Introduction to Survival Data Mining

Course Notes

Introduction to Survival Data Mining Course Notes was developed by Mike Patetta. Instructional design, editing, and production support was provided by the Learning Design and Development team.

Introduction to Survival Data Mining Course Notes

Book code E71405, course code ATESDMI, prepared date 03Dec2018. ATESDMI_001

To learn more…

For information about other courses in the curriculum, contact the SAS Education Division at 1-800-333-7660, or send e-mail to training@sas.com. You can also find this information on the web at http://support.sas.com/training/ as well as in the Training Course Catalog.

For a list of SAS books (including e-books) that relate to the topics covered in this course notes, visit https://www.sas.com/sas/books.html or call 1-800-727-0025. US customers receive free shipping to US addresses.

Lesson 1 Introduction to Survival Data Mining

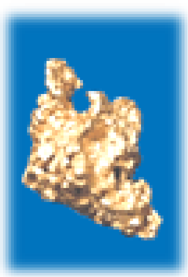
1.1 Introduction to Survival Data Mining

What Is Survival Analysis?

- *Survival analysis* is a class of statistical methods for which the outcome variable of interest is time until an event occurs.
- Time is measured from when an individual or organization first becomes a customer until the event occurs or until the end of the observation interval.
- In survival analysis, the basis of the analysis is tenure, or time at risk for the event, and not calendar time.

2

What Is Data Mining?



"Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner."

– David Hand

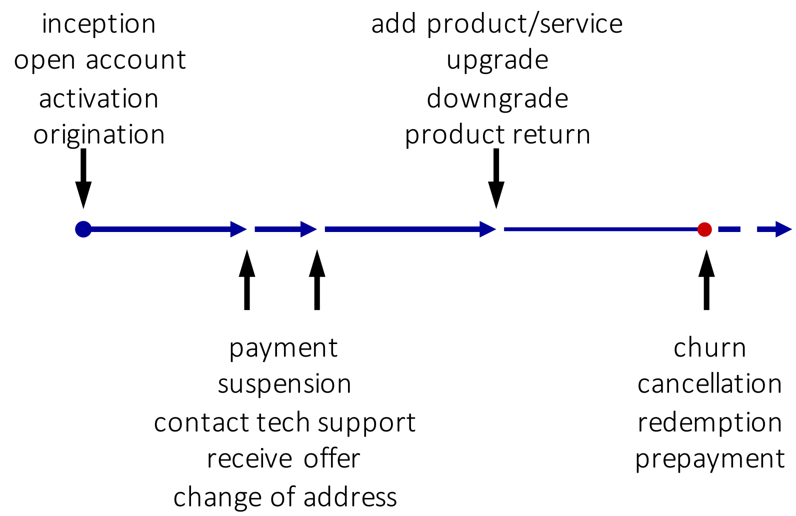
3

Survival Data Mining

Survival Analysis	+	*Data Mining*
Statistical methods for censored, time to event data		Knowledge discovery in opportunistic databases

4

Customer History Data



5

Time-Dependent Customer Outcomes

Customer retention applications

- cancellation of all products and services
- severe downgrade or extreme inactivity
- unprofitable behavior

Add-on selling applications

- acquisition of the target product or service
- more profitable behavior

Credit risk management applications

- charge-off
- loan termination

6

Extracting and Preparing the Relevant Data

Customer Data

- identifier: account number, person, household
- billing details: address, amount, payment method
- application information: demographics, credit history

Product Data

- features: description, category, rate plan
- status: start/stop date, cancellation reason, changes, suspensions, disconnections
- usage: activity, amount, duration, balance, payment

Contact Data

- marketing promotions: direct mail, call center
- customer service: technical support, billing inquiries

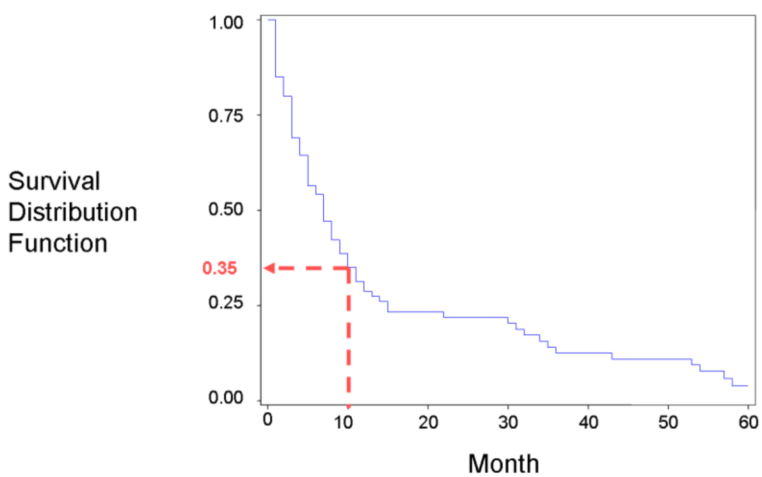
7

Data Structure

Customer	Tenure	Status
A	4.0	1 (event)
B	6.0	0 (censored)
C	3.0	0
D	5.0	1
E	3.0	0
F	3.0	1
G	2.0	1

8

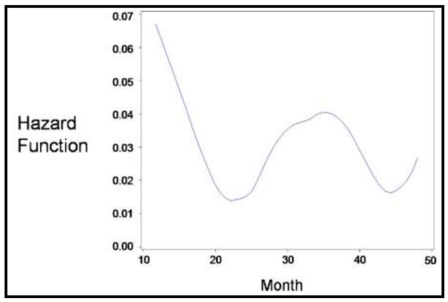
Survival Function for Continuous Time



9

Hazard Function for Continuous Time

- The hazard function is the instantaneous risk or potential that an event will occur at tenure *t*, given that the individual has survived up to tenure *t*.
- It takes the form of the expected number of events per interval of tenure.
- For tenure measured on a continuous scale, it is a rate, not a probability, that ranges from zero to infinity.



10

Hazard Function for Discrete Time

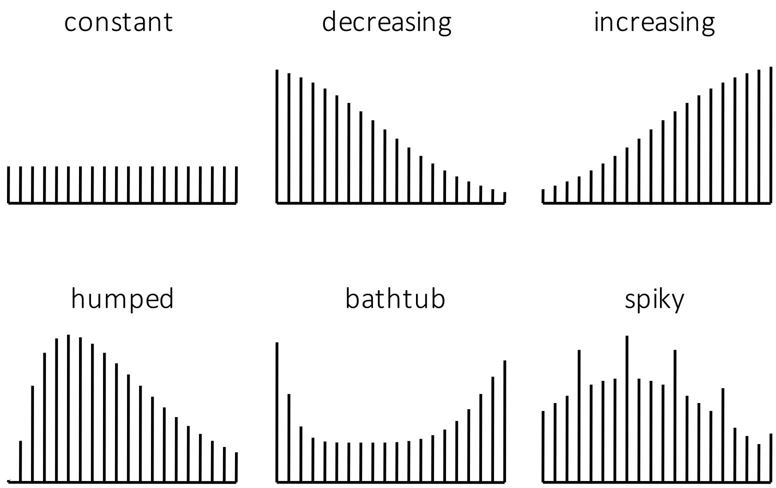
$$h(t) = \Pr(T = t \mid T \geq t)$$

= probability of having the event at tenure *t* given no prior occurrence of the event

$$= 1 - \left(\frac{S(t)}{S(t-1)}\right)$$

12

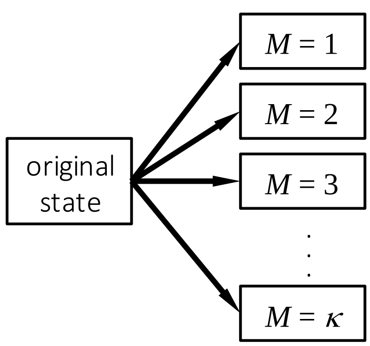
Hazard Shapes



13

Competing Risks

$$M \in \{1, 2, \ldots \kappa\}$$



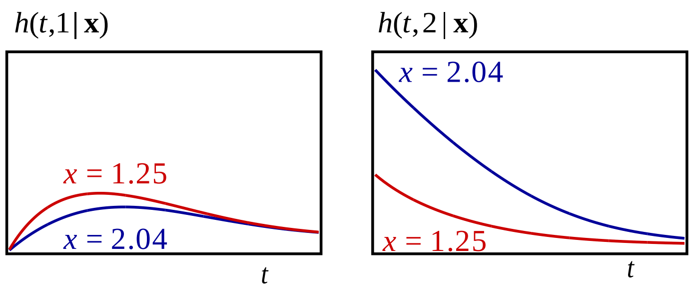
- multiple cancellation reasons (causes of death)
- voluntary or involuntary churn
- loan prepayment or default
- next product acquired
- outcome severity

14

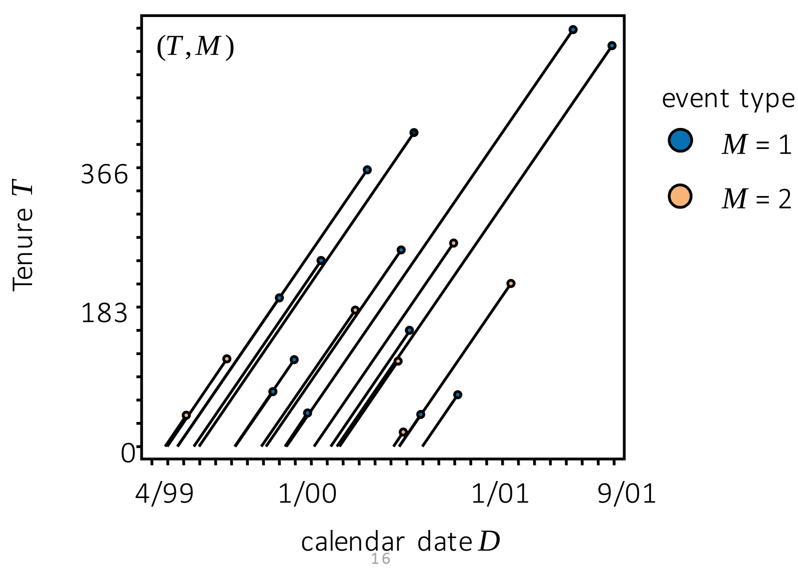
Sub-Hazard Function

$$h(t,m\,|\,\mathbf{x}) = \Pr(T = t, M = m\,|\,T \geq t, \mathbf{x})$$

= the conditional probability that an event of type *m* occurs at tenure *t*, given that an event of any type has not yet occurred and given the values of the covariates (also known as cause-specific hazard)



Event Time and Type



The Event-Time Distribution

Duration between start and event for continuous time

$$T = D^{(\text{event})} - D^{(\text{start})}$$

Discrete random variable

- smallest meaningful unit
- days, months, billing cycles
- many ties

$$T = D^{(\text{event})} - D^{(\text{start})} = 0, 1, 2, \ldots$$

Covariates

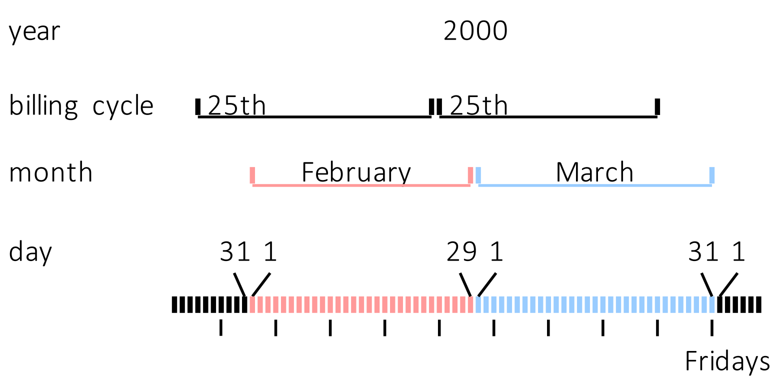
- time-independent covariates
- time-dependent covariates

$$\mathbf{x} = (x_1, \ldots, x_p)$$

$$\mathbf{x}(t) = (x_1(t), \ldots, x_p(t))$$

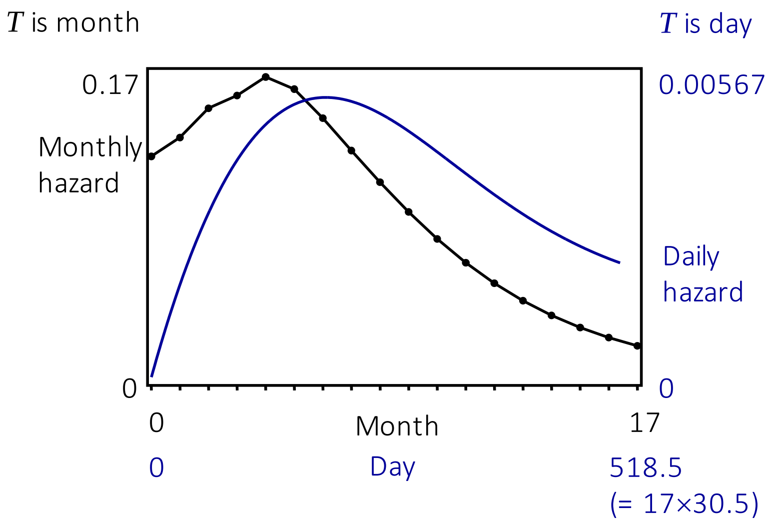
17

Grouped Dates



18

Grouping Dates Redefines the Hazards



19

Censoring and Truncation

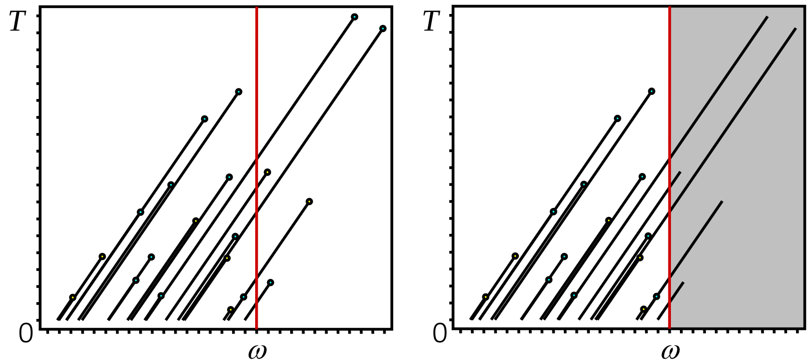
complete data $\{(t_i, m_i, \mathbf{x}_i)\}_{i=1}^{n}$

- The observed data are not simply realizations of the random variables (T, M).
- Survival data are incompletely observed.
- An observation is right-censored if the observation is terminated before the event occurs.
- An observation is left-truncated if the observation had the event before a certain time and the observation was omitted from the sample.
- Censoring is a property of the observation while truncation is a property of the sample.

20

Right (End-of-Study) Censoring

complete data $\{(t_i, m_i, \mathbf{x}_i)\}_{i=1}^{n}$ → incompletely observed data $\{(y_i, v_i, \mathbf{x}_i)\}_{i=1}^{n}$



ω = date that the extracted data was current
= censoring date, examination date, end of study

Observed Event Time and Type

complete data $\{(t_i, m_i, \mathbf{x}_i)\}_{i=1}^{n}$ → incompletely observed data $\{(y_i, v_i, \mathbf{x}_i)\}_{i=1}^{n}$

The incomplete data are realizations of the random variables Y and V. For those who have not had the event yet, the censored time Y is set equal to the observation limit C and the censored event type V is set equal to 0.

$$Y = \begin{cases} T & \text{if } T \le C \\ C & \text{if } T > C \end{cases} \qquad C = \omega - D^{(\text{start})}$$

$$V = \begin{cases} M & \text{if } T \le C \\ 0 & \text{if } T > C \end{cases}$$

22

Independent Censoring

How can the joint distribution of (T, M) be estimated from the incomplete data?

- The solution depends on the assumption of independent censoring: conditional on the covariates.
- (T, M) and C (hence, D(start)) are independent.

Loss-to-Follow-up Censoring

Database errors

- When event dates are missing or incomplete, the customers could be censored at an earlier trustworthy date.

Nuisance competing risks

- moving out of the coverage area of a service
- selling a loan before it terminates
- churn before upgrade

Avoiding Dependent Censoring

End-of-study censoring

- The sources of variation related to the inception date need to be identified and incorporated into the analysis.

Loss-to-follow-up censoring

- Treat it as a competing risk. Add a level to the event type V.
- Treat as censoring (V = 0).

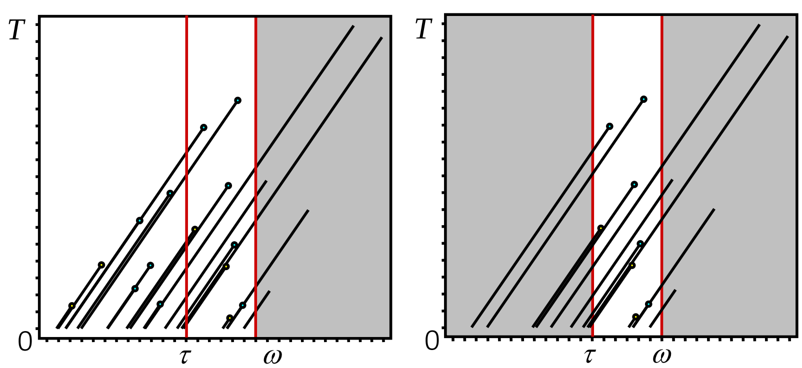
25

Left Truncation

right censored data

$\{(y_i, v_i, \mathbf{x}_i)\}_{i=1}^{n}$ → truncated

$\{(y_i, v_i, \mathbf{x}_i) : d_i^{(\text{event})} \geq \tau\}$



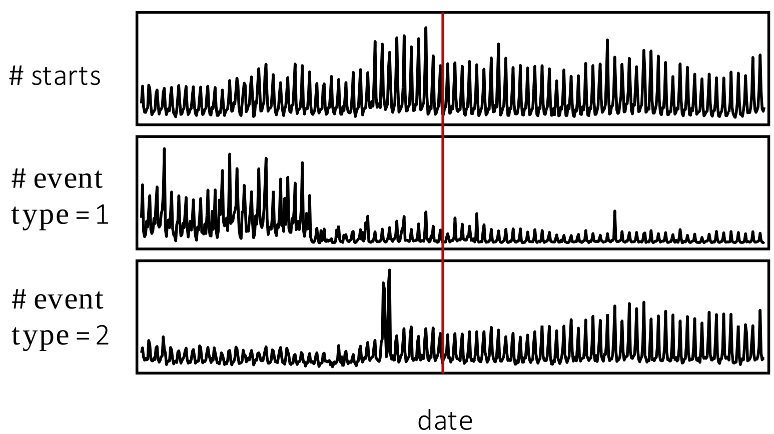
τ = truncation date

26

Long-Life Bias

- The truncated data is biased in favor of longer-lived customers.
- Among customers who originated at some date, the longer-lived would be the only ones remaining in a database of current or recent customers.
- To avoid the long-life bias, the history of the cases prior to the truncation date is excluded from the analysis.

Concentrate on Recent Events



Time Dependent Covariates

In addition to the censored event time and type, the observed data for each customer includes the values of the covariate vector. If the values of the covariates change over time, then the data for each customer consists of many individual time series.

$$\{(y_i, v_i, \mathbf{x}_i(\mathbf{y}_i))\}_{i=1}^{n}$$

$$\mathbf{x}_i(\mathbf{y}_i) = \begin{pmatrix} \mathbf{x}_i(0) \\ . \\ \mathbf{x}_i(y_i) \end{pmatrix} = \begin{pmatrix} x_{1i}(0) & \cdots & x_{pi}(0) \\ . & . & . \\ x_{1i}(y_i) & \cdots & x_{pi}(y_i) \end{pmatrix}$$

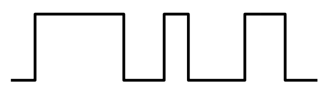
29

Varieties of Time Dependent Covariates

$$x(\mathbf{y}_i) = (x(0), \ldots, x(y_i))'$$

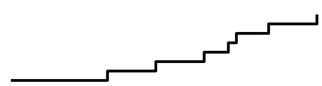
Binary indicators of (reversible) state transitions

- pay off a loan
- have an investment account



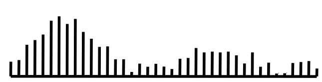
Steps or points representing (cumulative) counts of recurrent events

- report problem to tech support
- delinquent payment



Continuously varying quantities

- minutes of phone usage
- account balance



30

Discrete-Time Survival Models

The logit model is $\log\left[\frac{P_{it}}{1-P_{it}}\right] = \alpha_t + \beta_1 x_{it1} + \beta_2 x_{it2} + \ldots + \beta_k x_{itk}$

where P_{it} is the conditional probability that individual *i* has an event at time *t* given that an event has not already occurred to that individual. The parameter α_t is some function of time.

Discrete-Time Survival Models

- Data must be expanded so that each individual's survival history is broken down into a set of discrete time units that are treated as distinct observations.
- In each time interval there is a response indicating whether an event has occurred.
- The logit model provides estimates of conditional probabilities of each event occurring in each time unit.
- The covariates are allowed to vary over time from one time unit to another.
- For competing risks, the multinomial logit model can be used.

LOGISTIC Procedure

```
PROC LOGISTIC <options>;
      CLASS variables </options>;
      MODEL response = <effects></options>;
      FREQ variable;
      CODE <options>;
RUN;
```

33

Discrete-Time Survival Models

Strengths:

- The models can be easily fit in PROC LOGISTIC.
- The models are well suited to the challenging features of survival data mining problems such as
  - competing risks
  - truncated data
  - time-dependent covariates and time-varying effects of the covariates
  - irregular, nonlinear hazards.

Weaknesses:

- If the observation period is long relative to the width of the time intervals, the data set might become very large.

34

Logistic Regression Models with Competing Risks

$$\ln\left(\frac{h(t,m\,|\,\mathbf{x}(t))}{1-h(t\,|\,\mathbf{x}(t))}\right)=\eta(t,\mathbf{x}(t),\boldsymbol{\theta}_m) \qquad m=1,\ldots,\kappa$$

The generalized logit link function is the log of the odds of an event of type *m*.

Each competing risk has a separate model.

The parametric predictor function represents the effect of time and the covariates. The function has the same form but a different parameter vector for each competing risk.

35

Generalized Logit Link Function

Probability of Event Type *m* at Time *t*

$$\ln\left(\frac{\Pr(g_{it}=m\,|\,\mathbf{x}_i(t))}{\Pr(g_{it}=0\,|\,\mathbf{x}_i(t))}\right)=\ln\left(\frac{h(t,m\,|\,\mathbf{x}_i(t))}{1-h(t\,|\,\mathbf{x}_i(t))}\right)=\eta(t,\mathbf{x}_i(t),\boldsymbol{\theta}_m)$$

Probability of No Event prior to Time *t*

36

Expanded Data

$$\left\{(y_i, v_i, \mathbf{x}_i(\mathbf{y}_i))\right\}_{i=1}^{n} \rightarrow \left\{(t, g_{it}, \mathbf{x}_i(t)) : t = 0, \ldots, y_i \ \& \ i = 1, \ldots, n\right\}$$

Categorical target $g_{it} = v_i \cdot I\{t = y_i\}$

		t	g	$\mathbf{x}$
	i	0	0	$\mathbf{x}_i(0)$
	i	1	0	$\mathbf{x}_i(1)$
i^{th} customer	i	2	0	$\mathbf{x}_i(2)$
	.	.	.	.
	i	$y_i - 1$	0	$\mathbf{x}_i(y_i - 1)$
	i	y_i	v_i	$\mathbf{x}_i(y_i)$

37

Multinomial Likelihood for Censored Data

expanded data $\sum_{i=1}^{n}(y_i + 1)$ rows

categorical target $g_{it} = v_i \cdot I\{t = y_i\}$

$I\{g_{it} \neq 0\} = \sum_{m=1}^{\kappa} I\{g_{it} = m\}$

$$\prod_{i=1}^{n}\prod_{t=0}^{y_i} h(t,1 \mid \mathbf{x}_i(t))^{I\{g_{it}=1\}} \cdots h(t,\kappa \mid \mathbf{x}_i(t))^{I\{g_{it}=\kappa\}} (1 - h(t \mid \mathbf{x}_i(t)))^{1-I\{g_{it}\neq 0\}}$$

sub-hazard

overall hazard = $\sum_{m=1}^{\kappa} h(t,m \mid \mathbf{x}_i(t))$

38

Parametric Predictor Function

$$\eta\left(t, \mathbf{x}(t), \begin{pmatrix} \boldsymbol{\alpha}_m \\ \boldsymbol{\beta}_m \end{pmatrix}\right) = \alpha_{0m} + \psi(t, \boldsymbol{\alpha}_m) + \beta_{1m} x_1(t) + \ldots + \beta_{pm} x_p(t)$$

function of time

$$\psi(t, \boldsymbol{\alpha}_m)$$

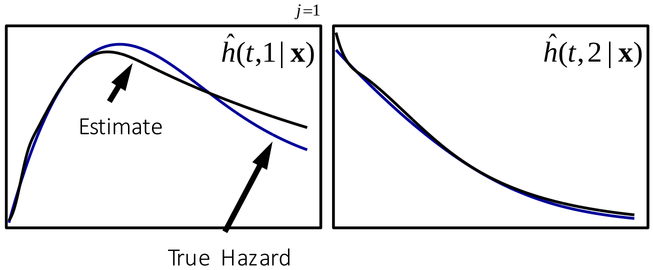
smooth trend
- polynomial
- regression spline
- neural network

spiky trend
- periodicities
- time zero effect
- discontinuities

Regression Spline Hazards

function of time

$$\psi(t, \boldsymbol{\alpha}) = \alpha_{00} + \alpha_0 t + \sum_{j=1}^{\#\{\text{knots}\}} \alpha_j \operatorname{csb}(t, k_j)$$

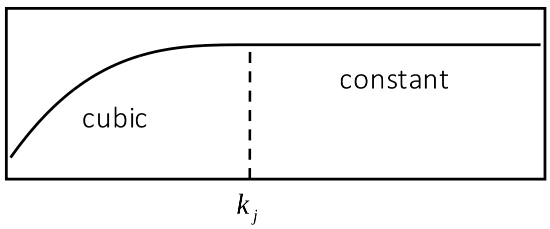


Cubic Spline Basis Functions

$$\psi(t,\boldsymbol{\alpha}) = \alpha_{00} + \alpha_0 t + \sum_{j=1}^{\#\{\text{knots}\}} \alpha_j \,\text{csb}(t,k_j)$$

$$\text{csb}(t,k_j) = I\{t > k_j\}(t-k_j)^3 - t^3 + 3k_j t^2 - 3k_j^2 t$$

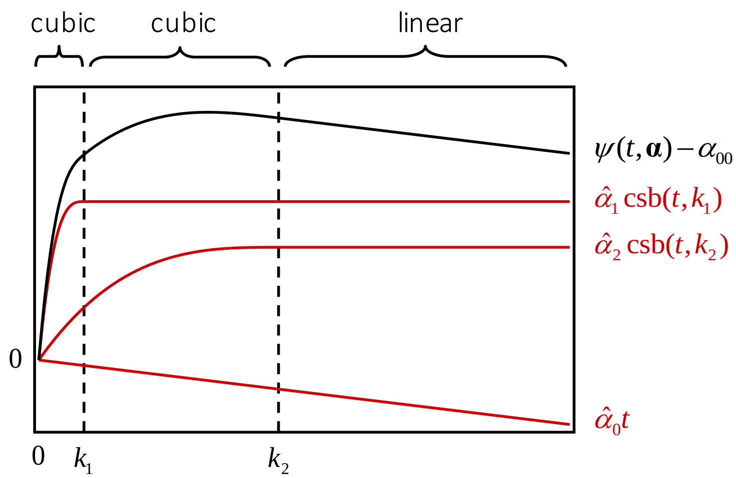
$$= \begin{cases} -t^3 + 3k_j t^2 - 3k_j^2 t & \text{if } t \le k_j \quad \text{cubic} \\ -k_j^3 & \text{if } t > k_j \quad \text{constant} \end{cases}$$



continuous: $\text{csb}(t,k_j)$, $\text{csb}'(t,k_j)$, $\text{csb}''(t,k_j)$

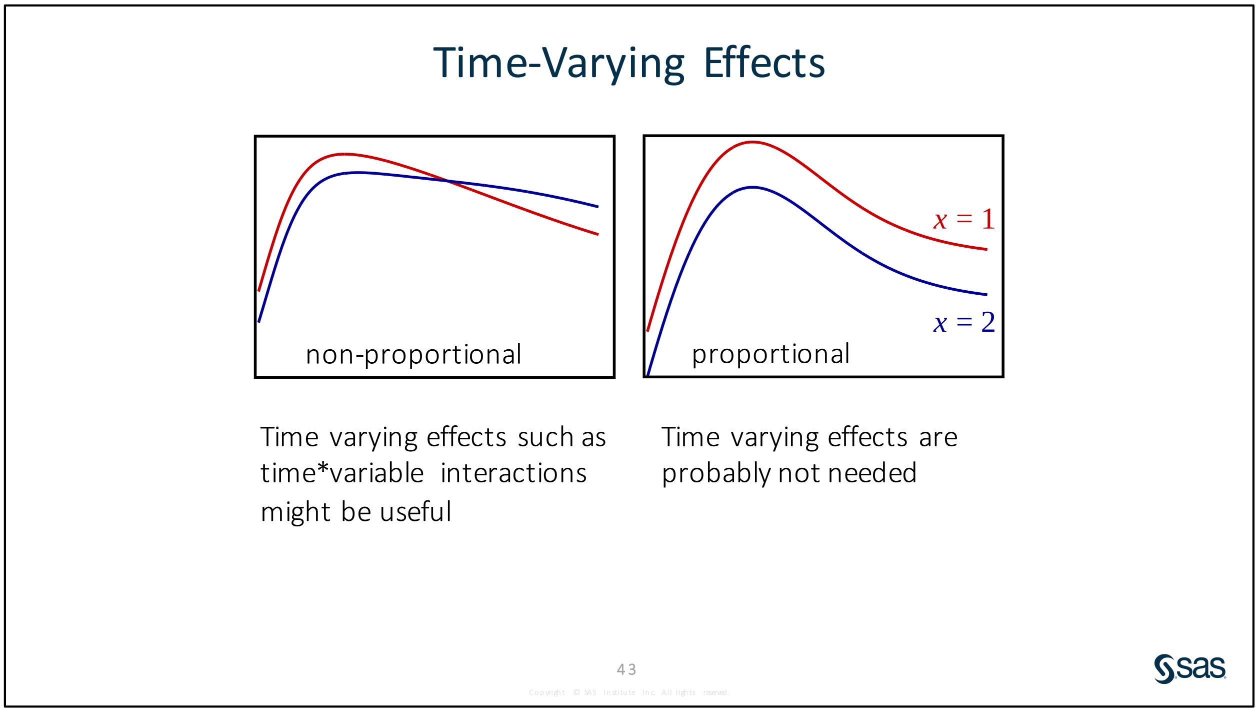
41

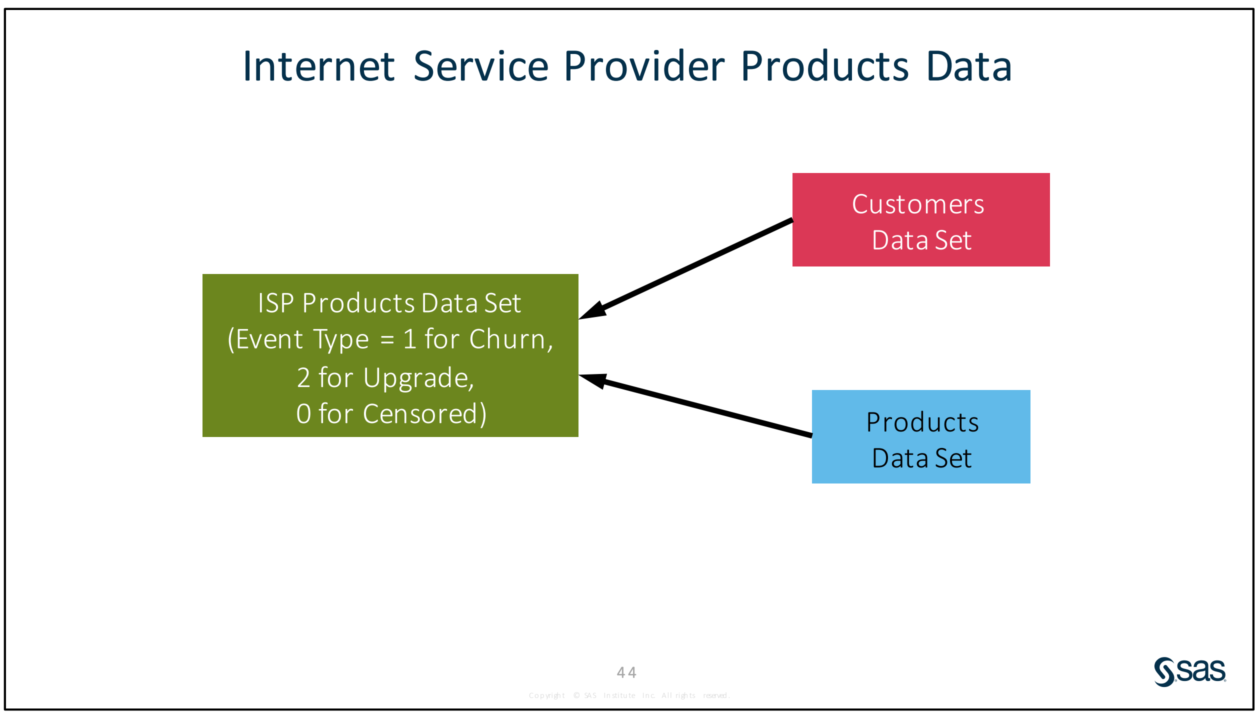
Cubic Spline with Linear End Piece



continuous: $\psi(k_j,\boldsymbol{\alpha})$, $\psi'(k_j,\boldsymbol{\alpha})$, $\psi''(k_j,\boldsymbol{\alpha})$

42





Example: An Internet service provider wants to predict when their customers upgrade their products. However, if a customer churned, then they are no longer at risk for upgrade. Consequently, churn prior to upgrade is considered a (nuisance) competing risk. The data has been expanded and the time dependent variables were created.

These are variables in the data set:

account_id	account identifier

office	geographic region coded as A-O
credit_card_payment	indicator of payment by credit card
telecom	indicator of whether business is in telecommunications field
financial	indicator of whether business is in financial field
computer	indicator of whether business is in computer field
health	indicator of whether business is in health field
legal	indicator of whether business is in legal field
table	indicates which database table the record originated from. The data was created by joining three database tables (initial products (INIT), disconnections (DISC), and product additions (PROD)).
event_date	is month and year of the event.
product_category	is the main product category. (DS-1 and DS-3 are upgrades.)
bandwidth	is the sub-category representing bandwidth.
quantity	is the number of products.
initial_date	date of initial product acquisition
upgrade_date	date high end access products were added
churn_date	date all current products were disconnected
event_time	number of months between the event or censoring date and inception
time	the time point the customer was observed at. The time points range from 0 to the event time
event_type	the event indicator which equals 1 for customers who churned before upgrading, 2 for customers who upgraded and 0 for censoring
number_dial	number of dial-up products that are present in the previous month
number_isdn	number of isdn products that are present in the previous month
number_dsl	number of dsl products that are present in the previous month
number_fds1	number of fds1 products that are present in the previous month
number_ds13	number of ds13 products that are present in the previous month

Note: The variables telecom, financial, computer, health, and legal are indicators of particular Standard Industrial Classification (SIC) codes.

Fitting Regression Spline Hazard Models

Example: Fit a regression spline hazard model with time-independent covariates, time-dependent covariates, and interactions involving time to accommodate the possible time-varying effects of these covariates. Write a text file of DATA step scoring code and finally display the fitted hazard functions.

```
%let knots=2 4 8;

data bmce.ExpandedISP_spline;
   set bmce.ExpandedISP;
   array k{3} _temporary_ (&knots);
   cubic_spline_b1=(time>k[1])*(time-k[1])**3-
       time**3+3*k[1]*time**2-3*k[1]**2*time;
   cubic_spline_b2=(time>k[2])*(time-k[2])**3-
       time**3+3*k[2]*time**2-3*k[2]**2*time;
   cubic_spline_b3=(time>k[3])*(time-k[3])**3-
       time**3+3*k[3]*time**2-3*k[3]**2*time;
run;
```

For simplicity, three knots, placed at the quartiles of the event time distribution, are assumed to be adequate for this data. A better fit could probably be found by selecting the knots from a larger set of candidate positions. The cubic spline basis functions are added to the expanded data in the DATA step. The knots are specified as a macro variable and read into a temporary array.

```
proc logistic data=bmce.ExpandedISP_spline;
   class office / param=ref;
   model event_category(ref='0')=office credit_card_payment telecom
         financial computer health legal number_dial number_isdn
         number_dsl number_fds1 number_ds13 time cubic_spline_b1-
         cubic_spline_b3 time*credit_card_payment time*number_dsl
         time*number_fds1 time*number_ds13 / link=glogit;
   code file="s:\workshop\model1.txt";
   title "Regression Spline Hazard Model with Time-Dependent "
         "Covariates";
run;
```

The logistic model includes the six time-independent covariates from the customer data and the five time-dependent product counts. In addition, the model includes four selected interactions involving time to accommodate the possibly time-varying effects of these covariates.

The CODE statement writes SAS DATA step code for computing predicted values of the fitted model either to a file or to a catalog entry. This code can then be included in a DATA step to score new data.

The covariate **office** is categorical and listed in the CLASS statement. The PARAM=REF option uses the set-to-zero parameterization for the dummy variables.

The key command for running multinomial logistic regression is the LINK=GLOGIT option in the MODEL statement. The categorical event indicator **event_category** is the target. The REF='0' option makes the censored class the reference level. Consequently, the generalized logits have the sub-hazards in the numerator and 1 minus the overall hazard in the denominator.

Regression Spline Hazard Model with Time-Dependent Covariates

The LOGISTIC Procedure

Model Information

Data Set	BMCE.EXPANDEDISP_SPLINE
Response Variable	event_category
Number of Response Levels	3
Model	generalized logit
Optimization Technique	Newton-Raphson

Number of Observations Read	85065
Number of Observations Used	85065

Response Profile

Ordered Value	event_ category	Total Frequency
1	0	82188
2	1	2655
3	2	222

Logits modeled use event_category=0 as the reference category.

Class Level Information

Class	Value	Design Variables													
office	A	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	B	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	C	0	0	1	0	0	0	0	0	0	0	0	0	0	0
	D	0	0	0	1	0	0	0	0	0	0	0	0	0	0
	E	0	0	0	0	1	0	0	0	0	0	0	0	0	0
	F	0	0	0	0	0	1	0	0	0	0	0	0	0	0
	G	0	0	0	0	0	0	1	0	0	0	0	0	0	0
	H	0	0	0	0	0	0	0	1	0	0	0	0	0	0
	I	0	0	0	0	0	0	0	0	1	0	0	0	0	0
	J	0	0	0	0	0	0	0	0	0	1	0	0	0	0
	K	0	0	0	0	0	0	0	0	0	0	1	0	0	0
	L	0	0	0	0	0	0	0	0	0	0	0	1	0	0
	M	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	N	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	O	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

	Intercept	Intercept and

Criterion	Only	Covariates
AIC	26710.328	25920.041
SC	26729.030	26555.921
-2 Log L	26706.328	25784.041

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	922.2865	66	<.0001
Score	1061.9440	66	<.0001
Wald	845.6360	66	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
office	28	80.2161	<.0001
credit_card_payment	2	20.9827	<.0001
telecom	2	54.7891	<.0001
financial	2	8.2339	0.0163
computer	2	13.7895	0.0010
health	2	13.7289	0.0010
legal	2	12.4037	0.0020
number_dial	2	0.9810	0.6123
number_isdn	2	1.0644	0.5873
number_dsl	2	154.9427	<.0001
number_fds1	2	37.8211	<.0001
number_ds13	2	69.8914	<.0001
time	2	0.2801	0.8693
cubic_spline_b1	2	0.6228	0.7324
cubic_spline_b2	2	0.4209	0.8102
cubic_spline_b3	2	0.1536	0.9261
credit_card_pay*time	2	36.3139	<.0001
number_dsl*time	2	12.0837	0.0024
number_fds1*time	2	8.4066	0.0149
number_ds13*time	2	10.7560	0.0046

Analysis of Maximum Likelihood Estimates

Parameter		event_ category	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1	-3.4636	0.1262	753.2292	<.0001
Intercept		2	1	-4.6491	0.3245	205.2844	<.0001
office	A	1	1	0.1795	0.1111	2.6113	0.1061
office	A	2	1	-0.4364	0.3611	1.4604	0.2269
office	B	1	1	0.1688	0.1091	2.3914	0.1220
office	B	2	1	-0.5385	0.3528	2.3305	0.1269
office	C	1	1	0.2086	0.1128	3.4188	0.0645
office	C	2	1	-0.1563	0.3445	0.2060	0.6500
office	D	1	1	0.3038	0.1215	6.2522	0.0124
office	D	2	1	0.5027	0.3273	2.3595	0.1245
office	E	1	1	0.2313	0.1306	3.1381	0.0765
office	E	2	1	-1.9504	1.0410	3.5107	0.0610

office	F	1	1	0.1735	0.1181	2.1580	0.1418
office	F	2	1	0.00935	0.3533	0.0007	0.9789
office	G	1	1	0.4353	0.1169	13.8674	0.0002
office	G	2	1	-1.2100	0.5573	4.7144	0.0299
office	H	1	1	0.1409	0.1179	1.4274	0.2322
office	H	2	1	0.1036	0.3503	0.0874	0.7675
office	I	1	1	0.1779	0.1366	1.6953	0.1929
office	I	2	1	-0.4105	0.4393	0.8728	0.3502
office	J	1	1	0.2289	0.1390	2.7119	0.0996
office	J	2	1	-0.4657	0.4981	0.8740	0.3498
office	K	1	1	-0.2071	0.1486	1.9434	0.1633
office	K	2	1	-0.5000	0.4715	1.1247	0.2889
office	L	1	1	0.3647	0.1477	6.0964	0.0135
office	L	2	1	-1.5322	1.0416	2.1638	0.1413
office	M	1	1	-0.00406	0.1678	0.0006	0.9807
office	M	2	1	-0.4592	0.5289	0.7538	0.3853
office	N	1	1	0.0765	0.1359	0.3171	0.5733
office	N	2	1	0.4371	0.3463	1.5931	0.2069
credit_card_payment		1	1	0.1909	0.0765	6.2239	0.0126
credit_card_payment		2	1	-1.8591	0.4854	14.6709	0.0001
telecom		1	1	0.000615	0.1124	0.0000	0.9956
telecom		2	1	1.3210	0.1785	54.7784	<.0001
financial		1	1	-0.3524	0.1281	7.5642	0.0060
financial		2	1	0.2528	0.3169	0.6367	0.4249
computer		1	1	-0.2407	0.0874	7.5869	0.0059
computer		2	1	0.4856	0.1971	6.0714	0.0137
health		1	1	-0.5445	0.1496	13.2408	0.0003
health		2	1	0.2456	0.3649	0.4530	0.5009
legal		1	1	-0.4069	0.1155	12.4029	0.0004
legal		2	1	-0.0171	0.3268	0.0028	0.9582
number_dial		1	1	-0.00849	0.0159	0.2853	0.5933
number_dial		2	1	0.0281	0.0340	0.6836	0.4083
number_isdn		1	1	-0.0482	0.0478	1.0205	0.3124
number_isdn		2	1	-0.0414	0.1909	0.0471	0.8282
number_dsl		1	1	-0.9443	0.0802	138.5310	<.0001
number_dsl		2	1	-0.9852	0.2394	16.9434	<.0001
number_fds1		1	1	-0.7320	0.1193	37.6342	<.0001
number_fds1		2	1	0.0853	0.2262	0.1422	0.7061
number_ds13		1	1	-0.8537	0.1198	50.7796	<.0001
number_ds13		2	1	0.3029	0.0697	18.9073	<.0001
time		1	1	0.00141	0.0133	0.0112	0.9157
time		2	1	-0.0256	0.0494	0.2681	0.6046
cubic_spline_b1		1	1	-0.0212	0.0409	0.2690	0.6040
cubic_spline_b1		2	1	0.0847	0.1434	0.3491	0.5546
cubic_spline_b2		1	1	-0.00574	0.00906	0.4015	0.5263
cubic_spline_b2		2	1	0.00460	0.0342	0.0181	0.8930
cubic_spline_b3		1	1	0.000277	0.000741	0.1398	0.7085
cubic_spline_b3		2	1	0.000349	0.00290	0.0145	0.9043
credit_card_pay*time		1	1	-0.0785	0.0131	35.8414	<.0001
credit_card_pay*time		2	1	-0.0868	0.1231	0.4965	0.4810
number_dsl*time		1	1	0.0393	0.0116	11.4476	0.0007
number_dsl*time		2	1	0.0354	0.0433	0.6677	0.4139
number_fds1*time		1	1	0.0418	0.0154	7.3796	0.0066
number_fds1*time		2	1	0.0336	0.0323	1.0827	0.2981
number_ds13*time		1	1	0.0414	0.0170	5.9080	0.0151
number_ds13*time		2	1	0.0387	0.0175	4.9243	0.0265

Odds Ratio Estimates

Effect		event_ category	Point Estimate	95% Wald Confidence Limits	
office	A vs O	1	1.197	0.963	1.488
office	A vs O	2	0.646	0.318	1.312
office	B vs O	1	1.184	0.956	1.466
office	B vs O	2	0.584	0.292	1.165
office	C vs O	1	1.232	0.988	1.537
office	C vs O	2	0.855	0.435	1.680
office	D vs O	1	1.355	1.068	1.719
office	D vs O	2	1.653	0.870	3.140
office	E vs O	1	1.260	0.976	1.628
office	E vs O	2	0.142	0.018	1.094
office	F vs O	1	1.189	0.944	1.499
office	F vs O	2	1.009	0.505	2.018
office	G vs O	1	1.545	1.229	1.943
office	G vs O	2	0.298	0.100	0.889
office	H vs O	1	1.151	0.914	1.451
office	H vs O	2	1.109	0.558	2.204
office	I vs O	1	1.195	0.914	1.561
office	I vs O	2	0.663	0.280	1.569
office	J vs O	1	1.257	0.957	1.651
office	J vs O	2	0.628	0.236	1.666
office	K vs O	1	0.813	0.608	1.088
office	K vs O	2	0.607	0.241	1.528
office	L vs O	1	1.440	1.078	1.924
office	L vs O	2	0.216	0.028	1.664
office	M vs O	1	0.996	0.717	1.384
office	M vs O	2	0.632	0.224	1.781
office	N vs O	1	1.080	0.827	1.409
office	N vs O	2	1.548	0.785	3.052
telecom		1	1.001	0.803	1.247
telecom		2	3.747	2.641	5.317
financial		1	0.703	0.547	0.904
financial		2	1.288	0.692	2.396
computer		1	0.786	0.662	0.933
computer		2	1.625	1.104	2.391
health		1	0.580	0.433	0.778
health		2	1.278	0.625	2.614
legal		1	0.666	0.531	0.835
legal		2	0.983	0.518	1.865
number_dial		1	0.992	0.961	1.023
number_dial		2	1.028	0.962	1.099
number_isdn		1	0.953	0.868	1.046
number_isdn		2	0.959	0.660	1.395
cubic_spline_b1		1	0.979	0.904	1.061
cubic_spline_b1		2	1.088	0.822	1.441
cubic_spline_b2		1	0.994	0.977	1.012
cubic_spline_b2		2	1.005	0.939	1.074
cubic_spline_b3		1	1.000	0.999	1.002
cubic_spline_b3		2	1.000	0.995	1.006

The hazard ratio for telecom for event type 2 indicates that customers in the business classified as telecommunications have 3.747 times the hazard of upgrading compared to customers not in a business classified as telecommunications.

A partial listing of the scoring code generated by PROC LOGISTIC is shown below. It is important to note that the code does not include the computations of the cubic spline basis functions. Furthermore, the variables with the predicted probabilities are **p_event_category1** for the probability of churning and **p_event_category2** for the probability of upgrade.

```
*****************************************;
** SAS Scoring Code for PROC Logistic;
*****************************************;
length I_event_category $ 12;
label I_event_category='Into: event_category' ;
label U_event_category='Unnormalized Into: event_category' ;
label P_event_category1='Predicted: event_category=1' ;
label P_event_category2='Predicted: event_category=2' ;
label P_event_category0='Predicted: event_category=0' ;

drop _LMR_BAD;
_LMR_BAD=0;

*** Check credit_card_payment for missing values;
if missing(credit_card_payment) then do;
   _LMR_BAD=1;
   goto _SKIP_000;
end;
```

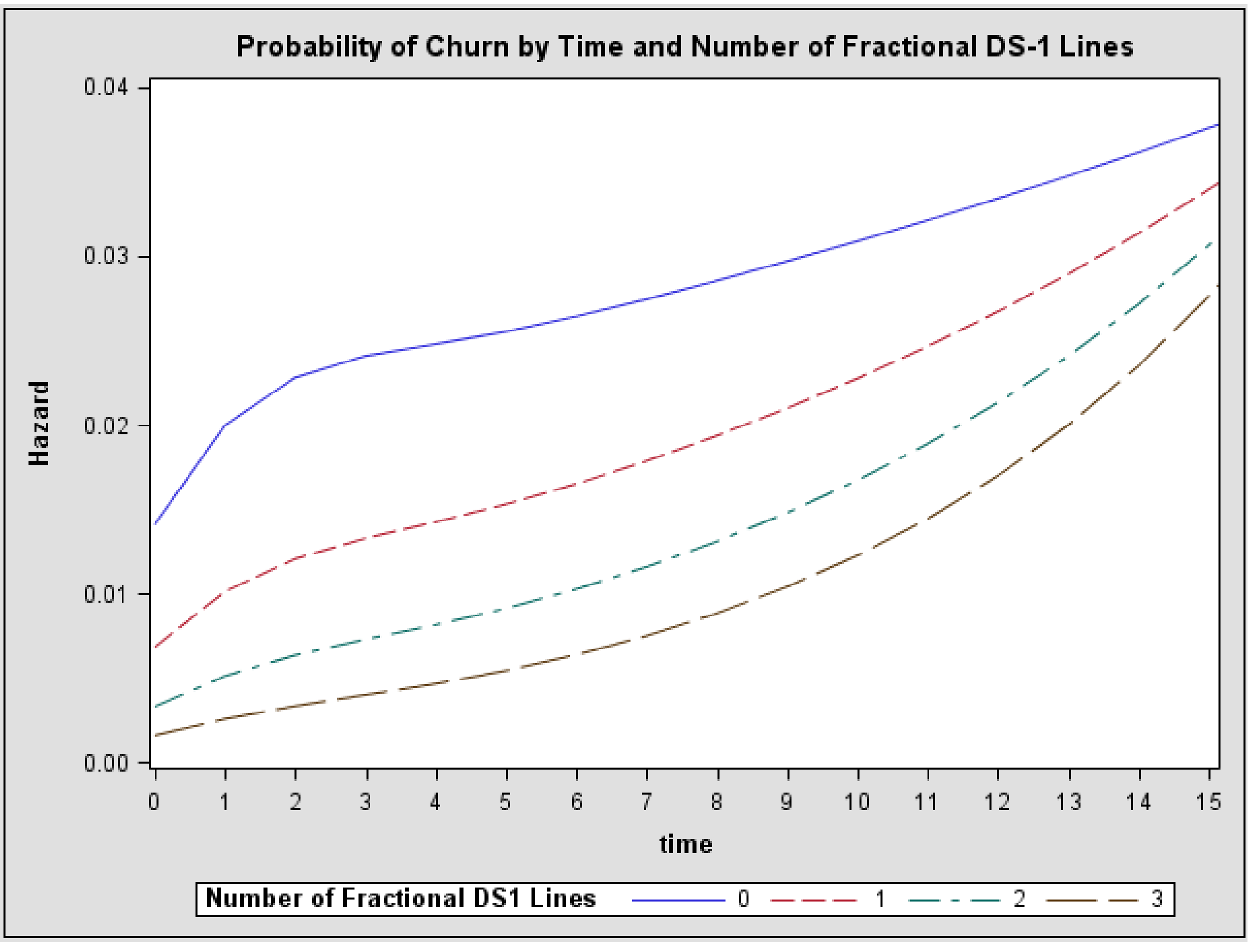
```
%let knots=2 4 8;

data work.plot;
   array knots{3} _temporary_ (&knots);
   label number_fds1="Number of Fractional DS1 Lines";
   do number_fds1=0,1,2,3;
     do time=0 to 16;
       cubic_spline_b1=(time>knots[1])*(time-knots[1])**3
        -time**3+3*knots[1]*time**2-3*knots[1]**2*time;
       cubic_spline_b2=(time>knots[2])*(time-knots[2])**3
        -time**3+3*knots[2]*time**2-3*knots[2]**2*time;
       cubic_spline_b3=(time>knots[3])*(time-knots[3])**3
        -time**3+3*knots[3]*time**2-3*knots[3]**2*time;
       office='B';
       credit_card_payment=0;
       telecom =0;
       financial=0;
       computer=0;
       health=0;
       legal=0;
       number_dial=1;
       number_isdn=0;
       number_dsl=1;
       number_ds13=0;
       %include "s:\workshop\model1.txt";
```

```
      output;
    end;
  end;
run;
```

The **plot** data set is fabricated and scored for making hazard plots. In this case, plots are made to depict the effect of prior fractional DS-1 access products on the hazard while the other covariates are held constant. The scoring code is added to the DATA step with the %include statement.

```
ods html style=default;
proc sgplot data=work.plot;
   series y=p_event_category1 x=time / group=number_fds1;
   yaxis label="Hazard";
   xaxis values=(0 to 15 by 1);
   title "Probability of Churn by Time and Number of Fractional DS-
          1 Lines";
run;
```

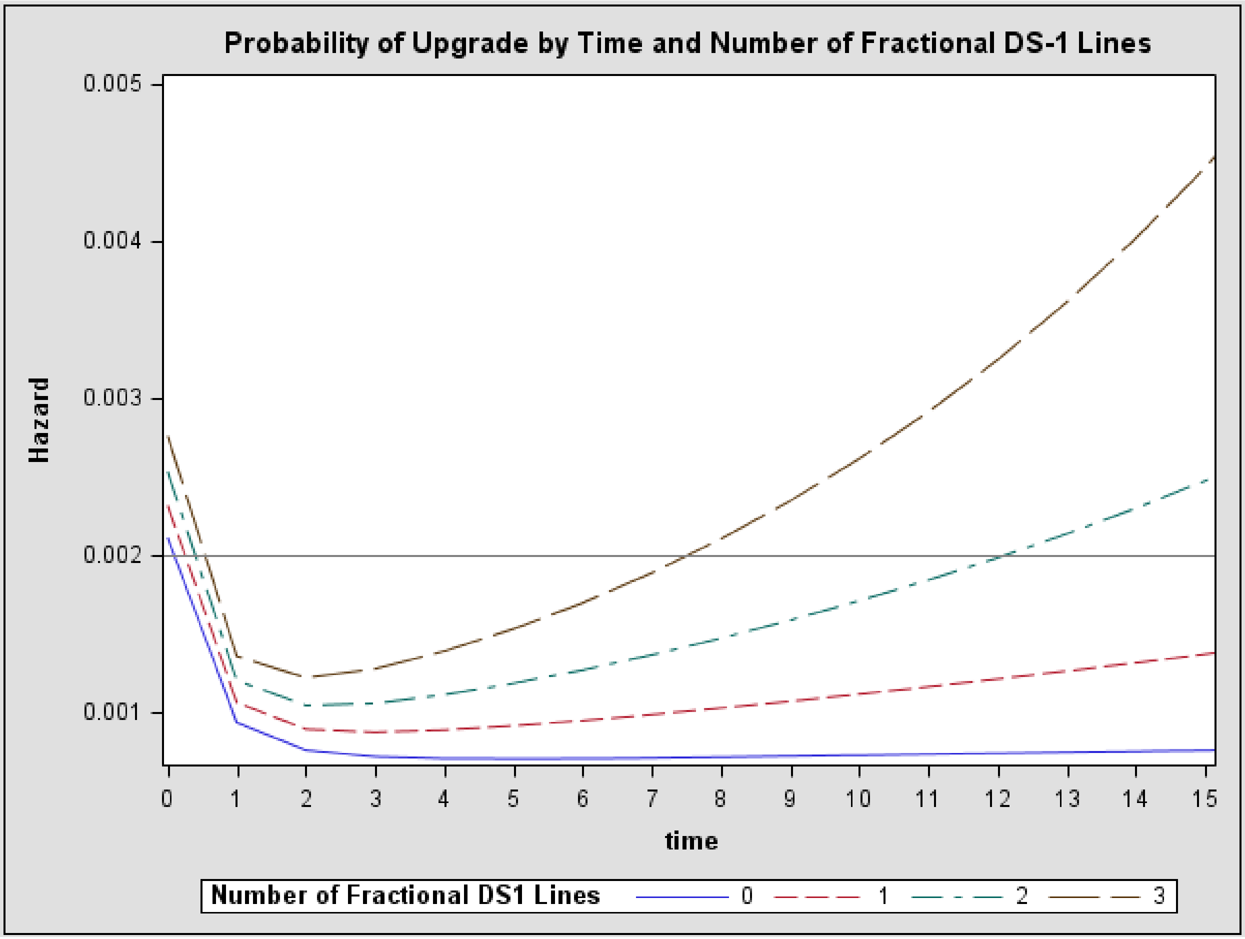
The graph is created in PROC SGPLOT. The SERIES statement creates a line plot and the GROUP= option generates separate lines for each number of fractional DS-1 lines.



The graph shows that the probability of churn over time is highest for customers with products with lower bandwidths than DS-1 (DIAL, DSL, and ISDN).

```
proc sgplot data=work.plot;
   series y=p_event_category2 x=time / group=number_fds1;
   yaxis label="Hazard";
   xaxis values=(0 to 15 by 1);
   refline 0.002;
   title "Probability of Upgrade by Time and Number of Fractional "
         "DS-1 Lines";
run;
```

The reference line is drawn on the graph based on business knowledge. For example, a profitable business decision might be to contact customers when their hazard is above the reference line.



The graph shows that the probability of upgrade over time is highest for customers with the highest bandwidths for DS-1 Access Products. Contacting customers when their hazard for upgrades is above 0.002 might lead to higher profits.

End of Demonstration

Wrap-Up

Thank you for attending our SAS seminar.

Instructor email: Mike.Patetta@sas.com

Course links:

https://support.sas.com/edu/schedules.html?ctry=us&crs=BMCE

https://support.sas.com/edu/schedules.html?ctry=us&crs=BDMSDM

46