RECONHECIMENTO E CLASSIFICAÇÃO DE PADRÕES: APLICAÇÃO DE METODOLOGIAS ESTATÍSTICAS EM CRÉDITO AO CONSUMIDOR

Inácio Andruski Guimarães*
Anselmo Chaves Neto**

RESUMO: A inadimplência é um dos maiores problemas, senão o maior, enfrentado pelas administradoras de cartão de crédito. No estudo deste problema foi criado o conceito de risco, que é essencialmente a probabilidade de não recebimento dos créditos por parte das administradoras de cartões. Alguns autores, CAOUETTE et al. (2000), e SILVA (1988), referem-se às técnicas estatísticas multivariadas como ferramentas poderosas na administração do risco envolvido na concessão de crédito pessoal. Este trabalho apresenta a construção e avaliação de regras de reconhecimento e classificação de padrões baseadas em duas técnicas multivariadas: a Função Discriminante Linear de Fisher (FDL) e a Regressão Logística (RL), para classificar clientes de cartão de crédito em um de dois grupos. A eficiência dos procedimentos é avaliada por meio do Método de Lachenbruch, Lachenbruch (1975).

PALAVRAS-CHAVE: Função Discriminante Linear de Fisher, Regressão Logística, Decisão de crédito.

ABSTRACT: The non-payment (breach of contract) is one of the major, if not the major, problem faced by administrators (companies, agencies) of credit. In studies of such problems it was created a risk concept, that is essentially the probability of not receiving the credits from the administrators. Some authors, Caouette et al. (2000) and Silva (1988), refer the multivariate analysis as a very powerful tool in the risk administration of conceding the personal credit. This work show the build and the evaluation of pattern recognition and classified rules based on the Discriminant Linear Function (DLF) and on the Logistic Regression (LR), to classify the clients of credit card in one of two groups. The efficiency of the procedures was evaluated by the Lachenbruch Method, Lachenbruch (1975).

* O autor é M.Sc.PPGMNE/UFPR-CEFET-PR, inacio@fesppr.br.
** O autor é Dr PPGMNE/DEST/UFPR, anselmo@est.ufpr.

KEY WORDS: Discriminant Linear Function, Logistic Regression, Credit Decision.

1. Introdução

1.1 Problema

A palavra "crédito" pode ter mais de um significado dependendo do contexto sob o qual esteja sendo tratada. Sob o ponto de vista, meramente empresarial, a concessão de crédito significa a transferência da posse de um bem ou de uma quantia em dinheiro, mediante a promessa de pagamento futuro. De acordo com este conceito, pode-se entender o crédito como a disposição de uma pessoa, física ou jurídica, com capacidade em obter dinheiro, produtos ou serviços, mediante compromisso de pagamento num determinado período de tempo.

Uma das maiores revoluções no crédito pessoal foi desencadeada pela criação do cartão de crédito. Também chamado "dinheiro de plástico", que é, antes de mais nada, um instrumento de crédito automático. Este sistema atingiu proporções que tornam obrigatória e permanente busca de técnicas que permitam o gerenciamento de um grande número de *portfólios* de empréstimos aos mais diversificados e descentralizados consumidores, de modo a obter, simultaneamente, um expressivo retorno para a administradora.

Inicialmente, é necessário distinguir, segundo enfoques estatísticos, o que é "risco" e "incerteza". O primeiro, existe quando a tomada de decisões é baseada em probabilidades objetivas para a estimação de diferentes resultados. Desta forma, a expectativa fundamenta-se em dados históricos, permitindo que as decisões sejam tomadas a partir de estimativas consideradas aceitáveis. A segunda, é observada quando não se tem à disposição os dados históricos acima mencionados. Isto exige do tomador de decisões uma certa dose de sensibilidade, baseada em observações altamente subjetivas.

No caso específico do crédito ao consumidor, as características observadas são:

- grande volume em pequenos montantes;
- processo de aprovação massificado;
- dados limitados e relativamente pobres;

- histórico de crédito do cliente disponível, mas, geralmente, incompleto. Na maior parte dos casos, limita-se ao passado negativo ou positivo do cliente;
- utilização de bases estatísticas para avaliação do desempenho do gerenciamento do *portfólio*.

O principal meio de controle do risco ou, pelo menos, o mais utilizado é o sistema de escore. Este sistema consiste, basicamente, em avaliar características do novo cliente, atribuindo um determinado valor a cada característica. Em seguida os dados obtidos são usados na elaboração de um escore. Com base no escore obtido pelo cliente toma-se a decisão de conceder, ou não, o crédito. Para tomar tal decisão, o escore é comparado a um valor previamente estabelecido, chamado *valor de corte*. É na obtenção deste último que reside a maior parte dos problemas enfrentados pelos profissionais envolvidos. A questão a ser resolvida, neste ponto, pode ser colocada da seguinte forma: "como obter um valor de corte confiável a ponto de evitar perdas para a empresa, tanto pela aceitação errada, de maus clientes, como pela rejeição, igualmente errada, de bons clientes?"

1.2 Objetivos

Este trabalho tem os seguintes objetivos:

- utilizar as técnicas estatísticas multivariadas denominadas Função Discriminante Linear de Fisher (FDL), *a* Regressão Logística (RL), e, ainda, o Método de Lachenbruch na identificação de variáveis que permitam evidenciar, com certa antecedência, situações de inadimplência por parte de clientes de uma administradora de cartões de crédito, a partir de informações cadastrais fornecidas pelos mesmos em propostas para adesão ao cartão de crédito administrado pela instituição.
- Programar um sistema de classificação de clientes usando a Função Discriminante Linear de Fisher que indique, por um Modelo de Regressão Logística, a probabilidade estimada de um cliente tornar-se inadimplente.
- Avaliar a eficiência dos modelos desenvolvidos, aplicando o Método de Lachenbruch.

1.3 Justificativa

A Análise Estatística Multivariada tem sido mencionada por alguns autores, como SILVA (1988) e CAOUETTE (2000), como uma ferramenta bastante poderosa na administração do risco de inadimplência existente na concessão de crédito. Uma aplicação é a previsão do risco que corre um banco ou a administradora de crédito, e a conseqüente busca de uma forma de controle deste risco, via obtenção de um valor de corte calculado com base nas características apuradas junto à base de dados da companhia.

2. Técnicas Estatísticas Multivariadas de Reconhecimento de Padrões e de Classificação

2.1 Introdução

As técnicas estatísticas da *Análise Discriminante* e *Regressão Logística* fazem parte do quadro de métodos quantitativos, tidos como os mais eficientes para a auxiliar a tomada de decisões. De acordo com SILVA (1988, p.93), "*O uso da estatística, da teoria das probabilidades, é um valioso instrumento para a tomada de decisão*". Também, segundo o mesmo autor, "*Outros recursos, como o Teorema de Bayes, Análise Fatorial e Pesquisa Operacional, por exemplo, têm sua aplicabilidade nas áreas de crédito*". Quando o analista de crédito tem a sua disposição uma regra de reconhecimento de padrões e classificação que indique, previamente, a chance de inadimplência de um futuro cliente, a decisão de concessão de crédito fica extraordinariamente facilitada. Dessa forma, esse profissional pode, então, incorporar argumentos quantitativos a outros subjetivos e decidir, com maior confiança e segurança, a concessão ou não do crédito.

2.2 Análise Discriminante

A Análise Discriminante é uma técnica estatística usada na resolução de problemas que envolvem a ***separação*** de conjuntos distintos de objetos ou observações e a ***alocação*** de novos objetos ou observações em um grupo específico. Integra o conjunto de

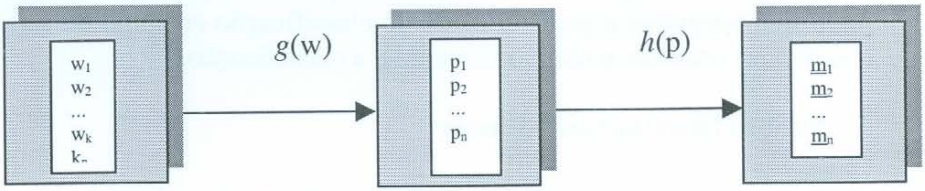
técnicas usadas no Reconhecimento de Padrões, juntamente com técnicas de Programação Matemática e, mais recentemente, Redes Neurais. O reconhecimento de padrões, de um modo geral, está presente em áreas tais como:

- classificação de empresas;
- processamento de sinais;
- análise de sinais eletrocardiográficos;
- reconhecimento de impressões digitais;
- elaboração de perfis de consumidores;
- diagnóstico médico preliminar; entre outras.

Um dos objetivos da Análise Discriminante é determinar a que grupo, dentre dois ou mais definidos, *a priori*, pertence um novo elemento, com base em características observadas para o mesmo. Cada característica constitui uma variável independente, contribuindo para a classificação. A Análise Discriminante combina estas variáveis em uma ou mais funções, de modo a determinar para cada elemento escores de classificação com base em um banco de dados dos grupos de indivíduos pré-definidos. Neste trabalho, os grupos definidos *a priori* são dois: um de "bons" clientes e outro de "maus" clientes, assim chamados os clientes com pagamento em dia e clientes em atraso, respectivamente. O problema básico no reconhecimento de padrões pode ser apresentado da seguinte forma: dado um vetor de dimensão n de medidas, $\underline{m}_i$, obter um método de inversão do mapeamento nas relações g e h, de modo a identificar a classe geradora das medidas a partir de $\underline{m}_i$. Este raciocínio é ilustrado na Figura 1.

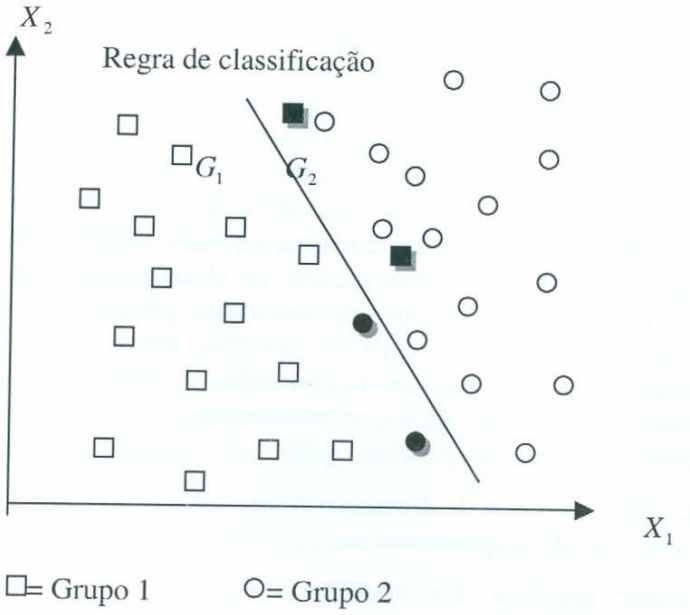
Espaço de classes (C), Espaço de padrões (P), Espaço de medidas (M).

Figura1 – Mapeamento das Relações g e m



Sejam, por exemplo, dois grupos de observações resultantes de classificação segundo um critério (portadores e não portadores de uma enfermidade, clientes adimplentes e inadimplentes, etc.) e, ainda, as variáveis X_1 e X_2 observadas para cada indivíduo pertencente a um dos grupos. Na Figura 2, tem-se o espaço discriminante.

Figura 2 – Espaço Discriminante



Pode-se observar que a regra de classificação, $f(X_1,X_2)$, separa os pontos em dois grupos de forma razoável, porém, há uma "mistura" de pontos. Esta "mistura" mostra a ocorrência de erros nas classificações. Desta forma, o objetivo é a obtenção de uma regra que minimize a probabilidade de classificação errônea, ou seja, que busque otimizar o reconhecimento e a classificação.

2.3 Função Discriminante Linear

Uma função discriminante linear tem a forma:

$$f(X) = Z = \beta_0 + \beta_1 X_1 + ... + \beta_i X_i + ... + \beta_{p-1} X_{p-1} = \beta_0 + \sum_{i=1}^{p-1} \beta_i X_i \quad (2.01)$$

onde f(X) = Z é o escore discriminante (variável dependente) e β_i , i = 0,1, ... ,p são os coeficientes da função nas variáveis X_i , i = 1,2, ..., p.

A função f(X) retorna um valor Z, para um novo padrão m_i, que funcionará como escore de classificação deste padrão. Assim, sejam Π_1 e Π_2, duas populações distintas sob um determinado critério, sendo X um vetor aleatório de dimensão p composto por medidas das características dos membros das populações (grupos) e com $f_1(\underline{X})$ e $f_2(\underline{X})$ as f.d.p's associadas. Sejam R_1, o conjunto de valores para os quais o elemento é classificado como pertencente a Π_1, e R_2, o conjunto de valores para os quais o elemento é classificado como pertencente a Π_2 e com $R_1 \cap R_2 = \varnothing$, dentro de um espaço amostral $\Omega = R_1 \cup R_2$ que corresponde ao conjunto de todas as possíveis observações X.

Na classificação para duas populações (1 e 2), considera-se as probabilidades de classificação errônea:

$$P(2|1) = P(\underline{X} \in \Pi_2 | \Pi_1) = \int_{R_2} f_1(\underline{X}) d\underline{X} \quad (2.02)$$

$$P(1|2) = P(\underline{X} \in \Pi_1 | \Pi_2) = \int_{R_1} f_2(\underline{X}) d\underline{X} \quad (2.03)$$

E as probabilidades de ocorrência *a priori* p_1 em Π_1 e p_2 em Π_2, tais que $p_1 + p_2 = 1$. É comum que regras de reconhecimento sejam avaliadas em termos de probabilidades de reconhecimento errado, conforme a Tabela 1, a seguir:

Tabela 1 – Matriz do Custo De Reconhecimento Errado

	RECONHECIDA COMO SENDO DE	
POPULAÇÃO VERDADEIRA	Π_1	Π_2
Π_1	0	C(2|1)
Π_2	C(1|2)	0

O custo esperado de reconhecimento errado (ECM) é dado por

$$ECM = c(2|1)p(2|1)p_1 + c(1|2)p(1|2)p_2 \qquad (2.08)$$

E uma boa função de reconhecimento deve ter um ECM mínimo. As regiões que minimizam ECM são definidas pelos valores de X que tornam válidas as desigualdades:

Para $$R_1 \therefore \frac{f_1(\underline{X})}{f_2(\underline{X})} \geq \frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1} \quad \Rightarrow \quad R_D \geq R_C.R_P$$

Para $$R_2 \therefore \frac{f_1(\underline{X})}{f_2(\underline{X})} < \frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1} \quad \Rightarrow \quad R_D < R_C.R_P \qquad (2.09)$$

onde R_D é a razão das densidades, R_C é a razão dos custos e R_P é a razão das probabilidades *a priori*.

A ***Função Discriminante Linear de Fisher*** foi o primeiro método estatístico de discriminação e apresenta boas propriedades para reconhecer padrões e classificá-los. No caso de duas populações, o ponto de partida de *Fisher* foi a transformação das observações multivariadas X's em observações univariadas y's , tais que as observações de cada uma das populações Π_1 e Π_2 sejam tão separadas quanto possível. A idéia mestra consiste em tomar a combinação linear de X para obter y. Assim, dadas as médias μ_{1y} e μ_{2y} dos y's obtidos a partir dos $\underline{X}$'s pertencentes a Π_1 e a Π_2, respectivamente, seleciona-se a combinação linear que maximiza a distância quadrática entre as médias dadas, com relação à

variabilidade dos *y*'s.

Sejam

$\underline{\mu}_1 = E(\underline{X} \mid \Pi_1)$ = Valor esperado de uma observação multivariada de Π_1 (2.10)

$\underline{\mu}_2 = E(\underline{X} \mid \Pi_2)$ = Valor esperado de uma observação multivariada de Π_2 (2.11)

$\Sigma = E(\underline{X} - \underline{\mu}_i)(\underline{X} - \underline{\mu}_i)'$ = Matriz de covariância, que se supõe igual para Π_1 e Π_2 (2.12)

e, também, a combinação linear

$Y = \underline{c}'_{1xp}\, \underline{X}_{px1}$. (2.13)

Substituindo (2.13) em (2.10) e (2.11) tem-se que

$\mu_{1y} = E(Y \mid \Pi_1) = E(\underline{c}'\underline{X} \mid \Pi_1) = \underline{c}' E(\underline{X} \mid \Pi_1) = \underline{c}'\underline{\mu}_1$ (2.14)

$\mu_{2y} = E(Y \mid \Pi_2) = E(\underline{c}'\underline{X} \mid \Pi_2) = \underline{c}' E(\underline{X} \mid \Pi_2) = \underline{c}'\underline{\mu}_2$ (2.15)

Além disso,

$V(Y) = \sigma^2 = V(\underline{c}'\underline{X}) = \underline{c}' V(\underline{X})\underline{c} = \underline{c}'\Sigma\underline{c}$ (2.16)

que, conforme comentário anterior, é a mesma para ambas as populações.

A melhor combinação linear é obtida da razão entre o quadrado da distância entre as médias e a variância de *Y*. Desta forma,

$$\frac{(\mu_{1y} - \mu_{2y})^2}{\sigma_y^2} = \frac{(\underline{c}'\underline{\mu}_1 - \underline{c}'\underline{\mu}_2)^2}{\underline{c}'\Sigma\underline{c}} = \frac{\underline{c}'(\underline{\mu}_1 - \underline{\mu}_2)(\underline{\mu}_1 - \underline{\mu}_2)'\underline{c}}{\underline{c}'\Sigma\underline{c}} = \frac{(\underline{c}'\underline{\delta})^2}{\underline{c}'\Sigma\underline{c}} \quad (2.17)$$

onde $\underline{\delta} = \underline{\mu}_1 - \underline{\mu}_2$.

A razão (2.17) é maximizada por

$\underline{c} = k\Sigma^{-1}\underline{\delta}$, $\forall\, k \neq 0$ (2.18)

Fazendo $k = 1$, e substituindo em (2.18), tem-se que

$\underline{c} = \Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$ (2.19)

Então

$Y = (\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}\underline{X}$ (2.20)

A expressão (2.20) é conhecida, de acordo com JOHNSON & WICHERN *(1988)*, como **Função Discriminante Linear de Fisher**, e tem a forma (2.01) sem o intercepto (β_0).

Seja, agora, o ponto médio m das médias das duas populações univariadas, obtidas a partir da transformação das populações multivariadas Π_1 e Π_2, e uma observação $\underline{X}_0$, onde

$$m = \frac{1}{2}\left(\mu_{1y} + \mu_{2y}\right) \tag{2.21}$$

Substituindo-se (2.14) e (2.15) em (2.21), obtém-se

$$m = \frac{1}{2}\left(\underline{c}'_1\,\underline{\mu}_1 + \underline{c}'_2\,\underline{\mu}_2\right) \tag{2.22}$$

Então,

$$m = \frac{1}{2}\left[\left(\underline{\mu}_1 - \underline{\mu}_2\right)'\Sigma^{-1}\underline{\mu}_1 + \left(\underline{\mu}_1 - \underline{\mu}_2\right)'\Sigma^{-1}\underline{\mu}_2\right] \tag{2.23}$$

$$m = \frac{1}{2}\left[\left(\underline{\mu}_1 - \underline{\mu}_2\right)'\Sigma^{-1}\left(\underline{\mu}_1 + \underline{\mu}_2\right)\right] \tag{2.24}$$

também,

$$E(y_0 \mid \Pi_1) - m \geq 0$$
$$E(y_0 \mid \Pi_2) - m < 0$$

Desta forma, se $\underline{X}_0 \in \Pi_1$, é de se esperar que y_0 seja no mínimo igual a m. De modo análogo, se $\underline{X}_0 \in \Pi_2$, o valor esperado para y_0 é menor que o ponto médio. Com isso, pode-se expressar a regra de classificação como:

- *Alocar* $\underline{X}_0$ *em* Π_1 *se* $y_0 - m \geq 0$.
- *Alocar* $\underline{X}_0$ *em* Π_2 *se* $y_0 - m < 0$.

Na realidade, os parâmetros $\underline{\mu}_1, \underline{\mu}_2$ e Σ não são conhecidos. Então trabalha-se com os seus estimadores.

Sejam n_1 observações da variável aleatória multivariada X, de dimensão p, que formam a matriz de dados X_1, de ordem n_1 x p, amostra da população Π_1.

$$X_1 = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \dots & \dots & \dots & \dots \\ X_{n_1 1} & X_{n_1 2} & \dots & X_{n_1 p} \end{bmatrix} = \begin{bmatrix} \underline{X}'_{11} \\ \underline{X}'_{21} \\ \dots \\ \underline{X}'_{n_1 1} \end{bmatrix}$$

e n_2 observações da variável aleatória multivariada X, de dimensão p, que formam a matriz de dados X_2, de ordem n_2 x p, amostra da população Π_2.

$$X_2 = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \dots & \dots & \dots & \dots \\ X_{n_2 1} & X_{n_2 2} & \dots & X_{n_2 p} \end{bmatrix} = \begin{bmatrix} \underline{X}'_{12} \\ \underline{X}'_{22} \\ \dots \\ \underline{X}'_{n_2 2} \end{bmatrix}$$

Os estimadores dos parâmetros são:

$$\overline{\underline{X}}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \underline{X}_{i1} \tag{2.25}$$

$$\overline{\underline{X}}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} \underline{X}_{i2} \tag{2.26}$$

$$S_1 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(\underline{X}_{i1} - \overline{\underline{X}}_1\right)\left(\underline{X}_{i1} - \overline{\underline{X}}_1\right)' \tag{2.27}$$

$$S_2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} \left(\underline{X}_{i2} - \overline{\underline{X}}_2\right)\left(\underline{X}_{i2} - \overline{\underline{X}}_2\right)' \tag{2.28}$$

Conforme já se sabe, considera-se a matriz de covariância como sendo a mesma para ambas as populações. Estima-se, então, a matriz de covariância comum, por

$$S_p = \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1).S_1 + (n_2 - 1).S_2\right] \tag{2.29}$$

que, pode-se demonstrar, é um estimador não tendencioso do parâmetro Σ.

Com base no exposto, a **Função Discriminante Linear de Fisher Amostral** pode ser apresentada como

$$y = \left(\overline{\underline{X}}_1 - \overline{\underline{X}}_2\right)' S_p^{-1} \underline{X} \tag{2.30}$$

A estimativa do ponto médio entre as médias amostrais univariadas, $\bar{y}_1 = \underline{c}'\underline{\bar{X}}_1$ e $\bar{y}_2 = \underline{c}'\underline{\bar{X}}_2$, é dada por

$$m = \frac{1}{2}\left(\bar{y}_1 + \bar{y}_2\right) \tag{2.31}$$

ou seja,

$$m = \frac{1}{2}\left[\left(\underline{\bar{X}}_1 - \underline{\bar{X}}_2\right)' S_p^{-1}\underline{\bar{X}}_1 + \left(\underline{\bar{X}}_1 - \underline{\bar{X}}_2\right)' S_p^{-1}\underline{\bar{X}}_2\right] \tag{2.32}$$

$$m = \frac{1}{2}\left(\underline{\bar{X}}_1 - \underline{\bar{X}}_2\right)' S_p^{-1}\left(\underline{\bar{X}}_1 + \underline{\bar{X}}_2\right) \tag{2.33}$$

Desta forma, a regra de classificação é apresentada como

- Alocar $\underline{X}_0$ em Π_1 se $y_0 = \left(\underline{\bar{X}}_1 - \underline{\bar{X}}_2\right)' S_p^{-1}\underline{X}_0 \geq m$
- Alocar $\underline{X}_0$ em Π_2 se $y_0 = \left(\underline{\bar{X}}_1 - \underline{\bar{X}}_2\right)' S_p^{-1}\underline{X}_0 < m$

2.5 Regressão Logística

2.5.1 Introdução

A regressão logística consiste, fundamentalmente, na busca de um modelo que permita relacionar uma variável *Y*, chamada "variável resposta", aos "fatores" $X_1,\ldots,X_{p-1}$, que, supõe-se, influenciam as ocorrências de um evento. A variável resposta deve ser do tipo dicotômica, assumindo apenas os valores 0 ou 1. Neste caso, existe interesse apenas na ocorrência, ou não, do evento em questão. No presente trabalho, 0 (zero) designa "cliente inadimplente" e 1 designa "cliente adimplente".

A situação aqui tratada, variável resposta dicotômica, não recomenda a aplicação do Modelo Linear Geral (MLG), basicamente por dois motivos:

1. O MLG pode gerar para a resposta valores fora do intervalo [0 , 1].
2. A variância dos resíduos não é constante.

Embora siga o mesmo raciocínio da regressão linear, a regressão logística apresenta, com relação à primeira, algumas

diferenças. A primeira diz respeito à relação entre a variável resposta e os fatores. No modelo linear supõe-se que a variável resposta, chamada também de ***dependente***, relaciona-se com os fatores, chamados ainda de variáveis ***independentes***, através do modelo dado por

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{p-1} X_{p-1} + \varepsilon = \beta_0 + \sum_{i=1}^{p-1} \beta_i X_i + \varepsilon \qquad (2.34)$$

que pode ser representado matricialmente como

$$Y = \beta' X + \varepsilon$$

considerando-se n observações de Y e das covariáveis X's. Os β's são parâmetros desconhecidos que devem ser estimados com base nos valores observados para Y e para os X's, e $\varepsilon \sim N_p(\underline{0}, \Sigma)$ é um *ruído aleatório* associado ao modelo.

No modelo logístico, a relação é dada por

$$Y = \frac{e^{\mu}}{1 + e^{\mu}} \qquad (2.35)$$

onde μ é dado por uma expressão da forma $\beta' X$.

Para a variável dependente Y relacionada com uma única variável independente X, ou com várias variáveis X's, a função é chamada *Sigmóide*, e o seu gráfico tem a forma da figura 2.2, a seguir. É fácil perceber que

$$\begin{cases} X \to -\infty \Rightarrow Y \to 0 \\ X \to +\infty \Rightarrow Y \to 1 \end{cases}$$

e, também, que

$$X = 0 \Rightarrow Y = \frac{1}{2}$$

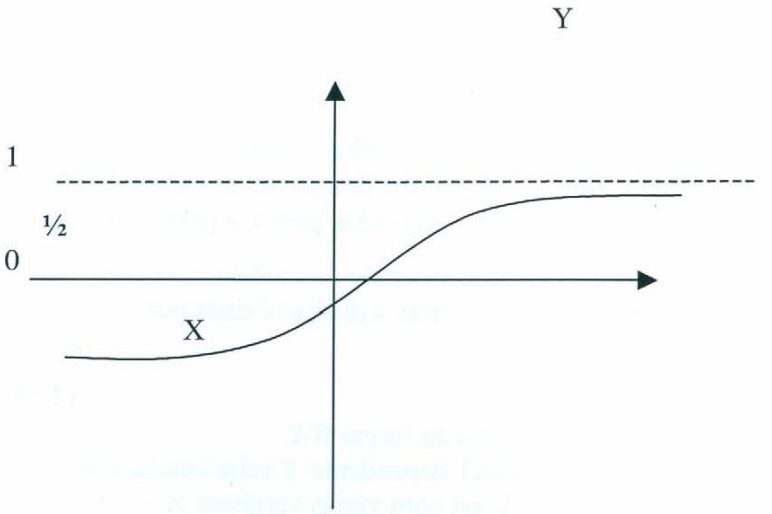
Em qualquer problema de regressão, a quantidade chave é o valor médio da variável dependente, dado o valor da variável independente. Tal quantidade será representada por $E(Y \mid X)$, que se lê "valor esperado para Y, dado o valor de X ". Na regressão linear parte-se da suposição que esta quantidade pode ser expressa por uma expressão da forma (2.34), isto é,

$$E(Y \mid X) = \beta_0 + \sum_{i=1}^{p-1} \beta_i X_i + \varepsilon \qquad (2.36)$$

Esta relação torna admissível a possibilidade de que

$E(Y \mid X)$ possa assumir qualquer valor para $X \in (-\infty, \infty)$.

Figura 3 – Gráfico da Função Sigmóide



Na regressão logística o que se tem é $0 \leq E(Y \mid X) \leq 1$, o que pode ser visualizado na figura 3. A fim de simplificar a notação, a quantidade $E(Y \mid X)$, quando referente ao modelo logístico, será representada por $P(X)$, não havendo, ressalte-se, nenhuma razão específica para o uso desta notação além da já mencionada. Desta forma,

$$P(X) = \frac{e^{\beta_0 + \sum_{i=1}^{p-1} \beta_i X_i}}{1 + e^{\beta_0 + \sum_{i=1}^{p-1} \beta_i X_i}} \qquad (2.37)$$

Uma suposição fundamental para inferências é a de que ε_i, em (2.34), possui distribuição normal com média 0 (zero) e variância constante. Segue que a distribuição da variável dependente é normal, com média $E(Y \mid X)$ e variância constante. Entretanto, isto não ocorre quando a variável dependente é dicotômica. Neste caso, pode-

se expressar o valor da mesma como

$$Y = P(X) + \varepsilon \tag{2.38}$$

onde ε pode assumir um de dois possíveis valores:

- Se Y = 1, então ε = 1 – *P(X)*, com probabilidade *P(X)*.
- Se Y = 0, então ε = - *P(X)*, com probabilidade 1 – *P(X)*.

Contudo, ε tem uma distribuição com média 0 e com a variância dada por

$$P(X)[1 - P(X)]$$

Já foi visto que

$$E(Y_j \mid X) = P(Y_j = 1) = \beta_0 + \sum_{i=1}^{p-1} \beta_i X_i = \theta_j$$

Também é sabido que

$$V(Y_j \mid X) = E(Y_j - E(Y_i))^2$$

ou, de outra forma,

$$V(Y_j) = (1 - \theta_j) P(Y_j = 1)(1 - \theta_j) + (0 - \theta_j) P(Y_j = 0)(0 - \theta_j)$$

que leva a

$$V(Y_j) = \theta_j(1 - \theta_j). \tag{2.39}$$

Desta forma, *V(Yi)* não é constante, o que invalida os testes de significância usuais, com o Modelo Linear Geral e resposta politômica. Uma dificuldade adicional reside no fato de que o Modelo Linear Geral fornece para *Y* valores que não pertencem ao intervalo [0 , 1].

Seja a função

$$\Pi(X_i) = [P(X_i)]^{Y_i} [1 - P(X_i)]^{1-Y_i} \tag{2.40}$$

O método de estimação que leva à função de mínimos quadrados no MLG é chamado Máxima Verossimilhança (MV), que, por sua vez, é a base para a abordagem ao MRL. Em linhas gerais, o Método da Máxima Verossimilhança fornece estimativas para os parâmetros que maximizam a probabilidade de obter o conjunto observado de dados. Para aplicar tal método deve-se, em primeiro lugar, construir a função chamada Função de Verossimilhança (FV). Os estimadores de máxima verossimilhança destes parâmetros são escolhidos entre aqueles que maximizam esta

função. A Função de Verossimilhança é dada por

$$\lambda(\beta)=\prod_{i=1}^{n} P(X_i) \tag{2.41}$$

A Máxima Verossimilhança implica que o estimador para β seja o valor que maximiza a expressão dada em (2.41). Contudo, é mais fácil, do ponto de vista matemático, trabalhar com o logaritmo da mesma. A expressão fica, então

$$L(\beta)=\ln[\lambda(\beta)]$$

$$L(\beta)=\ln\prod_{i=1}^{n} P(X_i)$$

$$L(\beta)=\sum_{i=1}^{n}\{Y_i \ln[P(X_i)]+(1-Y_i)\ln[1-P(X_i)]\}$$

$$L(\beta)=\sum_{i=1}^{n}\left\{Y_i \ln\frac{e^{\beta' X}}{1+e^{\beta' X}}+(1-Y_i)\ln\left[1-\frac{e^{\beta' X}}{1+e^{\beta' X}}\right]\right\}$$

$$L(\beta)=\sum_{i=1}^{n}\{Y_i(\beta' X)-\ln(1+e^{\beta' X})\} \tag{2.42}$$

Para obter β que maximiza $L(\beta)$ basta derivar a expressão em relação a β e igualar a zero as equações obtidas. As expressões resultantes são

$$\frac{\partial L(\beta)}{\partial \beta_0}=\sum_{i=1}^{n}\left(Y_i-\frac{e^{\beta' X}}{1+e^{\beta' X}}\right) \tag{2.43}$$

$$\frac{\partial L(\beta)}{\partial \beta_i}=\sum_{i=1}^{n}\left(X_i Y_i-\frac{X_i e^{\beta' X}}{1+e^{\beta' X}}\right) \tag{2.44}$$

No MLG as expressões (2.43) e (2.44) conduzem a um sistema de equações lineares, o que facilita em muito o cálculo dos estimadores para os parâmetros desconhecidos. No MRL as expressões são não lineares, requerendo, portanto, métodos especiais para a sua resolução. Tais métodos são iterativos, o que exige a utilização de *softwares* específicos, ou a construção de programas computacionais para implementação dos referidos métodos. Entre os referidos métodos iterativos, um dos mais conhecidos é o *Método de Newton*, que apresenta a vantagem de convergir rapidamente para a solução.

A função $L(\beta)$, para este caso, pode ser escrita na forma

$$L(\beta) = \sum_{i=1}^{n} \left[Y_i (\beta' X) - \ln\left(1 + e^{\beta' X}\right) \right] \tag{2.45}$$

Os estimadores procurados devem maximizar a função acima. As derivadas parciais de primeira ordem da função $L(\beta)$ são dadas por

$$\frac{\partial L(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} \left(Y_i - \frac{e^{\beta' X}}{1 + e^{\beta' X}} \right) = \sum_{i=1}^{n} \left[Y_i - P(X_i) \right] \tag{2.46}$$

$$\frac{\partial L(\beta)}{\partial \beta_i} = \sum_{i=1}^{n} \left(X_i Y_i - \frac{X_i e^{\beta' X}}{1 + e^{\beta' X}} \right) = \sum_{i=1}^{n} X_i \left[Y_i - P(X_i) \right] \tag{2.47}$$

As derivadas de segunda ordem da função $L(\beta)$ são

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} X_{ij}^2 P(X_i)\left[1 - P(X_i)\right] \tag{2.48}$$

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_u} = -\sum_{i=1}^{n} X_{ij} X_{iu} P(X_i)\left[1 - P(X_i)\right] \tag{2.49}$$

com j , u = 0 , 1 , 2 , ...

Com isto obtém-se a expressão

$$\beta_{k+1} = \beta_k + \left\{ I(\beta_k) \right\}^{-1} \left\{ \left[S(\beta_k) \right] \right\} \tag{2.50}$$

onde

$I(\beta_k) = X'VX$, sendo V a matriz diagonal de variâncias, dada por

$$V_{ii} = \hat{P}(X_i)\left[1 - \hat{P}(X_i)\right] = \frac{e^{\beta' X}}{1 + e^{\beta' X}} \left[1 - \frac{e^{\beta' X}}{1 + e^{\beta' X}} \right] = \frac{e^{\beta' X}}{\left(1 + e^{\beta' X}\right)^2} \tag{2.51}$$

e

$$S(\beta_k) = \begin{bmatrix} \dfrac{\partial L(\beta)}{\partial \beta_0} \\ \cdots \\ \dfrac{\partial L(\beta)}{\partial \beta_{p-1}} \end{bmatrix} \tag{2.52}$$

com aproximação inicial dada, neste trabalho, por $\beta_0' = [0 \quad \ldots \quad 0]$

2.6 Abordagem de Lachenbruch

Tão importante quanto a obtenção de uma boa regra de classificação é a determinação da eficiência da mesma. Uma regra que apresente uma taxa de erros elevada, pouca, ou nenhuma, utilidade terá. A *Abordagem de Lachenbruch,* LACHENBRUCH (1975), é uma forma de avaliar a eficiência da regra de classificação. Esta técnica segue os passos apresentados a seguir:

1. Escolher um dos grupos (amostras).
2. Retirar uma observação do grupo.
3. Construir uma função discriminante com as $n_1 - 1$ observações restantes do grupo escolhido e as n_2 observações do segundo grupo, ou seja, para $n_1 - 1 + n_2$ observações.
4. Classificar a observação retirada usando a função obtida anteriormente.
5. Realocar a observação descartada e repetir os passos 1 e 2 para todas as observações do primeiro grupo.
6. Repetir os passos 1 a 5 para o segundo grupo.
7. Finalmente, construir a regra de classificação com o total das $n = n_1 + n_2$ observações.

Assim obtém-se:

$$P(2|1) = \frac{n_{1|2}}{n_1} \tag{2.53}$$

$$P(1|2) = \frac{n_{2|1}}{n_2} \tag{2.54}$$

que são as probabilidades de classificação incorreta para cada um dos grupos e

$$\hat{E}(AER) = \frac{n_{1|2} + n_{2|1}}{n_1 + n_2} \tag{2.55}$$

que é a proporção total esperada de erro.

Desta forma, obtém-se uma regra de reconhecimento e classificação construída com as n observações amostrais e testada com todas as referidas observações, mas sempre com a observação em

teste fora do ajuste. Isto equivale a ter um grupo com n observações para o ajuste e outro grupo, também de tamanho n, para testar a eficiência do procedimento.

3. Material e Método

3.1 Amostra, Questionário e Características das Variáveis Estudadas

Neste trabalho, utilizou-se uma amostra com 707 observações, sendo 102 pertencentes ao grupo "0", de maus clientes, e 505 pertencentes ao grupo "1", de bons clientes. Foram estudadas as variáveis, normalmente, constantes em formulários de adesão utilizados pela instituição financeira fornecedora do dados.

3.2 Função Discriminante Linear e Regressão Logística

A matriz de dados de ordem n x p, onde n = 707 observações do vetor X , de dimensão p = 27 variáveis, foi utilizada para a obtenção de uma Função Discriminante Linear Amostral de Fisher, partindo-se da matriz de covariância estimada

$$S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \qquad (3.01)$$

onde:

n_1 = 102 observações do grupo "0" , de clientes inadimplentes; e

n_2 = 605 observações do grupo "1" , de clientes adimplentes.

A Função Discriminante Linear de Fisher Amostral obtida tem seus coeficientes dados na Tabela 2, e é ali comentada. A eficiência foi avaliada pelo Método de Lachenbruch, descrito no item 2.4. Os resultados da avaliação constam na Tabela 3. Algumas variáveis foram descartadas, por representarem combinações lineares de outras. Foram criadas algumas variáveis, chamadas sintéticas, cujos valores são obtidos a partir de variáveis correntes como, por exemplo, a diferença entre a renda do cliente e o limite de crédito e razão entre a renda e a idade.

A mesma amostra utilizada para o ajuste da Função

Discriminante Linear de Fisher foi, também, usada para modelar um ajuste logístico através do mesmo programa. A mesma matriz acima mencionada foi usada para a obtenção de um Modelo de Regressão Logística, utilizando as mesmas observações e calculando os Estimadores de Máxima Verossimilhança através do Método de Newton. A eficiência foi igualmente avaliada pelo Método de Lachenbruch.

3.3 Avaliação do Desempenho dos Modelos Ajustados Usando-se o Método de Lachenbruch

Tabela 2 – coeficientes da função discriminante Linear de Fisher

Variável	Coeficiente
Limite	0,0001
Sexo	- 2,5097
Tempo de residência	- 0,0806
Seguro automotivo	-1,0719
Seguro residencial	0,1594
Cartão segurado	2,2729
Seguro de vida	0,9609
Renda	- 0,0004
Idade	- 0,0166
Tempo no atual emprego	- 0,1010
Idade do cônjuge	- 0,0051
Telefone celular	7,1033
Estado civil	0,2520
Tipo do documento apresentado	- 0,2060
Escolaridade	- 0,7977
Tipo de residência	- 1,1765
Setor de atividade	- 0,9055
Resultado (método atual)	- 9,7641
Cep	0,0002

Tabela 3 – Coeficientes do Modelo de Regressão Logística

Variável	Coeficiente $(\hat{\beta})$
Limite	- 0,0006
Sexo	2,8811
Tempo de Residência	0,1643
Seguro automotivo	0,1260
Seguro residencial	- 0,3739
Cartão segurado	- 5,3658

Cont. Tabela 3

Seguro de Vida	- 0,7594
Renda	0,0012
Idade	0,0298
Tempo no Atual Emprego	0,2443
Idade do Cônjuge	0,0092
Telefone celular	- 6,6774
Estado civil	- 0,0615
Tipo do Documento Apresentado	0,7448
Escolaridade	0,9614
Tipo de residência	1,5737
Setor de atividade	- 0,8687
Resultado (método atual)	6,1844
Cep	- 0,0003
Constante	19,6195

As taxas de acerto para a Função Discriminante Linear de Fischer e para o Modelo de Regressão Logística obtidas, com e sem o acréscimo das variáveis sintéticas, são apresentadas na Tabela 4.

Tabela 4 – Resultados da Avaliação dos Modelos de Reconhecimento de Padrões Pelo Método de Lachenbruch

Modelo de Reconhecimento de Padrões	Variáveis correntes		Variáveis correntes e variáveis sintéticas	
	Grupo 0: "maus"	Grupo 1: "bons"	Grupo 0: "maus"	Grupo 1: "bons"
Função Discriminante Linear	92,16%	92,40%	69,61%	58,84%
Modelo de Regressão Logística	99,02%	99,83%	87,25%	98,84%

5. Conclusão

A utilização das técnicas multivariadas abordadas neste trabalho tem qualidades como, por exemplo, medir e avaliar a eficiência através de uma ferramenta científica utilizada em atividades de pesquisa, tais como Cardiologia, Sociologia e outras. Desta forma, a instituição de crédito tem a sua disposição uma ferramenta, que, aliada à informática, possui inegável confiabilidade quanto aos resultados apresentados. Não se deve esquecer, contudo, que esta

ferramenta necessita, para o seu desenvolvimento, e posterior aplicação, de um banco de dados da maior qualidade, a fim de aumentar ainda mais o seu potencial como ferramenta de auxílio à tomada de decisões. A diferença verificada para a eficiência dos dois procedimentos desenvolvidos não constitui empecilho à sua utilização. Ao contrário, possibilita ao tomador de decisões a escolha daquele que apresentar a maior eficiência, aumentando, desta forma, a segurança com relação à escolha efetuada, característica altamente recomendável face ao montante de recursos envolvidos nas operações de crédito. Tais atributos, desde que aliados à uma base de dados efetivamente confiável, podem conduzir a uma diminuição no custo dos financiamentos, o que deve ser um dos principais objetivos de uma instituição financeira.

6. Referências Bibliográficas

CAOUETTE, J. B.; ALTMAN, E. I. & NARAYANAN, P. **Gestão do Risco de Crédito**. O Próximo Grande Desafio Financeiro. São Paulo: Qualitymark Editora, 2000

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis.** Prentice Hall, 1988.

LACHENBRUCH, P. A. **Discriminant Analysis**. Hafner Press, 1975.

SILVA, J. P. **Análise e Decisão de Crédito**. São Paulo, Atlas: 1988

Recebido para publicação em 08/06/2001
Aceito para publicação em 31/10/2001