Beneva: Survival Models and Incomplete Data

Twelfth Industrial Problem Solving Workshop

G.Aflaki[1] É. Bizimana[2] A.Cwiling[3] F.Farhangian[4] J.H. McVittie[5] B. Monsia[6] J.F. Plante[1] J. Schulz[1] E. Tafolong[1]

[1]HEC - Université de Montréal

[2]Université Concordia

[3]MAP5, Université Paris Cité

[4]ETS - École de Technologie Supérieure

[5]University of Regina

[6]Université de Montréal

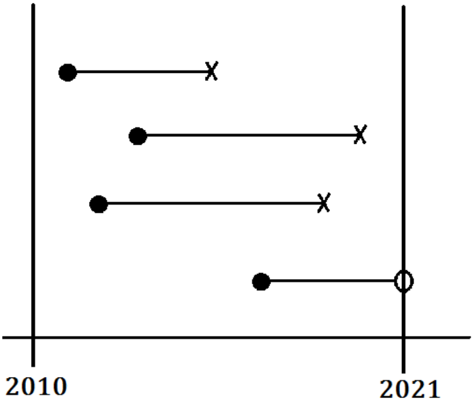
Outline

Outline

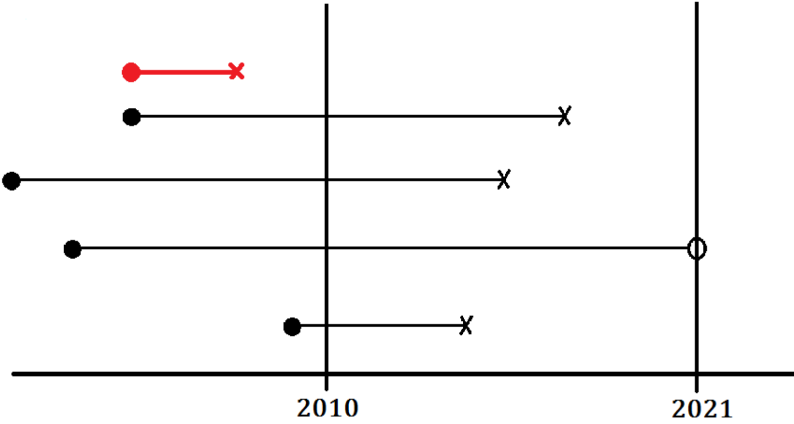
Beneva and Insurance Modelling

- Goal: modelling customer lifetimes (home insurance).
- The data:
  - Cross-section of "active" clients (*prevalent cohort*) as of 2010 + "new" clients (*incident cohort*) from 2010
  - All of these customers followed until the end of the study (2022)
  - For the "active" customers (*prevalent cohort*):
    - baseline covariates unavailable, only lagged covariates (measured at 2010)
  - For the "new" customers (*incident cohort*):
    - only baseline covariates are available

The data: incident cohort



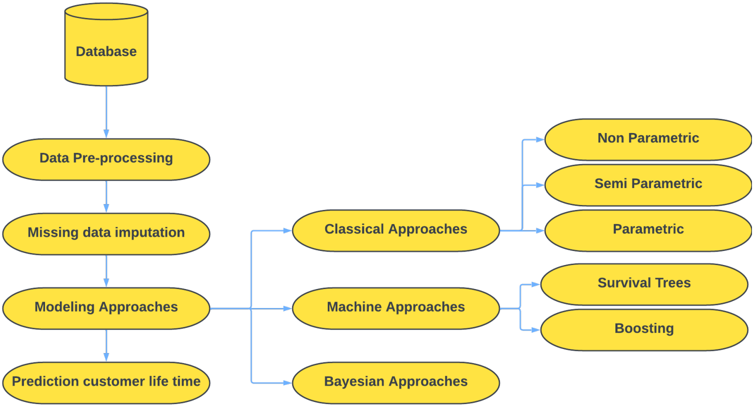
The data: prevalent cohort



The Problem

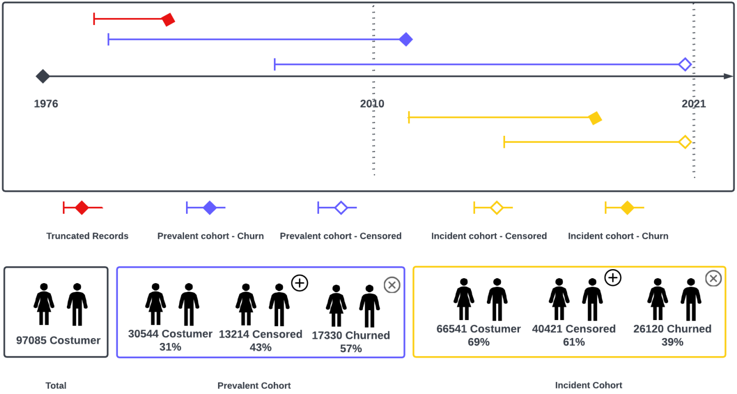
- The main difficulties with the analysis here:
  - Right censoring
  - Left truncation
  - Incomplete data: time-varying covariates only measured once, with missing data
  - Combination of prevalent + incident cohorts

Methodology overview



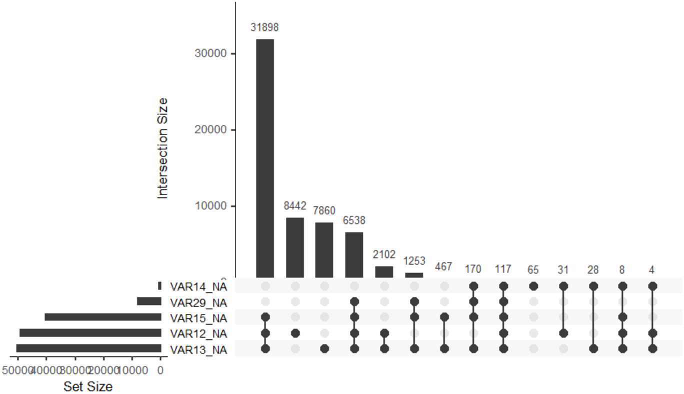
Outline

Exploratory data analysis



Exploratory data analysis

- 33 variables existed, 2 new variables added
  Baseline: 0: prevalent, 1: incident
  End Date: starting date and survived years aggregation
- 6 variables with missing values is detected that for 3 of these more or near 50% missing is reported.



Data Cleaning and Missing Imputation

- Records with end date before 2010 are removed
- Treatment for variables with missing is either
  Informative: 3 Variables
  Non-informative: 3 Variables
- Non-informative missing treatments:
  - complete-case analysis
  - imputation (single, multiple) (due to time constraint)
  - Joint Modelling

Limitations:

- covariates only measured at one point in time (2010 or baseline)
  - only observed either baseline covariates, or lagged covariates, but not both $\rightarrow$ unable to observe the evolution of the covariates
- all covariates are time-varying

Outline

Notation

$$\{(A_j, X_j, \delta_j) : T_j > A_j, j = 1, 2, \ldots, n\}$$

- $A_j = max\{2010 - O_j, 0\}$, where O_j is the entry time
- $X_j = \min\{T_j, C_j\}$, where T_j is the failure time and C_j is the censoring time (i.e., $C_j = 2022 - O_j$)
- $\delta_j = 1\{T_j < C_j\}$

Incident vs. Prevalent Cohorts

- Separate analyses:
  - models for the incident cohort (right censoring)
  - models for the prevalent cohort (right censoring + left-truncation $T_j > 2010 - O_j$)
  - not using information in both jointly
- Combined analysis:
  - truncation for prevalent cohort: $2010 - O_j$
  - truncation for incident cohort: 0
  - estimation reflects left-truncation + right-censoring
  - can be shown* that estimators based on combined data have nice properties

Non-parametric Model

Kaplan-Meier estimator:

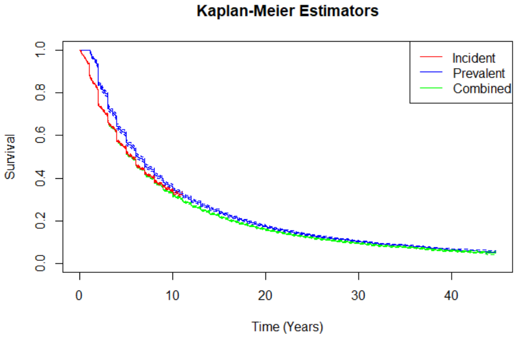
$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{m_i}\right)$$

where

- t_i are the distinct observed failure times, $t_1 < t_2 < \ldots < t_m$
- d_i are the number of failures at time t_i
- $m_i = \sum_j 1\{a_j \leq t_i \leq x_j\}$ (recall: a_j are the truncation times)

Non-parametric Model

- Incident cannot estimate past 12 years
- Prevalent and combined can estimate probabilities past 40 years
- Drops at integer times



Parametric Model

We assume the underlying failure times of all cohorts T_i have a common distribution function $F(\cdot;\boldsymbol{\theta}) = 1 - S(\cdot;\boldsymbol{\theta})$. We consider the Exponential, Weibull and Gamma distributions. The likelihood function for each cohort is as follows:

$$\mathcal{L}_{inc}(\boldsymbol{\theta}) \propto \prod_{i=1}^{n_{inc}} f(x_i;\boldsymbol{\theta})^{\delta_i} S(x_i;\boldsymbol{\theta})^{1-\delta_i}$$

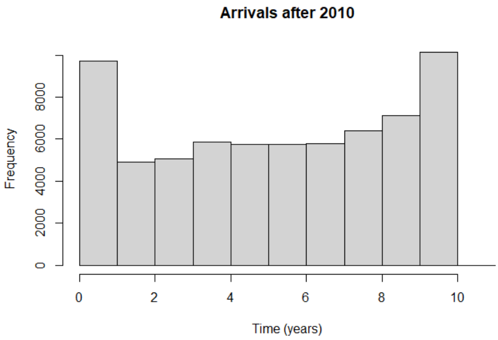
$$\mathcal{L}_{prev}(\boldsymbol{\theta}) \propto \prod_{k=1}^{n_{prev}} \frac{f(x_k;\boldsymbol{\theta})^{\delta_k} S(x_k;\boldsymbol{\theta})^{1-\delta_k}}{\mathbb{P}(T_k > A_k;\boldsymbol{\theta})}$$

The likelihood of the combined cohorts can be obtained from the combination of the two different likelihoods:

$$\mathcal{L}_{comb}(\boldsymbol{\theta}) = \mathcal{L}_{inc}(\boldsymbol{\theta}) \times \mathcal{L}_{prev}(\boldsymbol{\theta})$$

Parametric Model - Uniform Assumption

In the likelihood for the prevalent cohort, the denominator $\mathbb{P}(T_k > A_k; \theta)$ involves the random variable A_k. To handle it, we assume A_k follows the discrete uniform distribution.



Parametric Model - Uniform Assumption

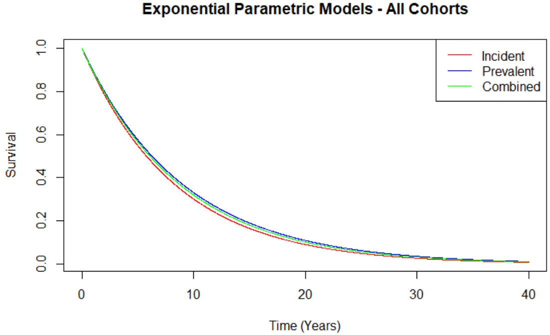
Due to the assumption, we can compute the denominator exactly.

$$\mathbb{P}(T > A; \theta) = \sum_{i=1}^{50} \mathbb{P}(T > A; \theta | A = i)\mathbb{P}(A = i) =$$

$$\frac{1}{50}\sum_{i=1}^{50} \mathbb{P}(T > i; \theta | A = i) = \frac{1}{50}\sum_{i=1}^{50} \mathbb{P}(T > i; \theta) = \frac{1}{50}\sum_{i=1}^{50} S(i; \theta)$$

Parametric Model - Results

- Can estimate survival probabilities for any horizon
- Smooth (not ideal for integer times)
- Combined curve lies between incident and prevalent curves



Semi-parametric Model

The Proportional Hazards (PH) model is a semi-parametric approach allowing the consideration of *covariates* **Z**.

$$\frac{\lambda(t|\mathbf{Z})}{\lambda_0(t)} = \exp(\mathbf{Z}\beta)$$

Partial likelihood for incident cohort:

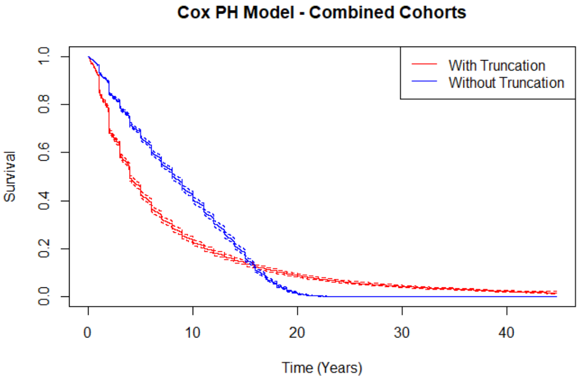
$$\mathcal{L}^{inc}(\beta) = \prod_{i=1}^{n_{inc}} \left(\frac{e^{\mathbf{Z}_i\beta}}{\sum_{j:X_i \le X_j} e^{\mathbf{Z}_j\beta}} \right)^{\delta_i}$$

Partial likelihood for prevalent cohort:

$$\mathcal{L}^{prev}(\beta) = \prod_{i=1}^{n_{prev}} \left(\frac{e^{\mathbf{Z}_i\beta}}{\sum_{j:A_j \le X_i \le X_j} e^{\mathbf{Z}_j\beta}} \right)^{\delta_i}$$

Semi-parametric Model

- With truncation: truncation time as usual
- Without truncation: truncation time considered as a covariate



Survival trees

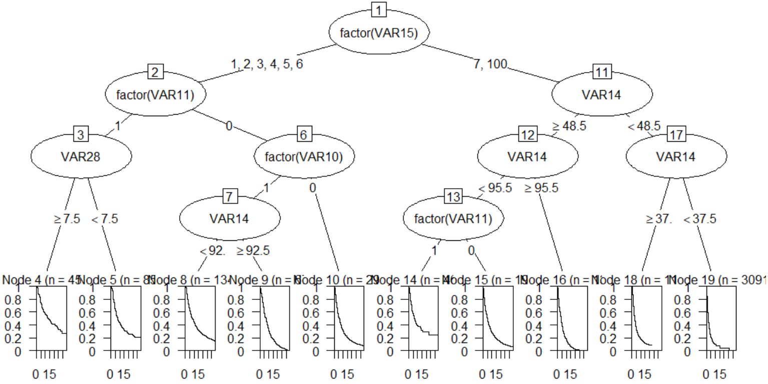
We are manipulating left-truncated right-censored data with missing information. Some approaches with random survival trees/forests:

- Random survival forest handling left-truncated right-censored data and imputing missing data simultaneously. $\rightarrow$ No package available yet.
- 1. Create a new covariate accounting for left truncation $A_j = max\{2010 - O_j, 0\}$, where O_j is the entry time.
  2. Random survival forest for right censored data which imputs missing data simultaneously. (Ishwaran, 2008)

  $\rightarrow$ Drawbacks: long running time + loss of information about left truncation *(clients who left Beneva prior to 2010 are not encapsulated)*
- 1. Impute missing data.
  2. Random survival tree/forest for left-truncated right-censored data. (Fu & Simonoff, 2017; Yao, 2022)

Survival trees

Imputed data set + survival tree for left-truncated right-censored data (Fu and Simonoff, 2017)



Boosting Indiviual Survival Distribution (ISD)

The heterogeneity of clients, coupled with the need to provide probabilistic estimates at several time points, has motivated the creation of several individual survival time distribution (ISD).

We use xgbse algorithm (an enhanced XGBoost ensemble model) for survival analysis to account for two properties:

- Prediction of survival curves for each client.
- Extrapolation over long-time horizon beyond the observational period.

Trees enables us to find the terminal leave for each client.

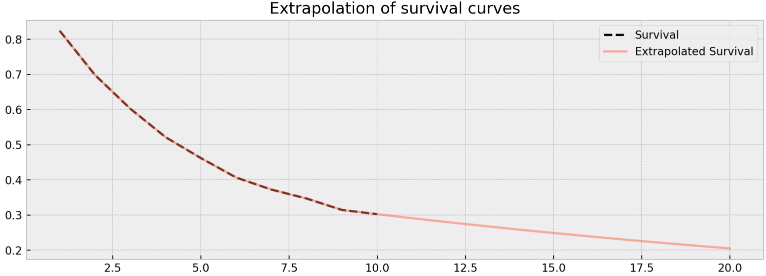
Boosting Individual Survival Distribution (ISD) : Results

The model output extrapolation over long time horizon illustrated bellow.

	Time					Extrapolation				
ID	1	2	...	9	10	11	12	....	19	20
0	0,98	0,94		0,72	0,72	0,72	0,72		0,72	0,72
1	0,89	0,85		0,43	0,43	0,43	0,43		0,43	0,43
2	0,79	0,59		0,26	0,26	0,26	0,26		0,26	0,26
3	0,82	0,70		0,31	0,30	0,29	0,28		0,21	0,20
4	0,97	0,81		0,42	0,40	0,38	0,36		0,24	0,23
5	0,87	0,66		0,25	0,17	0,11	0,07		0,00	0,00
6	0,98	0,89		0,62	0,62	0,62	0,62		0,62	0,62
7	0,98	0,85		0,37	0,33	0,28	0,25		0,10	0,08
8	0,86	0,74		0,36	0,34	0,32	0,29		0,18	0,17
9	0,77	0,69		0,33	0,26	0,21	0,16		0,03	0,02
10	0,91	0,80		0,36	0,32	0,28	0,24		0,10	0,08

Boosting Individual Survival Distribution (ISD) : Results

The model output extrapolation over long time horizon for given client is provided bellow.



Bayesian approach

There are several approaches within a Bayesian framework:

- Bayesian analysis for basic survival models (Weibull distribution, etc.) ;
- Bayesian analysis for regression models using complete covariable data (after imputation);
- Bayesian analysis for regression models that take account for missing covariables and the mechanism that describes the probability of missingness.

Challenges: There is no package that takes account for left truncation and right censoring data in our case: We have to implement these methods ourselves.

- We then implement the Weibull distribution model using Bayesian inference (Kundu & Mitra, 2016).

Weibull distribution based on left truncated and right censored data

- It is assumed that the lifetime has a Weibull distribution: $f(t) = \alpha\lambda t^{\alpha-1}\exp(-\lambda t^{\alpha}), t > 0$.
- Scale parameter λ follows Gamma distribution

$$\lambda \sim Gamma(a, b)$$

- Shape parameter α can be known and unknown
- Likelihood function is as follows:

$$L(\theta) = \prod_{i\in S_1} \{f(t_i;\theta)\}^{\delta_i}\{1 - F(t_i;\theta)\}^{(1-\delta_i)} \times$$

$$\prod_{i\in S_2} \left\{\frac{f(t_i,\theta)}{1 - F(\tau_{iL};\theta)}\right\}^{\delta_i} \left\{\frac{1 - F(t_i;\theta)}{1 - F(\tau_{iL};\theta)}\right\}^{1-\delta_i}$$

Known Alpha

For the sake of simplicity we will continue with α being known

- (fixed α as MLE from classical parametric model)
- Parameters, adaption is considered
- λ estimate when α is fixed is as follows:

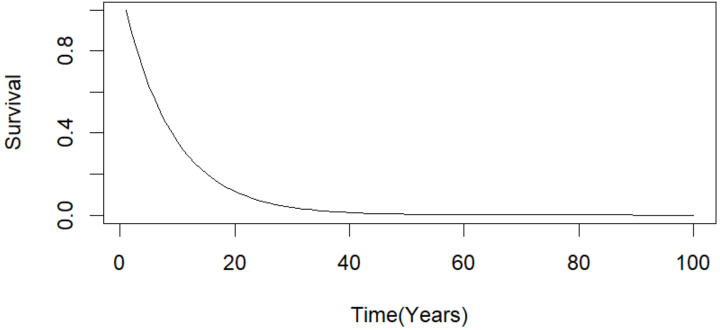
$$\hat{\lambda} = \frac{a + m}{b + \sum_{i \in S} t_i^{\alpha} - \sum_{i \in S_2} {\tau_{iL}}^{\alpha}}$$

- Survival function for $t > t_i$ is as:

$$S(t|t_i, \alpha, \lambda) = e^{-\lambda(t^{\alpha} - {t_i}^{\alpha})}$$

Prediction

Fixing the value for α and estimation λ based on Beneva's data set let us find survival function for any new individual (starting time 0)



Outline

Discussion

- This is a difficult problem!
  - The idea of combining prevalent + incident cohorts
  - The issue of missing time-varying covariates remains unsolved.
- We hope that our work offers some potential solutions / future paths to further explore for Beneva