0.753 0.729-0.781	0.489
9	0.184 0.147-0.23	0.184 0.161-0.24	0.380 0.327-0.433	0.379 0.347-0.447	0.746 0.711-0.769	0.748 0.725-0.775	0.452
10	0.179 0.142-0.222	0.180 0.159-0.235	0.370 0.318-0.421	0.369 0.337-0.434	0.74 0.706-0.763	0.744 0.722-0.771	0.226
11	0.173 0.138-0.218	0.175 0.155-0.225	0.356 0.306-0.405	0.357 0.325-0.420	0.733 0.698-0.755	0.738 0.717-0.764	0.101
12	0.169 0.133-0.211	0.171 0.151-0.223	0.35 0.301-0.400	0.353 0.320-0.412	0.729 0.695-0.751	0.734 0.714-0.761	0.056
13	0.172 0.139-0.215	0.174 0.154-0.225	0.354 0.305-0.403	0.358 0.325-0.415	0.731 0.700-0.753	0.736 0.715-0.762	0.070
14	0.17 0.140-0.214	0.172 0.152-0.221	0.35 0.303-0.400	0.354 0.323-0.411	0.727 0.697-0.751	0.732 0.714-0.759	0.064
15	0.171 0.141-0.215	0.169 0.149-0.218	0.353 0.307-0.401	0.354 0.321-0.408	0.728 0.700-0.751	0.731 0.713-0.757	0.210

GEE- generalized estimating equations, F-G – Fine-Gray model, 95% CI – 95% confidence intervals as percentiles form 1000 bootstrapped samples.

For the single-event approach, we also applied the method based on a reduction of the CIF to the case of one type of event. This led to the use of the $\hat{S}_{CIF}(t) = 1 - \hat{F}_{CIF}(t))$ estimator of the survival function, instead of the $\hat{S}_{KM}(t)$ estimator, in the calculation of the pseudo-observations. We observed that, for the pseudo-observations, the relation $\hat{S}_{KM}(t) = 1 - \hat{F}_{CIF}(t)$ does not hold. The differences between estimates were very low, but irregular. Figure 8 shows box-plots for the differences $\hat{S}_{KM}(t) - (1 - \hat{F}_{CIF}(t))$ for all the units at all time points. However, one should note that all the differences are very close to zero.

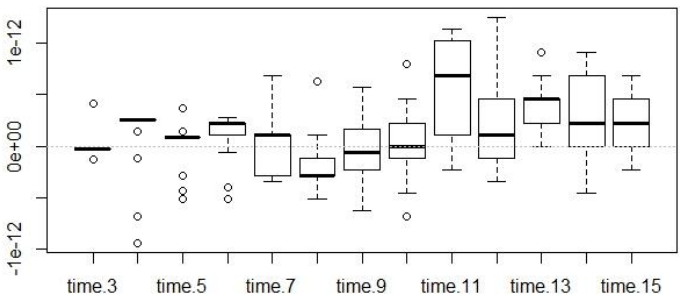

Figure 9. Distribution of differences between the KM estimator and the complement to CIF estimator for the survival function, both estimated by pseudo-observations

The GEE models for survival functions built for pseudo-observations based on both estimators gave exactly the same results. Thus, in spite of the above differences, the methods are fully interchangeable.

5. Conclusions

Pseudo-observations are a method that can be considered competitive with other survival analysis techniques. As shown in section 2, the values of pseudo-observations depend both on the type and time of event. Regression models for pseudo-observations correctly evaluate the whole survival curve, and the use of the log(-log) link function causes the GEE models for both single and competing approaches to simply mimic the results of the Cox PH and Fine-Gray models, respectively.

This observation is consistent with the results of earlier studies by other authors and argues against the use of a more cumbersome pseudo-values approach instead of more classic methods. However, because the independence matrix happened to be the best choice for the GEE model in all of the studies, it is suggested that pseudo-observations could be used as dependent variables in other methods for complete, independent data, such as classification trees.

In application to credit-risk assessment, competing-risks models had more discriminatory power than single-event models, which supports the use of competing-risks models in preference to models for single events. Further studies should focus on the variable-selection method that could be applied to the GEE models.

Acknowledgments

The authors gratefully acknowledge helpful feedback from an anonymous reviewer.

REFERENCES

AGRESTI, A., (2007). Logistic Regression, in An Introduction to Categorical Data Analysis, Second Edition, John Wiley & Sons, Inc., Hoboken, NJ, USA.

AKAIKE, H., (1974). A new look at the statistical model identification, IEEE Transactions on Automatic Control, 19 (6), pp. 716–723.

ANDERSEN, P. K., KLEIN, J. P., ROSTHØJ, S., (2003). Generalised linear models for correlated pseudo-observations, with applications to multi-state models, Biometrika, 90 (1), pp. 15–27.

ANDERSEN, P. K., PERME, M., (2010). Pseudo-observations in survival analysis, Statistical Methods in Medical Research 19 (1), pp. 71–99.

BINDER, N., GERDS, T. A., ANDERSEN, P. K., (2014). Pseudo-observations for competing risks with covariate dependent censoring, Lifetime data analysis, 20(2), pp. 303–315.

COX, D., (1972). Regression Models and Life-Tables, Journal of the Royal Statistical Society, Series B (Methodological), 34 (2), pp. 187–220

DELONG, E., DELONG, D., CLARKE-PEARSON, D., (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics 44, pp. 837–845.

DIRICK, L., CLAESKENS, G., BAESENS, B., (2017). Time to default in credit scoring using survival analysis: a benchmark study, J Oper Res Soc 6, pp. 652–655.

FINE, J., GRAY, R., (1999). A Proportional Hazards Model for the Subdistribution of a Competing Risk, Journal of the American Statistical Association, 94 (446), pp. 496–509.

HALLER, B., SCHMIDT, G., ULM, K., (2013). Applying competing risks regression models: an overview, Lifetime Data Anal 19, pp. 33–58.

HAND, D. J., (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve, Mach Learn 77, pp. 103–123.

HØJSGAARD, S., HALEKOH, U., YAN, J., (2005). The R Package geepack for Generalized Estimating Equations. Journal of Statistical Software, 15:2, pp. 1–11.

KLEIN, J., MOESCHBERGER, M., (2003). Survival Analysis: Techniques for Censored and Truncated Data, Statistics for Biology and Health, 2nd ed., Springer, New York.

KLEIN, J. P., ANDERSEN, P. K., (2005). Regression Modelling of Competing Risks Data Based on Pseudovalues of the Cumulative Incidence Function, Biometrics, 61 (1), pp. 223–229.

KUK, D., VARADHAN R., (2013). Model selection in competing risks regression, Statistics in Medicine 32, pp. 3077–3088.

LIANG, K., ZEGER, S., (1986). Longitudinal Data Analysis Using Generalized Linear Models. Biometrika, 73 (1), pp. 13–22.

MILLS, M., (2011). Introducing Survival and Event History Analysis, Sage, Los Angeles.

PINTILIE, M., (2006). Competing Risks: A Practical Perspective, Wiley.

VENABLES, W. N., RIPLEY, B. D., (2002). Modern Applied Statistics with S. Fourth edition, Springer.

WATKINS, J. G. T., VASNEV, A. L., GERLACH, R., (2014). Multiple Event Incidence and Duration Analysis for Credit Data Incorporating Non-Stochastic Loan Maturity, J. Appl. Econ., 29, pp. 627–648.