Spline based survival model for credit risk modeling

Authors: Sirong Luo*, Xiao Kong, Tingting Nie
European Journal of Operational Research, 2016

Presenter: Shuai Li

February 25, 2025

Guidelines

1. Introduction
2. Preliminary
3. Spline Based Multinomial Logistic Regression Survival Model (SMLRS)
4. Numerical Example - Comparison of SMLRS and Cox Models

Note: For the convenience of discussion, the No. of figures and tables in this slide are consistent with those in the original paper.

Introduction

Research Background:

- Consumer credit has been a driving force in the economy of most developed countries for the last 50 years.
- In 2015, US consumer debt reached $11.85 trillion, with $2.08 trillion in consumer credit.
- The rapid growth of consumer credit has increased the demand for more accurate credit scoring models to manage credit risk and maximize profitability.

Some Important Terms:

- **Default:** Failure to repay a loan or credit card debt.
- **Attrition:** Customer pay off their balance and close their accounts.
- **Censoring:** When the event of interest has not occurred by the end of the observation period.
- **Customer Lifetime Value (CLV)**: The total revenue a customer generates over their lifetime.
- **Hazard Function:** The probability of an event occurring at a given time, given that it has not occurred before.

Research Gap:

- **Existing Models:**
  - **Accelerated Failure Time (AFT) model:** Need to specify the survival time distribution.(Parametric model)
  - **Cox proportional hazards model:** Difficult to handle poor approximation, computationally time-consuming, and lacks flexibility in modeling the hazard function (Cox, 1972).(Semi-parametric model)
  - **Discrete time survival model:** Lack of flexibility in modeling the hazard function (Belloti and Crook, 2013).
- The existing literature and idea lack the ability to:
  - Test hypotheses about the shape of the hazard function.
  - Model noliner, irregular hazard functions, and spiky patterns in the hazard function.
  - Construct a more understandable and interpretable model.

Research Objectives:

- Develop a semi-parametric survival model using nonparametric regression splines to model the hazard function.
- Compare the performance of this new model with the widely used Cox model in terms of parameter estimation and prediction accuracy.

Research Contribution:

- Introduce a novel semi-parametric survival model for credit scoring.
- Show that the proposed model offers better prediction accuracy than the Cox model, especially in attrition models with low event rates.

Main Findings:

- The proposed model provides similar parameter estimates as the Cox model but outperforms it in predicting attrition, particularly in datasets with low event rates.
- This study advances survival modeling in credit scoring and offers a more efficient alternative for practical applications.

Preliminary

Parametric Survival Model (AFT Model)

- **Assumptions**:
  - Survival time (T) follows a distribution based on covariates (X).
  - Acceleration factor: $S(t|X) = S_0(te^{\beta X})$ (Klein and Moeschberger, 2003).
- **Limitations**:
  - Cannot incorporate time-dependent covariates.
  - Requires specifying the distribution, which may not fit the actual data.

Semi-parametric Cox Proportional Hazard Model

- **Formula**: $h(t|X) = h_0(t)e^{\beta X}$ (Cox, 1972)
- **Key Features**:
  - Does not require specifying a survival distribution.
  - Can handle time-dependent covariates.
- **Limitations**:
  - Struggles with ties in the data.
  - Computationally demanding with many time-dependent covariates.
- **Extensions:** Time-dependent Cox model, Stratified Cox model.
  - **Time-dependent Cox model**: $h(t|X(t)) = h_0(t)e^{\beta X(t)}$ (Bellotti and Crook, 2009).
  - **Different types of event**: $h(t, m|X(t)) = h_0(t, m)e^{\beta_m X(t)}, m = 1, 2$ (Allison, 2010).

Comparison and Motivation for The Study

- The proposed **discrete time survival model** addresses the limitations of both AFT and Cox models (Belloti and Crook, 2013).
- Seek for flexible modeling of hazard functions and incorporate time-dependent covariates without computational complexity.

Spline Based Multinomial Logistic Regression Survival Model (SMLRS)

Topic Overview:

- The model analyzes loan portfolios with competing risks.
- Time-dependent covariates and event types.
- Focus on default, attrition, and open status events.

Key Components of the Data: $\{t_i, d_i, m_i, x_i(t_i)\}$ for $i = 1, \ldots, n$

- **T**: Survival time from start date to event date.
- **M**: Event type: default (1), attrition (2), open status (0).
- **Censoring D**: If no event occurs by observation time, the data is censored.
- **Covariates** $x(t_i)$**:** Time-dependent.

Hazard, Survival, Density, and Cumulative Distribution Functions

- **Hazard Rate** $h(t, m)$:
  - Probability of event occurring at time t, given no prior event.
  - Defined as:
    - $h(t, m) = \frac{f(t,m)}{1-F(t-1,m)}$.
- **Survival Function** $S(t, m) = F(\infty, m) - F(t, m)$:
  - Probability of surviving past time t for event type m.
- **Define** $f(t, m)$, $F(t, m)$ as the density, cumulative distribution functions, respectively. We have:

$$f(t,m) = F(t,m) - F(t-1,m)$$

$$F(t,m) = Pr\{T \leq t, M = m | x(t)\} = \sum_{j=1}^{t} f(j,m)$$

(Go back to slide 15)

The Multinomial Logistic Regression Survival (MLRS) Model

- **Objective**: Predict the probability of different events occurring at each time point.
- **Model Structure**:
  - Use logistic regression for discrete-time survival data.
  - Competing risks modeled using multinomial logit approach.
- **Logistic Regression Form**:
  - $\ln\left(\frac{h(t,m)}{1-h(t)}\right) = \Theta(t, x(t), \xi_m)$ for $m = 1, 2$.

Incorporating Time-Dependent Covariates

- **Time-dependent covariates** are modeled just like time-independent ones.
- **Modeling Flexibility**:
  - The SMLRS model adapts to nonlinear and irregular hazard shapes.
  - **Regression Splines** are used to capture complex time effects.
- **Hazard Calculation**:
  - $h(t, m|x(t)) = \frac{\exp(\Theta(t,x(t),\xi_m))}{1+\exp(\Theta(t,x(t),\xi_1))+\exp(\Theta(t,x(t),\xi_2))}.$
- **Parametric Predictor Function**:
  - Define $\xi_m = \{\alpha_m, \beta_m\}$ for each event type m.
  - $\Theta(t, x(t), \xi_m) = \alpha_m' x(t) + S(t, \beta_m), m = 1, 2.$

Regression Splines for Hazard Function

- **Spline Function**:
  - A piecewise polynomial with smooth joins at knot points.
  - Flexible for modeling various types of trends in hazard functions (e.g., smooth or spiky).
- **Spline Formulation**:
  - $S(t, \beta_m) = \beta_0 m + \sum_{j=1}^{\kappa} \beta_{j,m} S(t, t_j)$.
  - **Cubic Splines** provide better flexibility than quadratic splines.
- If we use cubic spline function, it can be written as [1]:

$$\tilde{S}(t, t_j) = I(t > t_j)(t - t_j)^3 - t^3 + 3t_j t^2 - 3t_j^2 t = \begin{cases} -t_j^3 + 3t_j t^2 - 3t_j^2 t, & t \leq t_j \\ -t_j^3, & t > t_j \end{cases}$$

where t_j represents the knots or the points where the spline segments meet;$I\{.\}$ is the indicator function.

1. In Appendix B, we demonstrate how the regression spline function is derived.

Maximum Likelihood Estimation (MLE)

- **MLE Setup**:
  - The likelihood function for the multinomial logistic survival model is based on the joint probability distribution of the event time and type.
  - The **log-likelihood** is maximized to estimate the parameters of the model.

$$
\begin{aligned}
\Pi_{i=1}^{n} L_i &= \Pi_{i=1}^{n} Pr\{T = t_i, M = m_i | \vec{x}_i(t_i)\} \\
&= \Pi_{i=1}^{n} Pr\{T = t_i, M = m_i | T > t_i, \vec{x}_i(t_i)\} Pr\{T > t_i | \vec{x}_i(t_i)\} \\
&= \Pi_{i=1}^{n} h(t_i, 1 | \vec{x}_i(t_i))^{I\{m_i=1\}} h(t_i, 2 | \vec{x}_i(t_i))^{I\{m_i=2\}} \\
&\quad \times \Pi_{t=1}^{t_i - 1} (1 - h(t | \vec{x}_i(t_i))) \\
&= \Pi_{i=1}^{n} \Pi_{t=1}^{t_i} h(t_i, 1 | \vec{x}_i(t_i))^{I\{z_{it}=1\}} h(t_i, 2 | \vec{x}_i(t_i))^{I\{z_{it}=2\}} \\
&\quad \times (1 - h(t | \vec{x}_i(t_i)))^{I\{z_{it}=0\}}
\end{aligned}
$$

where $z_{it} = m_i * I\{t = t_i\}$.

We then have:

$$\Pi_{i=1}^{n} L_i = \Pi_{i=1}^{n} \Pi_{t=1}^{t_i} \times \left\{ \frac{\exp(\Theta(t_i, \vec{x}_i(t_i), \vec{\alpha}_1, \vec{\beta}_1))}{1 + \exp(\Theta(t_i, \vec{x}_i(t_i), \vec{\alpha}_1, \vec{\beta}_1)) + \exp(\Theta(t, \vec{x}_i(t_i), \vec{\alpha}_2, \vec{\beta}_2))} \right\}^{I\{z_{it}=1\}} * \left\{ \frac{\exp(\Theta(t_i, \vec{x}_i(t_i), \vec{\alpha}_2, \vec{\beta}_2))}{1 + \exp(\Theta(t, \vec{x}_i(t_i), \vec{\alpha}_1, \vec{\beta}_1)) + exp(\Theta(t_i, \vec{x}_i(t_i), \vec{\alpha}_2, \vec{\beta}_2))} \right\}^{I\{z_{it}=2\}} * \left\{ \frac{1}{1 + \exp(\Theta(t, \vec{x}_i(t_i), \vec{\alpha}_1, \vec{\beta}_1)) + \exp(\Theta(t_i, \vec{x}_i(t_i), \vec{\alpha}_2, \vec{\beta}_2))} \right\}^{I\{z_{it}=0\}}$$

- **Notice:** Once we estimate the hazard functions $h(t, m) = \frac{f(t,m)}{1-F(t-1,m)}$, we can calculate the $f(t, m)$and $F(t, m)$ recursively. (Go to slide 10)

Practical Considerations and Model Flexibility

- **Knots Selection**:
  - Knots can be placed at quantiles of the event times (e.g., 5th, 50th, and 90th percentiles).
  - **Harrell's Recommendation**: Use 3-6 knots for cubic splines (Harrell, 2001).
  - The knots can be more explicit by selecting specified interior points based on the empirical graph (Hastie et al., 2001).

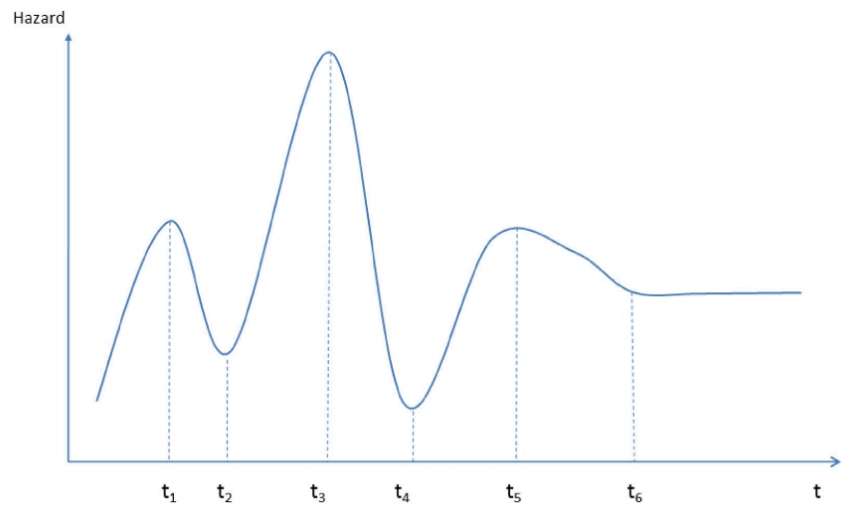


Fig. 1. Knots selection for regression spline.

Numerical Example - Comparison of SMLRS and Cox Models

Introduction to the Dataset

- **Source**: Real credit card performance data from a commercial bank.
- **Sample Size**: 30,000+ credit card accounts.
- **Timeframe**: Data spans over 20 months of monthly performance.
- **Covariates**: 8 time-independent covariates (e.g., APR, credit limit, delinquent status) and 1 time-dependent covariate (market interest rate).
- **Event Types**: Default (m=1) and attrition (m=2).
- **Data Split**: 70% for training, 30% for validation.

Table 2
The covariates in the dataset.

	Variable name	Description
1	Purchase_APR	Purchasing Annual Percent Rate (APR)
2	Cash_APR	Cash Annual Percent Rate (APR)
3	Credit_line	Credit limit
4	Pbad	Credit risk model score
5	Cash_balance	Cash balance
6	Dq_prior	The prior delinquent status
7	Outstanding	Total outstanding
8	Utilization	Card utilization
9	MarketRate	Market interest rate

Model Comparison

- **Objective**: Predict default and attrition risks.
- **Benchmark Model**: Cox Proportional Hazards (PHREG).
- **Proposed Model**: SMLRS (Survival Multi-Level Regression Spline).
- **Key Comparison**: Explanatory power & predictive performance.

Key Results - Default Model

- **Covariates Impact**: Both SMLRS and Cox models show similar sign and magnitude of parameter estimates.
  - **Purchase Rate & Cash Rate**: Positive impact on default risk.
  - **Credit Line**: Negative impact on default risk.
  - **Pbad Score**: Significant positive effect (default risk increases by 23.3% per unit increase).
  - **Market Rate**: Positive effect, greater impact on default risk than APR.
  - **Spline Terms**: Significant, especially spikes at t = 7, 9, 11.
- **C-Statistic**: SMLRS model slightly outperforms the Cox model.
  - 12 months: Improvement of 0.295%.
  - 18 months: Improvement of 0.332%.

Table 3
Parameter estimation for default model.

Parameter	SMLRS model		Cox model		
	Estimate	ProbChiSq	Estimate	ProbChiSq	HazardRatio
Intercept	−36.0940	<0.0001			
Purchase_APR	0.0153	0.0233	0.0174	0.0063	1.018
Cash_APR	0.3496	<0.0001	0.2992	<0.0001	1.349
Credit_line	−0.0719	<0.0001	−0.0681	<0.0001	0.934
Credit_line*Purchase_APR	−0.0024	<0.0001	−0.0028	<0.0001	0.997
Pbad	3.3023	<0.0001	3.1484	<0.0001	23.298
Cash_balance	0.0686	<0.0001	0.0341	0.0172	1.035
Dq_prior	0.5184	<0.0001	0.4453	<0.0001	1.561
Outstanding	0.1144	<0.0001	0.1115	<0.0001	1.118
Utilization	0.2941	<0.0001	0.2306	<0.0001	1.259
MarketRate	1.3410	<0.0001	1.2568	<0.0001	3.514
Marketrate*Cash_APR	−0.0559	<0.0001	−0.0477	<0.0001	0.953
CSB(t, 3)	0.1595	<0.0001			
CSB(t, 6)	−0.6847	<0.0001			
CSB(t, 7)	0.7147	<0.0001			
CSB(t, 9)	−0.2624	<0.0001			
CSB(t, 10)	0.0713	0.0032			
CSB(t, 17)	0.0016	<0.0001			
SPK(7)	−0.4237	<0.0001			
SPK(7)*Dq_prior	0.4220	<0.0001			
SPK(9)	−0.7633	<0.0001			
SPK(9)*Pbad	−1.7915	< 0.0001			
SPK(11)	0.1368	0.0386			
C statistics (12 months)	0.8726			0.8700	
C statistics (18 months)	0.8549			0.8521	

Table 5
Prediction comparison for Cox model and SMLRS model with different time windows.

AUC	Default model			Attrition model		
C Statistics	SMLRS	Cox model	Change (percent)	SMLRS	Cox model	Change (percent)
12 months	0.8729	0.8692	0.416	0.7350	0.7112	3.348
18 months	0.8538	0.8494	0.514	0.7125	0.6811	4.603

Key Results - Attrition Model

- **Covariates Impact**:
  - **Purchase Rate**: Positive effect (higher purchase rate increases attrition risk).
  - **Credit Line**: Negative effect (higher credit line reduces attrition risk).
  - **Utilization & Cash Balance**: Negative impact on attrition.
  - **Pbad Score**: Negative effect (higher Pbad reduces likelihood of attrition).
  - **Market Rate**: Negative effect (increasing market rate reduces attrition).
- **C-Statistic**: SMLRS outperforms Cox model by:
  - 12 months: 3.107% improvement.
  - 18 months: 4.620% improvement.
- **Challenges**: Attrition rate is lower than default, making the attrition model harder to build.

Table 4
Parameter estimation for attrition model.

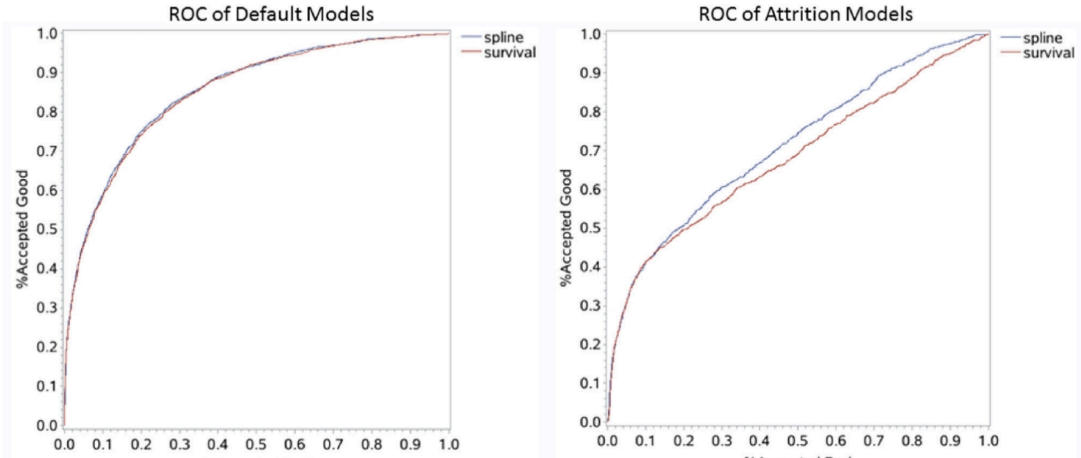
Parameter	SMLRS model		Cox model		
	Estimate	ProbChiSq	Estimate	ProbChiSq	HazardRatio
Intercept	−15.6183	<0.0001			
Purchase_APR	0.0175	0.0015	0.0183	0.0008	1.018
Credit_line	−0.0513	<0.0001	−0.0518	<0.0001	0.950
Pbad	−1.3274	<0.0001	−1.3063	<0.0001	0.271
Cash_balance	−0.1212	0.0530	−0.1173	0.0595	0.889
Dq_prior	0.5078	<0.0001	0.5025	<0.0001	1.653
Utilization	−1.4459	<0.0001	−1.4533	<0.0001	0.234
Marketrate	−0.3052	0.0515	−1.0936	0.0004	0.335
CSB(t, 4)	0.1218	<0.0001			
CSB(t, 6)	−0.1591	<0.0001			
CSB(t, 9)	0.0491	<0.0001			
CSB(t, 11)	−0.0119	<0.0001			
SPK(4)	0.2444	0.0022			
SPK(9)	−0.3834	0.0032			
SPK(19)	−0.2770	0.0316			
C statistics (12 months)	0.7502		0.7276		
C statistics (18 months)	0.7160		0.6844		

Table 5
Prediction comparison for Cox model and SMLRS model with different time windows.

AUC	Default model			Attrition model		
C Statistics	SMLRS	Cox model	Change (percent)	SMLRS	Cox model	Change (percent)
12 months	0.8729	0.8692	0.416	0.7350	0.7112	3.348
18 months	0.8538	0.8494	0.514	0.7125	0.6811	4.603

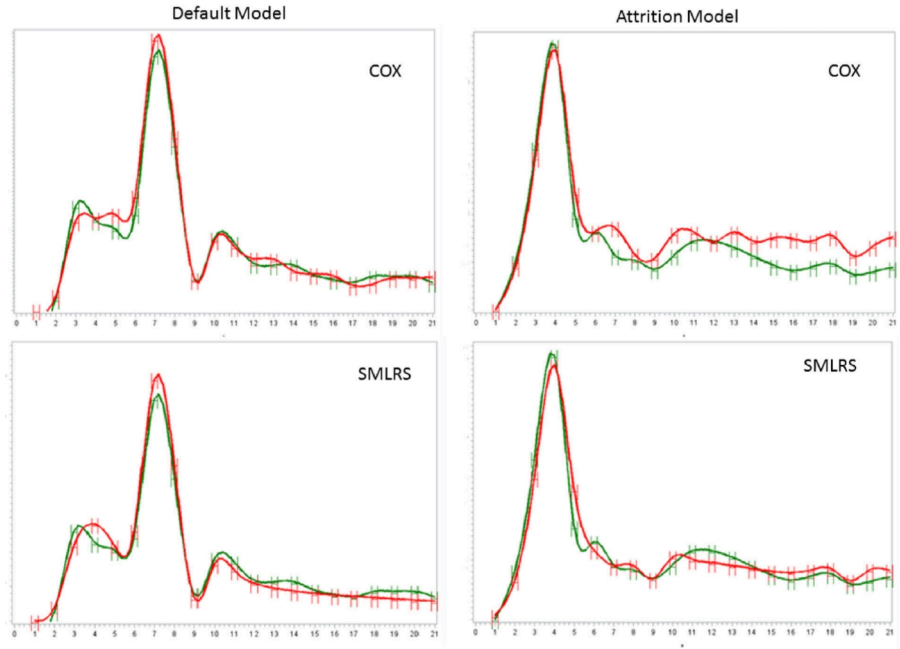
ROC Curve Comparison

- **ROC Curve Analysis**: The Receiver Operating Characteristic (ROC) curve was used to assess the classification performance of both models.
- **Results**:
  - The area under the curve (AUC) for SMLRS is consistently higher than that of the Cox model, indicating better classification performance.
  - SMLRS shows better discrimination power, particularly in the 12-month and 18-month prediction windows.



Prediction vs. Actual Hazard Functions

- **SMLRS vs. Cox Model**: Both models predict hazard functions for default and attrition events. Below is a visual comparison of the predicted vs. actual hazard functions.
- **Findings**: SMLRS provides a closer match to the actual hazard function, especially for time-dependent changes, compared to the Cox model.



Cost-based Performance Comparison

- **Objective**: Assess the financial performance based on a cost-benefit analysis of model predictions.
- **Method**: We calculate the cost of misclassifications (false positives and false negatives) and compare the net benefit of SMLRS and Cox models.
- **Findings**:
  - SMLRS outperforms the Cox model by achieving a better cost-to-benefit ratio.
  - The cost of misclassifying default and attrition events is lower in the SMLRS model due to its higher predictive accuracy.

Table 6
Cost based performance comparison for default model.

Time window	Cost ratio	Cut-off from training data			Cut-off from validation data		
	C_{FP}/C_{FN}	Cox	SMLRS	Change (percent)	Cox	SMLRS	Change (percent)
12 months	15	0.5880	0.5763	2.032	0.5876	0.5694	3.200
	20	0.6385	0.6132	4.139	0.6289	0.6042	4.093
	25	0.6637	0.6596	0.625	0.6480	0.6378	1.598
18 months	15	0.6278	0.6202	1.224	0.6179	0.6044	2.243
	20	0.6606	0.6516	1.381	0.6460	0.6266	3.098
	25	0.6646	0.6634	0.180	0.6594	0.6485	1.672

Summary & Conclusion

- **SMLRS vs Cox Model**: SMLRS shows better predictive performance in both default and attrition models, particularly with higher C-statistics in the attrition model.
- **Findings**:
  - **Product-related variables** affect both default and attrition.
  - **Customer behavior** has significant impact on risk predictions.
  - **Spline and spike terms** improve model fitting, especially for default risks.

Thank You !