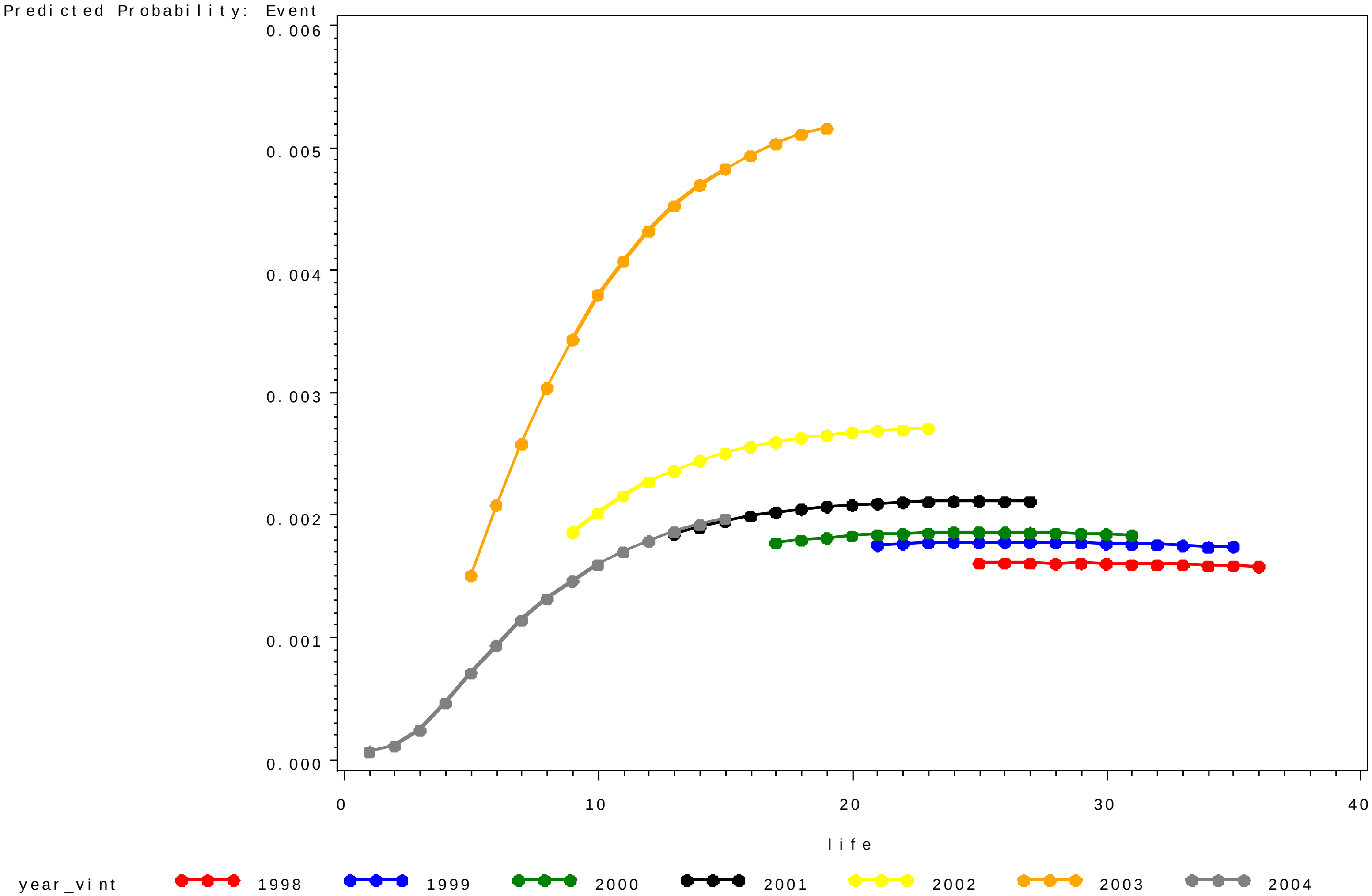
Time To Event Modeling

Time to Event Modeling: WHY?

Recognizes…

- the importance of time
- that your chance of default/churn/attrite/upsell depends not only on your attributes but also your tenure or your position in the typical customer life-cycle.

Examples…Hazard of loan default by 'Vintage'



Examples…Hazard of ESRD VS Kidney function

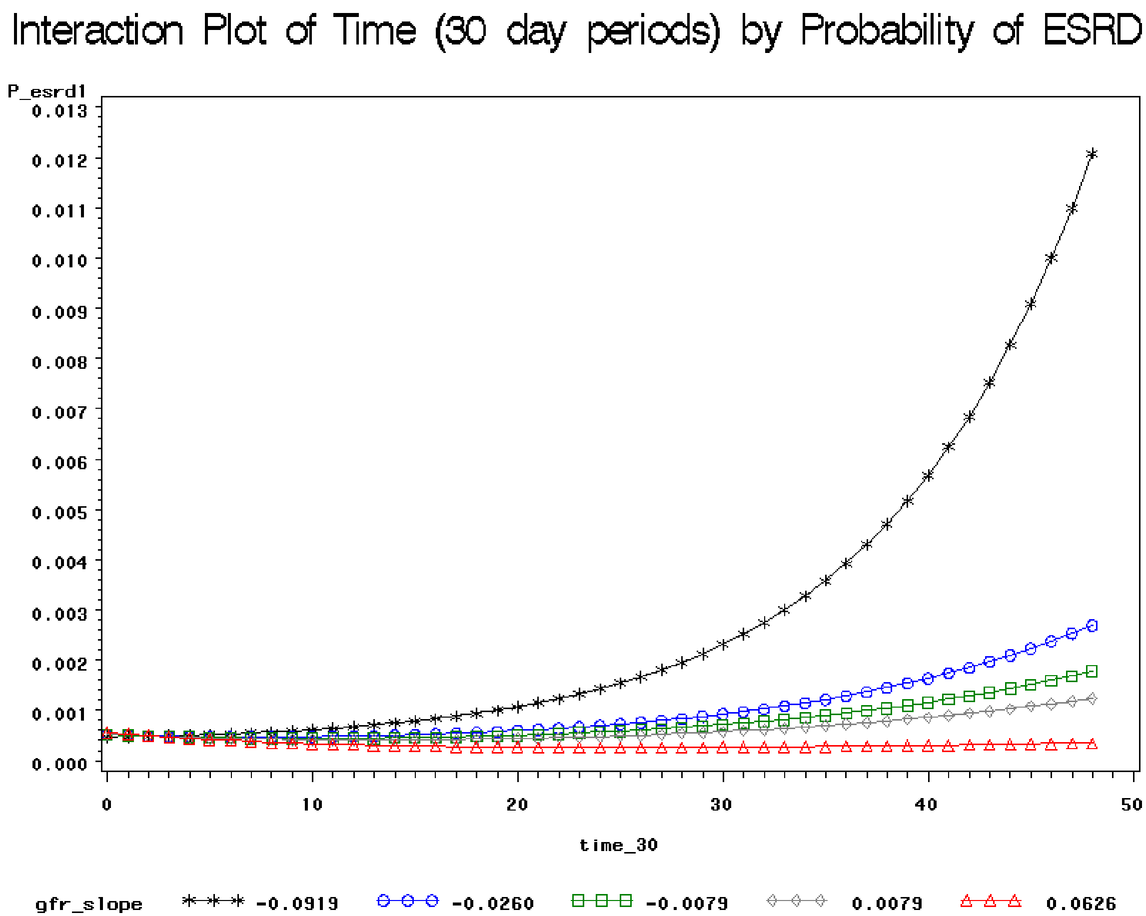
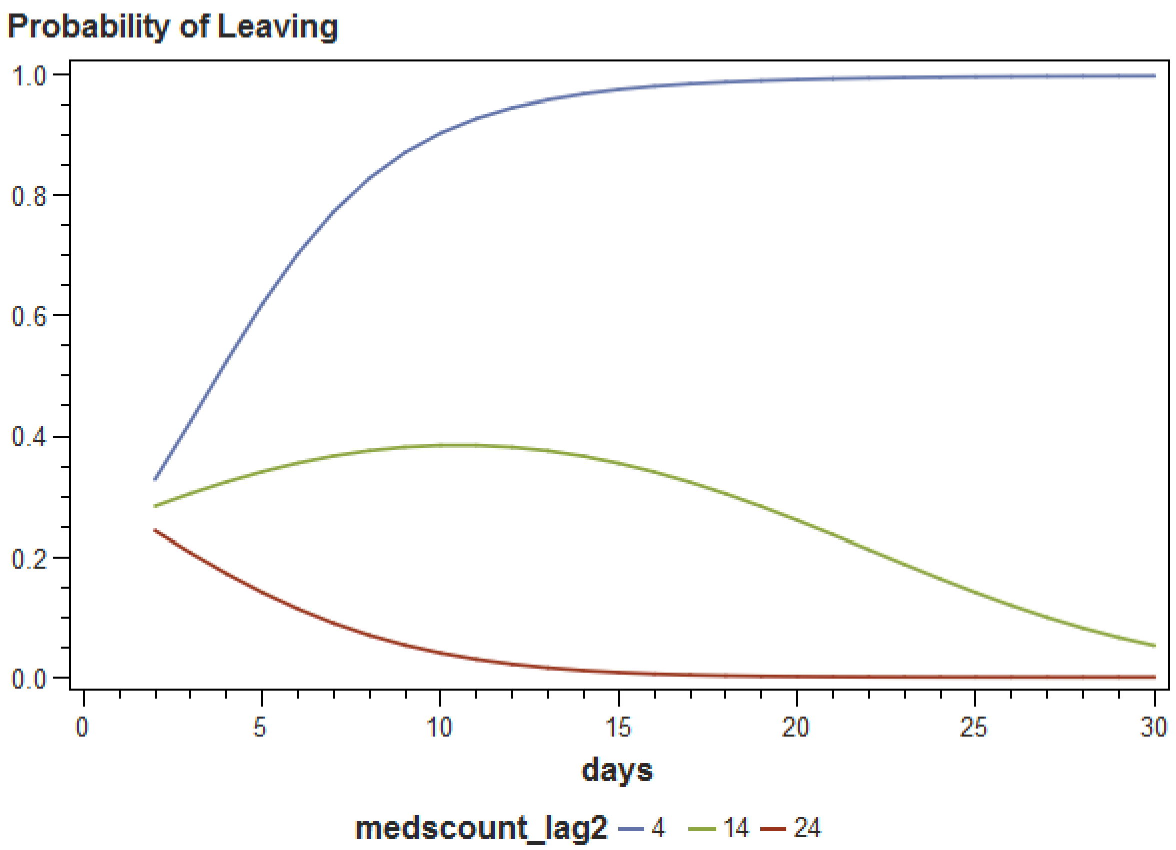
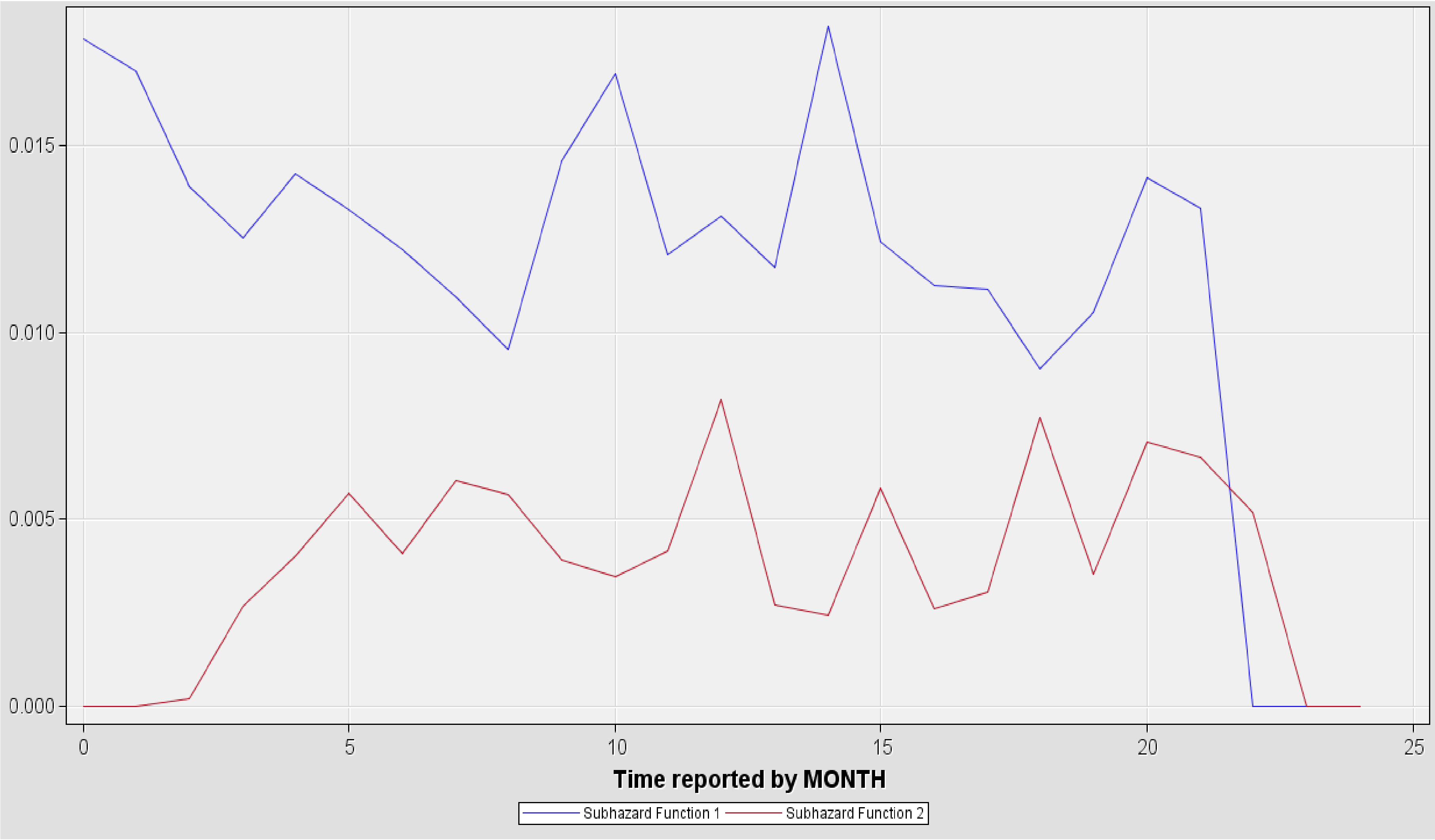


Figure 1. End Stage Renal Disease probability vs. Time . The gfr_slope values represent the 5th, 25th, 50th, 75th and 95th percentiles. Negative GFR slopes indicate declining kidney function.

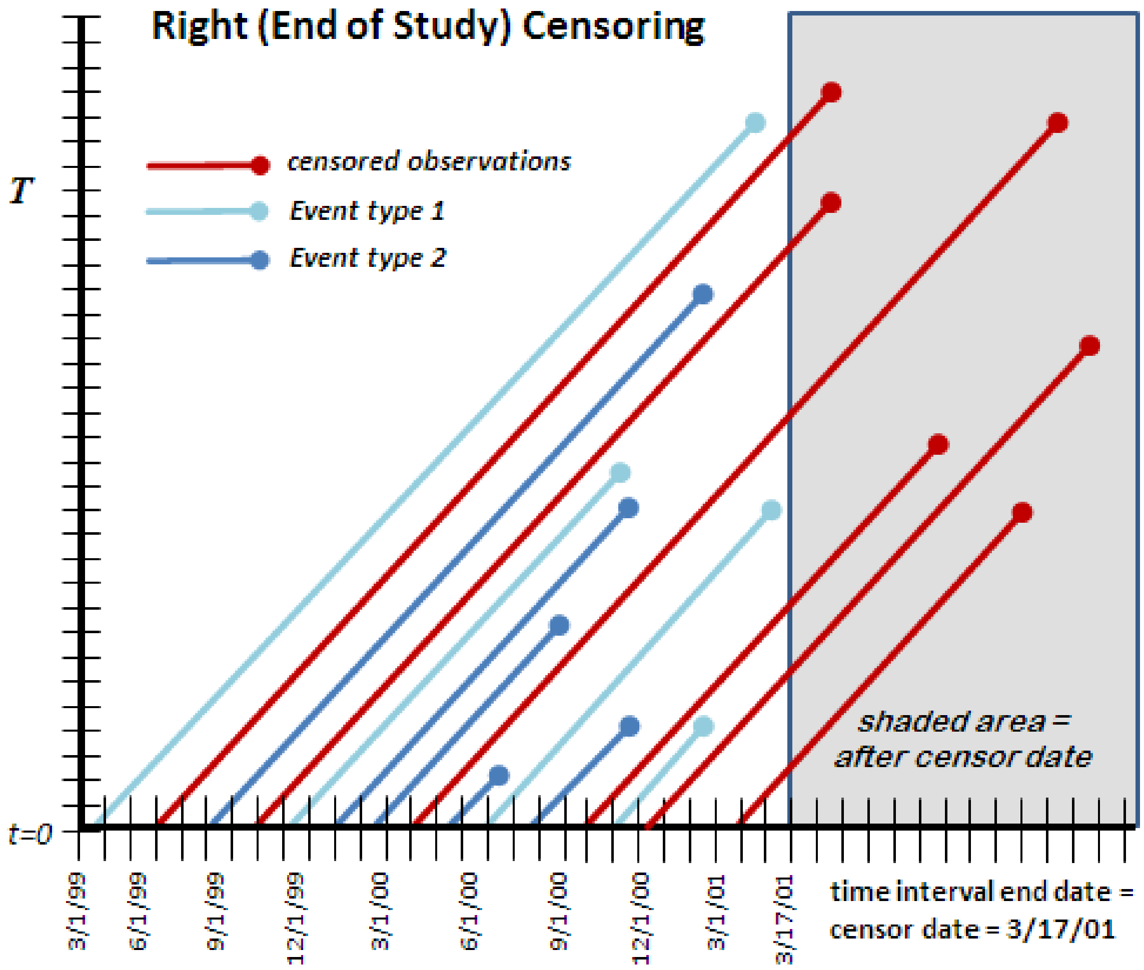
Examples…Hazard of leaving hospital VS Med counts



Hazard of Voluntary (1) and Involuntary (2) Churn

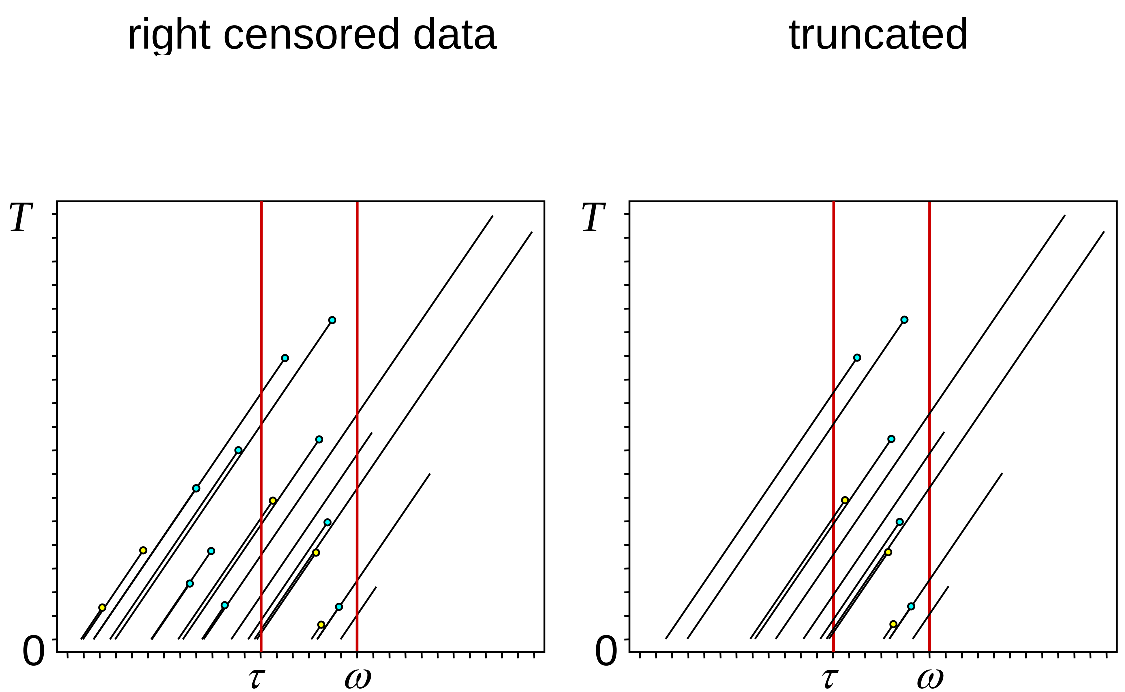


Characteristic of Survival Data: Right Censoring



Characteristic of Survival Data: Other Issues

Left Truncation



τ = truncation date

- Left-truncated data
- Competing Risks
- Time-dependent covariates
- Nonlinear Hazard functions

Traditional Approaches: The Cox Model

$$\log h_i(t) = \log h_0(t) + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}$$

$$h_i(t) = h_0(t) e^{\{\beta_1 X_{i1} + \ldots + \beta_k X_{ik}\}}$$

Baseline Hazard function – involves time but not predictor variables

Linear function of a set of predictor variables

What is a Hazard Function?

Conditional Probability

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t}$$

Interval of time

Instantaneous risk or potential

Discrete Time Logistic Hazards Model

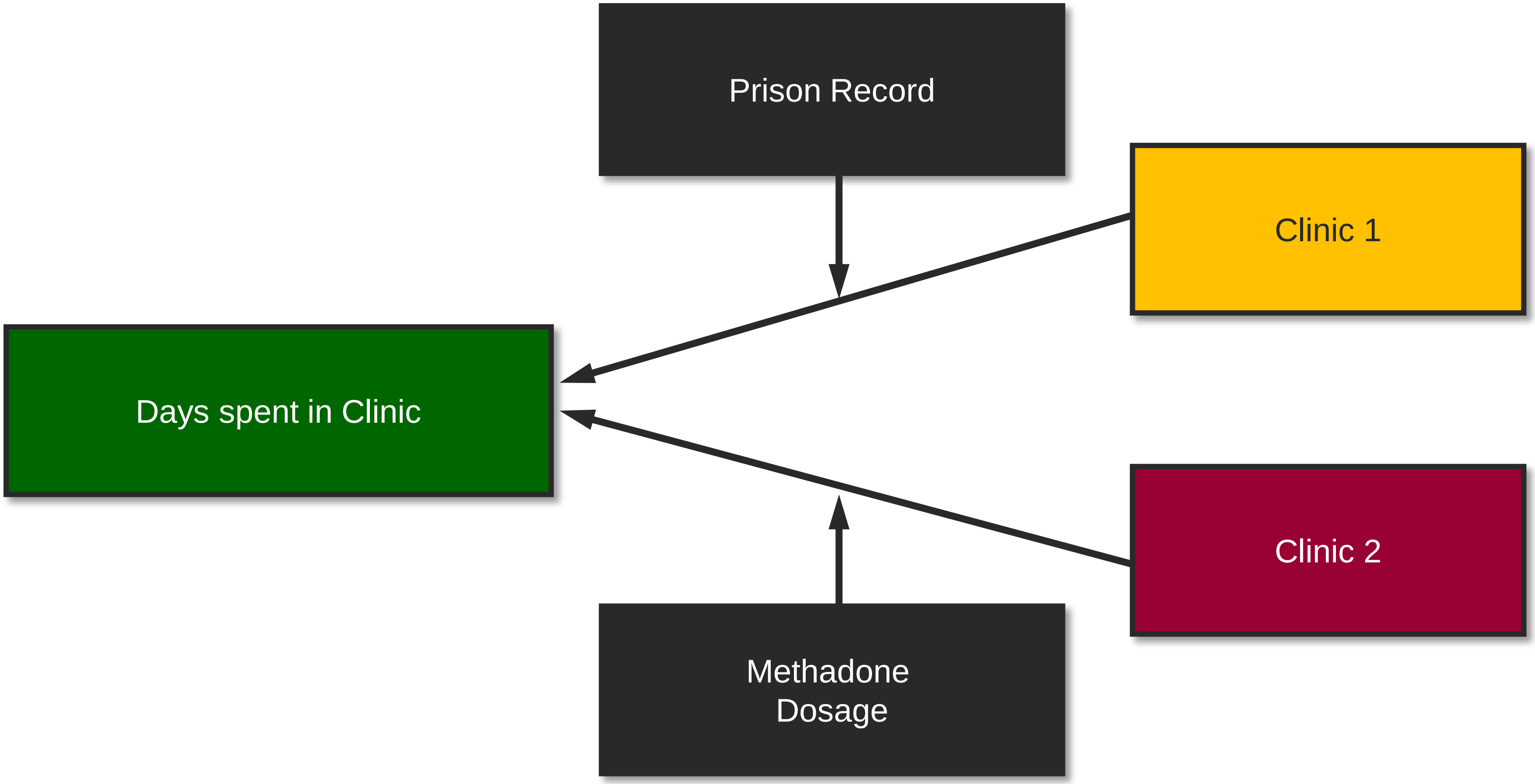
$$\ln\left(\frac{h(t,m\,|\,\mathbf{x}(t))}{1-h(t\,|\,\mathbf{x}(t))}\right)=\eta(t,\mathbf{x}(t),\boldsymbol{\theta}_m) \qquad m=1,\ldots,\kappa$$

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 same form but a different parameter vector for each competing risk.

Example…Methadone Treatment Data



Standard Data Structure and PHREG Code to Fit a Cox Model

PatientID	Clinic	Status	Time	Prison	Dose
98	1	1	237	0	45
99	1	1	517	0	70
100	1	1	749	0	70
101	1	1	150	1	80
102	1	1	465	0	65
103	2	1	708	1	60
104	2	0	713	0	50
105	2	0	146	0	50
106	2	1	450	0	55
109	2	0	555	0	80
110	2	1	460	0	50
111	2	0	53	1	60
113	2	1	122	1	60
114	2	1	35	1	40
118	2	0	532	0	70
119	2	0	684	0	65
120	2	0	769	1	70
121	2	0	591	0	70
122	2	0	769	1	40
123	2	0	609	1	100

```
proc phreg data=meth;
   model time*status(0)=clinic dose prison/rl;
run;
```

Transformed Data Structure & LOGISTIC Code

Patient	Clinic	Status	Time	Prison	Dose	days	target
201	0	1	127	0	20	126	0
201	0	1	127	0	20	127	1
202	0	1	7	1	40	1	0
202	0	1	7	1	40	2	0
202	0	1	7	1	40	3	0
202	0	1	7	1	40	4	0
202	0	1	7	1	40	5	0
202	0	1	7	1	40	6	0
202	0	1	7	1	40	7	1
203	0	1	29	1	60	1	0
203	0	1	29	1	60	2	0
203	0	1	29	1	60	3	0
203	0	1	29	1	60	4	0
203	0	1	29	1	60	5	0
203	0	1	29	1	60	6	0
203	0	1	29	1	60	7	0
203	0	1	29	1	60	8	0
203	0	1	29	1	60	9	0
203	0	1	29	1	60	10	0
203	0	1	29	1	60	11	0
203	0	1	29	1	60	12	0

```
proc logistic data=methadone desc;
model target=clinic dose prison days
/*days*days days*days*days*/;
run;
```

LOGISTIC vs. PHREG Output

The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
Clinic	1	-1.01069	0.21506	22.0853	<.0001	0.364	0.239	0.555
Dose	1	-0.03547	0.00639	30.8330	<.0001	0.965	0.953	0.977
Prison	1	0.32696	0.16742	3.8138	0.0508	1.387	0.999	1.925

The LOGISTIC Procedure

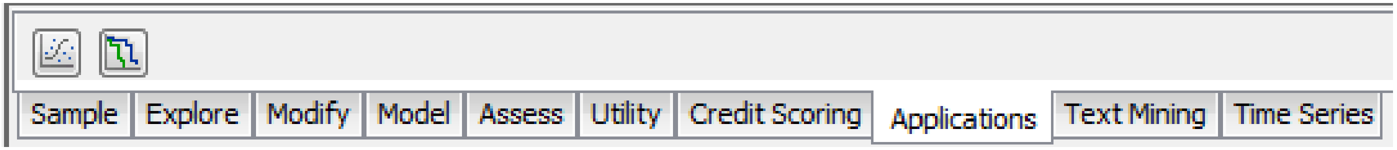
Analysis of Maximum Likelihood Estimates

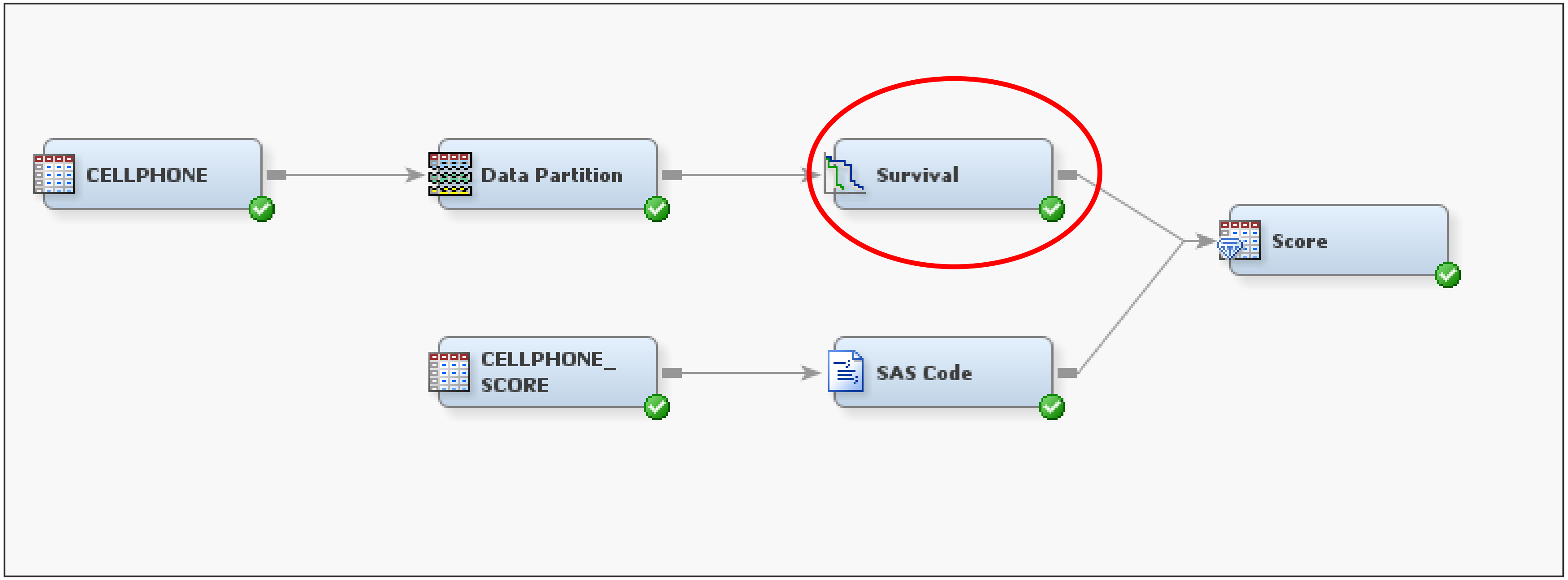
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.7778	0.3721	164.9102	<.0001
Clinic	1	-1.0303	0.2146	23.0609	<.0001
Dose	1	-0.0352	0.00632	31.0811	<.0001
Prison	1	0.3267	0.1666	3.8457	0.0499
days	1	0.00190	0.000372	26.1543	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Clinic	0.357	0.234	0.543
Dose	0.965	0.954	0.977
Prison	1.386	1.000	1.922
days	1.002	1.001	1.003

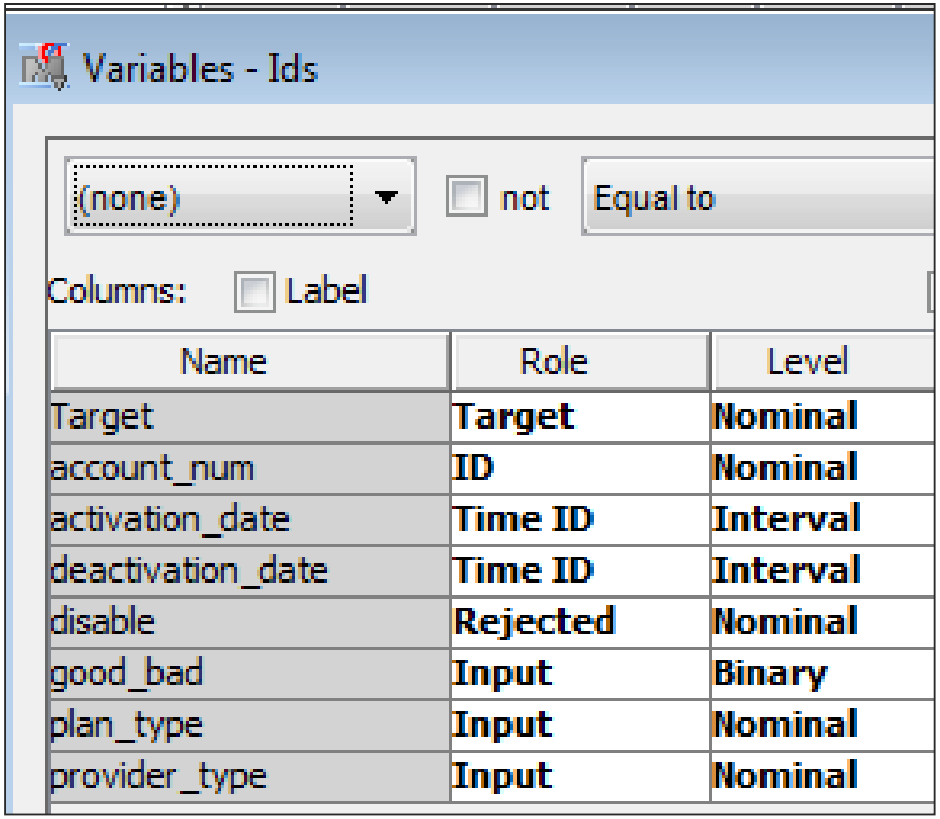
Predictive Survival Analysis in Enterprise Miner





Survival Node Requirements

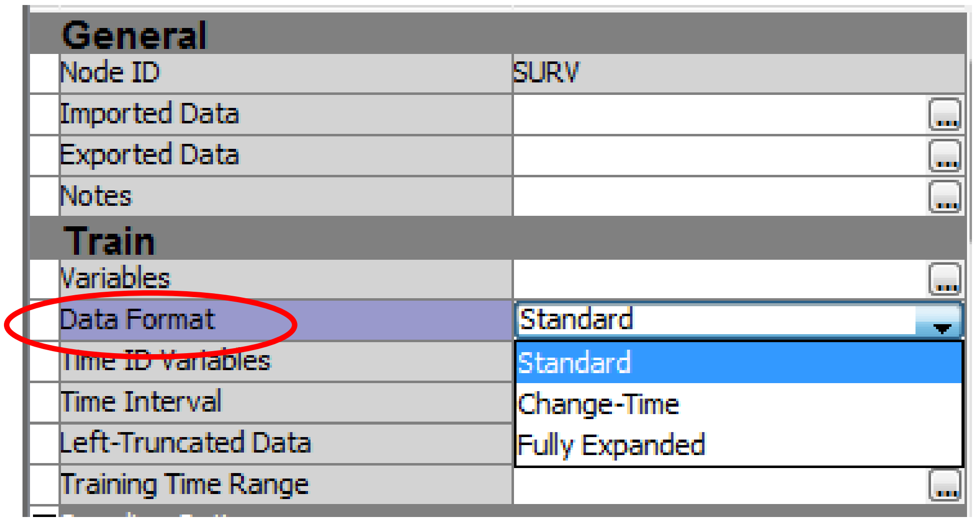
- The input data must have a unique ID variable (such as customer ID) for observations.
- At least two TIMEID variables are required. The first TIMEID variable maps to the inception, origin, or start date. The second TIMEID variable maps to the event date.
- At least one input variable is required for predictive hazard modeling using the Survival node.
- All input variables must be time independent **<u>prior to Version 12.3</u>**..
- There must be one numeric class target variable that represents the type of event that occurs on the event date.



Survival Node Version 12.3 and beyond

New versions now support three styles of data input…

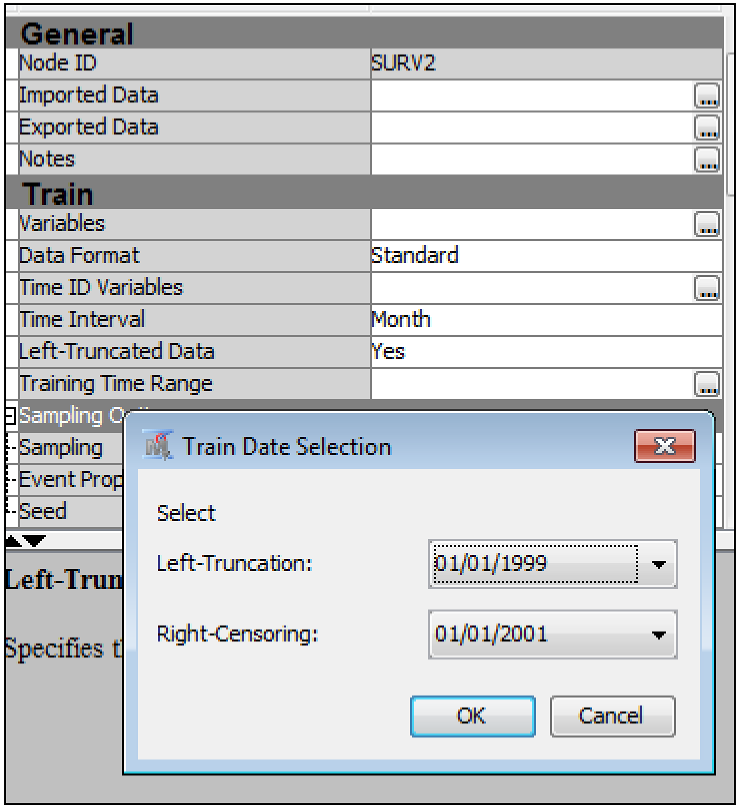
- ➢ Standard
- ➢ Change Time
- ➢ Fully Expanded



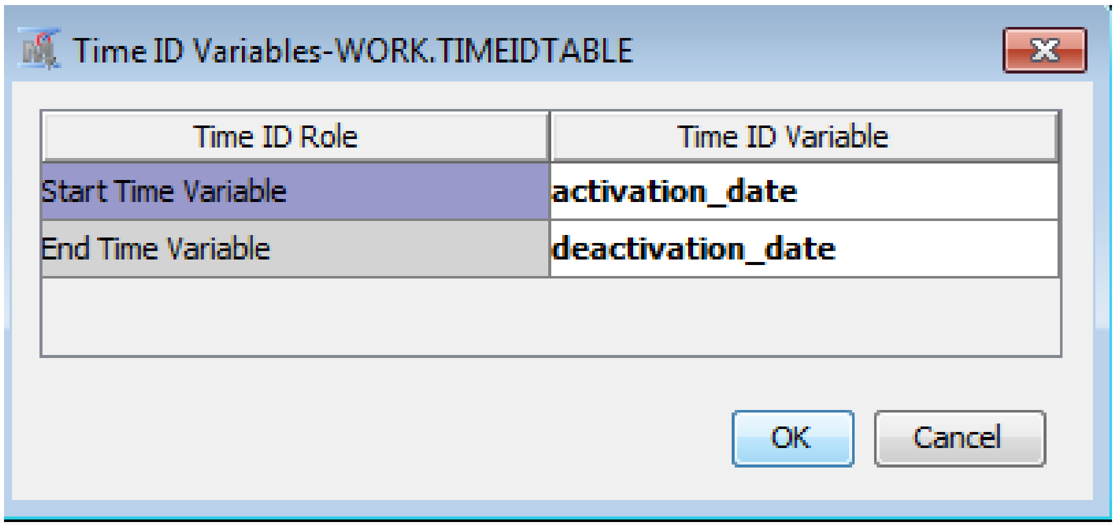
- ➢ Change Time and Fully Expanded formats allow for time dependent covariates.

Survival Node Version 12.3 and beyond

- New versions allow user specification of Left-truncation and Right-Censoring dates.



Survival Node: Standard Data Input



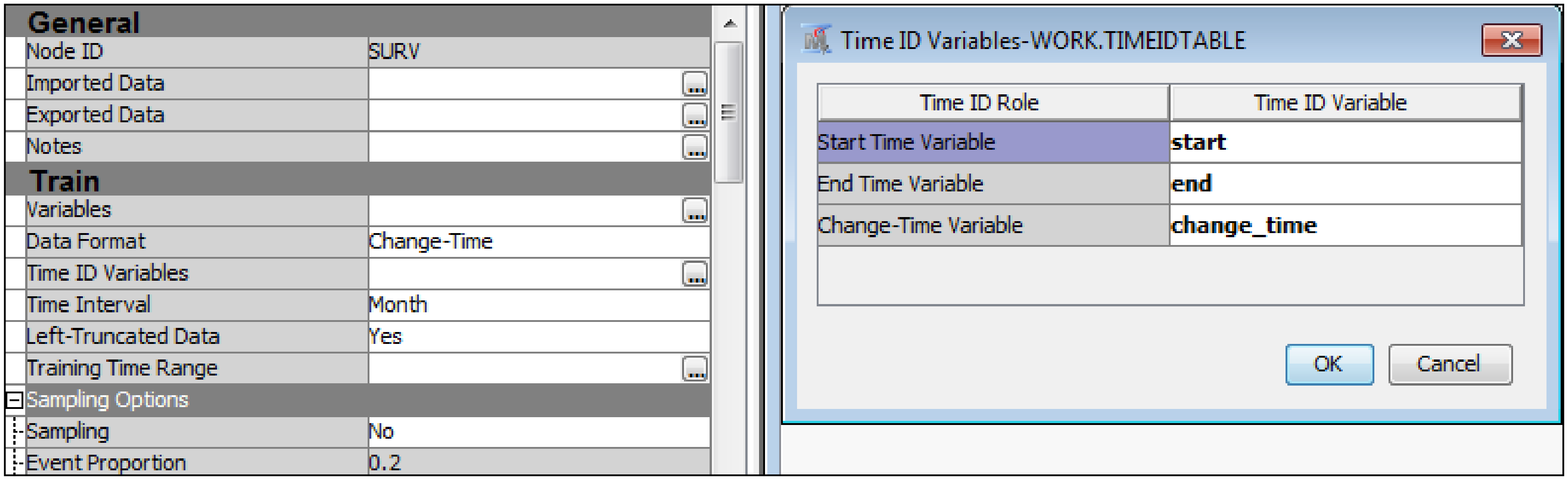
- Standard format requires a Start Date (Activation Date) and a “Censoring” Date (Deactivation Date).
- The Deactivation Date is set to a date value for events and missing for censored observations.
- By default EM choses the last event date in the data as the censoring date.

Survival Node: Standard Data Input

Obs #	account_num	Good Bad Credit Indicator	Provider Type	Type of Rate Plan	Disable Reason	Event Type	Activation Date	Deactivation Date
1	180437080184	1	PROV1	3		0	09/28/1999	.
2	180437283474	1	PROV1	1		0	01/09/2001	.
3	180437340410	0	PROV1	1		0	12/31/1999	.
4	180437356568	0	PROV2	1	DUE	2	12/22/1999	06/28/2000
5	180437356837	1	PROV3	1		0	04/17/2000	.
6	180437375280	1	PROV1	2	TRANSFER	1	08/16/1999	08/21/2000
7	180437392909	1	PROV3	1		0	07/26/1999	.
8	180437420657	0	PROV2	1		0	12/15/1999	.
9	180437433673	0	PROV1	3		0	11/21/2000	.
10	180437452331	0	PROV3	2		0	12/28/2000	.
11	180437466686	1	PROV3	3		0	07/15/2000	.
12	180437492423	1	PROV1	1		0	11/20/2000	.
13	180437494586	0	PROV1	2		0	08/29/2000	.
14	180437498878	0	PROV1	2		0	06/16/2000	.
15	180437499481	1	PROV2	1		0	07/03/1999	.
16	180437502892	1	PROV1	3		0	03/22/2000	.
17	180437507436	1	PROV1	1		0	07/02/1999	.
18	180437512268	0	PROV2	1	PAY	1	08/29/1999	07/13/2000
19	180437514966	1	PROV1	1	PAY	1	12/04/1999	06/09/2000

- Standard data contains one row per individual. Time dependent information cannot be modeled.
- EM creates fully expanded data before fitting the Logistic regression model.

Survival Node: ChangeTime Data Input (V12.3)



- Change Time Format requires three Time ID Roles: Start Time, EndTime, and Change-Time.

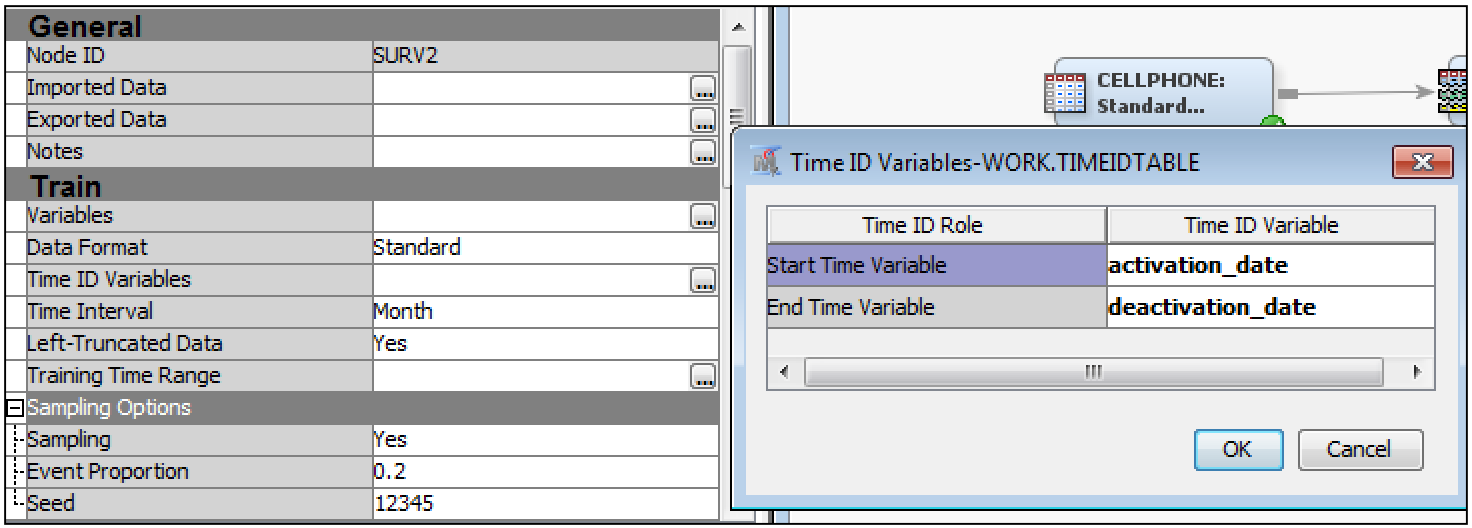
Survival Node: ChangeTime Data Input (V12.3)

SAMPSIO.CHURN_CHANGETIME

Obs #	customer_id ▲	promotions	num_complaints	churn	start	end	change_time
1	1	1	0	1	20May1988	10Jul1988	20May1988
2	1	1	5	1	20May1988	10Jul1988	27May1988
3	1	1	6	1	20May1988	10Jul1988	03Jun1988
4	1	1	8	1	20May1988	10Jul1988	10Jun1988
5	1	1	10	1	20May1988	10Jul1988	17Jun1988
6	2	1	0	1	10Nov1987	21Feb1988	10Nov1987
7	2	1	1	1	10Nov1987	21Feb1988	05Jan1988
8	3	1	0	0	27Jul1987	.	27Jul1987
9	3	1	1	0	27Jul1987	.	03Aug1987
10	4	1	0	0	17Jan1988	.	17Jan1988

- A row of data is added to a subject whenever an input variable value changes (time-dependent variable). The variable value is added and the Change Time variable indicates when the new values occurred.

Survival Node: Fully Expanded Data Input (V12.3)



- Fully expanded data requires two Time ID Roles: Start Time and End Time.

Survival Node: Fully Expanded Data Input

- Fully expanded data contains one row per each individual x time. Time dependent information can also be captured in this data format.
- Expanded data must also include a time index variable called, _t_ that is set to a role of Input.

SAMPSIO.CHURN_FULLYEXPANDED_WEEKLY

Obs #	customer_id	_t_	promotions	num_complaints	churn	start	end
1	1	0	1	0	1	20May1988	10Jul1988
2	1	1	1	5	1	20May1988	10Jul1988
3	1	2	1	6	1	20May1988	10Jul1988
4	1	3	1	8	1	20May1988	10Jul1988
5	1	4	1	10	1	20May1988	10Jul1988
6	1	5	1	10	1	20May1988	10Jul1988
7	1	6	1	10	1	20May1988	10Jul1988
8	1	7	1	10	1	20May1988	10Jul1988
9	1	8	1	10	1	20May1988	10Jul1988
10	2	0	1	0	1	10Nov1987	21Feb1988
11	2	1	1	0	1	10Nov1987	21Feb1988
12	2	2	1	0	1	10Nov1987	21Feb1988
13	2	3	1	0	1	10Nov1987	21Feb1988
14	2	4	1	0	1	10Nov1987	21Feb1988
15	2	5	1	0	1	10Nov1987	21Feb1988
16	2	6	1	0	1	10Nov1987	21Feb1988
17	2	7	1	0	1	10Nov1987	21Feb1988
18	2	8	1	1	1	10Nov1987	21Feb1988
19	2	9	1	1	1	10Nov1987	21Feb1988
20	2	10	1	1	1	10Nov1987	21Feb1988
21	2	11	1	1	1	10Nov1987	21Feb1988
22	2	12	1	1	1	10Nov1987	21Feb1988
23	2	13	1	1	1	10Nov1987	21Feb1988
24	2	14	1	1	1	10Nov1987	21Feb1988
25	2	15	1	1	1	10Nov1987	21Feb1988
26	3	0	1	0	0	27Jul1987	.
27	3	1	1	1	0	27Jul1987	.
28	3	2	1	1	0	27Jul1987	.
29	3	3	1	1	0	27Jul1987	.
30	3	4	1	1	0	27Jul1987	

Survival Node: Fully Expanded and ChangeTime

Property	Value
Training Time Range	...
Sampling Options	
Sampling	No
Event Proportion	0.2
Seed	12345
Regression Spline Model	
Covariate x Time Interactions	Include all
Covariates for Interactions	...
Stepwise Regression	Yes
Entry Significance Level	0.05
Stay Significance Level	0.05
Number of Knots	5
Knot Selection	No
Survival Validation	
Survival Validation Method	Default
Validation Score Date	...

- Fully expanded and ChangeTime formats can accommodate time dependent variables. They can optionally include Input (Covariate) X Time interaction terms.

Sampling and Partitioning Data

Oversampling

- ❖ The survival node allows for oversampling to a desired proportion of events since expanding the modeling event data to represent one customer record per unit time can quickly create very large input data tables that are impractical to use for modeling.
- ❖ The use can specify the event rate for oversampling.

Data Partition

- ❖ NOTE: If you are using Change Time or Expanded data formats then the Data Partition node must be configured to do Cluster based sampling with ID as the Cluster variable so that individual within each ID are not assigned to different data partitions.

Modeling Hazards

- The discrete event time represents the duration from the inception (start) time until the censoring date.
- The hazard function represents the conditional probability of an event at time *t* or, in other words, the probability of experiencing the event at time *t* given survival up to that time point.
- Cubic spline basis functions of discrete time are used as predictors in the multinomial logistic regression to model baseline hazards and subhazard.
- Transforming the event time function with cubic spline basis functions allows the hazard and sub-hazard functions to be more flexible. This results in a greater ability to detect and model customer behavior patterns.

Modeling Hazards: Cubic Spline Basis Functions

The cubic spline basis functions are segmented functions composed of polynomials, joined at knots, or points where the function makes a transformation. For example, a knot is the point at which one of the cubic spline basis functions changes from a cubic function to a constant function.

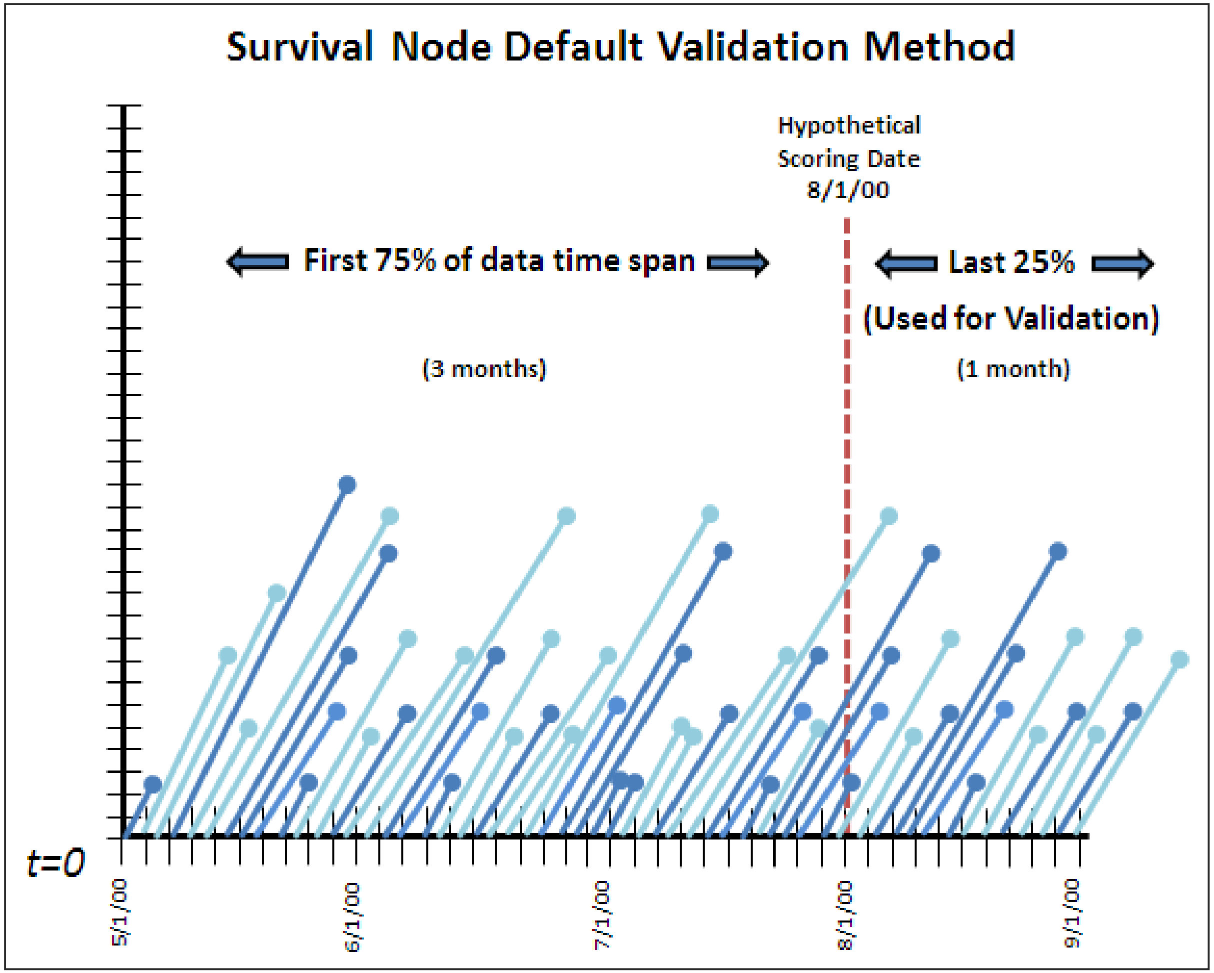
.

$$csb(t, k_j) = \begin{cases} -t^3 + 3k_j t^2 - 3k_j^2 t & if\ t \leq\ k_j \\ -k_j^3 & if\ t >\ k_j \end{cases}$$

where *j* is the number of knots and *k* is the value of the knot.

Model Validation

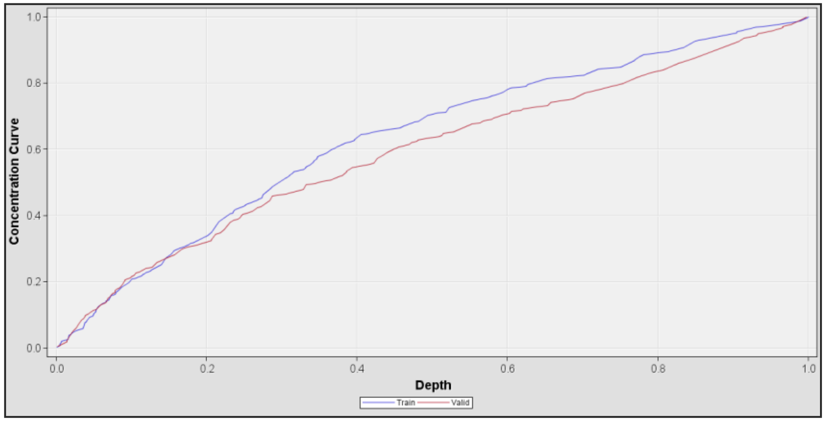
- Validation is internal to the survival node because of the use of a time dimension and the presence of right censoring that complicates assessment.
- By default, the last quarter of data are used to validate survival models in EM.
- K-S statistic, Benefit and Gini concentration ratio are reported for training and validation.

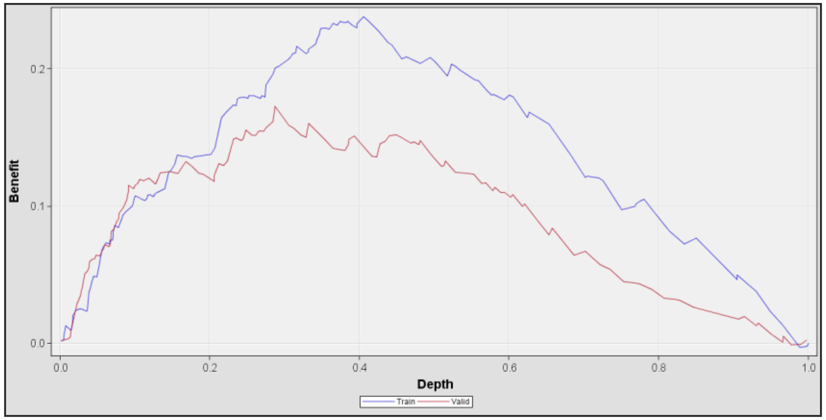


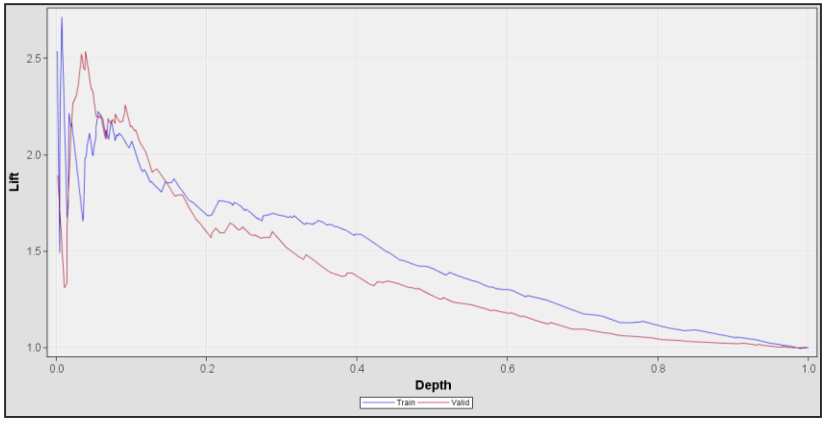
Model Validation

Model validation reports include the K-S , Lift ,Benefit, and Gini concentration ratio charts and statistics such as:

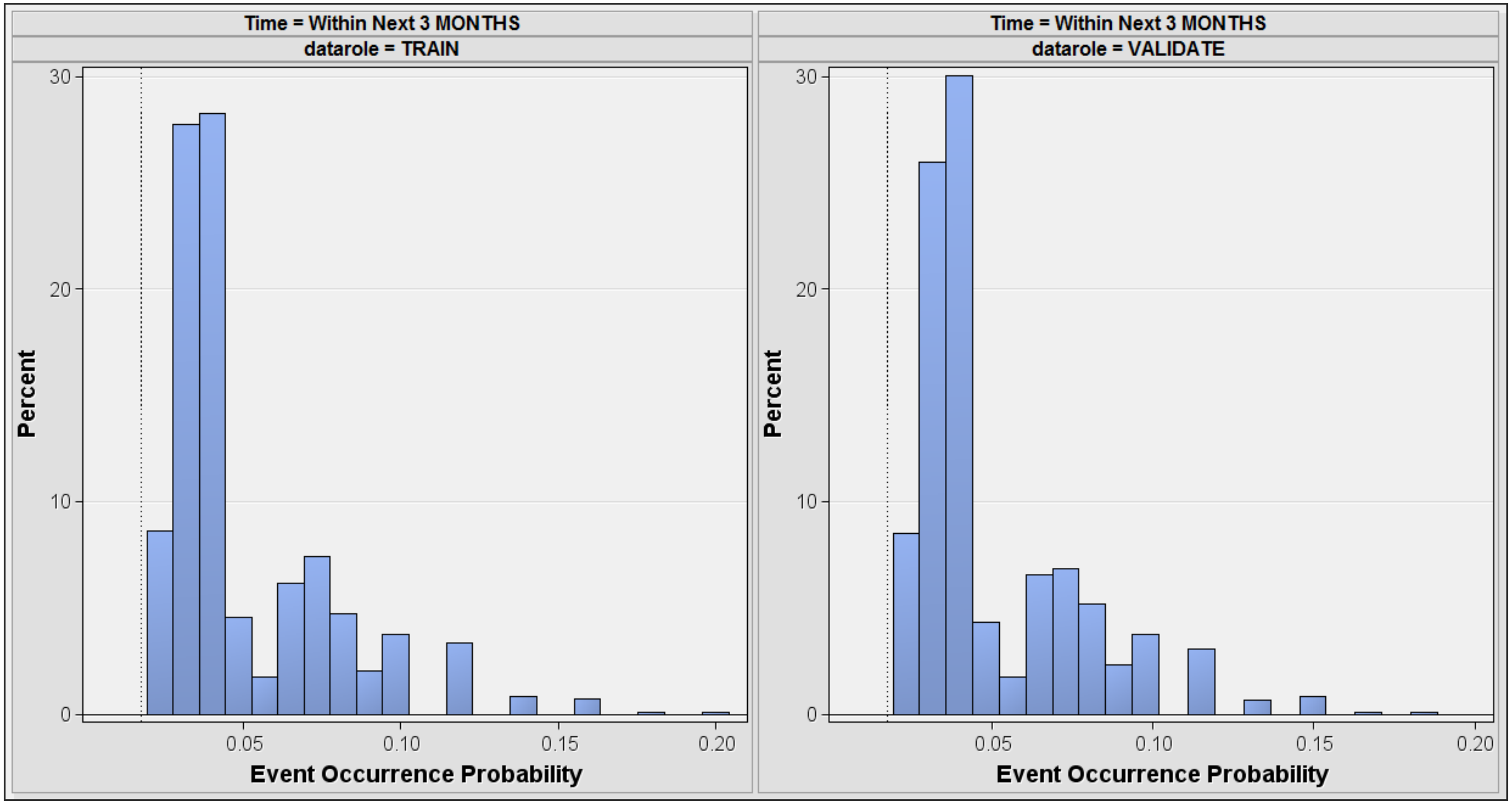
- **Benefit** the maximum benefit value
- **Lift** the lift at the maximum benefit value
- **Kolmogorov-Smirnov statistic** the maximum distance between the event and non-event distributions
- **Gini Concentration Ratio** twice the area between the concentration curve and the random model (represented by a 45–degree diagonal line).





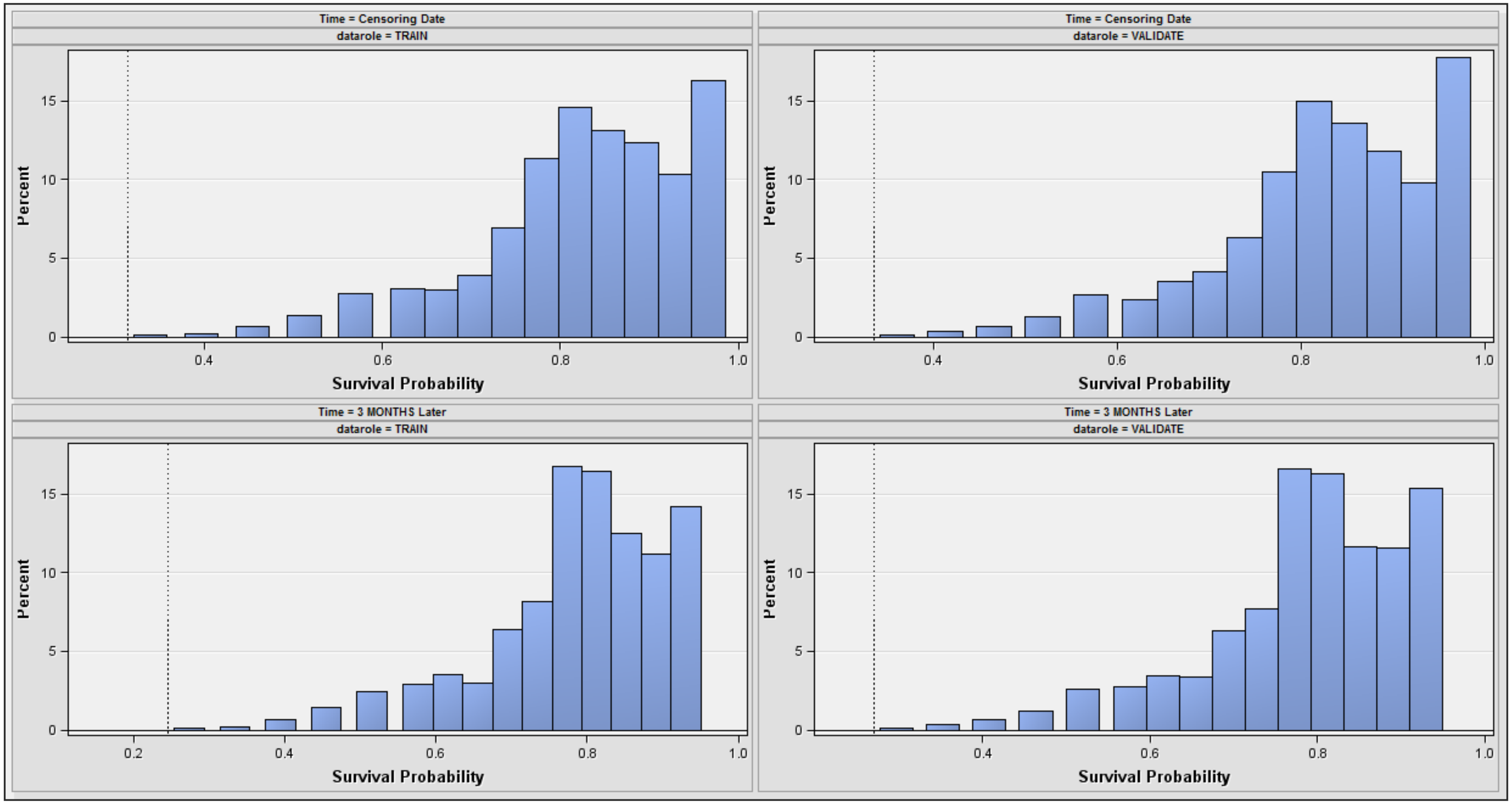


Default Results: Event and Survival Histograms



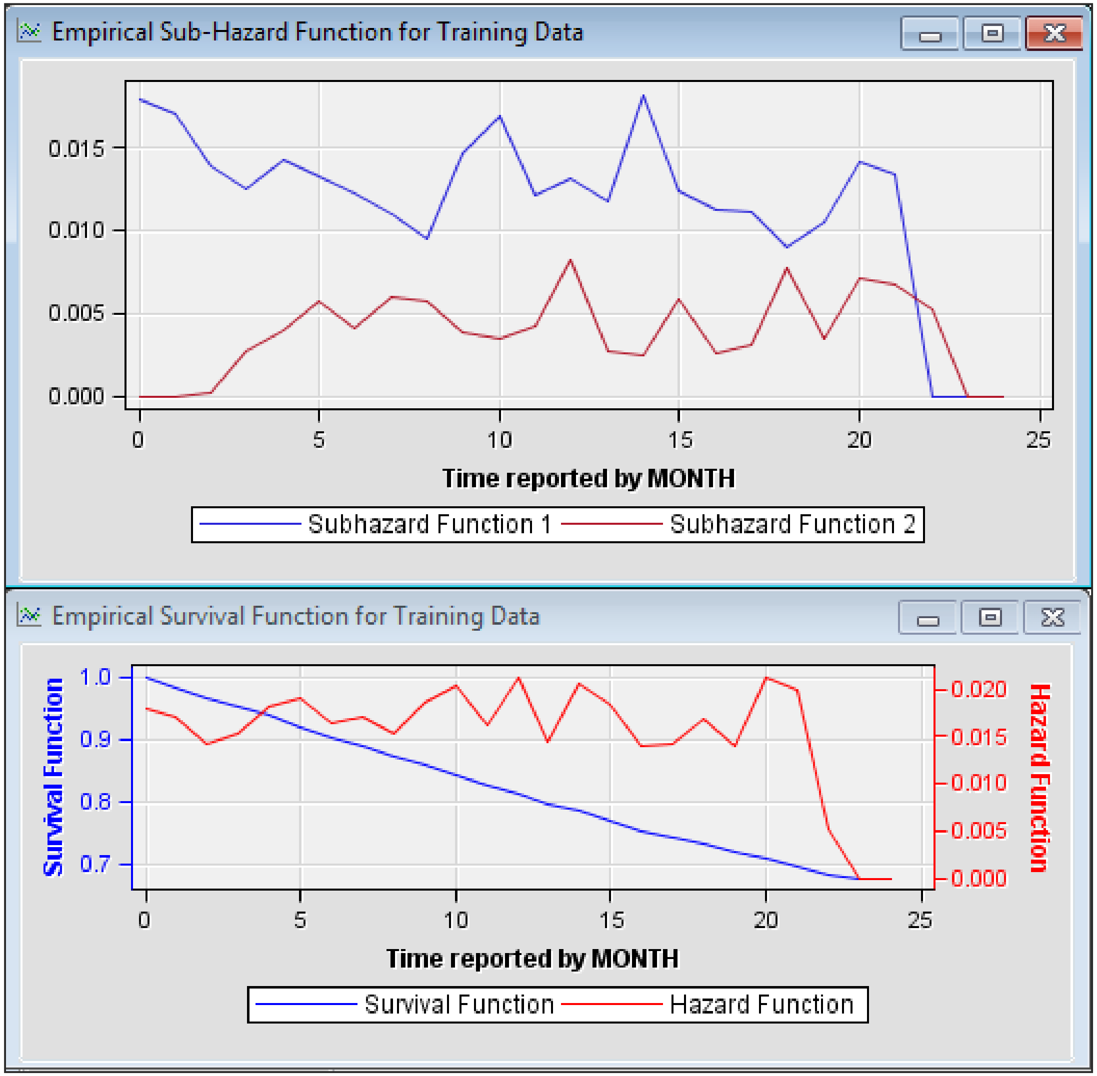
The Event Occurrence Probability histogram displays the distribution of the probabilities of having an event of interest occur within the next three time units.

Default Results: Event and Survival Histograms



The Survival Probability Histogram for three time units later displays the probabilities that a customer account will remain active during the three-month interval that follows the censor date.

Default Results: Hazard, Sub-Hazard and Survival Functions



Default Results: Nominal Logistic Regression

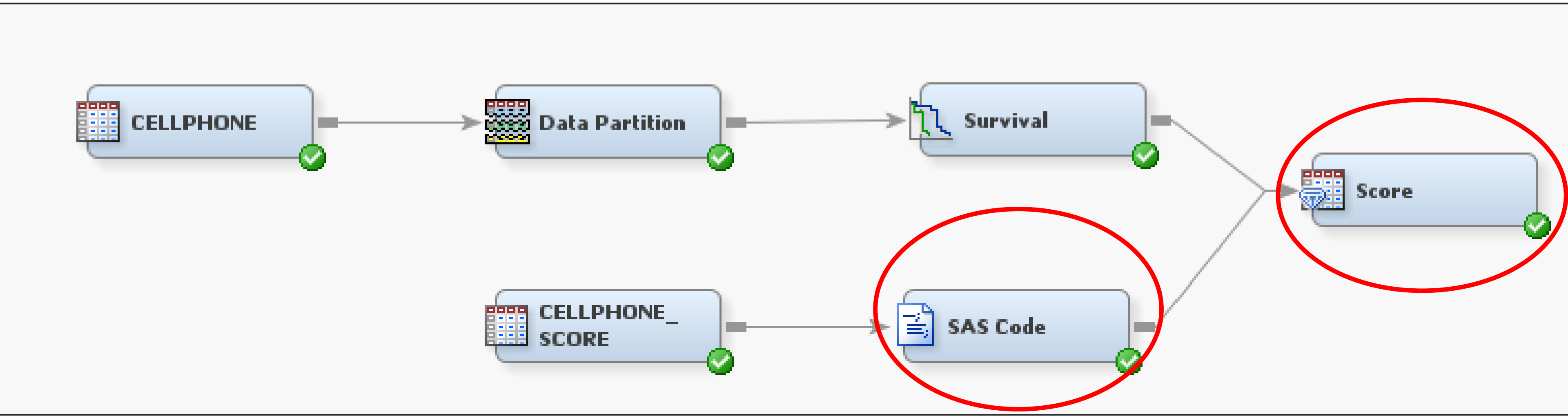
Output

Parameter		_g_	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	
Intercept		2	1	-27.0970	9.5285	8.09	0.0045	
Intercept		1	1	-3.7797	0.1507	629.25	<.0001	
_t_		2	1	0.1213	0.1213	1.00	0.3171	
_t_		1	1	-0.0871	0.0894	0.95	0.3298	
_csb1		2	1	-0.2800	0.1813	2.39	0.1225	
_csb1		1	1	0.00345	0.0114	0.09	0.7623	
_csb2		2	1	-0.0108	0.0174	0.38	0.5353	
_csb2		1	1	-0.00387	0.00710	0.30	0.5858	
_csb3		2	1	0.00617	0.0121	0.26	0.6097	
_csb3		1	1	0.00395	0.00672	0.35	0.5567	
_csb4		2	1	-0.00567	0.00843	0.45	0.5014	
_csb4		1	1	-0.00092	0.00521	0.03	0.8603	
_csb5		2	1	0.00218	0.00276	0.62	0.4297	
_csb5		1	1	-0.00029	0.00179	0.03	0.8710	
plan_type	1	2	1	-0.5270	0.1696	9.65	0.0019	
plan_type	1	1	1	-0.2843	0.0906	9.85	0.0017	
plan_type	2	2	1	-0.4019	0.2237	3.23	0.0724	
plan_type	2	1	1	-0.2012	0.1179	2.91	0.0878	
plan_type	3	2	0	0	.	.	.	
plan_type	3	1	0	0	.	.	.	
provider_type	PROV1	2	1	-0.2427	0.2180	1.24	0.2656	
provider_type	PROV1	1	1	-0.1442	0.1041	1.92	0.1659	
provider_type	PROV2	2	1	-0.5846	0.3175	3.39	0.0656	
provider_type	PROV2	1	1	0.00825	0.1323	0.00	0.9503	
provider_type	PROV3	2	1	0.3596	0.2353	2.34	0.1265	
provider_type	PROV3	1	1	-0.00811	0.1204	0.00	0.9463	
provider_type	PROV4	2	0	0	.	.	.	.
provider_type	PROV4	1	0	0	.	.	.	.

Odds Ratio Estimates

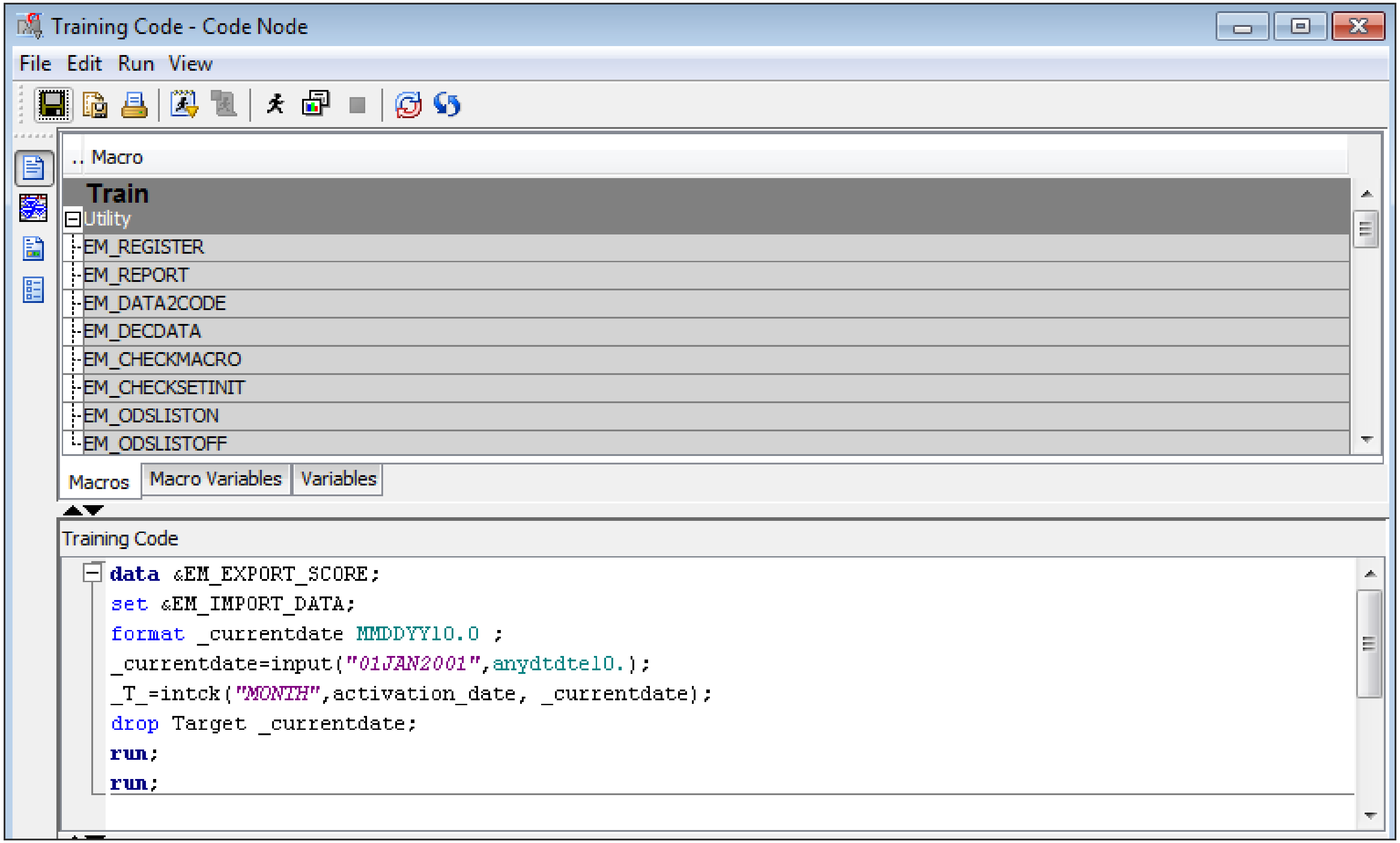
Effect		_g_	Point Estimate
_t_		2	1.129
_t_		1	0.917
_csb1		2	0.756
_csb1		1	1.003
_csb2		2	0.989
_csb2		1	0.996
_csb3		2	1.006
_csb3		1	1.004
_csb4		2	0.994
_csb4		1	0.999
_csb5		2	1.002
_csb5		1	1.000
plan_type	1 vs 3	2	0.590
plan_type	1 vs 3	1	0.753
plan_type	2 vs 3	2	0.669
plan_type	2 vs 3	1	0.818
provider_type	PROV1 vs PROV4	2	0.785
provider_type	PROV1 vs PROV4	1	0.866
provider_type	PROV2 vs PROV4	2	0.557
provider_type	PROV2 vs PROV4	1	1.008
provider_type	PROV3 vs PROV4	2	1.433
provider_type	PROV3 vs PROV4	1	0.992
good_bad	0 vs 1	2	12.846
good_bad	0 vs 1	1	1.322

Scoring



CELLPHONE
Data Partition
Survival
Score
CELLPHONE_
SCORE
SAS Code

Scoring



In order to score, a variable _T_ must be calculate._T_ is the time from inception until the current date used at scoring.

Scoring: Key Variables

EMWS1.Score_SCORE

account_num	Survival Probability at Censoring Time	Survival Probability at Future Time	Event Probability before or at the Future Time	Mean Residual Life RMRL
180437020551	0.85306	0.603516	0.292528	21.3372
180437142445	0.787119	0.713225	0.09388	36.8648
180437151668	0.965925	0.798461	0.173371	38.8358
180437162450	0.787218	0.711676	0.09596	33.1327
180437165776	0.805834	0.721911	0.104144	37.9031
180437202982	0.965925	0.798461	0.173371	38.8358
180437219430	0.919394	0.626161	0.318941	21.5986
180437242709	0.893228	0.734863	0.177296	35.5689
180437248254	0.879172	0.771805	0.122123	38.7427
180437257019	0.846757	0.727426	0.140927	33.4759
180437266960	0.879172	0.771805	0.122123	38.7427
180437271892	0.972711	0.836838	0.139685	42.7056
180437289947	0.810207	0.708241	0.125852	31.9505
180437294118	0.939539	0.810296	0.13756	41.7059
180437295658	0.846999	0.756632	0.106691	37.1773
180437306100	0.846999	0.756632	0.106691	37.1773
180437306154	0.746129	0.668654	0.103835	29.3842

- **Survival probability at future time**: the chance that a given current customer will still be a customer 3 months from the time that the model was trained (date specified in the scoring data).
- **Event prob. Before or at Future Time**: The chance of having the event within the forecast period (date specified in the scoring data).

Note: Future time is set in the **Default** and **Number of Forecast Intervals** property. The defaults depends on the time unit being modeled: Day=30, Week=4, Month=3, Quarter=4, Semi-Year=2, Year=1.

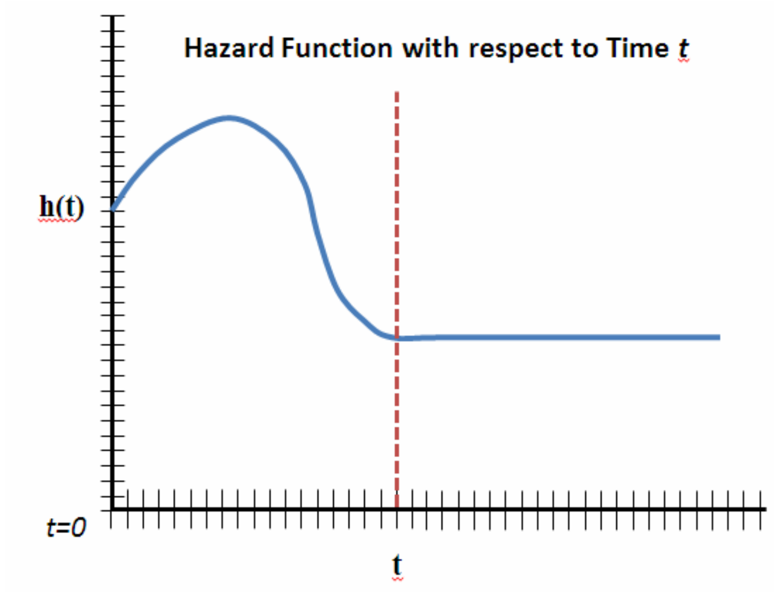
Thank You!

Lorne Rothman, PhD, P.Stat.
Principal Statistician
Lorne.Rothman@sas.com

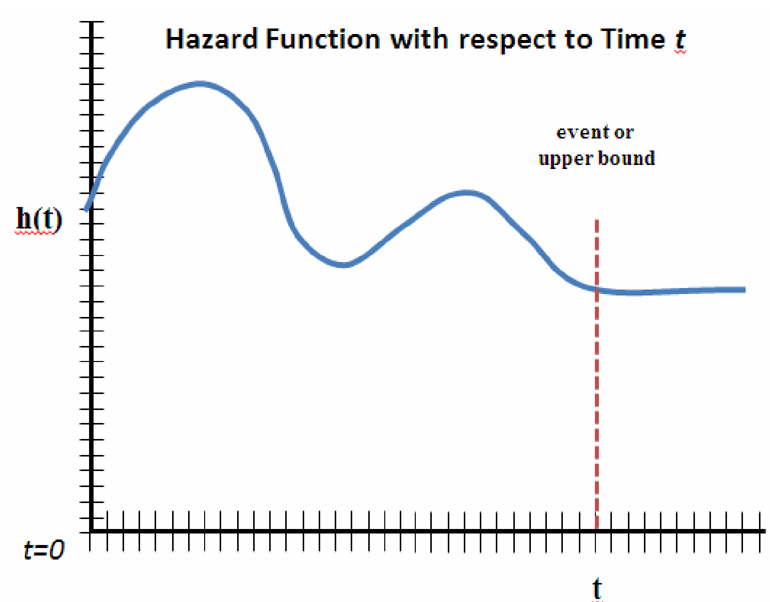
THE POWER TO KNOW®

Mean Residual Lifetime (Time remaining until an event will occur)

Constant Hazard Extrapolation



Restricted Mean Residual Life



Constant Hazard Extrapolation: from time *t* onward, the hazard function is constant from the final value.

Restricted Mean Residual Life: the hazard function continues trending until an event occurs, or until the maximum value for MRL is reached, whichever comes first. Once the maximum value for MRL is reached, the hazard is held constant from that point forward.