INTRODUCTION TO SURVIVAL ANALYSIS IN BUSINESS

By Jeff Morrison

Survival model provides not only the probability of a certain event to occur but also when it will occur ... survival probability can alert a company whether or not a specific account needs a special treatment … the analysis can be used effectively in retaining existing customers and acquiring new ones in various industries including telecommunication, banking and finance.

It has been said that *timing is everything* – especially in the area of romance. Well, the same may be true in business. Of course, the first thing that might come to mind is the stock market. If we only knew when a particular stock was going to jump in price, we could quit our day jobs! Unfortunately, that kind of information is often not available. What is available is a field of statistics called Survival Analysis. It deals with the timing of events with applications spanning medicine, law enforcement, banking, telecommunications, and a host of other industries. As the field of credit scoring is focused on predicting 'if' an account will become delinquent over a certain span of time, Survival Analysis can tell us 'when.' Survival Analysis is called different things in different industries — event history analysis, reliability analysis, time to failure, and even duration analysis. The purpose of this article is to give an introduction to the subject with an emphasis on how it can be used in banking and finance. However, some discussion will be devoted to Survival Analysis and customer retention — a topic that has seen much attention in the field of telecommunications.

SCORING CREDIT

Because survival analysis and credit scoring go hand in hand, let's start with a brief discussion of credit scoring. In a broad sense, credit scoring is the application of statistical techniques to determine if credit should be granted to a borrower. It involves collecting information on a set of accounts reflecting satisfactory (good) payment status for a particular period of time (called the observation window) and following their payment performance, say, for a period of one year. At the beginning of the observation window, we may know a great deal about the account — its time on books, how many times it went 30, 60, or 90 days delinquent, its credit limit, etc. We then apply a regression technique that yields a predicted value that will hopefully distinguish between accounts that will either pay or not pay in the year that follows. The choice of the 1-year window is optional depending on the application of the score. What is important to know is that the score does not attempt to describe 'when' the event will occur within the performance window. It only describes the likelihood of the event occurring during that 1-year block of time. To address the timing issue, a more exacting approach is needed — Survival Analysis.

JEFF MORRISON

Mr. Morrison is Vice-President of Modeling at SunTrust Bank. He has over 20 years of analytic experience in telecommunications, natural gas distribution and consumer credit. He has spoken at numerous conferences in the areas of quantitative analysis, econometric forecasting, and new product planning. He has widely published.

SURVIVAL ANALYSIS

Let's take a fictitious example in Residential Mortgage for illustrative purposes. Say you conduct a study gathering information on 15 year fixed rate mortgages that are in good standing in January 1995. You then track their monthly performance over a five-year period ending January 2000. In a credit scoring application, you would gather information available in January 1995 on each mortgage, and then determine 'if' there was a default anytime during the five years. However, in building a Survival model, you also record the timing of the default from the point of origin. In this case, because the point of origin is January 1995 we record the number of months until default since January 1995. The reference to the point of origin is essential in interpreting the survival probabilities, which we will discuss shortly. Table 1

shows a part of what you might have come up with.

The data in Table 1 needs some explanation. Default is assigned a value of "1" if the loan was observed to have defaulted during the five-year period. However, during the course of the study, you may have loans that were paid-off or sold for a number of reasons. The loan might have been sold to a third party, it could have been paid off early by the borrower, or it could have reached its maturity and been paid off as scheduled. For simplicity's sake here, we will assign these accounts a value of "0" under 'Default.' If these cases occur at random, then they are often referred to as randomly censored observations.

What about those accounts that did not default over 5 years and are still active? They could default in the future, but we are not able to observe the event. This is another example of censoring — more specifically, right censoring — a problem specifically suited for Survival Analysis. Censoring gives headaches to many other regression techniques often used in data analysis. The 'Default' value for these accounts is also assigned a value of "0."

Now we record the time to default since our origin of time. In our example, a variable called 'Time' is created to represent the survival time of the loan since January 1995. In the first account, the loan was observed to have defaulted at the 37th month past January 1995. Account number five has a value of "0" for 'Default' and a 'Time' = 60, indicating the default event was not observed over the five-year observation window. Maybe a default occurred with the loan after five years, or maybe it did not. For purposes of this study, we simply don't know. X1-X3 could be any variable describing the loan that might be predictive of default — product category, loan to value ratio, borrower's income level, geographic location, etc. or it could have a linkage with economic variables such as GDP or interest rates.

As you will see, what makes Survival Analysis unique is its ability to handle many different types of censoring that occurs in real world business situations. Your standard regression tools like linear, logistic, or probit regression are unable to handle this censoring issue. In the unlikely event that you have no censoring, you could correctly use linear regression and explain the survival time as a function of the explanatory variables — in this case X1-X3. However, in Survival Analysis applications, you rarely have the luxury of not dealing with censoring. If you have censored observations and mistakenly use linear regression to estimate the model, especially if the censoring occurs in abundance, then you will get bias and unreliable results — either underestimating or overestimating the parameters. Logistic and probit regressions are not set up to handle a continuous dependent variable such as time to default, but a dependent variable with only two possible outcomes. Survival Analysis, on the other hand, has a collection of mathematical techniques that can be correctly applied.

First, it is appropriate to mention what Survival Analysis calls the hazard rate. It is simply the risk of the event occurring over time. If an account has a high hazard rate, then its survival time is low. Likewise, if it has a low hazard rate, then its survival time is expected to be high. If the process generating the event you are studying (in this case, loan default) does not change over time, it is said to have a constant hazard rate. This type of model is the simplest to estimate, but is often rare in business. This is because risk often changes with time – just as the risk of dying increases with age. Although the idea of a hazard rate is account specific and varies across individuals, the graphs shown in this article were created by using the average values (sample means) of the data.

So, how do you know the form of the hazard rate? Is it constant, or does it take on a variety of shapes that are more complex? That's where the statistical procedures in the more sophisticated software packages

TABLE 1
EXAMPLE OF SURVIVAL DATA

	Default	Time	X1	X2	X3
1	1	37	1	33	4
2	0	14	2	21	3
3	1	55	3	59	2
4	1	44	2	76	1
5	0	60	3	24	3
6	0	60	2	22	2
7	1	32	1	28	2
8	1	38	1	49	2
9	0	60	3	76	2
10	1	55	2	59	2
11	0	60	3	45	5
12	0	22	3	21	4
13	0	60	3	49	2
14	0	60	1	23	4
15	1	55	1	55	2
16	0	60	2	29	3
17	1	44	1	49	2
18	0	60	2	45	5
19	0	30	2	24	3
20	0	60	1	30	3

TABLE 2
PARAMETER ESTIMATES BY MODEL

	A Exponential Model Estimates	B Weibull Model Estimates	C Lognormal Model Estimates
Intercept	3.4085	3.7485	3.7986
X1	0.7	0.2736	0.3121
X2	-0.0281	-0.0132	-0.0156
X3	0.5801	0.2283	0.153
SCALE	1	0.2792	0.3644
Log Likelihood	-66.59	-34.89	-34.7

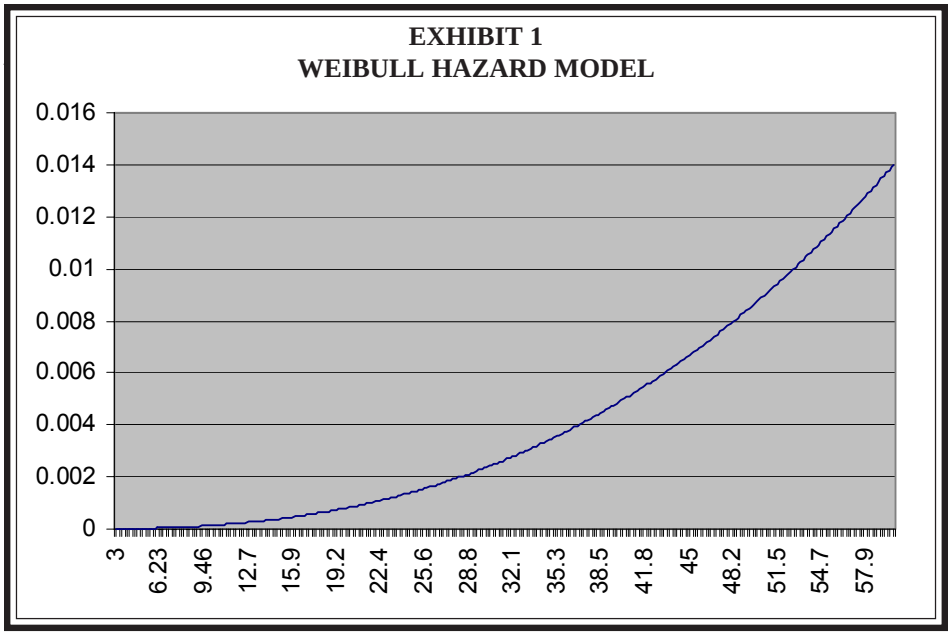
such as SAS® and S-PLUS can help. SAS® has procedures called PROCS that perform a variety of heavier computations associated with estimating models. In our example of loan data, we use PROC LIFEREG to estimate a Survival Model:

Proc lifereg data = survival; Model TIME * DEFAULT(0) = X1 X2 X3 /dist = exponential; Run;.

In the "model" statement, our variable called TIME (survival time) is considered the dependent variable, and DEFAULT is the event variable where a value of "0" in parenthesis indicates the censoring. Actually, SAS®'s LIFEREG procedure automatically calculates and uses the natural log of the time variable (log(time)) as the dependent variable. This ensures that the predicted values will be positive regardless of the values of the data and parameters that are estimated. X1, X2, and X3 are the variables we are going to use to predict time to default. At the end of the "model" statement, we specify the type of model (or distribution) to be used with the 'dist =' notation. Three distributional models will be discussed here: (A) Exponential, (B) Weibull and (C) Lognormal. Although other possible distributions can be used depending on your software, these are the most popular ones.

EXPONENTIAL DISTRIBUTION MODEL

Let's begin our discussion by using the Exponential distribution for our first model as shown in Table 2, column A. Among other things, Table 2 shows us that the Log Likelihood was found to be 66.59. Although not always definitive, we will use this measure to compare the fit of several other models to determine which is the best. We will consider the model to fit the data better if its Log Likelihood value is closer to 0. The parameter estimate, called SCALE, represents the estimate for the random disturbance term in the model. For the exponential model, the SCALE factor will always be one. Because the formulas for computing the hazard and survival rates are a little complex, the reader is referred to the Allison reference at the end of the article for further details. The important thing to note here is that, for the exponential model, the hazard rate will ALWAYS show up as a horizontal line over time. This is often an over-simplification of the risk dynamics in most real world applications.



WEIBULL DISTRIBUTION MODEL

Column B of Table 2 shows the results of using a Weibull distribution rather than the Exponential to estimate the model. For the Weibull model, the restrictions of being nothing more than a straight horizontal line are removed. The Weibull model allows us more flexibility than the previous one with a hazard rate that can trend upward or downward, increase at an increasing rate, or increase at a decreasing rate. It cannot, however, reach a peak and then go in the opposite direction. In our example, this model had a much better Log Likelihood (-34.89), indicating a better fit to the data. Exhibit 1 gives the graph of hazard rates based on this model. It reveals

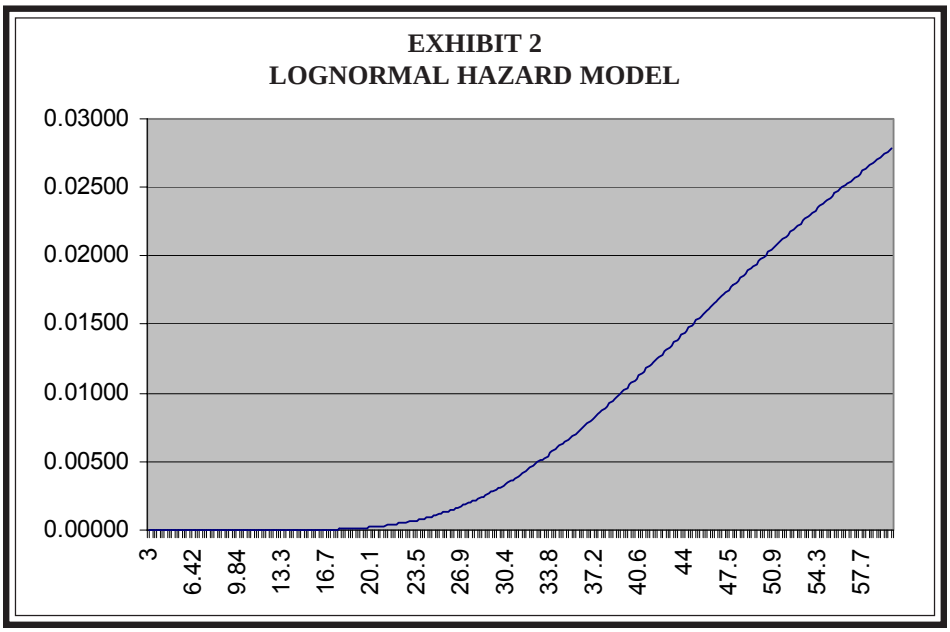


TABLE 3
SURVIVAL PROBABILITIES

	Default	Time	X1	X2	X3	Survival Probability LNORMAL 12 Months	Survival Probability LNORMAL 24 Months	Survival Probability LNORMAL 36 Months	Survival Probability LNORMAL 48 Months
1	1	37	1	33	4	100.00%	99.77%	95.70%	82.31%
2	0	14	2	21	3	100.00%	99.99%	99.62%	96.97%
3	1	55	3	59	2	100.00%	99.52%	93.05%	75.49%
4	1	44	2	76	1	99.36%	72.22%	30.05%	9.47%
5	0	60	3	24	3	100.00%	100.00%	99.97%	99.54%
6	0	60	2	22	2	100.00%	99.95%	98.62%	92.13%
7	1	32	1	28	2	100.00%	98.62%	86.22%	61.83%
8	1	38	1	49	2	99.93%	90.42%	57.67%	27.56%
9	0	60	3	76	2	99.99%	96.90%	77.43%	48.56%
10	1	55	2	59	2	99.99%	95.87%	73.33%	43.39%
11	0	60	3	45	5	100.00%	100.00%	99.96%	99.46%
12	0	22	3	21	4	100.00%	100.00%	100.00%	99.92%
13	0	60	3	49	2	100.00%	99.87%	97.17%	86.80%
14	0	60	1	23	4	100.00%	99.94%	98.40%	91.22%
15	1	55	1	55	2	99.84%	85.31%	47.50%	19.71%
16	0	60	2	29	3	100.00%	99.97%	98.99%	93.76%
17	1	44	1	49	2	99.93%	90.42%	57.67%	27.56%
18	0	60	2	45	5	100.00%	99.98%	99.34%	95.46%
19	0	30	2	24	3	100.00%	99.99%	99.44%	95.98%
20	0	60	1	30	3	100.00%	99.44%	92.29%	73.74%

that the estimated risk of default increases at an increasing rate over time.

LOG-NORMAL MODEL

In Column C of Table 2, we see the estimates of fitting a Lognormal distribution to the data. This type of model offers even more flexibility for the hazard curve than the previous ones. Unlike the Weibull model, the hazard curve for this distribution can actually have a hump. In other words, it can turn upward, peak, and then turn downward depending on the data. The actual shape of the curve depends on your data. As shown in Table 2, the Lognormal model has an estimated Log Likelihood of -34.70, the best so far – but not by much.

As shown in Exhibit 2, the Lognormal hazard function takes on the appearance of the beginning of an S-shaped curve, meaning that the risk of default is small for loans that have been around for less than 24 months since January 1995, but jumps up remarkably at the 2-year time frame. Depending on the data, the end of the S-curve could actually start to peak and then decline.

Now what? Well now we have the model (parameter estimates) to construct a probability of survival for each account since the point of origin. We can also create a probability to survive for any month within our 60-month window, or maybe venture out a little further with some degree of confidence. Therefore, for illustrative purposes, we will predict the probability of survival at the end of each year – 12 months, 24 months, 36 months, and 48 months.

If you were to compute an account level survival probability for each month in the study, you would see that almost every account had a probability of surviving close to 99% for the first few months since the origin of time. This is not surprising given the early shape of the hazard curve. To use this outcome in a systematic fashion in servicing the loans, you could set up a decision threshold such as a survival probability of .50, for instance. If the survival probability of an account for some month in the future falls below that threshold, the account could be flagged as a potential short timer and setup for some special treatment. Such a treatment might include limiting additional exposures to that customer as in home equity lines. In our example, the survival probabilities are given in Table 3.

Using the parameter estimates given in Table 2 (column C) for the Lognormal model, the probability of survival until a certain period can be easily computed. For example, the probability of survival until the 48th month for the first observation in

Table 3 is calculated as follows:
Step 1: Use the parameter estimates for X1, X2, X3, and the constant in Table 2 to calculate the linear predictor, LP:

LP = 3.7986 + .31212*(X1) - .01556*(X2) + .15299*(X3)
LP = 3.7986 + .31212*(1) -.01556*(33) + .15299*(4)
LP = 4.209

Step 2: Use the standard normal cumulative distribution function in Excel to complete the calculation for a specific time period, in this case month 48:

Survival Probability (Month: 48) =1-NORMSDIST(((LN(48))-4.209)/0.36443)

Survival Probability (Month: 48) = .8231 (Here scale = 0.36443)

The results of the model could also be used for new loans in setting prices. In other words, for a new applicant who has a predicted survival probability of only 3 years, the company may wish to offset the additional risk with a different term structure, mortgage insurance, or up front fees. Since most generic credit scores produce a probability of default over the next year, Survival Analysis allows the bank to exert additional flexibility in dealing with the potential risk of default.

OTHER APPLICATIONS OF SURVIVAL ANALYSIS

Finally, it may be worthwhile to note the application of Survival Analysis in customer retention. Given the intense level of competition in today's market, telecommunications companies are desperately looking for ways to get new customers and retain those they already have. This effort to keep existing customers is referred to as churn analysis – an additional field of study where Survival techniques can be applied with considerable value. In this environment, the company wishes to predict the survival times of their existing customers. The same issues of censoring apply here as in banking and finance. Once survival times have been predicted, then the company can devise customer retention strategies tailored to their customer's propensity to remain loyal. ■

REFERENCES

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

Lu, Janxiang. **Predicting Customer Churn in the Telecommunications Industry – An Application of Survival Analysis Using SAS®.** SUGI 27 (SAS® Users Group International), April 2002.

Pennington-Cross, Anthony. **Patterns of Default and Prepayment for Prime and Nonprime Mortgages.** Office of Federal Housing Enterprise & Oversight, Washington, D.C. March 2002.