Preparing for Modeling Requirements in Basel II

Part 1: Model Development

by Jeffrey S. Morrison

Gain the benefit of a regional bank's experience. In this first of four articles, Jeff Morrison discusses SunTrust Banks' approach to statistical modeling. Part II details steps taken to validate the model. Part III pulls it all together within a GUI software interface. Then Part IV moves into the realms of stress testing.

The Basel II Capital Accord, currently planned for implementation in 2007, sets out detailed analytic requirements for risk assessment that will be based on data collected by banks throughout the life cycle of the loan. The purpose of Basel II is to introduce a more risk-sensitive capital framework with incentives for good risk management practices. Many banks are examining or implementing models now to help enhance their risk management efforts. And it can get pretty confusing article can serve as a refresher.

Models

Remember that old statistics book in college and what you said about it?: "I'll never use that stuff in the real world!" Well, never say "Never." That old book and this

Let's start by defining the word *model*. Webster's more statistical definition of the word is "...a system of postulates, data, and inferences presented as a mathematical description of an entity or state of affairs." Basically, think of a model as a mathematical representation of reality. It's not going to be perfect, it will definitely be oversimplified, but the aim of such a representation is to gain insight into behavior so predictions can be made that are both reasonably accurate and directionally correct.

Quantitative models in consumer credit have been used for many years. Models developed from the application data on new accounts are called front-end or *application* models. These models do not use the prospective lender's payment history information for a potential new borrower because that information is simply not available. Once these accounts begin to become seasoned, different models can be developed to yield behavioral scores, that is, algorithms designed to include payment history as well as other factors associated with loan origination, geography, and the demographics of the borrower. In contrast, scores developed from pools of data typically obtained from credit bureaus are called *generic* models. These models reflect credit behavior across a variety of financial institutions and capitalize on the assumption that a consumer will exhibit behavior around some average risk level. Customized scores developed with the payment history of a *single* institution

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

can often outperform generic models because they are tailored to the specific credit issuer.

Models for Basel. Similar models may be developed for Basel. The models used in SunTrust's Risk Rating System have been built specifically for Basel II on a two-dimensional structure. The first dimension reflects the probability of default (PD) for the obligor. The second reflects the loss given default (LGD) associated with a particular loan or facility. Therefore, for each loan, the expected dollar loss is simply the product of the dollar Exposure at Default × PD × LGD.

Let's begin by looking at developing a PD model for the obligor and then move toward developing a facility-based model for LGD. We can construct these types of models for the commercial side of the business, but to make it simpler, think in terms of retail portfolios, such as residential mortgage, as you read further.

Typically, bank models for Basel requirements come in two flavors—vendor and custom. In the commercial world, models may have to come from vendors because only they have invested the resources to collect data robust enough for modeling. This is because the number of commercial defaults for any single bank in a given year is so small. Based on the sheer size of the loan volume, the retail side is just much riper for custom modeling, where a bank can use its own data and not rely on costly vendors. Even if a bank does not yet have enough historical data to develop a statistical model, it can begin with one derived from judgment and consensus until the more sophisticated models are available.

BECAUSE BASEL REQUIRES ALL LOANS TO BE RATED WITH THESE MODELS FOR A CERTAIN MINIMUM AMOUNT OF TIME BEFORE THE ADVANCED APPROACH MAY BE USED, INTEGRATING VENDOR AND CUSTOM SOLUTIONS INTO THE PROCESS SHOULD BEGIN AS SOON AS POSSIBLE.

Judgmental models are simply a set of rules that quantify assumptions about the portfolio's risk level without the use of statistical approaches. Examples might include a mapping of risk grades according to loan-to-value or debt-to-income ratios. Others might provide a rough mapping of FICO score bands to PD. Although judgmental models definitely have their place, the remainder of this article will focus on the development of statistical models that are reflected in both custom and vendor efforts. And because Basel requires all loans to be rated with these models for a certain minimum amount of time before the advanced approach may be used, integrating vendor and custom solutions into the process should begin as soon as possible.

The current school of thought surrounding the models mentioned in Basel is that banks should have separate models for the obligor and the facility. The obligor model should predict PD—usually defined as 90-plus days delinquent, or in foreclosure, bankruptcy, charge-off, repossession, or restructuring. Models on the facility side should predict LGD, or 1 minus the recovery rate.The recovery rate is simply the amount of dollars recovered divided by the dollars owed at the time of default.

Let the fun begin! As will be shown, the statistical approaches associated with these two applications are quite different. But first, here are a few simple definitions:

- Dependent variable—the variable you wish to predict (default versus nondefault or percent recovered).
- Independent variables—explanatory variables (LTV, debt burden, etc.) used to explain the dependent variable.
- Correlation—a number between -1 and 1 that measures the degree to which two variables are linearly related. A high correlation is a correlation near +1 or -1.
- Regression analysis—a family of statistical procedures that quantify the relationship between the dependent variable and a set of independent variables using historical data. There are many types of regressions.
- Parameter estimates—the set of weights produced by the regression used for prediction. One weight is used for each independent variable plus a constant value, sometimes called a *y-intercept*.

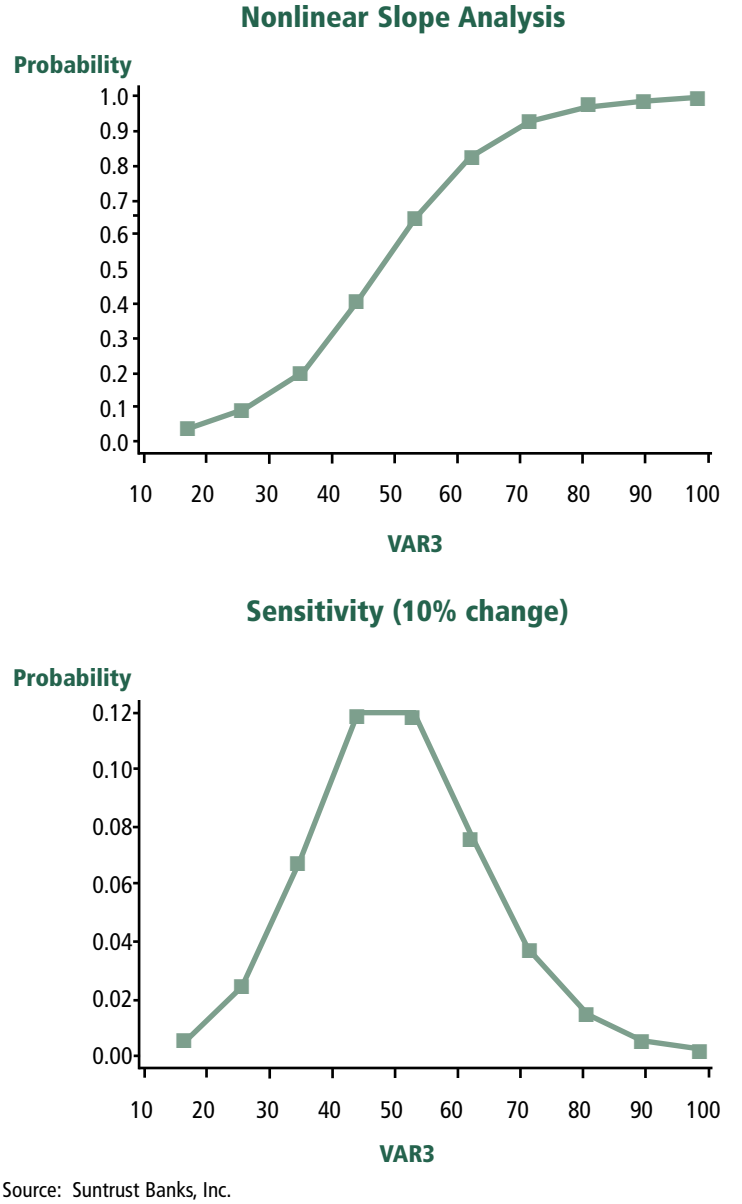
Regardless of the type of regression you use, all approaches allow you to determine which independent variables to include or leave out in the model. When you include an explanatory variable in a regression model, you generally will get back a parameter estimate. However, given some level of precision, this estimate might not be significantly different from zero and, therefore, should not be used. A measure called a *t-statistic* is produced by most regression packages; the t-statistic indicates whether a variable should be left out of the model. This is one of the primary advantages of using regression. Modeling is not an exact science, and because statisticians come from a wide range of backgrounds and experience, a number of modeling approaches or designs are possible that could work quite well. Nevertheless, the purpose of this article is to offer some general advice or rules of thumb that you can use to get a head start on the modeling process in your financial institution.

Obligor Models: Probability of Default

Since regression analysis is the primary mechanism for building statistical models, let's begin there. Many types of regression procedures exist. For predicting the probability of default, *logistic* regression is often recommended. Logistic regression is appropriate in cases where the dependent variable is binary—taking on one of two values. In this discussion, the dependent variable is an indicator of whether or not the loan went into default over a certain period of time—usually a year. If the objective is trying to predict the probability of default, then the dependent variable would have a value of 1 (for a default) or 0 (for a nondefault). Most statistical software packages will easily perform this type of regression.

Logistic regression has some interesting capabilities. First, the predicted values from the regression come out just the way you need them—as probabilities bounded between 0 and 1. So if you have a predicted value of 0.356, then that loan has a probability of default over the next 12 months of 35.6%. Second, logistic regression has the flexibility of capturing relationships that are nonlinear, such as LTV. Depending on your data, the relationship between LTV and the

Figure 1



Source: Suntrust Banks, Inc.

probability of default can be nonlinear or S-shaped. The top part of Figure 1 shows that the independent variable VAR3 (think of it as LTV) has the steepest slope at a value of 50, making the model most sensitive around this range. This is demonstrated in the bottom part of the graph where a 10% change in VAR3 values around 50 will lead to a 12% change in the probability of default—all other things remaining equal. Changes around VAR3 values that are much lower or higher than 50 would tend to have a significantly smaller impact on the probability of default. Note at the top part of the graph the probability of default plateaus as it approaches .90, where LTV is near 80.

OK. Now with a little understanding of logistic regression in your back pocket, let's prepare an instructional checklist for building a model. As an illustrative portfolio, think in terms of residential mortgages.

Checklist #1

Step 1: Define your dependent variable. Let's say a default is a loan that is 90-plus days delinquent, or in foreclosure, bankruptcy, charge-off, repossession, or restructuring. Code an indicator variable for this with a 1 (default) or a 0 (nondefault).

Step 2: Define the performance window. This is the amount of time over which the set of accounts can enter a default status. Let's choose one year.

Step 3: Find all the loans that were in nondefault status a year ago and track their performance over the following 12 months. If you have a huge portfolio, you may want to take a random sample. The sample size could vary widely. Around 25,000 observations may be a good number to build the model, and another 25,000 might be needed to test it. However, having a sufficient number of defaults is very important—the more the better. Attach the indicator from Step 1 and call it the dependent variable.

Step 4: At the *beginning* of the performance window (one year ago), select relevant variables that you think might be predictive of default over the next 12 months—LTV, age of the loan, loan type, number of times the loan is 30/60 days late, etc.

Step 5: *Look* at the data graphically, through frequency counts, averages, minimums, maximums, and correlations. Usually, you can examine your data from all of these perspectives with just a few commands in most software packages. As you don't want to include two independent variables in the regression that reflect duplicate information (i.e., too correlated with one another), look for these candidates. See which variables are correlated the most with the dependent variable. Look for wacky data—observations that have extremely high or extremely low values.

Step 6: Important—*handle missing data*. For each variable, determine the percentage of missing data. In general, if the percentage of missing values for a particular variable is greater than 30%, don't use it. Although this is an arbitrary cut-off value, the idea is to look for variables that have sufficient information content. For the remaining variables, substitute the average, or mean, of the values you do have as a proxy for missing a small amount of information. If no steps are taken to handle missing data, the regression software will automatically skip the record, and you could end up eliminating most of your data from the analysis.

Step 7: Estimate your model by running a logistic regression with a *stepwise* option. This simple feature will automatically remove any variables that are not statistically significant. The software does the work for you.

Step 8: Examine the *sign* of the parameter estimates. Does it make sense from a business standpoint? A negative sign means there is an inverse relationship between the variable and the probability of default. A positive sign means that as the value of the variable increases, so does the probability of default. Do *not* simply take the answers from the regression at their face value. Look at your results. If the sign is counterintuitive, then look at the data again to find out why.

Step 9: Produce the predicted probabilities from your model. Often this is done for you automatically. However, to show how it works, look at Figure 2. This is example code in SAS® that uses your parameter estimates to produce default probabilities.

Figure 2

Implementation Code

	code
6	IF VAR2 = . THEN VAR2 = 20
7	IF var 3 = . THEN VAR 3 = 42.492;
8	IF VAR4 = . THEN VAR4 = 2.66;
13	HSCORE =
14	1.8538568006 +
15	VAR2 * -0.145032377 +
16	VAR3 * 0.1081924412 +
17	VAR4 * -1.556902303;
18	HSCORE = 1 /(1+EXP(-(HSCORE)));

Figure 3

Probability of Default Calculation

A	B	C	D
	Parameter	**Value**	**B*C**
Intercept	1.8538	N/A	1.8538
Var2	-0.145	20	-2.9
Var3	0.10819	42	4.54398
Var4	-1.557	3	-4.671
Sum			-1.17322
Using exponential formula on Sum: Probability = 1/(1+exp(-sum))			**0.236273447**

The example code is called *implementation code*. In this case, we substituted the means of VAR2–VAR4 as proxies for missing information (shown as "="). Also, be sure to include numeric checks in case invalid values somehow find their way to your code. The "EXP" mathematical function at the end is what turns your answer into a probability. The variable, called HSCORE, would be your estimated PD. Assuming you have valid values for VAR2–VAR4 and no missing information, you could perform the calculations easily in Excel, as shown in Figure 3. The account in the example given has a one-year probability of default of 23.6%.

Facility Model: Loss Given Default

According to Moody's Investor Service, "There is no good framework for predicting the outcome of a default. This deficiency is so poignant because default outcomes are so broadly diverse. A defaulted loan may pay off essentially in full with accrued interest or it might pay off only five cents on the dollar. A resolution might complete by the next month or it might take four and one-half years."[1]

In other words, building a recovery model from a loss perspective, especially on the commercial side, is hard—much more so than building one for probability of default. This is because it's hard to get enough predictive data. The accuracy you achieve in using one particular statistical approach over another is secondary to obtaining enough good-quality data. Since there are so few commercial defaults, the time needed to collect default data may be substantial. By contrast, on the retail side you should experience a higher level of success because of the abundance of default data. Now we will examine two types of statistical approaches recommended for estimating loss given default or 1 minus the recovery rate.

Remember how we had to collect information on both defaulted and nondefaulted loans? In building a recovery model, we focus only on information related to defaults. For example, say you collected the following information on defaults from your residential mortgage portfolio:

- Percent dollars recovered.
- U.S. Census region / geography / zip code / MSA.
- Age of the loan at default.
- LTV.
- Indicator for type of bankruptcy.
- Amount of time in collections.
- Age of property.
- Change in property value as of one year ago.
- Average household income in that geography.
- FICO score.
- Index of leading economic indicators.
- Size of the outstanding balance.

Since the dependent variable (the recovery rate) is not a binary (0/1) variable as in our default model, logistic regression is not the appropriate approach. The recovery rate typically varies from 0-100%, depending on how you account for charges and fees. Given the shape of the distribution for this type of data, two statistical techniques are often used—linear regression and tobit[2] regression.

Perhaps the most popular type of regression is linear regression, which uses the method of *least squares* to compute the weights for the prediction equation. The name says it all. This technique produces a line that minimizes the squared differences between the actual and predicted values. Linear regression can only estimate a *linear* relationship between the independent variable and the recovery rate. Unfortunately, even if the relationship is really nonlinear (S-shaped or U-shaped), linear regression will provide only a linear approximation to the curve. This should not be of too much concern, since a good statistician knows some tricks to work around this limitation.

Tobit regression can be thought of as a hybrid between a linear regression model and logis-

tic regression's twin brother, probit regression. *Probit* regression is similar to its logistic sibling, but it is based on a slightly different S-shaped distribution. Tobit regression's edge over the other methods is that it was designed to handle cases where the dependent variable is clustered around limits such as 0. If there are many observations where the percent recovered was 0 (as in the case of consumer credit cards), then estimating the model using linear regression could produce biased, less accurate results.

Now armed with this information in your other back pocket, you are ready to build your recovery model. The good news here is that you can eliminate a step or two from what you did in your PD model. For example, there is no need to worry about a performance window since you are only dealing with defaulted loans. So here's your second checklist:

Checklist #2

Step 1: Define the dependent variable—percent dollars recovered. Since the recovery operation can be an ongoing process over a long period of time, part of the defining process is to set up a cutoff period for recovery transactions. For example, if you haven't collected any additional money in two years after the default, then you might assume the collection process is complete.

Step 2: At the time of default, add the explanatory variables that might be predictive.

Step 3: *Look* at the data graphically, through frequency counts, averages, minimums, maximums, and correlations.

Step 4: Important—*Handle missing data*. For each variable, determine the percentage of missing data.

Step 5: Estimate your model by running the appropriate regression with a stepwise option, if available.

Step 6: Examine the *sign* of the parameter estimates. Does it make sense from a business standpoint?

Step 7: Produce the estimates for LGD. If you are using linear regression, then your model may have predicted recovery rates that are negative or greater than 100%. You may want to manually set these equal to 0 and 100, respectively. In linear regression, there is no fancy EXP function needed. You simply multiply the parameters by the value of the variable and add them together along with the intercept. Figure 4 shows an example using linear regression in which the recovery rate for a particular account was calculated to be 58.31%. Therefore, the LGD for this loan would be 1−0.5831, or 41.69%.

Figure 4
Calculating Percent Recovered

A	B	C	D
	Parameter	Value	B*C
Intercept	41.770	N/A	41.77
Var4	-1.700	3	-5.10
Var5	-0.195	36	-7.02
Var8	-0.230	8	-1.84
Var9	30.500	1	30.50
sum			58.31

ONCE YOU'VE COLLECTED THE NECESSARY DATA TO ESTIMATE THESE TYPES OF MODELS, YOU WILL BE WELL UNDER WAY TO USING A MORE RISK-SENSITIVE APPROACH TO CAPITAL REQUIREMENTS—AN APPROACH THAT IS HOPEFULLY IN YOUR FAVOR.

Summary

Well, there you have it. You have been spared the statistical detail behind these modeling approaches—their assumptions, derivations, mathematical distributions, and words like *heteroscedasticity* and *multicollinearity*. Once you've collected the necessary data to estimate these types of models, you will be well under way to using a more risk-sensitive approach to capital requirements—an approach that is hopefully in your favor. In the next article, we will focus on model accuracy and the validation requirements needed to support Basel II. ❐

Contact Morrison at at
Jeff.Morrison@suntrust.com

Notes

1 Moody's Investor Service, Global Credit Research, Special Comments, November 2000.

2 If using tobit regression, see William H. Greene. *Econometric Analysis,* 2nd edition, 1993. Macmillan Publishing Company, New York, NY. This is useful as the prediction formula is more complex.