Competing risks models of default in the presence of early repayments

Ewa Wycinka

Department of Statistics
Faculty of Management
University of Gdansk

Edinburgh, 30.08.2017

Agenda

1. The idea of competing risks
2. Default and early repayment as competing risks
3. Some of the competing risks models
    1. Cause-specific hazard regression
    2. Subdistribution hazard regression
    3. Mixture model
    4. Vertical modeling
    5. Regression based on pseudo-observations
4. Data description
5. Results
6. Conclusions and further research

1. The idea of competing risks
First approach

Let's (T, C) be a bivariate random variable such as T is the continuous variable representing time of the first event and $C = i \ \ (i = 1, \ldots, p)$ is the discrete variable denoting the type of event. If time of observation for some units is shorter than time of the first event we meet the right censoring. In such situation $C = 0$ and T_c is the time at which the observation was censored. All we know is $T > T_c$. Due to the right censoring the variable (T, C) is only partially observable. We observe a pair $(\min(T, T_c), C)$. As the result the joint distribution of (T, C) is difficult to identify.

1. The idea of competing risks
First approach

Conditional distributions of the bivariate random variable can be expressed as

$$P(T = t|C = i) = \frac{P(T=t,C=i)}{P(C=i)}$$

and

$$P(C = i|T = t) = \frac{P(T=t,C=i)}{P(T=t)}$$

where $P(C = i)$ is the marginal distribution of type of event and $P(T = t)$ is the marginal distribution of the time of first event.

1. The idea of competing risks
First approach

Cumulative incidence function CIF (subdistribution) of the event i is the probability that event i will occur until time t

$$\mathrm{F_i(t) = P(T \leq t, C = i)}$$

Subhazard function

$$\tilde{h}_i(t) = \lim_{\partial t \to 0} \left(\frac{P(t<T \leq t+\partial t, C=i \mid T>t}{\partial t} \right) = \frac{f_i(t)}{S(t)},$$

is the hazard of event type i under condition that unit survived untill time t, being in risk of all types of events.

Sum of subhazards is equal to hazard function in marginal distribution of T.

1. The idea of competing risks
First approach

Hazard of subdistribution [Gray (1988)]:

$$h_i^*(t) = lim_{\partial t \to 0} \frac{P(t \leq T < t+\partial t, C=i | T \geq t \vee (T \leq t \wedge C \neq i))}{\partial t} = \frac{f_i(t)}{1-F_i(t)}$$

is the probability of the occurrence of event i in time interval ∂t, under condition that the entity hasn't experienced any event until time t or has experienced any other type of event before time t. Individuals failing before time t from any event different to i remain in the risk set for all future failure times.

CIF can be directly designated from hazard of subdistribution

$$\mathrm{F_i(t)} = 1 - \exp(-\int_0^t h_i^*(u)du)$$

1. The idea of competing risks
Second approach

Let's assume that T is a multivariate latent random variable $\boldsymbol{T} = (T_1, T_2, \ldots T_p)$ of p unobserved event times.
Distribution of $\boldsymbol{T}$: $G(\boldsymbol{T}) = P(\boldsymbol{T} \leq \mathbf{t})$ is the probability that each of T_i variables is lower or equal to $\boldsymbol{t} = (t_1, \mathrm{t}_2, \ldots, t_\mathrm{p})$.

For independent T_i variables $G(\boldsymbol{t}) = \prod_{i=1}^{p} \mathrm{G}_i(t_i)$, where $\mathrm{G}_i(t_i) = P(T_i \leq t_i)$ are marginal distributions.

Marginal distributions do not define the join distribution unless they are independent. This assumption cannot be verified because only first event $T = \min_i\{T_1, T_2, \ldots T_p\}$ can be observed.
Hazards of the marginal distribution (**cause-specific hazards**) are defined as

$$h(t) = -\frac{\partial ln\,(1-G_i(t))}{\partial t} = \frac{f_i(t)}{S_i(t)}$$

For the case of independence cause-specific hazards are equal to subhazards.

2. Default and early repayment as competing risks

- Default is a risk of interest whereas early repayment is a competing risk
- In some credit portfolios is excess of early repayments over defaults
- Time scale problem (tied observations)
- Regularly repaid loan or loan maturity is the source of heavy right censoring (at the end of follow-up period)

3. Some of the competing risks regression models

1. Cause-specific hazard regression [Prentice et al. 1978]
2. Subdistribution hazard regression [Fine and Gray 1999]
3. Mixture model [Larson and Dinse 1985]
4. Vertical modeling [Nicolaie et al. 2010]
5. Regression based on pseudo-observations [Andersen et al. 2003]

3.1. Cause-specific hazard regression

Cause-specific hazard can be modelled by semi-parametric Cox regression

$$h_i(t|\boldsymbol{X}) = h_{i0}(t)\exp\left(\sum_{k=1}^{m}\beta_k X_k\right)$$

In this approach all competing risks are assumed to be censored observations. Cause-specific hazard for event i is modelled as this event were the only possible one. The effect of covariates on the cause-specific hazard cannot be translated directly on the cumulative-incidence function.

3.2. Subdistribution hazard regression

Hazard of subdistribution can be modelled by Cox type regression

$$h_i{}^*(t|\boldsymbol{X}) = h_{i0}{}^*(t)\exp(\Sigma_{k=1}^{m}\beta_k{}^*X_k)$$

Cumulative incidence function can be directly estimated as:

$$\mathrm{F_i}(\mathrm{t}|\mathbf{X}) = 1 - \exp(-\int_0^t h_i^*(u|\boldsymbol{X})du$$

3.3. Mixture model

The joint distribution of event types is the product of marginal distribution of types of event and the conditional distribution of the times to the accordant event given the type of event.

$$P(T, C) = P(C)P(T|C)$$

$$P(C = 1|\boldsymbol{X}) = \frac{1}{1 + \exp(-(\alpha + \sum_{k=1}^{m} \pi_k X_k))}$$

$$\mathrm{F_i(t|\mathbf{X}) = P(T > t|X, C = i) = \exp\left(-\int_0^t h_i(u) exp\left(\sum_{k=1}^{m} {\beta_k}^* X_k\right) du\right)}$$

Expectation-maximization algorithm is used to estimate parameters of the model.

3.4. Vertical model

Joint distribution is the product of time to all types of events and conditional distribution for types of events given the event time.

$$P(T = t, C = i) = P(T = t) \cdot P(C = i | T = t)$$

$$\pi_i(t) = P(C = i | T = t)$$

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

$$\mathrm{F_i(t)} = \int_0^t \pi_i(u) h.(u) \exp(-\mathrm{H.}(u)) du$$

3.5. Regression based on pseudo-observations

Pseudo-observations are evaluated at a predefined series of timepoints $\tau = (\tau_1, \ldots, \tau_H)$

$$\hat{\theta}_{jh} = n\hat{F}(\tau_h) - (n-1)\hat{F}^{(k)}(\tau_h)$$

These pseudo-observations are dependent variables in the regression model

$$g\left(\hat{\theta}_{jh} \middle| x_{jh}\right) = \alpha_h + \sum_{k=1}^{m} \beta_{khj} X_{khj}$$

4. Data description

- 5000 consumer credit accounts from a 60-month retail loans portfolio of one of the Polish financial institutions.
- All loans have been observed for 24 months since loan origination.
- Default is defined as the first 90 days overdue.
- In the data set, there were:
  - 2274 creditors (45.5%) who repaid all 24 instalments (or have delay in payment shorter than 90 days),
  - 297 creditors (5.9%) who defaulted during the first 24 months
  - 2429 creditors (48.58%) who repaid the credit.
- Covariates - typical application characteristics such as: amount of credit, amount of the instalment, the purpose of the loan, age of the applicant, property, educational level (18 dummy variables representing 12 covariates).
- In all of the estimated models variable selection was performed using backward elimination selection.

5. Results

Table 1. Parameter estimates results from the Cox models

Model:	Cox model			
		95% CI		
Covariates	HR	Lower	Upper	p-value
X12	1,51	1,17	1,92	0,0015
X13	1,67	1,32	2,11	0,0000
X1_3	0,48	0,27	0,83	0,0086
X2_2	0,58	0,36	0,93	0,0251
X4_2	2,27	1,68	3,07	0,0000
X4_6	0,40	0,31	0,52	0,0000
X5_3	0,67	0,52	0,86	0,0018
X6_2	2,03	1,49	2,77	0,0000
X6_3	1,54	1,17	2,03	0,0019
X10_2	0,62	0,46	0,85	0,0028
X10_3	0,55	0,39	0,77	0,0006

Calculations with R (package: Survival)

5. Results (continued)

Table 2. Parameter estimates results from the Gray model

Model:	Gray model			
		95% CI		
Covariates	HR	Lower	Upper	p-value
X3	0,87	0,788	0,964	0,00740
X11	0,83	0,762	0,910	0,00005
X1_3	0,57	0,388	0,834	0,00380
X2_3	1,79	1,223	2,613	0,00270
X4_2	1,19	1,034	1,366	0,01500
X5_2	1,39	1,188	1,632	0,00004
X5_3	1,31	1,142	1,506	0,00013
X6_2	1,15	1,016	1,298	0,02600
X6_3	1,23	1,121	1,347	0,00001
X10_2	0,89	0,793	1,000	0,05000
X10_3	0,79	0,696	0,896	0,00026

Calculations with R (package: Survival)

5. Results (continued)

Table 3. Parameter estimates results from the mixture model

Components:	Mixture model (Event type - default)				Mixture model (Event time – to default)			
Covariates:	OR	95% CI		p-value	HR	95% CI		p-value
		Lower	Upper			Lower	Upper	
Intercept	7,48	4,26	13,15	0,0000	.	.	.	.
X3	9,08	4,77	17,31	0,0000	0,04	0,03	0,05	0,0000
X12	2,98	1,27	6,99	0,0062	0,03	0,02	0,03	0,0000
X13	4,46	1,97	10,08	0,0002	0,15	0,13	0,17	0,0000
X14	4,12	1,07	15,94	0,0200	0,35	0,29	0,41	0,0000
X2_2	4,51	2,59	7,85	0,0000	0,47	0,42	0,52	0,0000
X4_2	0,13	0,07	0,26	0,0000	0,44	0,37	0,52	0,0000
X4_6	2,86	1,54	5,30	0,0004	0,52	0,47	0,57	0,0000
X5_2	0,37	0,21	0,65	0,0003	0,76	0,68	0,85	0,0000
X6_3	0,22	0,12	0,39	0,0000	0,74	0,67	0,81	0,0000
X10_2	3,31	1,78	6,17	0,0001	0,82	0,76	0,89	0,0000

Calculations with R (package: NPMLEcmprsk)

5. Results (continued)

Table 4. Parameter estimates results from the vertical model

Components:	Vertical model (logit model)				Vertical model (Cox model)				
Covariates:	OR	95% CI		p-value	Covariates:	HR	95% CI		p-value
		Lower	Upper				Lower	Upper	
(Intercept)	0,07	0,03	0,15	0,0000	X3	0,92	0,85	0,99	0,0253
bs(T)1	1028,44	207,04	5108,64	0,0000	X11	0,93	0,88	1,00	0,0380
bs(T)2	0,00	0,00	0,00	0,0000	X1_3	0,79	0,67	0,92	0,0033
bs(T)3	2284,95	1105,14	4724,31	0,0000	X2_2	0,80	0,69	0,92	0,0025
X9	0,54	0,38	0,77	0,0005	X4_2	1,29	1,15	1,44	0,0000
X11	1,55	1,23	1,96	0,0002	X4_6	0,93	0,87	0,99	0,0258
X12	1,55	1,19	2,03	0,0013	X6_2	1,14	1,05	1,24	0,0020
X13	1,39	1,10	1,76	0,0059	X6_3	1,14	1,07	1,22	0,0001
X2_3	0,71	0,53	0,95	0,0228	X10_3	0,92	0,86	0,98	0,0079
X4_2	2,19	1,51	3,16	0,0000					
X4_6	0,49	0,37	0,64	0,0000					
X5_2	0,51	0,34	0,77	0,0013					
X5_3	0,34	0,24	0,48	0,0000					

Calculations with R (packages: survival, splines)

5. Results (continued)

Table 4. Parameter estimates results from the pseudo-values model

Covariates:	OR	95% CI		p-value	Covariates:	HR	95% CI		p-value
		Lower	Upper				Lower	Upper	
(Intercept)	0,01	0,00	0,01	0,0000	tpseudo=21	15,04	8,76	25,81	0,0000
tpseudo=4	1,48	1,12	1,97	0,0065	tpseudo=22	15,86	9,23	27,26	0,0000
tpseudo=5	2,19	1,47	3,24	0,0001	tpseudo=23	16,82	9,78	28,94	0,0000
tpseudo=6	2,93	1,88	4,57	0,0000	tpseudo=24	17,61	10,23	30,32	0,0000
tpseudo=7	3,82	2,38	6,13	0,0000	tpseudo=25	178,84	103,08	310,27	0,0000
tpseudo=8	4,75	2,91	7,75	0,0000	X3	1,19	1,05	1,35	0,0057
tpseudo=9	5,67	3,43	9,38	0,0000	X9	0,74	0,63	0,88	0,0004
tpseudo=10	6,74	4,04	11,25	0,0000	X11	1,18	1,06	1,31	0,0017
tpseudo=11	7,42	4,43	12,46	0,0000	X1.3	1,57	1,07	2,31	0,0205
tpseudo=12	8,39	4,98	14,15	0,0000	X2_3	0,60	0,40	0,88	0,0084
tpseudo=13	9,16	5,42	15,49	0,0000	X4_2	1,34	1,06	1,69	0,0148
tpseudo=14	10,26	6,04	17,42	0,0000	X4_6	0,82	0,74	0,91	0,0003
tpseudo=15	10,74	6,32	18,26	0,0000	X5_2	0,64	0,52	0,78	0,0000
tpseudo=16	11,53	6,77	19,65	0,0000	X5_3	0,61	0,52	0,72	0,0000
tpseudo=17	12,27	7,19	20,96	0,0000	X6_3	0,88	0,78	0,99	0,0387
tpseudo=18	12,90	7,54	22,05	0,0000					
tpseudo=19	13,54	7,91	23,19	0,0000					
tpseudo=20	14,24	8,31	24,41	0,0000					

Calculations with R (packages: geepack, pseudo)

5. Results (continued)

Table 3: Comparison of variables' selection by competing risks models

Variable	CoxPH	Gray	Mixture	Vertical modeling		Pseudo-values
				Cox	logit	
X3		+	+	+		+
X9					+	+
X11		+		+	+	+
X12	+		+		+	
X13	+		+		+	
X14			+			
X1_2						
X1_3		+		+		+
X2_2	+		+	+		
X2_3		+			+	+
X4_2	+	+	+	+	+	+
X4_6	+		+	+	+	+
X5_2		+	+		+	+
X5_3	+	+			+	+
X6_2	+	+		+		
X6_3	+	+	+	+		+
X10_2	+	+	+			
X10_3	+	+		+		

6. Conclusions and further research

- Cause-specific and subdistribution hazard regressions should be analysed together because:
  - In cause-specific hazard the effect of competing risk is not controled, therefore higher cause-specific hazard do not have to result in higher incidence rate
  - In subdistribution hazard regression the cumulative incidence function for the event of interest can changes also if units experience competing risk at time t. It could lead to improper interpretation of the covariates' effect
- Mixture models and vertical modeling are useful tool in exploration of data structure.

6. Conclusions and further research

- Vertical modeling adds informationto the analysis which event is more probable to occur given any event occurs. In this study relative risk for default was stable over time. It seems to be interesting to compare models' display in the credit portfolios with changing relative risk.
- Regression based on pseudo-observation needs finite number of time point what reflects the feature of the credit risk data, however presence of censored observations only in the last observed period has important impact on obtained results.

References

- Haller, B., Schmidt, G. & Ulm, K., Applying competing risks regression models: an overview, Lifetime Data Anal (2013) 19: 33. doi:10.1007/s10985-012-9230-8
- Klein J.P., Andersen P.K., Regression Modelling of Competing Risks Data Based on Pseudovalues of the Cumulative Incidence Function, Biometrics, 2005, Vol. 61, No. 1,
- Klein J.P. et al., SAS and R Functions to Compute Pseudo-values for Censored Data Regression, Comput Methods Program Biomed 2008; 89(3), 289-300
- Nicolaie M.A., Houwelingen H.C., Putter H., Vertical modelling: A pattern mixture approach for competing risks modelling, Statistics in Medicine, 2010, 29, 1190-1205
- Pintilie M., Competing Risks: A Practical Perspective, Wiley 2006
- Wycinka E., Time to default analysis in personal credit scoring, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu. 2015, 381, 527-536
- Wycinka E., Jurkiewicz T. , Mixture Cure Models in Prediction of Time to Default: Comparison with Logit and Cox Models (Pages 221-231) , „Contemporary Trends and Challenges in Finance. Proceedings from the 2nd Wroclaw International Conference in Finance, Editors: Jajuga K., Orlowski L., Staehr K. (Eds.), Springer 2017