Survival Analysis Workflow

assessing the impact of Macro-Economic Shocks on Credit Portfolios and predicting the Time of Default

(By Anupam Saha, Software Specialist – Analytics, SAS R&D (India) Pvt. Ltd., and Naeem Siddiqi, Senior Solutions Specialist, SAS Institute Inc.)

Introduction

The recent global credit crisis has reiterated the need for bankers to focus on macro-economic factors (such as inflation, house price index, and so on) that severely impact the default behavior of credit portfolios. At the micro level, the response to these economic changes varies with individual capacity and tolerance. For example, although accounts within the same credit score or rating group show homogeneous behavior in default performance, the individual borrowers respond to macroeconomic shocks in a different manner than others in the same segment. This results in scattered default distributions within the predefined performance window. Because defaults are scattered across time, assessing the macro economic impact on a credit portfolio is a challenging task.

In credit scoring, the traditional probability-of-default (PD) model is used to estimate the individual likelihood to default. This involves generating a sample of new or existing accounts at one point in time and monitoring the performance of those accounts across a specified performance window to determine their default or non-default status. It is generally assumed that the probability of default is constant through the entire performance window, and the outcome of the model is a probability of default value (PD) that is valid for the full window. Because the defaults are scattered across time in the performance window, PD varies with respect to time within the same window. Along with the PD, the time of default can be a key piece of information for bankers, allowing them to create better account management strategies. The traditional methodology makes it difficult to calculate PD by different time points and the time of default within the specified performance window.

In this paper, we propose a survival analysis workflow for two purposes:

- to observe the impact of macro-economic shocks on default behavior of credit exposures
- to predict the time of default

We intend to use traditional PD models in conjunction with a survival analysis technique (explained later in this paper) for overcoming the challenges associated with traditional PD models. In this paper, we explain a survival model called “Cox Proportional Hazard Model” that can be used to establish a relationship between macro-economic variables and the hazard

function (that is, the rate of change of probability of default at time *t*) at different time points in a observation window. The survival distribution of default can be estimated by this model. The survival function is used to determine the time of default. Further analysis is then performed to obtain the expected time-dependent default count distribution per segment. Finally, variances of macro-economic factors are used to stress-test the credit portfolio.

This paper also gives insights into the data requirements, SAS procedures for survival analysis, and the techniques to validate and monitor the workflow for accuracy and consistency.

Benefits of Using Survival Analysis in Credit Scoring

1. Modelers in banks are aware of the fact that macro-economic variables affect probabilities of default that are generated using models. These macro-economic variables can change at any time during the observation period. However, such variables, and in particular, the leading indicators, have not been included in the traditional logistic regression models used in banking. The impact of macro-economic variables on the PD value can be gauged by considering those variables as time-dependent covariates in the survival analysis model.

The survival function of default as well as the time of default can be derived from that survival model. Time of default here refers to the survival time of a particular account. Survival time is the time from when the account was opened to the date of default. If the account does not default by the end of the observation period, that observation can be considered as a censored event.

For analyzing the survival data, the hazard function gives the rate of change of probability of default at time *t*.

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

The probability of survival at time t can be given in terms of the hazard function:

$$S(t) = \exp\{-\int_0^t h(u)du\}$$

This is the probability of survival at time t. In credit scoring scenario, this is the probability that an account has not defaulted by time t after the account was opened (that is, 1-PD at time t). If the interval censoring is considered, then this indicates the probability that an account has not defaulted by time t after the start of the observation period.

There are several ways to generate the hazard function. We will use the Cox Proportional Hazard (Cox PH) model because it allows the inclusion of macro-economic variables as time-dependent covariates in the model. By Cox PH model,

$h_i(t) = \lambda_0(t) \exp(\beta_1 X_{i1}(t) + \beta_2 X_{i2}(t) + \ldots \beta_k X_{ik}(t))$

where,

i is a subscript for observation i=1(1)n.

$h_i(t)$ = Hazard of the *i*th individual at time *t.*

$\lambda_0(t)$ = Baseline hazard function.

$X_{i1}(t) \ldots \ldots X_{ik}(t)$ are the k time-dependent covariates of the *i*th individual at time *t*.

$\beta_1 \ldots\ldots \beta_k$ are the regression parameters associated with the covariates, estimated by the Partial Likelihood method in Cox model.

For applying Cox PH model in the credit scoring environment, macro-economic variables should be captured at account default date, prior to default date (any lag period can be considered), at account open date, and at regular time intervals during the observation period. The change in macro-economic values between the account open date and the default date or prior to the default date (considering any lag period, for example, 6 months, 3 months, and so on) can also be considered as a time-dependent covariate in the model. The time of default and the default indicators in the observation window should also be captured. The survival probabilities for each of the individuals at failure time points can be estimated by using the baseline survival function and the parameter estimates $\tilde{\beta}$ of β by maximum likelihood estimation and the macro-economic values at those time points. If a user changes the macro-economic values, the user can get the different survival probabilities of those accounts at default time points. In this way, the Cox PH model can help to see the impact of those variables on the probability of default.

2. Survival analysis can be used to determine the probability of default across time in an observation window. It is also possible to derive the survival function of default for future time points. This represents an advantage over the traditional PD-estimation methods, which limit the prediction to a predefined observation window. By generating PDs across time, a risk manager can adjust account treatment strategies according to an account's risk level at that point in time, rather than rely on a single, long-term PD forecast that might not represent the account's current risk levels.

Figure 1 shows an example of the probability of default across time in an observation window, as generated by using survival analysis. Figure 2 shows an example of the survival function of default across time in that observation window. Both the graphs are based on the same sample data. From Figure 1 and Figure 2, it is evident that as time passes, the probability of default increases, and the corresponding survival function decreases. At the early stages of the performance window, the change in the macro-economic values might not be significant to

cause default events. However, as time passes, those economic conditions might vary considerably, which might force the accounts to default. Thus, it indicates that at an early stage of the performance window, the probability of default is low, and the survival probability is high. However, as time passes, due to changes in economic conditions, the probability of default might increase, and the survival probability might decrease.

Figure 1:

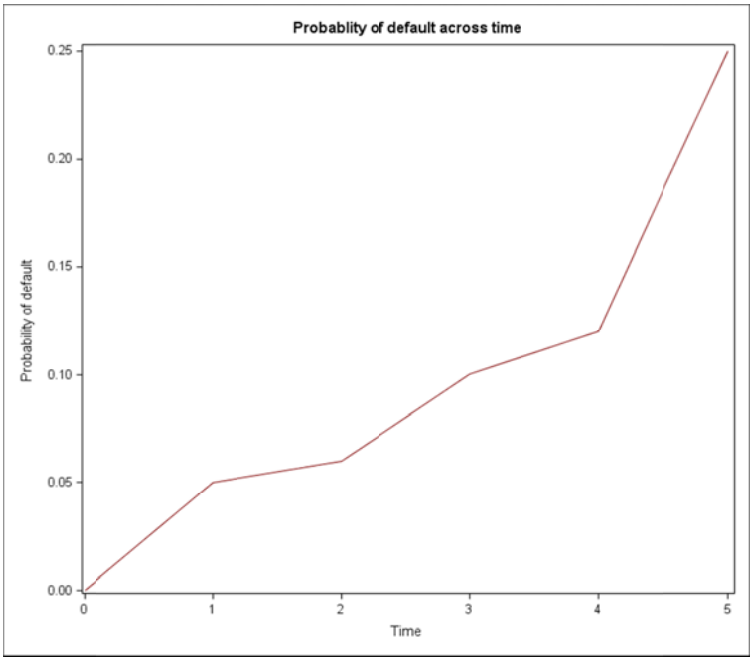
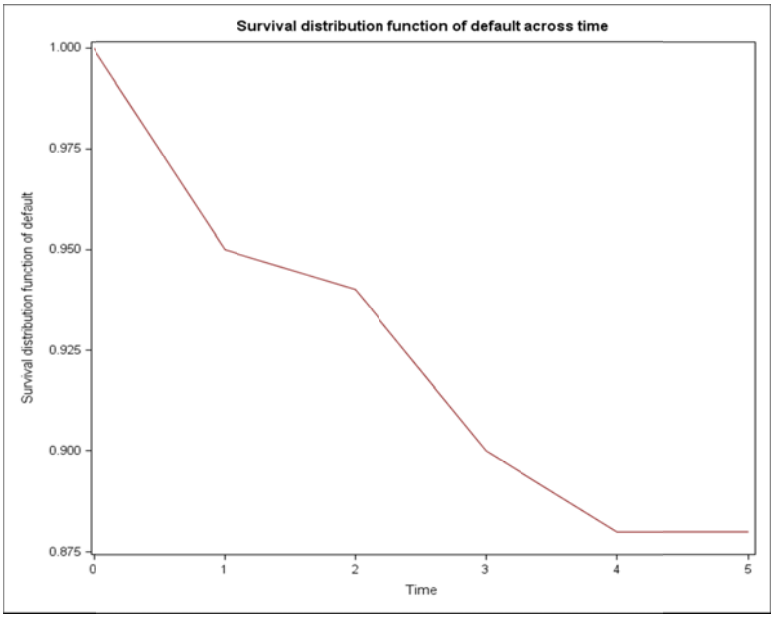


Figure 2:



3. Bankers generally build several segmented models for any given portfolio, to generate better forecasts for specific populations in that portfolio. As an extension of the above, a user can build separate survival models for the various segments in the population to generate further insight into how the PDs of the various sub-populations vary across time. Figure 3 illustrates how the PDs of five separate subpopulations within a portfolio vary across time. Again, this gives the risk manager deeper insights into account behavior than when compared to one simple PD prediction for the whole performance window. Figure 4 shows how the different survival functions of default for the various segments can be compared.

Figure 3 and Figure 4 are based on the same sample data. Let us assume that the subpopulations are based on the scores of the accounts. Segment A has lowest-score accounts—that is, the highest-risk-profile accounts. This means that for the accounts in that segment, the probability of default is higher compared to the accounts that belong to higher-score segments. Segment E has the highest-score accounts—that is, lowest-risk-profile accounts.

Figure 3 shows the trend of PD values for segments across time. In the figure, it can be observed that, for segment A, the distribution of PD values across time is high in all the segments. However, for segment E, the distribution of PD values across time is low. In the Figure 4, it can be observed that for segment A, survival function of default is low in all the segments. However, for segment E, the survival function of default is high.

Figure 3:

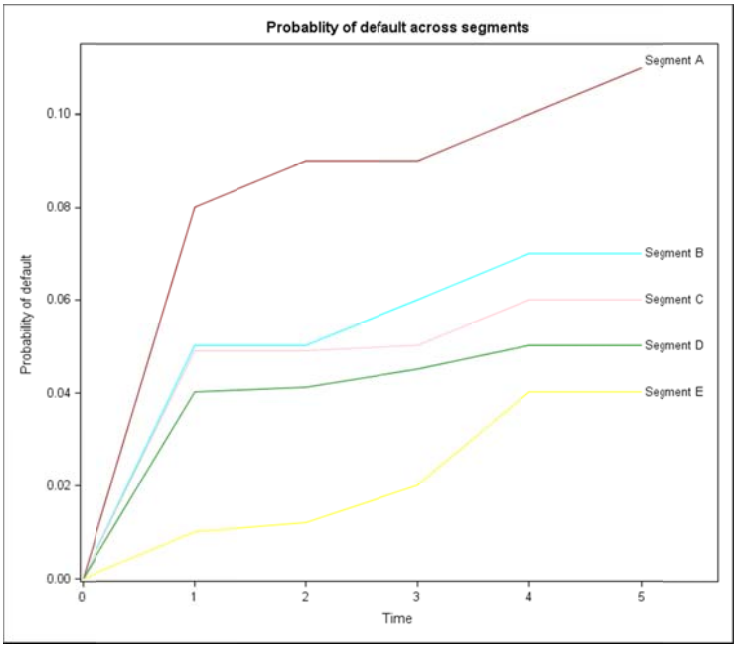
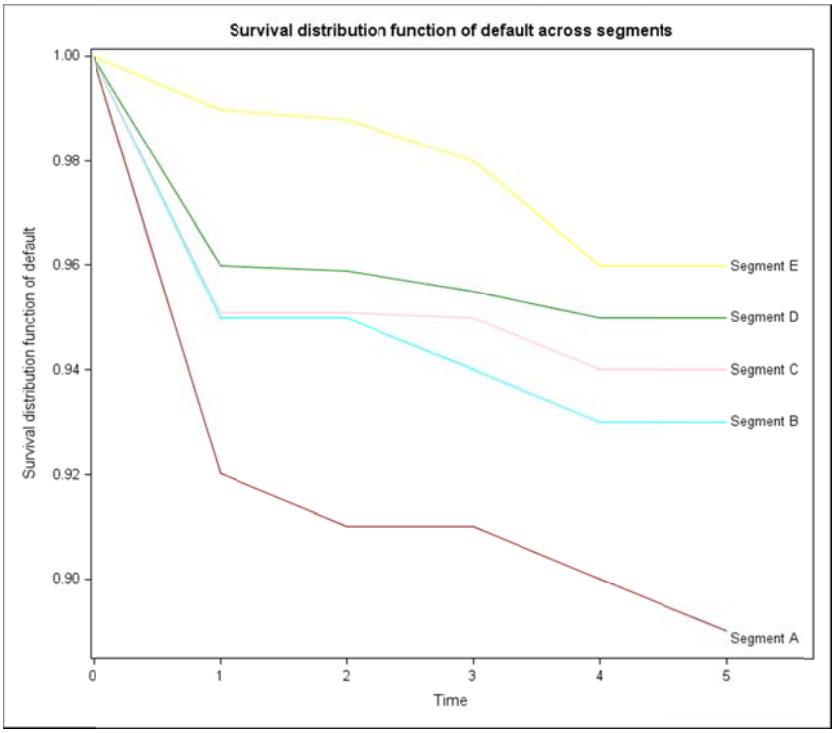


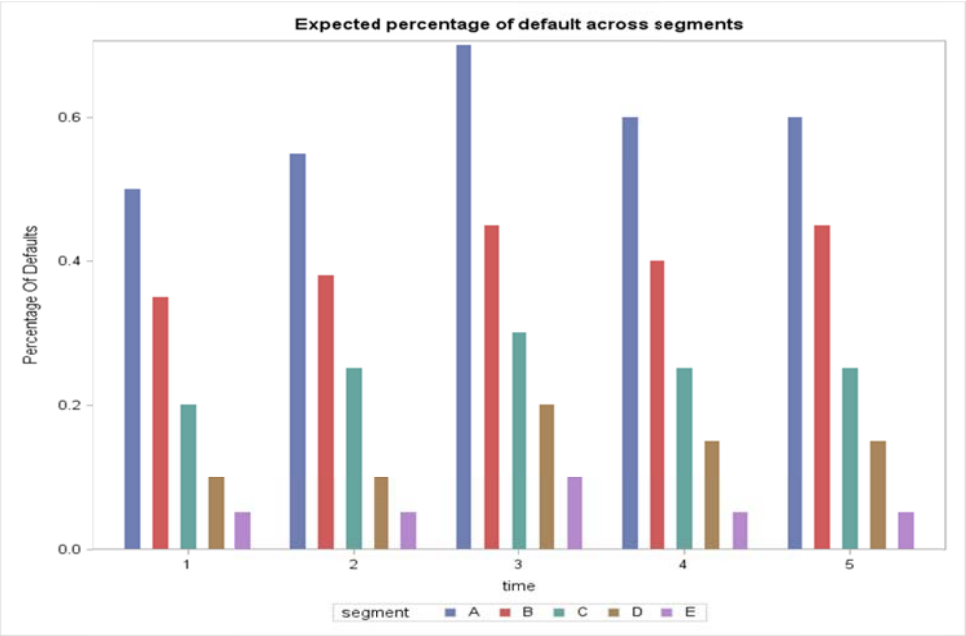
Figure 4:



4. Using the survival functions generated by the survival models, it is possible to derive the time of default for each account, within all the segments. We can then use these predictions to derive the expected number of defaults for each segment in consecutive future time points. Figure 5 shows an example of such an analysis for five pools across five time points. Those who monitor the observation of portfolios will recognize this chart as being similar to vintage or cohort analysis used widely in the industry.

Figure 5 is based on sample data. Let us assume that segments are based on the scores of the accounts. Segment A has lowest-score accounts—that is, the highest-risk-profile accounts. This means that for the accounts in that segment, the probability of default is higher compared to the accounts that belong to higher-score segments. Segment E has the highest-score accounts—that is, the lowest-risk-profile accounts. From Figure 5, it is evident that, for a particular time point, the percentage of default is relatively decreasing from segment A to segment E.

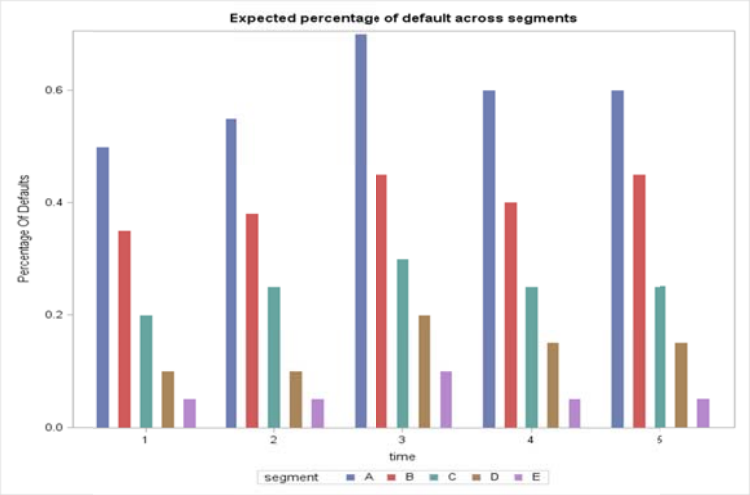
Figure 5:



5. Because macro-economic variables are being used as time-dependent covariates here, the impact of changes in macro-economic conditions on the predicted default rates and the time of default can be recalculated at future time points. Figure 6 shows an example of how the changes in default behavior, which result due to changes in five selected macro-economic variables, can be monitored.

Let us assume that five different survival models have been built for five different segments. Figure 6 shows the expected percentage of default across future time points for five segments based on certain macro-economic values.

Figure 6:



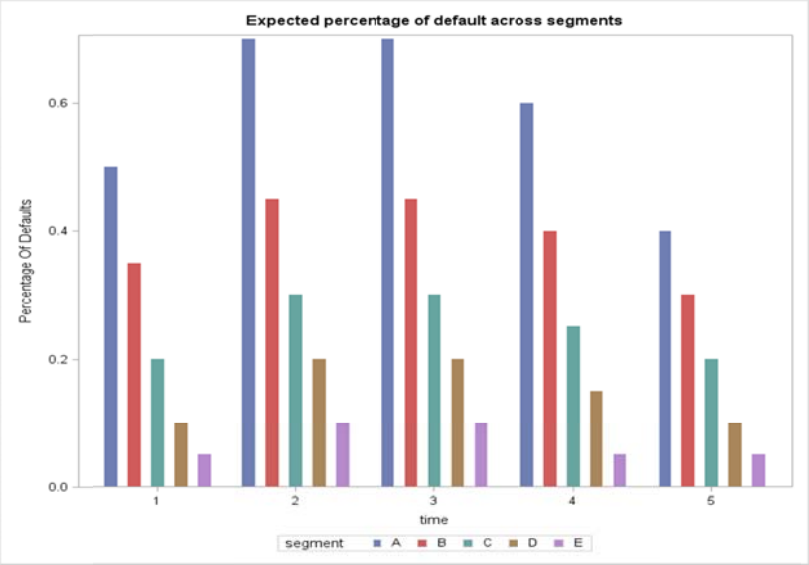
Now, a user might be interested to see what impact the changes in macro-economic values has on this sort of report. Now, let us assume that the user has the capability to perform iteration on the macro-economic values for future time points from any time series model. Now, if the updated macro-economic values are significantly different from the previous values, the user might be interested to see its impact on the default behavior (Table 1).

Table 1:

Time	Updated Macro-economic Values
1	<value>
2	<value>
3	<value>
4	<value>
5	<value>

The user can then recalculate the survival function of default for each account within all segments for those future time points, for updated macro-economic values (Table 1). Then, from the survival functions, it is possible to predict which accounts will default and at what time. It is also possible to generate same kind of report (Figure 6) that defines the expected percentage of default distribution across time for the five segments. This report (Figure 7) appears different from the earlier one (Figure 6). This sort of analysis can help bankers to see the variation in percentage of default across time (within the sub-populations), with any changes in the macro-economic values.

Figure 7:



Process Flow of Survival Analysis in Credit Scoring

Data requirements

There are no limitations on the number or the nature of macro-economic variables that can be used for this exercise. The exact list will depend on the availability of data in each market and whether there is a business buy-in for using the data. In general, leading indicators such as the following are considered. Although the raw value of each indicator can be used, changes in the indicator over a 3 to 6 months period can be a better predictor of the future, as it incorporates the direction in which that indicator is moving.

- inflation rate
- interest rates and bond yields
- unemployment rate and employment growth rate
- housing price index
- consumer price index
- average personal income growth rate, growth rate
- stock market index
- housing starts and similar statistics such as building permits issued
- GDP
- money supply

- manufacturing output
- corporate profits and inventory levels

In addition to the macro-economic variables above, demographic (postal code, residential status, age, and so on) and behavioral variables (account-related variables) can also be considered. If the segmentation was based on these variables, then for the Survival model, only the macro-economic variables should be considered.

The default time of each defaulted account should be captured in the modeling data. If the account does not default in the outcome period, then the entire outcome period should be considered as survival time for that account. The default indicator should also be considered in the modeling data.

Macro-economic variables should be captured at account default date, prior to default date (any lag period can be considered), at account open date, and at regular time intervals during the observation period. The change in macro-economic values between the account open date and the default date or prior to the default date (considering any lag period, for example, 6 months, 3 months, and so on) can also be considered as time-dependent covariate in the model. For the survival model, interval censoring can be considered. If the defaults are captured for a specific time period, then the censoring technique can be considered as interval censoring.

Table 2 shows the structure of a sample modeling analytical base table (ABT) for using survival analysis in a credit-scoring environment. Consider a sample data consisting of the following:

1. ID (account identification number)
2. Time (time of default of the account in a predefined observation period)
3. Default (censoring status where 1 = default and 0 = censored)
4. Consider that a time-dependent covariate is captured at default dates cov1 – cov6 (a covariate value at six default times in the observation period). cov_i denotes a covariate value at the i^{th} time and i=1(1)6.

For instance, the first account defaults at time3. Therefore, up to time 3, the covariate values have been captured. The second account defaults at time 4. Therefore, up to time 4, the covariate values have been captured. The third account does not default in the outcome period. Therefore, up to time 6, its covariate values can be captured. The third account will be considered a censored event.

Table2:

Account id	time	default	cov1	cov2	cov3	cov4	cov5	cov6
1	3	1	<value>	<value>	<value>			
2	4	1	<value>	<value>	<value>	<value>		
3	6	0	<value>	<value>	<value>	<value>	<value>	<value>

Study the impact of macro-economic variables on default behavior (using Cox model by PROC PHREG)

Many types of models have been used for survival analysis. In this paper, the Cox proportional hazard model has been mentioned. In SAS, the PHREG procedure performs regression analysis of the survival data based on the Cox proportional hazard model.

1. To handle time-dependent covariate in PHREG procedure, the counting process style of input is used. The advantage of this format that user can easily get the residual and influence statistics. In this format, for each record, multiple records are crated, one record for each distinct pattern of the time-dependent measurements. Each record contains a *T1* value and a *T2* value, representing the time interval *T1*and *T2* during which the values of the explanatory variables remain unchanged. Each record also contains the censoring status at *T2*.

Following is an example of a modeling data set that is created in counting process style of input format. PHREG can be run on this type of data set.

Table 3:

Account id	Time	default	T1	T2	status	covariate
1	3	1	0	1	0	<value>
1	3	1	1	2	0	<value>
1	3	1	2	3	1	<value>

2. Consider that we have a scoring data set, and we want to find the predicted survival probability of default across time for the accounts in that data set.

   Table 4:

Account id	time	covariate
1	1	<value>
1	2	<value>
1	3	<value>
1	4	<value>

PROC PHREG can be run on the modeling data set (Table 3) and can generate survival probability for the accounts in the scoring data set (Table 4). Following is a sample SAS code for generating survival probabilities.

```
Data Zerocov;
    covariate= 0;
Run;

proc sort data = scoredata;
   by time;
run;

proc phreg data= inputdata;
   ods output ParameterEstimates = para_est FitStatistics = fit_stat ;
   model (T1,T2)*Status(0)=covariate;
   baseline covariates=Zerocov  out=Basesurv Survival=Basesurvival;
   id ID Time Default;
run;

proc sql;
      select Estimate into:Estimate from para_est ;
quit;

data Scoredata(drop=lastS);

    retain lastS 1;

    merge Scoredata Basesurv(rename=(t2=time)drop= covariate);

    by time;

    if missing(Basesurvival) then Basesurvival=lastS;

    else lastS= Basesurvival;

    survival= Basesurvival ** exp(&Estimate*covariate);

 run;
```

This way, it is possible to generate survival probability for each account across time. The following table (Table 5) shows the structure of the data set that contains the survival probabilities for each account across time:

Table 5:

Account id	time	survival probability
1	1	<value>
1	2	<value>
1	3	<value>
1	4	<value>
1	5	<value>
1	6	<value>

A user can change the covariate values for future time points and can get different survival probability for each account for those time points. In this way, it is possible to see the impact of macro-economic variables on default behavior.

3. One advantage of using the counting process formulation is that you can easily obtain various residuals and influence statistics. To do this, you can use the following code:

```
proc phreg data= inputdata;

  model (T1,T2)*Status(0)=covariate;
  output out= dev_est    xbeta = est resmart=  marting ;

run;
```

In the dev_est data set, the variable "marting" contains martingale residuals that can help the user proceed on residual analysis of the model.

Different Survival Methods

The counting process formulation enables PROC PHREG to fit a superset of the Cox model, which is known as the multiplicative hazards model.

A. The Multiplicative Hazards Model

Consider a set of *n* accounts so that the counting process $N_i \equiv \{N_i(t), t \geq 0\}$ for the *i*th account indicates the number of observed events that occur over time *t*. The sample paths of the counting process N_i are step functions that have jumps of size $+1$, with $N_i(0) = 0$. Let $\boldsymbol{\beta}$ denote

the vector of unknown regression coefficients. The multiplicative hazards function $\Lambda(t, \mathbf{Z}_i(t))$ for N_i is denoted by

$$Y_i(t)d\Lambda(t, \mathbf{Z}_i(t)) = Y_i(t)\exp(\boldsymbol{\beta}'\mathbf{Z}_i(t))d\Lambda_0(t)$$

where,

$Y_i(t)$ indicates whether the ith account defaults at time t (specifically, $Y_i(t) = 1$ if defaults, otherwise $Y_i(t) = 0$).

$\mathbf{Z}_i(t)$ is the vector of explanatory variables for the ith account at time t.

$\Lambda_0(t)$ is an unspecified baseline hazard function.

B. Piece-wise Constant Baseline Hazard Model

In **Piece-wise Constant Baseline Hazard Model**, the period is divided into certain intervals. Then, considering a baseline hazard value for each of the intervals, the baseline cumulative hazard function is generated.

Considering single failure time variable case, let $\{(t_i, \mathbf{x}_i, \delta_i), i = 1, 2, \ldots, n\}$ be the observed data. Let $a_0 = 0 < a_1 < \ldots < a_{J-1} < a_J = \infty$ be a partition of the time axis.

When hazards are in original scale, the hazard function for account i is denoted by

$$h(t|\mathbf{x}_i; \boldsymbol{\theta}) = h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}_i)$$

where,

$$h_0(t) = \lambda_j a_{j-1} \leq t < a_j (j = 1, \ldots, J)$$

The baseline cumulative hazard function is denoted by

$$H_0(t) = \sum_{j=1}^{J} \lambda_j \Delta_j(t)$$

where,

$$\Delta_j(t) = \begin{cases} 0 & t < a_{j-1} \\ t - a_{j-1} & a_{j-1} \le t < a_j \\ a_j - a_{j-1} & t \ge a_j \end{cases}$$

For using counting process style of input, let $\{((s_j, t_i], \mathbf{x}_i, \delta_i), i = 1, 2, \ldots, n\}$ be the observed data. Let $a_0 = 0 < a_1 < \ldots < a_k$ be a partition of the time axis, where $a_k > t_i$ for all $i = 1, 2, \ldots, n$.

Then, $\Delta_j(t_i)$ should be replaced with

$$\Delta_j((s_i, t_i]) = \begin{cases} 0 & t_i < a_{j-1} \vee s_i > a_j \\ t_i - \max(s_i, a_{j-1}) & a_{j-1} \le t_i < a_j \\ a_j - \max(s_i, a_{j-1}) & t_i \ge a_j \end{cases}$$

the formulation for the single failure time variable applies.

The BAYES statement invokes a Bayesian analysis of the Cox model or the piecewise constant baseline hazard model (also known as the piecewise exponential model). The following code can be used.

```
proc phreg data= inputdata;
  ods output ParameterEstimates = para_est ;
  model (T1,T2)*Status(0)=covariate;
  bayes seed=1 piecewise=hazard (n =interval);

run;
```

The code specifies the number of intervals with constant baseline hazard rates. PROC PHREG partitions the time axis into the given number of intervals with approximately equal number of events in each interval.

Algorithm for Predicting the Time of Default

After a model has been created, it can be used to predict defaults. By using the Cox PH model, the survival function of default for an account can be generated.

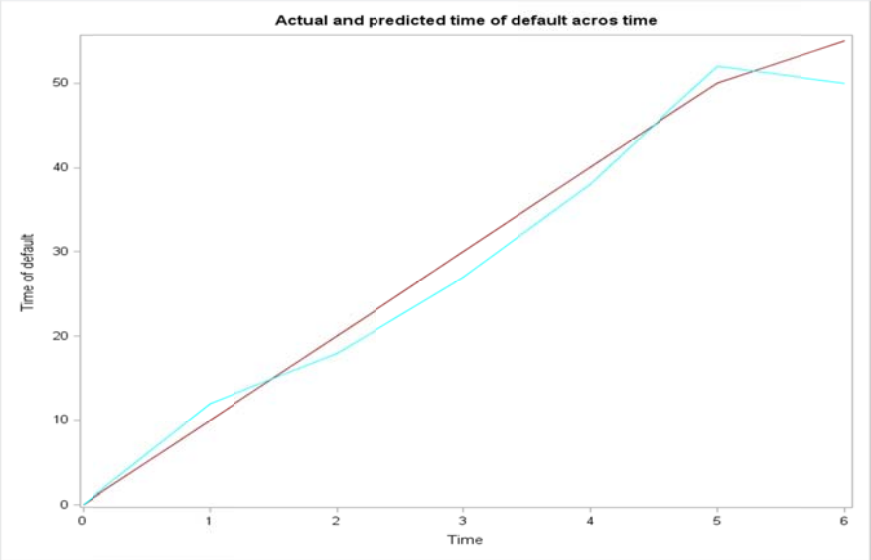
The time point at which the survival probability is zero can be considered as the time of default. Theoretically, the survival probability can be zero. However, practically, it might not be possible to get a zero probability of survival. To compensate this, a user can predefine a threshold value. Now, if the predicted survival probability is less than or equal to the threshold value, then that time might be considered the time of default. The threshold value can be based on business judgment. It can vary industry to industry.

Generally, the survival time of an account is measured from the date when the account was opened. If the account never defaults, then the survival time can be measured from when the account was opened to the end of the observation period. If the interval censoring is considered for the survival model, then the survival time can be measured from when the observation period starts to the date of default. If the account does not default then survival time would be the entire observation period.

Validation Technique for Predicting Time of Default

1. A standard practice in any predictive modeling project is to validate the outcome of the modeling process. For the validation process, a validation data set from some point in the past can be used. The survival function is applied on this data set to generate scores (predicted time to default) based on a specified threshold value. The actual time of default of an account is known for the development data set. The time of default of an account can be predicted. In an ideal scenario, the predicted time of default for an account should be close to the actual time of default.

Figure 8:



This graph is based on sample data. The actual number of defaults and the predicted number of defaults of accounts are plotted across time. In an ideal case, both the plots should be close to each other.

In addition to a visual confirmation, some statistics can be generated to measure the difference between the actual and the predicted observations.

A measure can be calculated as:

$$\sum_{i=1}^{n} (Actual\ time\ of\ default\ of\ the\ ith\ account - Predicted\ time\ of\ default\ of\ ith\ account)$$

Where, n is the total number of accounts in the data.

The absolute difference of the actual and the predicted time of default can be considered. The discriminatory power of a time of default model gets better as the above measure gets closer to zero, indicating a closer match between the predicted and actual values.

Some other statistics (for example, Brier score) can also be generated in this scenario. A non-parametric test, such as median test, can also be performed to compare the median actual time of default with the median predicted time of default.

Acknowledgements

I would like to thank Mrs. Billie Anderson, Mr. Ying So, Mr. David Park, Mr. Rajesh Khanwelkar, and the SAS Credit Scoring for Banking team.

References

Paul D. Allison. Survival Analysis using SAS A Practical Guide

SAS/STAT 9.2 User's Guide

Elisa T Lee, John Wenyu Wang. Statistical methods for survival data analysis

John P Klein, Melvin L. M. Moeschberger. Survival analysis: techniques for censored and truncated data

Shenyang Guo. Survival Analysis

J D Kalbfleisch, Ross L. prentice. The statistical analysis of failure time data

Miller R.G. 1981 "Survival Analysis." New York Wiley

Peto R. 1973. "Experimental Survival curves for interval-Censored Data." Applied Statistics 22:86-93

Alan B. Cantor. SAS Survival Analysis Techniques

Jeraid F. Lawless. Statistical models and methods for lifetime data

David W.Hosmer, stanleylemeshow. Applied Survival Analysis: Regression Modeling of Time to Event Data (Wiley Series in Probability and Statistics)