

# Preparativos para Atender aos Requisitos de Modelagem do Novo Acordo da Basileia

## Parte 1: Desenvolvimento de Modelo

A experiência de um banco regional pode vir a ser muito útil. Neste primeiro artigo de uma série de quatro, Jeffrey S. Morrison discute a abordagem do SunTrust Bank à modelagem estatística. A Parte II trata em detalhes dos passos para validar o modelo. A Parte III reúne tudo em uma interface de software GUI. Finalmente, a Parte IV ingressa nos domínios dos testes de estresse.

O Novo Acordo de Capital da Basileia, cuja adoção está prevista para 2007, estabelece requisitos analíticos detalhados para a avaliação de risco, baseados em dados coletados pelos bancos durante todo o ciclo de vida do empréstimo. O objetivo do Novo Acordo da Basileia é introduzir uma estrutura de capital mais sensível ao risco, incentivando as boas práticas de gerenciamento de risco. Muitos bancos estão estudando ou implantando modelos de apoio para o gerenciamento de risco, mas o processo é complexo.

Jeffrey S. Morrison

## Preparing for Basel II Modeling Requirements

### Part 1: Model Development

*Gain the benefit of a regional bank's experience. In this first of four articles, Jeff Morrison discusses SunTrust Bank's 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 realm 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 to help enhance their risk management efforts. And it can get pretty confusing.*

## 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 article can serve as a refresher.

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 and 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, al-

## Modelos

Alguém ainda se lembra do livro de estatística usado na faculdade e o que pensava dele? "Nunca vou usar estes ensinamentos na prática!" Jamais se deve fazer essa afirmação. O seu velho livro mais este artigo podem servir como revisão.

Vamos começar pela definição da palavra *modelo*. A definição mais estatística encontrada no dicionário Webster é "... um sistema de postulados, dados e inferências apresentado como descrição matemática de uma entidade ou de um estado de coisas". Basicamente, os modelos devem ser considerados como representações matemáticas da realidade. Porém a representação não será perfeita mas, com certeza, simplista. Entretanto seu objetivo é proporcionar um *insight* sobre o comportamento, de maneira a permitir previsões razoavelmente precisas e que apontem na direção certa.

Modelos quantitativos vêm sendo usados há muitos anos no crédito ao consumidor. Aqueles desenvolvidos com base nos dados de solicitação de abertura de novas contas são chamados de modelos de linha de frente, ou de *solicitação*. Esses modelos não empregam o histórico de pagamentos, que o credor tem sobre o possível devedor, simplesmente porque esses dados ainda não existem. Quando essas contas começarem a amadurecer, diferentes modelos podem ser desenvolvidos para produzir *scores* comportamentais, ou seja, algoritmos concebidos para incluir tan-

Banks are implementing models to help enhance their risk management.

Os bancos estão implantando modelos de apoio para o gerenciamento de risco.

to o histórico de pagamentos quanto outros fatores associados à origem do empréstimo e aos dados geográficos e demográficos do devedor. Por outro lado, *scores* desenvolvidos com base em *pools* de dados, normalmente obtidos de *credit bureaus*, são chamados de modelos *genéricos*. Esses modelos refletem o comportamento em crédito em uma grande variedade de instituições financeiras e se baseiam na premissa de que o consumidor apresentará comportamento associado a algum nível médio de risco. *Scores* customizados desenvolvidos com o histórico de pagamentos de *uma única* instituição podem, muitas vezes, apresentar desempenho superior ao dos modelos genéricos por que são feitos sob medida para cada emitente de crédito individual.

**Modelos para o Acordo da Basileia.** Modelos similares podem ser desenvolvidos para o Acordo da Basileia. Os modelos usados no Sistema de *Rating* de Risco do SunTrust foram construídos especificamente para o Novo Acordo sobre uma estrutura bidimensional. A primeira dimensão reflete a probabilidade de inadimplência (PI) do devedor. A segunda, a perda em caso de inadimplência (PCI) associada a um empréstimo ou a uma linha de crédito. Assim, para cada empréstimo, a perda esperada em dólares é, simplesmente, o produto da Exposição em Dólares no Momento da Inadimplência X PI X PCI.

Vamos começar com o desenvolvimento de um

*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 payment history of a single institution 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 X PD X 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*

*Modelos de julgamento subjetivo são simples conjuntos de regras.*

*Judgmental models are simply a set of rules.*

*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 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.*

*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.*

modelo de PI do devedor e, depois, passar para o desenvolvimento de um modelo de PCI, baseado numa linha de crédito. Podemos construir esses modelos para o lado comercial da atividade, mas, para simplificar, daqui por diante vamos pensar em termos de carteiras de varejo, como hipotecas residenciais.

Em geral, os modelos que os bancos criam para os requisitos da Basiléia são de dois tipos — comprados de fornecedores e customizados. No mundo do crédito comercial, os modelos podem precisar ser comprados, porque só os fornecedores desses modelos investiram os recursos necessários para coletar dados suficientemente robustos para fins de modelagem. Isso porque o número de inadimplências comerciais, de qualquer banco individual num ano qualquer, é muito pequeno. Por causa do enorme volume de empréstimos, o varejo se presta muito mais à modelagem customizada, em que cada banco pode usar seus próprios dados, sem depender dos custosos fornecedores. Mesmo que um banco ainda não tenha dados históricos suficientes para desenvolver um modelo estatístico, pode começar com um modelo baseado no bom senso e no julgamento subjetivo, até que os mais sofisticados estejam disponíveis.

Os modelos de julgamento subjetivo são simples conjuntos de regras que quantificam premissas adotadas sobre o nível de risco da carteira, sem uso de abordagens estatísticas. Por exemplo, um mapeamento dos graus de risco, segundo índices de empréstimo/valor (LTV) ou dívida/renda. Outros modelos poderiam fornecer um mapeamento das faixas de *score* FICO em relação à PI. Embora os modelos de julgamento subjetivo tenham sua razão de

ser, o restante deste artigo tratará do desenvolvimento de modelos estatísticos, refletidos nos esforços tanto customizados quanto oferecidos pelos vendedores. Como o Acordo exige que todos os empréstimos sejam submetidos a *rating* por esses modelos, por um prazo mínimo, antes que a abordagem avançada possa ser usada, a integração das soluções de fornecedores e customizadas ao processo deve começar o quanto antes.

A atual escola de pensamento, com referência aos modelos mencionados no Acordo, sustenta que os bancos devem ter modelos separados para o devedor e para a linha. O modelo do devedor deve prever a PI — normalmente definida como inadimplência superior a 90 dias, ou presença de execução, falência, baixa, busca e apreensão ou reestruturação. Os modelos do lado da linha devem prever a PCI, ou 1 menos a taxa de recuperação. A taxa de recuperação é simplesmente o montante recuperado em dólares, dividido pelo montante devido no momento da inadimplência.

**Deixe a diversão começar!** Como veremos, as abordagens estatísticas associadas aos modelos de PI e PCI são bem diferentes. Todavia, nos antecipando, aqui vão algumas definições simples.

>Variável dependente — a variável que se deseja prever (inadimplência/adimplência ou porcentagem recuperada).

>Variáveis independentes — as variáveis explicativas (LTV, encargo da dívida, etc.) usadas para explicar a variável dependente.

>Correlação — um número entre -1 e 1 que mede o grau de relação linear entre duas variáveis. Quanto mais próxima de +1 ou -1, mais alta a correlação.

>Análise de regressão — uma família de procedimentos estatísticos que quantificam a relação en-

*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 PD and LGD models 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 y-intercept.*

*Regardless of the type of regression you use, all approaches allow you to determine which inde-*

pendent 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 a 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—

tre a variável dependente e um conjunto de variáveis independentes, através de dados históricos. Há muitos tipos de regressão.

>Estimativas de parâmetros — o conjunto de ponderações produzido pela regressão usada na previsão. Usa-se uma ponderação para cada variável independente, mais um valor constante, por vezes chamado de *interseção y*.

**Regression analysis is the primary mechanism for building statistical models.**

**O principal mecanismo para a construção de modelos estatísticos é a análise de regressão.**

Independentemente do tipo de regressão utilizado, todas as abordagens permitem determinar quais variáveis independentes o modelo deverá incluir ou desconsiderar. Ao incluir uma variável explicativa num modelo de regressão, normalmente se obtém uma estimativa de parâmetro. Contudo, dado um determinado nível de precisão, esta estimativa pode não ser significativamente diferente de zero e, nesse caso, não deve ser usada. A maioria dos pacotes de regressão produz uma medida chamada *estatística t*, que indica se uma variável deve ou não ser excluída do modelo. Esta é uma das principais vantagens do uso de uma regressão. A modelagem não é uma ciência exata e, como os estatísticos

vêm de um amplo espectro de formações e experiências, são possíveis diversas abordagens ou desenhos de modelagem, que podem funcionar muito bem. Ainda assim, o objetivo deste artigo é oferecer conselhos genéricos, ou regras de bolso que podem ser usadas para adiantar o processo de modelagem em sua instituição financeira.

## Modelos do Devedor: Probabilidade de Inadimplência

Como a análise de regressão é o principal mecanismo para a construção de modelos estatísticos, vamos começar por ela. Há muitos tipos de procedimentos de regressão. Para prever a probabilidade de inadimplência, costuma-se recomendar a regressão logística. Esta é apropriada nos casos em que a variável dependente é binária — assumindo um dentre dois valores. Nesse caso, a variável dependente indica se o empréstimo entrou ou não em inadimplência, num determinado período de tempo — geralmente um ano. Se o objetivo é tentar prever a probabilidade de inadimplência, a variável dependente deve ter valor 1 (para inadimplência) ou 0 (para adimplência). A maioria dos pacotes de *software* estatístico faz esse tipo de regressão com facilidade.

A regressão logística permite fazer coisas interessantes. Primeiro, os valores previstos pela regressão já vêm na forma em que precisamos deles — como probabilidades limitadas entre 0 e 1. Assim, se temos um valor previsto de 0,356, então a probabilidade de inadimplência para o empréstimo em questão no decorrer dos próximos 12 meses será de 35,6%. Em segundo lugar, a regressão logística tem flexibilidade para captar relações não-lineares, como o LTV.

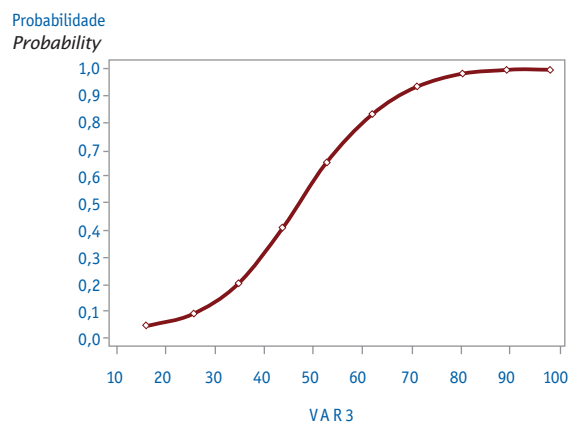
Dependendo dos dados, a relação entre o LTV e a probabilidade de inadimplência pode ser não-linear, ou tem forma de S. A parte superior da Figura 1 mostra que a variável independente VAR3 (que pode ser encarada como o LTV) tem inclinação mais acentuada, no valor de 50, fazendo com que o modelo seja mais sensível nessa faixa. Isso se encontra demonstrado na parte inferior do gráfico,

Figura 1

Figura 1

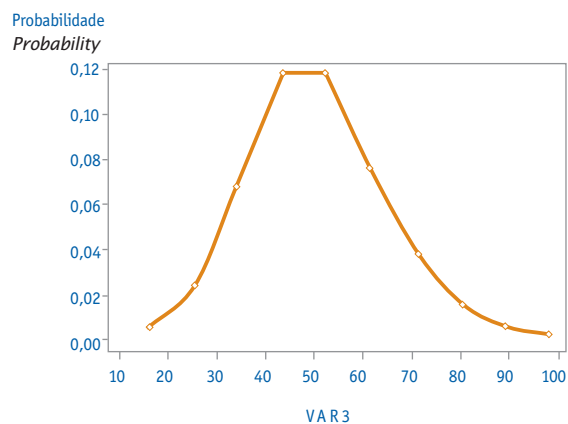
### Análise de Inclinação Não-Linear

#### Nonlinear Slope Analysis



### Sensibilidade (variação de 10%)

#### Sensitivity (10% change)



Fonte: SunTrust Bank, Inc.  
Source: SunTrust Bank, Inc.

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 pack-

ages 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 probability of default can be non-linear 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 that the probability of default plateaus as it approaches 0.90, 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,*

*onde uma variação de 10% da VAR3, em torno de 50, leva a uma variação de 12% da probabilidade de inadimplência — em igualdade das demais condições. Variações em torno dos valores da VAR3, que sejam muito maiores ou menores do que 50, tenderiam a ter impacto significativamente menor sobre a probabilidade de inadimplência. Observe, na parte superior do gráfico, que a probabilidade de inadimplência atinge seu pico ao aproximar-se de 0,90, onde o LTV é próximo de 80.*

Muito bem. Armados de um pouco de conhecimento sobre a regressão logística, vamos preparar uma *checklist* para a construção de um modelo. No que se refere à carteira usada como exemplo, pense nela em termos de hipotecas residenciais.

## Checklist nº 1

**Passo 1:** Defina sua variável dependente. Vamos dizer que seja inadimplência um empréstimo com mais de 90 dias em delinqüência ou que apresente execução, falência, baixa, busca e apreensão ou reestruturação. Vamos codificar uma variável indicativa para isto, usando 1 (inadimplência) ou 0 (adimplência).

**Passo 2:** Defina a janela de aplicação. Esse é o prazo durante o qual o conjunto de contas pode entrar em estado inadimplente. Vamos escolher um ano.

**Passo 3:** Identifique todos os empréstimos que estejam regulares há um ano e acompanhe seu desempenho nos 12 meses seguintes. Se sua carteira for muito grande, pode ser bom usar uma amostra aleatória. O tamanho da amostra pode variar muito. Cerca de 25.000 observações podem ser um bom volume para construir o modelo e para testá-lo podem ser necessárias mais 25.000. Mas

ter um número suficiente de inadimplências é muito importante — quanto mais, melhor. Aplique o indicador do Passo 1 e chame-o de variável dependente.

Passo 4: No começo da janela de aplicação (ano anterior), selecione variáveis relevantes que, em sua opinião, possam prever a inadimplência nos 12 meses seguintes — LTV, a idade do empréstimo, o tipo do empréstimo, número de vezes que o empréstimo ficou 30/60 dias em atraso, etc.

Passo 5: Avalie graficamente os dados por meio de contagens de frequência, médias, mínimos, máximos e correlações. Normalmente, é possível examinar seus dados de todas essas perspectivas por meio de apenas alguns comandos da maioria dos pacotes de *software*. Como não é desejável incluir na regressão duas variáveis que reflitam informações equivalentes (ou seja, excessiva correlação entre si), procure por essas candidatas. Veja quais variáveis são as mais correlacionadas com a variável dependente. Procure por dados estranhos — observações com valores extremamente altos ou baixos.

Passo 6: Importante — *trate dos dados extraviados*. Determine, para cada variável, a porcentagem de dados desaparecidos. De maneira geral, se a porcentagem de valores perdidos, para uma variável específica, for superior a 30%, não a utilize. Embora este seja um valor de corte arbitrário, a idéia é procurar por variáveis que contenham suficientes informações. Com referência às variáveis que restarem, use a média dos valores de que dispõe para substituir a pequena quantidade de informação ausente. Se não forem tomadas medidas para lidar com os dados que faltam, o *software* de regressão saltará automaticamente os registros afetados. Você

*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 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 may be predictive of default over the next 12 months — LTV, the 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 cutoff 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 were 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 estimate. Does it make sense from a business standpoint? A negative sign means that 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 values. 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 you model. Often this is done for you automatically. However, to show how it works, look at Figure 2. This is an example code in SAS® that uses your parameter estimates to produce default probabilities.*

pode evitar o fato, eliminando da análise a maior parte dos seus dados.

**Passo 7:** Estime seu modelo rodando uma regressão logística, com a opção “stepwise”. Essa função removerá automaticamente quaisquer variáveis que não sejam estatisticamente significativas. O software faz todo o trabalho por você.

**Passo 8:** Examine o sinal da estimativa de parâmetro. Ele faz sentido do ponto de vista dos negócios? Um sinal negativo significa que há uma relação inversa, entre a variável e a probabilidade de inadimplência. Um sinal positivo significa que a probabilidade de inadimplência aumenta com o valor da variável. Não aceite apenas resultados da regressão como certos. Analise-os. Se o sinal vai contra sua intuição, analise novamente os dados para descobrir o porquê.

**Passo 9:** Produza as probabilidades previstas com seu modelo. Em muitos casos, isso é feito automaticamente. Entretanto, para entender como funciona, veja a Figura 2. Trata-se de um exemplo de código do SAS® que usa suas estimativas de parâmetros para produzir probabilidades de inadimplência.

Figure 2

Figura 2

### Código de Implementação

#### Implementation Code

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

Esse exemplo é chamado de código de implementação. Neste caso, usamos as médias de VAR2 – VAR4 para substituir as informações perdidas (representadas como “=”). Além disso, lembre-se de incluir verificações numéricas, para o caso de valores inválidos acabarem entrando no código. A função matemática “EXP”, no final, é o que converte a resposta numa probabilidade. A variável, chamada de HSCORE, seria a PI estimada. Admitindo que tenhamos valores válidos para VAR2 – VAR 4 e não haja informações extraviadas, os cálculos podem ser facilmente realizados no Excel, como mostra a Figura 3. A conta do exemplo tem probabilidade de inadimplência, em um ano, de 23,6%.

*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 into 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 – VAR 4 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%.*

Figura 3

### Cálculo da Probabilidade de Inadimplência

A	B	C	D
	Parâmetro	Valor	B x C
Interseção	1,8538	N/D	1,8538
Var2	-0,145	20	-2,9
Var3	0,10819	42	4,54398
Var4	-1,557	3	-4,671
Soma			-1,17322
Aplicação da fórmula exponencial à soma: Probabilidade = 1/ (1+exp(-soma))			0.236273447

Figure 3

### Calculating Probability of Default

A	B	C	D
	Parameter	Value	B x C
Intercept	1,8538	N/D	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

### Modelo da Linha: Perda em Caso de Inadimplência

Segundo a Moody’s Investor Service, “não há uma boa estrutura para prever o resultado da inadimplência. Essa deficiência é tão grande porque os resultados possíveis da inadimplência são muito diversos. Um empréstimo inadimplente pode render essencialmente 100%, inclusive os juros acumulados, ou pode pagar apenas cinco

### Facility Model: Loss Given Default

*According to Moody’s Investor Service, “there is no good framework for predicting the outcome of 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 be the next month or it might take four and one-half years.”<sup>1</sup>*

*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 the 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.*

cents por dólar. Uma resolução pode levar de um mês a quatro anos e meio”<sup>1</sup>.

Em outras palavras, é difícil construir um modelo de recuperação do ponto de vista das perdas, especialmente do lado do crédito comercial — muito mais do que construir um modelo de probabilidade de inadimplência. Isso porque não é fácil conseguir dados preditivos em quantidade suficiente. A precisão obtida com o uso de uma abordagem estatística qualquer é secundária, em relação à obtenção de uma quantidade suficiente de dados de boa qualidade. Como há muito poucas inadimplências comerciais, o prazo necessário para coletar dados de inadimplência pode ser considerável. Por outro lado, no varejo, você deve ser mais bem sucedido por causa da abundância de dados de inadimplência. Agora examinaremos dois tipos de abordagem estatística recomendados para estimar a perda em caso de inadimplência, ou 1 menos a taxa de recuperação.

Lembre-se de como tivemos que coletar informações, tanto sobre os empréstimos inadimplentes quanto sobre os adimplentes? Ao construir um modelo de recuperação, nos concentramos apenas nas informações ligadas às inadimplências. Por exemplo, digamos que você tenha coletado as seguintes informações sobre inadimplências em sua carteira de hipotecas residenciais.

- >Porcentagem de dólares recuperados.*
- >Região do U.S. Census/localidade/código postal /MSA.*
- >Idade do empréstimo quando da inadimplência.*
- >LTV.*
- >Indicador de tipo de falência.*
- >Há quanto tempo está em cobrança.*
- >Idade do imóvel.*

>Variação do valor do imóvel em comparação com o ano anterior.

>Renda familiar média da localidade.

>Score FICO.

>Índice de indicadores econômicos antecipados.

>Saldo devedor.

Como a variável independente (a taxa de recuperação) não é binária (0 ou 1) em nosso modelo de inadimplência, a regressão logística não é a abordagem apropriada. A taxa de recuperação varia de 0% a 100%, dependendo de como são contabilizados os encargos e as tarifas. Dado o formato da distribuição para este tipo de dado, costumam ser usadas duas abordagens estatísticas — regressão linear e regressão *tobit*.<sup>2</sup>

A regressão linear é, talvez, o tipo mais popular de regressão e usa o método dos mínimos quadrados para calcular as ponderações da equação preditiva. O nome diz tudo: essa técnica produz uma linha que minimiza o quadrado da diferença entre os valores reais e previstos. A regressão linear só é capaz de estimar uma relação linear, entre a variável independente e a taxa de recuperação. Infelizmente, mesmo que a reação seja efetivamente não-linear (em forma de S ou de U), a regressão linear só proporciona uma aproximação linear dessa curva. Isso não deve ser um grande problema, já que os bons estatísticos conhecem truques que lhes permite contornar esta limitação.

A regressão *tobit* pode ser encarada como um híbrido, entre um modelo de regressão linear e a irmã gêmea da regressão logística, a regressão *probit*. A regressão *probit* é semelhante à logística, mas se baseia numa distribuição em S, um pouco diferente. A vantagem da regressão

*>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 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 regression.<sup>2</sup>*

*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 logistic regression's twin brother, probit regression. Probit regression is similar to its logistic sibling, but is based on a slightly different S-shaped dis-*

*tribution. 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 percentage 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*

*tobit, sobre os demais métodos, é o fato de que ela foi criada para lidar com casos em que a variável dependente se aglomera em torno de limites como 0. Se houver muitas observações em que a porcentagem recuperada tenha sido 0 (como no caso de cartões de crédito de consumidores), usar uma regressão linear para estimar o modelo pode produzir resultados distorcidos e menos precisos.*

*Armado dessas informações, você agora está pronto para construir seu modelo de recuperação. A boa notícia é que é possível eliminar um ou dois passos em relação ao que fizemos com o modelo de PI. Por exemplo, não é preciso haver preocupação com a janela de aplicação porque só estamos lidando com empréstimos inadimplentes. Então aqui vai sua segunda checklist:*

## **Checklist Nº 2**

*Passo 1: Defina a variável dependente — porcentagem de dólares recuperados. Como a operação de recuperação pode ser um processo que perdura por um longo prazo, parte do processo de definição é estabelecer um prazo-limite para as transações de recuperação. Por exemplo, se você não recebeu mais nada, dois anos depois da inadimplência, pode admitir que o processo de cobrança esteja completo.*

*Passo 2: No momento da inadimplência, some as variáveis explicativas que podem ser preditivas.*

*Passo 3: Analise graficamente os dados por meio de frequências, médias, mínimos, máximos e correlações.*

*Passo 4: Importante — Trate os dados ausentes. Para qualquer variável, determine a porcentagem de dados ausentes.*

*Passo 5: Estime seu modelo rodando a regres-*

são apropriada com a função “stepwise”, caso haja.

Passo 6: Examine o  *sinal* das estimativas de parâmetros. É racional sob o ponto de vista do negócio?

Passo 7: Produza as estimativas de PCI. Se estiver usando regressão linear, seu modelo pode ter previstas taxas de recuperação negativas ou maiores do que 100%. Pode ser bom determinar manualmente que estas sejam iguais a 0 e 100, respectivamente. Na regressão linear, não há necessidade da função EXP. Basta multiplicar os parâmetros pelo valor da variável e somá-los uns com os outros e com a interseção. A Figura 4 mostra um exemplo que usa regressão linear, em que a taxa de recuperação de uma conta específica foi calculada em 58,31%. Assim, a PCI deste emprestimo seria de  $1 - 0,5831$ , ou 41,69%.

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%.

Figura 4

### Cálculo da Porcentagem Recuperada

A	B	C	D
	Parâmetro	Valor	B x C
Interseção	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
Soma			58,31

Figure 4

### Calculating Percent Recovered

A	B	C	D
	Parameter	Value	B x 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

## Sumário

Pois bem. Você foi poupado dos detalhes estatísticos que envolveram as abordagens de modelagem — suas premissas, derivações, distribuições matemáticas e palavras como *heteroscedasticidade* e *multicolinearidade*. Uma vez coletados dados suficientes para estimar

## Summary

Well, there you have it. You have been spared the statistical details behind these modeling approaches — their assumptions, derivations, mathematical distributions, and words like *heteroscedasticity* and *multicollinearity*. Once you’ve collected the necessary data to es-

*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.*

## 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. The reference is useful, as the prediction formula presented is more complex.

---

©2003 RMA. Jeff Morrison is vice president, Credit Metrics – PRISM Team, at SunTrust Banks Inc., Atlanta, Georgia. Contact Morrison at: [Jeff.Morrison@suntrust.com](mailto:Jeff.Morrison@suntrust.com) RMA - Risk Management Association is an international association of financial services professionals. For membership information, e-mail [acauley@rmahq.org](mailto:acauley@rmahq.org); to subscribe to The RMA Journal, visit [www.rmahq.org/Ed\\_Opps/pubs/journalad.htm](http://www.rmahq.org/Ed_Opps/pubs/journalad.htm)

esses tipos de modelo, você estará preparado para o uso de uma abordagem aos requisitos de capital mais sensível ao risco que, esperamos, lhe será favorável. No próximo artigo, nos concentraremos na precisão do modelo e nos requisitos de validação necessários para o Novo Acordo da Basileia.

## Notas

1. Moody's Investor Service, Global Credit Research, *Special Comments*, Novembro de 2000.

2. Se utilizar a regressão *tobit*, ver William H. Greene, *Econometric Analysis*, 2ª edição, 1993, Macmillan Publishing Company, New York, NY. A referência é útil, uma vez que a fórmula de previsão apresentada é mais complexa.

---

©2003 RMA. Jeff Morrison é vice-presidente de Medidas de Crédito — Equipe PRISM, do SunTrust Banks Inc., Atlanta, Georgia. Entre em contato com Morrison no endereço [Jeff.Morrison@suntrust.com](mailto:Jeff.Morrison@suntrust.com) A RMA - Risk Management Association é uma associação internacional de serviços financeiros profissionais. Para informações, e-mail [acauley@rmahq.org](mailto:acauley@rmahq.org); Para assinar The RMA Journal visite o site [www.rmahq.org/Ed\\_Opps/pubs/journalad.htm](http://www.rmahq.org/Ed_Opps/pubs/journalad.htm)