Luiz Paulo Fávero
Patrícia Belfiore
Fabiana Lopes da Silva
Betty Lilian Chan

análise de dados

MODELAGEM MULTIVARIADA PARA TOMADA DE DECISÕES

CAMPUS

CAPÍTULO 12

Regressão Logística e Regressão Logística Multinomial

> Aquilo que os homens, de fato, querem não é o conhecimento, mas a certeza.
>
> BERTRAND RUSSELL

AO FINAL DESTE CAPÍTULO, VOCÊ SERÁ CAPAZ DE:

- Identificar as situações passíveis de utilização destas técnicas.
- Entender as propriedades da regressão logística e explicar sua popularidade.
- Calcular o risco que representa cada variável independente na função logística.
- Descrever como a função logística pode ser interpretada em função da chance (*odds*).
- Interpretar os parâmetros da função logística.
- Entender e aplicar a técnica de regressão logística multinomial e saber diferenciá-la da regressão logística.
- Compreender as principais diferenças entre as técnicas de regressão logística e de regressão logística multinomial e as demais técnicas de dependência.

12.1. APRESENTAÇÃO DO CAPÍTULO

Este capítulo descreve o modelo de regressão logística binária e multinomial, apresentando seus fundamentos teóricos e, em paralelo, oferecendo ao pesquisador aplicações práticas para facilitar a compreensão e utilização das técnicas.

Um pesquisador pode estar interessado, por exemplo, na determinação da probabilidade de insolvência de determinado cliente de lojas de eletroeletrônicos em função de suas características sociodemográficas. Um médico, por outro lado, pode desejar investigar se a probabilidade de um ataque cardíaco pode ser predita em função de características sanguíneas e emocionais, sexo e estilo de vida do paciente. Por fim, uma operadora de telefonia móvel pode querer saber a probabilidade de mudança de plano por parte dos clientes que compõem sua carteira, em função de características como nível de escolaridade, renda, sexo, estado civil, número de filhos e tempo de relacionamento com a operadora. Todos esses são exemplos que podem ser modelados por meio da técnica de regressão logística.

Os três maiores objetivos deste capítulo são: (1) introduzir a natureza, a filosofia e as condições das técnicas de regressão logística e de regressão logística multinomial; (2) apresentar a aplicação das técnicas; e (3) discutir os resultados obtidos.

12.2. UMA INTRODUÇÃO À REGRESSÃO LOGÍSTICA

Segundo Corrar, Paulo e Dias Filho (2007), a técnica de regressão logística foi desenvolvida por volta da década de 1960, em resposta ao desafio de realizar predições ou explicar a ocorrência de determinados fenômenos quando a variável dependente fosse de natureza binária. Um dos primeiros estudos que contribuíram para conferir notoriedade à técnica foi o Framingham Heart Study, realizado com a colaboração da Universidade de Boston, que tinha como objetivo identificar fatores que propiciam doenças cardiovasculares em uma amostra de 5.209 indivíduos com idades entre 30 e 60 anos residentes na cidade de Framingham, Massachusetts. Com a utilização da regressão logística, vários fatores foram identificados como sendo de risco, como hipertensão arterial, taxas altas de colesterol, tabagismo, obesidade, diabetes e sedentarismo.

A regressão logística é uma técnica estatística utilizada para descrever o comportamento entre uma variável dependente binária e variáveis independentes métricas ou não métricas. Ou seja, destina-se a investigar o efeito das variáveis pelas quais os indivíduos, objetos ou sujeitos estão expostos sobre a probabilidade de ocorrência de determinado evento de interesse.

Daí a popularidade do uso desta técnica, pois há uma infinidade de eventos de interesse que poderiam ser modelados pela regressão logística, tais como a ocorrência de uma doença (ciências biomédicas), de uma inadimplência (análise de crédito), de um sinistro (seguros), da compra de um bem (marketing), entre outros. Recentemente, a regressão logística vem sendo muito utilizada no desenvolvimento de *Credit Scoring*.

Por exemplo, suponha que uma seguradora esteja interessada em investigar qual a probabilidade de uma pessoa falecer dado que é ou não fumante. Neste sentido, o evento de interesse seria a morte (variável dependente), cuja ocorrência poderia ser representada por 1, enquanto a não-ocorrência poderia ser denotada por 0. Analogamente, a variável explicativa (fumante ou não) também poderia ser representada por 1 e 0, respectivamente. Adicionalmente, podem ser introduzidas no modelo algumas variáveis de controle que podem estar, de alguma forma, relacionadas ao evento de interesse, como, por exemplo, a idade, o sexo, a prática ou não de esportes.

Cabe ainda destacar que a grande vantagem da regressão logística diante das outras técnicas, como a análise discriminante estudada no Capítulo 11, reside na flexibilidade de seus pressupostos, o que amplia sua aplicabilidade.

A função logística, $f(Z) = \frac{1}{1+e^{-(Z)}}$, assume valores entre 0 e 1, para qualquer Z entre $-\infty$ e $+\infty$, conforme mostrado na Figura 12.1. Assim, a popularidade desta técnica advém não apenas da possibilidade de prever a ocorrência de eventos de interesse, mas também da capacidade de apresentar a probabilidade de sua ocorrência, sendo esta uma limitação inerente à análise discriminante, a qual fornece o valor de um *score* em vez da probabilidade.

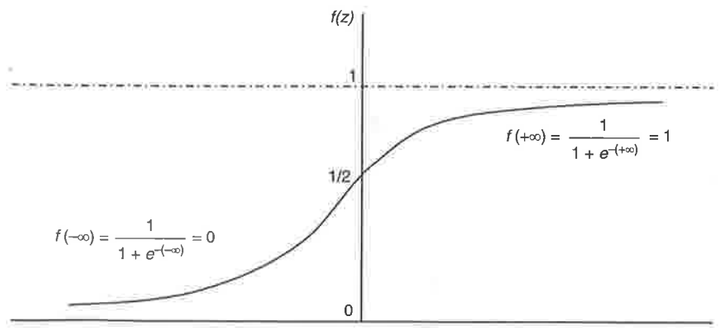


Figura 12.1: Função logística.

12.3. FUNDAMENTOS CONCEITUAIS

A regressão logística é uma técnica desenvolvida na década de 1960 para investigar a relação entre variáveis explicativas, métricas e não métricas e uma variável dependente categórica binária. Diferentemente da regressão múltipla, a regressão logística não pressupõe a existência de homogeneidade de variância e normalidade dos resíduos.

Isto é, a regressão logística destina-se a aferir a probabilidade de ocorrência de um evento e a identificar características dos elementos pertencentes a cada grupo determinado pela variável categórica.

A função logística se apresenta como uma curva em formato de "S", cujos valores se situam entre 0 e 1, representando a probabilidade de ocorrência do evento de interesse.

12.3.1 Função Logística

Um modelo é definido como logístico se a função segue a seguinte equação:

$$f(Z) = \frac{1}{1+e^{-(Z)}} \tag{12.1}$$

Sendo Z:

$$Z = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_k \tag{12.2}$$

Em que p indica a probabilidade de ocorrência de determinado evento de interesse, X representa o vetor de variáveis explicativas (ou independentes) e α e β os parâmetros do modelo. O termo $ln(p/1\text{-}p)$ é chamado de *logit* e o termo $(p/1\text{-}p)$ representa a chance (*odds*) de ocorrência do evento de interesse. Por exemplo:

- Se $p = 0{,}50$, a chance de ocorrência do evento será de 1 (1 para 1);
- Se $p = 0{,}75$, a chance de ocorrência do evento será de 3 (3 para 1).

Logo, é fácil definir que a probabilidade de ocorrência de um evento de interesse é $p = (odds/1+odds)$.

Portanto, substituindo (12.2) em (12.1), tem-se:

$$f(Z) = \frac{1}{1+e^{-(\alpha+\sum\beta_i X_i)}} \tag{12.3}$$

Simplificadamente, a função $f(Z)$ pode ser entendida como a probabilidade de a variável dependente ser igual a 1, dado o comportamento das variáveis explicativas $X_1, X_2, \ldots, X_k$. Ou seja, matematicamente, pode ser representada como segue:

$$P(1) = f(Y = 1 | X_1, X_2, \ldots X_k) = \frac{1}{1+e^{-(\alpha+\sum\beta_i X_i)}} \tag{12.4}$$

Em outras palavras, se isolarmos p na Expressão (12.2), chegaremos à Expressão (12.4). Como α e β são parâmetros desconhecidos, é necessário estimá-los para a determinação da probabilidade de ocorrência do evento de interesse. A notação utilizada para designar parâmetros estimados é "^" (chapéu), e no caso específico, $\hat{a}$ e $\hat{\beta}$, respectivamente. O método utilizado para estimar tais parâmetros é o de máxima verossimilhança, cuja estimação será apresentada no Apêndice A deste capítulo.

Em outras palavras, o objetivo de estimar tais parâmetros é encontrar uma função logística de tal maneira que as ponderações das variáveis explicativas permitam estabelecer a importância de cada variável para a ocorrência do evento de interesse, bem como calcular a probabilidade de ocorrência desse evento.

Assim, suponha que a probabilidade de um cliente adquirir a assinatura de uma revista por mala direta seja dada pela seguinte equação:

$$prob(event) = \frac{1}{1+e^{-(1,143+0,452X_1+0,029X_2-0,242X_3)}}$$

Sendo:
X_1 = *sexo (1 para feminino e 0 para masculino);*
X_2 = *idade;*
X_3 = *estado civil (1 para solteiro e 0 para casado).*

Para uma pessoa do sexo feminino, com 40 anos de idade e casada, a probabilidade de adquirir a assinatura da revista é:

$$prob(event) = \frac{1}{1+e^{-(1,143+0,452.1+0,029.40-0,242.0)}} = 0,47$$

Sob as mesmas condições, mas se fosse do sexo masculino, a probabilidade seria calculada como segue:

$$prob(event) = \frac{1}{1+e^{-(1,143+0,452.0+0,029.40-0,242.0)}} = 0,02$$

Neste sentido, a razão do risco (*risk ratio*), em função do sexo, é dada por:

$$\hat{RR} = \frac{0,47}{0,02} = 27,59$$

Isso significa que uma mulher teria uma probabilidade de quase 28 vezes maior de adquirir a assinatura da revista que um homem.

O que foi feito aqui é uma forma de se estimar diretamente a chance em função da mudança no atributo. Entretanto, tal prática só é possível quando forem especificadas todas as variáveis independentes e quando a análise se concentrar em cada observação.

Mas, quando isso não for possível, situação que representa a maior parte dos casos na prática, é necessário estimar a razão do risco de forma indireta, por intermédio da razão da chance, usualmente denominada de *odds ratio*. No início da década de 1990, houve uma grande discussão no mundo acadêmico a respeito da melhor tradução do termo *odds ratio* para a língua espanhola, como pode ser encontrado em Rigau (1990), Martín (1990) e Tapia e Nieto (1993). Não nos ateremos a essa discussão, uma vez que utilizaremos, como tradução para este termo, o conceito de chance.

Outro conceito importante é o *Risk Odds Ratio*, representado por *ROR* neste capítulo, o qual é calculado pela razão do *odds* (chance) entre dois grupos (R_0 e R_1), aplicando-se o modelo logístico:

$$ROR_{R_1,R_0} = \frac{odds\, R_1}{odds\, R_0} \tag{12.5}$$

Tendo em vista a Equação (12.4) de cálculo da probabilidade de ocorrência do evento de interesse, tem-se que o *odds* é dado por:

$$odds\, R_1 = \frac{P(R_1)}{1-P(R_1)} = \frac{\frac{1}{1+e^{-(\alpha+\sum\beta_i X_i)}}}{\frac{e^{-(\alpha+\sum\beta_i X_i)}}{1+e^{-(\alpha+\sum\beta_i X_i)}}} = e^{(\alpha+\sum\beta_i X_i)} \tag{12.6}$$

$$odds\, R_0 = \frac{P(R_0)}{1-P(R_0)} = \frac{\frac{1}{1+e^{-(\alpha+\sum\beta_i X_i)}}}{\frac{e^{-(\alpha+\sum\beta_i X_i)}}{1+e^{-(\alpha+\sum\beta_i X_i)}}} = e^{(\alpha+\sum\beta_i X_i)} \tag{12.7}$$

Assim, *Risk Odds Ratio* (*ROR*) será calculado da seguinte maneira:

$$ROR_{R_1,R_0} = \frac{\frac{P(R_1)}{1-P(R_1)}}{\frac{P(R_0)}{1-P(R_0)}} \tag{12.8}$$

$$ROR_{R_1,R_0} = \frac{odds\, R_1}{odds\, R_0} = \frac{e^{\alpha+\sum\beta_i X_{1i}}}{e^{\alpha+\sum\beta_i X_{0i}}} \tag{12.9}$$

Para:

$$a = e^{\alpha+\sum\beta_i X_{1i}} \text{ e } b = e^{\alpha+\sum\beta_i X_{0i}} \tag{12.10}$$

Então:

$$ROR_{R_1,R_0} = \frac{odds\, R_1}{odds\, R_0} = \frac{e^{\alpha+\sum\beta_i X_{1i}}}{e^{\alpha+\sum\beta_i X_{0i}}} = e^{a-b} = e^{\sum_{i=1}^{k}\beta(X_{1i}-X_{0i})} \tag{12.11}$$

O conceito de *odds* é fundamental para se determinar o modelo *logit*, também denotado de Z, que consiste no logaritmo do *odds* (chance), conforme segue:

$$Z = logit = \ln \underbrace{\left[\frac{P(R_1)}{1-P(R_1)}\right]}_{Odds\ R_1} = \ln\left[\frac{P(R_1)}{P(R_0)}\right] = \ln[e^{(\alpha+\sum\beta_i X_i)}] = \alpha + \sum\beta_i X_i \tag{12.12}$$

Sendo $P(R_1)$ a probabilidade de ocorrência do evento de interesse, ou seja, a probabilidade de classificar o indivíduo no grupo R_1, podemos calculá-la da seguinte maneira:

$$P(R_1) = \frac{1}{1+e^{-(\alpha+\sum\beta_i X_i)}} \tag{12.13}$$

A partir desta função, é possível entender o significado de α e β_i, sendo o primeiro o logaritmo natural da chance quando todas as variáveis explicativas são nulas e o segundo representa a mudança no logaritmo natural da chance dada a variação de uma unidade na variável *X*.

12.3.2 Premissas da Regressão Logística

A regressão logística assume as seguintes premissas:

- relação linear entre o vetor das variáveis explicativas *X* e a variável dependente *Y*;
- valor esperado dos resíduos é igual a zero;
- ausência de heterocedasticidade;
- ausência de multicolinearidade.

Como podemos observar, diferentemente da análise de regressão múltipla e da análise discriminante, a regressão logística não pressupõe normalidade dos resíduos. Isso acaba representando, quando da aplicação daquelas técnicas, uma grande limitação, uma vez que, na presença de muitas variáveis dicotômicas, este pressuposto acaba sendo violado. Além disso, a regressão logística também não pressupõe homogeneidade de variância e a redução do número de pressupostos torna-a preferível em muitas situações práticas.

Quando tivermos, em um mesmo modelo, variáveis explicativas com escalas de mensuração qualitativa e quantitativa, a premissa de normalidade multivariada não será atendida na análise discriminante. Nesses casos, o pesquisador pode optar pelo uso da regressão logística, uma vez que esta não faz nenhuma consideração sobre a distribuição das variáveis explicativas (SHARMA, 1996).

12.3.3 Medidas de Ajuste do Modelo de Regressão Logística

Dentre as diferenças entre a regressão múltipla e a regressão logística, destaca-se a forma de estimação dos parâmetros. Enquanto a primeira baseia-se no método dos mínimos quadrados, a fim de minimizar os desvios quadráticos, a segunda consiste no método de máxima verossimilhança, ou seja, busca maximizar a probabilidade (verossimilhança) de que um evento ocorra.

Em função dessa diferença, a medida de ajuste do modelo também difere. Assim, em relação à regressão logística, Hair, Anderson, Tatham e Black (2005) explicam que "a medida geral do quão bem o modelo se ajusta, semelhante ao valor das somas de quadrados de erros ou resíduos para regressão múltipla, é dada pelo valor de verossimilhança (na verdade, é –2 vezes o logaritmo do valor da verossimilhança e é chamado de $-2LL$ ou –2log verossimilhança)". Neste sentido, quanto menor o valor de $-2LL$, melhor é a adequação do modelo. Ou seja, quando a verossimilhança for 1, indicando o ajuste perfeito, o valor de $-2LL$ é zero.

Cabe ainda destacar as seguintes medidas de ajustamento:

- Pseudo R^2 (R^2 *logit*):

$$R^2_{logit} = \frac{-2LL_0 - (-2LL_\beta)}{-2LL_0}$$

- Cox & Snell R^2 (medida semelhante ao R^2 da regressão linear múltipla):

$$R^2_{CS} = 1 - \left(\frac{L_0}{L_\beta}\right)^{\frac{2}{N}}$$

$$R^2_{CS_{MÁX}} = 1 - (L_0)^{\frac{2}{N}}$$

- Nagelkerke R^2:

$$\widetilde{R}^2_N = \frac{R^2_{CS}}{R^2_{CS_{MÁX}}}$$

- Teste Qui-quadrado: avalia a existência de diferenças significativas entre o esperado e o observado.
- Hosmer-Lemeshow *Goodness-of-fit Test:* testa se as classificações previstas para cada grupo são iguais às observadas, por meio da estratificação das observações em faixas (decis) e da aplicação de um teste Qui-quadrado (χ^2) para avaliar se há diferenças significativas entre as frequências observadas e esperadas em cada faixa, de forma similar ao já aplicado quando do estudo da análise de correspondência no Capítulo 8.

Para analisar o poder preditivo do modelo, é usual a utilização da tabela de classificação. Para elaboração dessa tabela, é necessário o estabelecimento de um ponto de corte, ora denotado por *c* (*classification cutoff*), cujos valores de probabilidades acima deste ponto indicam a presença do evento de interesse (ocorrência de sinistro, inadimplência, entre outros) e os valores abaixo desse ponto indicam ausência, conforme representada pela Figura 12.2 a seguir.

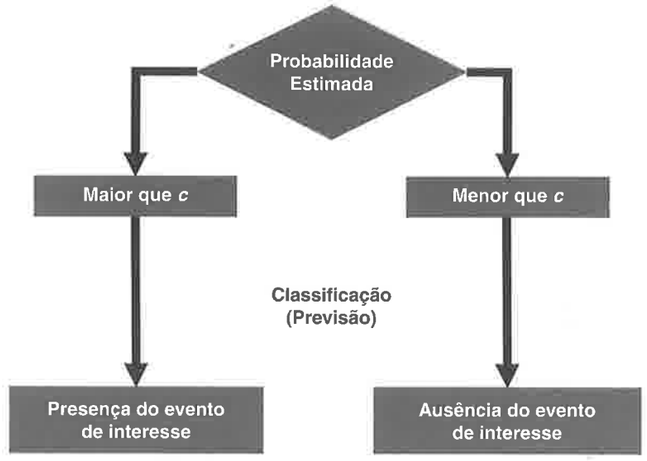


Figura 12.2: Ponto de corte (*c*).

Assim, se o evento de interesse, por exemplo, para uma seguradora, for a ocorrência de sinistro, seria possível, a partir do estabelecimento do ponto de corte, comparar a classificação prevista *versus* a observada, como ilustrado na Tabela 12.1.

Tabela 12.1: Tabela de Classificação

Observado	Predito		
	Ocorrência de sinistro	Não-ocorrência de sinistro	Total
Ocorrência de sinistro	25	7	32
Não-ocorrência de sinistro	5	163	168
Total	30	170	200

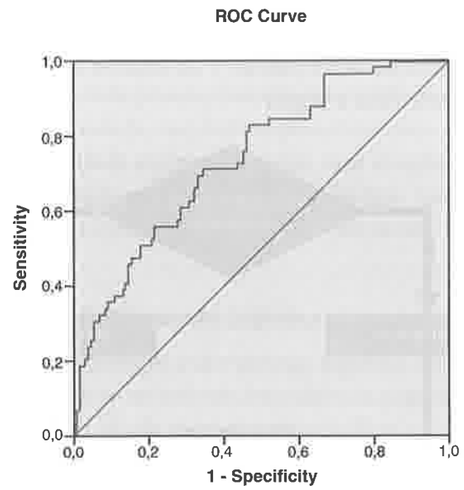
É usual o cálculo da sensitividade (verdadeiro positivo) e da especificidade (verdadeiro negativo), como segue:

- Sensitividade = 25/32 = 78% (percentual de acerto dos casos de ocorrência do evento de interesse, no caso, sinistro).
- Especificidade = 163/168 = 97% (percentual de acerto dos casos em que não ocorreram o sinistro).
- Percentual do Acerto do Modelo = (25+163)/200 = 94%.

Se, para cada ponto de corte (*c*), fosse calculada a sensitividade e a especificidade, seria possível construir o gráfico conhecido como Curva ROC (*Receiver Operating Characteristic*), conforme exemplificado a no Gráfico 12.1.

Quanto maior a área abaixo da Curva ROC, maior é a capacidade de o modelo discriminar os grupos sinistrados (evento de interesse) dos não sinistrados. Mas, quanto mais próxima a Curva ROC estiver da reta diagonal, pior é o poder discriminatório do modelo.

Gráfico 12.1: Exemplo de Curva ROC



Uma referência usual em relação à área da curva ROC é apresentada no Quadro 12.1 a seguir:

Quadro 12.1: Área Abaixo da Curva ROC

Área abaixo da curva ROC	Interpretação
Menor ou igual a 0,5	Não há discriminação
Entre 0,7 e 0,8	Discriminação aceitável
Maior que 0,8	Discriminação excelente

Outra medida de qualidade de ajuste do modelo é o K-S (Kolmogorov-Smirnov), que mede o grau de segregação dos dois grupos (sinistrados e não sinistrados), cujo valor pode ser interpretado conforme o apresentado no Quadro 12.2.

Quadro 12.2: Tabela de Qualidade do Ajuste do Modelo (K-S)

K – S	Interpretação
Menor que 30	Baixa Discriminação
De 30 a 50	Boa Discriminação
Maior que 50	Ótima Discriminação

12.4. REGRESSÃO LOGÍSTICA: UM EXEMPLO PRÁTICO

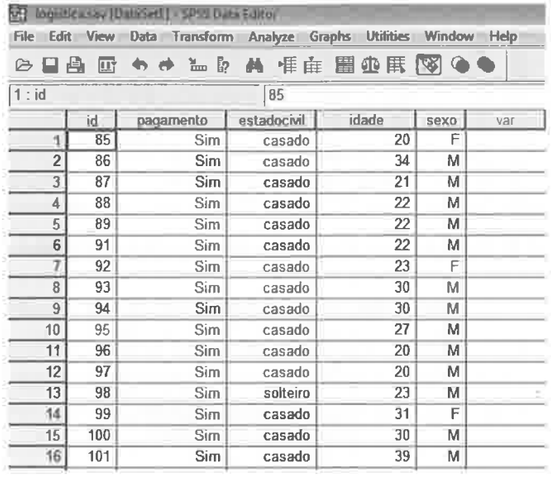
Suponha que uma empresa varejista ABC esteja interessada em identificar o perfil de clientes em atraso que deixariam o *status* de inadimplente diante de uma ação de cobrança. Dessa maneira, a empresa busca direcionar melhor seus esforços de cobrança, tendo em vista o elevado custo desta atividade (telemarketing, *mailing*, entre outras atividades) se comparado ao *ticket* médio de sua operação.

O banco de dados em SPSS está disponível no arquivo **Logistica.sav** e apresenta as seguintes variáveis:

- *id*: código de identificação do cliente;

- *pagamento*: variável dependente indicativa do cliente que, dada a ocorrência de dias de atraso, volta a pagar as prestações mediante um esforço de cobrança (y=1) e clientes que se tornam inadimplentes por mais de 360 dias (y=0);
- *estadocivil*: casado (0) ou solteiro (1);
- *idade*;
- *sexo*: feminino (0) ou masculino (1).

A rotina em SAS para este exemplo encontra-se no Apêndice B deste capítulo.



	id	pagamento	estadocivil	idade	sexo	var
1	85	Sim	casado	20	F	
2	86	Sim	casado	34	M	
3	87	Sim	casado	21	M	
4	88	Sim	casado	22	M	
5	89	Sim	casado	22	M	
6	91	Sim	casado	22	M	
7	92	Sim	casado	23	F	
8	93	Sim	casado	30	M	
9	94	Sim	casado	30	M	
10	95	Sim	casado	27	M	
11	96	Sim	casado	20	M	
12	97	Sim	casado	20	M	
13	98	Sim	solteiro	23	M	
14	99	Sim	casado	31	F	
15	100	Sim	casado	30	M	
16	101	Sim	casado	39	M	

Figura 12.3: Apresentação da base de dados.

O banco de dados contém uma amostra de 180 observações, sendo que 130 tendem a sair do *status* de inadimplente, dado um esforço de cobrança (y=1) e 50 continuariam no *status* de inadimplente (y=0).

Cabe ressaltar que, em muitas situações reais, o banco de dados contém valor(es) *missing* para determinada(s) variável(is). Neste sentido, como há o interesse em calcular as probabilidades de cada indivíduo, é necessário criar alternativas para tratar os valores *missing*. Uma das alternativas mais utilizadas é a categorização das variáveis contínuas, criando-se faixas (estabelecidas, por exemplo, em função de *odds*), inclusive para valores *missing*. A determinação de faixas para cada variável também auxilia no contorno do problema de multicolinearidade. Neste exemplo, não foi criada nenhuma faixa de valores para a variável *idade*, embora também seja possível esse tipo de tratamento.

Para proceder à análise de regressão logística, clique em **Analyze → Regression → Binary Logistic**, conforme Figura 12.4.

Em seguida, considere como variável dependente o pagamento e como variáveis independentes o estado civil, a idade e o sexo, conforme indicado na Figura 12.5. Por ora, considere como **Method** a opção **Enter**, que executa o modelo com todas as variáveis selecionadas pelo pesquisador. Existem outras opções, como a **Forward Wald**, em que o *software* automaticamente propõe a melhor solução com a inclusão de variáveis significantes e exclusão de variáveis não significantes.

Em **Categorical**, especifique quais variáveis são categóricas. Neste caso, *estadocivil* e *sexo*. Ao indicá-las, o *software* SPSS automaticamente cria variáveis *dummy* para inclusão no modelo, oferecendo ao pesquisador a possibilidade de escolha da referência (**Last** ou **First**), indicada pelo valor 0. Após essa escolha, clique em **Change** (Figura 12.6).

Em **Save**, marque as opções **Probabilities** e **Group Membership** (Figura 12.7).

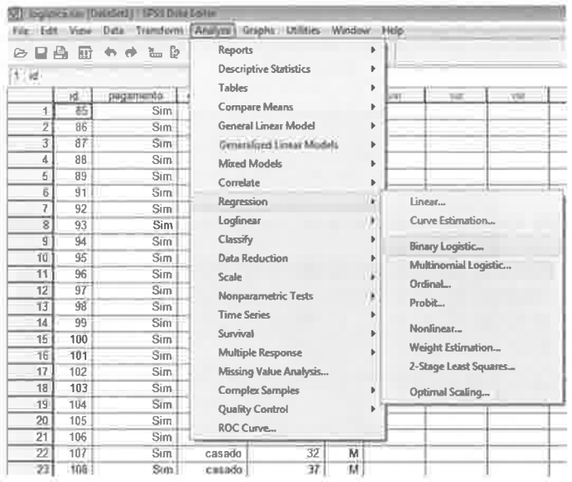


Figura 12.4: Procedimento para elaboração de regressão logística no SPSS.

Figura 12.5: Parametrização das variáveis.

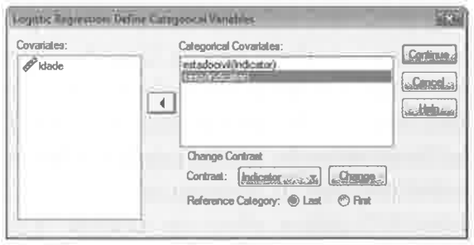


Figura 12.6: Variáveis categóricas.

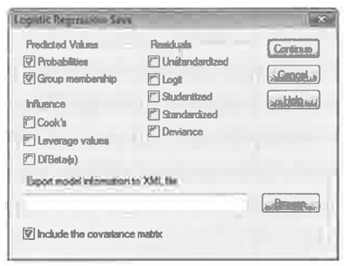


Figura 12.7: Menu **Save**.

No menu **Options**, marque as opções **Classification plots**, **Hosmer-Lemeshow goodness-of-fit** e **CI for exp(B)**, de acordo com a Figura 12.8.

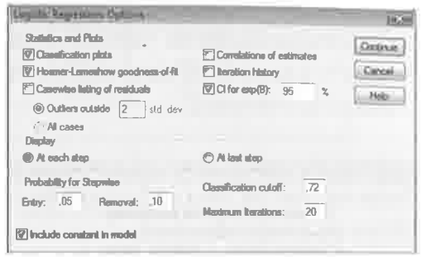


Figura 12.8: Menu **Options**.

Cabe esclarecer que, para o desenvolvimento da regressão logística, o ideal é utilizar uma amostra equilibrada, ou seja, 50% da amostra representada pela ocorrência do evento de interesse (Y=1) e 50% não (Y=0). Por isso, o valor *default* do **Classification cutoff** (ponto de corte *c* para segregação dos grupos) é 0,50. Entretanto, como na amostra analisada foi utilizada uma proporção de 0,72 (=130/180) da variável de interesse, preenchemos o **Classification cutoff** com 0,72, mas nada impede que o analista/pesquisador proponha outro ponto de corte, conforme o risco que queira assumir em relação aos erros tipo I e tipo II.

Por vezes, os dados estão muito distantes de uma população equilibrada, e o analista/pesquisador força uma amostra equilibrada (ou quase equilibrada) para a aplicação da regressão logística. Quando a proporção da amostra de desenvolvimento é diferente da proporção real da população, é conveniente ajustar a probabilidade proveniente do modelo à proporção real. Anderson (1982) e Potts e Patetta (1999) apresentam maiores detalhes sobre esta questão.

Anderson (1982) propõe a seguinte fórmula de correção do intercepto:

$$b_0 corrigido = b_0 calcculado + \ln\left(\frac{\pi_1}{\pi_2} \cdot \frac{n_2}{n_1}\right)$$

Sendo:

- π_1 = proporção de eventos de interesse na população;
- π_2 = proporção da não-ocorrência do evento de interesse na população;
- n_1 = número de eventos de interesse na amostra utilizada para desenvolvimento do modelo;
- n_2= número de observações da não-ocorrência do evento de interesse na amostra utilizada para desenvolvimento do modelo.

Potts e Patetta (1999) propõem que o ajuste seja efetuado diretamente na probabilidade, como segue:

$$p^* = \frac{p\pi_1}{p\pi_1 + (1-p)\pi_2}$$

Em que:

- p^* é o valor corrigido da probabilidade e p é seu valor estimado pelo modelo.
- $\pi_1 = \dfrac{\%evento\ de\ interesse\ na\ população}{\%evento\ de\ interesse\ na\ amostra}$
- $\pi_2 = \dfrac{\%não\,evento\ de\ interesse\ na\ população}{\%não\,evento\ de\ interesse\ na\ amostra}$

Como no exemplo proposto, para fins de simplificação, não foi realizada uma re-amostragem de maneira a torná-la equilibrada, não realizaremos os mencionados ajustes na probabilidade.

Por fim, clique em **Continue** e em **OK**.

Na Tabela 12.2, é apresentada a categorização da variável dependente, sendo considerado 0 a não-ocorrência do evento de interesse e 1 a ocorrência do evento de interesse que, no caso, é a saída do *status* de inadimplente, dado um esforço de cobrança.

Tabela 12.2: Codificação da Variável Dependente

Dependent Variable Encoding

Original Value	Internal Value
Não	0
Sim	1

Na Tabela 12.3, é apresentada a codificação utilizada pelo SPSS para as variáveis categóricas independentes, sendo esta a base a ser utilizada na equação do modelo. Ou seja, para a variável *sexo*, o valor 1 representa o sexo feminino e o valor 0 o sexo masculino, e para a variável *estadocivil*, o valor 1 indica casado e o valor 0 indica solteiro.

Tabela 12.3: Codificação das Variáveis Independentes Categóricas

Categorical Variables Codings

		Frequency	Parameter coding (1)
sexo	F	86	1,000
	M	94	,000
estadocivil	casado	151	1,000
	solteiro	29	,000

Em seguida, o SPSS mostra os resultados do *Block 0: Beginning Block*, que retrata o modelo logístico com apenas o intercepto. Embora se trate de um resultado aparentemente pouco interessante para o objetivo em questão, é uma etapa do processo que mostra as variáveis que não entraram no modelo (no caso, todas). Este resultado já sugere qual a próxima variável mais relevante para o modelo com base no *score*, caso a seleção seja o método *forward* em vez do método *Enter* utilizado.

Tabela 12.4: Importância Relativa das Variáveis Independentes

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	estadocivil(1)	52,089	1	,000
		idade	16,582	1	,000
		sexo(1)	11,347	1	,001
	Overall Statistics		64,036	3	,000

A próxima etapa apresenta os resultados do *Block 1: Method=Enter*. O primeiro passo é testar se os coeficientes em conjunto são significativos para o modelo, por intermédio da distribuição Qui-quadrado, que representa um teste análogo ao teste *F* para a regressão múltipla estudada no Capítulo 10.

Tabela 12.5: Teste de Significância dos Coeficientes do Modelo

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	66,057	3	,000
	Block	66,057	3	,000
	Model	66,057	3	,000

Com base na Tabela 12.5, observa-se que os coeficientes em conjunto são estatisticamente significativos, ou seja, há pelo menos um coeficiente diferente de zero ao nível de significância de 5%.

Na Tabela 12.6, são apresentados os resultados do ajuste do modelo. A estatística $-2LL$ não possui, normalmente, uma interpretação direta, mas influencia no resultado do teste Qui-quadrado anterior. As medidas de Cox & Snell e Nagelkerke são semelhantes ao R^2 da regressão, porém, usualmente, esta última é uma medida preferível em relação à primeira em função do valor máximo que pode atingir, neste caso, 1, facilitando a interpretação do pesquisador. Neste exemplo, o modelo proposto apresenta um poder explicativo de 44,3%.

Outro teste muito usual para verificar o ajuste do modelo é o de Hosmer-Lemeshow (Tabelas 12.7 e 12.8), cuja aplicação consiste na comparação entre os eventos observados e os esperados, com base na divisão da base de dados em 10 grupos, sendo analisado o número de eventos para cada categoria da variável dependente. Este teste refere-se à aplicação de um teste Qui-quadrado (χ^2) para avaliar se há diferenças significativas entre as frequências observadas e esperadas em cada faixa, de forma similar ao já aplicado

Tabela 12.6: Ajuste do Modelo

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	146,647[a]	,307	,443

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

Tabela 12.7: Teste de Hosmer-Lemeshow

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4,951	8	,763

Tabela 12.8: Grupos do Teste de Hosmer-Lemeshow

Contingency Table for Hosmer and Lemeshow Test

		pagamento = Não		pagamento = Sim		
		Observed	Expected	Observed	Expected	Total
Step 1	1	16	16,280	2	1,720	18
	2	10	10,173	6	5,827	16
	3	6	6,760	12	11,240	18
	4	8	5,400	11	13,600	19
	5	3	3,057	13	12,943	16
	6	2	2,675	16	15,325	18
	7	0	1,971	16	14,029	16
	8	2	1,608	15	15,392	17
	9	2	1,329	18	18,671	20
	10	1	,747	21	21,253	22

quando do estudo da análise de correspondência no Capítulo 8. Porém, para este caso, talvez o pesquisador deseje que os resultados do teste não rejeitem a hipótese nula, uma vez que há o interesse de que as frequências observadas e esperadas da tabela de contingência sejam próximas.

O teste de Hosmer-Lemeshow sugere, para o exemplo em questão, que não há diferenças significativas entre as frequências previstas e as observadas, ao nível de significância de 5%, tendo em vista que o valor de Sig. foi de 0,763. Cabe, porém, ressaltar que a aplicabilidade deste teste é limitada, já que seus resultados são mais consistentes quando a amostra apresenta grandes dimensões.

A tabela de classificação mostrada a seguir é uma forma de visualizar quanto o modelo classifica corretamente os eventos, com base no ponto de corte *c* estabelecido inicialmente como sendo 0,72. É importante lembrar que outros pontos de corte podem ser aplicados conforme o interesse do pesquisador em relação aos erros do tipo I e do tipo II.

Tabela 12.9: Tabela de Classificação

Classification Table[a]

			Predicted		
			pagamento		Percentage
	Observed		Não	Sim	Correct
Step 1	pagamento	Não	36	14	72,0
		Sim	30	100	76,9
	Overall Percentage				75,6

a. The cut value is ,720

Assim, o percentual de acerto dos clientes em atraso que continuariam como inadimplentes é de 72% (=36/(36+14)) e os que deixariam o *status* de inadimplentes diante de uma ação de cobrança seria de 76,9% (=100/(100+30)). O percentual de acerto global do modelo é de 75,6% (=(36+100)/(36+14 +30+100)).

A Tabela 12.10 apresentada na sequência mostra os resultados dos parâmetros estimados, que se mostraram significativos uma vez que todos os valores de Sig. (estatística de Wald – análoga ao teste *t* da regressão múltipla) foram inferiores ao nível de significância de 5%, ou seja, podem ser considerados estatisticamente diferentes de 0. A finalidade da estatística de Wald é verificar se cada um dos parâmetros do modelo é significativamente diferente de 0.

Tabela 12.10: Resultados do Modelo

Variables in the Equation

								95,0% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	estadocivil(1)	2,951	,583	25,627	1	,000	19,124	6,101	59,947
	idade	,116	,044	6,868	1	,009	1,123	1,030	1,225
	sexo(1)	-1,301	,439	8,801	1	,003	,272	,115	,643
	Constant	-3,616	1,163	9,659	1	,002	,027		

a. Variable(s) entered on step 1: estadocivil, idade, sexo.

Para calcularmos as probabilidades de cada indivíduo, precisamos lembrar os critérios apresentados por meio da Tabela 12.3 de codificação das variáveis independentes. Assim, para melhor entendermos como foram calculadas as probabilidades, faz-se necessário um exemplo.

Consideremos a primeira observação da base de dados, que apresenta as seguintes características:

- *Y* (*pagamento*) = 1 (sim);
- *estadocivil* = casado (codificação interna do SPSS = 1);
- *idade* = 20 anos;
- *sexo* = feminino (codificação interna do SPSS = 1).

O pesquisador pode achar estranho o fato de a categoria *casado* da variável *estadocivil* apresentar valor 1 e seu valor no banco de dados estar 0. Essa confusão não pode ocorrer. Na verdade, o valor 0 do banco de dados é apenas um *label* (rótulo), que poderia ser 10, 20 ou 30. Porém, o valor 1 para esta categoria foi atribuído quando de sua inserção no modelo e determinação da categoria de referência (que optamos por **Last**) para a criação da variável *dummy*.

Com base na Equação (12.4), temos:

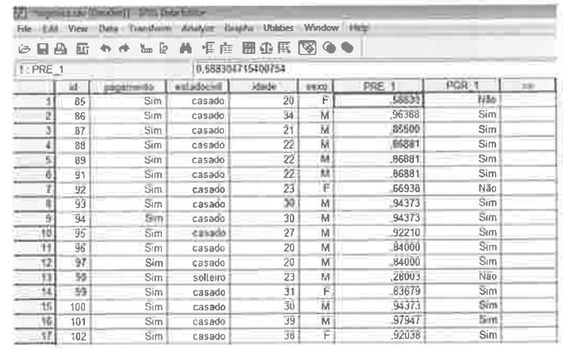
$$P(Y=1)=\frac{1}{1+e^{-(\alpha+\sum\beta_i X_i)}}=\frac{1}{1+e^{-(-3,616+2,951.1+0,116.20-1,301.1)}}=0,58830$$

Como a probabilidade de 0,58830 (*vide* Figura 12.9) é inferior ao ponto de corte *c* de 0,72, esta observação é classificada no grupo Y=0, ou seja, pertence ao grupo daqueles que continuariam inadimplentes, dado um esforço de cobrança.

O entendimento dos coeficientes do modelo é fundamental. Quando o coeficiente for maior que zero, maior será a probabilidade de ocorrência do evento de interesse, e vice-versa. Observe, por exemplo, que o coeficiente dos indivíduos casados é de 0,116, o que significa que a probabilidade de ocorrência do evento de interesse é aumentada por um fator de 1,123 em relação aos solteiros.

É importante, ainda, destacar que, se os coeficientes do modelo estatístico não fizerem sentido lógico, talvez haja problema de multicolinearidade entre as variáveis independentes, cabendo ao analista/pesquisador investigar tal efeito.

A Figura 12.9 apresenta as probabilidades de ocorrência do evento para cada observação. Note, na primeira linha, o valor de 0,58830 calculado anteriormente e o respectivo *status* desse indivíduo em função do *cutoff* de 0,72.



	id	pagamento	estadocivil	idade	sexo	PRE_1	PGR_1
1	85	Sim	casado	20	F	,58830	Não
2	86	Sim	casado	34	M	,96368	Sim
3	87	Sim	casado	21	M	,86500	Sim
4	88	Sim	casado	22	M	,86881	Sim
5	89	Sim	casado	22	M	,86881	Sim
6	91	Sim	casado	22	M	,86881	Sim
7	92	Sim	casado	23	F	,66938	Não
8	93	Sim	casado	30	M	,94373	Sim
9	94	Sim	casado	30	M	,94373	Sim
10	95	Sim	casado	27	M	,92210	Sim
11	96	Sim	casado	20	M	,84000	Sim
12	97	Sim	casado	20	M	,84000	Sim
13	98	Sim	solteiro	23	M	,28003	Não
14	99	Sim	casado	31	F	,83679	Sim
15	100	Sim	casado	30	M	,94373	Sim
16	101	Sim	casado	39	M	,97947	Sim
17	102	Sim	casado	38	F	,92038	Sim

Figura 12.9: Probabilidades estimadas e classificação prevista.

Porém, em quanto é alterada a probabilidade? Na verdade, depende de onde se inicia o cálculo. Por exemplo, se a chance de saída do *status* de inadimplente é da ordem de 1 para 100, o incremento de 1,123 resulta em:

$$p=\frac{chance}{1+chance}=\frac{(1/100)}{1+(1/100)}=\frac{0,01}{1,01}=0,009$$

$$p = \frac{chance}{1+chance} = \frac{0{,}01123}{1{,}01123} = 0{,}011$$

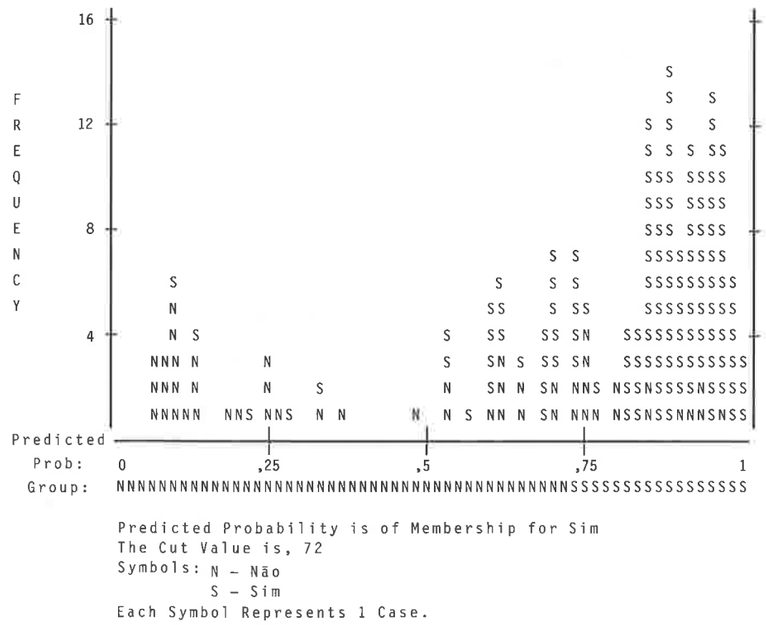
Por outro lado, se a chance de saída do *status* de inadimplente é da ordem de 1 para 1 (probabilidade inicial de 50%), o incremento de 1,123 resulta em:

$$p = \frac{chance}{1+chance} = \frac{(1/1)}{1+(1)} = \frac{1}{2} = 0{,}500$$

$$p = \frac{chance}{1+chance} = \frac{1{,}123}{2{,}123} = 0{,}529$$

No Gráfico 12.2, é apresentada a distribuição de frequência das probabilidades de ocorrência do evento de interesse em relação aos pontos de corte, o que auxilia no estabelecimento de outros pontos de corte.

Gráfico 12.2: Ponto de Corte



Uma análise usual relativa à qualidade do ajuste do modelo é a Curva ROC (*Receiver Operating Characteristic*). Para tanto, clique em **Graphs ROC Curve**. Em **Test Variable**, inclua as probabilidades previstas (variável *PRE_1*) e em **State Variable** inclua a variável dependente (*pagamento*). Em **Value of State Variable**, digite o valor 1, que representa o evento de interesse. Marque as opções **ROC Curve With diagonal reference line** e **Standard error and confidence interval**, conforme ilustrado na Figura 12.10.

Quanto mais distante a Curva ROC estiver da diagonal, melhor será o poder discriminatório do modelo. Conforme apresentado na Tabela 12.11, como a área abaixo da Curva ROC é de 0,846, pode-se dizer, com base no Quadro 12.1, que o modelo proposto apresenta um poder de discriminação considerado excelente.

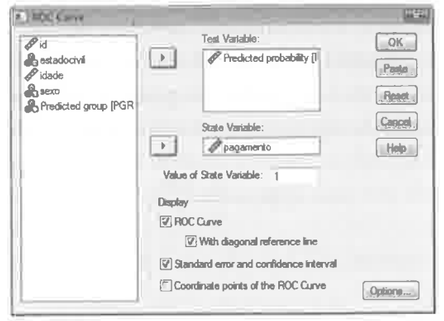


Figura 12.10: Procedimento para elaboração da curva ROC no SPSS.

Tabela 12.11: Área Abaixo da Curva ROC

Area Under the Curve

Test Result Variable(s): Predicted probability

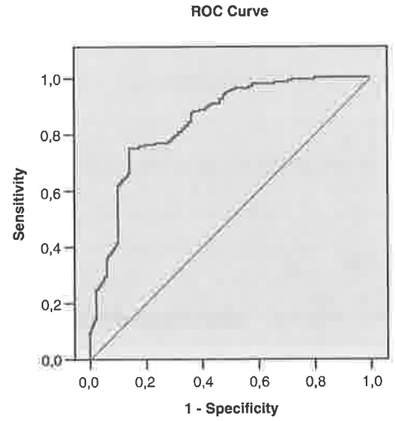
Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,846	,034	,000	,779	,912

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Gráfico 12.3: Curva ROC



Diagonal segments are produced by ties.

As premissas do modelo de regressão logística, relacionadas em seção específica deste capítulo, devem ser testadas seguindo a mesma lógica proposta no capítulo de análise de regressão, cujos testes não foram repetidos aqui. Além disso, para o desenvolvimento do exemplo hipotético proposto, foi utilizado o método *Enter*, mas há alternativas de seleção automática das variáveis relevantes pelo próprio *software*.

Por fim, alguns métodos diagnósticos podem ser elaborados a fim de se avaliar a influência relativa de cada observação da amostra no ajuste do modelo. Desta forma, pode-se solicitar, no menu **Save** da regressão logística, as medidas de distância de *leverage* (hi) e de Cook. Quanto mais próximas de 0 forem estas medidas, melhor, pois há menor influência nos parâmetros por indivíduo.

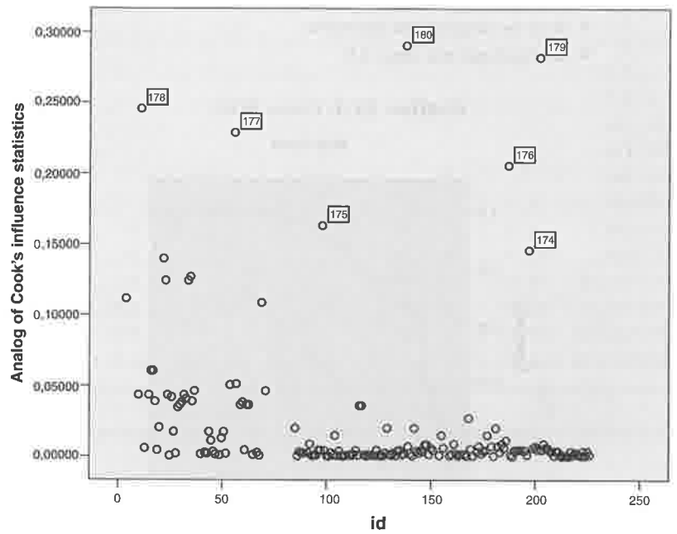
A distância de Cook (*Cook's distance*) é comumente utilizada para se estimar a influência de determinada observação em modelos de regressão, medindo o efeito quando da eliminação de uma observação qualquer. Assim, observações com resíduos elevados ou grandes *leverages* podem distorcer os resultados e a acurácia do modelo de regressão. A expressão da distância de Cook é:

$$CD_i = \frac{Z_i^2 \, x \, h_i}{(1-h_i)}$$

Em que:
CD: *Cook's distance*;
Z_i: resíduos padronizados;
h_i: *leverage*.

Um gráfico de pontos (**Graphs Scatter/Dot**) pode ser elaborado, com as distâncias de Cook no eixo Y e o *id* de cada observação no eixo X. O Gráfico 12.4 a seguir apresenta essa plotagem para os dados do nosso exemplo, com destaque para as observações que apresentam maior influência relativa no ajuste do modelo e que, portanto, merecem maior atenção quando da elaboração de modelos de regressão. Na prática, recomenda-se que observações com distância de Cook próximas de 1 ou até maiores do que este valor sejam eliminadas da amostra.

Gráfico 12.4: Método Diagnóstico – Distância de Cook



Ressalta-se que este método diagnóstico não se aplica somente a modelos de regressão logística, mas também a modelos de regressão linear simples e múltipla.

12.5. UMA INTRODUÇÃO À REGRESSÃO LOGÍSTICA MULTINOMIAL

A regressão logística multinomial trata de um modelo de regressão logística que permite que a variável categórica dependente apresente mais de duas categorias, as quais, por sua vez, podem ser de natureza nominal (ex.: preferência do consumidor por marca de veículo: VW, GM, Fiat, Ford) ou ordinal (ex.: muito satisfeito, satisfeito ou não satisfeito).

Cabe esclarecer que, embora seja possível modelar a variável dependente nominal ou ordinal pela regressão logística multinomial, existe um modelo de regressão logística ordinal que trata especificamente da segunda situação e que não será abordado neste livro.

Na regressão logística multinomial, uma das categorias da variável dependente deve ser escolhida como referência, a fim de compará-la com as demais, e esta escolha pode ou não ser arbitrária, conforme desejo e orientação do pesquisador. É importante notar que isso não altera a forma do modelo, mas apenas o modo de interpretar os parâmetros.

Retornando à função *logit* apresentada na Equação (12.12), temos:

$$Z = logit = \ln\left[\frac{P(Resultado=1|X)}{P(Resultado=0|X)}\right] = \alpha + \Sigma\beta_i X_i \qquad (12.14)$$

A regressão logística multinomial com, por exemplo, três categorias na variável dependente (0, 1 e 2) e com a suposição de que a categoria de referência seja zero, seria dada por:

$$Z = logit = \ln\left[\frac{P(Resultado=1|X)}{P(Resultado=0|X)}\right] = \alpha_1 + \Sigma\beta_{1i} X_{1i} \qquad (12.15)$$

$$Z = logit = \ln\left[\frac{P(Resultado=2|X)}{P(Resultado=0|X)}\right] = \alpha_2 + \Sigma\beta_{2i} X_{2i} \qquad (12.16)$$

Ou seja, a expressão de probabilidade (12.14) agora é apresentada na forma de duas novas expressões que calculam as probabilidades de ocorrência de determinados fenômenos em relação a um fenômeno de referência.

Para entender melhor como esta técnica funciona, faz-se necessário um exemplo prático.

12.6. REGRESSÃO LOGÍSTICA MULTINOMIAL: UM EXEMPLO PRÁTICO

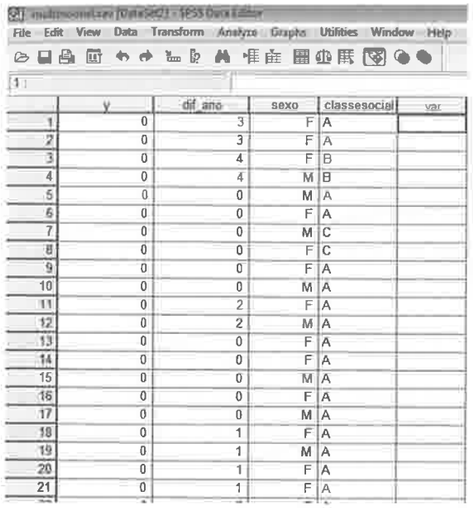
Suponha que um banco de financiamento de uma montadora de automóveis esteja interessado em investir em uma campanha de marketing direto e, para tanto, precisa identificar em sua base de dados o perfil dos clientes que (a) não desejam trocar de carro; (b) desejam trocar de carro, mas pagariam à vista; e (c) desejam trocar de carro, mas financiariam o pagamento. Neste sentido, o modelo apresenta uma variável com as seguintes categorias:

- $y = 0$ para clientes que não desejam trocar de carro;
- $y = 1$ para clientes que desejam trocar de carro, mas pagariam à vista;
- $y = 2$ para clientes que desejam trocar de carro, mas financiariam o pagamento.

As variáveis explicativas disponíveis no banco de dados são:

- *dif_ano*: corresponde à diferença entre o ano base e o ano do veículo;
- *sexo*: sendo 0 para indicar o sexo feminino e 1 para indicar o sexo masculino;
- *classesocial*: classes A, B, C.

Este exemplo, cuja base de dados utilizada apresenta 89 observações e é denominada de **Multinomial.sav** (conforme Figura 12.11), será desenvolvido por meio do *software* SPSS, sendo a rotina em SAS disponibilizada no Apêndice C deste capítulo.



	y	dif_ano	sexo	classesocial	var
1	0	3	F	A	
2	0	3	F	A	
3	0	4	F	B	
4	0	4	M	B	
5	0	0	M	A	
6	0	0	F	A	
7	0	0	M	C	
8	0	0	F	C	
9	0	0	F	A	
10	0	0	M	A	
11	0	2	F	A	
12	0	2	M	A	
13	0	0	F	A	
14	0	0	F	A	
15	0	0	M	A	
16	0	0	F	A	
17	0	0	M	A	
18	0	1	F	A	
19	0	1	M	A	
20	0	1	F	A	
21	0	1	F	A	

Figura 12.11: Base de dados.

Para proceder à rotina de regressão logística multinomial, clique em **Analyze → Regression → Multinomial Logistic**, conforme ilustrado na Figura 12.12.

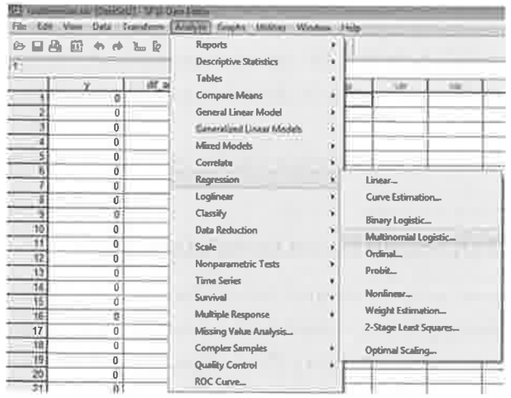


Figura 12.12: Regressão logística multinomial.

Em seguida, como variável dependente, selecione *y*. Em **Reference Category**, o pesquisador deve escolher a categoria de referência da variável dependente. Neste caso, foi selecionado 0 (**Custom**), já que temos a intenção de comparar os clientes com propensão para aquisição de novo veículo, seja o pagamento à vista ou financiado, com os clientes que não têm intenção de trocar o veículo.

O campo **Factor(s)** deve ser preenchido com eventuais variáveis categóricas. No caso, foram selecionadas as variáveis *sexo* e *classesocial*. O campo **Covariate(s)** é utilizado para designar variáveis métricas, sendo, no nosso exemplo, representadas apenas pela variável *dif_ano*.

Intencionalmente, neste exemplo, não foi selecionada nenhuma interação entre as variáveis dependentes, mas, se assim desejar o pesquisador, basta selecionar o menu **Model** e incluir a opção. Além disso, neste item, foi incluído o intercepto no modelo, mas, caso ele não se mostre significativo, basta desmarcá-lo na caixa de diálogo do menu **Model**.

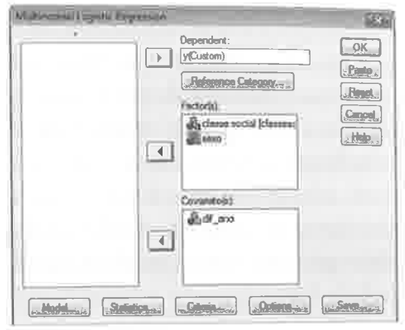


Figura 12.13: Seleção de variáveis.

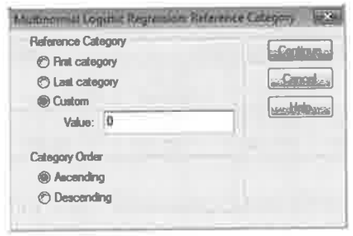


Figura 12.14: Categoria de referência.

Em **Statistics**, selecione as opções indicadas na Figura 12.15.
Em **Save**, marque todas as opções, conforme indicado pela Figura 12.16.

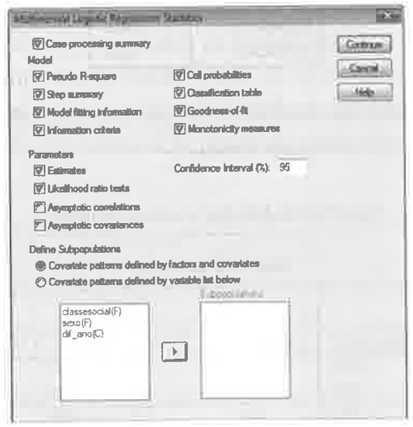


Figura 12.15: Seleção das estatísticas.

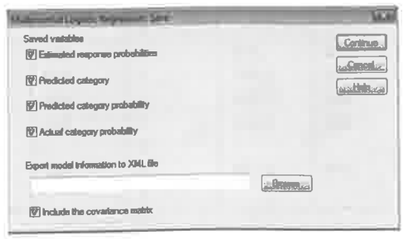


Figura 12.16: Menu **Save**.

Por fim, clique em **OK**.

Na tabela *Model Fitting Information*, são disponibilizadas as informações de AIC (*Akaike Information Criterion*) e BIC (*Bayesian Information Criterion*), critérios de informação que levam em consideração simultaneamente o grau de ajuste do modelo e a parcimônia, sendo indicadores utilizados na comparação de modelos. Como, no caso específico, não foram testados outros modelos de regressão multinomial com outras variáveis ou interação entre as variáveis, não se tem base de comparação acerca de qual o melhor modelo.

Ainda na Tabela 12.12, é mostrado o teste de significância dos coeficientes do modelo final, sendo análogo ao teste F utilizado na análise de regressão. Assim, conforme pode ser observado nesta tabela, o modelo se mostrou significativo ao nível de 5%.

Tabela 12.12: Critério de Informação

Model Fitting Information

	Model Fitting Criteria			Likelihood Ratio Tests		
Model	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	142,834	147,812	138,834			
Final	119,197	144,083	99,197	39,638	8	,000

Em seguida, são apresentados os coeficientes pseudo R^2, que indicam quanto da variação na variável de interesse é capturada pelo modelo, sendo o poder explicativo apenas modesto neste exemplo.

Tabela 12.13: Pseudo R^2

Pseudo R-Square

Cox and Snell	,359
Nagelkerke	,404
McFadden	,203

Na Tabela 12.14, são apresentados os resultados dos parâmetros estimados, sendo a referência o valor de y=0, ou seja, clientes que não trocariam de carro.

Tabela 12.14: Parâmetros Estimados

Parameter Estimates

y[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
1	Intercept	-,853	1,020	,698	1	,403			
	dif_ano	,422	,223	3,591	1	,058	1,525	,986	2,359
	[classesocial=A]	,001	,893	,000	1	,999	1,001	,174	5,759
	[classesocial=B]	,107	,918	,014	1	,907	1,113	,184	6,726
	[classesocial=C]	0[b]	.	.	0	.	.	.	.
	[sexo=0]	,131	,559	,055	1	,815	1,140	,381	3,411
	[sexo=1]	0[b]	.	.	0	.	.	.	.
2	Intercept	,772	,986	,613	1	,434			
	dif_ano	,672	,262	6,607	1	,010	1,959	1,173	3,271
	[classesocial=A]	-2,504	,888	7,956	1	,005	,082	,014	,466
	[classesocial=B]	-2,076	,888	5,462	1	,019	,125	,022	,715
	[classesocial=C]	0[b]	.	.	0	.	.	.	.
	[sexo=0]	-1,382	,678	4,159	1	,041	,251	,066	,948
	[sexo=1]	0[b]	.	.	0	.	.	.	.

a. The reference category is: 0.

b. This parameter is set to zero because it is redundant.

Com base na tabela anterior, os coeficientes das variáveis explicativas não se mostraram significativos, ao nível de 5%, quando comparamos os clientes que trocariam de carro e pagariam à vista (y=1) com os clientes que não trocariam de carro (y=0). Entretanto, quando comparamos os clientes que comprariam e financiariam o automóvel (y=2) com os clientes que não trocariam de carro (y=0), os coeficientes se mostraram significativos, salvo o coeficiente linear. Neste caso, a escolha das variáveis explicativas deve levar em consideração se o pesquisador pretende ou não manter as mesmas categorias da variável y. Em caso afirmativo, os valores dos coeficientes angulares devem ser mantidos. Além disso, observa-se que o coeficiente linear não foi significativo para nenhuma das comparações. Assim, talvez seja melhor excluí-lo do modelo e elaborar o teste novamente. Para fins didáticos, prosseguiremos com a análise dos demais resultados.

As probabilidades de y=0, y=1 e y=2 foram salvas na base de dados, conforme ilustrado na Figura 12.17.

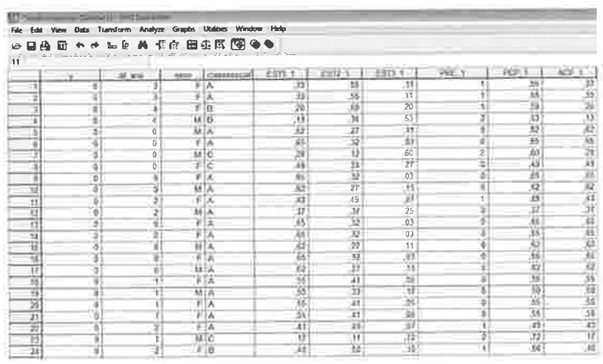


Figura 12.17: Valores preditos.

Para o entendimento de como foram calculadas as probabilidades, é preciso considerar os resultados da Tabela 12.14. Além disso, é importante lembrar que os resultados da referida tabela comparam y=1 com y=0 e y=2 com y=0. Em outras palavras, adaptando a Equação (12.11), pode-se calcular o *Risk Odds Ratio* (*ROR*) pela razão do *odds* (chance) entre dois grupos (y=1 e y=0) e pela razão do *odds* entre dois grupos (y=2 e y=0), como segue:

$$g(1) = ROR_{y=1,\, y=0} = \frac{odds(y=1)}{odds(y=0)} = e^{\alpha_1 + \sum \beta_{1i} X_{1i}} \tag{12.17}$$

$$g(2) = ROR_{y=2,\, y=0} = \frac{odds(y=2)}{odds(y=0)} = e^{\alpha_2 + \sum \beta_{2i} X_{2i}} \tag{12.18}$$

Assim, o cálculo de probabilidades deve levar em consideração as probabilidades individuais denotadas pela taxa de risco (*Risk Odds Ratio*):

$$P(y=0) = \frac{P(0)}{P(0)+P(1)+P(2)} \tag{12.19}$$

A fórmula anterior é análoga a:

$$P(y=0) = \frac{g(0)}{g(0)+g(1)+g(2)} \tag{12.20}$$

$$P(y=1) = \frac{g(1)}{g(0)+g(1)+g(2)} \tag{12.21}$$

$$P(y=2) = \frac{g(2)}{g(0)+g(1)+g(2)} \tag{12.22}$$

Como $g(0)$ é igual a 1, por definição, pois é a relação de *odds* de $y = 0$ com ele mesmo, então:

$$P(y=0) = \frac{1}{1+g(1)+g(2)} \tag{12.23}$$

Para ilustrar o cálculo de probabilidades, tomamos como exemplo o primeiro registro do banco de dados, o qual apresenta as seguintes características:

- $y = 0$;
- *dif_ano* = 3;
- *sexo* = F;
- *classesocial* = A.

Como base na Tabela 12.3, pode-se calcular $g(1)$ e $g(2)$ como segue:

$$g(1) = e^{-0,85257+0,421888.3+0,130636+0,000751} = 1,72372$$
$$g(2) = e^{0,771973+0,672414.3-1,38225-2,50375} = 0,333944$$

Assim, pode-se calcular as probabilidades $P(0)$, $P(1)$ e $P(2)$, indicadas pelo SPSS, de *EST1_1*, *EST2_1* e *EST3_1*, respectivamente, da seguinte maneira:

$$P(y=0) = \frac{1}{1+1,72372+0,333944} = 0,32704705$$

$$P(y=1) = \frac{1,72372}{1+1,72372+0,333944} = 0,56373749$$

$$P(y=2) = \frac{0,333944}{1+1,72372+0,333944} = 0,10921546$$

Como a maior probabilidade é 0,56373749, referente a $P(y=1)$, então estima-se que a primeira observação pertença à categoria 1, ou seja, ao grupo daqueles que comprariam outro veículo, mas pagariam à vista. Portanto, a variável *PRE_1* fornecida pelo SPSS é igual a 1. Portanto, neste caso, o PCP (*predicted category probability*) é de 0,56373749, embora o ACP (*actual category probability*) seja de 0,32704705, tendo em vista que a categoria original deste registro era 0 (daqueles que não comprariam outro veículo).

Por fim, a seguir é apresentada a tabela de classificação, mostrando o grau de acerto do modelo para cada categoria e a *performance* geral, que é de 63,3%.

Tabela 12.15: Classificação Prevista *versus* Observada

Classification

	Predicted			
Observed	0	1	2	Percent Correct
0	16	8	5	55,2%
1	5	20	5	66,7%
2	5	2	23	76,7%
Overall Percentage	29,2%	33,7%	37,1%	66,3%

12.7. RELAÇÃO COM OUTRAS TÉCNICAS

A regressão logística é utilizada para prever o comportamento de uma variável dependente categórica binária. É importante observar que não se trata de um problema objeto de investigação da regressão múltipla, já que este último apresenta, como pressupostos, a homogeneidade de variância e a normalidade dos resíduos, entre outros.

Duas outras técnicas também são muito utilizadas para prever o comportamento de variáveis categóricas: a análise discriminante (estudada no Capítulo 11) e a análise de sobrevivência, ou Modelo de Riscos Proporcionais (a ser estudado no Capítulo 15). Embora a análise discriminante seja uma técnica robusta para tal fim, ela requer a assunção de inúmeras premissas para a validade do modelo, restringindo, portanto, as situações passíveis de sua utilização. Além disso, trata-se de uma técnica que não dispõe diretamente da probabilidade de ocorrência do evento de interesse. Na prática, muitos problemas podem ser modelados tanto por meio da análise discriminante quanto por meio da regressão logística. Porém, quando tivermos em um mesmo modelo variáveis explicativas com escalas de mensuração qualitativa e quantitativa, a premissa de normalidade multivariada não será atendida na análise discriminante.

A análise de sobrevivência, por sua vez, diferencia-se da técnica tratada neste capítulo por considerar o tempo para a ocorrência do evento de interesse, que não é objetivo de investigação da regressão logística.

Cabe ainda ressaltar que, na prática, muitas são as situações em que as três técnicas são aplicáveis, sendo necessário avaliar qual o melhor modelo que retrata a realidade subjacente.

12.8. CONSIDERAÇÕES FINAIS

A regressão logística é uma técnica multivariada de dependência destinada a identificar as variáveis mais significativas para previsão da ocorrência de determinado evento de interesse, provendo inclusive a probabilidade de sua ocorrência. Trata-se de uma técnica muito difundida em diversos campos do conhecimento humano, principalmente em função da facilidade de sua aplicação e da flexibilização de seus pressupostos, se comparados a outras técnicas, como regressão múltipla e análise discriminante.

Segundo Hair, Anderson, Tatham e Black (2005), a regressão logística pode ser preferida em relação à análise discriminante por diversas razões. Primeiramente, como já discutido, "a análise discriminante depende estritamente de se atenderem as suposições de normalidade multivariada e de igualdade de matrizes de variância-covariância nos grupos – suposições que não são atendidas em muitas situações. A regressão logística não depende dessas suposições rígidas e é muito mais robusta quando tais pressupostos não são satisfeitos, o que torna sua aplicação apropriada em muito mais situações. Segundo, mesmo quando os pressupostos são satisfeitos, muitos pesquisadores preferem a regressão logística por ser similar à regressão. Ambas têm testes estatísticos diretos e a habilidade de incorporar efeitos não-lineares e uma vasta gama de diagnósticos. Por estas e outras razões, a regressão logística é equivalente à análise discriminante de dois grupos e pode ser mais adequada em muitas situações".

Embora sua forma mais simples comporte apenas variáveis dependentes dicotômicas, há variantes, como a regressão logística multinomial, que admitem mais de duas possibilidades como variável resposta.

A popularidade das técnicas de regressão logística e de regressão logística multinomial vem crescendo de forma bastante perceptível nos últimos anos em diversas áreas do conhecimento, já que têm se mostrado eficazes na solução de problemas que envolvem a escolha de um evento de interesse e quando há o desejo de se investigar a probabilidade de ocorrência deste evento e quais as variáveis representativas para sua explicação. O estudo de probabilidades de sucesso ou fracasso no lançamento de novos produtos, de propensão à inadimplência por parte de consumidores quando da aquisição de crédito, de suscetibilidade à falência de empresas de pequeno e médio porte, de surgimento de novas doenças por

parte de pessoas sedentárias, de escolha de um plano de saúde por famílias de baixa renda, entre outros, são apenas alguns poucos exemplos, entre tantos, que podem ser modelados por pesquisadores com o uso destas técnicas.

12.9. EXERCÍCIOS – APLICAÇÃO DE BANCOS DE DADOS

1. Refaça a análise de regressão multinomial com o arquivo **Multinomial.sav**, excluindo o intercepto. Analise os resultados.
2. Utilizando o arquivo **Multinomial.sav**, refaça a análise de regressão logística, selecionando somente os registros cujos valores de *y* são 0 ou 2. Compare os resultados com os obtidos ao longo do capítulo.
3. Refaça a análise de regressão logística com base no arquivo **Logistica.sav** pelo método *forward* Wald. Analise e compare os resultados.
4. O arquivo **Bankloan.sav** apresenta oito variáveis referentes a 500 pessoas que são clientes de uma financeira. Por meio de uma regressão logística, pede-se:
   a) Quais variáveis são significativas para se elaborar uma boa previsão de risco de *default*?
   b) Elabore novamente, sem as variáveis que apresentaram problemas de significância (estatística de Wald);
   c) Interprete os *outputs* da técnica;
   d) Elabore o diagnóstico dos resíduos e da distância de Cook;
   e) Elabore uma curva ROC e interprete-a;
   f) Calcule a probabilidade de *default* de um indivíduo com as seguintes características:
      - Idade = 40 anos
      - Nível de educação = 3
      - Emprego atual = 3 anos
      - Endereço atual = 5 anos
      - Renda familiar anual (em milhares) = $ 60,00
      - Endividamento = 17%
      - Dívida do cartão de crédito (em milhares) = $ 0,70
      - Outras dívidas (em milhares) = $ 3,00

12.10. RESUMO

A regressão logística é uma técnica estatística utilizada para descrever a relação entre uma variável dependente categórica binária e variáveis independentes métricas ou não métricas.

Tal relação é expressa em termos de probabilidades de ocorrência diante das variáveis pelas quais os objetos ou sujeitos estão expostos, sendo possível abstrair a taxa de risco de cada variável explicativa no modelo.

O principal objetivo desta técnica é encontrar uma função logística, formada por meio de ponderações das variáveis (atributos), cuja resposta permita estabelecer a probabilidade de ocorrência de determinado evento e a importância das variáveis (peso) para essa ocorrência.

A regressão logística requer que as variáveis do vetor X sejam quantitativas ou dicotômicas, que a relação entre o vetor *X* e a variável *Y* seja linear, que o valor esperado dos resíduos seja zero e que haja ausência de heterocedasticidade e de multicolinearidade.

Assim, a regressão logística destina-se a aferir a probabilidade de ocorrência de um evento e a identificar características dos elementos pertencentes a cada um dos dois grupos especificados pela variável categórica binária.

A regressão logística multinomial, por sua vez, trata de um modelo de regressão logística que permite que a variável categórica dependente apresente mais de duas categorias que podem ser de natureza nominal ou ordinal.

Na regressão logística multinomial, uma das categorias da variável dependente deve ser escolhida como referência. Dessa forma, o pesquisador pode ter interesse no cálculo de probabilidades de dois ou mais fenômenos (categorias) em relação a um fenômeno representado por uma categoria de referência.

12.11. QUESTÕES COMPLEMENTARES

a) Defina regressão logística. Qual é o objetivo da técnica?
b) Qual a diferença entre regressão logística e regressão logística multinomial?
c) Indique três situações de pesquisa em que poderia ser utilizada a técnica regressão logística.
d) Dê dois exemplos de situações em que poderia ser utilizada a técnica de regressão logística multinomial.
e) Defina brevemente *odds* e curva ROC.
f) Quais são as vantagens e desvantagens da utilização da técnica de regressão logística diante da análise discriminante?
g) Em que consiste o teste de Hosmer-Lemeshow?
h) Para que serve o método diagnóstico que utiliza a distância de Cook?