Preparing for Basel II

Common Problems, Practical Solutions

Part 4: Time to Default

by Jeffrey S. Morrison

Previous articles in this series have focused on the problems of missing data, model-building strategies, and special challenges in model validation—topics often associated with modeling retail portfolios. This article moves a bit beyond Basel, offering additional tools to further enhance risk and account management strategies. Building a quantitatively based framework—beyond any Basel II requirements—is just plain old good risk management.

Banks opting for advanced status under Basel II are expected to have, among other things, models for PD (probability of default) and LGD (loss given default). Some of these models will be provided by outside vendors, while others will be more custom driven. Previous articles in *The RMA Journal* have discussed how to build both PD and LGD models[1]. A PD model will let you know the probability of a loan defaulting sometime in the next year. Looking at the modeling results, you might think that loans with an estimated PD of .95 will default more quickly than ones with a PD of .45 or lower. *The PD model, however, was never designed to address this issue.* Such a model makes no attempt at describing the exact timing of default—either within the year or further down the road. Wouldn't it be nice to know—at booking or at any time during the life of the loan—when the default might occur? Obviously, if the loan were expected to default in the next few months it would not be worth your while to book it. But what if you knew that the probability of default for a particular loan might stay low for the first three years but increase dramatically thereafter? Would it be worthwhile then to know? Under what conditions might you book it?

Modeling time to default requires nothing but a slight variation to a statistical technique already discussed in previous articles. Although building a good

 Jeff Morrison is vice president, Credit Metrics—PRISM Team, at SunTrust Banks, Inc., Atlanta, Georgia.

time-to-default model can prove challenging, your success with PD and LGD models, coupled with a comprehensive and well-understood database, should give you a great head start. If successful, your organization will have a mechanism to help predict when an account will default—a prediction made early enough that it might prevent the loss from occurring, mitigate its impact if it does occur, or at the very least provide insight into the profitability, pricing, or term structure of the loan. For illustration purposes and to keep things simple, we will use a hypothetical dataset with generically labeled variables called X1, X2, and X3. In reality, these might be predictors such as LTV, debt to income, fixed-versus-variable interest rate, term of loan, and so forth.

Let's start with a story. Three economists went deer hunting. The first economist saw a deer and shot to the left of the animal, missing it by about two feet. The second economist shot to the right of the deer, missing it by about the same margin. Immediately afterwards and with unbridled enthusiasm, the third economist stood up and shouted, "We got him, we got him!!" The idea here is that economists who focus on predicting when an event will occur (such as a downturn in the economy) realize the difficulty of their task. They are happy simply when they come close. Whereas previous articles in this series looked at methods of predicting *if* an account will default sometime within a 12-month window of time, we now consider a more formidable task—*the timing of default*.

Censored Data and Survival Analysis

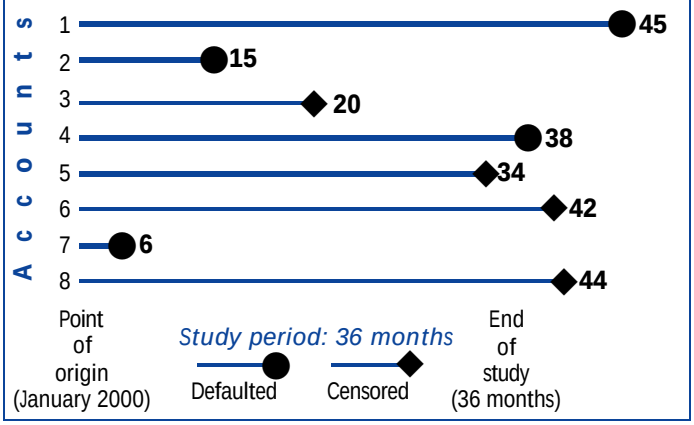
The task is more formidable for two reasons. First, obtaining a certain level of accuracy across time is inherently far more difficult than in simpler classification methods. In other words, it's much harder to explain various degrees of gray than to explain an object being either black or white. Therefore, a higher standard of predictive data may be required. The second reason deals with *censoring*—a data issue where important information is not available or is present only outside the study period. Given that the data collection efforts for any study cover a specific period, there will be many accounts whose final default status you will not know. In fact, those accounts will make up the majority of your data. This seemingly small detail dramatically affects not only the accuracy of the predictive model, but also the choice of the statistical tool.

Predicting time to default is part of a larger field of study called *survival analysis*. Much of the progress in this field came from World War II military research, where the aim was to predict time to failure of military equipment. Combined with recent advances in computer technology and the rapid testing of new drugs in clinical trials, survival analysis has experienced a dramatic resurgence. Like many other fields of study, survival analysis has its own terminology. One term is *hazard rate*. With respect to credit risk, the hazard rate is the chance that an account will default in some increment of time, divided by the chance that the account has "survived" up to that point. If an account has a high hazard rate, then its survival time is small. If it has a low hazard rate, then its survival time is large. These hazard rates can have different shapes over time. The good news is that they can be accounted for in the modeling process.

Perhaps the easiest way to understand time-to-default modeling is to look at some data. Figure 1 shows eight hypothetical retail loans selected at random. Assume for each account we have three variables to predict time to default—X1, X2, and X3. Further assume that this information does not change over time. The starting point of the study is the *point of origin*—that snapshot in time where we begin tracking account status and performance. In our example, let's use January 2000 as the point of origin. If we measure time in months since the point

Figure 1

Data-Censoring Concept



of origin, then a survival time of 12 means that our account experienced payment default 12 months after January 2000, or January 2001.

Figure 1 shows the tracking of accounts over a 36-month window. If the point of origin is January 2000, then the end of the study period is January 2004. Past this period we would not know whether an account has defaulted. If we do not know, then those accounts represent censored data. Censoring also can occur before the end of the study period. Account 3 could have been closed due to a death event. Account 5 could have been closed at the request of the customer for normal attrition reasons. In this framework, if an account has not defaulted, then its value is considered censored. Therefore, in our Figure 1 example, all loans are considered censored except for accounts 2 and 7, both of which defaulted.

With Figure 1 in mind, let's turn these eight accounts into data appropriate for modeling. Figure 2 shows the creation of the dependent variable (survival_time) for our time-to-default model. As mentioned previously, the dependent variable is what we want to predict. However, in survival modeling, a variable is required to identify censoring. Here we name it *censor*. Note that there are no survival times greater than 36 months because that's the end of our study period—our censoring limit. Censor is a 0 / 1 variable indicating whether that observation represents a censored account. A censored variable is assigned a value of 0. Otherwise, it gets a value of 1.

LGD Models versus Time-to-Default Models

You may consider it good news that the same basic type of regression technique recommended as a possible approach for LGD modeling also can be used, with some variations, for time-to-default modeling. In the economic literature, this type of approach is called a *tobit* model. In our LGD model, the dependent variable was the percentage of dollars recovered (or not recovered). For a time-to-default model, the dependent variable is the number of months since the point of origin, or how long the account has "survived."

There are a number of ways to estimate a time-to-default model in SAS or any other statistical software package, depending on your assumptions and whether or not your predictor variables change over time (time-dependent attributes). Let's keep things simple here and work with a SAS procedure called PROC LIFEREG, which estimates a general class of models called accelerated-failure-time (AFT) models. This procedure, however, does not allow your predictor variables to change over time, which can often enhance your model's predictive power. That capability, provided by a more complex procedure in SAS called PROC PHREG, is outside the scope of this discussion. For purposes of our discussion, let's assume the values of the predictor variables were captured at the beginning of the study period and remain unchanged over the three-year period. Figures 3 and 4 show the SAS code necessary to estimate both a LGD and a time-to-default model. We'll use X1, X2, and X3 as predictor variables. Notice how similar the SAS code looks between the two types of models.

In Figure 3, our tobit model for predicting LGD uses the SAS keyword *LOWER* to account for a clustering around accounts that show no recovery. As mentioned in previous articles, having a number of loans clustered at a zero recovery rate could cause modeling problems if you were to use linear regression. The tobit model was recommended as one solution. This clustering issue around a lower bound (zero in our LGD case) is similar to our censoring problem in the time-to-default model.

As shown in Figure 4, the SAS code for time to default looks almost the same, except it requires the name of the censoring variable and a designation of a "0" or "1" to represent the censoring value.

Figure 2

Modeling Data

Observation	Dependent Variable "Survival Time"	Censoring Variable "Censor"	X1	X2	X3
1	36	0	.72	30	20
2	15	1	.56	45	19
3	20	0	.90	23	45
4	36	0	.90	12	76
5	34	0	.75	4	56
6	36	0	.88	36	12
7	6	1	.90	61	44
8	36	0	.90	22	11

The Hazard Function

In time-to-default modeling, one of the first tasks is to determine the shape of the hazard rate (hazard function)—that is, which distribution to use.

Figure 3

SAS Code for LGD (Tobit Regression)

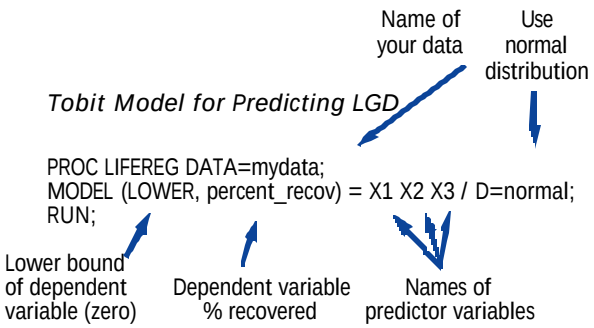
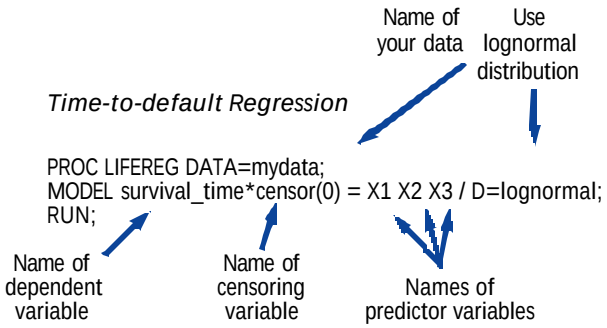


Figure 4

SAS code for Time for Default



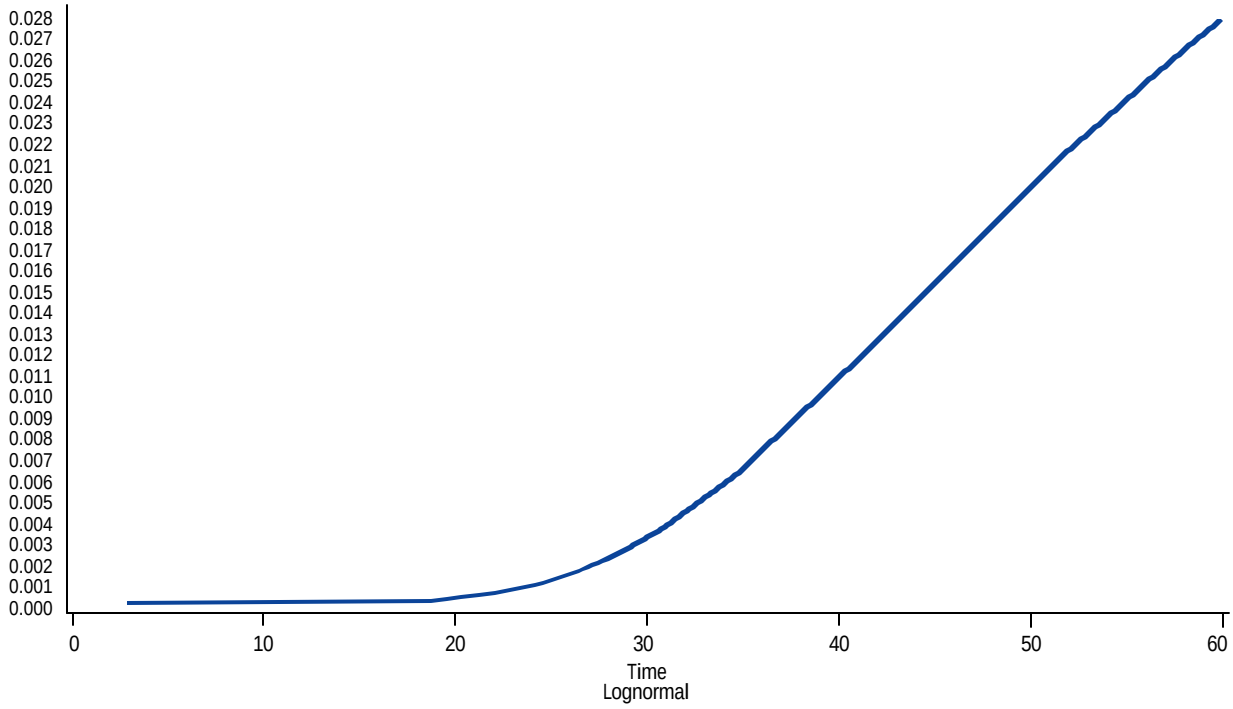
To a large extent, this can be done by running the model with a variety of distributions and picking the one with the best fit. One way of doing this is by looking at the log-likelihood value, a statistical measure produced by most software packages. The model with the log-likelihood value closest to zero wins. In Figure 4, we show the use of the lognormal distribution to model the hazard rate. By using this particular distribution, we can even account for a more complex hazard rate that can turn upward, peak, and even turn downward, depending on the data. A typical lognormal hazard rate is shown in Figure 5. Although the concept really applies more to the behavior of individual accounts, the hazard rate in Figure 5 is calculated using the sample averages of the data for illustrative purposes.

Prediction and Application

The statistical output of our time-to-default model looks very similar to any other regression model—a set of coefficients for each predictor variable, an intercept, and an estimate of something called *SCALE*. For the lognormal model, this scale estimate simply stretches or compresses the hazard

Figure 5

Typical Lognormal Hazard Function



rate. Once the model's coefficients have been estimated, we can do two very useful things.

1. We can make predictions of time to default measured from our point of origin. Although the coding of how these coefficients are used in the prediction formulas is not listed here, it can readily be found in various references.[2] Attention could be focused on those accounts that were predicted to default within the next six months, such as treatment letters or a review of their payment status.
2. We can specify a survival time threshold and calculate the probability that each account would survive up until that point in the future. If you knew there was an 80% chance that your account will survive three more years, you may want to grant some (but maybe not all) increases in their credit lines.

In addition, if there is a contract involved, then this predictive information could be helpful in setting up new terms of the loan.

Summary

The good thing about implementing the modeling requirements for Basel II is that they provide an excellent foundation for establishing sound risk management practices. These practices come at all stages of the credit life cycle—from booking a loan to possible default and collections. Time-to-default modeling allows the institution to build on this foundation while offering some distinct advantages:

- It deals with censored data.
- It avoids the inflexibility of having to choose a fixed period (one year, for example) to measure performance, as in a PD model.
- It allows greater flexibility in incorporating economic changes (time-dependent characteristics) into the scoring system.
- It forecasts default levels as a function of time according to a variety of hazard rates.
- It can use much of the same predictive information already found in existing corporate data warehouses that are used for management reporting, credit-scoring applications, and Basel II compliance.

Time-to-default techniques can also be fine-tuned to account for more than just the time to default. For example, in residential mortgages, survival analysis has been used repeatedly to simultaneously account for time to prepayment. This is done under the topic of competing risks—something that offers a great deal of flexibility and practical application in the banking industry. In summary, modeling time to default can allow banks a more structured mechanism for taking standard risk management practices to a new level, moving toward a more quantitative approach to full profitability analysis—the ultimate bottom line. ❐

Contact Morrison by e-mail at
Jeff.Morrison@suntrust.com.

Notes

1 Morrison, Jeffrey S., "Preparing for Modeling Requirements in Basel II Part 1: Model Development," The RMA Journal , May 2003.

2 The following publications provide more information on coefficient coding for prediction models:

Allison, Paul D., Survival Analysis Using the SAS® System: A Practical Guide, SAS® Institute Inc., Cary, North Carolina, 1995.

Harrell, Frank E. Jr., Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis, Springer-Verlag, New York, Inc., 2001.

Edelman, Thomas et al., Credit Scoring and Its Applications, Society for Industrial and Applied Mathematics, 2002.