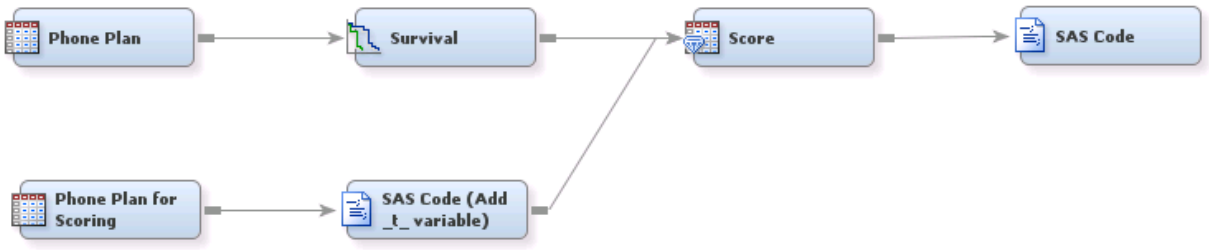
Survival Analysis with Time-Varying Covariates Using SAS® Enterprise Miner™



Data:

This examples uses SAMPSIO.PHONEPLAN_TVC as both the training data and the score data. This data set contains information about business customers for a telephone and internet service provider. In these data, the event of interest is an upgrade to a high-end plan, with customer churn being a competing risk. Thus, the variable event is the target, which is coded as 0 for customers who haven't upgraded or churned, 1 for customers who churned before upgrading, and 2 for customers who have upgraded to the high-end plan. Inputs (covariates) are region, paywithcc, variables with the prefix 'n_' (which represent the number of products that a customer has in a category), and variables with the prefix 'cum_' (which represent the cumulative number of times a customer added a product in a category). The 'n_' and 'cum_' variables can vary over time for a customer, which is the reason for multiple observations per customer ID (acctno). This data set contains three Time ID variables: initdate, which contains the customer's start date; eventdate, which contains the date of a customer event (either upgrade or churn); and changedate, which contains the date that any of the covariates changes value for a customer. Other than the time-varying covariates and the changedate variable, the variables in the data set stay constant across customer ID.

Goal:

The goal is to determine when a customer is most likely to upgrade. Estimating when customers upgrade (as opposed to whether they upgrade) enables you to view upgrade trends in time and make decisions accordingly. This example examines the use of the Survival node with time-varying covariates incorporated to potentially improve the model.

Flow:

The Phone Plan node (an Input Data node that contains the training data) is connected to a Survival node.

Beginning in SAS Enterprise Miner 12.3, the Survival node supports time-varying covariates when your data are suitably formatted. The default properties of the Survival node do not accommodate time-varying covariates, so you need to change the following properties in order to make use of this feature.

Change the **Data Format** property (new in SAS Enterprise Miner 12.3) from the default (**Standard)** to either **Change-Time** or **Fully Expanded**. In previous versions the Survival node accepted only **Standard** for the data format; **Standard** does not allow for the use of time-varying covariates. The **Change-Time** and **Fully Expanded** data formats both support time-varying covariates, with multiple observations allowed per ID. For this example, the **Data Format** property should be set to **Change-Time**, because this particular data set has a third Time ID variable, which indicates the date that the value of any of the covariates changed. When more than two Time ID variables are present, you need to click the ellipses next to the **Time ID Variables** property to manually indicate which Time ID variables correspond to the start time, event time, and change time. For this particular example, the appropriate Time ID variables have already been identified, with InitDate as the Start Time Variable, EventDate as the End Time Variable, and ChangeDate as the Change-Time Variable.

The default value of the **Covariate x Time Interaction** property is **Do not include**. You can change this value to either **Include all** or **Include selected**. The **Include all** option includes a covariate-by-time interaction term for each input variable. This example uses the **Include selected** option so that interactions between time and region or paywithcc are not included.

After you have chosen **Include selected** for the **Covariate x Time Interaction** property, you need to specify the covariates that you want to use in interaction terms with time. Click the ellipses next to the property **Covariates for Interactions**. In this example, the six time-varying covariates are already included for covariate-by-time interactions. If the variables are not already selected, you would do the following when the **Terms** dialog box appears:

1. Click a variable in the box titled **Variables**.
2. Click the right arrow button to move the variable into the box titled **Term**.
3. Click **Save** to indicate that you want this variable to be used in a covariate-by-time interaction.
4. After you have selected all the variables you want, click **OK**.

In this example, the remaining properties have been left as default. If you want more control over the time range of the data that you are considering, you can set the **Left-Truncated Data** property to **Yes**, and then click the ellipses next to **Training Time Range** to set the start and end dates for training.

After you have run the flow, you can view the results for the Survival node to see various plots and tables that summarize the created survival model. Expand the Output window to see the ODS output from the various SAS procedures that were run, including the Type 3 Analysis of Effects table that is produced by the DMREG procedure. Here you can verify that the correct interaction terms between time and the selected covariates (_t_ x cum_dial_add and so on) were included in the regression model.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
_t_	2	0.1935	0.9078
_csb1	2	49.8561	<.0001
_csb2	2	15.9441	0.0003
_csb3	2	6.4603	0.0396
_csb4	2	2.8682	0.2383
_csb5	2	1.6027	0.4487
paywithcc	2	16.9875	0.0002
region	28	43.5949	0.0305
cum_dial_add	2	54.5587	<.0001
cum_dsl_add	2	21.0697	<.0001
cum_isdn_add	2	0.4587	0.7951
n_dial	2	7.2741	0.0263
n_dsl	2	7.0783	0.0290
n_isdn	2	7.6273	0.0221
_t_*cum_dial_add	2	28.7120	<.0001
_t_*cum_dsl_add	2	4.1598	0.1249
_t_*cum_isdn_add	2	1.3526	0.5085
_t_*n_dial	2	2.5064	0.2856
_t_*n_dsl	2	4.1543	0.1253
_t_*n_isdn	2	4.8937	0.0866

You can now make predictions on new data by applying the score code that the Survival node generates. In order to apply score code from the Survival node, your score data set must have, in addition to the inputs and start date variable that were present in the training data, a variable _t_ that represents the number of time intervals between the start date and the current date. In this example, _t_ is in terms of months because months was the time interval used when training.

Note that only one set of values for the covariates can be used in scoring, so there should only be one observation per ID—even if your training data are in change-time or fully expanded format and contains multiple observations per ID.

For the purpose of illustration, the score data used in this example are created from the training data. The Phone Plan for Scoring node (an Input Data node) represents the same data set, SAMPSIO.PHONEPLAN_TVC, as the Phone Plan node (also an Input Data node) represents. The Phone Plan for Scoring node is then connected to the SAS Code (Add _t_ variable) node, which enables you to incorporate custom SAS code into the flow. You can click the ellipses next to the **Code Editor** property to view the following SAS code:

```
data &EM_EXPORT_SCORE;
      set &EM_IMPORT_DATA;
      by acctno;
      if last.acctno;
      format _currentdate MMDDYY10.0;
      _currentdate = input("31DEC2000", anydtdte10.);
      _t_ = intck("MONTH",initdate, _currentdate);
      drop Target _currentdate;
run;
```

Note that 31DEC2000 is used as the current date here for calculating _t_. The macro variable EM_EXPORT_SCORE, which is used in the DATA statement, resolves to the name of the data set with Role=Score exported from this node. Only the last record for each ID (acctno) is kept in the score data.

Next in the flow, both the SAS Code (Add _t_ variable) node and the Survival node are connected to a Score node. The Score node uses the previously modeled hazard and survival functions to score the score data, enabling you to inspect which customers in the score data are most likely or least likely to upgrade during the time that spans the desired number of forecast intervals. You can specify the number of forecast intervals under the property **Number of Forecast Intervals** (on the Survival node) if you choose **No** for the property **Default Forecast Intervals**.

In this example, the second SAS Code node is used to further inspect which customers have the lowest hazard rate for upgrade at the forecast interval. This example uses SAS code to print the 100 customers from the scoring data who have the lowest hazard rate (those least likely to upgrade in the forecast time interval) in the results of the SAS Code node. You can edit this code to display whatever statistic you want.