The prediction of time to default for personal loans using mixture cure models: including macro-economic factors.

L. Dirick[1] T. Bellotti[2]
G. Claeskens[1] B. Baesens[1,3]

[1] Faculty of Economics and Business, KU Leuven, Belgium

[2] Department of Mathematics, Imperial College, London, UK

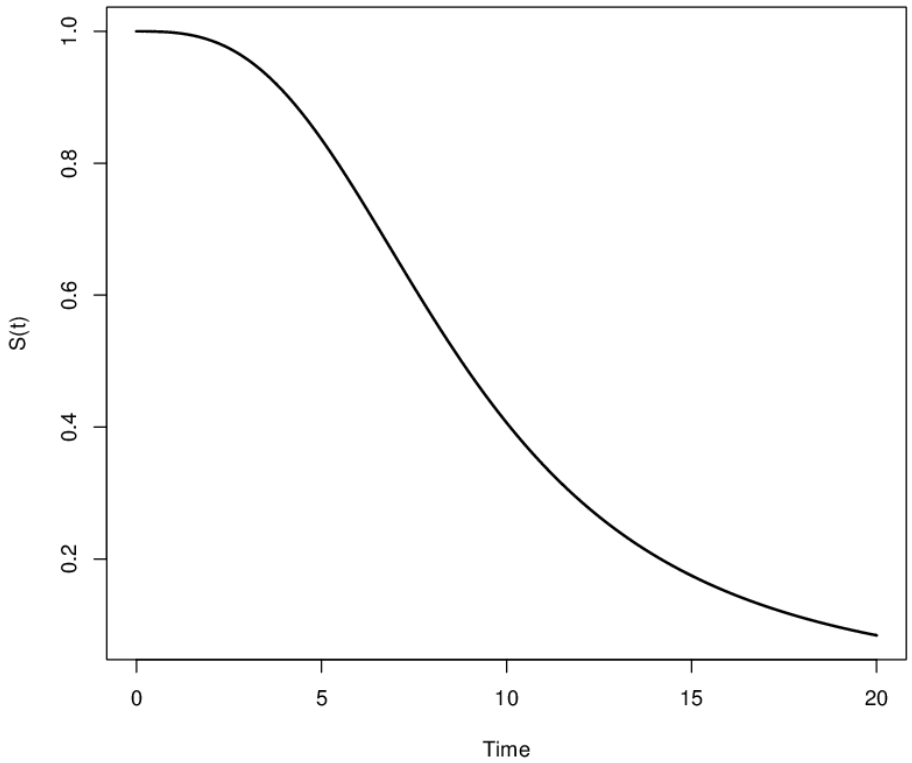
[3] School of Management, University of Southampton, UK

Credit scoring & credit control XIV conference, Edinburgh

Introduction

Why survival analysis?

- Basel II accord: more accurate credit risk calculations needed.
- Survival analysis: event of interest = default.
  - compute profitability over a customer's lifetime.
  - intuitive: possibility to model time aspect.



Survival analysis: issues

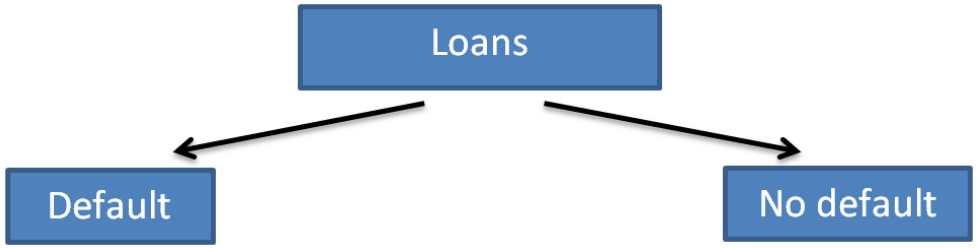
- Idea of survival in credit risk context $\neq$ medical survival.
- "The event might not be observed, but will take place eventually".
- For credit risk events such as default, this is not the case.

$\Rightarrow$ Solution: a mixture cure model to model *long-term survivors*.

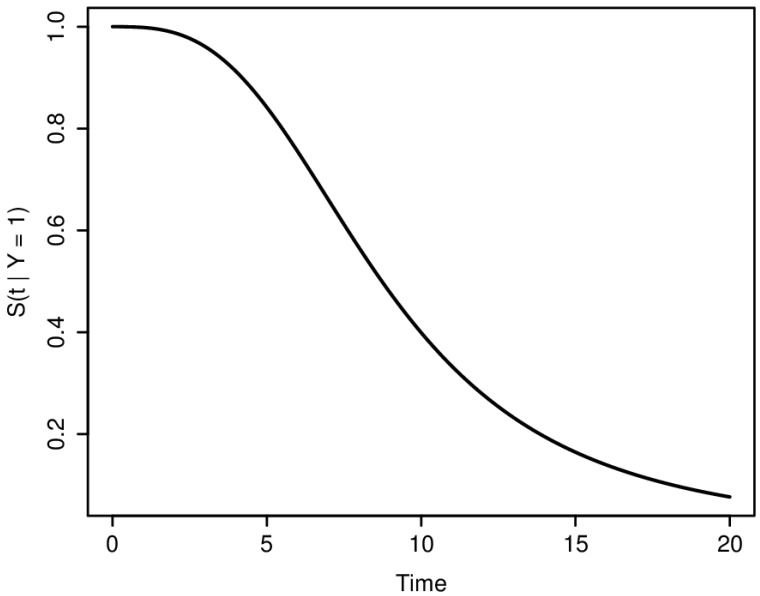
The classical mixture cure model

Given covariate vector $\boldsymbol{z}$, susceptibility indicator Y

$$S(t \mid \boldsymbol{z}) = \pi(\boldsymbol{z}) S(t \mid Y = 1, \boldsymbol{z}) + 1 - \pi(\boldsymbol{z})$$



prob$= \pi(\boldsymbol{z})$ prob$= 1 - \pi(\boldsymbol{z})$



The data structure . . .

in a classical mixture cure model (fixed loan term data, here 36 months).

subject	time	δ	z_1	. . .	z_m
1	22	1	"married"	. . .	3
2	36	0	"divorced"	. . .	2
⋮	⋮	⋮	⋮	⋱	⋮
n	7	0	"widowed"	. . .	6

With time-varying covariates (TVCs), in our paper macro-economic factors

subject	time	δ	z_1	. . .	z_m	$x_1(t{=}1)$	. . .	$x_1(t{=}36)$	. . .	$x_l(t{=}1)$	. . .	$x_l(t{=}36)$
1	22	1	"married"	. . .	3	0.3	. . .	0.5	. . .	3	. . .	1
2	36	0	"divorced"	. . .	2	0.7	. . .	0.3	. . .	2	. . .	2
⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋱	⋮	⋱	⋮	⋱	⋮
n	7	0	"widowed"	. . .	6	0.2	. . .	0.9	. . .	4	. . .	3

δ: 1 if default, 0 if no default OR not yet observed.

Data restructuring for inclusion of TVCs

KU LEUVEN

subject	interval	time	λ	δ	z_1	...	z_m	$x_1(t)$	...	$x_l(t)$
1	$[0,1]$	22	0	1	"married"	...	3	0.3	...	3
1	$[1,2]$	22	0	1	"married"	...	3	0.6	...	5
1	$\vdots$	22	0	1	"married"	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
1	$[21,22]$	22	1	1	"married"	...	3	0.5	...	1
2	$[0,1]$	36	0	0	"divorced"	...	2	0.7	...	2
2	$[1,2]$	36	0	0	"divorced"	...	2	0.1	...	3
2	$\vdots$	36	0	0	"divorced"	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
2	$[35,36]$	36	0	0	"divorced"	...	2	0.3	...	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	$[0,1]$	4	0	0	"widowed"	...	6	0.2	...	4
n	$[1,2]$	4	0	0	"widowed"	...	6	0.1	...	1
n	$\vdots$	4	0	0	"widowed"	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	$[6,7]$	4	0	0	"widowed"	...	6	0.9	...	3

Look at each subject on a monthly basis

λ: time-specific censoring indicator when TVCs in a mixture cure model

Model Formulation

KU LEUVEN

Model:

$$\boxed{S(t \mid \boldsymbol{z}, \boldsymbol{x}(t)) = \pi(\boldsymbol{z}) S(t \mid Y = 1, \boldsymbol{z}, \boldsymbol{x}(t)) + 1 - \pi(\boldsymbol{z})}$$

$\pi(\boldsymbol{z}) = P(Y = 1 \mid \boldsymbol{z}) =$ the *Incidence model component*, gives proportion of accounts susceptible to default, given $\boldsymbol{z} = (z_1, \ldots, z_m)$.

$$\log\left(\frac{\pi(\boldsymbol{z})}{1 - \pi(\boldsymbol{z})}\right) = \boldsymbol{z}'\boldsymbol{b} \Leftrightarrow \pi(\boldsymbol{z}') = \frac{\exp(\boldsymbol{z}'\boldsymbol{b})}{1 + \exp(\boldsymbol{z}'\boldsymbol{b})}$$

$S(t \mid Y = 1, \boldsymbol{z}, \boldsymbol{x}(t)) =$ the *Latency model component*, a conditional survival function, given $\boldsymbol{z} = (z_1, \ldots, z_m)$ **AND** $\boldsymbol{x}(t) = (x_1(t), \ldots, x_l(t))$

$$S(t \mid Y = 1, \boldsymbol{z}, \boldsymbol{x}(t)) \quad = \quad \exp\left(-\exp(\beta'\boldsymbol{z} + \beta_t'\boldsymbol{x}(t)) \int_0^t h_0(u \mid Y = 1) du\right),$$

About susceptibility Y...

KU LEUVEN

We are talking about susceptibility, however, susceptibility is not fully observed.

When $\delta = 0$: will default eventually occur or not?

$$Y = \begin{cases} 0 \text{ when an account is non-susceptible to default} \\ 1 \text{ when the account is susceptible to default} \end{cases}$$

Three possible states:

$$\begin{cases} \delta = 1 \text{ and } Y = 1 \\ \delta = 0 \text{ and } Y = 1 \\ \delta = 0 \text{ and } Y = 0 \end{cases}$$

δ fully observed, but Y only observed when $\delta = 1$.
Incomplete information: iterative EM-algorithm.

The likelihood function

KU LEUVEN

For each time interval j from observation i, following information: $\mathbf{O}_{ij} = (\mathbf{x}_i(t{=}j), \mathbf{z}_i, t_i, \delta_i, \lambda_{i,j})$. The complete likelihood is given by:

$$L(\mathbf{b}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\mathbf{z}_i)^{Y_i} (1 - \pi(\mathbf{z}_i))^{(1-Y_i)}$$
$$\times \prod_{j=1}^{k_i} h(t_i \mid Y_i = 1, \mathbf{x}_i)^{\lambda_{i,j} Y_i} S(t_i \mid Y_i = 1, \mathbf{x}_i)^{Y_i}$$

then the complete-data $\log L_c(\boldsymbol{b}, \boldsymbol{\beta}, \boldsymbol{\beta}_t \mid \boldsymbol{z})$ is the sum of

$$\log L_{inc}(\boldsymbol{b} \mid \boldsymbol{z}, Y) = \sum_{i=1}^{n} (1 - Y_i)\big(1 - \pi(\boldsymbol{z}_i)\big) + Y_i \, \pi(\boldsymbol{z}_i)$$

$$\log L_{lat}(\boldsymbol{\beta}, \boldsymbol{\beta}_t \mid \boldsymbol{z}, \boldsymbol{x}(t), Y) = \sum_{i=1}^{n} \sum_{j=1}^{k_i} Y_i \lambda_{i,j} \log h(t{=}j \mid Y_i{=}1, \boldsymbol{z}_i, \boldsymbol{x}_i(t{=}j))$$
$$+ Y_i \log S(t{=}j \mid Y_i{=}1, \boldsymbol{z}_i, \boldsymbol{x}_i(t{=}j)).$$

The EM-algorithm

KU LEUVEN

The p-th Expectation Step

Conditional probability of i-th observation being susceptible

Survival contributions of subject i: TVC-values at censoring time (t_i)

$$w_i^{(p)} = E(Y_i \mid \boldsymbol{\Theta}^{(p)}, \mathbf{O}) = \begin{cases} \dfrac{\pi(\boldsymbol{z}_i) S(t_i \mid Y_i = 1, \boldsymbol{z}_i, \boldsymbol{x}_i(t_i))}{\pi(\boldsymbol{z}_i) S(t_i \mid Y_i = 1, \boldsymbol{z}_i, \boldsymbol{x}_i(t_i)) + (1 - \pi(\boldsymbol{z}_i))} & \text{if } \delta_i = 0 \\ 1 & \text{if } \delta_i = 1. \end{cases}$$

The p-th Maximization Step

Maximizing $Q(\Theta^{(p+1)} \mid \Theta^{(p)}) = E[\log L_c(\boldsymbol{b}, \boldsymbol{\beta}, \boldsymbol{\beta}_t \mid \boldsymbol{z})]$,

parameter update $\hat{\Theta}^{(p+1)} = \left(\hat{\boldsymbol{b}}^{(p+1)}, \hat{\boldsymbol{\beta}}^{(p+1)}, \hat{\boldsymbol{\beta}}_t^{(p+1)}, \hat{\boldsymbol{S}}_0^{(p+1)}\right)$.

Denote $R(t_j)$ the individuals at risk in interval j, then

$$\hat{S}_0^{(p+1)}(t) = \exp\left(- \sum_{j : t_j \leq t} \frac{\sum_{i \epsilon R(t_j)}^{n} \lambda_{i,j}}{\sum_{i \epsilon R(t_j)}^{n} w_i^{(p)} \exp(\boldsymbol{\beta}'^{(p)} \boldsymbol{z}_i + \boldsymbol{\beta}_t'^{(p)} \boldsymbol{x}_i(t = j))} \right).$$

repeat until convergence.

Simulation settings

KU LEUVEN

- 100 replications and $n = 1000$.
- Event times exponentially distributed, rate= 0.7.
- Censoring times exponentially distributed, rate= 0.1/0.2 (Setting 1/2).
- Simulating TVCs: used method proposed by Austin (2012).
- Resulting censoring and unsusceptibility %:
  - → 32.9 and 22.7 for Setting 1
  - → 86.1 and 80.7 for Setting 2

Covariate distributions

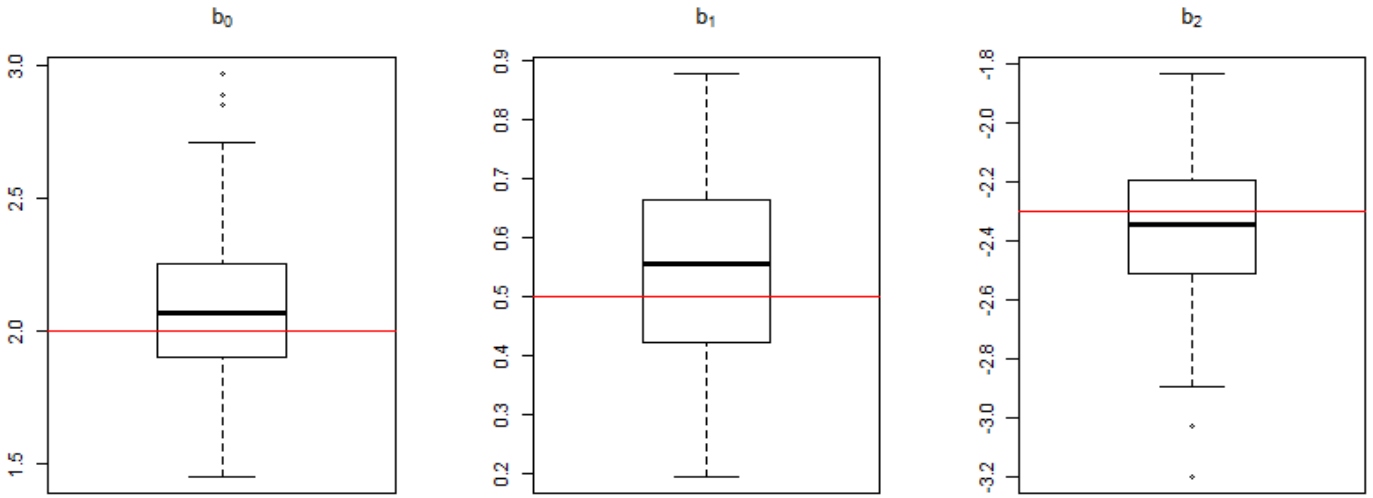
variable	z_1	z_2	$x_1(t)$	$x_2(t)$
Distr.	N(1.5, 0.6)	bin(1, 0.5)	$N(2, 0.5)$	$N(0.8, 0.5)$

Parameter values of the true model

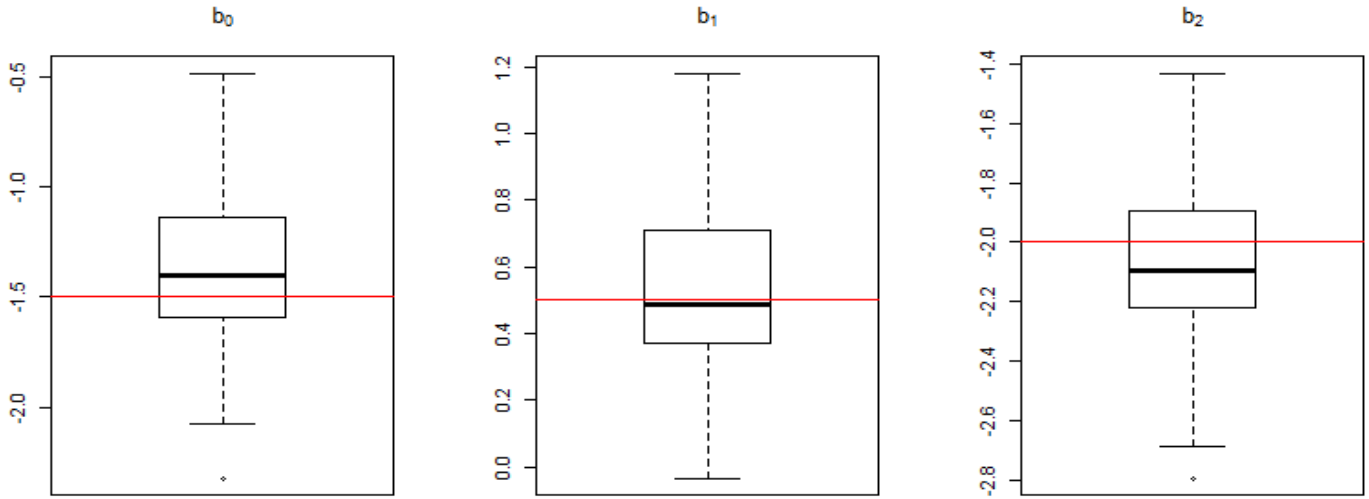
parameter	b_0	b_1	b_2	β_1	β_2	β_{t1}	β_{t2}
Setting 1	2	0.5	-2.3	-1.2	1	1	-0.7
Setting 2	-1.5	0.5	2	-1.2	1	1	-0.7

Boxplots *b*-parameters

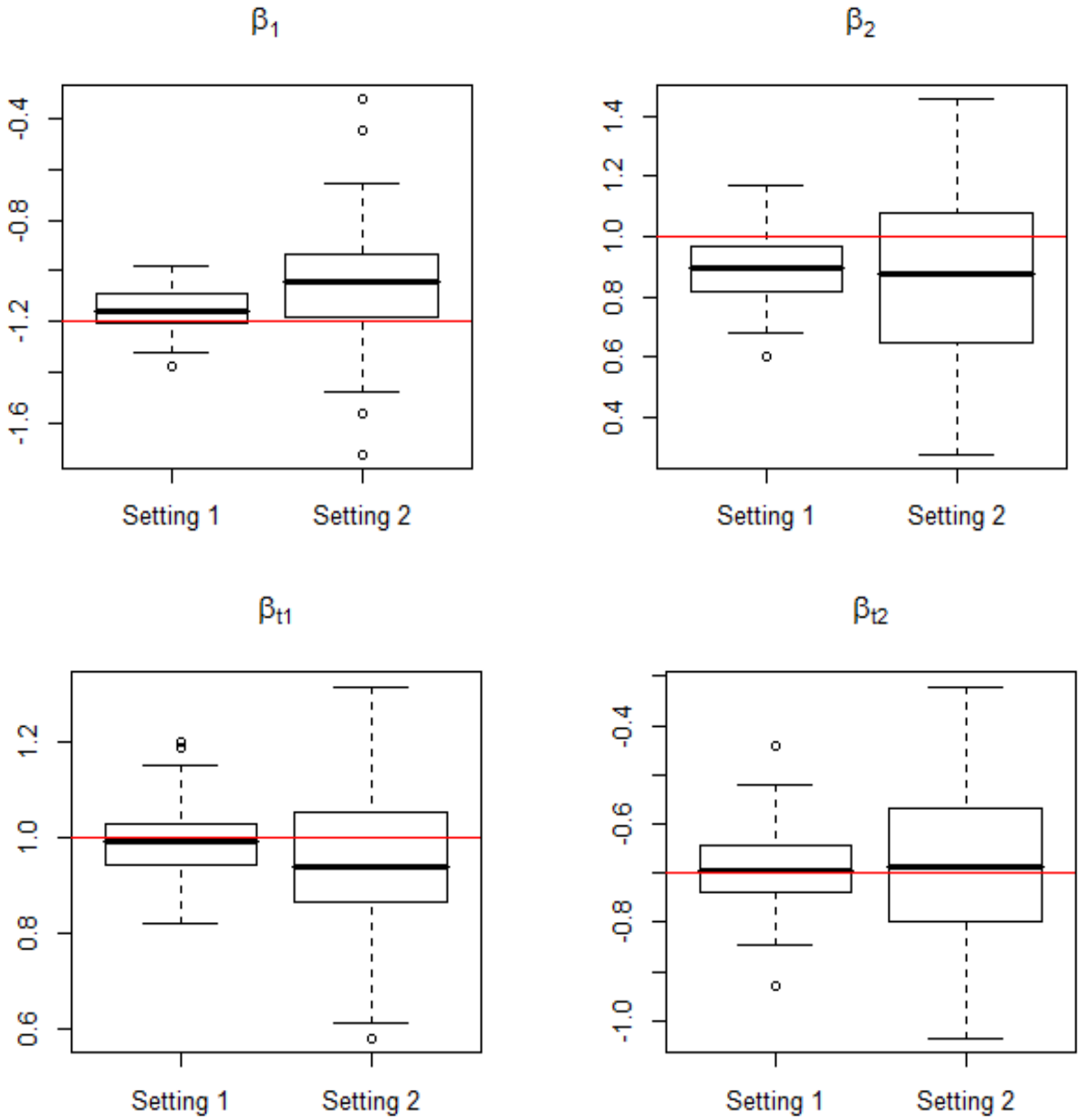
- Setting 1



- Setting 2



Boxplots β-parameters



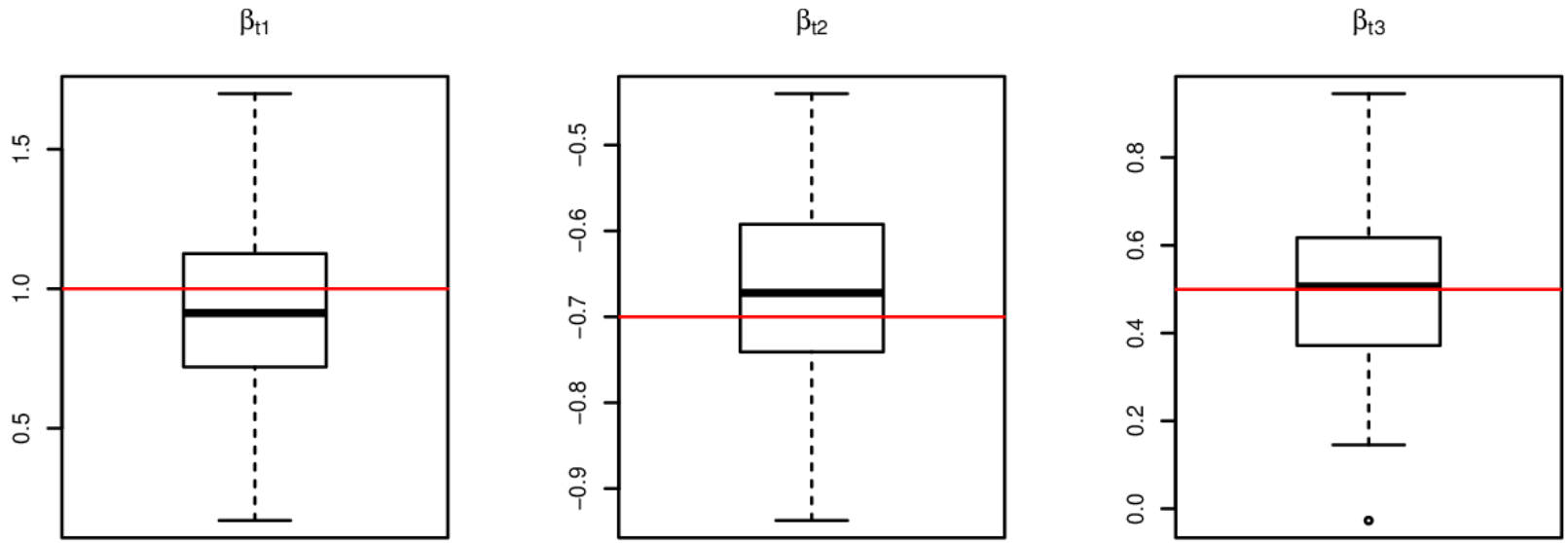
Setting 3

Test for correlated TVCs. Setting 3 = Setting 1, but ...

$$x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix}; \quad \mu = \begin{pmatrix} 2 \\ 0.8 \\ -0.7 \end{pmatrix}; \quad \Sigma = \begin{pmatrix} 0.7 & 0.8 & 0.8 \\ 0.8 & 1.2 & 0.8 \\ 0.8 & 0.8 & 1.0 \end{pmatrix}.$$

correlation matrix for the time-dependent covariates:

$$\rho = \begin{pmatrix} 1 & 0.873 & 0.956 \\ 0.873 & 1 & 0.730 \\ 0.956 & 0.730 & 1 \end{pmatrix}.$$



Data example

KU LEUVEN

- Data from a Belgian financial institution and add macro-economic factor-information from the National Bank of Belgium.
- Loans ran between 2004 and 2014.
- 20 000 observations, loan term= 36 months.
- 7 time-fixed covariates are included.
- We looked at 20 different models (and the model without TVCs), each time including 3 or 4 of the 6 macro-economic factors.

	Description
z_1	Annual income (per 1000)
z_2	age
z_3	Monthly child allowance (Y/N)
z_4	Number of years at current address
z_5	Total employment years
z_6	Application score
z_7	Mortgage on real estate (Y/N)

	Description	lag
TVC 1	Market interest rate	6
TVC 2	BEL 20 index	0
TVC 3	Consumer confidence	0
TVC 4	Gross Domestic product	0
TVC 5	Inflation rate	6
TVC 6	Unemployment rate	6

Parameter estimates

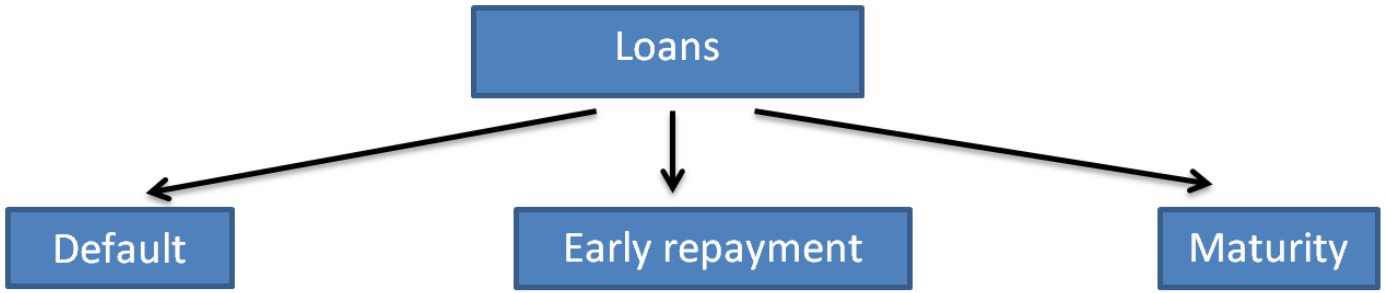
AICcd		(int)	z_1	z_2	z_3	z_4	z_5	z_6	z_7	IR	BEL 20	cons conf	GDP	inflation
no TVC	***b***	2.492	0.016	0.005	-0.076	-0.023	-0.039	-1.083	-1.319					
		(0.015)	(0.002)	(0.001)	(0.033)	(0.001)	(0.002)	(0.038)	(0.026)					
30127		***	***	***	**	***	***	***	***					
	β		-0.000	-0.011	-0.074	-0.024	-0.023	-0.615	-0.488					
			(0.005)	(0.003)	(0.077)	(0.004)	(0.006)	(0.104)	(0.092)					
			ns	***	ns	***	***	***	***					
best	***b***	3.55	0.046	0.058	0.192	0.009	-0.07	-1.406	-2.153					
		(0.085)	(0.002)	(0.002)	(0.037)	(0.002)	(0.002)	(0.04)	(0.033)					
26685		***	***	***	***	***	***	***	***					
	β		-0.011	-0.026	-0.188	-0.036	-0.016	-0.519	-0.119			-0.003	-0.011	0.001
			(0.005)	(0.003)	(0.076)	(0.004)	(0.006)	(0.101)	(0.077)			(0.005)	(0.021)	(0.041)
			*	***	*	***	**	***	ns			ns	ns	ns
2nd best	***b***	2.861	0.035	0.034	0.067	-0.003	-0.059	-1.23	-1.829					
		(0.079)	(0.002)	(0.001)	(0.036)	(0.002)	(0.002)	(0.039)	(0.032)					
27744		***	***	***	.	*	***	***	***					
	β		-0.008	-0.021	-0.133	-0.031	-0.016	-0.543	-0.188	0.071	-0.14		0.009	-0.033
			(0.005)	(0.003)	(0.077)	(0.004)	(0.005)	(0.1)	(0.075)	(0.036)	(0.057)		(0.022)	(0.041)
			.	***	.	***	**	***	*	.	*		ns	ns

- Lower AICcd for all models with macro-economic factors.
- Interest rate and BEL 20 index only significant TVCs.

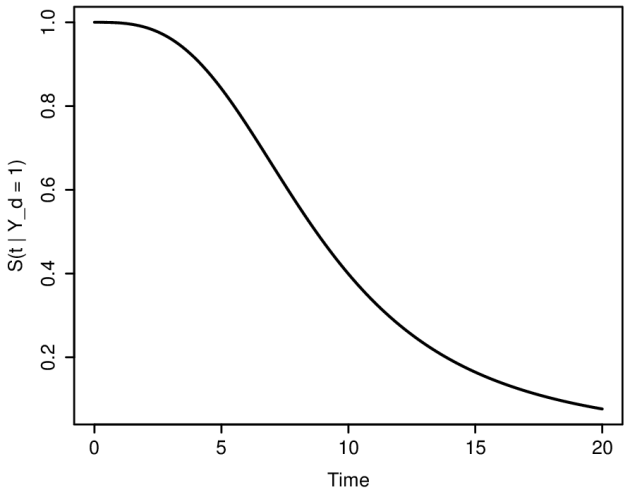
Extension: multiple events

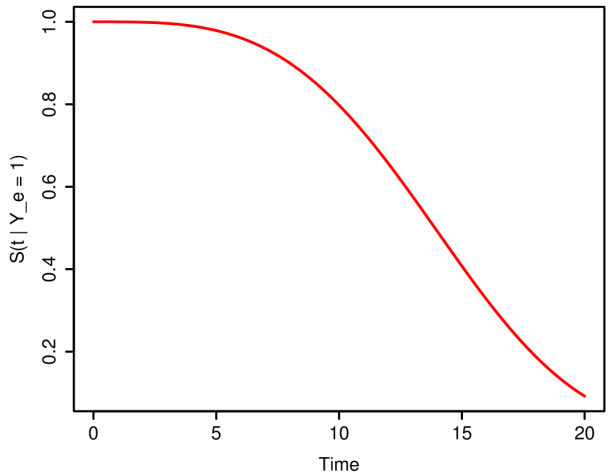
Given $\boldsymbol{z}, \boldsymbol{x}(t)$, default indicator Y_d, early repayment indicator Y_e

$$S(t \mid \boldsymbol{x}) = \pi_d(\boldsymbol{z}) S_d(t \mid Y_d = 1, \boldsymbol{z}, \boldsymbol{x}(t)) + \pi_e(\boldsymbol{z}) S_e(t \mid Y_e = 1, \boldsymbol{z}, \boldsymbol{x}(t)) + (1 - \pi_d(\boldsymbol{z}) - \pi_e(\boldsymbol{z}))$$



prob $= \pi_d(\boldsymbol{z})$ prob $= \pi_e(\boldsymbol{z})$ prob $= (1 - \pi_d(\boldsymbol{z}) - \pi_d(\boldsymbol{z}))$





Multiple events: data example

KU LEUVEN

d/e		(int)	z_1	z_2	z_3	z_4	z_5	z_6	z_7	IR	BEL 20	GDP	inflation
d	b	-0.033	0.006	-0.014	-0.103	-0.037	-0.047	-1.246	-1.208				
		0.163	0.004	0.003	0.078	0.004	0.005	0.1	0.083				
		ns	ns	***	ns	***	***	***	***				
	β		-0.001	-0.001	-0.049	-0.007	-0.006	-0.198	-0.107	0.039	0.011	0.007	0.029
			0.005	0.003	0.079	0.004	0.006	0.103	0.086	0.038	0.061	0.022	0.042
			ns	ns	ns	.	ns	.	ns	ns	ns	ns	ns
e	b	0.695	-0.007	-0.017	0.008	-0.017	-0.014	-0.741	-0.217				
		0.095	0.002	0.001	0.034	0.002	0.002	0.041	0.034				
		***	***	***	ns	***	***	***	***				
	β		0.002	-0.001	0.037	-0.002	-0.001	-0.176	-0.136	0.01	-0.056	0.023	0.022
			0.002	0.001	0.036	0.002	0.002	0.042	0.035	0.019	0.025	0.009	0.023
			ns	ns	ns	ns	ns	***	***	ns	*	*	ns

- More complex model: significance of some parameters affected
- In general, early repayment affected by BEL 20 index and GDP growth

Discussion

KU LEUVEN

- AICcd showed that including TVCs can lead to better model fit.
- Only limited number of TVCs are significant, eg. unemployment. insignificance may be due to selection bias.
- Information might be lost because monthly averaging: look at weekly default and TVCs.

References

- Dirick, L., Claeskens, G. and Baesens,B. (2015). An Akaike information criterion for multiple event mixture cure models. *Eur. J. Oper. Res.*, **241**:449–457.
- Tong, E. N. C., Mues, C., and Thomas, L. C. (2012). Mixture cure models in credit scoring: if and when borrowers default. *Eur. J. Oper. Res.*, **218**, 132–139.
- Cavanaugh, J. E. and Shumway, R. H. (1998). An Akaike information criterion for model selection in the presence of incomplete data. *J. Stat. Plan. Infer.*, **67**, 45–65.
- Stepanova, M. and Thomas, L. C. (2002). Survival analysis methods for personal loan data. *Operations Research*, **50 (2)**, 277–289.
- Watkins, J. G. T., Vasnev, A. L., and Gerlach, R. (2013). Multiple event incidence and duration analysis for credit data incorporating non-stochastic loan maturity. *J. Appl. Econ.*
- Bellotti, T. and Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis.*J. Oper. Res. Soc.*, 60(12):16991707.
- Andersen, P. K. (1992). Repeated assessment of risk factors in survival analysis. *Stat. Methods Med Res*, **1(3)**, 297–315.
- Austin, P. C. (2012). Generating survival times to simulate cox proportional hazards models with time-varying covariates. *Stat. Med.*, **31(29)**:3946–3958.