Using Survival Analysis to Model Credit Risk

3/30/21 12:00 PM

Speaker: Lore Dirick (Flatiron School)
Date: 03/30/2021

Credit Risk Modeling

Loss = PD * EAD * LGD

- PD: Probability of Default
- EAD: Exposure at Default (total value a bank is exposed to when a loan is default)
- LGD: Loss given Default

Logistic Regression:

- Prediction on probability of each individual
- Sum over each individuals only

Survival Analysis:

- Sum over each time, each individual
- So more accurate

How is PD typically modeled?

- Binary Classification
- Data collection: grant loans to all, wait until repaid, 2 possible statuses

Survival Analysis

Medical Data (Original case)

	Used to create KM curve		Predictors in model			
	Time of last patient update (day)	Status	Hormonal therapy	age	menopause	...
Patient 1	403	1	0	55	1	
Patient 2	2237	0	0	64	1	
Patient 3	42	0	1	37	0	
...						

- Status = 0: it could still happen, but we just haven't observed it yet

Credit Risk Data

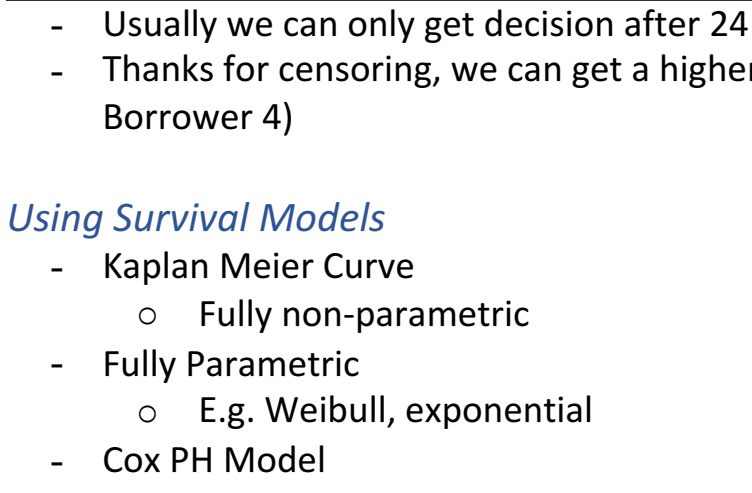
	Default	Time (months)	Years at current Address	children	Age	...
Borrower 1	1	8	3	1	30	
Borrower 2	0	24	2	0	39	
Borrower 3	0	24	22	2	75	
Borrower 4	0	13	10	1	62	
...						

- Usually we can only get decision after 24 months (2 year loan paid in full)
- Thanks for censoring, we can get a higher n (censoring allows to bring more data, i.e. Borrower 4)

Using Survival Models

- Kaplan Meier Curve
  - Fully non-parametric
- Fully Parametric
  - E.g. Weibull, exponential
- Cox PH Model
  - Semi-parametric model
    - Non-parametric baseline hazard
    - Parametric component
  - Hazards between individuals are proportional

Modeling Outcome



Survival Analysis	Typical Classification Models
Multiple PD estimates, 1 for each repayment period	1 PD estimate
More data because of censoring	Censored data cannot be included

Extensions:

- Cure models
  - A large fraction of population would never default, so adjust the curve to converge to a value larger than 0