

# Modelos de Regressão Logística Binária e Multinomial

*Nos campos da observação, a chance favorece apenas a mente preparada.*

Louis Pasteur

Ao final deste capítulo, você terá condições de:

- Estabelecer as circunstâncias a partir das quais os modelos de regressão logística binária e multinomial podem ser utilizados.
- Diferenciar a probabilidade de ocorrência de um evento da chance de ocorrência de um evento.
- Entender a estimação pelo método de máxima verossimilhança.
- Avaliar os resultados dos testes estatísticos pertinentes aos modelos logísticos.
- Elaborar intervalos de confiança dos parâmetros do modelo para efeitos de previsão.
- Elaborar a análise de sensibilidade e entender os conceitos de *cutoff*, eficiência global do modelo, sensibilidade e especificidade.
- Interpretar a curva de sensibilidade e a curva *ROC*.
- Elaborar modelos de regressão logística binária e multinomial em Microsoft Office Excel®, Stata Statistical Software® e IBM SPSS Statistics Software® e interpretar seus resultados.

## 13.1. INTRODUÇÃO

Os modelos de regressão logística, embora bastante úteis e de fácil aplicação, ainda são pouco utilizados em muitas áreas do conhecimento humano. Embora o desenvolvimento de softwares e o incremento da capacidade de processamento dos computadores tenham propiciado a sua aplicação de forma mais direta, muitos pesquisadores ainda desconhecem as suas utilidades e, sobretudo, as condições para que seu uso seja correto.

Diferentemente da tradicional técnica de regressão estimada por meio de métodos de mínimos quadrados, em que a variável dependente apresenta-se de forma quantitativa e devem ser obedecidos alguns pressupostos, conforme estudamos no capítulo anterior, as técnicas de regressão logística são utilizadas quando o fenômeno a ser estudado apresenta-se de forma qualitativa e, portanto, representado por uma ou mais variáveis *dummy*, dependendo da quantidade de possibilidades de resposta (categorias) desta variável dependente.

Imagine, por exemplo, que um pesquisador tenha interesse em avaliar a probabilidade de ocorrência de infarto em executivos do mercado financeiro, com base em suas características físicas (peso, cintura abdominal), em seus hábitos alimentares e em seus hábitos de saúde (exercícios físicos, tabagismo). Um segundo pesquisador deseja avaliar a chance de consumidores que adquirem bens duráveis num determinado período tornarem-se inadimplentes, em função da renda, do estado civil e da escolaridade de cada um deles. Note que o infarto ou a inadimplência são as variáveis dependentes nos dois casos e seus eventos podem ou não ocorrer, em função das variáveis explicativas inseridas nos respectivos modelos e, portanto, são variáveis qualitativas dicotômicas que representam cada um dos fenômenos em estudo. Nosso intuito é o de estimar a **probabilidade de ocorrência** destes fenômenos e, para tanto, faremos uso da **regressão logística binária**.

Imagine ainda que um terceiro pesquisador tenha o interesse em estudar a probabilidade de obtenção de crédito por parte de empresas de micro e pequeno porte, em função de suas características financeiras e operacionais. Sabe-se que cada empresa poderá receber crédito integral sem restrição, crédito com restrição ou não receber

crédito algum. Neste caso, a variável dependente que representa o fenômeno é também qualitativa, porém oferece três possibilidades de resposta (categorias), e portanto, para estimarmos as probabilidades de ocorrência das alternativas propostas, deveremos fazer uso da **regressão logística multinomial**.

Logo, se um fenômeno em estudo se apresentar por meio de apenas e tão somente duas categorias, será representado por apenas uma única variável *dummy*, em que a primeira categoria será a de referência e indicará o não evento de interesse (*dummy* = 0) e a outra categoria indicará o evento de interesse (*dummy* = 1), e estaremos lidando com a técnica de regressão logística binária. Por outro lado, se o fenômeno em estudo apresentar mais de duas categorias como possibilidades de ocorrência, precisaremos inicialmente definir a categoria de referência para, a partir daí, elaborar a técnica de regressão logística multinomial.

Ao se ter uma variável qualitativa como fenômeno a ser estudado, fica inviável a estimação do modelo por meio do método de mínimos quadrados ordinários estudado no capítulo anterior, uma vez que esta variável dependente não apresenta média e variância e, portanto, não há como minimizar a somatória dos termos de erro ao quadrado sem que seja feita uma incoerente ponderação arbitrária. Como a inserção desta variável dependente em softwares de modelagem é feita com base na digitação de valores que representam cada uma das possibilidades de resposta, é comum que haja um esquecimento sobre a definição dos rótulos (*labels*) das categorias correspondentes a cada um dos valores digitados e, portanto, é possível que um pesquisador desavisado ou iniciante estime o modelo por meio da regressão por mínimos quadrados, inclusive obtendo *outputs*, uma vez que o software interpretará aquela variável dependente como sendo quantitativa. **Isso é um erro grave, porém infelizmente mais comum do que parece!** As técnicas de regressão logística binária e multinomial são elaboradas com base na **estimação por máxima verossimilhança**, a ser estudada nas seções 13.2.1 e 13.3.1, respectivamente.

Analogamente ao que foi discutido no capítulo anterior, os modelos de regressão logística são definidos com base na teoria subjacente e na experiência do pesquisador, de modo que seja possível estimar o modelo desejado, analisar os resultados obtidos por meio de testes estatísticos e elaborar previsões.

Neste capítulo, trataremos dos modelos de regressão logística binária e multinomial, com os seguintes objetivos: (1) introduzir os conceitos sobre regressão logística; (2) apresentar a estimação por máxima verossimilhança; (3) interpretar os resultados obtidos e elaborar previsões; e (4) apresentar a aplicação das técnicas em Excel, Stata e SPSS. Inicialmente, será elaborada a solução em Excel de um exemplo concomitantemente à apresentação dos conceitos e à sua resolução manual. Após a introdução dos conceitos serão apresentados os procedimentos para a elaboração das técnicas no Stata e no SPSS, mantendo o padrão adotado no livro.

## 13.2. O MODELO DE REGRESSÃO LOGÍSTICA BINÁRIA

A regressão logística binária tem como objetivo principal estudar a probabilidade de ocorrência de um evento definido por  $Y$  que se apresenta na forma qualitativa dicotômica ( $Y = 1$  para descrever a ocorrência do evento de interesse e  $Y = 0$  para descrever a ocorrência do não evento), com base no comportamento de variáveis explicativas. Desta forma, podemos definir um vetor de variáveis explicativas, com respectivos parâmetros estimados, da seguinte forma:

$$Z_i = \alpha + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \dots + \beta_k \cdot X_{ki} \quad (13.1)$$

em que  $Z$  é conhecido por **logito**,  $\alpha$  representa a constante,  $\beta_j$  ( $j = 1, 2, \dots, k$ ) são os parâmetros estimados de cada variável explicativa,  $X_j$  são as variáveis explicativas (métricas ou *dummies*) e o subscrito  $i$  representa cada observação da amostra ( $i = 1, 2, \dots, n$ , em que  $n$  é o tamanho da amostra). É importante ressaltar que  $Z$  não representa a variável dependente, denominada por  $Y$ , e o nosso objetivo neste momento é definir a expressão da **probabilidade  $p_i$**  de ocorrência do evento de interesse para cada observação, em função do logito  $Z_i$ , ou seja, em função dos parâmetros estimados para cada variável explicativa. Para tanto, devemos definir o conceito de **chance** de ocorrência de um evento, também conhecida por **odds**, da seguinte forma:

$$\text{chance (odds)}_{Y_i=1} = \frac{p_i}{1 - p_i} \quad (13.2)$$

Imagine que tenhamos o interesse em estudar o evento “aprovação na disciplina de Cálculo”. Se, por exemplo, a probabilidade de um determinado aluno ser aprovado nesta disciplina for de 80%, a sua chance de ser aprovado será de 4 para 1 ( $0,8 / 0,2 = 4$ ). Se a probabilidade de outro aluno ser aprovado na mesma disciplina for de 25%, dado que tem estudado muito menos que o primeiro aluno, a sua chance de ser aprovado será de 1 para 3

(0,25 / 0,75 = 1/3). Apesar de estarmos acostumados cotidianamente a usar o termo **chance** como sinônimo de **probabilidade**, seus conceitos são diferentes!

A regressão logística binária define o logito  $Z$  como o logaritmo natural da chance, de modo que:

$$\ln(\text{chance}_{Y_i=1}) = Z_i \quad (13.3)$$

de onde vem que:

$$\ln\left(\frac{p_i}{1-p_i}\right) = Z_i \quad (13.4)$$

Como o nosso intuito é definir uma expressão para a probabilidade de ocorrência do evento em estudo em função do logito, podemos matematicamente isolar  $p_i$  a partir da expressão (13.4), da seguinte maneira:

$$\frac{p_i}{1-p_i} = e^{Z_i} \quad (13.5)$$

$$p_i = (1-p_i) \cdot e^{Z_i} \quad (13.6)$$

$$p_i \cdot (1 + e^{Z_i}) = e^{Z_i} \quad (13.7)$$

E, portanto, temos que:

**Probabilidade de ocorrência do evento:**

$$p_i = \frac{e^{Z_i}}{1 + e^{Z_i}} = \frac{1}{1 + e^{-Z_i}} \quad (13.8)$$

**Probabilidade de ocorrência do não evento:**

$$1 - p_i = 1 - \frac{e^{Z_i}}{1 + e^{Z_i}} = \frac{1}{1 + e^{Z_i}} \quad (13.9)$$

Obviamente, a soma das expressões (13.8) e (13.9) é igual a 1.

A partir da expressão (13.8), podemos elaborar uma tabela com valores de  $p$  em função dos valores de  $Z$ . Como  $Z$  varia de  $-\infty$  a  $+\infty$ , iremos, apenas para efeitos didáticos, utilizar valores inteiros entre -5 e +5. A Tabela 13.1 traz estes valores.

**Tabela 13.1** Probabilidade de ocorrência de um evento ( $p$ ) em função do logito  $Z$ .

$p_i = \frac{1}{1 + e^{-Z_i}}$	$Z_i$
0,0067	-5
0,0180	-4
0,0474	-3
0,1192	-2
0,2689	-1
0,5000	0
0,7311	1
0,8808	2
0,9526	3
0,9820	4
0,9933	5

A partir da Tabela 13.1, podemos elaborar um gráfico de  $p = f(Z)$ , como o apresentado na Figura 13.1. Por meio deste gráfico, podemos verificar que as probabilidades estimadas, em função dos diversos valores assumidos por  $Z$ , situam-se entre 0 e 1, o que foi garantido quando se impôs que o logito fosse igual ao logaritmo natural da chance. Assim, dados os parâmetros estimados do modelo e os valores de cada uma das variáveis explicativas para uma dada observação  $i$ , podemos calcular o valor de  $Z_i$  e, por meio da curva logística apresentada na Figura 13.1 (também conhecida por curva  $S$ , ou sigmoide), estimar a probabilidade de ocorrência do evento em estudo para esta determinada observação  $i$ .

A partir das expressões (13.1) e (13.8), podemos definir a expressão geral da probabilidade estimada de ocorrência de um evento que se apresenta na forma dicotômica para uma observação  $i$  da seguinte forma:

$$p_i = \frac{1}{1 + e^{-(\alpha + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \dots + \beta_k \cdot X_{ki})}} \quad (13.10)$$

O que a regressão logística binária estima, portanto, não são os valores previstos da variável dependente, mas, sim, a probabilidade de ocorrência do evento em estudo para cada observação. Partiremos, então, para a estimação propriamente dita dos parâmetros do logito, por meio da apresentação de um exemplo elaborado inicialmente em Excel.

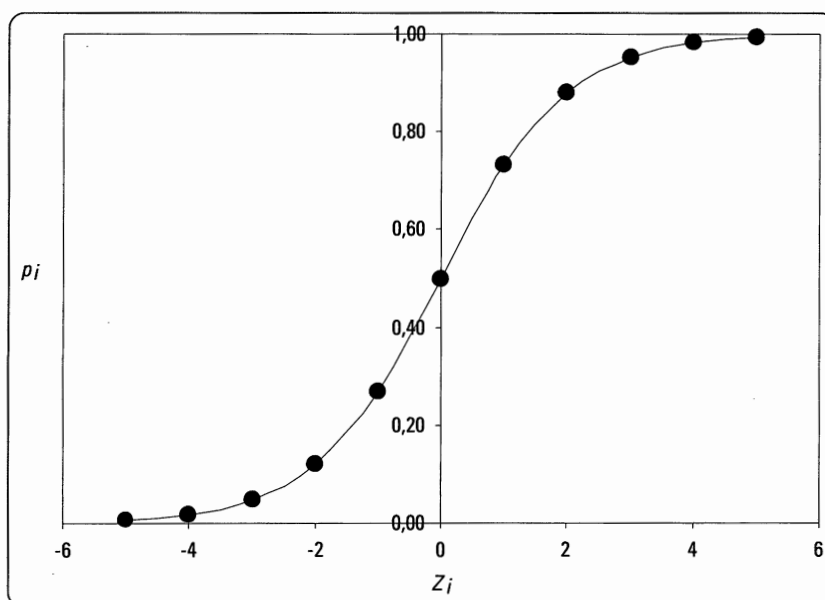


Figura 13.1 Gráfico de  $p = f(Z)$ .

### 13.2.1. Estimação do modelo de regressão logística binária por máxima verossimilhança

Apresentaremos os conceitos pertinentes à **estimação por máxima verossimilhança** por meio de um exemplo similar ao desenvolvido ao longo do capítulo anterior. Entretanto, agora a variável dependente será qualitativa e dicotômica.

Imagine que o nosso curioso professor, que já explorou consideravelmente os efeitos de determinadas variáveis explicativas sobre o tempo de deslocamento de um grupo de alunos até a escola, por meio da técnica de regressão múltipla, tenha agora o interesse em investigar se estas mesmas variáveis explicativas influenciam a probabilidade de um aluno chegar atrasado à aula. Ou seja, o fenômeno em questão a ser estudado apresenta somente duas categorias (chegar ou não atrasado) e o evento de interesse refere-se a *chegar atrasado*.

Sendo assim, o professor elaborou uma pesquisa com 100 alunos da escola onde leciona, questionando se cada um deles chegou ou não atrasado naquele dia. Perguntou também sobre a distância percorrida no trajeto (em quilômetros), o número de semáforos pelos quais cada um passou, o período em que foi realizado o trajeto (manhã ou tarde) e como cada um se considera em termos de perfil ao volante (calmo, moderado ou agressivo). Parte do banco de dados elaborado encontra-se na Tabela 13.2.

**Tabela 13.2** Exemplo: atraso (sim ou não) x distância percorrida, quantidade de semáforos, período do dia para o trajeto até a escola e perfil ao volante.

Estudante	Chegou atrasado à escola ( $Y_i$ )	Distância percorrida até a escola (quilômetros) ( $X_{1i}$ )	Quantidade de semáforos ( $X_{2i}$ )	Período do dia ( $X_{3i}$ )	Perfil ao volante ( $X_{4i}$ )
Gabriela	Não	12,5	7	manhã	calmo
Patrícia	Não	13,3	10	manhã	calmo
Gustavo	Não	13,4	8	manhã	moderado
Letícia	Não	23,5	7	manhã	calmo
Luiz Ovídio	Não	9,5	8	manhã	calmo
Leonor	Não	13,5	10	manhã	calmo
Dalila	Não	13,5	10	manhã	calmo
Antônio	Não	15,4	10	manhã	calmo
Júlia	Não	14,7	10	manhã	calmo
Mariana	Não	14,7	10	manhã	calmo
...					
Filomena	Sim	12,8	11	tarde	agressivo
...					
Estela	Sim	1,0	13	manhã	calmo

Para a variável dependente, como o evento de interesse refere-se a *chegar atrasado*, esta categoria apresentará valores iguais a 1, ficando a categoria *não chegar atrasado* com valores iguais a 0.

Seguindo o que foi definido no capítulo anterior em relação às variáveis explicativas qualitativas, a categoria de referência da variável correspondente ao período do dia será *tarde*, ou seja, as células do banco de dados com esta categoria assumirão valores iguais a 0, ficando as células com a categoria *manhã* com valores iguais a 1. Já a variável *perfil ao volante* deverá ser transformada em duas *dummies* (variáveis *perfil2* para a categoria *moderado* e *perfil3* para a categoria *agressivo*), já que definiremos a categoria *calmo* como sendo a referência.

Desta forma, a Tabela 13.3 apresenta parte do banco de dados final a ser utilizado para a estimação do modelo de regressão logística binária.

O banco de dados completo pode ser acessado por meio do arquivo **Atrasado.xls**.

**Tabela 13.3** Substituição das categorias das variáveis qualitativas pelas respectivas variáveis *dummy*.

Estudante	Chegou atrasado à escola ( <i>Dummy</i> Sim = 1; Não = 0) ( $Y_i$ )	Distância percorrida até a escola (quilômetros) ( $X_{1i}$ )	Quantidade de semáforos ( $X_{2i}$ )	Período do dia <i>Dummy per</i> ( $X_{3i}$ )	Perfil ao volante <i>Dummy perfil2</i> ( $X_{4i}$ )	Perfil ao volante <i>Dummy perfil3</i> ( $X_{5i}$ )
Gabriela	0	12,5	7	1	0	0
Patrícia	0	13,3	10	1	0	0
Gustavo	0	13,4	8	1	1	0
Letícia	0	23,5	7	1	0	0
Luiz Ovídio	0	9,5	8	1	0	0
Leonor	0	13,5	10	1	0	0
Dalila	0	13,5	10	1	0	0
Antônio	0	15,4	10	1	0	0
Júlia	0	14,7	10	1	0	0
Mariana	0	14,7	10	1	0	0
...						
Filomena	1	12,8	11	0	0	1
...						
Estela	1	1,0	13	1	0	0

Desta forma, o logito cujos parâmetros queremos estimar é definido da seguinte maneira:

$$Z_i = \alpha + \beta_1 \cdot \text{dist}_i + \beta_2 \cdot \text{sem}_i + \beta_3 \cdot \text{per}_i + \beta_4 \cdot \text{perfil2}_i + \beta_5 \cdot \text{perfil3}_i$$

e a probabilidade estimada de que um determinado estudante chegue atrasado pode ser escrita da seguinte forma:

$$p_i = \frac{1}{1 + e^{-(\alpha + \beta_1 \cdot \text{dist}_i + \beta_2 \cdot \text{sem}_i + \beta_3 \cdot \text{per}_i + \beta_4 \cdot \text{perfil2}_i + \beta_5 \cdot \text{perfil3}_i)}}$$

Como não faz sentido definirmos o termo de erro para cada observação, dado que a variável dependente apresenta-se na forma dicotômica, não há como estimarmos os parâmetros da equação de probabilidade por meio da minimização da somatória dos quadrados dos resíduos, como fizemos quando da elaboração das técnicas tradicionais de regressão. Neste caso, portanto, faremos uso da função de verossimilhança a partir da qual será elaborada a estimação por máxima verossimilhança. Segundo Sharma (1996), a estimação por máxima verossimilhança é a técnica mais popular de estimação dos parâmetros de modelos de regressão logística.

Em decorrência deste fato, é importante inclusive mencionar, com relação aos pressupostos estudados para os modelos de regressão estimados por mínimos quadrados ordinários, que o pesquisador deve se preocupar apenas com o pressuposto da ausência de multicolinearidade das variáveis explicativas quando da estimação de modelos de regressão logística.

Na regressão logística binária, a variável dependente segue uma **distribuição de Bernoulli**, ou seja, o fato de determinada observação  $i$  ter incidido ou não no evento de interesse pode ser considerado como um ensaio de Bernoulli, em que a probabilidade de ocorrência do evento é  $p_i$  e a probabilidade de ocorrência do não evento é  $(1 - p_i)$ , conforme estudamos no Capítulo 5. De maneira geral, analogamente à expressão (5.25) daquele capítulo, podemos escrever que a probabilidade de ocorrência de  $Y_i$ , podendo  $Y_i$  ser igual a 1 ou igual a 0, é dada por:

$$p(Y_i) = p_i^{Y_i} \cdot (1 - p_i)^{1 - Y_i} \quad (13.11)$$

Para uma amostra com  $n$  observações, podemos definir a função de verossimilhança (*likelihood function*) como sendo:

$$L = \prod_{i=1}^n [p_i^{Y_i} \cdot (1 - p_i)^{1 - Y_i}] \quad (13.12)$$

de onde vem, com base nas expressões (13.8) e (13.9), que:

$$L = \prod_{i=1}^n \left[ \left( \frac{e^{Z_i}}{1 + e^{Z_i}} \right)^{Y_i} \cdot \left( \frac{1}{1 + e^{Z_i}} \right)^{1 - Y_i} \right] \quad (13.13)$$

Como, na prática, é mais conveniente se trabalhar com o logaritmo da função de verossimilhança, podemos chegar à seguinte função, também conhecida por *log likelihood function*:

$$LL = \sum_{i=1}^n \left\{ \left[ (Y_i) \cdot \ln \left( \frac{e^{Z_i}}{1 + e^{Z_i}} \right) \right] + \left[ (1 - Y_i) \cdot \ln \left( \frac{1}{1 + e^{Z_i}} \right) \right] \right\} \quad (13.14)$$

E agora cabe uma pergunta: **Quais os valores dos parâmetros do logito que fazem com que o valor de LL da expressão (13.14) seja maximizado?** Esta importante questão é a chave central para a elaboração da estimação por máxima verossimilhança (ou *maximum likelihood estimation*) em modelos de regressão logística binária, e pode ser respondida com o uso de ferramentas de programação linear, a fim de que sejam estimados os parâmetros  $\alpha, \beta_1, \beta_2, \dots, \beta_k$  com base na seguinte função-objetivo:

$$LL = \sum_{i=1}^n \left\{ \left[ (Y_i) \cdot \ln \left( \frac{e^{Z_i}}{1 + e^{Z_i}} \right) \right] + \left[ (1 - Y_i) \cdot \ln \left( \frac{1}{1 + e^{Z_i}} \right) \right] \right\} = \text{máx} \quad (13.15)$$

Iremos resolver este problema com o uso da ferramenta **Solver** do Excel e utilizando os dados do nosso exemplo. Para tanto, devemos abrir o arquivo **AtrasadoMáximaVerossimilhança.xls**, que servirá de auxílio para o cálculo dos parâmetros.

Neste arquivo, além da variável dependente e das variáveis explicativas, foram criadas três novas variáveis, que correspondem, respectivamente, ao logito  $Z_i$ , à probabilidade de ocorrência do evento de interesse  $p_i$  e ao logaritmo da função de verossimilhança  $LL_i$  para cada observação. A Tabela 13.4 mostra parte dos resultados quando os parâmetros  $\alpha$ ,  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$ ,  $\beta_4$  e  $\beta_5$  forem iguais a 0.

A Figura 13.2 apresenta parte das observações presentes no arquivo **AtrasadoMáximaVerossimilhança.xls**, já que algumas delas foram aqui ocultadas por conta do número total ser igual a 100.

**Tabela 13.4** Cálculo de  $LL$  quando  $\alpha = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ .

Estudante	$Y_i$	$X_{1i}$	$X_{2i}$	$X_{3i}$	$X_{4i}$	$X_{5i}$	$Z_i$	$p_i$	$LL_i$ $(Y_i) \cdot \ln(p_i) + (1 - Y_i) \cdot \ln(1 - p_i)$
Gabriela	0	12,5	7	1	0	0	0	0,5	-0,69315
Patrícia	0	13,3	10	1	0	0	0	0,5	-0,69315
Gustavo	0	13,4	8	1	1	0	0	0,5	-0,69315
Letícia	0	23,5	7	1	0	0	0	0,5	-0,69315
Luiz Ovídio	0	9,5	8	1	0	0	0	0,5	-0,69315
Leonor	0	13,5	10	1	0	0	0	0,5	-0,69315
Dalila	0	13,5	10	1	0	0	0	0,5	-0,69315
Antônio	0	15,4	10	1	0	0	0	0,5	-0,69315
Júlia	0	14,7	10	1	0	0	0	0,5	-0,69315
Mariana	0	14,7	10	1	0	0	0	0,5	-0,69315
...									
Filomena	1	12,8	11	0	0	1	0	0,5	-0,69315
...									
Estela	1	1,0	13	1	0	0	0	0,5	-0,69315
<b>Somatória</b>	$LL = \sum_{i=1}^{100} [(Y_i) \cdot \ln(p_i)] + [(1 - Y_i) \cdot \ln(1 - p_i)]$								<b>-69,31472</b>

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Estudante	Atrasado (Y)	Distância (X <sub>1</sub> )	Semáforos (X <sub>2</sub> )	Período (X <sub>3</sub> )	Perfil2 (X <sub>4</sub> )	Perfil3 (X <sub>5</sub> )	Z <sub>i</sub>	p <sub>i</sub>	LL <sub>i</sub>			
2	Gabriela	0	12,5	7	1	0	0	0	0,5	-0,69315			
3	Patrícia	0	13,3	10	1	0	0	0	0,5	-0,69315			
4	Gustavo	0	13,4	8	1	1	0	0	0,5	-0,69315			
5	Letícia	0	23,5	7	1	0	0	0	0,5	-0,69315			
6	Luiz Ovídio	0	9,5	8	1	0	0	0	0,5	-0,69315			
7	Leonor	0	13,5	10	1	0	0	0	0,5	-0,69315			
8	Dalila	0	13,5	10	1	0	0	0	0,5	-0,69315			
9	Antônio	0	15,4	10	1	0	0	0	0,5	-0,69315			
10	Júlia	0	14,7	10	1	0	0	0	0,5	-0,69315			
11	Mariana	0	14,7	10	1	0	0	0	0,5	-0,69315			
12	Roberto	0	13,7	10	1	0	0	0	0,5	-0,69315			
13	Renata	0	11	10	1	0	0	0	0,5	-0,69315			
14	Guilherme	0	18,4	10	1	0	0	0	0,5	-0,69315			
15	Rodrigo	0	11	11	1	1	0	0	0,5	-0,69315			
16	Giulia	0	11	10	1	0	0	0	0,5	-0,69315			
17	Felipe	0	12	7	1	1	0	0	0,5	-0,69315			
18	Karina	0	14	10	1	0	1	0	0,5	-0,69315			
19	Pietro	0	11,2	10	1	0	0	0	0,5	-0,69315			
20	Cecilia	0	13	10	1	0	0	0	0,5	-0,69315			
21	Gisele	0	12	6	1	0	0	0	0,5	-0,69315			
22	Elaine	0	17	10	1	0	1	0	0,5	-0,69315			
23	Kamal	0	12	9	1	0	0	0	0,5	-0,69315			
24	Rodolfo	0	12	10	1	1	0	0	0,5	-0,69315			
25	Pilar	0	13	5	0	0	0	0	0,5	-0,69315			
26	Vivian	0	11,7	10	0	0	0	0	0,5	-0,69315			
27	Danielle	0	17	10	0	0	0	0	0,5	-0,69315			
28	Juliana	0	14,4	10	0	1	0	0	0,5	-0,69315			
101	Estela	1	1	13	1	0	0	0	0,5	-0,69315			
102													
103													
											Somatória LL <sub>i</sub> -69,31472		

**Figura 13.2** Dados do arquivo **AtrasadoMáximaVerossimilhança.xls**.

Definir Objetivo:

Para: ☒ Máx. ☐ Mín. ☐ Valor de:

Alterando Células Variáveis:

Sujeito às Restrições:

☐ Tornar Variáveis Irestritas Não Negativas

Selecionar um Método de Solução:

Método de Solução

Selecione o mecanismo GRG Não Linear para Problemas do Solver suaves e não lineares. Selecione o mecanismo LP Simplex para Problemas do Solver lineares. Selecione o mecanismo Evolutionary para problemas do Solver não suaves.

Ajuda Resolver Fechar

Figura 13.3 Solver – Maximização da somatória do logaritmo da função de verossimilhança.

Como podemos verificar, quando  $\alpha = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ , o valor da somatória do logaritmo da função de verossimilhança é igual a  $-69,31472$ . Entretanto, deve haver uma combinação ótima de valores dos parâmetros, de modo que a função-objetivo apresentada na expressão (13.15) seja obedecida, ou seja, que o valor da somatória do logaritmo da função de verossimilhança seja o máximo possível.

Seguindo a lógica proposta por Belfiore e Fávero (2012), vamos então abrir a ferramenta **Solver** do Excel. A função-objetivo está na célula J103, que é a nossa célula de destino e que deverá ser maximizada. Além disso, os parâmetros  $\alpha$ ,  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$ ,  $\beta_4$  e  $\beta_5$ , cujos valores estão nas células M3, M5, M7, M9, M11 e M13, respectivamente, são as células variáveis. A janela do **Solver** ficará como mostra a Figura 13.3.

Ao clicarmos em **Resolver** e em **OK**, obteremos a solução ótima do problema de programação linear. A Tabela 13.5 mostra parte dos resultados obtidos.

Logo, o valor máximo possível da somatória do logaritmo da função de verossimilhança é  $LL_{\max} = -29,06568$ . A resolução deste problema gerou as seguintes estimativas dos parâmetros:

$$\begin{aligned}\alpha &= -30,202 \\ \beta_1 &= 0,220 \\ \beta_2 &= 2,767\end{aligned}$$





E, portanto, a expressão da probabilidade estimada de que um estudante  $i$  chegue atrasado pode ser escrita da seguinte forma:

$$P_i = \frac{1}{1 + e^{-(30,202 + 0,220 \cdot dist_i + 2,767 \cdot sem_i - 3,653 \cdot per_i + 1,346 \cdot perfil2_i + 2,914 \cdot perfil3_i)}}$$

Desta maneira, cabe agora a proposição de algumas interessantes perguntas:

**Qual é a probabilidade média estimada de se chegar atrasado à escola ao se deslocar 17 quilômetros e passar por 10 semáforos, tendo feito o trajeto de manhã e sendo considerado agressivo ao volante?**

**Em média, em quanto se altera a chance de se chegar atrasado à escola ao se adotar um percurso 1 quilômetro mais longo, mantidas as demais condições constantes?**

**Um aluno considerado agressivo apresenta, em média, uma chance maior de chegar atrasado do que outro considerado calmo? Se sim, em quanto é incrementada esta chance, mantidas as demais condições constantes?**

Antes de respondermos a estas importantes questões, precisamos verificar se todos os parâmetros estimados são estatisticamente significantes a um determinado nível de confiança. Se não for este o caso, precisaremos re-estimar o modelo final, a fim de que o mesmo apresente apenas parâmetros estatisticamente significantes para, a partir de então, ser possível a elaboração de inferências e previsões.

Portanto, tendo sido elaborada a estimação por máxima verossimilhança dos parâmetros da equação de probabilidade de ocorrência do evento, partiremos para o estudo da significância estatística geral do modelo obtido, bem como das significâncias estatísticas dos próprios parâmetros, de forma análoga ao realizado quando do estudo dos modelos tradicionais de regressão no capítulo anterior. É importante mencionar que no apêndice deste capítulo faremos uma breve apresentação dos modelos de regressão probit que podem ser utilizados alternativamente aos modelos de regressão logística binária para os casos em que a curva de probabilidades de ocorrência de determinado evento ajustar-se mais adequadamente à função densidade de probabilidade acumulada da distribuição normal padrão.

### 13.2.2. Significância estatística geral do modelo e dos parâmetros da regressão logística binária

Se, por exemplo, elaborarmos um gráfico linear da nossa variável dependente (*atrasado*) em função da variável referente ao número de semáforos (*sem*), perceberemos que as estimativas do modelo não são capazes de se ajustar de maneira satisfatória ao comportamento da variável dependente, dado que esta é uma *dummy*. O gráfico da Figura 13.5a apresenta este comportamento. Por outro lado, se o modelo de regressão logística binária for elaborado e forem plotadas as estimativas das probabilidades de se chegar atrasado para cada observação da nossa amostra, em função especificamente do número de semáforos pelos quais cada estudante passa, perceberemos que o ajuste é bem mais adequado ao comportamento da variável dependente (curva S, ou sigmoide), com valores estimados limitados entre 0 e 1 (Figura 13.5b).

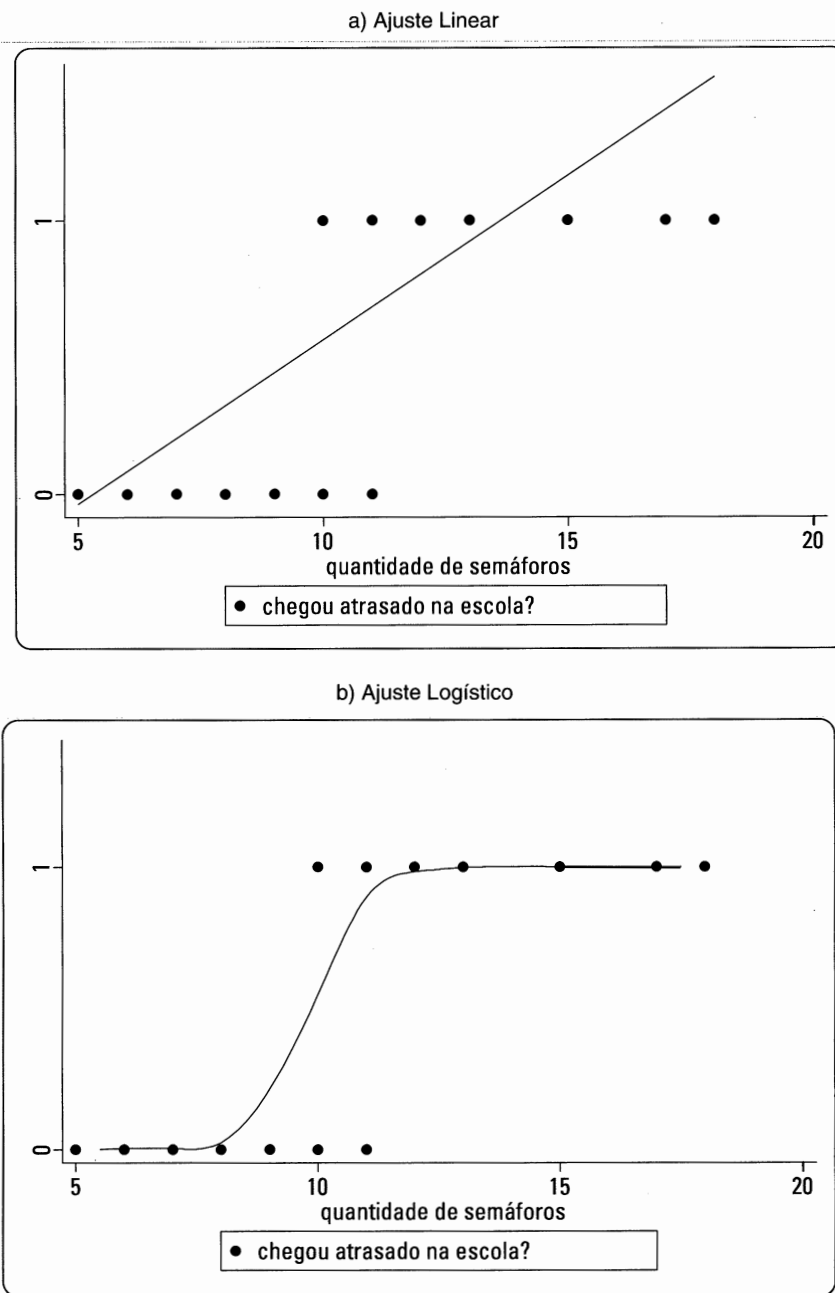
Portanto, como a variável dependente é qualitativa, não faz sentido discutirmos o percentual de sua variância que é explicado pelas variáveis preditoras, ou seja, em modelos de regressão logística não há um coeficiente de ajuste  $R^2$  como nos modelos tradicionais de regressão estimados pelo método de mínimos quadrados ordinários. Entretanto, muitos pesquisadores apresentam, em seus trabalhos, um coeficiente conhecido por **pseudo  $R^2$  de McFadden**, cuja expressão é dada por:

$$pseudo R^2 = \frac{-2 \cdot LL_0 - (-2 \cdot LL_{máx})}{-2 \cdot LL_0} \quad (13.16)$$

e cuja utilidade é bastante limitada e restringe-se a casos em que o pesquisador tiver interesse em comparar dois ou mais modelos distintos, dado que um dos diversos critérios existentes para a escolha do modelo é o critério de maior pseudo  $R^2$  de McFadden.

No nosso exemplo, conforme já discutimos na seção anterior e já calculamos por meio do **Solver** do Excel,  $LL_{máx}$ , que é o valor máximo possível da somatória do logaritmo da função de verossimilhança, é igual a -29,06568.

Já  $LL_0$  representa o valor máximo possível da somatória do logaritmo da função de verossimilhança para um modelo conhecido por **modelo nulo**, ou seja, para um modelo que só apresenta a constante  $\alpha$  e nenhuma variável explicativa. Por meio do mesmo procedimento elaborado na seção anterior, porém agora utilizando o arquivo **AtrasadoMáximaVerossimilhançaModeloNulo.xls**, obteremos  $LL_0 = -67,68585$ . As Figuras 13.6 e 13.7 mostram, respectivamente, a janela do **Solver** e parte dos resultados obtidos pela modelagem neste arquivo.



**Figura 13.5** Ajustes linear e logístico da variável dependente em função da variável *sem*.

Logo, com base na expressão (13.16), obteremos:


$$\text{pseudo } R^2 = \frac{-2.(-67,68585) - [(-2.(-29,06568))]}{-2.(-67,68585)} = 0,5706$$

Conforme discutimos, um maior pseudo  $R^2$  de McFadden pode ser utilizado como critério para escolha de um modelo em detrimento de outro. Entretanto, conforme iremos estudar na seção 13.2.4, há outro critério mais adequado à escolha do melhor modelo, o qual se refere à maior área abaixo da curva ROC.

Muitos pesquisadores também utilizam o pseudo  $R^2$  de McFadden como um indicador de desempenho do modelo escolhido, independentemente da comparação com outros modelos, porém a sua interpretação exige muitos cuidados e, por vezes, há a inevitável tentação em associá-lo, erroneamente, com percentuais de variância da variável dependente. Como iremos estudar na seção 13.2.4, o melhor indicador de desempenho de um


✕

## Parâmetros do Solver

**Definir Objetivo:**  

**Para:** ☒ **Máx.** ☐ **Mín.** ☐ **Valor de:**

**Alterando Células Variáveis:**



**Sujeito às Restrições:**

Adicionar


Alterar

Excluir

Redefinir Tudo

Carregar/Salvar

☐ **Tornar Variáveis Irrestritas Não Negativas**

**Selecionar um Método de Solução:**  

Opções

**Método de Solução**

Selecione o mecanismo GRG Não Linear para Problemas do Solver suaves e não lineares. Selecione o mecanismo LP Simplex para Problemas do Solver lineares. Selecione o mecanismo Evolutionary para problemas do Solver não suaves.

Ajuda

Resolver

Fechar

**Figura 13.6** Solver – Maximização da somatória do logaritmo da função de verossimilhança para o modelo nulo.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Estudante	Atrasado (Y)	Distância (X <sub>1</sub> )	Semáforos (X <sub>2</sub> )	Período (X <sub>3</sub> )	Perfil2 (X <sub>4</sub> )	Perfil3 (X <sub>5</sub> )	Z <sub>1</sub>	P <sub>1</sub>	LL <sub>1</sub>			
2	Gabriela	0	12,5	7	1	0	0	0,36397	0,59000	-0,89160			
3	Patrícia	0	13,3	10	1	0	0	0,36397	0,59000	-0,89160			
4	Gustavo	0	13,4	8	1	1	0	0,36397	0,59000	-0,89160			
5	Letícia	0	23,5	7	1	0	0	0,36397	0,59000	-0,89160			
6	Luiz Ovídio	0	9,5	8	1	0	0	0,36397	0,59000	-0,89160			
7	Leonor	0	13,5	10	1	0	0	0,36397	0,59000	-0,89160			
8	Dallia	0	13,5	10	1	0	0	0,36397	0,59000	-0,89160			
9	Antônio	0	13,4	10	1	0	0	0,36397	0,59000	-0,89160			
10	Júlia	0	14,7	10	1	0	0	0,36397	0,59000	-0,89160			
11	Mariana	0	14,7	10	1	0	0	0,36397	0,59000	-0,89160			
12	Roberto	0	13,7	10	1	0	0	0,36397	0,59000	-0,89160			
13	Renato	0	11	10	1	0	0	0,36397	0,59000	-0,89160			
14	Guilherme	0	18,4	10	1	0	0	0,36397	0,59000	-0,89160			
15	Rodrigue	0	11	11	1	1	0	0,36397	0,59000	-0,89160			
16	Giulia	0	11	10	1	0	0	0,36397	0,59000	-0,89160			
17	Felipe	0	12	7	1	1	0	0,36397	0,59000	-0,89160			
18	Karina	0	14	10	1	0	1	0,36397	0,59000	-0,89160			
19	Pietro	0	11,2	10	1	0	0	0,36397	0,59000	-0,89160			
20	Cecilia	0	13	10	1	0	0	0,36397	0,59000	-0,89160			
21	Grisele	0	12	6	1	0	0	0,36397	0,59000	-0,89160			
22	Elaine	0	17	10	1	0	1	0,36397	0,59000	-0,89160			
23	Kamell	0	12	9	1	0	0	0,36397	0,59000	-0,89160			
24	Rodolfo	0	12	10	1	1	0	0,36397	0,59000	-0,89160			
25	Pilar	0	13	5	0	0	0	0,36397	0,59000	-0,89160			
26	Vivian	0	11,7	10	0	0	0	0,36397	0,59000	-0,89160			
27	Daniella	0	17	10	0	0	0	0,36397	0,59000	-0,89160			
28	Juliana	0	14,4	10	0	1	0	0,36397	0,59000	-0,89160			
101	Estete	1	1	13	1	0	0	0,36397	0,59000	-0,89160			
102													
103													

α      0,354

Soma total LL<sub>1</sub> -67,68585

**Figura 13.7** Obtenção dos parâmetros quando da maximização de  $LL$  pelo Solver – modelo nulo.

modelo de regressão logística binária refere-se à eficiência global do modelo, que é definida com base na determinação de um *cutoff*, cujos conceitos também serão estudados na mesma seção.

Embora a utilidade do pseudo  $R^2$  de McFadden seja limitada, softwares como o Stata e o SPSS fazem seu cálculo e o apresentam em seus respectivos *outputs*, conforme veremos nas seções 13.4 e 15.5, respectivamente.

Analogamente ao procedimento apresentado no capítulo anterior, inicialmente iremos estudar a significância estatística geral do modelo que está sendo proposto. O teste  $\chi^2$  propicia condições à verificação da significância do modelo, uma vez que suas hipóteses nula e alternativa, para um modelo geral de regressão logística, são, respectivamente:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1: \text{existe pelo menos um } \beta_j \neq 0$$

Enquanto o teste  $F$  é utilizado para modelos de regressão em que a variável dependente apresenta-se na forma quantitativa, o que gera a decomposição de variância (tabela ANOVA) estudada no capítulo anterior, o teste  $\chi^2$  é mais adequado para modelos estimados pelo método de máxima verossimilhança, como os modelos de regressão logística.

O teste  $\chi^2$  propicia ao pesquisador uma verificação inicial sobre a existência do modelo que está sendo proposto, uma vez que, se todos os parâmetros estimados  $\beta_j$  ( $j = 1, 2, \dots, k$ ) forem estatisticamente iguais a 0, o comportamento de alteração de cada uma das variáveis  $X$  não influenciará em absolutamente nada a probabilidade de ocorrência do evento em estudo. A estatística  $\chi^2$  tem a seguinte expressão:

$$\chi^2 = -2.(LL_0 - LL_{\max}) \quad (13.17)$$

Voltando ao nosso exemplo, temos que:

$$\chi^2_{5g.l.} = -2.[-67,68585 - (-29,06568)] = 77,2403$$

Para 5 graus de liberdade (número de variáveis explicativas consideradas na modelagem, ou seja, número de parâmetros  $\beta$ ), temos, por meio da Tabela D do apêndice do livro, que o  $\chi^2_c = 11,070$  ( $\chi^2$  crítico para 5 graus de liberdade e para o nível de significância de 5%). Desta forma, como o  $\chi^2$  calculado  $\chi^2_{\text{cal}} = 77,2403 > \chi^2_c = 11,070$ , podemos rejeitar a hipótese nula de que todos os parâmetros  $\beta_j$  ( $j = 1, 2, \dots, 5$ ) sejam estatisticamente iguais a zero. Logo, pelo menos uma variável  $X$  é estatisticamente significativa para explicar a probabilidade de ocorrência do evento em estudo e teremos um modelo de regressão logística binária estatisticamente significativa para fins de previsão.

Softwares como o Stata e o SPSS não oferecem o  $\chi^2_c$  para os graus de liberdade definidos e um determinado nível de significância. Todavia, oferecem o nível de significância do  $\chi^2_{\text{cal}}$  para estes graus de liberdade. Desta forma, em vez de analisarmos se  $\chi^2_{\text{cal}} > \chi^2_c$ , devemos verificar se o nível de significância do  $\chi^2_{\text{cal}}$  é menor do que 0,05 (5%) a fim de darmos continuidade à análise de regressão. Assim:

Se *valor-P* (ou *P-value* ou *Sig.  $\chi^2_{\text{cal}}$*  ou *Prob.  $\chi^2_{\text{cal}}$* )  $< 0,05$ , existe pelo menos um  $\beta_j \neq 0$ .

O nível de significância do  $\chi^2_{\text{cal}}$  pode ser obtido no Excel por meio do comando **Fórmulas**  $\rightarrow$  **Inserir Função**  $\rightarrow$  **DIST.QUI**, que abrirá uma caixa de diálogo conforme mostra a Figura 13.8.

Análogo ao teste  $F$ , o teste  $\chi^2$  avalia a significância conjunta das variáveis explicativas, não definindo qual ou quais destas variáveis consideradas no modelo são estatisticamente significantes para influenciar a probabilidade de ocorrência do evento.

Desta forma, é preciso que o pesquisador avalie se cada um dos parâmetros do modelo de regressão logística binária é estatisticamente significativo e, neste sentido, a **estatística  $z$  de Wald** será importante para fornecer a significância estatística de cada parâmetro a ser considerado no modelo. A nomenclatura  $z$  refere-se ao fato de que a distribuição desta estatística é a distribuição normal padrão. As hipóteses do teste  $z$  de Wald para o  $\alpha$  e para cada  $\beta_j$  ( $j = 1, 2, \dots, k$ ) são, respectivamente:

$$H_0: \alpha = 0$$

$$H_1: \alpha \neq 0$$

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

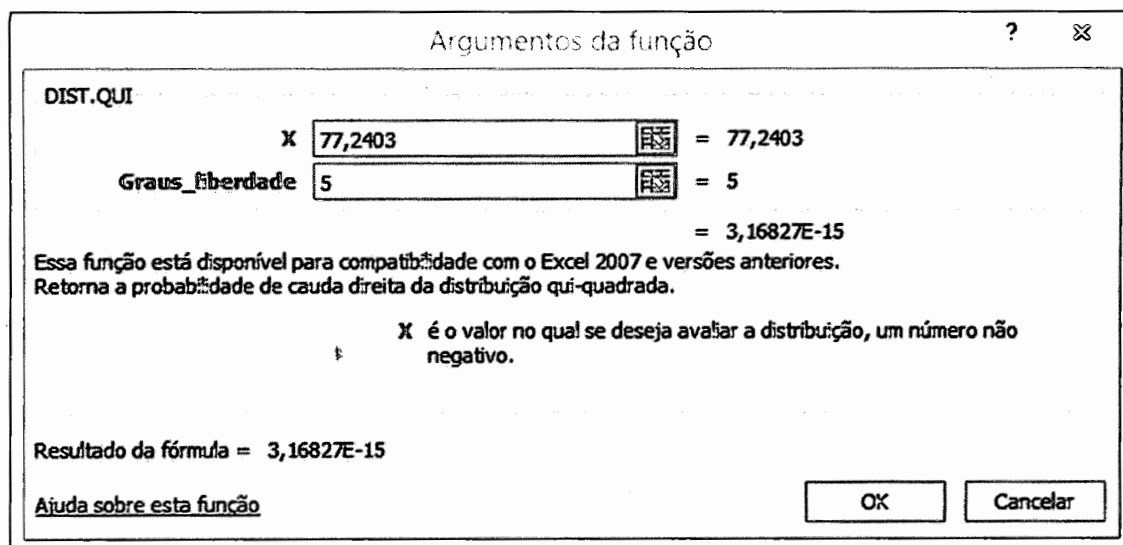


Figura 13.8 Obtenção do nível de significância de  $\chi^2$  (comando **Inserir Função**).

As expressões para o cálculo das estatísticas  $z$  de Wald de cada parâmetro  $\alpha$  e  $\beta_j$  são dadas, respectivamente, por:

$$z_{\alpha} = \frac{\alpha}{s.e.(\alpha)} \quad (13.18)$$

$$z_{\beta_j} = \frac{\beta_j}{s.e.(\beta_j)}$$

em que *s.e.* significa o erro-padrão (*standard error*) de cada parâmetro em análise. Dada a complexidade do cálculo dos erros-padrão de cada parâmetro, não o faremos neste momento, porém recomendamos a leitura de Engle (1984). Os valores de *s.e.* de cada parâmetro, para o nosso exemplo, são:

$$\begin{aligned} s.e.(\alpha) &= 9,981 \\ s.e.(\beta_1) &= 0,110 \\ s.e.(\beta_2) &= 0,922 \\ s.e.(\beta_3) &= 0,878 \\ s.e.(\beta_4) &= 0,748 \\ s.e.(\beta_5) &= 1,179 \end{aligned}$$

Logo, como já calculamos as estimativas dos parâmetros, temos que:

$$z_{\alpha} = \frac{\alpha}{s.e.(\alpha)} = \frac{-30,202}{9,981} = -3,026$$

$$z_{\beta_1} = \frac{\beta_1}{s.e.(\beta_1)} = \frac{0,220}{0,110} = 2,000$$

$$z_{\beta_2} = \frac{\beta_2}{s.e.(\beta_2)} = \frac{2,767}{0,922} = 3,001$$

$$z_{\beta_3} = \frac{\beta_3}{s.e.(\beta_3)} = \frac{-3,653}{0,878} = -4,161$$

$$z_{\beta_4} = \frac{\beta_4}{s.e.(\beta_4)} = \frac{1,346}{0,748} = 1,799$$

$$z_{\beta_5} = \frac{\beta_5}{s.e.(\beta_5)} = \frac{2,914}{1,179} = 2,472$$

Após a obtenção das estatísticas  $z$  de Wald, o pesquisador pode utilizar a tabela de distribuição da curva normal padrão para obtenção dos valores críticos a um dado nível de significância e verificar se tais testes rejeitam ou não a hipótese nula.

Para o nível de significância de 5%, temos, por meio da Tabela E do apêndice do livro, que o  $z_c = -1,96$  para a cauda inferior (probabilidade na cauda inferior de 0,025 para a distribuição bicaudal) e  $z_c = 1,96$  para a cauda superior (probabilidade na cauda superior também de 0,025 para a distribuição bicaudal).

Os valores de  $z_c$  para o nível de significância de 5% podem ser obtidos no Excel por meio do comando **Fórmulas → Inserir Função → INV.NORMP**, sendo que o pesquisador deverá digitar uma probabilidade de 2,5% para a obtenção de  $z_c$  para a cauda inferior e 97,5% para a obtenção de  $z_c$  para a cauda superior, conforme mostram, respectivamente, as Figuras 13.9 e 13.10.

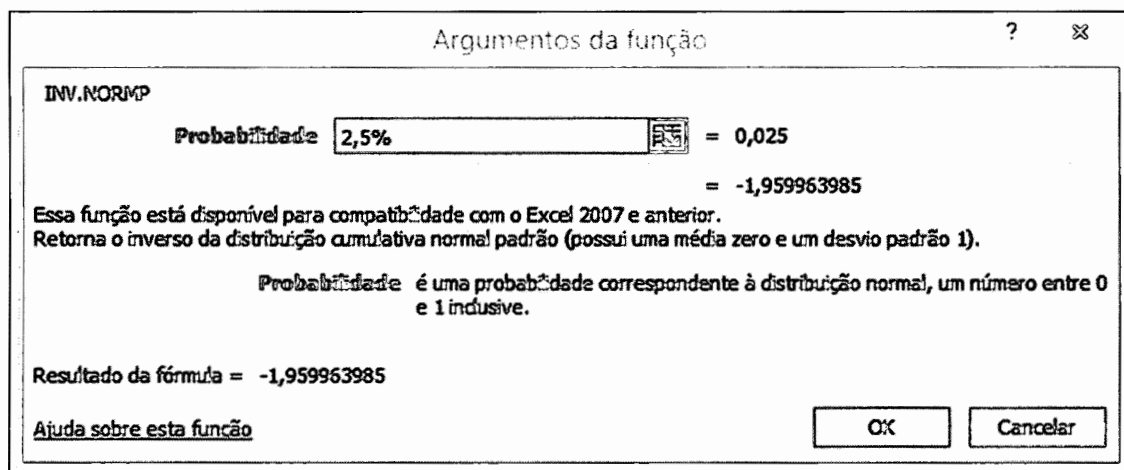


Figura 13.9 Obtenção de  $z_c$  para a cauda inferior (comando **Inserir Função**).

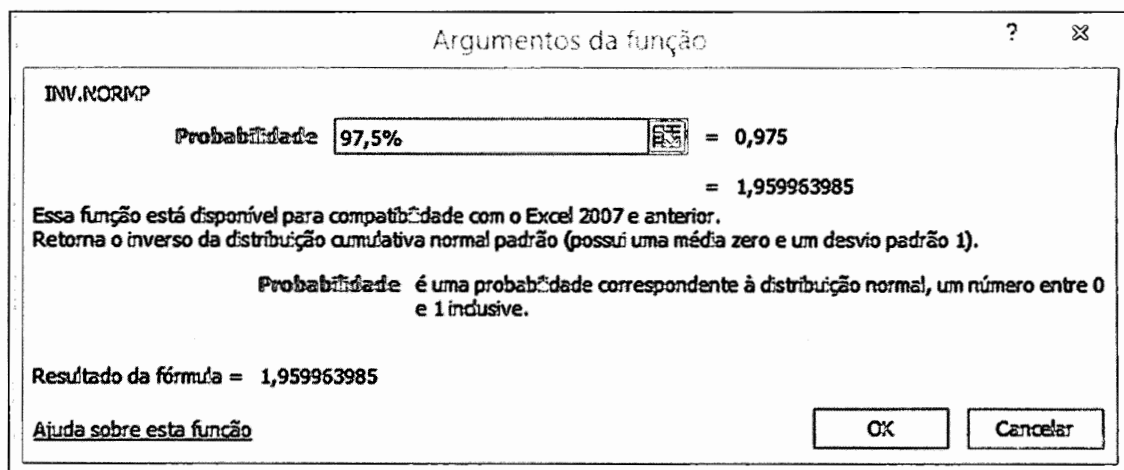


Figura 13.10 Obtenção de  $z_c$  para a cauda superior (comando **Inserir Função**).

Apenas a estatística  $z$  de Wald do parâmetro  $\beta_4$  apresentou valor entre  $-1,96$  e  $1,96$ , o que indica, ao nível de significância de 5%, que, para este caso, não houve rejeição da hipótese nula, ou seja, este parâmetro não pode ser considerado estatisticamente diferente de zero.

Como no caso do teste  $\chi^2$ , os pacotes estatísticos também oferecem os valores dos níveis de significância dos testes  $z$  de Wald, o que facilita a decisão, já que, com 95% de nível de confiança (5% de nível de significância), teremos:

Se *valor-P* (ou *P-value* ou *Sig.  $z_{cal}$*  ou *Prob.  $z_{cal}$* )  $< 0,05$  para  $\alpha$ ,  $\alpha \neq 0$

e

Se *valor-P* (ou *P-value* ou *Sig.  $z_{cal}$*  ou *Prob.  $z_{cal}$* )  $< 0,05$  para determinada variável explicativa  $X$ ,  $\beta \neq 0$ .

Desta forma, como  $-1,96 < z_{\beta_4} = 1,799 < 1,96$ , veremos que o *valor-P* da estatística  $z$  de Wald da variável *perfil2* será maior do que 0,05.

A não rejeição da hipótese nula para o parâmetro  $\beta_4$ , ao nível de significância de 5%, indica que a correspondente variável *perfil2* não é estatisticamente significativa para aumentar ou diminuir a probabilidade de se chegar atrasado à escola na presença das demais variáveis explicativas e, portanto, poderá ser excluída do modelo final.

Neste momento, iremos fazer a exclusão manual desta variável, a fim de obtermos o modelo final. Entretanto, é importante ressaltar que a exclusão manual de uma variável pode fazer com que outra inicialmente significativa passe a apresentar um parâmetro não significativo, e este problema tende a piorar tanto quanto maior for o número de variáveis explicativas no banco de dados. O contrário também pode ocorrer, ou seja, não se recomenda que haja a exclusão manual simultânea de duas ou mais variáveis cujos parâmetros, num primeiro momento, não se mostrarem estatisticamente diferentes de zero, uma vez que um determinado parâmetro  $\beta$  pode tornar-se estatisticamente diferente de zero, mesmo inicialmente não sendo, ao se eliminar da análise outra variável cujo parâmetro  $\beta$  também não se mostrava estatisticamente diferente de zero. Felizmente estes fenômenos não ocorrem neste exemplo e, assim, optamos por excluir manualmente a variável *perfil2*. Isto será comprovado quando estimarmos este modelo de regressão logística binária por meio do procedimento *Stepwise* nos softwares Stata (seção 13.4) e SPSS (seção 13.5).

Assim, vamos abrir o arquivo **AtrasadoMáximaVerossimilhançaModeloFinal.xls**. Note que agora o cálculo do logito ( $Z$ ) não leva mais em consideração o parâmetro da variável *perfil2*, excluída da modelagem. As Figuras 13.11 e 13.12 mostram, respectivamente, a janela do **Solver** e parte dos resultados obtidos pela modelagem por meio deste último arquivo.

Logo, para o modelo final, temos que  $LL_{máx} = -30,80079$ . Antes de partirmos para a definição da expressão final da probabilidade de ocorrência do evento em estudo, precisamos definir se o novo modelo estimado (modelo final) apresenta perda na qualidade do ajuste em relação ao modelo completo estimado com todas as variáveis explicativas. Para tanto, o **teste de razão de verossimilhança (likelihood-ratio test)**, que verifica a adequação do ajuste do modelo completo em comparação com o ajuste do modelo final, pode ser utilizado, apresentando a seguinte expressão:

$$\chi^2_{1g.l.} = -2 \cdot (LL_{\text{modelo final}} - LL_{\text{modelo completo}}) \quad (13.19)$$

Para os dados do nosso exemplo, temos que:

$$\chi^2_{1g.l.} = -2 \cdot [-30,80079 - (-29,06568)] = 3,4702$$

Logo, para 1 grau de liberdade, temos, por meio da Tabela D do apêndice do livro, que o  $\chi^2_c = 3,841$  ( $\chi^2$  crítico para 1 grau de liberdade e para o nível de significância de 5%). Desta forma, como o  $\chi^2$  calculado  $\chi^2_{cal} = 3,4702 < \chi^2_c = 3,841$ , não rejeitamos a hipótese nula do teste de razão de verossimilhança, ou seja, a estimação do modelo final com a exclusão da variável *perfil2* não alterou a qualidade do ajuste, ao nível de significância de 5%, o que faz com que este modelo seja preferível em relação ao modelo completo estimado com todas as variáveis explicativas.

Nas seções 13.4 e 13.5 apresentaremos, por meio dos softwares Stata e SPSS, respectivamente, outro teste muito usual para verificação da qualidade de ajuste do modelo final, conhecido por **teste de Hosmer-Lemeshow**. Segundo Ayçaguer e Utra (2004), ao se dividir a base de dados em 10 grupos pelos decis das probabilidades estimadas pelo modelo final para cada observação, este teste avalia, por meio da elaboração de um teste  $\chi^2$ , se existem



**Parâmetros do Solver**

Definir Objetivo:

Para: ☒ Máx. ☐ Mín. ☐ Valor de:

Alterando Células Variáveis:

Sujeito às Restrições:

☐ Tornar Variáveis Irrestritas Não Negativas

Selecionar um Método de Solução:

**Método de Solução**  
 Selecione o mecanismo GRG Não Linear para Problemas do Solver suaves e não lineares. Selecione o mecanismo LP Simplex para Problemas do Solver lineares. Selecione o mecanismo Evolutionary para problemas do Solver não suaves.

**Figura 13.11** Solver – Maximização da somatória do logaritmo da função de verossimilhança para o modelo final.

diferenças significativas entre as frequências observadas e esperadas do número de observações em cada um dos 10 grupos e, caso tais diferenças não sejam estatisticamente significativas, a um determinado nível de significância, o modelo estimado não apresentará problemas em relação à qualidade do ajuste proposto.

Sendo assim, retornaremos à análise dos resultados da estimação do modelo final, e a resolução deste novo problema gerou as seguintes estimativas finais dos parâmetros:

$$\alpha = -30,935$$

$$\beta_1 = 0,204$$

$$\beta_2 = 2,920$$

$$\beta_3 = -3,776$$

$$\beta_5 = 2,459$$

com os respectivos erros-padrão:

$$s.e. (\alpha) = 10,636$$

$$s.e. (\beta_1) = 0,101$$

$$s.e. (\beta_2) = 1,011$$

$$s.e. (\beta_3) = 0,847$$

$$s.e. (\beta_5) = 1,139$$

	A	B	C	D	E	F	G	H	I	J	K	L
1	Estudante	Atrasado (Y)	Distância (X <sub>1</sub> )	Semáforos (X <sub>2</sub> )	Período (X <sub>3</sub> )	Perfil3 (X <sub>4</sub> )	Z <sub>i</sub>	p <sub>i</sub>	LL <sub>i</sub>			
2	Gabriela	0	12,5	7	1	0	-11,717409	0,00001	-0,00001			
3	Patrícia	0	13,3	10	1	0	-2,79341709	0,05768	-0,05941			
4	Gustavo	0	13,4	8	1	0	-8,61344032	0,00018	-0,00018			
5	Letícia	0	23,5	7	1	0	-9,47159036	0,00008	-0,00008			
6	Luiz Ovídio	0	9,5	8	1	0	-9,40968511	0,00008	-0,00008			
7	Leonor	0	13,5	10	1	0	-2,75258402	0,05994	-0,06181			
8	Dallia	0	13,5	10	1	0	-2,75258402	0,05994	-0,06181			
9	Antônio	0	15,4	10	1	0	-2,36466989	0,08591	-0,08982			
10	Júlia	0	14,7	10	1	0	-2,50758562	0,07533	-0,07832			
11	Mariana	0	14,7	10	1	0	-2,50758562	0,07533	-0,07832			
12	Roberto	0	13,7	10	1	0	-2,71175096	0,06228	-0,06431			
13	Renata	0	11	10	1	0	-3,26299735	0,03686	-0,03756			
14	Guilherme	0	18,4	10	1	0	-1,7521739	0,14777	-0,15990			
15	Rodrigo	0	11	11	1	0	-0,34277747	0,41513	-0,53637			
16	Giulia	0	11	10	1	0	-3,26299735	0,03686	-0,03756			
17	Felipe	0	12	7	1	0	-11,8194917	0,00001	-0,00001			
18	Karina	0	14	10	1	1	-0,19140737	0,45229	-0,60202			
19	Pietro	0	11,2	10	1	0	-3,22216428	0,03834	-0,03909			
20	Cecília	0	13	10	1	0	-2,85466669	0,05444	-0,05598			
21	Glisele	0	12	6	1	0	-14,7397115	0,00000	0,00000			
22	Elaine	0	17	10	1	1	0,42108863	0,60374	-0,92569			
23	Kamal	0	12	9	1	0	-5,9790519	0,00252	-0,00253			
24	Rodolfo	0	12	10	1	0	-3,05883202	0,04484	-0,04587			
25	Pilar	0	13	5	0	0	-13,6794265	0,00000	0,00000			
26	Vivian	0	11,7	10	0	0	0,656258	0,65842	-1,07417			
27	Danielle	0	17	10	0	0	1,73833425	0,85048	-1,90029			
28	Juliana	0	14,4	10	0	0	1,20750439	0,76986	-1,46905			
101	Estela	1	1	13	1	0	3,45600899	0,96941	-0,03107			
102												
103												
							Somatória LL <sub>i</sub>	-30,80079				

Figura 13.12 Obtenção dos parâmetros quando da maximização de LL pelo Solver - modelo final.

e as seguintes estatísticas  $z$  de Wald:

$$z_{\alpha} = \frac{\alpha}{s.e.(\alpha)} = \frac{-30,935}{10,636} = -2,909$$

$$z_{\beta_1} = \frac{\beta_1}{s.e.(\beta_1)} = \frac{0,204}{0,101} = 2,020$$

$$z_{\beta_2} = \frac{\beta_2}{s.e.(\beta_2)} = \frac{2,920}{1,011} = 2,888$$

$$z_{\beta_3} = \frac{\beta_3}{s.e.(\beta_3)} = \frac{-3,776}{0,847} = -4,458$$

$$z_{\beta_5} = \frac{\beta_5}{s.e.(\beta_5)} = \frac{2,459}{1,139} = 2,159$$

com todos os valores de  $z_{cal} < -1,96$  ou  $> 1,96$  e, portanto, com valores- $P$  das estatísticas  $z$  de Wald  $< 0,05$ .

O modelo final ainda apresenta as seguintes estatísticas:

$$pseudo R^2 = \frac{-2.(-67,68585) - [(-2.(-30,80079))]}{-2.(-67,68585)} = 0,5449$$

$$\chi^2_{4g.l.} = -2.[-67,68585 - (-30,80079)] = 73,77012 > \chi^2_{4g.l.} = 9,48773$$

Desta forma, podemos escrever o logito  $Z_i$  como segue:

$$Z_i = -30,935 + 0,204.dist_i + 2,920.sem_i - 3,776.per_i + 2,459.perfil3_i$$

com a seguinte expressão final de probabilidade estimada de que um estudante  $i$  chegue atrasado à escola:

$$p_i = \frac{1}{1 + e^{-(30,935 + 0,204 \cdot dist_i + 2,920 \cdot sem_i - 3,776 \cdot per_i + 2,459 \cdot perfil3_i)}}$$

Estes parâmetros e respectivas estatísticas também serão obtidos por meio do procedimento *Stepwise* quando da estimação do modelo de regressão logística binária no Stata e no SPSS.

Com base na estimação da função probabilística, um curioso pesquisador pode, por exemplo, desejar elaborar um gráfico das probabilidades estimadas de cada aluno chegar atrasado à escola (coluna H do arquivo do modelo final no Excel) em função do número de semáforos pelos quais cada um passa no percurso (coluna D no Excel). A Figura 13.13 apresenta este gráfico e, ao contrário do gráfico da Figura 13.5b, que oferece um ajuste logístico determinístico (apenas valores iguais a 0 ou 1 para a variável dependente), este novo gráfico apresenta um ajuste logístico probabilístico.

Com base na Figura 13.13, que também apresenta a curva logística ajustada à nuvem de pontos que representam as probabilidades estimadas para cada observação, podemos verificar que, enquanto a probabilidade de se chegar atrasado à escola é muito baixa quando se passa por até 8 semáforos ao longo do trajeto, esta probabilidade passa ser bastante elevada quando se é obrigado a passar por 11 ou mais semáforos no percurso.

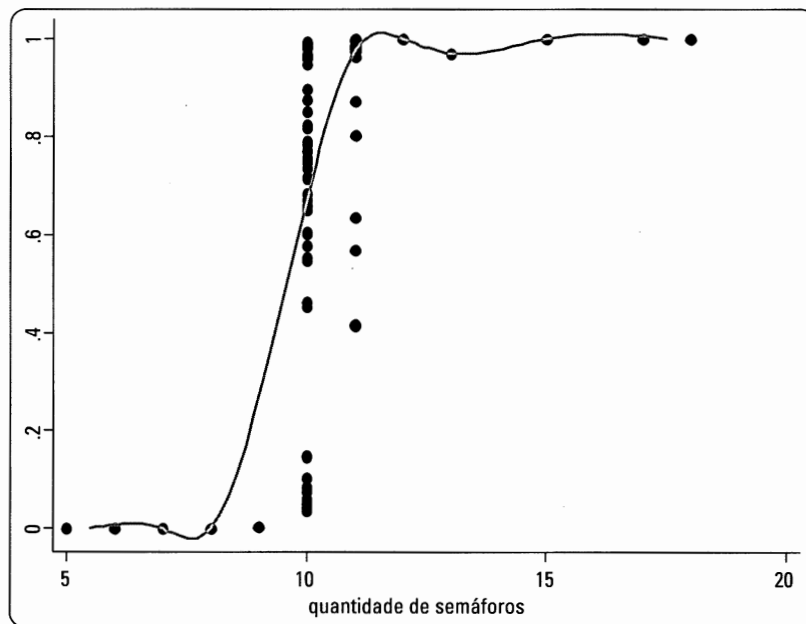


Figura 13.13 Ajuste logístico probabilístico em função da variável *sem*.

Aprofundando a análise da função probabilística, podemos retornar às nossas três importantes perguntas, respondendo uma de cada vez:

**Qual é a probabilidade média estimada de se chegar atrasado à escola ao se deslocar 17 quilômetros e passar por 10 semáforos, tendo feito o trajeto de manhã e sendo considerado agressivo ao volante?**

Fazendo uso da última expressão de probabilidade e substituindo os valores fornecidos nesta equação, teremos:

$$p = \frac{1}{1 + e^{[-30,935 + 0,204 \cdot (17) + 2,920 \cdot (10) - 3,776 \cdot (1) + 2,459 \cdot (1)]}} = 0,603$$

Logo, a probabilidade média estimada de se chegar atrasado à escola é, nas condições informadas, igual a 60,3%.

**Em média, em quanto se altera a chance de se chegar atrasado à escola ao se adotar um percurso 1 quilômetro mais longo, mantidas as demais condições constantes?**

Para respondermos a esta questão, devemos recorrer à expressão (13.3), que poderá ser escrita da seguinte forma:

$$chance_{Y=1} = e^{Z_i} \quad (13.20)$$

de modo que, mantidas as demais condições constantes, a chance de se chegar atrasado à escola ao se adotar um trajeto 1 quilômetro mais longo é:

$$chance_{Y=1} = e^{0,204} = 1,226$$

Logo, a chance é multiplicada por um fator de 1,226, ou seja, mantidas as demais condições constantes, a chance de se chegar atrasado à escola ao se adotar um trajeto 1 quilômetro mais longo é, em média, 22,6% maior.

**Um aluno considerado agressivo apresenta, em média, uma chance maior de chegar atrasado do que outro considerado calmo? Se sim, em quanto é incrementada esta chance, mantidas as demais condições constantes?**

Como  $\beta_5$  é positivo, podemos afirmar que a probabilidade de um aluno considerado agressivo chegar atrasado é maior do que um aluno considerado calmo, fato que também é comprovado quando se analisa a chance, dado que, se  $\beta_5 > 0$ , logo  $e^{\beta_5} > 1$ , ou seja, a chance será maior de chegar atrasado quando se é agressivo ao volante em relação a ser calmo. Isso comprova, mais uma vez, que a agressividade no volante não leva a nada!

Mantidas as demais condições constantes, a chance de chegar atrasado quando se é agressivo ao volante em relação a ser calmo é dada por:

$$chance_{Y=1} = e^{2,459} = 11,693$$

Logo, a chance é multiplicada por um fator de 11,693, ou seja, mantidas as demais condições constantes, a chance de se chegar atrasado à escola quando se é agressivo ao volante em relação a ser calmo é, em média, 1.069,3% maior.

Vale comentar que não há diferenças na probabilidade de se chegar atrasado à escola quando se é considerado moderado ou calmo, dado que o parâmetro  $\beta_4$  (referente à categoria *moderado*) apresentou-se estatisticamente igual a zero, ao nível de significância de 5%.

Conforme podemos perceber, estes cálculos utilizaram sempre as estimativas médias dos parâmetros. Partiremos agora para o estudo dos intervalos de confiança destes parâmetros.

### 13.2.3. Construção dos intervalos de confiança dos parâmetros do modelo de regressão logística binária

Os intervalos de confiança dos coeficientes da expressão (13.10), para os parâmetros  $\alpha$  e  $\beta_j$  ( $j = 1, 2, \dots, k$ ), ao nível de confiança de 95%, podem ser escritos, respectivamente, da seguinte forma:

$$\alpha \pm 1,96 \cdot [s.e.(\alpha)] \quad (13.21)$$

$$\beta_j \pm 1,96 \cdot [s.e.(\beta_j)]$$

em que, conforme vimos, 1,96 é o  $z_c$  para o nível de confiança de 95% (nível de significância de 5%).

Desta maneira, podemos elaborar a Tabela 13.6, que traz os coeficientes estimados dos parâmetros da expressão de probabilidade de ocorrência do evento de interesse do nosso exemplo, com os respectivos erros-padrão, as estatísticas  $z$  de Wald e os intervalos de confiança para o nível de significância de 5%.

Esta tabela é igual à que obteremos quando da elaboração da modelagem no Stata e no SPSS por meio do procedimento *Stepwise*. Como base nos intervalos de confiança dos parâmetros, podemos escrever as expressões dos limites inferior (mínimo) e superior (máximo) da probabilidade estimada de que um estudante  $i$  chegue atrasado à escola, com 95% de confiança. Assim, teremos:

$$p_{i_{\min}} = \frac{1}{1 + e^{-(51,782 + 0,006 \cdot dist_i + 0,938 \cdot sem_i - 5,436 \cdot per_i + 0,227 \cdot perfil3_i)}}$$

$$p_{i_{\max}} = \frac{1}{1 + e^{-(10,088 + 0,402 \cdot dist_i + 4,902 \cdot sem_i - 2,116 \cdot per_i + 4,691 \cdot perfil3_i)}}$$

Com base na expressão (13.20), o intervalo de confiança da chance de ocorrência do evento de interesse para cada parâmetro  $\beta_j$  ( $j = 1, 2, \dots, k$ ), ao nível de confiança de 95%, pode ser escrito da seguinte forma:

$$e^{\beta_j \pm 1,96[s.e.(\beta_j)]} \quad (13.22)$$

Note que não apresentamos a expressão do intervalo de confiança da chance para o parâmetro  $\alpha$ , uma vez que só faz sentido discutirmos a mudança na chance de ocorrência do evento em estudo quando é alterada em uma unidade, por exemplo, determinada variável explicativa do modelo, mantidas as demais condições constantes.

Para os dados do nosso exemplo e com base nos valores da Tabela 13.6, vamos, então, elaborar a Tabela 13.7, que apresenta os intervalos de confiança da chance (*odds*) de ocorrência do evento de interesse para cada parâmetro  $\beta_j$ .

**Tabela 13.6** Cálculo dos intervalos de confiança dos parâmetros.

Parâmetro	Coeficiente	Erro-Padrão (s.e.)	z	Intervalo de Confiança (95%)	
				$\alpha - 1,96[s.e.(\alpha)]$ $\beta_j - 1,96[s.e.(\beta_j)]$	$\alpha + 1,96[s.e.(\alpha)]$ $\beta_j + 1,96[s.e.(\beta_j)]$
$\alpha$ (constante)	-30,935	10,636	-2,909	-51,782	-10,088
$\beta_1$ (variável <i>dist</i> )	0,204	0,101	2,020	0,006	0,402
$\beta_2$ (variável <i>sem</i> )	2,920	1,011	2,888	0,938	4,902
$\beta_3$ (variável <i>per</i> )	-3,776	0,847	-4,458	-5,436	-2,116
$\beta_5$ (variável <i>perfil3</i> )	2,459	1,139	2,159	0,227	4,691

**Tabela 13.7** Cálculo dos intervalos de confiança da chance (*odds*) para cada parâmetro  $\beta_j$ .

Parâmetro	Chance ( <i>Odds</i> )	Intervalo de Confiança da Chance (95%)	
	$e^{\beta_j}$	$e^{\beta_j - 1,96[s.e.(\beta_j)]}$	$e^{\beta_j + 1,96[s.e.(\beta_j)]}$
$\beta_1$ (variável <i>dist</i> )	1,226	1,006	1,495
$\beta_2$ (variável <i>sem</i> )	18,541	2,555	134,458
$\beta_3$ (variável <i>per</i> )	0,023	0,004	0,120
$\beta_5$ (variável <i>perfil3</i> )	11,693	1,254	109,001

Estes valores também poderão ser obtidos por meio do Stata e do SPSS, conforme mostraremos, respectivamente, nas seções 13.4 e 13.5.

Conforme já discutido no capítulo anterior, se o intervalo de confiança de determinado parâmetro contiver o zero (ou da chance contiver o 1), o mesmo será considerado estatisticamente igual a zero para o nível de confiança com que o pesquisador estiver trabalhando. Se isso acontecer com o parâmetro  $\alpha$ , recomenda-se que nada seja alterado na modelagem, uma vez que tal fato é decorrente da utilização de amostras pequenas, e uma amostra maior poderia resolver este problema. Por outro lado, se o intervalo de confiança de um parâmetro  $\beta_j$  contiver o zero, este será excluído do modelo final quando da elaboração do procedimento *Stepwise*. Embora não tenha sido mostrado aqui, o intervalo de confiança do parâmetro estimado para a variável *perfil2* conteve o zero já que, como discutido, seu valor de  $z_{cal}$  situou-se entre -1,96 e 1,96 e, portanto, tal variável foi excluída do modelo final.

Conforme também já discutido, a rejeição da hipótese nula para determinado parâmetro  $\beta$ , a um especificado nível de significância, indica que a correspondente variável  $X$  é significativa para explicar a probabilidade de ocorrência do evento de interesse e, conseqüentemente, deve permanecer no modelo final. Podemos, portanto, concluir que a decisão pela exclusão de determinada variável  $X$  em um modelo de regressão logística pode ser realizada por meio da análise direta da estatística  $z$  de Wald de seu respectivo parâmetro  $\beta$  (se  $-z_c < z_{cal} < z_c \rightarrow \text{valor-}P > 0,05 \rightarrow$  não podemos rejeitar que o parâmetro seja estatisticamente igual a zero) ou por meio da análise do intervalo de confiança (se o mesmo contiver o zero). O Quadro 13.1 apresenta os critérios de inclusão ou exclusão de parâmetros  $\beta_j$  ( $j = 1, 2, \dots, k$ ) em modelos de regressão logística.

**QUADRO 13.1** Decisão de inclusão de parâmetros  $\beta_j$  em modelos de regressão logística.

Parâmetro	Estatística $z$ de Wald (para nível de significância $\alpha$ )	Teste $z$ (análise do <i>valor-P</i> para nível de significância $\alpha$ )	Análise pelo Intervalo de Confiança	Decisão
$\beta_j$	$-z_{\alpha/2} < z_{cal} < z_{\alpha/2}$	$\text{valor-P} > \text{nível de sig. } \alpha$	O intervalo de confiança contém o zero	Excluir o parâmetro do modelo
	$z_{cal} > z_{\alpha/2}$ ou $z_{cal} < -z_{\alpha/2}$	$\text{valor-P} < \text{nível de sig. } \alpha$	O intervalo de confiança não contém o zero	Manter o parâmetro no modelo

Obs.: O mais comum em ciências sociais aplicadas é a adoção do nível de significância  $\alpha = 5\%$ .

#### 13.2.4. Cutoff, análise de sensibilidade, eficiência global do modelo, sensibilidade e especificidade

Estimado o modelo de probabilidade de ocorrência do evento, vamos agora definir o conceito de *cutoff*, a partir do qual será possível classificar, no nosso exemplo, as observações com base nas probabilidades estimadas de cada uma delas. Voltemos à expressão de probabilidade estimada para o modelo final:

$$p_i = \frac{1}{1 + e^{-(30,935 + 0,204 \cdot \text{dist}_i + 2,920 \cdot \text{sem}_i - 3,776 \cdot \text{per}_i + 2,459 \cdot \text{perfil}_i)}}$$

Calculados os valores de  $p_i$ , por meio do arquivo **AtrasadoMáximaVerossimilhançaModeloFinal.xls**, vamos elaborar uma tabela com algumas das observações da nossa amostra. A Tabela 13.8 traz os valores de  $p_i$  para 10 observações escolhidas aleatoriamente, apenas para fins didáticos.

O *cutoff*, que nada mais é do que um ponto de corte que o pesquisador escolhe, é definido para que sejam classificadas as observações em função das suas probabilidades calculadas e, desta forma, é utilizado quando há o intuito de se elaborarem previsões de ocorrência do evento para observações não presentes na amostra com base nas probabilidades das observações presentes na amostra.

Assim, se determinada observação não presente na amostra apresentar uma probabilidade de incidir no evento maior do que o *cutoff* definido, espera-se que haja a incidência do evento e, portanto, será classificada como *evento*. Por outro lado, se a sua probabilidade for menor do que o *cutoff* definido, espera-se que haja a incidência do não evento e, portanto, será classificada como *não evento*.

De maneira geral, podemos estipular o seguinte critério:

Se  $p_i > \text{cutoff} \rightarrow$  a observação  $i$  deverá ser classificada como *evento*.

Se  $p_i < \text{cutoff} \rightarrow$  a observação  $i$  deverá ser classificada como *não evento*.

Como a expressão de probabilidade é estimada com base nas observações presentes na amostra, a classificação para outras observações não presentes inicialmente na amostra leva em consideração a consistência do comportamento dos estimadores e, portanto, para efeitos inferenciais, a amostra deve ser significativa e representativa do comportamento populacional, como em qualquer modelo de dependência confirmatório.<sup>1</sup>

<sup>1</sup>Vale a pena mencionar que, ao longo de todo este capítulo, estamos considerando que a relação entre a proporção de observações definidas como evento e a proporção de observações definidas como não evento na amostra em estudo seja idêntica à correspondente relação existente na população, já que, por vezes, não se conhece essa relação. Se, entretanto, ela for conhecida e significativamente diferente da considerada na amostra em análise, a probabilidade estimada de ocorrência do evento em estudo para determinada observação da amostra pode ser consideravelmente diferente da observada na população em geral.

Neste sentido, para que o modelo possa ser aplicado a uma população cuja proporção de observações definidas como evento é substancialmente diferente daquela utilizada em sua estimação, é necessário que seja aplicada uma correção no valor do intercepto estimado no modelo amostral. Conforme sugere Anderson (1982) e discutem Brito e Assaf Neto (2007), pode ser utilizada a seguinte expressão para que o intercepto seja corrigido:

$$\alpha_{\text{corrigido}} = \alpha_{\text{estimado}} + \ln \left( \frac{\Pi_1 \cdot n_0}{\Pi_0 \cdot n_1} \right)$$

em que  $\Pi_1$  e  $\Pi_0$  representam, respectivamente, a proporção de observações definidas como evento e a proporção de observações definidas como não evento na população em geral, e  $n_0$  e  $n_1$  representam, respectivamente, a quantidade de observações definidas como não evento e a quantidade de observações definidas como evento na amostra em estudo, sendo  $n_0 + n_1 = n$  (tamanho da amostra).

**Tabela 13.8** Valores de  $p_i$  para 10 observações.

Observação	$p_i$
Adelino	0,05444
Carolina	0,67206
Cristina	0,55159
Eduardo	0,81658
Cintia	0,64918
Raimundo	0,05340
Emerson	0,04484
Raquel	0,56702
Rita	0,85048
Leandro	0,46243

**Tabela 13.9** Real incidência do evento e classificação para 10 observações com  $cutoff = 0,5$ .

Observação	Evento	$p_i$	Classificação $Cutoff = 0,5$
Adelino	Não	0,05444	Não
Carolina	Não	0,67206	Sim
Cristina	Não	0,55159	Sim
Eduardo	Não	0,81658	Sim
Cintia	Não	0,64918	Sim
Raimundo	Não	0,05340	Não
Emerson	Não	0,04484	Não
Raquel	Não	0,56702	Sim
Rita	Sim	0,85048	Sim
Leandro	Sim	0,46243	Não

O *cutoff* serve para que o pesquisador avalie a real incidência do evento para cada observação e a compare com a expectativa de que cada observação incida, de fato, no evento. Com isto feito, será possível avaliar a taxa de acerto do modelo com base nas próprias observações presentes na amostra e, por inferência, assumir que tal taxa de acerto se mantenha quando houver o intuito de avaliar a incidência do evento para outras observações não presentes na amostra (previsão).

Com base nos dados das observações apresentadas na Tabela 13.8, e escolhendo-se, por exemplo, um *cutoff* de 0,5, podemos definir que:

Se  $p_i > 0,5 \rightarrow$  a observação  $i$  deverá ser classificada como *evento*.

Se  $p_i < 0,5 \rightarrow$  a observação  $i$  deverá ser classificada como *não evento*.

A Tabela 13.9 traz, para cada uma das 10 observações escolhidas ao acaso, a real incidência do evento e a respectiva classificação com base na definição do *cutoff*.

**Tabela 13.10** Tabela de classificação para 10 observações ( $cutoff = 0,5$ ).

	Incidência Real do Evento	Incidência Real do Não Evento
Classificado como Evento	1	5
Classificado como Não Evento	1	3

Logo, podemos elaborar uma nova tabela de classificação (Tabela 13.10), ainda com base apenas nestas 10 observações, a fim de avaliarmos se as observações foram corretamente classificadas com um *cutoff* de 0,5.

Em outras palavras, para estas 10 observações, apenas uma delas foi evento e apresentou uma probabilidade maior do que 0,5, ou seja, foi evento e de fato foi classificada como tal (classificada corretamente). Outras 3 observações também foram classificadas corretamente, ou seja, não foram evento e de fato não foram classificadas como evento. Por outro lado, 6 observações foram classificadas de forma incorreta, ou seja, enquanto uma foi evento, embora tenha apresentado uma probabilidade menor do que 0,5 e, portanto, não foi classificada como evento, outras 5 não foram evento mas apresentaram probabilidades estimadas maiores do que 0,5 e, consequentemente, foram classificadas como evento.

Para a nossa amostra de 100 observações, podemos elaborar a Tabela 13.11, que traz a classificação completa para um *cutoff* de 0,5. Esta tabela será também obtida por meio da modelagem no Stata e no SPSS.

**Tabela 13.11** Tabela de classificação para a amostra completa (*cutoff* = 0,5).

	Incidência Real do Evento	Incidência Real do Não Evento
Classificado como Evento	56	11
Classificado como Não Evento	3	30

Para a amostra completa, podemos verificar que 86 observações foram classificadas corretamente, para um *cutoff* de 0,5, sendo que 56 delas foram evento e de fato foram classificadas como tal, e outras 30 não foram evento e não foram classificadas como evento com este *cutoff*. Entretanto, 14 observações foram classificadas incorretamente, sendo que 3 foram evento mas não foram classificadas como tal e 11 não foram evento mas foram classificadas como tendo sido.

Esta análise, conhecida por **análise de sensibilidade**, gera classificações que dependem da escolha do *cutoff*. Mais adiante, faremos alterações no *cutoff*, de modo a mostrar que as quantidades de observações classificadas, respectivamente, como *evento* ou *não evento* mudarão.

Neste momento, definiremos os conceitos de **eficiência global do modelo**, **sensitividade** e **especificidade**.

A **eficiência global do modelo** corresponde ao percentual de acerto da classificação para um determinado *cutoff*. Para o nosso exemplo, a eficiência global do modelo é calculada da seguinte forma:

$$EGM = \frac{56 + 30}{100} = 0,8600$$

Logo, para um *cutoff* de 0,5, 86,00% das observações são classificadas corretamente. Conforme mencionado na seção 13.2.2, a eficiência global do modelo, para um determinado *cutoff*, é bem mais adequada para se avaliar o desempenho da modelagem do que o pseudo  $R^2$  de McFadden, uma vez que a variável dependente apresenta-se na forma qualitativa dicotômica.

A **sensitividade** diz respeito ao percentual de acerto, para um determinado *cutoff*, considerando-se apenas as observações que de fato são evento. Logo, no nosso exemplo o denominador para o cálculo da sensibilidade é 59, e sua expressão é dada por:

$$Sensitividade = \frac{56}{59} = 0,9492$$

Assim, para um *cutoff* de 0,5, 94,92% das observações que são evento são classificadas corretamente.

Já a **especificidade**, por outro lado, refere-se ao percentual de acerto, para um dado *cutoff*, considerando-se apenas as observações que não são evento. No nosso exemplo, a sua expressão é dada por:

$$Especificidade = \frac{30}{41} = 0,7317$$



Desta forma, 73,17% das observações que não são evento são classificadas corretamente, ou seja, para um *cutoff* de 0,5, apresentam probabilidades de ocorrência do evento menores do que 50%.

Obviamente, a eficiência global do modelo, a sensibilidade e a especificidade mudam quando é alterado o valor do *cutoff*. A Tabela 13.12 apresenta uma nova classificação para as observações da amostra, considerando-se um *cutoff* de 0,3. Para este caso, teremos o seguinte critério de classificação:

Se  $p_i > 0,3 \rightarrow$  a observação  $i$  deverá ser classificada como *evento*.

Se  $p_i < 0,3 \rightarrow$  a observação  $i$  deverá ser classificada como *não evento*.

**Tabela 13.12** Tabela de classificação para a amostra completa (*cutoff* = 0,3).

	<b>Incidência Real do Evento</b>	<b>Incidência Real do Não Evento</b>
Classificado como Evento	57	13
Classificado como Não Evento	2	28
	<b>Eficiência Global do Modelo</b>	<b>0,8500</b>
	<b>Sensitividade</b>	<b>0,9661</b>
	<b>Especificidade</b>	<b>0,6829</b>

Em comparação aos valores obtidos para um *cutoff* de 0,5, podemos perceber, neste caso (*cutoff* de 0,3), que, enquanto a sensibilidade apresenta um pequeno aumento, a especificidade é reduzida de forma um pouco mais acentuada, o que resulta, no âmbito geral, numa redução percentual da eficiência global do modelo.

Vamos agora fazer mais uma alteração no *cutoff*, que passará, por exemplo, a ser 0,7. Para esta nova situação, teremos o seguinte critério de classificação:

Se  $p_i > 0,7 \rightarrow$  a observação  $i$  deverá ser classificada como *evento*.

Se  $p_i < 0,7 \rightarrow$  a observação  $i$  deverá ser classificada como *não evento*.

A Tabela 13.13 traz esta nova classificação, com os cálculos da eficiência global do modelo, da sensibilidade e da especificidade.

**Tabela 13.13** Tabela de classificação para a amostra completa (*cutoff* = 0,7).

	<b>Incidência Real do Evento</b>	<b>Incidência Real do Não Evento</b>
Classificado como Evento	47	5
Classificado como Não Evento	12	36
	<b>Eficiência Global do Modelo</b>	<b>0,8300</b>
	<b>Sensitividade</b>	<b>0,7966</b>
	<b>Especificidade</b>	<b>0,8780</b>

Neste caso, verificamos outro comportamento, ou seja, enquanto a sensibilidade apresenta uma redução considerável, a especificidade aumenta. Podemos inclusive perceber que a taxa de acerto para aqueles que são evento passa a ser menor do que a taxa de acerto para os que não são evento. Entretanto, a eficiência geral do modelo, com *cutoff* de 0,7, também apresenta uma redução percentual em relação ao modelo com *cutoff* de 0,5.

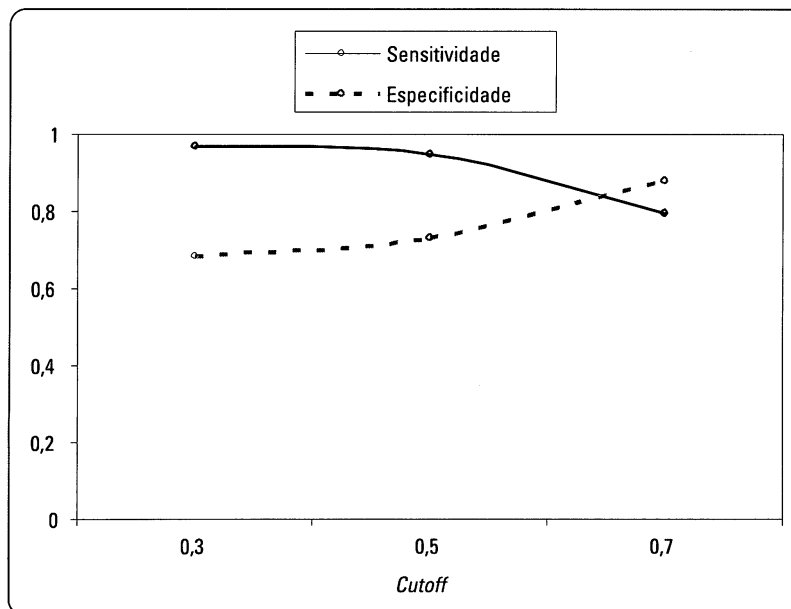
Esta análise de sensibilidade pode ser feita com qualquer valor de *cutoff* entre 0 e 1, o que permite que o pesquisador possa tomar uma decisão no sentido de definir um *cutoff* que atenda aos seus objetivos de previsão. Se, por exemplo, o objetivo for o de maximizar a eficiência global do modelo, pode ser utilizado um determinado *cutoff* que, como sabemos, poderá gerar valores de sensibilidade ou de especificidade não maximizados. Se, por outro lado, o objetivo for o de maximizar a sensibilidade, ou seja, a taxa de acerto para aqueles que são evento, poderá ser definido outro *cutoff* que não necessariamente aquele que maximizará a eficiência global do modelo.

Por fim, de maneira análoga, se houver o intuito de maximizar a taxa de acerto para as observações que não são evento (especificidade), outro *cutoff* ainda poderá ser definido.

Em outras palavras, a análise de sensibilidade é elaborada com base na teoria subjacente a cada estudo e leva em consideração as escolhas desejadas pelo pesquisador em termos de previsão de ocorrência do evento para observações não presentes na amostra, sendo, portanto, uma análise gerencial e estratégica sobre o fenômeno que se está investigando.

Em trabalhos acadêmicos e em relatórios gerenciais de diversas organizações, é comum que sejam apresentados e discutidos alguns gráficos da análise de sensibilidade. Os mais comuns são os conhecidos por **curva de sensibilidade** e **curva ROC (Receiver Operating Characteristic)**, que apresentam finalidades distintas. Enquanto a curva de sensibilidade é um gráfico que apresenta os valores da sensibilidade e da especificidade em função dos diversos valores de *cutoff*, a curva ROC é um gráfico que apresenta a variação da sensibilidade em função de  $(1 - \text{especificidade})$ .

Para os dados calculados no nosso exemplo, apresentamos a curva de sensibilidade (Figura 13.14) e a curva ROC (Figura 13.15). Embora não estejam completas, já que foram utilizados apenas três valores de *cutoff* (0,3, 0,5 e 0,7), tais curvas já permitem que sejam elaboradas algumas análises.

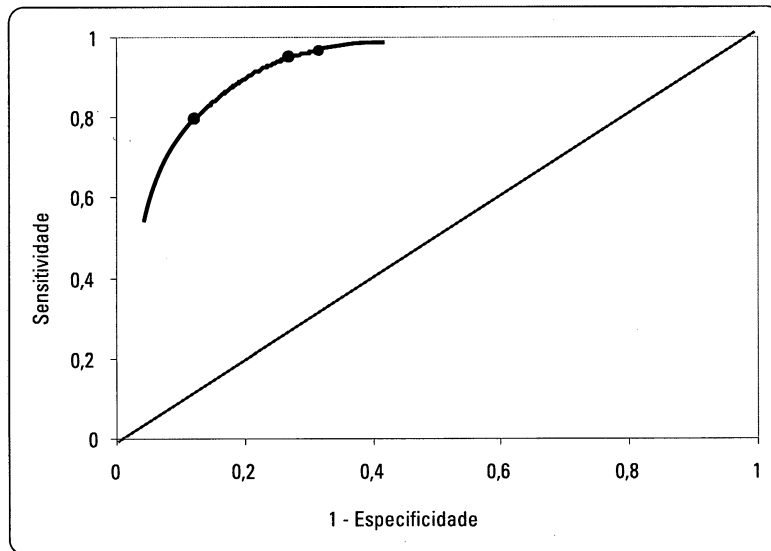


**Figura 13.14** Curva de sensibilidade para três valores de *cutoff*.

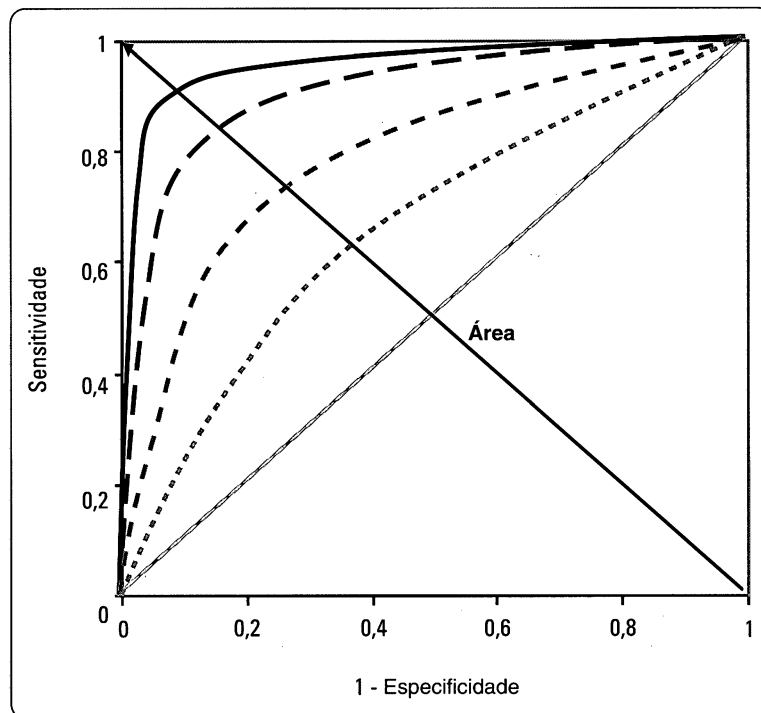
Por meio da curva de sensibilidade, podemos verificar que é possível definir o *cutoff* que iguala a sensibilidade com a especificidade, ou seja, o *cutoff* que faz com que a taxa de acerto de previsão para aqueles que serão evento seja igual à taxa de acerto para aqueles que não serão evento. É importante mencionar, contudo, que este *cutoff* não garante que a eficiência global do modelo seja a máxima possível.

Além disso, a curva de sensibilidade permite que o pesquisador avalie o *trade off* entre sensibilidade e especificidade quando da alteração do *cutoff*, já que, em muitos casos, conforme discutido, o objetivo da previsão pode ser o de aumentar a taxa de acerto para aqueles que serão evento sem que haja uma perda considerável de taxa de acerto para aqueles que não serão evento.

A curva ROC mostra o comportamento propriamente dito do *trade off* entre sensibilidade e especificidade e, ao trazer, no eixo das abscissas, os valores de  $(1 - \text{especificidade})$ , apresenta formato convexo em relação ao ponto  $(0, 1)$ . Desta forma, um determinado modelo com maior área abaixo da curva ROC apresenta maior eficiência global de previsão, combinadas todas as possibilidades de *cutoff* e, assim, a sua escolha deve ser preferível quando da comparação com outro modelo com menor área abaixo da curva ROC. Em outras palavras, se um pesquisador desejar, por exemplo, incluir novas variáveis explicativas na modelagem, a comparação do desempenho global dos modelos poderá ser elaborada com base na área abaixo da curva ROC, já que, quanto maior a sua convexidade



**Figura 13.15** Curva ROC para três valores de *cutoff*.



**Figura 13.16** Critério de escolha do modelo com maior área abaixo da curva ROC.

em relação ao ponto (0, 1), maior a sua área (maior sensibilidade e maior especificidade) e, conseqüentemente, melhor o modelo estimado para efeitos de previsão. A Figura 13.16 apresenta, de forma ilustrativa, este conceito.

Segundo Swets (1996), a curva ROC (*Receiver Operating Characteristic*), possui este nome porque compara o comportamento de alteração de duas características operacionais do modelo (sensitividade e especificidade). Foi primeiramente desenvolvida e utilizada por engenheiros na Segunda Guerra Mundial quando do estudo para detecção de objetos inimigos em batalhas. Na sequência, foi logo introduzida na Psicologia para a investigação das detecções perceptuais de determinados estímulos e, atualmente, é bastante utilizada em campos da Medicina, como a radiologia, e em diversos campos das ciências sociais aplicadas, como Economia e Finanças. Neste caso específico, é consideravelmente utilizada em modelos de gestão de risco de crédito e de probabilidade de *default*.

Nas seções 13.4 e 13.5 apresentaremos a curva de sensibilidade e a curva ROC elaboradas por meio dos softwares Stata e SPSS, respectivamente, com todas as possibilidades de valores de *cutoff* entre 0 e 1 para o modelo final estimado, inclusive com o cálculo da respectiva área abaixo da curva ROC.

### 13.3. O MODELO DE REGRESSÃO LOGÍSTICA MULTINOMIAL

Quando a variável dependente que representa o fenômeno em estudo é qualitativa, porém oferece mais de duas possibilidades de resposta (categorias), devemos fazer uso da regressão logística multinomial para estimarmos as probabilidades de ocorrência de cada alternativa. Para tanto, precisamos definir inicialmente a categoria de referência.

Imaginemos uma situação em que a variável dependente se apresenta na forma qualitativa com três categorias possíveis de resposta (0, 1 ou 2). Se a categoria de referência escolhida for a categoria 0, teremos duas outras possibilidades de evento em relação a esta categoria, que serão representadas pelas categorias 1 e 2 e, dessa forma, serão definidos dois vetores de variáveis explicativas com os respectivos parâmetros estimados, ou seja, dois logitos, como segue:

$$Z_{i_1} = \alpha_1 + \beta_{11} \cdot X_{1i} + \beta_{21} \cdot X_{2i} + \dots + \beta_{k1} \cdot X_{ki} \quad (13.23)$$

$$Z_{i_2} = \alpha_2 + \beta_{12} \cdot X_{1i} + \beta_{22} \cdot X_{2i} + \dots + \beta_{k2} \cdot X_{ki} \quad (13.24)$$

em que o número do logito aparece agora no subscrito de cada parâmetro a ser estimado.

Assim, de maneira genérica, se a variável dependente que representa o fenômeno em estudo apresentar  $M$  categorias de resposta, o número de logitos estimados será  $(M - 1)$  e, a partir dos mesmos, poderemos estimar as probabilidades de ocorrência de cada uma das categorias. A expressão geral do logito  $Z_{i_m}$  ( $m = 0, 1, \dots, M - 1$ ) para um modelo em que a variável dependente assume  $M$  categorias de resposta é:

$$Z_{i_m} = \alpha_m + \beta_{1m} \cdot X_{1i} + \beta_{2m} \cdot X_{2i} + \dots + \beta_{km} \cdot X_{ki} \quad (13.25)$$

em que  $Z_{i_0} = 0$  e, portanto,  $e^{Z_{i_0}} = 1$ .

Até o presente momento, neste capítulo, estávamos trabalhando com duas categorias e, conseqüentemente, apenas um logito  $Z_i$ . Dessa forma, as probabilidades de ocorrência do não evento e do evento eram calculadas, respectivamente, por meio das seguintes expressões:

**Probabilidade de ocorrência do não evento:**

$$1 - p_i = \frac{1}{1 + e^{Z_i}} \quad (13.26)$$

**Probabilidade de ocorrência do evento:**

$$p_i = \frac{e^{Z_i}}{1 + e^{Z_i}} \quad (13.27)$$

Já para três categorias, e com base nas expressões (13.23) e (13.24), podemos estimar a probabilidade de ocorrência da categoria de referência 0 e as probabilidades de ocorrência dos dois eventos distintos, representados pelas categorias 1 e 2. Dessa forma, as expressões dessas probabilidades podem ser escritas da seguinte forma:

**Probabilidade de ocorrência da categoria 0 (referência):**

$$p_{i_0} = \frac{1}{1 + e^{Z_{i_1}} + e^{Z_{i_2}}} \quad (13.28)$$

**Probabilidade de ocorrência da categoria 1:**

$$p_{i_1} = \frac{e^{Z_{i_1}}}{1 + e^{Z_{i_1}} + e^{Z_{i_2}}} \quad (13.29)$$

**Probabilidade de ocorrência da categoria 2:**

$$p_{i_2} = \frac{e^{Z_{i_2}}}{1 + e^{Z_{i_1}} + e^{Z_{i_2}}} \quad (13.30)$$

de modo que a soma das probabilidades de ocorrência dos eventos, representados pelas distintas categorias, será sempre 1.

Na forma completa, as expressões (13.28), (13.29) e (13.30) podem ser escritas, respectivamente, como segue:

$$p_{i_0} = \frac{1}{1 + e^{(\alpha_1 + \beta_{11} \cdot X_{1i} + \beta_{21} \cdot X_{2i} + \dots + \beta_{k1} \cdot X_{ki})} + e^{(\alpha_2 + \beta_{12} \cdot X_{1i} + \beta_{22} \cdot X_{2i} + \dots + \beta_{k2} \cdot X_{ki})}} \quad (13.31)$$

$$p_{i_1} = \frac{e^{(\alpha_1 + \beta_{11} \cdot X_{1i} + \beta_{21} \cdot X_{2i} + \dots + \beta_{k1} \cdot X_{ki})}}{1 + e^{(\alpha_1 + \beta_{11} \cdot X_{1i} + \beta_{21} \cdot X_{2i} + \dots + \beta_{k1} \cdot X_{ki})} + e^{(\alpha_2 + \beta_{12} \cdot X_{1i} + \beta_{22} \cdot X_{2i} + \dots + \beta_{k2} \cdot X_{ki})}} \quad (13.32)$$

$$p_{i_2} = \frac{e^{(\alpha_2 + \beta_{12} \cdot X_{1i} + \beta_{22} \cdot X_{2i} + \dots + \beta_{k2} \cdot X_{ki})}}{1 + e^{(\alpha_1 + \beta_{11} \cdot X_{1i} + \beta_{21} \cdot X_{2i} + \dots + \beta_{k1} \cdot X_{ki})} + e^{(\alpha_2 + \beta_{12} \cdot X_{1i} + \beta_{22} \cdot X_{2i} + \dots + \beta_{k2} \cdot X_{ki})}} \quad (13.33)$$

De maneira geral, para um modelo em que a variável dependente assume  $M$  categorias de resposta, podemos escrever a expressão das probabilidades  $p_{i_m}$  ( $m = 0, 1, \dots, M-1$ ) da seguinte forma:

$$p_{i_m} = \frac{e^{Z_{i_m}}}{\sum_{m=0}^{M-1} e^{Z_{i_m}}} \quad (13.34)$$

Analogamente ao procedimento elaborado nas seções 13.2.1, 13.2.2 e 13.2.3, iremos agora estimar os parâmetros das expressões (13.23) e (13.24) por meio de um exemplo. Iremos também avaliar a significância estatística geral do modelo e dos parâmetros, bem como estimar os seus intervalos de confiança a um determinado nível de significância. Para tanto, faremos uso novamente, neste momento, do Excel.

### 13.3.1. Estimação do modelo de regressão logística multinomial por máxima verossimilhança

Apresentaremos os conceitos pertinentes à estimação por máxima verossimilhança dos parâmetros do modelo de regressão logística multinomial por meio de um exemplo similar ao desenvolvido ao longo da seção anterior.

Imagine, agora, que o nosso incansável professor não esteja interessado somente em estudar o que leva os alunos a chegarem ou não atrasados à escola. Neste momento, ele deseja saber também se os alunos chegam atrasados à primeira aula ou à segunda aula. Em outras palavras, o professor agora tem o interesse em investigar se algumas variáveis relativas ao trajeto dos alunos até a escola influenciam a probabilidade de não se chegar atrasado ou de se chegar atrasado à primeira aula ou à segunda aula. Logo, a variável dependente passa a ter três categorias: *não chegar atrasado*, *chegar atrasado à primeira aula* e *chegar atrasado à segunda aula*.

Sendo assim, o professor elaborou uma pesquisa com os mesmos 100 alunos da escola onde leciona, porém a realizou em outro dia. Como alguns alunos já estavam um pouco cansados de responder a tantas perguntas ultimamente, o professor, além da variável referente ao fenômeno a ser estudado, resolveu perguntar apenas sobre a distância (*dist*) e sobre o número de semáforos (*sem*) pelos quais cada um havia passado naquele dia ao se deslocar para a escola. Parte do banco de dados elaborado encontra-se na Tabela 13.14.

Conforme podemos verificar, a variável dependente assume agora três distintos valores, que nada mais são do que rótulos (*labels*) referentes a cada uma das três categorias de resposta ( $M = 3$ ). É comum, infelizmente, que pesquisadores principiantes elaborem modelos, por exemplo, de regressão múltipla, assumindo que a variável dependente é quantitativa, já que apresenta números em sua coluna. Conforme já discutido na seção anterior, **isso é um erro grave!**

O banco de dados completo deste novo exemplo encontra-se no arquivo **AtrasadoMultinomial.xls**.

**Tabela 13.14** Exemplo: atraso (não, sim à primeira aula ou sim à segunda aula) x distância percorrida e quantidade de semáforos.

Estudante	Chegou atrasado à escola (Não = 0; Sim à primeira aula = 1; Sim à segunda aula = 2) ( $Y_i$ )	Distância percorrida até a escola (quilômetros) ( $X_{1i}$ )	Quantidade de semáforos ( $X_{2i}$ )
Gabriela	2	20,5	15
Patrícia	2	21,3	18
Gustavo	2	21,4	16
Letícia	2	31,5	15
Luiz Ovídio	2	17,5	16
Leonor	2	21,5	18
Dalila	2	21,5	18
Antônio	2	23,4	18
Júlia	2	22,7	18
Mariana	2	22,7	18
...			
Rodrigo	1	16,0	16
...			
Estela	0	1,0	13

As expressões dos logitos que desejamos estimar são, portanto:

$$Z_{i_1} = \alpha_1 + \beta_{11} \cdot \text{dist}_i + \beta_{21} \cdot \text{sem}_i$$

$$Z_{i_2} = \alpha_2 + \beta_{12} \cdot \text{dist}_i + \beta_{22} \cdot \text{sem}_i$$

que se referem, respectivamente, aos eventos 1 e 2 apresentados na Tabela 13.14. Note que o evento representado pelo rótulo 0 refere-se à categoria de referência.

Logo, com base nas expressões (13.31), (13.32) e (13.33), podemos escrever as expressões das probabilidades estimadas de ocorrência de cada evento correspondente a cada categoria da variável dependente. Sendo assim, temos:

$$p_{i_0} = \frac{1}{1 + e^{(\alpha_1 + \beta_{11} \cdot \text{dist}_i + \beta_{21} \cdot \text{sem}_i)} + e^{(\alpha_2 + \beta_{12} \cdot \text{dist}_i + \beta_{22} \cdot \text{sem}_i)}}$$

$$p_{i_1} = \frac{e^{(\alpha_1 + \beta_{11} \cdot \text{dist}_i + \beta_{21} \cdot \text{sem}_i)}}{1 + e^{(\alpha_1 + \beta_{11} \cdot \text{dist}_i + \beta_{21} \cdot \text{sem}_i)} + e^{(\alpha_2 + \beta_{12} \cdot \text{dist}_i + \beta_{22} \cdot \text{sem}_i)}}$$

$$p_{i_2} = \frac{e^{(\alpha_2 + \beta_{12} \cdot \text{dist}_i + \beta_{22} \cdot \text{sem}_i)}}{1 + e^{(\alpha_1 + \beta_{11} \cdot \text{dist}_i + \beta_{21} \cdot \text{sem}_i)} + e^{(\alpha_2 + \beta_{12} \cdot \text{dist}_i + \beta_{22} \cdot \text{sem}_i)}}$$

em que  $p_{i_0}$ ,  $p_{i_1}$  e  $p_{i_2}$  representam, respectivamente, a probabilidade de que um estudante  $i$  não chegue atrasado (categoria 0), a probabilidade de que um estudante  $i$  chegue atrasado à primeira aula (categoria 1) e a probabilidade de que um estudante  $i$  chegue atrasado à segunda aula (categoria 2).

Para estimarmos os parâmetros das expressões de probabilidade, faremos novamente uso da estimação por máxima verossimilhança. Genericamente, na regressão logística multinomial, em que a variável dependente segue uma **distribuição binomial**, uma observação  $i$  pode incidir num determinado evento de interesse, dados  $M$  eventos possíveis, conforme estudamos no Capítulo 5, e, portanto, a probabilidade de ocorrência  $P_{i_m}$  ( $m = 0, 1, \dots, M - 1$ ) deste específico evento pode ser escrita da seguinte maneira:

$$p(Y_{im}) = \prod_{m=0}^{M-1} (p_{i_m})^{y_{im}} \quad (13.35)$$

Para uma amostra com  $n$  observações, podemos definir a função de verossimilhança (*likelihood function*) da seguinte forma:

$$L = \prod_{i=1}^n \prod_{m=0}^{M-1} (p_{im})^{Y_{im}} \quad (13.36)$$

de onde vem, a partir da expressão (13.34), que:

$$L = \prod_{i=1}^n \prod_{m=0}^{M-1} \left( \frac{e^{Z_{im}}}{\sum_{m=0}^{M-1} e^{Z_{im}}} \right)^{Y_{im}} \quad (13.37)$$

Analogamente ao procedimento adotado quando do estudo da regressão logística binária, iremos aqui trabalhar com o logaritmo da função de verossimilhança, o que faz com que cheguemos à seguinte função, também conhecida por *log likelihood function*:

$$LL = \sum_{i=1}^n \sum_{m=0}^{M-1} \left[ (Y_{im}) \cdot \ln \left( \frac{e^{Z_{im}}}{\sum_{m=0}^{M-1} e^{Z_{im}}} \right) \right] \quad (13.38)$$

E, portanto, podemos elaborar uma importante questão: **Dadas  $M$  categorias da variável dependente, quais os valores dos parâmetros dos logitos  $Z_{im}$  ( $m = 0, 1, \dots, M - 1$ ) representados pela expressão (13.25) que fazem com que o valor de  $LL$  da expressão (13.38) seja maximizado?** Esta fundamental questão é a chave central para a elaboração da estimação dos parâmetros da regressão logística multinomial por máxima verossimilhança (ou *maximum likelihood estimation*), e pode ser respondida com o uso de ferramentas de programação linear, a fim de que seja solucionado o problema com a seguinte função-objetivo:

$$LL = \sum_{i=1}^n \sum_{m=0}^{M-1} \left[ (Y_{im}) \cdot \ln \left( \frac{e^{Z_{im}}}{\sum_{m=0}^{M-1} e^{Z_{im}}} \right) \right] = \text{máx} \quad (13.39)$$

Voltando ao nosso exemplo, iremos resolver este problema com o uso da ferramenta **Solver** do Excel. Para tanto, devemos abrir o arquivo **AtrasadoMultinomialMáximaVerossimilhança.xls**, que servirá de auxílio para o cálculo dos parâmetros.

Neste arquivo, além da variável dependente e das variáveis explicativas, foram criadas três variáveis  $Y_{im}$  ( $m = 0, 1, 2$ ) referentes às três categorias da variável dependente, e este procedimento deve ser feito a fim de que possa ser válida a expressão (13.35). Estas variáveis foram criadas com base no critério apresentado na Tabela 13.15.

Além disso, outras seis novas variáveis também foram criadas e correspondem, respectivamente, aos logitos  $Z_{i_0}$  e  $Z_{i_2}$ , às probabilidades  $p_{i_0}$ ,  $p_{i_1}$  e  $p_{i_2}$  e ao logaritmo da função de verossimilhança  $LL_i$  para cada observação. A Tabela 13.16 mostra parte dos resultados obtidos quando todos os parâmetros forem iguais a 0.

**Tabela 13.15** Critério para criação das variáveis  $Y_{im}$  ( $m = 0, 1, 2$ ).

$Y_i$	$Y_{i0}$	$Y_{i1}$	$Y_{i2}$
0	1	0	0
1	0	1	0
2	0	0	1

**Tabela 13.16** Cálculo de  $LL$  quando  $\alpha_1 = \beta_{11} = \beta_{21} = \alpha_2 = \beta_{12} = \beta_{22} = 0$ .

Estudante	$Y_i$	$Y_{i0}$	$Y_{i1}$	$Y_{i2}$	$X_{1i}$	$X_{2i}$	$Z_{i1}$	$Z_{i2}$	$p_{i0}$	$p_{i1}$	$p_{i2}$	$LL_i$ $\sum_{m=0}^2 [(Y_{im}) \cdot \ln(p_{im})]$
Gabriela	2	0	0	1	20,5	15	0	0	0,33	0,33	0,33	-1,09861
Patrícia	2	0	0	1	21,3	18	0	0	0,33	0,33	0,33	-1,09861
Gustavo	2	0	0	1	21,4	16	0	0	0,33	0,33	0,33	-1,09861
Letícia	2	0	0	1	31,5	15	0	0	0,33	0,33	0,33	-1,09861
Luiz Ovídio	2	0	0	1	17,5	16	0	0	0,33	0,33	0,33	-1,09861
Leonor	2	0	0	1	21,5	18	0	0	0,33	0,33	0,33	-1,09861
Dalila	2	0	0	1	21,5	18	0	0	0,33	0,33	0,33	-1,09861
Antônio	2	0	0	1	23,4	18	0	0	0,33	0,33	0,33	-1,09861
Júlia	2	0	0	1	22,7	18	0	0	0,33	0,33	0,33	-1,09861
Mariana	2	0	0	1	22,7	18	0	0	0,33	0,33	0,33	-1,09861
...												
Rodrigo	1	0	1	0	16,0	16	0	0	0,33	0,33	0,33	-1,09861
...												
Estela	0	1	0	0	1,0	13	0	0	0,33	0,33	0,33	-1,09861
Somatória	$LL = \sum_{i=1}^{100} \sum_{m=0}^2 [(Y_{im}) \cdot \ln(p_{im})]$											-109,86123

Apenas para efeitos didáticos, apresentamos a seguir o cálculo de  $LL$  de uma observação em que  $Y_i = 2$  e quando todos os parâmetros forem iguais a zero:

$$\begin{aligned}
 LL_i &= \sum_{m=0}^2 [(Y_{im}) \cdot \ln(p_{im})] = (Y_{i0}) \cdot \ln(p_{i0}) + (Y_{i1}) \cdot \ln(p_{i1}) + (Y_{i2}) \cdot \ln(p_{i2}) \\
 &= (0) \cdot \ln(0,33) + (0) \cdot \ln(0,33) + (1) \cdot \ln(0,33) = -1,09861
 \end{aligned}$$

A Figura 13.17 apresenta parte das observações presentes no arquivo **AtrasadoMultinomialMáximaVerossimilhança.xls**.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Estudante	Atrasado (Y)	$Y_0$	$Y_1$	$Y_2$	Distância (X <sub>1</sub> )	Semáforos (X <sub>2</sub> )	$Z_{1i}$	$Z_{2i}$	$p_0$	$p_1$	$p_2$	$LL_i$			
2	Gabriela	2	0	0	1	20,5	15	0	0	0,33333	0,33333	0,33333	-1,09861			
3	Patrícia	2	0	0	1	21,3	18	0	0	0,33333	0,33333	0,33333	-1,09861	$\alpha_1$	0,0000	
4	Gustavo	2	0	0	1	21,4	16	0	0	0,33333	0,33333	0,33333	-1,09861			
5	Letícia	2	0	0	1	31,5	15	0	0	0,33333	0,33333	0,33333	-1,09861	$\beta_{11}$	0,0000	
6	Luiz Ovídio	2	0	0	1	17,5	16	0	0	0,33333	0,33333	0,33333	-1,09861			
7	Leonor	2	0	0	1	21,5	18	0	0	0,33333	0,33333	0,33333	-1,09861	$\beta_{21}$	0,0000	
8	Dalila	2	0	0	1	21,5	18	0	0	0,33333	0,33333	0,33333	-1,09861			
9	Antônio	2	0	0	1	23,4	18	0	0	0,33333	0,33333	0,33333	-1,09861	$\alpha_2$	0,0000	
10	Júlia	2	0	0	1	22,7	18	0	0	0,33333	0,33333	0,33333	-1,09861			
11	Mariana	2	0	0	1	22,7	18	0	0	0,33333	0,33333	0,33333	-1,09861	$\beta_{12}$	0,0000	
12	Roberto	2	0	0	1	21,7	18	0	0	0,33333	0,33333	0,33333	-1,09861			
13	Renata	2	0	0	1	19,0	18	0	0	0,33333	0,33333	0,33333	-1,09861	$\beta_{22}$	0,0000	
14	Guilherme	2	0	0	1	26,4	18	0	0	0,33333	0,33333	0,33333	-1,09861			
15	Rodrigo	1	0	1	0	16,0	16	0	0	0,33333	0,33333	0,33333	-1,09861			
16	Giulia	2	0	0	1	19,0	18	0	0	0,33333	0,33333	0,33333	-1,09861			
17	Felipe	2	0	0	1	20,0	15	0	0	0,33333	0,33333	0,33333	-1,09861			
18	Karina	2	0	0	1	22,0	18	0	0	0,33333	0,33333	0,33333	-1,09861			
19	Pietro	2	0	0	1	19,2	18	0	0	0,33333	0,33333	0,33333	-1,09861			
20	Cecília	2	0	0	1	21,0	18	0	0	0,33333	0,33333	0,33333	-1,09861			
21	Osísele	2	0	0	1	20,0	14	0	0	0,33333	0,33333	0,33333	-1,09861			
22	Elaine	1	0	1	0	22,0	15	0	0	0,33333	0,33333	0,33333	-1,09861			
23	Kamal	2	0	0	1	20,0	17	0	0	0,33333	0,33333	0,33333	-1,09861			
24	Rodolfo	2	0	0	1	20,0	18	0	0	0,33333	0,33333	0,33333	-1,09861			
25	Pilar	2	0	0	1	21,0	13	0	0	0,33333	0,33333	0,33333	-1,09861			
26	Vivian	1	0	1	0	16,7	15	0	0	0,33333	0,33333	0,33333	-1,09861			
27	Danielle	0	1	0	0	17,0	10	0	0	0,33333	0,33333	0,33333	-1,09861			
28	Juliana	0	1	0	0	14,4	10	0	0	0,33333	0,33333	0,33333	-1,09861			
101	Estela	0	1	0	0	1,0	13	0	0	0,33333	0,33333	0,33333	-1,09861			
102																
103														Somatória $LL_i$	-109,86123	

**Figura 13.17** Dados do arquivo **AtrasadoMultinomialMáximaVerossimilhança.xls**.



Conforme discutimos na seção 13.2.1, aqui também deve haver uma combinação ótima de valores dos parâmetros, de modo que a função-objetivo apresentada na expressão (13.39) seja obedecida, ou seja, que o valor da somatória do logaritmo da função de verossimilhança seja o máximo possível. Recorreremos novamente ao **Solver** do Excel para resolver este problema.

A função-objetivo está na célula M103, que será a nossa célula de destino e que deverá ser maximizada. Os parâmetros  $\alpha_1, \beta_{11}, \beta_{21}, \alpha_2, \beta_{12}$  e  $\beta_{22}$ , cujos valores estão nas células P3, P5, P7, P9, P11 e P13, respectivamente, são as células variáveis. A janela do **Solver** ficará conforme mostra a Figura 13.18.

Ao clicarmos em **Resolver** e em **OK**, obteremos a solução ótima do problema de programação linear. A Tabela 13.17 mostra parte dos valores obtidos.

Parâmetros do Solver

Definir Objetivo:

Para: ☒ Máx. ☐ Mín. ☐ Valor de:

Alterando Células Variáveis:

Sujeito às Restrições:

☐ Tornar Variáveis Irrestritas Não Negativas

Selecionar um Método de Solução:

Método de Solução  
Selecione o mecanismo GRG Não Linear para Problemas do Solver suaves e não lineares. Selecione o mecanismo LP Simplex para Problemas do Solver lineares. Selecione o mecanismo Evolutionary para problemas do Solver não suaves.

**Figura 13.18** Solver – Maximização da somatória do logaritmo da função de verossimilhança para o modelo de regressão logística multinomial.

O valor máximo possível da somatória do logaritmo da função de verossimilhança é  $LL_{\max} = -24,51180$ . A resolução deste problema gerou as seguintes estimativas dos parâmetros:

$$\begin{aligned}\alpha_1 &= -33,135 \\ \beta_{11} &= 0,559 \\ \beta_{21} &= 1,670 \\ \alpha_2 &= -62,292 \\ \beta_{12} &= 1,078 \\ \beta_{22} &= 2,895\end{aligned}$$

**Tabela 13.17** Valores obtidos quando da maximização de  $LL$  para o modelo de regressão logística multinomial.

Estudante	$Y_i$	$Y_{i0}$	$Y_{i1}$	$Y_{i2}$	$X_{1i}$	$X_{2i}$	$Z_{i1}$	$Z_{i2}$	$p_{i0}$	$p_{i1}$	$p_{i2}$	$LL_i$ $\sum_{m=0}^2 [(Y_{im}) \cdot \ln(p_{im})]$
Gabriela	2	0	0	1	20,5	15	3,37036	3,23816	0,01799	0,52341	0,45860	-0,77959
Patrícia	2	0	0	1	21,3	18	8,82883	12,78751	0,00000	0,01873	0,98127	-0,01891
Gustavo	2	0	0	1	21,4	16	5,54391	7,10441	0,00068	0,17346	0,82586	-0,19133
Letícia	2	0	0	1	31,5	15	9,51977	15,10301	0,00000	0,00375	0,99625	-0,00375
Luiz Ovídio	2	0	0	1	17,5	16	3,36367	2,89778	0,02082	0,60162	0,37756	-0,97402
Leonor	2	0	0	1	21,5	18	8,94064	13,00323	0,00000	0,01691	0,98308	-0,01706
Dalila	2	0	0	1	21,5	18	8,94064	13,00323	0,00000	0,01691	0,98308	-0,01706
Antônio	2	0	0	1	23,4	18	10,00281	15,05262	0,00000	0,00637	0,99363	-0,00639
Júlia	2	0	0	1	22,7	18	9,61149	14,29758	0,00000	0,00914	0,99086	-0,00918
Mariana	2	0	0	1	22,7	18	9,61149	14,29758	0,00000	0,00914	0,99086	-0,00918
...												
Rodrigo	1	0	1	0	16,0	16	2,52511	1,27985	0,05852	0,73104	0,21044	-0,31329
...												
Estela	0	1	0	0	1,0	13	0	-10,87168	-23,58594	0,99998	0,00002	0,00000
<b>Somatória</b>	$LL = \sum_{i=1}^{100} \sum_{m=0}^2 [(Y_{im}) \cdot \ln(p_{im})]$											<b>-24,51180</b>

e, desta forma, os logitos  $Z_{i1}$  e  $Z_{i2}$  podem ser escritos da seguinte forma:

$$Z_{i1} = -33,135 + 0,559 \cdot \text{dist}_i + 1,670 \cdot \text{sem}_i$$

$$Z_{i2} = -62,292 + 1,078 \cdot \text{dist}_i + 2,895 \cdot \text{sem}_i$$

A Figura 13.19 apresenta parte dos resultados obtidos pela modelagem no arquivo **AtrasadoMultinomial-MáximaVerossimilhança.xls**.

Com base nas expressões dos logitos  $Z_{i1}$  e  $Z_{i2}$ , podemos escrever as expressões das probabilidades de ocorrência de cada uma das categorias da variável dependente, como segue:

**Probabilidade de um estudante  $i$  não chegar atrasado (categoria 0):**

$$p_{i0} = \frac{1}{1 + e^{(-33,135 + 0,559 \cdot \text{dist}_i + 1,670 \cdot \text{sem}_i)} + e^{(-62,292 + 1,078 \cdot \text{dist}_i + 2,895 \cdot \text{sem}_i)}}$$

**Probabilidade de um estudante  $i$  chegar atrasado à primeira aula (categoria 1):**

$$p_{i1} = \frac{e^{(-33,135 + 0,559 \cdot \text{dist}_i + 1,670 \cdot \text{sem}_i)}}{1 + e^{(-33,135 + 0,559 \cdot \text{dist}_i + 1,670 \cdot \text{sem}_i)} + e^{(-62,292 + 1,078 \cdot \text{dist}_i + 2,895 \cdot \text{sem}_i)}}$$

**Probabilidade de um estudante  $i$  chegar atrasado à segunda aula (categoria 2):**

$$p_{i2} = \frac{e^{(-62,292 + 1,078 \cdot \text{dist}_i + 2,895 \cdot \text{sem}_i)}}{1 + e^{(-33,135 + 0,559 \cdot \text{dist}_i + 1,670 \cdot \text{sem}_i)} + e^{(-62,292 + 1,078 \cdot \text{dist}_i + 2,895 \cdot \text{sem}_i)}}$$

Tendo sido elaborada a estimação por máxima verossimilhança dos parâmetros das equações de probabilidade de ocorrência de cada uma das categorias da variável dependente, podemos elaborar a classificação das

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Estudante	Atrasado (Y)	$Y_0$	$Y_1$	$Y_2$	Distância ( $X_1$ )	Semáforos ( $X_2$ )	$Z_{11}$	$Z_{12}$	$P_0$	$P_{11}$	$P_{12}$	$LL_i$			
2	Gabriela	2	0	0	1	20,5	15	3,36938	3,23724	0,01801	0,52339	0,45860	-0,77957			
3	Patrícia	2	0	0	1	21,3	18	8,82617	12,78452	0,00000	0,01874	0,98126	-0,01892	$\alpha_1$		-33,135
4	Gustavo	2	0	0	1	21,4	16	5,54223	7,10263	0,00068	0,17347	0,82585	-0,19134	$\beta_{11}$		0,559
5	Leticia	2	0	0	1	31,5	15	9,51650	15,09930	0,00000	0,00375	0,99625	-0,00376	$\beta_{21}$		1,670
6	Lutz Ovídio	2	0	0	1	17,5	16	3,36280	2,89699	0,02084	0,60159	0,37757	-0,97399	$\alpha_2$		-62,292
7	Leonor	2	0	0	1	21,5	18	8,93793	13,00019	0,00000	0,01692	0,98308	-0,01707	$\beta_{12}$		1,078
8	Dallia	2	0	0	1	21,5	18	8,93793	13,00019	0,00000	0,01692	0,98308	-0,01707	$\beta_{22}$		2,895
9	Antônio	2	0	0	1	23,4	18	9,99971	15,04909	0,00000	0,00637	0,99363	-0,00639			
10	Júlia	2	0	0	1	22,7	19	9,60853	14,29423	0,00000	0,00914	0,99086	-0,00918			
11	Mariana	2	0	0	1	22,7	18	9,60853	14,29423	0,00000	0,00914	0,99086	-0,00918			
12	Roberto	2	0	0	1	21,7	18	9,04970	13,21586	0,00000	0,01527	0,98472	-0,01539			
13	Renata	2	0	0	1	19,0	18	7,54086	10,30427	0,00003	0,05933	0,94064	-0,06120			
14	Ouilherme	2	0	0	1	26,4	18	11,67620	18,28420	0,00000	0,00135	0,99865	-0,00135			
15	Rodrigo	1	0	1	0	16,0	16	2,52456	1,27944	0,05855	0,73099	0,21046	-0,31355			
16	Giulia	2	0	0	1	19,0	18	7,54086	10,30427	0,00003	0,05933	0,94064	-0,06120			
17	Felipe	2	0	0	1	20,0	15	3,08997	2,69805	0,02644	0,58097	0,39260	-0,93497			
18	Karina	2	0	0	1	22,0	18	9,21735	13,53937	0,00000	0,01310	0,98690	-0,01319			
19	Pietro	2	0	0	1	19,2	18	7,65263	10,51994	0,00003	0,05379	0,94618	-0,05532			
20	Cecília	2	0	0	1	21,0	18	8,65852	12,46100	0,00000	0,02183	0,97817	-0,02207			
21	Glisele	2	0	0	1	20,0	14	1,42006	-0,19681	0,16782	0,69434	0,13784	-1,98166			
22	Elaine	1	0	1	0	22,0	15	4,20762	4,85479	0,00509	0,34188	0,65303	-1,07330			
23	Kamal	2	0	0	1	20,0	17	6,42978	8,48777	0,00018	0,11323	0,88659	-0,12037			
24	Rodolfo	2	0	0	1	20,0	18	8,09969	11,38264	0,00001	0,03616	0,96383	-0,03684			
25	Pilar	2	0	0	1	21,0	13	0,30898	-2,01330	0,40071	0,54578	0,05351	-2,92782			
26	Vítor	1	0	1	0	16,7	15	1,24583	-0,86056	0,20413	0,70953	0,08633	-0,34315			
27	Danielle	0	1	0	0	17,0	10	-6,93606	-15,01136	0,99903	0,00097	0,00000	-0,00097			
28	Juliana	0	1	0	0	14,4	10	-8,38901	-17,81512	0,99977	0,00023	0,00000	-0,00023			
101	Estela	0	1	0	0	1,0	13	-10,86760	-23,58068	0,99998	0,00002	0,00000	-0,00002			
102																
103																
																Somatória $LL_i$ -24,51180

**Figura 13.19** Obtenção dos parâmetros da regressão logística multinomial quando da maximização de  $LL$  pelo Solver.

observações e definir a **eficiência global do modelo de regressão logística multinomial**. Diferentemente da regressão logística binária, em que a classificação é elaborada com base na definição de um *cutoff*, na regressão logística multinomial a classificação de cada observação é feita com base na maior probabilidade entre aquelas calculadas ( $P_{i0}$ ,  $P_{i1}$  ou  $P_{i2}$ ). Assim, por exemplo, como a observação 1 (Gabriela) apresentou  $P_{i0} = 0,018$ ,  $P_{i1} = 0,523$  e  $P_{i2} = 0,459$ , devemos classificá-la como categoria 1, ou seja, por meio do nosso modelo espera-se que a Gabriela chegue atrasada à primeira aula. Entretanto, podemos verificar que, na verdade, esta aluna chegou atrasada à segunda aula e, portanto, para este caso, não obtivemos um acerto.

A Tabela 13.18 apresenta a classificação para a nossa amostra completa, com ênfase para os percentuais de acerto para cada categoria da variável dependente, destacando também a eficiência global do modelo (percentual total de acerto).

Por meio da análise desta tabela, podemos verificar que o modelo apresenta um percentual total de acerto de 89,0%. Entretanto, o modelo apresenta um maior percentual de acerto (95,9%) para os casos em que houver indicação de que não ocorrerá atraso ao se chegar à escola. Por outro lado, quando houver indícios de que um aluno chegará atrasado à primeira aula, o modelo terá um percentual de acerto menor (75,0%).

Partiremos agora para o estudo da significância estatística geral do modelo obtido, bem como das significâncias estatísticas dos próprios parâmetros, como fizemos na seção 13.2.

**Tabela 13.18** Tabela de classificação para a amostra completa.

Observado	Classificado			
	Não chegou atrasado	Chegou atrasado à primeira aula	Chegou atrasado à segunda aula	Percentual de Acerto
Não chegou atrasado	47	2	0	95,9%
Chegou atrasado à primeira aula	1	12	3	75,0%
Chegou atrasado à segunda aula	0	5	30	85,7%
Eficiência Global do Modelo				89,0%

### 13.3.2. A significância estatística geral do modelo e dos parâmetros da regressão logística multinomial

Assim como na regressão logística binária estudada na seção 13.2, a modelagem da regressão logística multinomial também oferece as estatísticas referentes ao pseudo  $R^2$  de McFadden e ao  $\chi^2$ , cujos cálculos são elaborados, respectivamente, com base nas expressões (13.16) e (13.17), sendo aqui novamente reproduzidas:

$$\text{pseudo } R^2 = \frac{-2.LL_0 - (-2.LL_{\max})}{-2.LL_0} \quad (13.40)$$

$$\chi^2 = -2.(LL_0 - LL_{\max}) \quad (13.41)$$

Enquanto o pseudo  $R^2$  de McFadden, conforme já discutido na seção 13.2.2, é bastante limitado em termos de informação sobre o ajuste do modelo, podendo ser utilizado quando o pesquisador tiver interesse em comparar dois modelos distintos, a estatística  $\chi^2$  propicia que seja realizado um teste para verificação da existência propriamente dita do modelo proposto, uma vez que, se todos os parâmetros estimados  $\beta_{jm}$  ( $j = 1, 2, \dots, k; m = 1, 2, \dots, M - 1$ ) forem estatisticamente iguais a 0, o comportamento de alteração de cada uma das variáveis explicativas não influenciará em absolutamente nada as probabilidades de ocorrência dos eventos representados pelas categorias da variável dependente. As hipóteses nula e alternativa do teste  $\chi^2$ , para um modelo geral de regressão logística multinomial, são, respectivamente:

$$H_0: \beta_{11} = \beta_{21} = \dots = \beta_{k1} = \beta_{12} = \beta_{22} = \dots = \beta_{k2} = \beta_{1, M-1} = \beta_{2, M-1} = \dots = \beta_{k, M-1} = 0$$

$$H_1: \text{existe pelo menos um } \beta_{jm} \neq 0$$

Voltando ao nosso exemplo, temos que  $LL_{\max}$ , que é o valor máximo possível da somatória do logaritmo da função de verossimilhança, é igual a -24,51180. Para o cálculo de  $LL_0$ , que representa o valor máximo possível da somatória do logaritmo da função de verossimilhança para um modelo que só apresenta as constantes  $\alpha_1$  e  $\alpha_2$  e nenhuma variável explicativa, faremos novamente uso do **Solver**, por meio do arquivo **AtrasadoMultinomialMáximaVerossimilhançaModeloNulo.xls**. As Figuras 13.20 e 13.21 mostram, respectivamente, a janela do **Solver** e parte dos resultados obtidos pela modelagem neste arquivo.

Com base no modelo nulo, temos  $LL_0 = -101,01922$  e, dessa forma, podemos calcular as seguintes estatísticas:

$$\text{pseudo } R^2 = \frac{-2.(-101,01922) - [(-2.(-24,51180))]}{-2.(-101,01922)} = 0,7574$$

$$\chi^2_{4 g.l.} = -2.[-101,01922 - (-24,51180)] = 153,0148$$

Para 4 graus de liberdade (número de parâmetros  $\beta$ , já que há duas variáveis explicativas e dois logitos), temos, por meio da Tabela D do apêndice do livro, que o  $\chi^2_c = 9,488$  ( $\chi^2$  crítico para 4 graus de liberdade e para o nível de significância de 5%). Dessa forma, como o  $\chi^2$  calculado  $\chi^2_{\text{cal}} = 153,0148 > \chi^2_c = 9,488$ , podemos rejeitar a hipótese nula de que todos os parâmetros  $\beta_{jm}$  ( $j = 1, 2; m = 1, 2$ ) sejam estatisticamente iguais a zero. Logo, pelo menos uma variável  $X$  é estatisticamente significativa para explicar a probabilidade de ocorrência de pelo menos um dos eventos em estudo. Da mesma forma que o discutido na seção 13.2.2, podemos definir o seguinte critério:

Se *valor-P* (ou *P-value* ou *Sig.  $\chi^2_{\text{cal}}$*  ou *Prob.  $\chi^2_{\text{cal}}$* )  $< 0,05$ , existe pelo menos um  $\beta_{jm} \neq 0$ .

Além da significância estatística geral do modelo, é necessário verificarmos a significância estatística de cada parâmetro, por meio da análise das respectivas estatísticas  $z$  de Wald, cujas hipóteses nulas e alternativa são, para os parâmetros  $\alpha_m$  ( $m = 1, 2, \dots, M - 1$ ) e  $\beta_{jm}$  ( $j = 1, 2, \dots, k; m = 1, 2, \dots, M - 1$ ), respectivamente:

$$H_0: \alpha_m = 0$$

$$H_1: \alpha_m \neq 0$$

$$H_0: \beta_{jm} = 0$$

$$H_1: \beta_{jm} \neq 0$$

**Parâmetros do Solver**

Definir Objetivo:

Para: ☒ Máx. ☐ Mín. ☐ Valor de:

Alterando Células Variáveis:

Sujeito às Restrições:

☐ Tomar Variáveis Irestritas Não Negativas

Selecionar um Método de Solução:

**Método de Solução**  
 Seleciona o mecanismo GRG Não Linear para Problemas do Solver suaves e não lineares. Seleciona o mecanismo LP Simplex para Problemas do Solver lineares. Seleciona o mecanismo Evolutionary para problemas do Solver não suaves.

**Botões de Ação:** Adicionar, Alterar, Excluir, Redefinir Tudo, Carregar/Salvar, Opções, Ajuda, Resolver, Fechar

**Figura 13.20** Solver – Maximização da somatória do logaritmo da função de verossimilhança para o modelo nulo da regressão logística multinomial.

As estatísticas  $z$  de Wald são obtidas com base na expressão (13.18), porém, mantendo o padrão do exposto na seção 13.2.2, não faremos os cálculos dos erros-padrão de cada parâmetro que, para o nosso exemplo, são:

$$s.e. (\alpha_1) = 12,183$$

$$s.e. (\beta_{11}) = 0,243$$

$$s.e. (\beta_{21}) = 0,577$$

$$s.e. (\alpha_2) = 14,675$$

$$s.e. (\beta_{12}) = 0,302$$

$$s.e. (\beta_{22}) = 0,686$$

Logo, como já elaboramos as estimativas dos parâmetros, temos que:

$$z_{\alpha_1} = \frac{\alpha_1}{s.e.(\alpha_1)} = \frac{-33,135}{12,183} = -2,720$$

$$z_{\beta_{11}} = \frac{\beta_{11}}{s.e.(\beta_{11})} = \frac{0,559}{0,243} = 2,300$$

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Estudante	Atrasado (Y)	Y <sub>0</sub>	Y <sub>1</sub>	Y <sub>1</sub>	Distância (X <sub>1</sub> )	Semáforos (X <sub>2</sub> )	Z <sub>1</sub>	Z <sub>2</sub>	P <sub>0</sub>	P <sub>1</sub>	P <sub>2</sub>	LL <sub>1</sub>			
2	Gabriela	2	0	0	1	20,5	15	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
3	Patricia	2	0	0	1	31,5	18	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982	a <sub>1</sub>		-1,719
4	Gustavo	2	0	0	1	21,4	16	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
5	Leandro	2	0	0	1	31,5	18	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982	a <sub>2</sub>		-0,825
6	Luiz Ovídio	2	0	0	1	17,5	16	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
7	Lauro	2	0	0	1	21,5	18	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
8	Dália	2	0	0	1	21,5	18	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
9	Adriano	2	0	0	1	28,4	18	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
10	Júlia	2	0	0	1	22,7	18	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
11	Marcelo	2	0	0	1	22,7	18	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
12	Roberto	2	0	0	1	21,7	18	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
13	Isabela	2	0	0	1	19,0	18	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
14	Ogullherme	2	0	0	1	26,4	18	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
15	Alcides	1	0	1	0	15,0	16	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
16	Glória	2	0	0	1	19,0	18	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
17	Adriano	2	0	0	1	20,0	18	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
18	Karina	2	0	0	1	22,0	18	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
19	Paulo	2	0	0	1	19,2	18	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
20	Cecília	2	0	0	1	21,0	18	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
21	João	2	0	0	1	20,0	18	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
22	Elaine	1	0	1	0	22,0	15	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
23	Carolina	2	0	0	1	20,0	18	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
24	Roberto	2	0	0	1	20,0	18	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
25	Paula	2	0	0	1	21,0	18	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
26	Vivian	1	0	1	0	16,7	15	-1,11923	-0,33647	0,49000	0,16000	0,35000	-1,04982			
27	Carolina	0	1	0	0	17,0	18	-1,11923	-0,33647	0,49000	0,16000	0,35000	-0,71935			
28	Juliana	0	1	0	0	14,4	10	-1,11923	-0,33647	0,49000	0,16000	0,35000	-0,71935			
101	Somatória	0	1	0	0	1	19	-1,11923	-0,33647	0,49000	0,16000	0,35000	-0,71935			
102																
103																
	</															

$$p_1 = \frac{e^{[-33,135+0,559.(17)+1,670.(15)]}}{1 + e^{[-33,135+0,559.(17)+1,670.(15)]} + e^{[-62,292+1,078.(17)+2,895.(15)]}} = 0,722$$

Logo, a probabilidade média estimada de se chegar atrasado à primeira aula é, nas condições informadas, igual a 72,2%.

**Em média, em quanto se altera a chance de se chegar atrasado à primeira aula, em relação a não chegar atrasado à escola, ao se adotar um percurso 1 quilômetro mais longo, mantidas as demais condições constantes?**

Para respondermos a esta questão, vamos novamente recorrer à expressão (13.3), que poderá ser escrita da seguinte forma:

$$\text{chance}_{Y_{i1}=1} = e^{Z_{i1}} \quad (13.42)$$

de modo que, mantidas as demais condições constantes, a chance de se chegar atrasado à primeira aula em relação a não chegar atrasado à escola, ao se adotar um trajeto 1 quilômetro mais longo, é:

$$\text{chance}_{Y_{i1}=1} = e^{0,559} = 1,749$$

Logo, a chance é multiplicada por um fator de 1,749, ou seja, mantidas as demais condições constantes, a chance de se chegar atrasado à primeira aula em relação a não chegar atrasado, ao se adotar um trajeto 1 quilômetro mais longo, é, em média, 74,9% maior. Em modelos de regressão logística multinomial, a chance (*odds ratio*) também é chamada de **razão de risco relativo** (*relative risk ratio*).

**Em média, em quanto se altera a chance de se chegar atrasado à segunda aula, em relação a não chegar atrasado, ao se passar por 1 semáforo a mais no percurso até a escola, mantidas as demais condições constantes?**

Neste caso, como o evento de interesse refere-se à categoria *chegar atrasado à segunda aula*, a expressão da chance passa a ser:

$$\text{chance}_{Y_{i2}=1} = e^{2,895} = 18,081$$

Logo, a chance é multiplicada por um fator de 18,081, ou seja, mantidas as demais condições constantes, a chance de se chegar atrasado à segunda aula em relação a não chegar atrasado, ao se passar por 1 semáforo a mais no percurso até a escola, é, em média, 1.708,1% maior.

Conforme podemos perceber, estes cálculos utilizaram sempre as estimativas médias dos parâmetros. Como fizemos na seção 13.2, partiremos agora para o estudo dos intervalos de confiança destes parâmetros.

### 13.3.3. Construção dos intervalos de confiança dos parâmetros do modelo de regressão logística multinomial

Os intervalos de confiança dos parâmetros estimados em uma regressão logística multinomial também são calculados por meio da expressão (13.21) apresentada na seção 13.2.3. Logo, ao nível de confiança de 95%, podem ser definidos, para os parâmetros  $\alpha_m$  ( $m = 1, 2, \dots, M-1$ ) e  $\beta_{jm}$  ( $j = 1, 2, \dots, k; m = 1, 2, \dots, M-1$ ), respectivamente, da seguinte forma:

$$\alpha_m \pm 1,96 \cdot [s.e.(\alpha_m)] \quad (13.43)$$

$$\beta_{jm} \pm 1,96 \cdot [s.e.(\beta_{jm})]$$

em que 1,96 é o  $z_c$  para o nível de significância de 5%.

Para os dados do nosso exemplo, a Tabela 13.19 apresenta os coeficientes estimados dos parâmetros  $\alpha_m$  ( $m = 1, 2$ ) e  $\beta_{jm}$  ( $j = 1, 2; m = 1, 2$ ) das expressões das probabilidades de ocorrência dos eventos de interesse, com os respectivos erros-padrão, as estatísticas  $z$  de Wald e os intervalos de confiança para o nível de significância de 5%.

Como já sabíamos, nenhum intervalo de confiança contém o zero e, com base nos seus valores, podemos escrever as expressões dos limites inferior (mínimo) e superior (máximo) das probabilidades estimadas de ocorrência de cada uma das categorias da variável dependente.

**Tabela 13.19** Cálculo dos intervalos de confiança dos parâmetros da regressão logística multinomial.

Parâmetro	Coeficiente	Erro-Padrão (s.e.)	z	Intervalo de Confiança (95%)	
				$\alpha_m - 1,96 \cdot [s.e.(\alpha_m)]$ $\beta_{jm} - 1,96 \cdot [s.e.(\beta_{jm})]$	$\alpha_m + 1,96 \cdot [s.e.(\alpha_m)]$ $\beta_{jm} + 1,96 \cdot [s.e.(\beta_{jm})]$
$\alpha_1$ (constante)	-33,135	12,183	-2,720	-57,014	-9,256
$\beta_{11}$ (variável <i>dist</i> )	0,559	0,243	2,300	0,082	1,035
$\beta_{21}$ (variável <i>sem</i> )	1,670	0,577	2,894	0,539	2,800
$\alpha_2$ (constante)	-62,292	14,675	-4,244	-91,055	-33,529
$\beta_{12}$ (variável <i>dist</i> )	1,078	0,302	3,570	0,486	1,671
$\beta_{22}$ (variável <i>sem</i> )	2,895	0,686	4,220	1,550	4,239

**Intervalo de Confiança (95%) da probabilidade estimada de um estudante *i* não chegar atrasado (categoria 0):**

$$p_{i0\min} = \frac{1}{1 + e^{(-57,014 + 0,082 \cdot dist_i + 0,539 \cdot sem_i)} + e^{(-91,055 + 0,486 \cdot dist_i + 1,550 \cdot sem_i)}}$$

$$p_{i0\max} = \frac{1}{1 + e^{(-9,256 + 1,035 \cdot dist_i + 2,800 \cdot sem_i)} + e^{(-33,529 + 1,671 \cdot dist_i + 4,239 \cdot sem_i)}}$$

**Intervalo de Confiança (95%) da probabilidade estimada de um estudante *i* chegar atrasado à primeira aula (categoria 1):**

$$p_{i1\min} = \frac{e^{(-57,014 + 0,082 \cdot dist_i + 0,539 \cdot sem_i)}}{1 + e^{(-57,014 + 0,082 \cdot dist_i + 0,539 \cdot sem_i)} + e^{(-91,055 + 0,486 \cdot dist_i + 1,550 \cdot sem_i)}}$$

$$p_{i1\max} = \frac{e^{(-9,256 + 1,035 \cdot dist_i + 2,800 \cdot sem_i)}}{1 + e^{(-9,256 + 1,035 \cdot dist_i + 2,800 \cdot sem_i)} + e^{(-33,529 + 1,671 \cdot dist_i + 4,239 \cdot sem_i)}}$$

**Intervalo de Confiança (95%) da probabilidade estimada de um estudante *i* chegar atrasado à segunda aula (categoria 2):**

$$p_{i2\min} = \frac{e^{(-91,055 + 0,486 \cdot dist_i + 1,550 \cdot sem_i)}}{1 + e^{(-57,014 + 0,082 \cdot dist_i + 0,539 \cdot sem_i)} + e^{(-91,055 + 0,486 \cdot dist_i + 1,550 \cdot sem_i)}}$$

$$p_{i2\max} = \frac{e^{(-33,529 + 1,671 \cdot dist_i + 4,239 \cdot sem_i)}}{1 + e^{(-9,256 + 1,035 \cdot dist_i + 2,800 \cdot sem_i)} + e^{(-33,529 + 1,671 \cdot dist_i + 4,239 \cdot sem_i)}}$$

Analogamente ao elaborado na seção 13.2.3, podemos definir a expressão dos intervalos de confiança das chances (*odds* ou *relative risk ratios*) de ocorrência de cada um dos eventos representados pelo subscrito *m* ( $m = 1, 2, M - 1$ ) em relação à ocorrência do evento representado pela categoria 0 (referência) para cada parâmetro  $\beta_{jm}$  ( $j = 1, 2, \dots, k; m = 1, 2, \dots, M - 1$ ), ao nível de confiança de 95%, da seguinte forma:

$$e^{\beta_{jm} \pm 1,96 [s.e.(\beta_{jm})]} \quad (13.44)$$

Para os dados do nosso exemplo, e a partir dos valores calculados na Tabela 13.19, vamos elaborar a Tabela 13.20, que apresenta os intervalos de confiança das chances (*odds* ou *relative risk ratios*) de ocorrência de cada um dos eventos em relação ao evento de referência para cada parâmetro  $\beta_{jm}$  ( $j = 1, 2; m = 1, 2$ ).

Estes valores também serão obtidos por meio da modelagem no software Stata, a ser apresentada na próxima seção.



**Tabela 13.20** Cálculo dos intervalos de confiança das chances (*odds* ou *relative risk ratios*) para cada parâmetro  $\beta_{jm}$ .

Evento	Parâmetro	Chance (Odds)	Intervalo de Confiança da Chance (95%)	
		$e^{\beta_{jm}}$	$e^{\beta_{jm} - 1,96 \cdot [s.e.(\beta_{jm})]}$	$e^{\beta_{jm} + 1,96 \cdot [s.e.(\beta_{jm})]}$
Chegar atrasado à primeira aula	$\beta_{11}$ (variável <i>dist</i> )	1,749	1,085	2,817
	$\beta_{21}$ (variável <i>sem</i> )	5,312	1,715	16,453
Chegar atrasado à segunda aula	$\beta_{12}$ (variável <i>dist</i> )	2,939	1,625	5,318
	$\beta_{22}$ (variável <i>sem</i> )	18,081	4,713	69,363

### 13.4. ESTIMAÇÃO DE MODELOS DE REGRESSÃO LOGÍSTICA BINÁRIA E MULTINOMIAL NO SOFTWARE STATA

O objetivo desta seção não é o de discutir novamente todos os conceitos inerentes às estatísticas dos modelos de regressão logística binária e multinomial, porém propiciar ao pesquisador uma oportunidade de elaboração dos mesmos exemplos explorados ao longo do capítulo por meio do Stata Statistical Software®. A reprodução de suas imagens nesta seção tem autorização da StataCorp LP®.

#### 13.4.1. Regressão logística binária no software Stata

Voltando então ao primeiro exemplo, lembremos que um professor tinha o interesse em avaliar se a distância percorrida, a quantidade de semáforos, o período do dia em que se dava o trajeto e o perfil dos alunos ao volante influenciavam o fato de se chegar ou não atrasado à escola. Já partiremos para o banco de dados final construído pelo professor por meio dos questionamentos elaborados ao seu grupo de 100 estudantes. O banco de dados encontra-se no arquivo **Atrasado.dta** e é exatamente igual ao apresentado parcialmente na Tabela 13.2.

```
. desc

obs:      100
vars:      6
size:      2,600 (99.9% of memory free)

-----
variable name  storage  display  value  variable label
              type   format   label
-----
estudante      str11   %11s
atrasado       byte    %8.0g   atrasado   chegou atrasado à escola?
dist           float   %9.0g   distância  distância percorrida até a escola (km)
sem            byte    %8.0g   quantidade quantidade de semáforos
per            byte    %8.0g   per        período do dia
perfil         float   %9.0g   perfil     perfil ao volante
Sorted by:
```

**Figura 13.22** Descrição do banco de dados **Atrasado.dta**.

Inicialmente, podemos digitar o comando **desc**, que faz com que seja possível analisarmos as características do banco de dados, como o número de observações, o número de variáveis e a descrição de cada uma delas. A Figura 13.22 apresenta este primeiro *output* do Stata.

A variável dependente, que se refere ao fato de se chegar ou não atrasado à escola, é qualitativa e possui apenas duas categorias, já rotuladas no banco de dados como *dummy* (Não = 0; Sim = 1). O comando **tab** oferece a distribuição de frequências de uma variável qualitativa, com destaque para a quantidade de categorias. Se o pesquisador tiver dúvidas sobre o número de categorias, poderá recorrer facilmente a este comando. A Figura 13.23 apresenta a distribuição de frequências da variável dependente *atrasado*.

É comum que se discuta sobre a necessidade de igualdade de frequências entre a categoria de referência e a categoria que representa o evento de interesse quando da estimação de modelos de regressão logística binária.

```
. tab atrasado
```

chegou   atrasado à escola?	Freq.	Percent	Cum.
Não	41	41.00	41.00
Sim	59	59.00	100.00
Total	100	100.00	

Figura 13.23 Distribuição de frequências da variável *atrasado*.

O fato de as frequências não serem iguais afetará a probabilidade de ocorrência do evento de interesse para cada observação da amostra, apresentada por meio da expressão (13.11), e, conseqüentemente, o respectivo logaritmo da função de verossimilhança. Entretanto, como o nosso objetivo é estimar um modelo de probabilidade de ocorrência do evento de interesse com base na maximização da somatória do logaritmo da função de verossimilhança para toda a amostra, respeitando as características do próprio banco de dados, **não há a necessidade de que as frequências das duas categorias sejam iguais.**

Com relação às variáveis explicativas qualitativas, a variável *per* também possui apenas duas categorias que, no banco de dados, já estão rotuladas como *dummy* (manhã = 1; tarde = 0). Por outro lado, a variável *perfil* possui três categorias e, portanto, será preciso que criemos ( $n - 1 = 2$ ) *dummies*. O comando `xi i.perfil` nos fornecerá estas duas *dummies*, nomeadas pelo Stata de `_Iperfil_2` e `_Iperfil_3`. Enquanto as Figuras 13.24 e 13.25 apresentam, respectivamente, as distribuições de frequência das variáveis *per* e *perfil*, a Figura 13.26 apresenta o procedimento para a criação das duas *dummies* a partir da variável *perfil*.

```
. tab per
```

periodo do dia	Freq.	Percent	Cum.
tarde	62	62.00	62.00
manhã	38	38.00	100.00
Total	100	100.00	

Figura 13.24 Distribuição de frequências da variável *per*.

```
. tab perfil
```

perfil ao volante	Freq.	Percent	Cum.
calmo	54	54.00	54.00
moderado	33	33.00	87.00
agressivo	13	13.00	100.00
Total	100	100.00	

Figura 13.25 Distribuição de frequências da variável *perfil*.

```
. xi i.perfil
```

i.perfil	_Iperfil 1-3	(naturally coded; _Iperfil 1 omitted)
----------	--------------	---------------------------------------

Figura 13.26 Criação das duas *dummies* a partir da variável *perfil*.

Vamos, então, à modelagem propriamente dita. Para tanto, devemos digitar o seguinte comando:

```
logit atrasado dist sem per _Iperfil_2 _Iperfil_3
```

O comando `logit` elabora uma regressão logística binária estimada por máxima verossimilhança. Se o pesquisador não informar o nível de confiança desejado para a definição dos intervalos dos parâmetros estimados, o

padrão será de 95%. Entretanto, se o pesquisador desejar alterar o nível de confiança dos intervalos dos parâmetros para, por exemplo, 90%, deverá digitar o seguinte comando:

```
logit atrasado dist sem per _Iperfil_2 _Iperfil_3, level(90)
```

Iremos seguir com a análise mantendo o nível padrão de confiança dos intervalos dos parâmetros, que é de 95%. Os *outputs* encontram-se na Figura 13.27 e são exatamente iguais aos calculados na seção 13.2.

Como a regressão logística binária faz parte do grupo de modelos conhecidos por **Modelos Lineares Generalizados** (*Generalized Linear Models*), e como a variável dependente apresenta uma distribuição de Bernoulli, conforme discutido na seção 13.2.1, a estimação apresentada na Figura 13.27 também poderia ter sido igualmente obtida por meio da digitação do seguinte comando:

```
glm atrasado dist sem per _Iperfil_2 _Iperfil_3, family(bernoulli)
```

. logit atrasado dist sem per _Iperfil_2 _Iperfil_3						
Iteration 0:	log likelihood =	-67.685855				
Iteration 1:	log likelihood =	-34.976399				
Iteration 2:	log likelihood =	-30.442925				
Iteration 3:	log likelihood =	-29.076531				
Iteration 4:	log likelihood =	-29.065694				
Iteration 5:	log likelihood =	-29.06568				
Iteration 6:	log likelihood =	-29.06568				
Logistic regression			Number of obs	=	100	
			LR chi2(5)	=	77.24	
			Prob > chi2	=	0.0000	
Log likelihood = -29.06568			Pseudo R2	=	0.5706	
atrasado	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
dist	.2201793	.1097042	2.01	0.045	.0051629	.4351956
sem	2.766715	.9216722	3.00	0.003	.960271	4.57316
per	-3.653351	.8781353	-4.16	0.000	-5.374464	-1.932237
_Iperfil_2	1.346041	.7477467	1.80	0.072	-.1195153	2.811598
_Iperfil_3	2.914474	1.178805	2.47	0.013	.6040581	5.22489
_cons	-30.20028	9.981061	-3.03	0.002	-49.7628	-10.63776
Note: 0 failures and 2 successes completely determined.						

Figura 13.27 Outputs da regressão logística binária no Stata.

Inicialmente, podemos verificar que os valores máximos do logaritmo da função de verossimilhança para o modelo completo e para o modelo nulo são, respectivamente, -29,06565 e -67,68585, e são exatamente aqueles calculados e apresentados nas Figuras 13.4 e 13.7, respectivamente. Assim, fazendo uso da expressão (13.17), temos que:

$$\chi^2_{5g.l.} = -2.[-67,68585 - (-29,06568)] = 77,24 \text{ com valor } -P(\text{ou Prob. } \chi^2_{\text{cal}}) = 0,000.$$

Logo, com base no teste  $\chi^2$ , podemos rejeitar a hipótese nula de que todos os parâmetros  $\beta_j$  ( $j = 1, 2, \dots, 5$ ) sejam estatisticamente iguais a zero ao nível de significância de 5%, ou seja, pelo menos uma variável  $X$  é estatisticamente significativa para explicar a probabilidade de ocorrência do fato de se chegar atrasado à escola.

Embora o pseudo  $R^2$  de McFadden, conforme discutido, apresente bastante limitação em relação à sua interpretação, o Stata o calcula, com base na expressão (13.16), exatamente como fizemos na seção 13.2.2.

$$\text{pseudo } R^2 = \frac{-2.(-67,68585) - [(-2.(-29,06568))]}{-2.(-67,68585)} = 0,5706$$

Por meio da maximização do logaritmo da função de verossimilhança, estimamos os parâmetros do modelo, que são exatamente iguais àqueles apresentados na Figura 13.4. Entretanto, conforme discutimos na seção 13.2.2, a variável *\_Iperfil\_2* (parâmetro  $\beta_4$ ) não se mostrou estatisticamente significativa para aumentar ou diminuir a probabilidade de se chegar atrasado à escola na presença das demais variáveis explicativas, ao nível de significância de 5%, uma vez que  $-1,96 < z_{\beta_4} = 1,80 < 1,96$  e, portanto, o *valor-P* da estatística *z* de Wald apresentou um valor maior do que 0,05.

A não rejeição da hipótese nula para o parâmetro  $\beta_4$ , ao nível de significância de 5%, obriga-nos a estimar o modelo de regressão logística binária por meio do procedimento *Stepwise*. Antes, porém, da elaboração deste procedimento, vamos salvar os resultados do modelo completo. Para tanto, devemos digitar o seguinte comando:

```
lrtest, saving(0)
```

Este comando salva as estimativas dos parâmetros do modelo completo, a fim de que seja possível elaborarmos, adiante, um teste para verificação da adequação do ajuste do modelo completo em comparação com o ajuste do modelo final estimado por meio do procedimento *Stepwise*.

Vamos, então, elaborar o procedimento *Stepwise* propriamente dito, por meio da digitação do seguinte comando, em que é definido o nível de significância do teste *z* de Wald a partir do qual as variáveis explicativas serão excluídas do modelo final.

```
stepwise, pr(0.05): logit atrasado dist sem per _Iperfil_2 _Iperfil_3
```

Os *outputs* do modelo final encontram-se na Figura 13.28.

Analogamente, a estimação apresentada na mesma figura também poderia ter sido obtida por meio do seguinte comando:

```
stepwise, pr(0.05): glm atrasado dist sem per _Iperfil_2 _Iperfil_3,
family(bernoulli)
```

. stepwise, pr(0.05): logit atrasado dist sem per _Iperfil_2 _Iperfil_3						
begin with full model						
p = 0.0718 >= 0.0500 removing _Iperfil_2						
Logistic regression			Number of obs		= 100	
			LR chi2(4)		= 73.77	
			Prob > chi2		= 0.0000	
Log likelihood = -30.800789			Pseudo R2		= 0.5449	
atrasado	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
dist	.2041463	.1011603	2.02	0.044	.0058758	.4024168
sem	2.920114	1.010796	2.89	0.004	.9389897	4.901238
per	-3.776301	.8466794	-4.46	0.000	-5.435762	-2.11684
_Iperfil_3	2.459067	1.139451	2.16	0.031	.2257837	4.692351
_cons	-30.93335	10.63625	-2.91	0.004	-51.78001	-10.08668
Note: 0 failures and 2 successes completely determined.						

Figura 13.28 Outputs da regressão logística binária com procedimento *Stepwise* no Stata.

Antes de analisarmos estes novos *outputs*, vamos elaborar o teste de razão de verossimilhança (*likelihood-ratio test*) que, conforme discutimos na seção 13.2.2, verifica a adequação do ajuste do modelo completo em comparação com o ajuste do modelo final estimado por meio do procedimento *Stepwise*. Para tanto, devemos digitar o seguinte comando:

```
lrtest
```

```
. lrtest
Likelihood-ratio test                LR chi2(1) =      3.47
(Assumption: . nested in LRTEST_0)   Prob > chi2 =    0.0625
```

**Figura 13.29** Teste de razão de verossimilhança para verificação da qualidade do ajuste do modelo final.

cujo resultado encontra-se na Figura 13.29 e é exatamente igual ao calculado manualmente por meio da expressão (13.19).

$$\chi^2_{1,g.l.} = -2.[-30,80079 - (-29,06568)] = 3,47 \text{ com valor } -P \text{ (ou Prob. } \chi^2_{cal}) > 0,05.$$

Por meio da análise do teste de razão de verossimilhança, podemos verificar que a estimação do modelo final com a exclusão da variável *\_Iperfil\_2* não alterou a qualidade do ajuste, ao nível de significância de 5%, fazendo com que o modelo estimado por meio do procedimento *Stepwise* seja preferível em relação ao modelo completo estimado com todas as variáveis explicativas.

Outro teste bastante usual para verificação da qualidade de ajuste do modelo final é o teste de Hosmer-Lemeshow, cujo princípio consiste em dividir a base de dados em 10 partes por meio dos decis das probabilidades estimadas pelo último modelo gerado e, a partir de então, elaborar um teste  $\chi^2$  para verificar se existem diferenças significativas entre as frequências observadas e esperadas do número de observações em cada um dos 10 grupos. Para elaborar este teste no Stata, devemos digitar o seguinte comando:

**estat gof, group(10) table**

em que o termo **gof** refere-se à expressão *goodness-of-fit*, ou seja, qualidade do ajuste.

O *output* deste teste encontra-se na Figura 13.30.

Os resultados apresentados nesta figura mostram os grupos formados pelos decis das probabilidades estimadas e as quantidades observadas e esperadas de observações por grupo, assim como o resultado do teste  $\chi^2$  que, para 8 graus de liberdade, não rejeita a hipótese nula de que as frequências esperadas e observadas sejam iguais, ao nível de significância de 5%. Portanto, o modelo final estimado não apresenta problemas em relação à qualidade do ajuste proposto.

Em relação a este modelo final estimado (Figura 13.28), todas as estatísticas apresentadas, os parâmetros estimados com respectivos intervalos de confiança, os erros-padrão e as estatísticas *z* de Wald são exatamente iguais aos

```
. estat gof, group(10) table
Logistic model for atrasado, goodness-of-fit test

(Table collapsed on quantiles of estimated probabilities)
+-----+
| Group | Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
+-----+-----+-----+-----+-----+-----+
| 1 | 0.0376 | 0 | 0.1 | 10 | 9.9 | 10 |
| 2 | 0.0555 | 0 | 0.5 | 10 | 9.5 | 10 |
| 3 | 0.2815 | 2 | 0.8 | 8 | 9.2 | 10 |
| 4 | 0.6423 | 5 | 5.4 | 5 | 4.6 | 10 |
| 5 | 0.7416 | 6 | 6.8 | 4 | 3.2 | 10 |
+-----+-----+-----+-----+-----+-----+
| 6 | 0.8087 | 9 | 7.8 | 1 | 2.2 | 10 |
| 7 | 0.8850 | 7 | 8.5 | 3 | 1.5 | 10 |
| 8 | 0.9719 | 10 | 9.4 | 0 | 0.6 | 10 |
| 9 | 0.9884 | 10 | 9.8 | 0 | 0.2 | 10 |
| 10 | 1.0000 | 10 | 10.0 | 0 | 0.0 | 10 |
+-----+-----+-----+-----+-----+-----+

number of observations =      100
number of groups      =       10
Hosmer-Lemeshow chi2(8) =       6.34
Prob > chi2           =     0.6091
```

**Figura 13.30** Teste de Hosmer-Lemeshow para verificação da qualidade do ajuste do modelo final.

calculados para o modelo final nas seções 13.2.2 e 13.2.3. Assim, para este modelo, temos que  $LL_{\max} = -30,80079$  e, portanto:

$$pseudo R^2 = \frac{-2.(-67,68585) - [(-2.(-30,80079))]}{-2.(-67,68585)} = 0,5449$$

$$\chi^2_{4g.l.} = -2.[-67,68585 - (-30,80079)] = 73,77 \text{ com valor } -P \text{ (ou Prob. } \chi^2_{cal}) = 0,000.$$

Como a estimação do modelo final foi elaborada por meio do procedimento *Stepwise* com nível de significância de 5%, obviamente todos os valores das estatísticas *z* de Wald são menores do que -1,96 ou maiores do que 1,96 e, portanto, todos os seus valores-*P* são menores do que 0,05.

Desta forma, como base nos *outputs* da Figura 13.28, podemos escrever a expressão final de probabilidade estimada de que um estudante *i* chegue atrasado à escola da seguinte forma:

$$p_i = \frac{1}{1 + e^{-(30,933 + 0,204 \cdot dist_i + 2,920 \cdot sem_i - 3,776 \cdot per_i + 2,459 \cdot \_Iperfil\_3_i)}}$$

e, dessa maneira, podemos retornar à primeira pergunta feita ao final da seção 13.2.2:

**Qual é a probabilidade média estimada de se chegar atrasado à escola ao se deslocar 17 quilômetros e passar por 10 semáforos, tendo feito o trajeto de manhã e sendo considerado agressivo ao volante?**

O comando **mf**x permite que o pesquisador responda esta pergunta diretamente. Assim, devemos digitar o seguinte comando:

**mf**x, at(dist=17 sem=10 per=1 \_Iperfil\_3=1)

Obviamente, o termo **\_Iperfil\_2 = 0** não precisa ser incluído no comando **mf**x, já que a variável **\_Iperfil\_2** não está presente no modelo final. O *output* é apresentado na Figura 13.31, por meio do qual podemos chegar à resposta de 0,603 (60,3%), que é exatamente igual àquela calculada manualmente na seção 13.2.2.

Ainda por meio da Figura 13.28, podemos escrever as expressões dos limites inferior (mínimo) e superior (máximo) da probabilidade estimada de que um estudante *i* chegue atrasado à escola, com 95% de confiança. Assim, teremos:

$$p_{i\min} = \frac{1}{1 + e^{-(51,780 + 0,006 \cdot dist_i + 0,938 \cdot sem_i - 5,436 \cdot per_i + 0,226 \cdot \_Iperfil\_3_i)}}$$

$$p_{i\max} = \frac{1}{1 + e^{-(10,087 + 0,402 \cdot dist_i + 4,901 \cdot sem_i - 2,116 \cdot per_i + 4,692 \cdot \_Iperfil\_3_i)}}$$

Pequenas diferenças na terceira casa decimal em relação aos parâmetros apresentados na seção 13.2.2 devem-se a critérios de arredondamento.

. mfx, at(dist=17 sem=10 per=1 _Iperfil_3=1)							
Marginal effects after logit							
y = Pr(atrasado) (predict)							
= .6037341							
variable	dy/dx	Std. Err.	z	P> z	[ 95% C.I. ]		x
dist	.0488398	.02476	1.97	0.049	.00031	.09737	17
sem	.6986059	.2811	2.49	0.013	.147657	1.24955	10
per*	-.3814532	.21615	-1.76	0.078	-.805109	.042203	1
_Iperf~3*	.4884655	.22979	2.13	0.034	.038084	.938847	1
(*) dy/dx is for discrete change of dummy variable from 0 to 1							

**Figura 13.31** Cálculo da probabilidade estimada para valores das variáveis explicativas – comando **mf**x.

Enquanto o comando **logit** faz com que o Stata apresente os coeficientes dos parâmetros estimados da expressão de probabilidade de ocorrência do evento, o comando **logistic** faz com que o software apresente as chances de ocorrência do evento de interesse ao se alterar em uma unidade a correspondente variável explicativa, mantidas as demais condições constantes. Desta forma, vamos digitar o seguinte comando:

```
logistic atrasado dist sem per _Iperfil_2 _Iperfil_3
```

Os *outputs* são apresentados na Figura 13.32.

```
. logistic atrasado dist sem per _Iperfil_2 _Iperfil_3
```

Logistic regression		Number of obs	=	100
		LR chi2(5)	=	77.24
		Prob > chi2	=	0.0000
Log likelihood = -29.06568		Pseudo R2	=	0.5706

	atrasado	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
dist		1.2463	.1367244	2.01	0.045	1.005176 1.545265
sem		15.9063	14.6604	3.00	0.003	2.612404 96.84966
per		.0259042	.0227474	-4.16	0.000	.0046334 .1448239
_Iperfil_2		3.842186	2.872982	1.80	0.072	.8873505 16.63648
_Iperfil_3		18.43911	21.73612	2.47	0.013	1.829528 185.8407

Note: 0 failures and 2 successes completely determined.

**Figura 13.32** *Outputs* da regressão logística binária no Stata – comando **logistic** para obtenção das *odds ratios*.

A única diferença entre os *outputs* da Figura 13.32 (comando **logistic**) e aqueles apresentados na Figura 13.27 (comando **logit**) é que, agora, o Stata apresenta as *odds ratios* de cada variável explicativa, calculadas com base na expressão (13.3). No mais, podemos perceber que as estatísticas *z* de Wald e os seus respectivos *valores-P* são exatamente os mesmos daqueles apresentados na Figura 13.27 e, desta forma, faz sentido elaborarmos, também para o comando **logistic**, o procedimento *Stepwise*. Assim, vamos digitar o seguinte comando:

```
stepwise, pr(0.05): logistic atrasado dist sem per _Iperfil_2  
_Iperfil_3
```

Os *outputs* encontram-se na Figura 13.33.

Analogamente, os *outputs* desta figura são os mesmos daqueles apresentados na Figura 13.28, à exceção das *odds ratios*.

```
. stepwise, pr(0.05): logistic atrasado dist sem per _Iperfil_2 _Iperfil_3
begin with full model
p = 0.0718 >= 0.0500 removing _Iperfil_2
```

Logistic regression		Number of obs	=	100
		LR chi2(4)	=	73.77
		Prob > chi2	=	0.0000
Log likelihood = -30.800789		Pseudo R2	=	0.5449

	atrasado	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
dist		1.226478	.1240708	2.02	0.044	1.005893 1.495435
sem		18.5434	18.7436	2.89	0.004	2.557396 134.4562
per		.0229073	.0193951	-4.46	0.000	.0043579 .1204115
_Iperfil_3		11.6939	13.32463	2.16	0.031	1.253305 109.1094

Note: 0 failures and 2 successes completely determined.

**Figura 13.33** *Outputs* da regressão logística binária com procedimento *Stepwise* no Stata – comando **logistic** para obtenção das *odds ratios*.

As estimações apresentadas nas Figuras 13.32 e 13.33 também poderiam ter sido obtidas por meio dos seguintes comandos, respectivamente:

```
glm atrasado dist sem per _Iperfil_2 _Iperfil_3, family(bernoulli)
eform
stepwise, pr(0.05): glm atrasado dist sem per _Iperfil_2 _Iperfil_3,
family(bernoulli) eform
```

em que o termo **eform** do comando **glm** equivale ao comando **logistic**.

Sendo assim, podemos retornar às duas últimas perguntas elaboradas ao final da seção 13.2.2:

**Em média, em quanto se altera a chance de se chegar atrasado à escola ao se adotar um percurso 1 quilômetro mais longo, mantidas as demais condições constantes?**

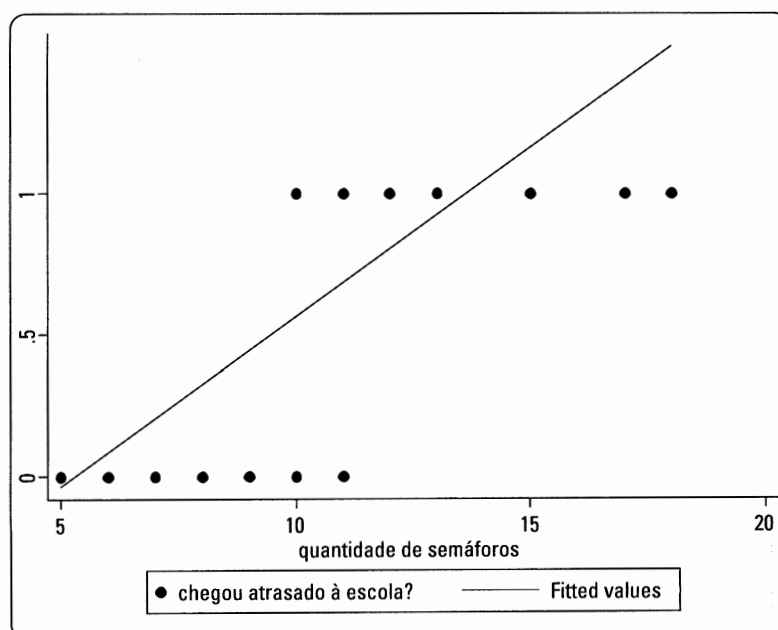
**Um aluno considerado agressivo apresenta, em média, uma chance maior de chegar atrasado do que outro considerado calmo? Se sim, em quanto é incrementada esta chance, mantidas as demais condições constantes?**

As respostas agora podem ser dadas de maneira direta, ou seja, enquanto a chance de se chegar atrasado à escola ao se adotar um trajeto 1 quilômetro mais longo é, em média e mantidas as demais condições constantes, multiplicada por um fator de 1,226 (chance 22,6% maior), a chance de se chegar atrasado à escola quando se é agressivo ao volante em relação a ser calmo é, em média e também mantidas as demais condições constantes, multiplicada por um fator de 11,693 (chance 1.069,3% maior). Estes valores são exatamente os mesmos daqueles calculados manualmente ao final da seção 13.2.2.

Estimado o modelo probabilístico, podemos, por meio do comando **predict phat**, gerar uma nova variável (*phat*) no banco de dados. Esta nova variável corresponde aos valores esperados (previstos) de probabilidade de ocorrência do evento para cada observação, calculados com base nos parâmetros estimados na última modelagem efetuada.

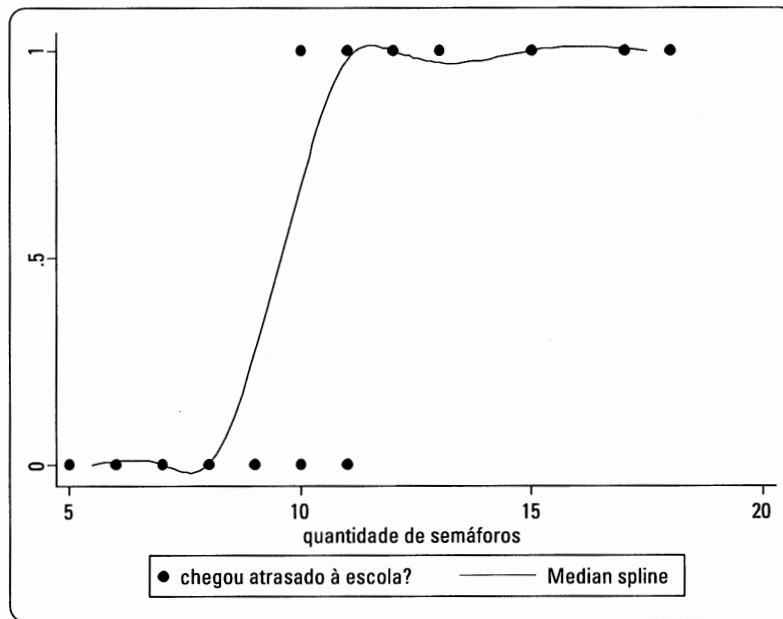
Apenas para fins didáticos, podemos elaborar três gráficos distintos que relacionam a variável dependente e a variável *sem*. Estes gráficos são apresentados nas Figuras 13.34, 13.35 e 13.36, e os comandos para a obtenção de cada um deles são, respectivamente, os seguintes:

```
graph twoway scatter atrasado sem || lfit phat sem
graph twoway scatter atrasado sem || mspline phat sem
graph twoway scatter phat sem || mspline phat sem
```

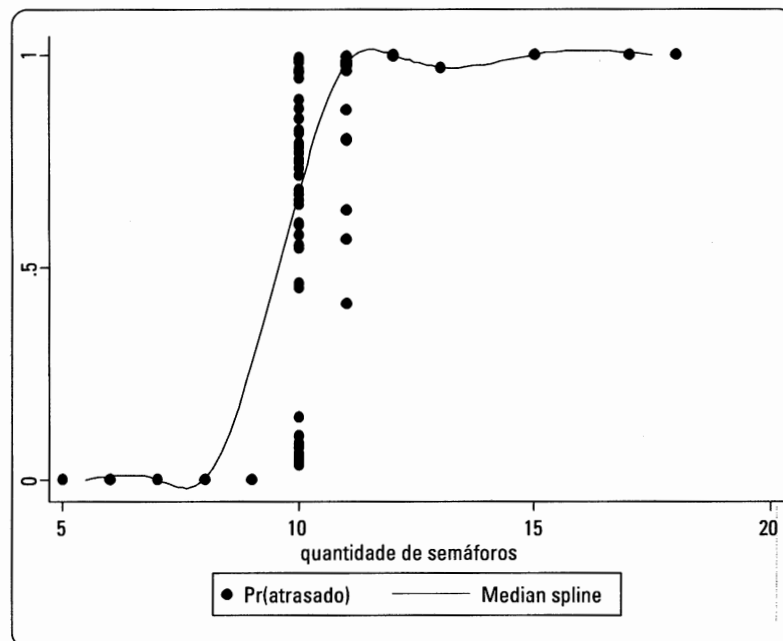


**Figura 13.34** Ajuste linear entre a variável dependente e a variável *sem*.





**Figura 13.35** Ajuste logístico determinístico entre a variável dependente e a variável *sem*.



**Figura 13.36** Ajuste logístico probabilístico entre a variável dependente e a variável *sem*.

Enquanto o gráfico da Figura 13.34 apresenta apenas o ajuste linear entre a variável dependente e a variável *sem*, o que não traz muitos benefícios à análise, o gráfico da Figura 13.35 traz o ajuste logístico com base nas probabilidades estimadas, porém ainda apresentando a variável dependente de forma dicotômica, o que faz com que este gráfico seja chamado de ajuste logístico determinístico. Por fim, o gráfico da Figura 13.36, embora similar ao anterior, mostra como as probabilidades de ocorrência do evento de interesse se comportam em função de alterações na variável *sem*, sendo, portanto, chamado de ajuste logístico probabilístico.

Com base no modelo final estimado, podemos agora elaborar a análise de sensibilidade do modelo proposto, de acordo com o apresentado na seção 13.2.4. Para tanto, devemos digitar o seguinte comando:

**estat class**

```
. estat class
```

Logistic model for atrasado

Classified	True D	~D	Total
+	56	11	67
-	3	30	33
Total	59	41	100

Classified + if predicted Pr(D) >= .5  
True D defined as atrasado != 0

Sensitivity	Pr( +  D)	94.92%
Specificity	Pr( -  ~D)	73.17%
Positive predictive value	Pr( D  +)	83.58%
Negative predictive value	Pr( ~D  -)	90.91%
False + rate for true ~D	Pr( +  ~D)	26.83%
False - rate for true D	Pr( -  D)	5.08%
False + rate for classified +	Pr( ~D  +)	16.42%
False - rate for classified -	Pr( D  -)	9.09%
Correctly classified		86.00%

Figura 13.37 Análise de sensibilidade (*cutoff* = 0,5).

Iniciaremos a análise de sensibilidade com um *cutoff* de 0,5. Ressalta-se que o comando **estat class** já apresenta, como padrão, um *cutoff* de 0,5. O *output* gerado encontra-se na Figura 13.37, que corresponde exatamente à Tabela 13.11.

Logo, conforme discutimos na seção 13.2.4, podemos verificar que 86 observações foram classificadas corretamente, para um *cutoff* de 0,5, sendo que 56 delas foram evento e de fato foram classificadas como tal, e outras 30 não foram evento e não foram classificadas como evento, para este *cutoff*. Entretanto, 14 observações foram classificadas incorretamente, sendo que 3 foram evento mas não foram classificadas como tal e 11 não foram evento mas foram classificadas como tendo sido.

O Stata também oferece em seus *outputs* a eficiência global do modelo, denominada *Correctly Classified* (percentual total de acerto da classificação), a sensibilidade, ou *Sensitivity* (percentual de acerto considerando-se apenas as observações que de fato foram evento) e a especificidade, ou *Specificity* (percentual de acerto considerando-se apenas as observações que não foram evento), para um *cutoff* de 0,5. Assim sendo, temos, respectivamente:

$$EGM = \frac{56 + 30}{100} = 0,8600$$

$$Sensitividade = \frac{56}{59} = 0,9492$$

$$Especificidade = \frac{30}{41} = 0,7317$$

A tabela da Figura 13.37 também pode ser obtida por meio da digitação da seguinte sequência de comandos, cujos *outputs* encontram-se na Figura 13.38:

```
gen classatrasado = 1 if phat>=0.5
replace classatrasado=0 if classatrasado==.
tab classatrasado atrasado
```

```
. gen classatrasado = 1 if phat>=0.5
(33 missing values generated)

. replace classatrasado=0 if classatrasado==.
(33 real changes made)

. tab classatrasado atrasado
```

classatrasado	chegou atrasado à escola?		Total
	Não	Sim	
0	30	3	33
1	11	56	67
Total	41	59	100

Figura 13.38 Obtenção por sequência de comandos da tabela de classificação ( $cutoff = 0,5$ ).

```
. estat class, cutoff(0.3)

Logistic model for atrasado
```

Classified	True		Total
	D	~D	
+	57	13	70
-	2	28	30
Total	59	41	100

```
Classified + if predicted Pr(D) >= .3
True D defined as atrasado != 0
```

Sensitivity	Pr( +  D)	96.61%
Specificity	Pr( -  ~D)	68.29%
Positive predictive value	Pr( D  +)	81.43%
Negative predictive value	Pr( ~D  -)	93.33%
False + rate for true ~D	Pr( +  ~D)	31.71%
False - rate for true D	Pr( -  D)	3.39%
False + rate for classified +	Pr( ~D  +)	18.57%
False - rate for classified -	Pr( D  -)	6.67%
Correctly classified		85.00%

Figura 13.39 Análise de sensibilidade ( $cutoff = 0,3$ ).

As Figuras 13.39 e 13.40 apresentam as análises de sensibilidade do modelo para valores de  $cutoff$  iguais a 0,3 e 0,7, e suas tabelas de classificação correspondem, respectivamente, às Tabelas 13.12 e 13.13 apresentadas na seção 13.2.4. Os comandos para obtenção das Figuras 13.39 e 13.40 são, respectivamente:

```
estat class, cutoff(0.3)
```

```
estat class, cutoff(0.7)
```

Como os valores de  $cutoff$  variam entre 0 e 1, torna-se operacionalmente impossível a elaboração de análises de sensibilidade para cada  $cutoff$ . Sendo assim, faz sentido, neste momento, que sejam elaboradas a curva de sensibilidade e a curva *ROC* (*Receiver Operating Characteristic*) para todas as possibilidades de  $cutoff$ . Os comandos para a elaboração de cada uma delas são, respectivamente:

```
lsens
```

```
lroc
```

```
. estat class, cutoff(0.7)
```

Logistic model for atrasado

Classified	True		Total
	D	~D	
+	47	5	52
-	12	36	48
Total	59	41	100

Classified + if predicted Pr(D) >= .7  
True D defined as atrasado != 0

Sensitivity	Pr( +  D)	79.66%
Specificity	Pr( -  ~D)	87.80%
Positive predictive value	Pr( D  +)	90.38%
Negative predictive value	Pr( ~D  -)	75.00%
False + rate for true ~D	Pr( +  ~D)	12.20%
False - rate for true D	Pr( -  D)	20.34%
False + rate for classified +	Pr( ~D  +)	9.62%
False - rate for classified -	Pr( D  -)	25.00%
Correctly classified		83.00%

Figura 13.40 Análise de sensibilidade (*cutoff* = 0,7).

Enquanto as Figuras 13.14 e 13.15 (seção 13.2.4) apresentavam apenas parte das curvas completas de sensibilidade e ROC (naquela oportunidade, foram plotadas considerando-se apenas três valores de *cutoff*), as Figuras 13.41 e 13.42 apresentam, respectivamente, estas curvas completas.

A análise da curva de sensibilidade (Figura 13.41) permite que cheguemos a um valor aproximado de *cutoff* que iguale a sensibilidade à especificidade, e esse *cutoff*, para o nosso exemplo, é aproximadamente igual a 0,67. O maior problema que podemos perceber na curva de sensibilidade refere-se ao comportamento da curva de especificidade. Enquanto a curva de sensibilidade apresenta percentuais de acerto de classificação para a maioria dos valores de *cutoff* (até aproximadamente 0,65), o mesmo não pode ser dito em relação ao comportamento da curva de especificidade,

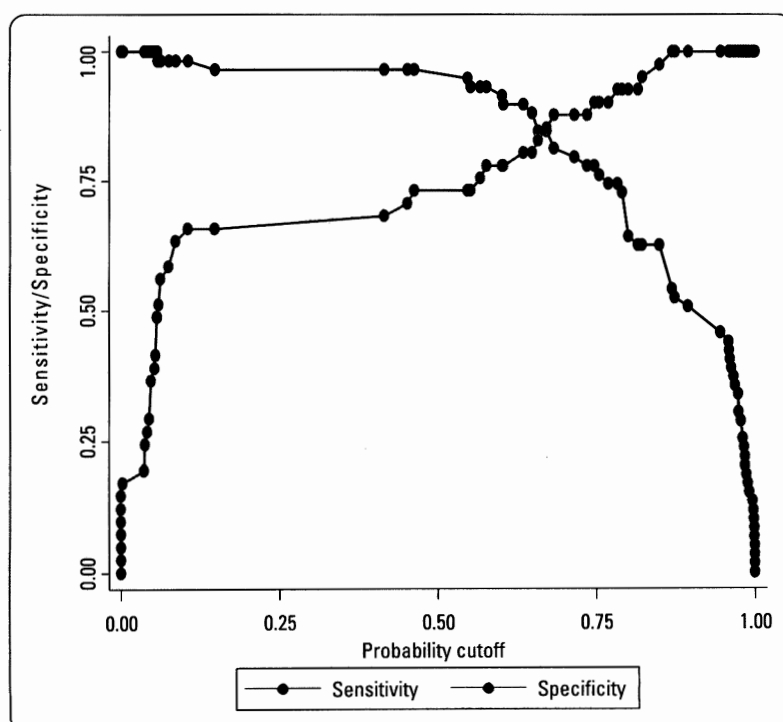
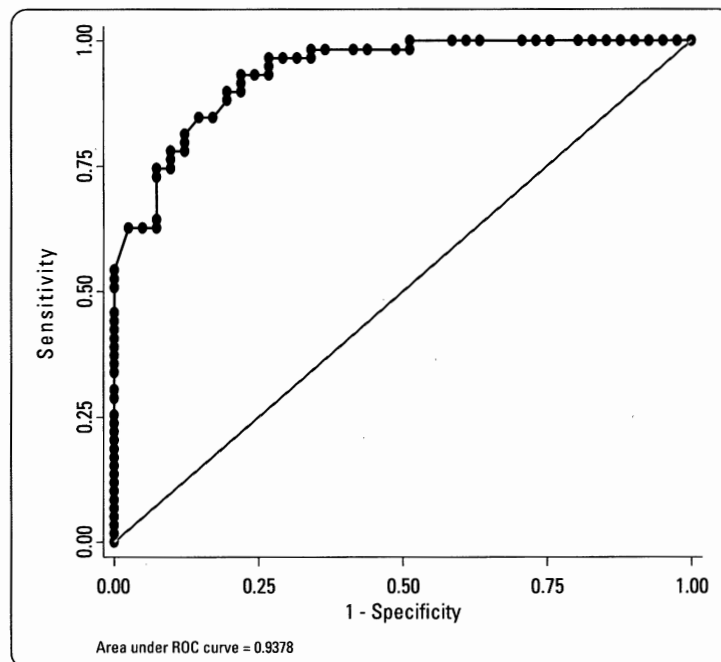


Figura 13.41 Curva de sensibilidade.



**Figura 13.42** Curva ROC.

que apresenta percentuais altos de acerto apenas para um intervalo bem pequeno de *cutoffs* (apenas para *cutoffs* maiores do que aproximadamente 0,75). Em outras palavras, enquanto o percentual de acerto para aqueles que serão evento é alto, quase que independentemente do *cutoff* que se use, o percentual de acerto daqueles que não serão evento só será alto para poucos valores de *cutoff*, o que poderá prejudicar a eficiência global de acerto do modelo para efeitos de previsão. Este modelo, portanto, é bom para prever se um aluno chegará de fato atrasado à escola, porém não apresenta o mesmo desempenho para se prever o não evento, ou seja, caso haja a indicação de que um aluno não chegará atrasado à escola. Quando houver esta última indicação, portanto, o modelo cometerá mais erros de previsão para a maioria dos valores de *cutoff*!

Assim sendo, embora tenhamos um modelo com alta eficiência global e com variáveis explicativas estatisticamente significantes para compor as expressões das probabilidades de ocorrência do evento e do não evento, poderíamos sugerir a inclusão de novas variáveis explicativas a fim de que, eventualmente, melhore o caráter de previsibilidade daqueles que não chegarão atrasados à escola e, desta forma, a eficiência global do modelo, com o consequentemente aumento da área abaixo da curva ROC. Embora isso seja verdade, é importante frisarmos que, para o nosso exemplo, a área abaixo da curva ROC é de 0,9378 (Figura 13.42), o que é considerado muito bom para efeitos de previsão!

### 13.4.2. Regressão logística multinomial no software Stata

O exemplo da seção 13.3 possui, como fenômeno a ser estudado, uma variável qualitativa com três categorias (*não chegou atrasado*, *chegou atrasado à primeira aula* ou *chegou atrasado à segunda aula*). O banco de dados encontra-se no arquivo **AtrasadoMultinomial.dta** e é exatamente igual ao apresentado parcialmente na Tabela 13.14. Seguindo o mesmo procedimento adotado na seção 13.4.1, iremos inicialmente digitar o comando **desc**, a fim de analisarmos as características do banco de dados, como o número de observações, o número de variáveis e a descrição de cada uma delas. A Figura 13.43 apresenta estas características.

Neste exemplo, apenas duas variáveis explicativas foram consideradas (*dist* e *sem*), sendo ambas quantitativas. A Figura 13.44 apresenta a distribuição de frequências das categorias da variável dependente *atrasado*, que foi obtida por meio da digitação do seguinte comando:

```
tab atrasado
```

```
. desc
```

obs:	100
vars:	4
size:	2,700 (99.9% of memory free)

---

variable name	storage type	display format	value label	variable label
estudante	str11	%11s		
atrasado	float	%31.0g	atrasado	chegou atrasado à escola?
dist	float	%9.0g		distância percorrida até a escola (km)
sem	float	%9.0g		quantidade de semáforos

---

Sorted by:

Figura 13.43 Descrição do banco de dados **AtrasadoMultinomial.dta**.

```
. tab atrasado
```

chegou atrasado à escola?	Freq.	Percent	Cum.
não chegou atrasado	49	49.00	49.00
chegou atrasado à primeira aula	16	16.00	65.00
chegou atrasado à segunda aula	35	35.00	100.00
Total	100	100.00	

Figura 13.44 Distribuição de frequências da variável *atrasado*.

Feitas estas considerações iniciais, partiremos para a modelagem propriamente dita da regressão logística multinomial. Para tanto, vamos digitar o seguinte comando:

```
mlogit atrasado dist sem
```

Os *outputs* encontram-se na Figura 13.45.

```
. mlogit atrasado dist sem
```

Iteration 0:	log likelihood =	-101.01922
Iteration 1:	log likelihood =	-42.107305
Iteration 2:	log likelihood =	-37.136795
Iteration 3:	log likelihood =	-28.8332
Iteration 4:	log likelihood =	-25.379085
Iteration 5:	log likelihood =	-24.540694
Iteration 6:	log likelihood =	-24.511848
Iteration 7:	log likelihood =	-24.511801
Iteration 8:	log likelihood =	-24.511801

Multinomial logistic regression	Number of obs	=	100
	LR chi2(4)	=	153.01
	Prob > chi2	=	0.0000
	Pseudo R2	=	0.7574

Log likelihood =	-24.511801
------------------	------------

atrasado	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
não_chegou~o	(base outcome)				
chegou_atr~a					
dist	.558829	.2433023	2.30	0.022	.0819653 1.035693
sem	1.669908	.5768518	2.89	0.004	.5392991 2.800516
_cons	-33.13523	12.18349	-2.72	0.007	-57.01444 -9.256017
chegou_atr~a					
dist	1.078369	.3023893	3.57	0.000	.4856968 1.671041
sem	2.894861	.6859786	4.22	0.000	1.550368 4.239354
_cons	-62.29224	14.67499	-4.24	0.000	-91.05468 -33.52979

Figura 13.45 *Outputs* da regressão logística multinomial no Stata.

Como podemos perceber por meio da análise desta figura, a categoria adotada como referência pelo Stata é a com maior frequência, ou seja, a categoria *não chegou atrasado*, conforme podemos verificar pela Figura 13.44. Coincidentemente, esta é a categoria que realmente desejamos que seja a referência e, portanto, nada precisará ser feito em relação a uma eventual mudança da categoria de referência antes da estimação do modelo. Entretanto, caso um pesquisador tenha o interesse em alterar a categoria de referência para, por exemplo, a categoria *chegou atrasado à segunda aula*, deverá digitar o seguinte comando:

**mlogit atrasado dist sem, b(2)**

Seguiremos com a análise dos *outputs* obtidos na Figura 13.45.

Inicialmente, podemos verificar que os valores máximos do logaritmo da função de verossimilhança para o modelo completo e para o modelo nulo são, respectivamente, -24,51180 e -101,01922, exatamente aqueles calculados e apresentados nas Figuras 13.19 e 13.21, respectivamente. Assim, fazendo uso da expressão (13.41), temos que:

$$\chi^2_{4_{g.l.}} = -2.[-101,01922 - (-24,51180)] = 153,01 \text{ com valor } -P(\text{ou Prob. } \chi^2_{cal}) = 0,000.$$

Logo, com base no teste  $\chi^2$ , podemos rejeitar a hipótese nula de que todos os parâmetros  $\beta_{jm}$  ( $j = 1, 2; m = 1, 2$ ) sejam estatisticamente iguais a zero ao nível de significância de 5%, ou seja, pelo menos uma variável  $X$  é estatisticamente significativa para explicar a probabilidade de ocorrência de pelo menos um dos eventos em estudo.

O Stata também apresenta o pseudo  $R^2$  de McFadden, cujo cálculo é feito com base na expressão (13.40), exatamente como fizemos na seção 13.3.2.

$$\text{pseudo } R^2 = \frac{-2.(-101,01922) - [-2.(-24,51180)]}{-2.(-101,01922)} = 0,7574$$

Como podemos verificar, todas as estatísticas  $z$  de Wald apresentam valores menores do que  $z_c = -1,96$  ou maiores do que  $z_c = 1,96$ , conforme já havíamos discutido na seção 13.3.2. Sendo assim, ainda com base nos *outputs* da Figura 13.45, podemos escrever as expressões finais das probabilidades médias estimadas de ocorrência de cada uma das três categorias da variável dependente, assim como as respectivas expressões dos limites inferior (mínimo) e superior (máximo) destas probabilidades estimadas, com 95% de confiança:

**Probabilidade de um estudante  $i$  não chegar atrasado (categoria 0):**

$$p_{i_0} = \frac{1}{1 + e^{(-33,135 + 0,559 \cdot \text{dist}_i + 1,670 \cdot \text{sem}_i)} + e^{(-62,292 + 1,078 \cdot \text{dist}_i + 2,895 \cdot \text{sem}_i)}}$$

**Intervalo de Confiança (95%) da probabilidade estimada de um estudante  $i$  não chegar atrasado (categoria 0):**

$$p_{i_{0\min}} = \frac{1}{1 + e^{(-57,014 + 0,082 \cdot \text{dist}_i + 0,539 \cdot \text{sem}_i)} + e^{(-91,055 + 0,486 \cdot \text{dist}_i + 1,550 \cdot \text{sem}_i)}}$$

$$p_{i_{0\max}} = \frac{1}{1 + e^{(-9,256 + 1,035 \cdot \text{dist}_i + 2,800 \cdot \text{sem}_i)} + e^{(-33,529 + 1,671 \cdot \text{dist}_i + 4,239 \cdot \text{sem}_i)}}$$

**Probabilidade de um estudante  $i$  chegar atrasado à primeira aula (categoria 1):**

$$p_{i_1} = \frac{e^{(-33,135 + 0,559 \cdot \text{dist}_i + 1,670 \cdot \text{sem}_i)}}{1 + e^{(-33,135 + 0,559 \cdot \text{dist}_i + 1,670 \cdot \text{sem}_i)} + e^{(-62,292 + 1,078 \cdot \text{dist}_i + 2,895 \cdot \text{sem}_i)}}$$

**Intervalo de Confiança (95%) da probabilidade estimada de um estudante  $i$  chegar atrasado à primeira aula (categoria 1):**

$$p_{i_{1\min}} = \frac{e^{(-57,014 + 0,082 \cdot \text{dist}_i + 0,539 \cdot \text{sem}_i)}}{1 + e^{(-57,014 + 0,082 \cdot \text{dist}_i + 0,539 \cdot \text{sem}_i)} + e^{(-91,055 + 0,486 \cdot \text{dist}_i + 1,550 \cdot \text{sem}_i)}}$$

$$p_{i_{\max}} = \frac{e^{(-9,256+1,035 \cdot \text{dist}_i + 2,800 \cdot \text{sem}_i)}}{1 + e^{(-9,256+1,035 \cdot \text{dist}_i + 2,800 \cdot \text{sem}_i)} + e^{(-33,529+1,671 \cdot \text{dist}_i + 4,239 \cdot \text{sem}_i)}}$$

**Probabilidade de um estudante *i* chegar atrasado à segunda aula (categoria 2):**

$$p_{i_2} = \frac{e^{(-62,292+1,078 \cdot \text{dist}_i + 2,895 \cdot \text{sem}_i)}}{1 + e^{(-33,135+0,559 \cdot \text{dist}_i + 1,670 \cdot \text{sem}_i)} + e^{(-62,292+1,078 \cdot \text{dist}_i + 2,895 \cdot \text{sem}_i)}}$$

**Intervalo de Confiança (95%) da probabilidade estimada de um estudante *i* chegar atrasado à segunda aula (categoria 2):**

$$p_{i_{2\min}} = \frac{e^{(-91,055+0,486 \cdot \text{dist}_i + 1,550 \cdot \text{sem}_i)}}{1 + e^{(-57,014+0,082 \cdot \text{dist}_i + 0,539 \cdot \text{sem}_i)} + e^{(-91,055+0,486 \cdot \text{dist}_i + 1,550 \cdot \text{sem}_i)}}$$

$$p_{i_{2\max}} = \frac{e^{(-33,529+1,671 \cdot \text{dist}_i + 4,239 \cdot \text{sem}_i)}}{1 + e^{(-9,256+1,035 \cdot \text{dist}_i + 2,800 \cdot \text{sem}_i)} + e^{(-33,529+1,671 \cdot \text{dist}_i + 4,239 \cdot \text{sem}_i)}}$$

Estimadas as expressões das probabilidades, vamos criar, no banco de dados, três variáveis correspondentes às expressões das probabilidades médias de ocorrência de cada um dos eventos, por meio da digitação dos seguintes comandos:

**Criação da variável referente à probabilidade de um estudante *i* não chegar atrasado (categoria 0):**

```
gen pi0 = (1) / (1 + (exp(-33.13523 + .558829*dist + 1.669908*sem))
+ (exp(-62.29224 + 1.078369*dist + 2.894861*sem)))
```

**Criação da variável referente à probabilidade de um estudante *i* chegar atrasado à primeira aula (categoria 1):**

```
gen pi1 = (exp(-33.13523 + .558829*dist + 1.669908*sem)) / (1
+ (exp(-33.13523 + .558829*dist + 1.669908*sem)) + (exp(-62.29224
+ 1.078369*dist + 2.894861*sem)))
```

**Criação da variável referente à probabilidade de um estudante *i* chegar atrasado à segunda aula (categoria 2):**

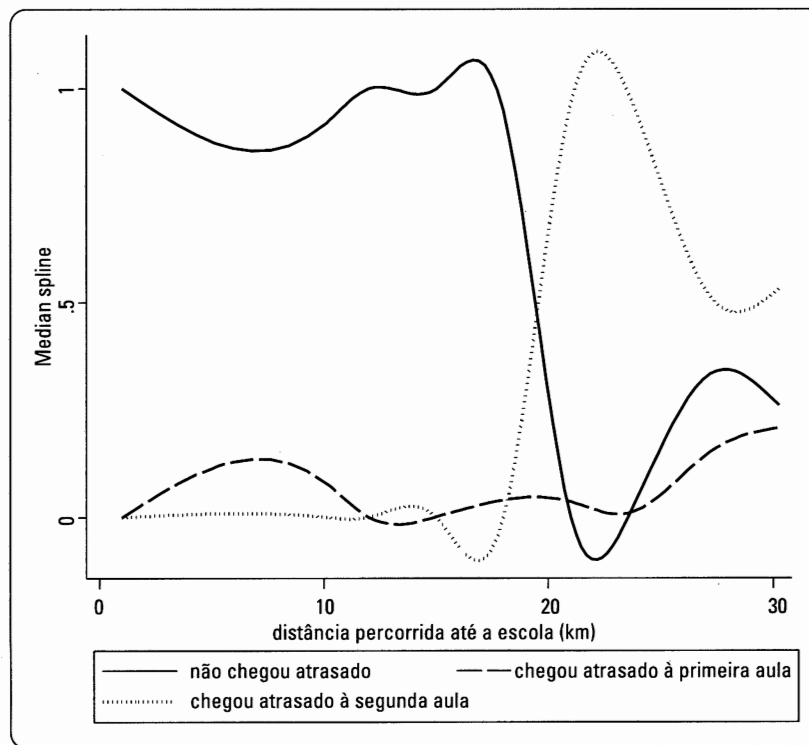
```
gen pi2 = (exp(-62.29224 + 1.078369*dist + 2.894861*sem)) / (1
+ (exp(-33.13523 + .558829*dist + 1.669908*sem)) + (exp(-62.29224
+ 1.078369*dist + 2.894861*sem)))
```

Podemos verificar que estas novas variáveis (*pi0*, *pi1* e *pi2*) são idênticas àquelas obtidas quando da elaboração da Figura 13.19 obtida pelo **Solver** do Excel (naquele caso, as variáveis presentes nas colunas J, K e L, respectivamente). Geradas estas novas variáveis, teremos condições de elaborar dois interessantes gráficos, a partir dos quais algumas conclusões podem ser obtidas. Enquanto o primeiro gráfico (Figura 13.46) mostra o comportamento das probabilidades de ocorrência de cada um dos eventos em função da distância percorrida até a escola, o segundo gráfico (Figura 13.47) mostra o comportamento destas probabilidades em função da quantidade de semáforos pelos quais cada um é obrigado a passar. Os comandos para elaboração destes gráficos são, respectivamente:

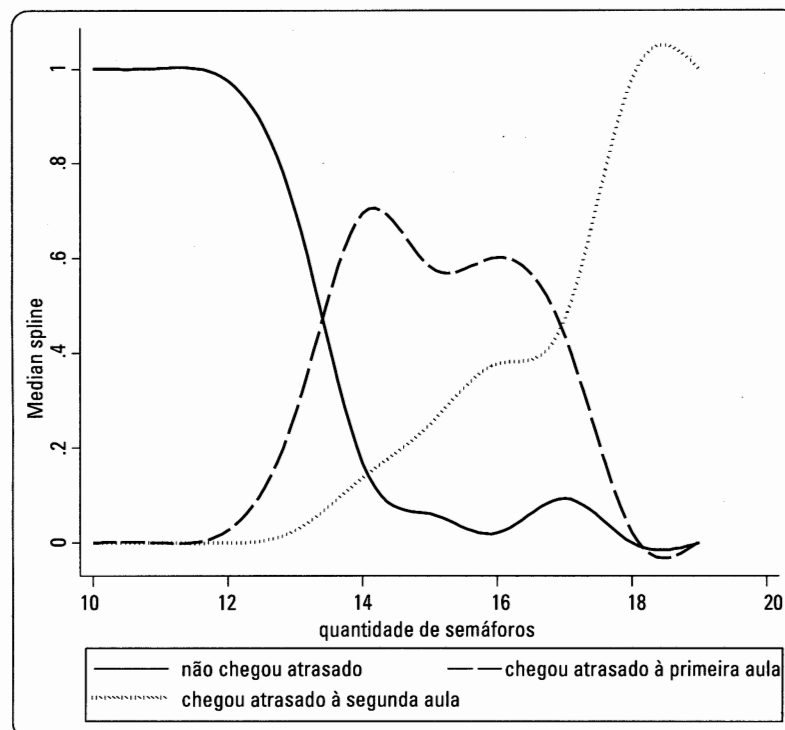
```
graph twoway mspline pi0 dist || mspline pi1 dist || mspline pi2
dist ||, legend(label(1 "não chegou atrasado") label(2 "chegou atrasado
à primeira aula") label(3 "chegou atrasado à segunda aula"))
```

```
graph twoway mspline pi0 sem || mspline pi1 sem || mspline pi2 sem ||,
legend(label(1 "não chegou atrasado") label(2 "chegou atrasado à
primeira aula") label(3 "chegou atrasado à segunda aula"))
```





**Figura 13.46** Probabilidades de ocorrência de cada evento x distância percorrida.



**Figura 13.47** Probabilidades de ocorrência de cada evento x quantidade de semáforos.

Por meio do gráfico da Figura 13.46, podemos verificar que há diferenças nas probabilidades de se chegar atrasado à primeira ou à segunda aula em relação a não se chegar atrasado, ao se variar a distância percorrida até a escola. Podemos perceber que, até aproximadamente 20 quilômetro de distância, as diferenças nas probabilidades de se chegar atrasado à primeira ou à segunda aula são pequenas, porém as maiores diferenças ocorrem para a probabilidade de não se chegar atrasado, que é bem maior. Por outro lado, uma distância maior que aproximadamente 20 quilômetros de percurso passa a fazer com que a probabilidade de se chegar atrasado à segunda aula aumente consideravelmente em relação à probabilidade de se chegar atrasado à primeira aula. Além disso, a partir desta distância, a probabilidade de não se chegar atrasado à escola cai consideravelmente. Isso explica o fato de a variável *dist* ter sido estatisticamente significativa, ao nível de significância de 5%, para os dois logitos do modelo, tendo sido considerada referência a categoria correspondente a não se chegar atrasado. Podemos também notar, independentemente da distância percorrida, que a probabilidade de se chegar atrasado à primeira aula é sempre baixa, e quase não apresenta alterações consideráveis com a mudança da distância. Desta forma, se, por exemplo, elaborássemos uma regressão logística com apenas duas categorias (binária), sendo o evento de interesse representado pela categoria correspondente a se chegar atrasado à primeira aula (*dummy* = 1), verificaríamos que a variável *dist* não seria estatisticamente significativa, ao nível de significância de 5%, para explicar a probabilidade de se chegar atrasado à primeira aula, como já comprovado por meio da análise do gráfico da Figura 13.46.

Já a análise da Figura 13.47, que mostra as diferenças nas probabilidades de se chegar atrasado à primeira ou à segunda aula em relação a não se chegar atrasado, ao se variar a quantidade de semáforos que são ultrapassados no trajeto até a escola, podemos verificar que, até uma quantidade de aproximadamente 12 semáforos, a probabilidade de se chegar atrasado à escola é praticamente nula. Porém, a partir desta quantidade, a probabilidade de se chegar atrasado passa a subir consideravelmente, com destaque para a probabilidade de se chegar atrasado à primeira aula. Entretanto, para quantidades superiores a aproximadamente 17 semáforos, a probabilidade de se chegar atrasado à segunda aula passa a ser a maior entre as três possibilidades de ocorrência de evento, ficando quase que absoluta com quantidades superiores a 18 semáforos. O comportamento destas probabilidades explica o fato de a variável *sem* ter sido estatisticamente significativa, ao nível de significância de 5%, para os dois logitos do modelo, tendo sido considerada referência a categoria correspondente a não se chegar atrasado, ou seja, para explicar o comportamento das probabilidades de ocorrência de cada uma das três categorias da variável dependente.

Por fim, mas não menos importante, vamos elaborar, assim como fizemos na seção 13.4.1, o modelo solicitando que sejam fornecidas as chances de ocorrência de cada um dos eventos de interesse ao se alterar em uma unidade a correspondente variável explicativa, mantidas as demais condições constantes. Em modelos de regressão logística multinomial, conforme discutimos na seção 13.3.2, a chance (*odds ratio*) também é chamada de razão de risco relativo (*relative risk ratio*). Desta forma, devemos digitar o seguinte comando:

```
mlogit atrasado dist sem, rrr
```

em que o termo **rrr** refere-se exatamente à expressão *relative risk ratio*. Os *outputs* estão apresentados na Figura 13.48.

Os *outputs* da Figura 13.48 são os mesmos daqueles apresentados na Figura 13.45, à exceção das *relative risk ratios*. Desta forma, podemos retornar às duas últimas perguntas elaboradas ao final da seção 13.3.2:

**Em média, em quanto se altera a chance de se chegar atrasado à primeira aula, em relação a não chegar atrasado à escola, ao se adotar um percurso 1 quilômetro mais longo, mantidas as demais condições constantes?**

**Em média, em quanto se altera a chance de se chegar atrasado à segunda aula, em relação a não chegar atrasado, ao se passar por 1 semáforo a mais no percurso até a escola, mantidas as demais condições constantes?**

As respostas agora podem ser dadas de maneira direta, ou seja, enquanto a chance de se chegar atrasado à primeira aula em relação a não chegar atrasado à escola, ao se adotar um trajeto 1 quilômetro mais longo, é, em média e mantidas as demais condições constantes, multiplicada por um fator de 1,749 (74,9% maior), a chance de se chegar atrasado à segunda aula em relação a não chegar atrasado, ao se passar por 1 semáforo a mais no percurso até a escola, é, em média, multiplicada por um fator de 18,081 (1.708,1% maior), também mantidas as demais condições constantes. Estes valores são exatamente os mesmos daqueles calculados manualmente ao final da seção 13.3.2.

```
. mlogit atrasado dist sem, rrr
```

Iteration 0: log likelihood = -101.01922  
Iteration 1: log likelihood = -42.107305  
Iteration 2: log likelihood = -37.136795  
Iteration 3: log likelihood = -28.8332  
Iteration 4: log likelihood = -25.379085  
Iteration 5: log likelihood = -24.540694  
Iteration 6: log likelihood = -24.511848  
Iteration 7: log likelihood = -24.511801  
Iteration 8: log likelihood = -24.511801

Multinomial logistic regression

Log likelihood = -24.511801

Number of obs	=	100
LR chi2(4)	=	153.01
Prob > chi2	=	0.0000
Pseudo R2	=	0.7574

atrasado	RRR	Std. Err.	z	P> z	[95% Conf. Interval]
não_chegou~o   (base outcome)					
-----+-----					
chegou_atr~a					
dist	1.748624	.4254441	2.30	0.022	1.085418 2.817057
sem	5.311678	3.064051	2.89	0.004	1.714804 16.45314
-----+-----					
chegou_atr~a					
dist	2.93988	.8889883	3.57	0.000	1.625307 5.3177
sem	18.08099	12.40317	4.22	0.000	4.713203 69.36305
-----+-----					

Figura 13.48 Outputs da regressão logística multinomial no Stata – *relative risk ratios*.

A capacidade do Stata para a estimação de modelos e a elaboração de testes estatísticos é enorme, porém acreditamos que o que foi exposto aqui é considerado obrigatório para pesquisadores que tenham a intenção de aplicar, de forma correta, as técnicas de regressão logística binária e multinomial.

Partiremos agora para a resolução dos mesmos exemplos por meio do SPSS.

### 13.5. ESTIMAÇÃO DE MODELOS DE REGRESSÃO LOGÍSTICA BINÁRIA E MULTINOMIAL NO SOFTWARE SPSS

Apresentaremos agora o passo a passo para a elaboração dos nossos exemplos por meio do IBM SPSS Statistics Software®. A reprodução de suas imagens nesta seção tem autorização da International Business Machines Corporation®.

Nosso objetivo não é discutir novamente os conceitos inerentes às técnicas, nem tampouco repetir aquilo que já foi explorado nas seções anteriores. O maior objetivo desta seção é o de propiciar ao pesquisador uma oportunidade de elaborar as técnicas de regressão logística binária e multinomial no SPSS, dada a facilidade de manuseio e a didática com que o software realiza as suas operações e se coloca perante o usuário. A cada apresentação de um *output*, faremos menção ao respectivo resultado obtido quando da elaboração das técnicas por meio do Excel e do Stata, a fim de que o pesquisador possa compará-los e, desta forma, decidir qual software utilizar, em função das características de cada um e da própria acessibilidade para uso.

#### 13.5.1. Regressão logística binária no software SPSS

Seguindo a mesma lógica proposta quando da aplicação dos modelos por meio do software Stata, já partiremos para o banco de dados construído pelo professor a partir dos questionamentos feitos a cada um de seus 100 estudantes. Os dados encontram-se no arquivo **Atrasado.sav** e, após o abrí-los, vamos inicialmente clicar em **Analyze → Regression → Binary Logistic...**. A caixa de diálogo da Figura 13.49 será aberta.

Devemos selecionar a variável *atrasado* e incluí-la na caixa **Dependent**. As demais variáveis devem ser simultaneamente selecionadas e inseridas na caixa **Covariates**. Manteremos, neste primeiro momento, a opção pelo **Method: Enter**. O procedimento *Enter*, ao contrário do procedimento *Stepwise* (no SPSS, a regressão logística binária utiliza procedimento análogo conhecido por *Forward Wald*), inclui todas as variáveis na estimação, mesmo aquelas cujos parâmetros sejam estatisticamente iguais a zero, e corresponde exatamente ao procedimento padrão elaborado pelo Excel (modelo completo apresentado na Figura 13.4) e também pelo Stata quando se aplica

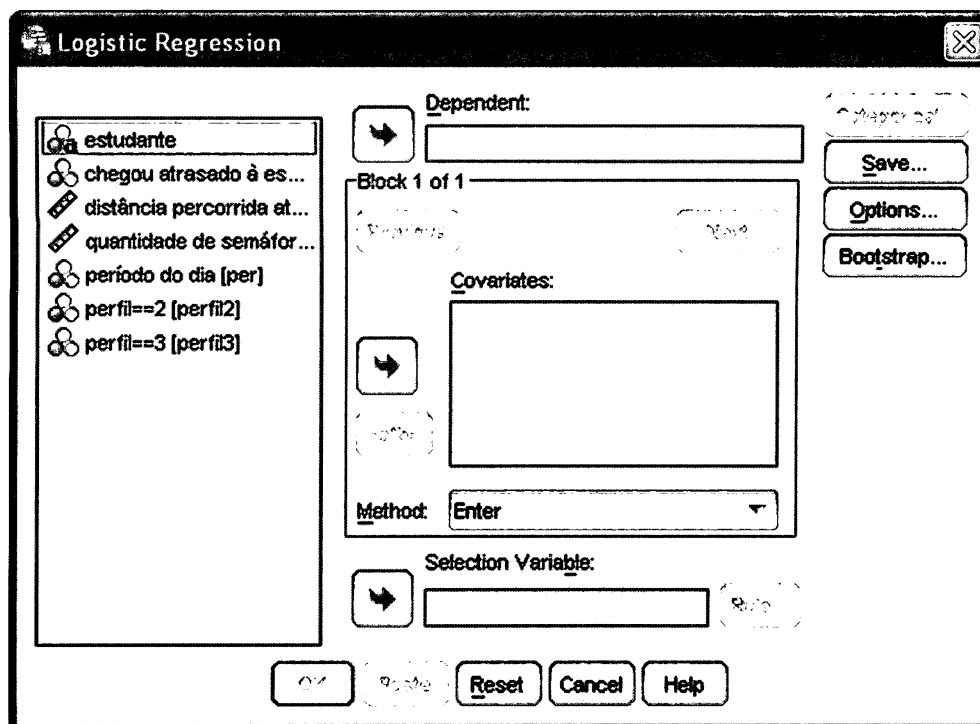


Figura 13.49 Caixa de diálogo para elaboração da regressão logística binária no SPSS.

diretamente o comando `logit`. A Figura 13.50 apresenta a caixa de diálogo do SPSS, com a definição da variável dependente e das variáveis explicativas a serem inseridas no modelo.

Caso o banco de dados não tivesse apresentado as variáveis *dummy* correspondentes às categorias da variável *perfil*, poderíamos selecionar o botão **Categorical...** e incluir a variável original (*perfil*) nesta opção, inclusive

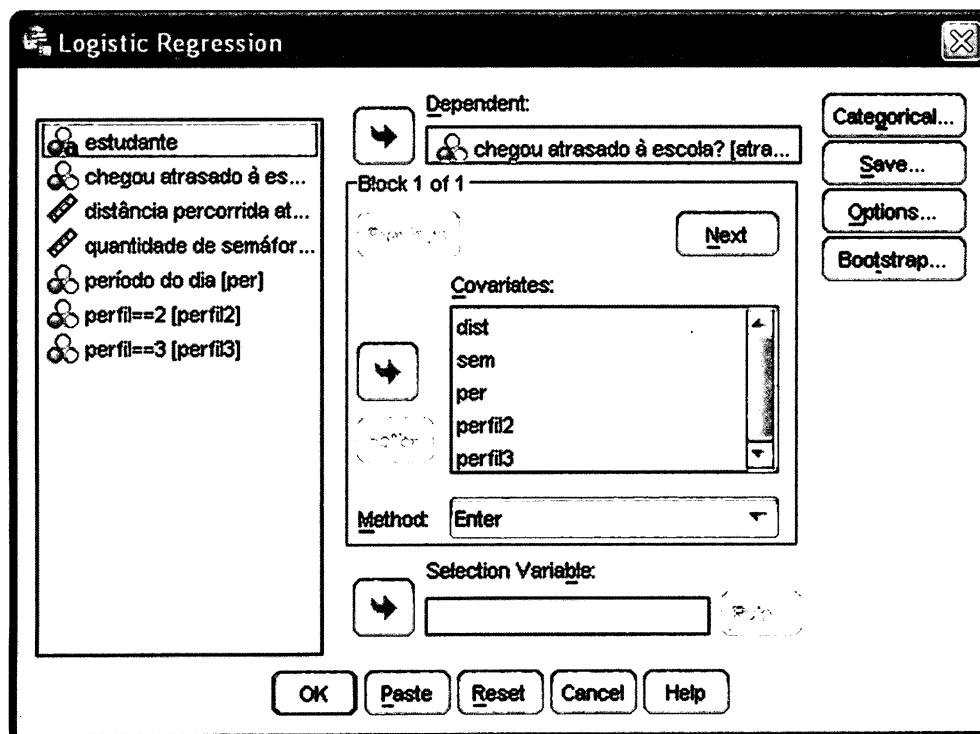


Figura 13.50 Caixa de diálogo para elaboração da regressão logística binária no SPSS com inclusão da variável dependente e das variáveis explicativas e seleção do procedimento *Enter*.

com a definição da categoria de referência. Como já temos as duas *dummies* (*perfil2* e *perfil3*), não há a necessidade de que este procedimento seja feito.

No botão **Options...**, selecionaremos apenas as opções **Iteration history** e **CI for exp(B)**, que correspondem, respectivamente, ao histórico do procedimento de iteração para a maximização da somatória do logaritmo da função de verossimilhança e aos intervalos de confiança das *odds ratios* de cada parâmetro. A caixa de diálogo que é aberta, ao clicarmos nesta opção, está apresentada na Figura 13.51, já com a seleção das mencionadas opções.

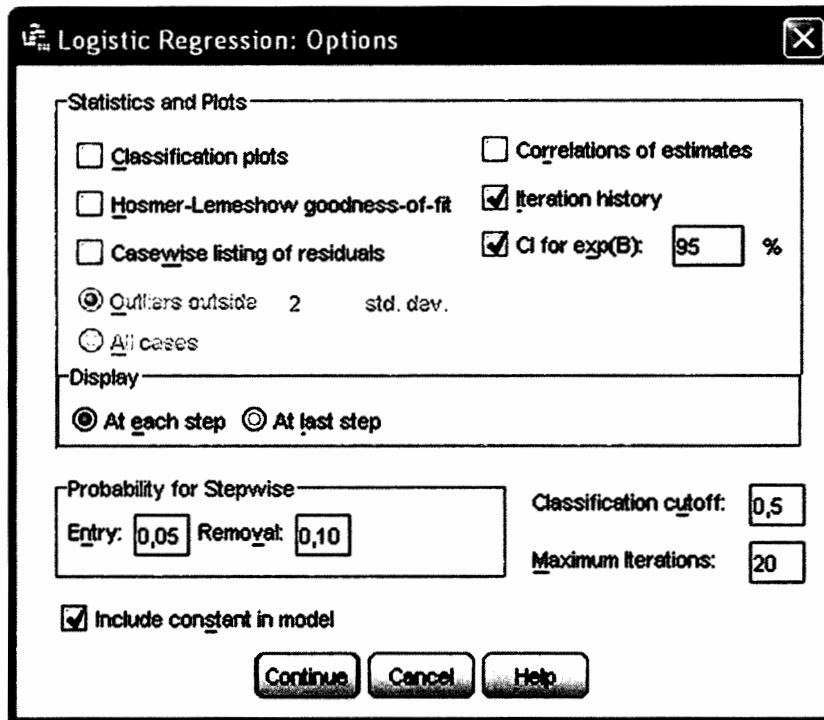


Figura 13.51 Opções para elaboração da regressão logística binária no SPSS.

Podemos notar, por meio da Figura 13.51, que o *cutoff* padrão utilizado pelo SPSS é igual a 0,5, porém é nesta caixa de diálogo que o pesquisador pode alterá-lo para o valor que desejar, a fim de elaborar classificações das observações existentes na base de dados e previsões para outras observações. Na caixa de diálogo do botão **Options...**, podemos ainda impor que o parâmetro  $\alpha$  seja igual a zero (ao desabilitarmos a opção **Include constant in equation**) e alterar o nível de significância a partir do qual o parâmetro de determinada variável explicativa pode ser considerado estatisticamente igual a zero (teste  $z$  de Wald) e, portanto, esta variável deverá ser excluída do modelo final quando da elaboração do procedimento *Stepwise*. Manteremos o padrão de 5% para os níveis de significância e deixaremos a constante no modelo (opção **Include constant in equation** selecionada).

Vamos agora selecionar **Continue** e **OK**. Os *outputs* gerados estão apresentados na Figura 13.52.

Esta figura traz apenas os resultados obtidos mais importantes para a análise da regressão logística binária. Não iremos novamente analisar todos os *outputs* gerados, uma vez que podemos verificar que são exatamente iguais àqueles obtidos quando da estimação da regressão logística binária no Excel e no Stata. Vale a pena comentar que, enquanto o Stata apresenta o cálculo do valor máximo obtido da somatória do logaritmo da função de verossimilhança, o SPSS apresenta o dobro deste valor, e com sinal invertido. Assim, enquanto obtivemos *LL* de -67,68585 para o modelo nulo (conforme pode ser verificado pelas Figuras 13.7 e 13.27) e de -29,06568 para o modelo completo (Figuras 13.4 e 13.27), o SPSS apresenta um valor de  $-2LL$  igual a 135,372 para o modelo nulo (*initial*) e igual a  $-2LL$  igual a 58,131 para o modelo completo.

A outra diferença entre os *outputs* gerados pelo Stata e pelo SPSS diz respeito ao pseudo  $R^2$ . Enquanto o Stata apresenta o já calculado pseudo  $R^2$  de McFadden, o SPSS apresenta o pseudo  $R^2$  de Cox & Snell e o pseudo  $R^2$  de Nagelkerke, cujos cálculos podem ser obtidos, respectivamente, por meio das expressões (13.45) e (13.46).

## Block 1: Method = Enter

Iteration History<sup>a,b,c,d</sup>

Iteration		-2 Log likelihood	Coefficients				
			Constant	dist	sem	per	perfil2
Step 1	1	75,870	-3,561	,059	,339	-2,094	,764
	2	65,970	-8,640	,100	,799	-2,696	1,116
	3	60,185	-17,902	,148	1,647	-3,028	1,249
	4	58,287	-26,614	,204	2,432	-3,439	1,326
	5	58,133	-29,795	,219	2,727	-3,630	1,347
	6	58,131	-30,193	,220	2,766	-3,653	1,346
	7	58,131	-30,200	,220	2,767	-3,653	1,346
	8	58,131	-30,200	,220	2,767	-3,653	1,346

a. Method: Enter

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 135,372

d. Estimation terminated at iteration number 8 because parameter estimates changed by less than ,001.

## Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	77,240	5	,000
	Block	77,240	5	,000
	Model	77,240	5	,000

## Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	58,131 <sup>a</sup>	,538	,725

a. Estimation terminated at iteration number 8 because parameter estimates changed by less than ,001.

Classification Table<sup>a</sup>

		Predicted		
		chegou atrasado à escola?		Percentage Correct
Observed		Não	Sim	
Step 1	chegou atrasado à escola?			
	Não	31	10	75,6
	Sim	4	55	93,2
Overall Percentage				86,0

a. The cut value is ,500

## Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
								Lower	Upper
Step 1 <sup>a</sup>	dist	,220	,110	4,028	1	,045	1,246	1,005	1,545
	sem	2,767	,922	9,011	1	,003	15,906	2,612	96,850
	per	-3,653	,878	17,309	1	,000	,026	,005	,145
	perfil2	1,346	,748	3,240	1	,072	3,842	,887	16,636
	perfil3	2,914	1,179	6,113	1	,013	18,439	1,830	185,841
	Constant	-30,200	9,981	9,155	1	,002	,000		

a. Variable(s) entered on step 1: dist, sem, per, perfil2, perfil3.

Figura 13.52 Outputs da regressão logística binária no SPSS – procedimento Enter.

$$pseudo R^2_{Cox \& Snell} = 1 - \left( \frac{e^{LL_0}}{e^{LL}} \right)^{\frac{2}{N}} \quad (13.45)$$

$$pseudo R^2_{Nagelkerke} = \frac{1 - \left( \frac{e^{LL_0}}{e^{LL}} \right)^{\frac{2}{N}}}{1 - \left( e^{LL_0} \right)^{\frac{2}{N}}} = \frac{pseudo R^2_{Cox \& Snell}}{1 - \left( e^{LL_0} \right)^{\frac{2}{N}}} \quad (13.46)$$

Portanto, para o nosso exemplo, temos que:

$$pseudo R^2_{Cox \& Snell} = 1 - \left( \frac{e^{LL_0}}{e^{LL}} \right)^{\frac{2}{N}} = 1 - \left( \frac{e^{-67,68585}}{e^{-29,06568}} \right)^{\frac{2}{100}} = 0,538$$

$$pseudo R^2_{Nagelkerke} = \frac{pseudo R^2_{Cox \& Snell}}{1 - \left( e^{LL_0} \right)^{\frac{2}{N}}} = \frac{0,538}{1 - \left( e^{-67,68585} \right)^{\frac{2}{100}}} = 0,725$$

Analogamente ao pseudo  $R^2$  de McFadden, estas duas novas estatísticas apresentam limitações para a análise do poder preditivo do modelo e, portanto, recomenda-se, conforme já discutido, que seja elaborada a análise de sensibilidade para esta finalidade.

Os demais resultados são iguais aos obtidos manualmente pelo Excel (seção 13.2) e pelo Stata (seção 13.4). Entretanto, como o parâmetro da variável *perfil2* não se mostrou estatisticamente diferente de zero, ao nível de significância de 5%, partiremos para a estimação do modelo final por meio do procedimento *Forward Wald* (*Stepwise*). Para elaborarmos este procedimento, devemos selecionar a opção **Method: Forward: Wald** na caixa de diálogo principal da regressão logística binária no SPSS, conforme mostra a Figura 13.53.

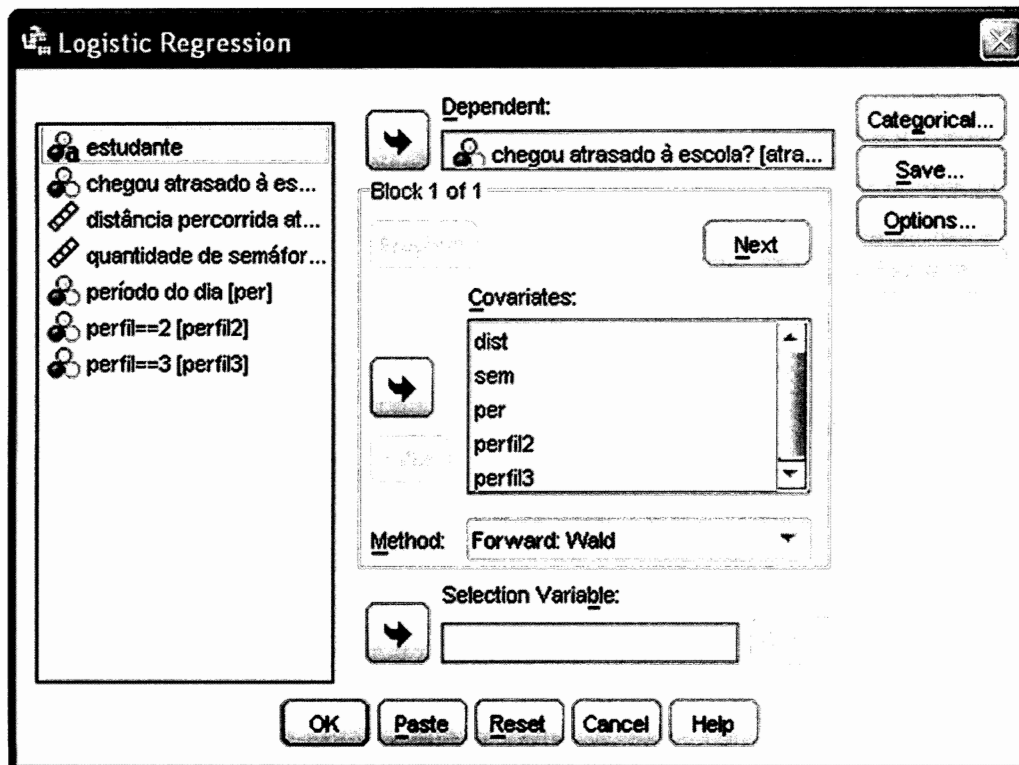
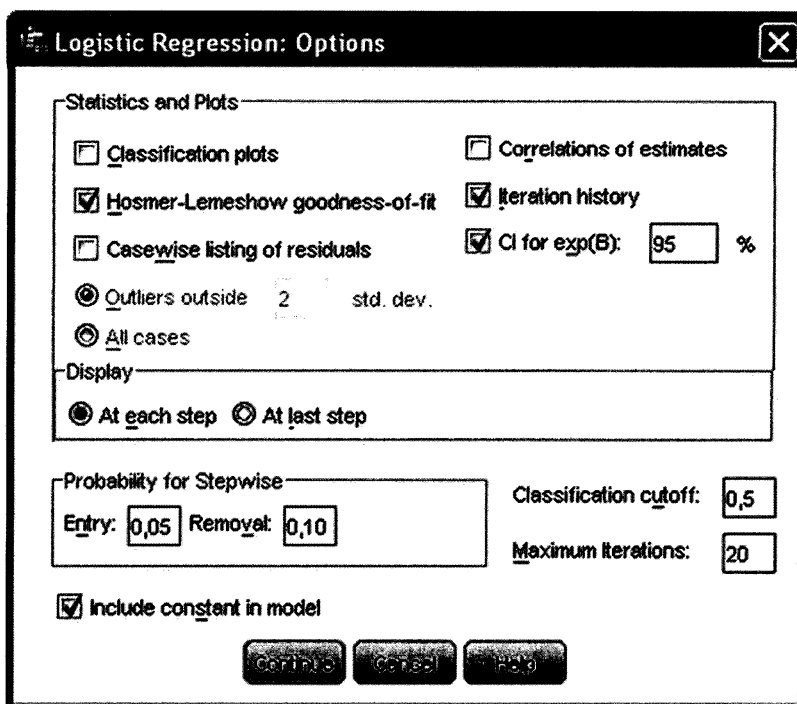


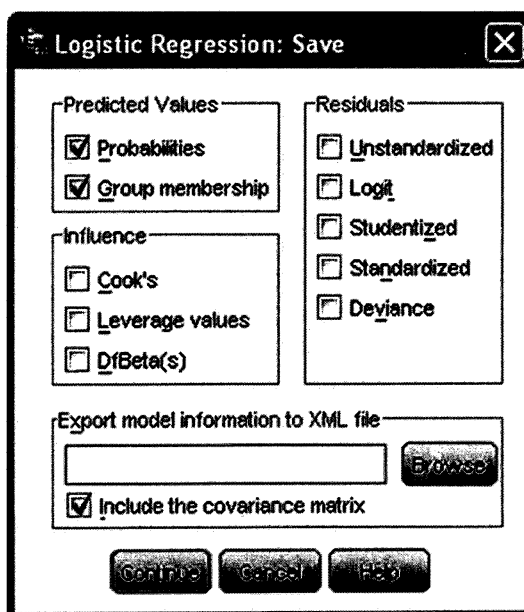
Figura 13.53 Caixa de diálogo com seleção do procedimento *Forward Wald*.

No botão **Options...**, além das opções já marcadas anteriormente, selecionaremos agora também a opção **Hosmer-Lemeshow goodness-of-fit**, conforme mostra a Figura 13.54. Feito isso, devemos clicar em **Continue**.



**Figura 13.54** Seleção do teste de Hosmer-Lemeshow para verificação da qualidade do ajuste do modelo final.

O botão **Save...**, por fim, permite que sejam geradas, no próprio banco de dados original, as variáveis referentes à probabilidade estimada de ocorrência do evento e a classificação de cada observação, com base na sua probabilidade estimada e no *cutoff* definido anteriormente. Dessa forma, ao clicarmos nesta opção, será aberta uma caixa de diálogo, conforme mostra a Figura 13.55. Devemos marcar as opções **Probabilities** e **Group membership** (em **Predicted Values**).



**Figura 13.55** Caixa de diálogo para criação das variáveis referentes à probabilidade estimada de ocorrência do evento e a classificação de cada observação.



Ao clicarmos em **Continue** e, na sequência, em **OK**, novos *outputs* são gerados, conforme mostra a Figura 13.56. Note que, além dos *outputs*, são criadas duas novas variáveis no banco de dados original, chamadas de *PRE\_1* e *PGR\_1*, que correspondem, respectivamente, às probabilidades estimadas de ocorrência do evento e às respectivas classificações, com base no *cutoff* de 0,5. Note que a variável *PRE\_1* é exatamente igual àquela apresentada na coluna *p<sub>i</sub>* da Figura 13.12 gerada pelo Excel e à variável *phat* gerada pelo Stata após a estimação do modelo apresentado na Figura 13.28.

O primeiro *output* gerado (**Iteration History**) apresenta os valores correspondentes à função de verossimilhança em cada passo da modelagem elaborada por meio do procedimento *Forward Wald*, que equivale ao procedimento *Stepwise*. Verificamos que o valor final de  $-2LL$  é igual a 61,602, ou seja,  $LL = -30,801$ , que é exatamente igual ao valor obtido quando da modelagem no Excel (Figura 13.12) e no Stata (Figura 13.28). O *output* **Model Summary** também apresenta esta estatística, baseada na qual é possível calcular a estatística  $\chi^2$ , cujo teste avalia a

**Block 1: Method = Forward Stepwise (Wald)**

**Iteration History<sup>a,b,c,d,e</sup>**

Iteration		-2 Log likelihood	Coefficients				
			Constant	per	sem	perfil3	dist
Step 1	1	92,166	1,355	-2,618			
	2	91,097	1,623	-3,097			
	3	91,090	1,648	-3,136			
	4	91,090	1,649	-3,137			
Step 2	1	84,812	-1,771	-2,379	,297		
	2	77,467	-5,995	-2,848	,744		
	3	74,614	-11,204	-3,041	1,266		
	4	73,486	-16,979	-3,143	1,839		
	5	73,329	-20,096	-3,212	2,150		
	6	73,327	-20,519	-3,223	2,192		
	7	73,327	-20,525	-3,223	2,193		
	8	73,327	-20,525	-3,223	2,193		
Step 3	1	81,283	-1,934	-2,338	,299	,976	
	2	72,501	-6,132	-2,920	,739	1,722	
	3	68,633	-12,193	-3,243	1,346	2,166	
	4	66,804	-19,909	-3,475	2,110	2,453	
	5	66,438	-25,179	-3,658	2,636	2,626	
	6	66,428	-26,190	-3,707	2,738	2,668	
	7	66,428	-26,217	-3,709	2,740	2,670	
	8	66,428	-26,217	-3,709	2,740	2,670	
Step 4	1	79,252	-3,180	-2,256	,335	,992	,061
	2	69,542	-8,421	-2,829	,821	1,607	,102
	3	63,854	-17,425	-3,165	1,651	1,957	,150
	4	61,832	-26,316	-3,557	2,471	2,274	,195
	5	61,607	-30,211	-3,746	2,848	2,430	,204
	6	61,602	-30,913	-3,775	2,918	2,458	,204
	7	61,602	-30,933	-3,776	2,920	2,459	,204
	8	61,602	-30,933	-3,776	2,920	2,459	,204

a. Method: Forward Stepwise (Wald)

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 135,372

d. Estimation terminated at iteration number 4 because parameter estimates changed by less than ,001.

e. Estimation terminated at iteration number 8 because parameter estimates changed by less than ,001.

**Figura 13.56** Outputs da regressão logística binária no SPSS – procedimento *Forward Wald*.

Omnibus Tests of Model Coefficients				
		Chi-square	df	Sig.
Step 1	Step	44,281	1	,000
	Block	44,281	1	,000
	Model	44,281	1	,000
Step 2	Step	17,763	1	,000
	Block	62,045	2	,000
	Model	62,045	2	,000
Step 3	Step	6,899	1	,009
	Block	68,943	3	,000
	Model	68,943	3	,000
Step 4	Step	4,827	1	,028
	Block	73,770	4	,000
	Model	73,770	4	,000

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	91,090 <sup>a</sup>	,358	,482
2	73,327 <sup>b</sup>	,462	,623
3	66,428 <sup>b</sup>	,498	,672
4	61,602 <sup>b</sup>	,522	,703

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than ,001.

b. Estimation terminated at iteration number 8 because parameter estimates changed by less than ,001.

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
1	,000	0	
2	,542	4	,969
3	,531	5	,991
4	6,341	8	,609

Figura 13.56 (cont).

existência de pelo menos um parâmetro estatisticamente significativo para explicar a probabilidade de ocorrência do evento em estudo. O **output Omnibus Tests of Model Coefficients** apresenta esta estatística ( $\chi^2 = 73,77$ , Sig.  $\chi^2 = 0,000 < 0,05$ ), já calculada manualmente na seção 13.2.2 e também já apresentada na Figura 13.28, por meio da qual podemos rejeitar a hipótese nula de que todos os parâmetros  $\beta_j$  ( $j = 1, 2, \dots, 5$ ) sejam estatisticamente iguais a zero, ao nível de significância de 5%. Logo, pelo menos uma variável  $X$  é estatisticamente significativa para explicar a probabilidade de se chegar atrasado à escola e, portanto, temos um modelo de regressão logística binária estatisticamente significativa para fins de previsão.

Na sequência, são apresentados os resultados do teste de Hosmer-Lemeshow (**Hosmer and Lemeshow Test**) e a respectiva tabela de contingência que mostra, a partir dos grupos formados pelos decis das probabilidades estimadas, as frequências esperadas e observadas de observações por grupo. Por meio da análise do resultado do teste (para o passo 4,  $\chi^2 = 6,341$ , Sig.  $\chi^2 = 0,609 > 0,05$ ), já apresentado também por meio da Figura 13.30 quando da sua elaboração no Stata, não podemos rejeitar a hipótese nula de que as frequências esperadas e observadas sejam iguais, ao nível de significância de 5% e, portanto, o modelo final estimado não apresenta problemas em relação à qualidade do ajuste proposto.

A **Classification Table** apresenta a evolução, passo a passo, da classificação das observações. Para o modelo final (passo 4), obtivemos um valor de especificidade igual a 73,2%, de sensibilidade igual a 94,9% e uma eficiência global do modelo igual a 86,0%, para um *cutoff* de 0,5. Tais valores correspondem àqueles obtidos pela Tabela 13.11 e também já apresentados na Figura 13.37. A tabela de classificação cruzada (ou *crosstabulation*) pode também ser diretamente obtida ao clicarmos em **Analyze → Descriptive Statistics → Crosstabs....** Na caixa de diálogos que é aberta, devemos inserir a variável *PGR\_1* (*Predicted group*) em **Row(s)** e a variável *atrasado*, em **Column(s)**. Na sequência, devemos clicar em **OK**. Enquanto a Figura 13.57 mostra esta caixa de diálogo, a Figura 13.58 apresenta a tabela de classificação cruzada propriamente dita.

Contingency Table for Hosmer and Lemeshow Test

		chegou atrasado à escola? = Não		chegou atrasado à escola? = Sim		Total
		Observed	Expected	Observed	Expected	
Step 1	1	31	31,000	7	7,000	38
	2	10	10,000	52	52,000	62
Step 2	1	8	7,977	0	,023	8
	2	22	22,381	4	3,619	26
	3	2	1,633	2	2,367	4
	4	9	8,697	35	35,303	44
	5	0	,294	11	10,706	11
	6	0	,018	7	6,982	7
Step 3	1	8	7,994	0	,006	8
	2	20	20,366	2	1,634	22
	3	4	3,637	4	4,363	8
	4	9	8,658	28	28,342	37
	5	0	,145	7	6,855	7
	6	0	,193	10	9,807	10
	7	0	,007	8	7,993	8
Step 4	1	10	9,923	0	,077	10
	2	10	9,521	0	,479	10
	3	8	9,214	2	,786	10
	4	5	4,588	5	5,412	10
	5	4	3,244	6	6,756	10
	6	1	2,189	9	7,811	10
	7	3	1,513	7	8,487	10
	8	0	,587	10	9,413	10
	9	0	,196	10	9,804	10
	10	0	,026	10	9,974	10

Figura 13.56 (cont).

Voltando à análise dos *outputs* da Figura 13.56, o procedimento *Forward Wald (Stepwise)* elaborado pelo SPSS mostra o passo a passo dos modelos que foram elaborados, partindo da inclusão da variável mais significativa (maior estatística  $z$  de Wald entre todas as explicativas) até a inclusão daquela com menor estatística  $z$  de Wald, porém ainda com  $\text{Sig. } z < 0,05$ . Tão importante quanto a análise das variáveis incluídas no modelo final é a análise da lista de variáveis excluídas (**Variables not in the Equation**). Assim, podemos verificar que, ao se incluir no modelo 1 apenas a variável explicativa *per*, a lista de variáveis excluídas apresenta todas as demais. Se, para o primeiro passo, houver alguma variável explicativa que tenha sido excluída, mas que se apresenta de forma significativa ( $\text{Sig. } z < 0,05$ ), como ocorre, por exemplo, para a variável *sem*, esta variável será incluída no modelo no passo seguinte (modelo 2). E assim sucessivamente, até que a lista de variáveis excluídas não apresente mais nenhuma variável com  $\text{Sig. } z < 0,05$ . A variável remanescente nesta lista, para o nosso exemplo, é a variável *perfil2*, conforme já discutimos quando da elaboração da regressão no Excel e no Stata, e o modelo final (modelo 4 do procedimento *Forward Wald*), que é exatamente aquele já apresentado nas Figuras 13.12 e 13.28, conta com as variáveis explicativas *dist*, *sem*, *per* e *perfil3*. Desta forma, com base no *output Variables in the Equation* (passo 4) da Figura 13.56, podemos escrever a expressão final de probabilidade estimada de que um estudante  $i$  chegue atrasado à escola:

$$p_i = \frac{1}{1 + e^{-(30,933 + 0,204 \cdot \text{dist}_i + 2,920 \cdot \text{sem}_i - 3,776 \cdot \text{per}_i + 2,459 \cdot \text{perfil3}_i)}}$$

O *output Variables in the Equation* apresenta também as *odds ratios* de cada parâmetro estimado (**Exp(B)**), que correspondem àquelas obtidas por meio do comando **logistic** do Stata (Figura 13.33), com os respectivos intervalos de confiança. Caso desejássemos obter os intervalos de confiança dos parâmetros, ao invés daqueles

Classification Table<sup>a</sup>

Observed			Predicted		
			chegou atrasado à escola?		Percentage Correct
			Não	Sim	
Step 1	chegou atrasado à escola?	Não	31	10	75,6
		Sim	7	52	88,1
	Overall Percentage				83,0
Step 2	chegou atrasado à escola?	Não	30	11	73,2
		Sim	4	55	93,2
	Overall Percentage				85,0
Step 3	chegou atrasado à escola?	Não	28	13	68,3
		Sim	2	57	96,6
	Overall Percentage				85,0
Step 4	chegou atrasado à escola?	Não	30	11	73,2
		Sim	3	56	94,9
	Overall Percentage				86,0

a. The cut value is ,500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
								Lower	Upper
Step 1 <sup>a</sup>	per	-3,137	,543	33,427	1	,000	,043	,015	,126
	Constant	1,649	,345	22,797	1	,000	5,200		
Step 2 <sup>b</sup>	sem	2,193	,925	5,618	1	,018	8,959	1,462	54,910
	per	-3,223	,642	25,188	1	,000	,040	,011	,140
	Constant	-20,525	9,297	4,874	1	,027	,000		
Step 3 <sup>c</sup>	sem	2,740	1,086	6,365	1	,012	15,491	1,843	130,201
	per	-3,709	,805	21,215	1	,000	,025	,005	,119
	perfil3	2,670	1,142	5,469	1	,019	14,433	1,541	135,217
	Constant	-26,217	10,906	5,779	1	,016	,000		
Step 4 <sup>d</sup>	dist	,204	,101	4,073	1	,044	1,226	1,006	1,495
	sem	2,920	1,011	8,346	1	,004	18,543	2,557	134,456
	per	-3,776	,847	19,893	1	,000	,023	,004	,120
	perfil3	2,459	1,139	4,657	1	,031	11,694	1,253	109,109
	Constant	-30,933	10,636	8,458	1	,004	,000		

a. Variable(s) entered on step 1: per.  
b. Variable(s) entered on step 2: sem.  
c. Variable(s) entered on step 3: perfil3.  
d. Variable(s) entered on step 4: dist.

Variables not in the Equation

				Score	df	Sig.
Step 1	Variables	dist		,996	1	,318
		sem		9,170	1	,002
		perfil2		2,206	1	,137
		perfil3		4,669	1	,031
	Overall Statistics			21,729	4	,000
Step 2	Variables	dist		4,904	1	,027
		perfil2		1,157	1	,282
		perfil3		5,955	1	,015
	Overall Statistics			14,154	3	,003
Step 3	Variables	dist		4,099	1	,043
		perfil2		3,221	1	,073
	Overall Statistics			7,336	2	,026
Step 4	Variables	perfil2		3,459	1	,063
	Overall Statistics			3,459	1	,063

Figura 13.56 (cont).

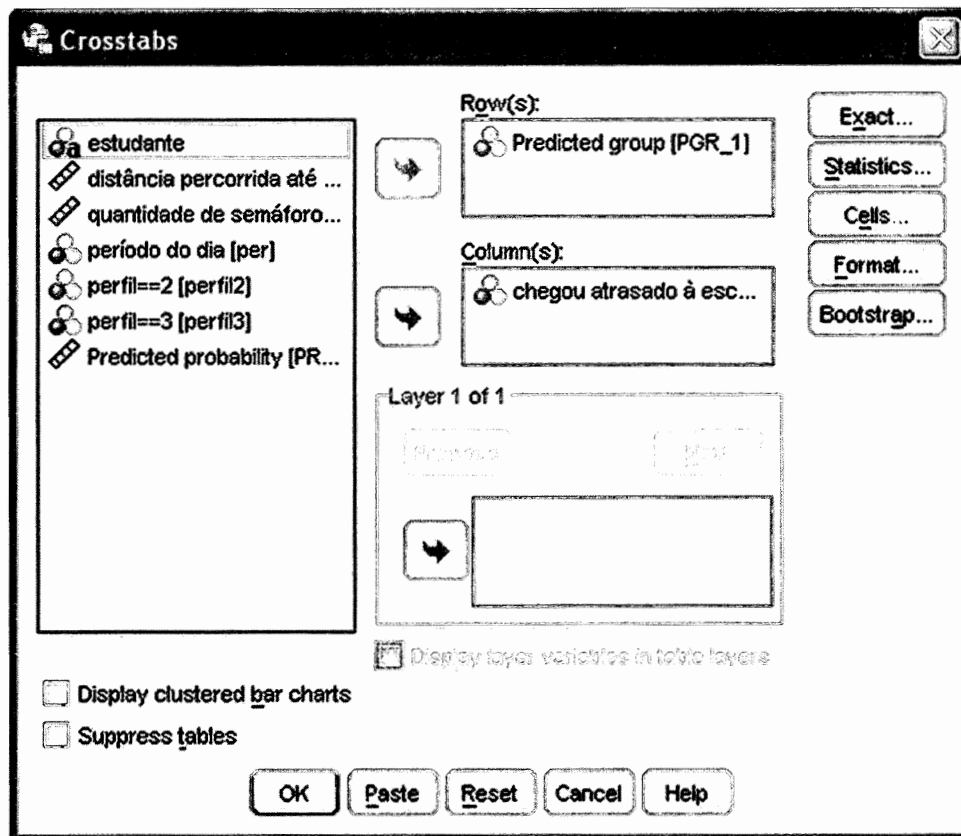


Figura 13.57 Caixa de diálogo para elaboração da tabela de classificação cruzada.

Predicted group \* chegou atrasado à escola? Crosstabulation

Count

		chegou atrasado à escola?		Total
		Não	Sim	
Predicted group	Não	30	3	33
	Sim	11	56	67
Total		41	59	100

Figura 13.58 Tabela de classificação cruzada.

referentes às chances, não deveríamos ter marcado a opção **CI for exp(B)** na caixa de diálogo **Options...** (Figura 13.54).

Por fim, vamos elaborar a curva ROC no SPSS. Para tanto, após a estimação do modelo final, devemos clicar em **Analyze → ROC Curve...** Uma caixa de diálogo como a apresentada na Figura 13.59 será aberta. Devemos inserir a variável *PRE\_1* (*Predicted probability*) em **Test Variable** e a variável *atrasado* em **State Variable**, com valor igual a 1 no campo **Value of State Variable**. Além disso, em **Display**, devemos clicar nas opções **ROC Curve** e **With diagonal reference line**. Na sequência, devemos clicar em **OK**.

A curva ROC elaborada encontra-se na Figura 13.60.

Conforme já discutimos quando da análise da Figura 13.42, a área abaixo da curva ROC, de 0,938, é considerada muito boa para definir a qualidade do modelo em termos de previsão de ocorrência do evento para novas observações.

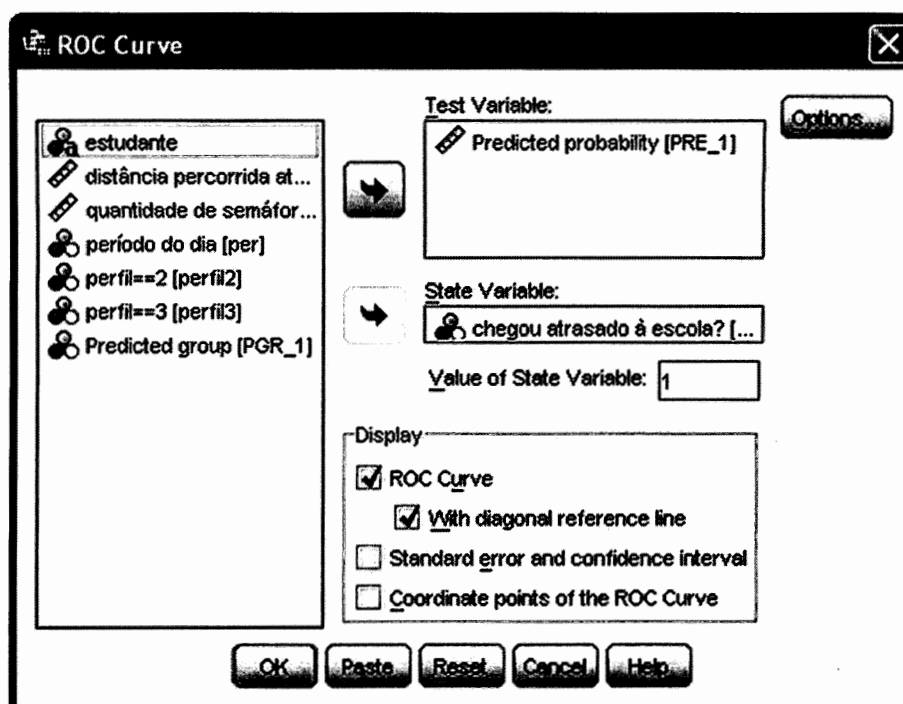
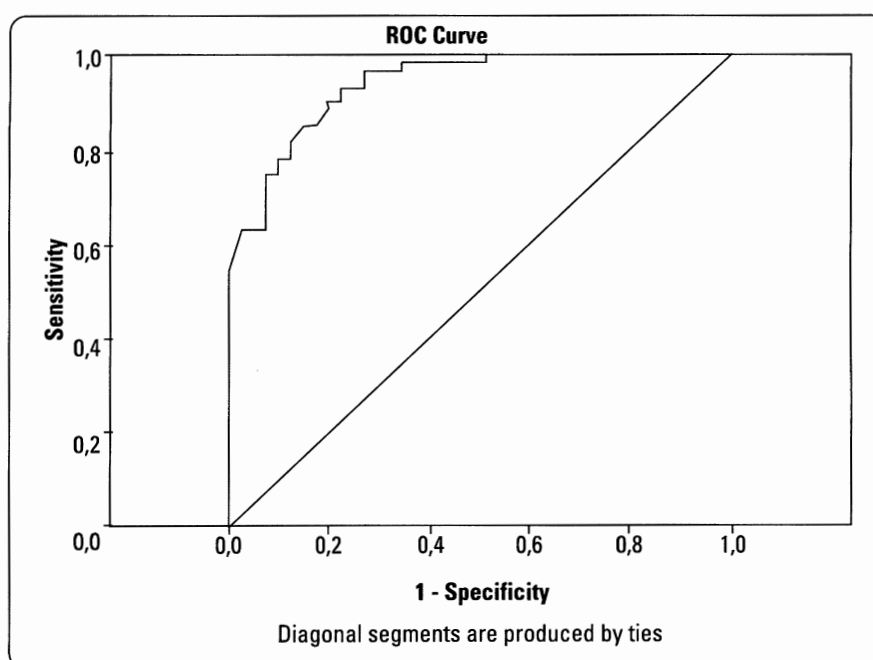


Figura 13.59 Caixa de diálogo para elaboração da curva ROC.



#### Area Under the Curve

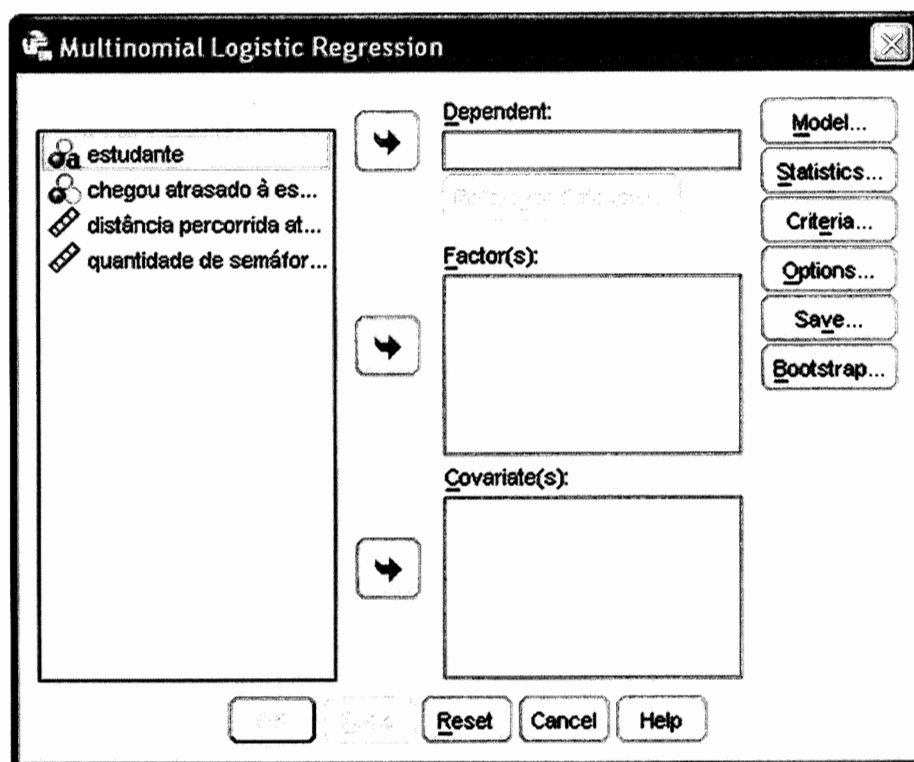
Test Result Variable  
(s): Predicted  
probability

Area
,938

Figura 13.60 Curva ROC.

### 13.5.2. Regressão logística multinomial no software SPSS

Vamos agora elaborar a modelagem da regressão logística multinomial no SPSS, por meio do mesmo exemplo utilizado nas seções 13.3 e 13.4.2. Os dados encontram-se no arquivo **AtrasadoMultinomial.sav** e, após o abrí-los, vamos inicialmente clicar em **Analyze → Regression → Multinomial Logistic...**. A caixa de diálogo da Figura 13.61 será aberta.



**Figura 13.61** Caixa de diálogo para elaboração da regressão logística multinomial no SPSS.

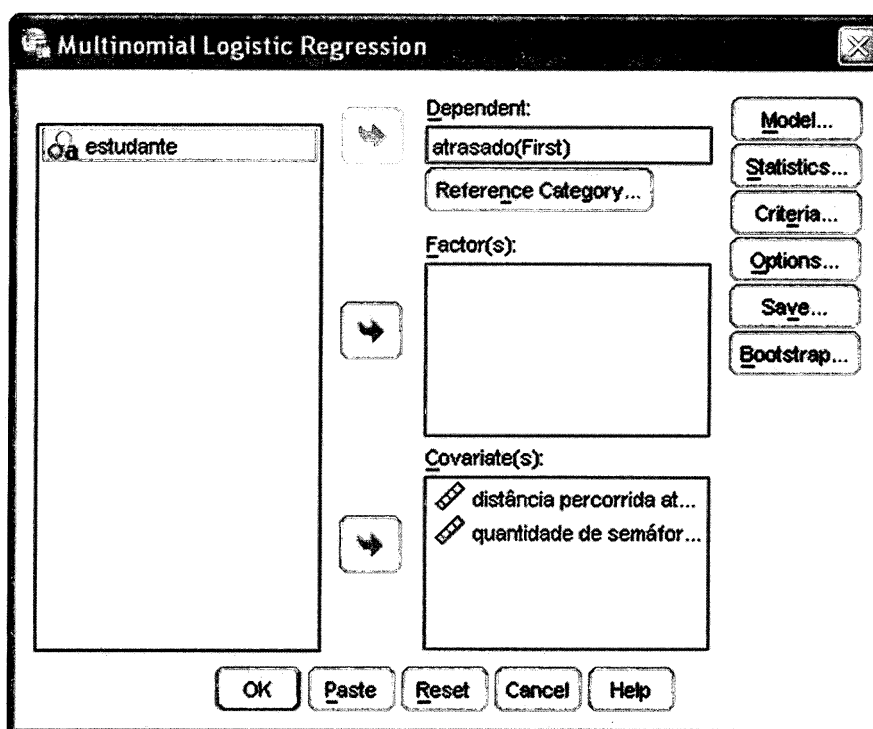
Vamos incluir a variável *atrasado* em **Dependent** e as variáveis explicativas quantitativas *dist* e *sem* na caixa **Covariate(s)**. A caixa **Factor(s)** deverá ser sempre preenchida com variáveis explicativas qualitativas, fato que não se aplica neste nosso exemplo. A Figura 13.62 apresenta esta caixa de diálogo devidamente preenchida.

Note que devemos definir a categoria de referência da variável dependente. Desta forma, em **Reference Category...**, devemos selecionar a opção **First Category**, uma vez que a categoria *não chegou atrasado* apresenta valores iguais a zero no banco de dados (Figura 13.63). Poderíamos também ter selecionado a opção **Custom**, com **Value** igual a 0. Esta última opção é mais utilizada quando o pesquisador tiver interesse em fazer com que determinada categoria intermediária da variável dependente seja a categoria de referência do modelo.

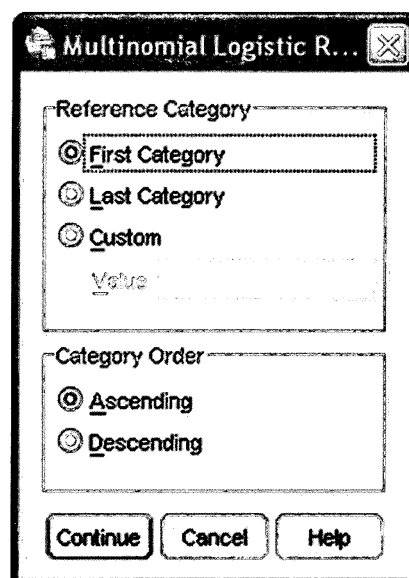
Após clicarmos em **Continue**, podemos dar sequência ao procedimento para elaboração da modelagem. No botão **Statistics...**, devemos clicar nas opções **Case processing summary** e, em **Model**, devemos marcar as opções **Pseudo R-square**, **Step summary**, **Model fitting information** e **Classification table**. Por fim, em **Parameters**, devemos marcar a opção **Estimates**. A Figura 13.64 mostra esta caixa de diálogo.

Por fim, após clicarmos em **Continue**, devemos selecionar o botão **Save...**. Nesta caixa de diálogo, vamos selecionar as opções **Estimated response probabilities** e **Predicted category**, conforme mostra a Figura 13.65. Este procedimento faz com que sejam geradas, para cada observação da amostra, as probabilidades de ocorrência de cada uma das três categorias da variável dependente e a classificação esperada de cada observação definida com base nestas probabilidades. Logo, serão geradas quatro novas variáveis no banco de dados (*EST1\_1*, *EST2\_1*, *EST3\_1* e *PRE\_1*).

Na sequência, vamos clicar em **Continue** e em **OK**. Os *outputs* gerados encontram-se na Figura 13.66.



**Figura 13.62** Caixa de diálogo para elaboração da regressão logística multinomial no SPSS com inclusão da variável dependente e das variáveis explicativas.



**Figura 13.63** Definição da categoria de referência da variável dependente.

Por meio destes *outputs*, podemos inicialmente verificar, com base no teste  $\chi^2$  ( $\chi^2 = 153,01$ ,  $\text{Sig. } \chi^2 = 0,000 < 0,05$  apresentado no *output Model Fitting Information*), que a hipótese nula de que todos os parâmetros  $\beta_{jm}$  ( $j = 1, 2; m = 1, 2$ ) sejam estatisticamente iguais a zero pode ser rejeitada ao nível de significância de 5%, ou seja, pelo menos uma variável  $X$  é estatisticamente significativa para explicar a probabilidade de ocorrência de pelo menos um dos eventos em estudo. Já o *output Pseudo R-Square* apresenta, diferentemente da regressão logística binária, o pseudo  $R^2$  de McFadden. O valor desta estatística, assim como o da estatística  $\chi^2$ , é exatamente igual àquele calculado manualmente na seção 13.3.2 e apresentado na Figura 13.45 quando da estimação do modelo no Stata.



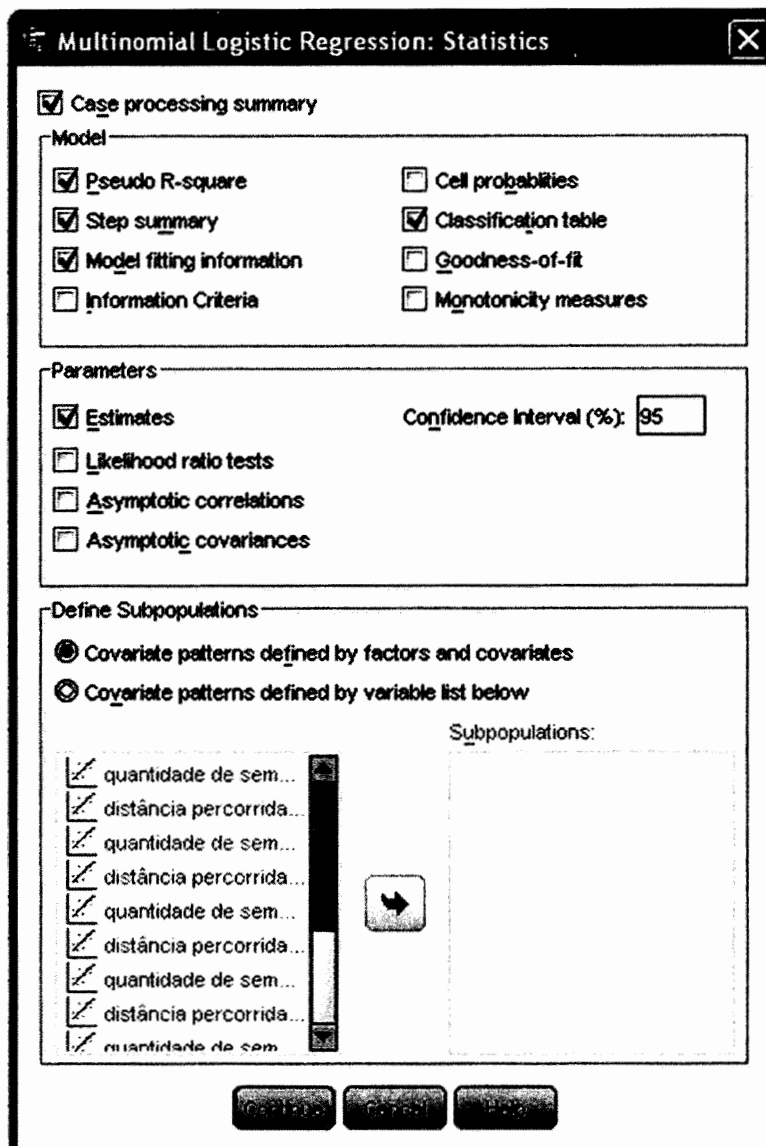


Figura 13.64 Caixa de diálogo para seleção das estatísticas da regressão logística multinomial.

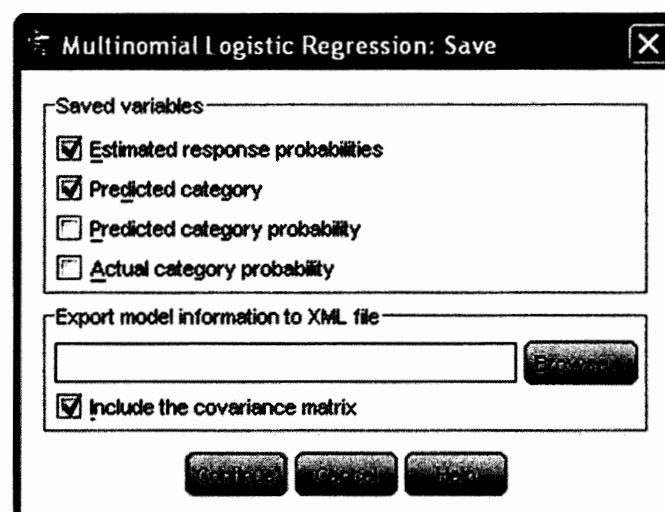


Figura 13.65 Caixa de diálogo para criação das variáveis referentes às probabilidades estimadas de ocorrência de cada categoria e a classificação de cada observação.

**Model Fitting Information**

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	199,841			
Final	46,826	153,015	4	,000

**Pseudo R-Square**

Cox and Snell	,783
Nagelkerke	,903
McFadden	,757

**Parameter Estimates**

		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
chegou atrasado à escola? <sup>a</sup>									
chegou atrasado à primeira aula	Intercept	-33,135	12,183	7,397	1	,007			
	dist	,559	,243	5,276	1	,022	1,749	1,085	2,817
	sem	1,670	,577	8,380	1	,004	5,312	1,715	16,453
chegou atrasado à segunda aula	Intercept	-62,292	14,675	18,018	1	,000			
	dist	1,078	,302	12,718	1	,000	2,940	1,625	5,318
	sem	2,895	,686	17,809	1	,000	18,081	4,713	69,363

a. The reference category is: não chegou atrasado.

**Classification**

Observed	Predicted			
	não chegou atrasado	chegou atrasado à primeira aula	chegou atrasado à segunda aula	Percent Correct
não chegou atrasado	47	2	0	95,9%
chegou atrasado à primeira aula	1	12	3	75,0%
chegou atrasado à segunda aula	0	5	30	85,7%
Overall Percentage	48,0%	19,0%	33,0%	89,0%

**Figura 13.66** Outputs da regressão logística multinomial no SPSS.

O modelo final pode ser obtido por meio do *output* **Parameter Estimates** e é exatamente igual ao apresentado na Figura 13.19 e obtido por meio do comando **mlogit** do Stata (Figura 13.45). Com base neste *output*, podemos escrever as expressões das probabilidades médias estimadas de ocorrência de cada um dos eventos representados pelas categorias da variável dependente, a saber:

**Probabilidade de um estudante  $i$  não chegar atrasado (categoria 0):**

$$p_{i_0} = \frac{1}{1 + e^{(-33,135 + 0,559 \cdot dist_i + 1,670 \cdot sem_i)} + e^{(-62,292 + 1,078 \cdot dist_i + 2,895 \cdot sem_i)}}$$

**Probabilidade de um estudante  $i$  chegar atrasado à primeira aula (categoria 1):**

$$p_{i_1} = \frac{e^{(-33,135 + 0,559 \cdot dist_i + 1,670 \cdot sem_i)}}{1 + e^{(-33,135 + 0,559 \cdot dist_i + 1,670 \cdot sem_i)} + e^{(-62,292 + 1,078 \cdot dist_i + 2,895 \cdot sem_i)}}$$

**Probabilidade de um estudante  $i$  chegar atrasado à segunda aula (categoria 2):**

$$p_{i_2} = \frac{e^{(-62,292 + 1,078 \cdot dist_i + 2,895 \cdot sem_i)}}{1 + e^{(-33,135 + 0,559 \cdot dist_i + 1,670 \cdot sem_i)} + e^{(-62,292 + 1,078 \cdot dist_i + 2,895 \cdot sem_i)}}$$

Este mesmo *output* apresenta também as *relative risk ratios* (**Exp(B)**) de cada parâmetro estimado, as quais correspondem àquelas obtidas por meio do comando **rrr** do Stata (Figura 13.48), com os respectivos intervalos de confiança.

Por fim, a tabela de classificação (*output* **Classification**) mostra, com base na maior probabilidade estimada ( $p_{i0}$ ,  $p_{i1}$  ou  $p_{i2}$ ) de cada observação, a classificação prevista e a observada para cada categoria da variável dependente. Desta forma, conforme já apresentado por meio da Tabela 13.18, chegamos a um modelo que apresenta um percentual total de acerto de 89,0% (eficiência global), possuindo um percentual de acerto de 95,9% quando houver indicação de que não ocorrerá atraso ao se chegar à escola, de 75,0% quando houver indicação de que haverá atraso na primeira aula e de 85,7% quando o modelo indicar que haverá atraso na segunda aula.

### 13.6. CONSIDERAÇÕES FINAIS

A estimação por máxima verossimilhança, embora ainda pouco conhecida por parte de um grande número de pesquisadores, é bastante útil para que se estimar parâmetros quando determinada variável dependente apresenta-se, por exemplo, na forma qualitativa.

A situação mais adequada para a aplicação de modelos de regressão logística binária acontece quando o fenômeno que se deseja estudar apresenta-se na forma dicotômica e o pesquisador tem a intenção de estimar uma expressão de probabilidade de ocorrência do evento definido dentre as duas possibilidades em função de determinadas variáveis explicativas. O modelo de regressão logística binária pode ser considerado um caso particular do modelo de regressão logística multinomial, cuja variável dependente também se apresenta na forma qualitativa, porém agora com mais de duas categorias de evento e, para cada categoria, será estimada uma expressão de probabilidade de sua ocorrência.

O desenvolvimento de qualquer modelo de dependência deve ser feito por meio do correto e consciente uso do software escolhido para a modelagem, com base na teoria subjacente e na experiência e na intuição do pesquisador.

### 13.7. EXERCÍCIOS

1. Uma empresa de concessão de crédito para consumo a pessoas físicas tem o intuito de avaliar a probabilidade de que seus clientes não cumpram com seus compromissos de pagamento (probabilidade de *default*). Por meio de uma base de dados com 2.000 observações que são os próprios clientes da companhia que obtiveram crédito recentemente, a empresa pretende estimar um modelo de regressão logística binária utilizando, como variáveis explicativas, a idade, o sexo (feminino = 0; masculino = 1) e a renda mensal (R\$) de cada indivíduo. A variável dependente refere-se ao *default* propriamente dito (não *default* = 0; *default* = 1). Os arquivos **Default.sav** e **Default.dta** trazem estes dados e, por meio da estimação do modelo de regressão logística binária, pede-se:

- a. Analise o nível de significância do teste  $\chi^2$ . Pelo menos uma das variáveis (*idade*, *sexo* e *renda*) é estatisticamente significativa para explicar a probabilidade de *default*, ao nível de significância de 5%?
- b. Se a resposta do item anterior for sim, analise o nível de significância de cada variável explicativa (testes *z* de Wald). Cada uma delas é estatisticamente significativa para explicar a probabilidade de *default*, ao nível de significância de 5%?
- c. Qual a equação final estimada para a probabilidade média de *default*?
- d. Em média, os indivíduos do sexo masculino tendem a apresentar maior probabilidade de *default* ao adquirirem crédito para consumo, mantidas as demais condições constantes?
- e. Em média, os indivíduos com maior idade tendem a apresentar maior probabilidade de *default* ao adquirirem crédito para consumo, mantidas as demais condições constantes?
- f. Qual a probabilidade média estimada de *default* de um indivíduo do sexo masculino, com 37 anos e com renda mensal de R\$6.850,00?
- g. Em média, em quanto se altera a chance de ser *default* ao se aumentar a renda em uma unidade, mantidas as demais condições constantes?
- h. Qual a eficiência global do modelo, para um *cutoff* de 0,5? E a sensibilidade e a especificidade, para este mesmo *cutoff*?

2. Com o intuito de estudar a fidelidade de clientes, um grupo supermercadista realizou uma pesquisa com 3.000 consumidores no momento em que o pagamento de suas respectivas compras estava sendo transacionado. Como a fidelidade de determinado consumidor pode ser medida com base no seu retorno ao estabelecimento, com compra efetuada, dentro de um ano da data da compra anterior, torna-se fácil o seu monitoramento por meio do acompanhamento do seu CPF. Assim, se o CPF de determinado consumidor estiver na base de dados da loja, porém não ocorre compra alguma com este mesmo CPF no período de um ano, este consumidor será classificado como *sem fidelidade ao estabelecimento*. Por outro lado, se o CPF de outro consumidor que também esteja na base de dados da loja é identificado em outra compra com intervalo de menos de um ano em relação à compra anterior, ele será classificado com a categoria *fidelidade ao estabelecimento*. A fim de estipular os critérios que elevam a probabilidade de que um consumidor apresente fidelidade ao estabelecimento, o grupo supermercadista coletou as seguintes variáveis de cada um dos 3.000 consumidores, na sequência os monitorando por um período de um ano da data daquela específica compra:

Variável	Descrição
<i>id</i>	Variável que substitui o CPF por motivos de confidencialidade. É uma variável <i>string</i> , varia de 0001 a 3000 e não será utilizada na modelagem.
<i>fidelidade</i>	Variável dependente binária correspondente ao fato de o consumidor retornar ou não à loja para efetuar nova compra em um período menor do que um ano (Não = 0; Sim = 1).
<i>sexo</i>	Sexo do consumidor (feminino = 0; masculino = 1).
<i>idade</i>	Idade do consumidor (anos).
<i>atendimento</i>	Variável qualitativa com 5 categorias correspondentes à percepção do nível de atendimento prestado pelo estabelecimento na compra atual (péssimo = 1; ruim = 2; regular = 3; bom = 4; ótimo = 5).
<i>sortimento</i>	Variável qualitativa com 5 categorias correspondentes à percepção de qualidade e variedade do sortimento de produtos ofertados pelo estabelecimento quando da compra atual (péssimo = 1; ruim = 2; regular = 3; bom = 4; ótimo = 5).
<i>acessibilidade</i>	Variável qualitativa com 5 categorias correspondentes à percepção de qualidade da acessibilidade ao estabelecimento, como estacionamento e acesso à área de vendas (péssimo = 1; ruim = 2; regular = 3; bom = 4; ótimo = 5).
<i>preço</i>	Variável qualitativa com 5 categorias correspondentes à percepção de preços ofertados dos produtos em relação à concorrência quando da compra atual (péssimo = 1; ruim = 2; regular = 3; bom = 4; ótimo = 5).

Por meio da análise do banco de dados presente nos arquivos **Fidelidade.sav** e **Fidelidade.dta**, pede-se:

- Quando da estimação do modelo completo de regressão logística binária com todas as variáveis explicativas do indivíduo (*sexo* e *idade*) e todas as  $(n - 1)$  *dummies* correspondentes às  $n$  categorias de cada uma das variáveis qualitativas, algumas destas categorias mostraram-se estatisticamente não significantes para explicar a probabilidade de ocorrência do evento (fidelidade ao estabelecimento varejista), ao nível de significância de 5%?
- Se a resposta do item anterior for sim, estime a expressão de probabilidade de ocorrência do evento por meio do procedimento *Stepwise*.
- Qual a eficiência global do modelo, com um *cutoff* de 0,5?
- Desejando estabelecer um critério que iguale a probabilidade de acerto daqueles que apresentarão fidelidade ao estabelecimento varejista à probabilidade de acerto daqueles que não apresentarão fidelidade, o diretor de marketing da empresa analisou a curva de sensibilidade do modelo. Qual o *cutoff* aproximado que iguala estas duas probabilidades de acerto?
- Para o modelo final estimado, em relação a um atendimento considerado péssimo, como se comportam, em média, as chances de se ter fidelidade ao estabelecimento por parte de consumidores que respondem ruim, regular, bom e ótimo para este quesito, mantidas as demais condições constantes?
- Elabore novamente o item anterior, porém agora utilizando separadamente as variáveis *sortimento*, *acessibilidade* e *preço*.
- Com base na análise das chances, o estabelecimento deseja investir em uma única variável perceptual para aumentar a probabilidade de que os consumidores tornem-se fiéis, fazendo com que deixem de ter percepções péssimas e passem, com maior frequência, a apresentar percepções ótimas sobre este quesito. Qual seria esta variável?

3. O Ministério da Saúde de determinado país deseja implementar uma campanha para melhorar os índices de colesterol LDL (mg/dL) dos cidadãos por meio do incentivo à prática de exercícios físicos e à redução do tabagismo e, para tanto, realizou uma pesquisa com 2.304 indivíduos, em que foram levantadas as seguintes variáveis:

Variável	Descrição
<i>colesterol</i>	Índice de colesterol LDL (mg/dL).
<i>cigarro</i>	Variável <i>dummy</i> correspondente ao fato de o indivíduo fumar ou não (não fuma = 0; fuma = 1).
<i>esporte</i>	Número de vezes em que pratica atividades físicas semanalmente.

Como se sabe que o índice de colesterol é posteriormente classificado segundo valores de referência, o Ministério da Saúde tem por intuito alertar a população sobre os benefícios trazidos pelo hábito de se praticar atividades físicas e pela abstinência do cigarro para a melhora da classificação. Desta forma, a variável *colesterol* será transformada para a variável *colestquali*, descrita a seguir, que apresenta 5 categorias e será a variável dependente do modelo cujos resultados serão divulgados pelo Ministério da Saúde.

Variável	Descrição
<i>colestquali</i>	Classificação do índice de colesterol LDL (mg/dL), a saber: <ul style="list-style-type: none"> <li>• Muito elevado: superior a 189 mg/dL (categoria de referência);</li> <li>• Elevado: de 160 a 189 mg/dL;</li> <li>• Limítrofe: de 130 a 159 mg/dL;</li> <li>• Subótimo: de 100 a 129 mg/dL;</li> <li>• Ótimo: inferior a 100 mg/dL.</li> </ul>

O banco de dados desta pesquisa encontra-se nos arquivos **Colestquali.sav** e **Colestquali.dta** e, por meio da estimação de um modelo de regressão logística multinomial com as variáveis *cigarro* e *esporte* como explicativas, pede-se:

- Apresente a tabela de frequências das categorias da variável dependente.
- Por meio da estimação de um modelo de regressão logística multinomial, é possível verificar que pelo menos uma das variáveis explicativas é estatisticamente significativa para compor a expressão de probabilidade de ocorrência de pelo menos uma das classificações propostas para o índice de colesterol LDL, ao nível de significância de 5%?
- Quais as equações finais estimadas para as probabilidades médias de ocorrência das classificações propostas para o índice de colesterol LDL?
- Quais as probabilidades de ocorrência de cada uma das classificações propostas para um indivíduo que não fuma e pratica atividades esportivas apenas uma vez por semana?
- Com base no modelo estimado, elabore um gráfico da probabilidade de ocorrência de cada evento representado pela variável dependente em função do número de vezes em que são realizadas atividades físicas semanalmente. A partir de qual periodicidade semanal de realização de atividades esportivas aumenta-se consideravelmente a probabilidade de que os índices de colesterol LDL passem a ser subótimos ou ótimos?
- Em média, em quanto se altera a chance de se ter um índice de colesterol considerado elevado, em relação a um nível considerado muito elevado, ao se aumentar em uma unidade o número de vezes em que são realizadas atividades físicas semanais, mantidas as demais condições constantes?
- Em média, em quanto se altera a chance de se ter um índice de colesterol considerado ótimo, em relação a um nível considerado subótimo, ao se deixar de fumar, mantidas as demais condições constantes?
- Elabore a tabela de classificação com base na probabilidade estimada de cada observação da amostra (classificação prevista e observada para cada categoria da variável dependente).
- Qual a eficiência global do modelo? Qual o percentual de acerto para cada categoria da variável dependente?

# Modelos de regressão probit

## A) Breve Introdução

Os **modelos de regressão probit**, cujo nome se refere à contração de *probability unit*, podem ser utilizados **alternativamente aos modelos de regressão logística binária**, para os casos em que a curva de probabilidades de ocorrência de determinado evento ajusta-se mais adequadamente à **função densidade de probabilidade acumulada da distribuição normal padrão**.

A ideia da regressão probit foi inicialmente concebida por Bliss (1934a, 1934b) que, ao realizar experimentos com o intuito de descobrir um eficaz pesticida contra insetos que se alimentavam de folhas de uva, acabou por representar graficamente a resposta dos insetos para diferentes níveis de concentração do pesticida. Como a relação encontrada entre a dose de pesticida e o tempo de resposta seguia uma **função sigmoide** (ou curva S), Bliss optou, naquela ocasião, por transformar a curva sigmoide dose-resposta em uma expressão linear, seguindo o já conhecido modelo de regressão linear. Duas décadas depois, Finney (1952), apoiando-se nas ideias e nos experimentos de Bliss, fez relevantes contribuições ao publicar um livro intitulado “*Probit Analysis*”. Ainda hoje, os modelos de regressão probit são muito utilizados para a compreensão de relações dose-resposta, quando a respectiva curva de probabilidades de ocorrência do evento de interesse, inicialmente representado por uma variável binária, seguir uma função sigmoide.

A **variável dependente segue uma distribuição de Bernoulli** e, portanto, a expressão da função-objetivo (logaritmo da função de verossimilhança) que tem por intuito estimar os parâmetros  $\alpha, \beta_1, \beta_2, \dots, \beta_k$  de determinado modelo de regressão probit é exatamente a mesma da expressão (13.15) deduzida neste capítulo para um modelo de regressão logística binária, dada por:

$$LL = \sum_{i=1}^n \{[(Y_i) \cdot \ln(p_i)] + [(1 - Y_i) \cdot \ln(1 - p_i)]\} = \text{máx} \quad (13.47)$$

O que varia, portanto, entre os modelos de regressão logística binária e os modelos de regressão probit é a expressão das probabilidades de ocorrência do evento de interesse  $p_i$ . Conforme estudamos, na regressão logística binária a expressão de  $p_i$ , que apresenta **distribuição logística**, é dada por:

$$p_i = \frac{1}{1 + e^{-Z_i}} = \frac{1}{1 + e^{-(\alpha + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \dots + \beta_k \cdot X_{ki})}} \quad (13.48)$$

Já para a regressão probit, a expressão das probabilidades de ocorrência do evento de interesse, que apresentam distribuição normal padrão acumulada, pode ser expressa por:

$$p_i = \Phi(Z_i) = \Phi(\alpha + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \dots + \beta_k \cdot X_{ki}) \quad (13.49)$$

em que  $\Phi$  representa a própria função densidade de probabilidade acumulada da distribuição normal padrão. Nesse sentido, a expressão (13.49) pode ser escrita conforme segue:

$$p_i = \int_{-\infty}^{Z_i} \frac{1}{\sqrt{2\pi}} \cdot e^{\left(-\frac{1}{2} \cdot Z^2\right)} dZ \quad (13.50)$$

que, para facilidade de cálculo, pode ser reescrita da seguinte maneira:

$$p_i = \frac{1}{2} + \frac{1}{2} \cdot \left( 1 - e^{-\frac{2 \cdot Z_i^2}{\pi}} \right)^{\frac{1}{2}} \text{ para } Z \geq 0 \quad (13.51)$$

e

$$p_i = 1 - \left[ \frac{1}{2} + \frac{1}{2} \cdot \left( 1 - e^{-\frac{2 \cdot Z_i^2}{\pi}} \right)^{\frac{1}{2}} \right] \text{ para } Z < 0 \quad (13.52)$$

A partir das expressões (13.48), (13.51) e (13.52), podemos elaborar a Tabela 13.21, que apresenta valores de  $p$  em função de valores de  $Z$  variando de  $-5$  a  $+5$  e torna possível a comparação entre as curvas logística (logit) e probit de probabilidades. Note que os valores de  $p$  na coluna referente à regressão logit são exatamente iguais aos já calculados e apresentados na Tabela 13.1. Caso o pesquisador opte por elaborar esta tabela no Excel, poderá fazer uso da função **=DIST.NORMP.N(Z; 1)** para determinar os valores de  $p$  na coluna referente à regressão probit.

**Tabela 13.21** Probabilidade de ocorrência de um evento ( $p$ ) em função de  $Z$  para os modelos de regressão logit e probit.

$Z_i$	Regressão Logit	Regressão Probit
	$p_i$	
-5	0,01	0,00
-4	0,02	0,00
-3	0,05	0,00
-2	0,12	0,02
-1	0,27	0,16
0	0,50	0,50
1	0,73	0,84
2	0,88	0,98
3	0,95	1,00
4	0,98	1,00
5	0,99	1,00

A partir da Tabela 13.21, podemos elaborar um gráfico de  $p = f(Z)$ , como o apresentado na Figura 13.67. Por meio deste gráfico, podemos verificar que, embora as probabilidades estimadas em função dos diversos valores assumidos por  $Z$  situam-se entre 0 e 1 para ambos os casos, parâmetros distintos serão estimados pelos modelos logit e probit, visto que diferentes valores de  $Z$  são necessários para que se chegue à mesma probabilidade de ocorrência do evento de interesse para determinada observação  $i$ .

Conforme podemos observar pelo gráfico da Figura 13.67, as funções logit e probit não são consideravelmente distintas, principalmente para valores de  $Z$  em torno de zero, sendo que os parâmetros estimados em cada caso seguem a relação  $\alpha, \beta_{logit} \approx 1,6 \cdot [\alpha, \beta_{probit}]$ , conforme discute Amemiya (1981). Essa relação também será por nós comprovada em exemplo a ser elaborado na próxima seção.

#### **Nesse sentido, para determinado banco de dados, qual modelo é melhor? O logit ou o probit?**

Conforme aponta Finney (1952), a opção pela escolha do modelo probit, em detrimento do modelo logit, dá-se, em tese, pela aderência da curva de probabilidades de ocorrência do evento de interesse à distribuição normal padrão acumulada. Na prática, entretanto, a decisão pode ser tomada com base em quatro critérios, cujos conceitos já foram discutidos ao longo deste capítulo:

- modelo com mais alto valor do logaritmo da função de verossimilhança;
- modelo com maior pseudo  $R^2$  de McFadden;
- modelo com mais alto nível de significância do teste de Hosmer-Lemeshow (menor estatística  $\chi^2$  deste teste);
- modelo com maior área abaixo da curva ROC.

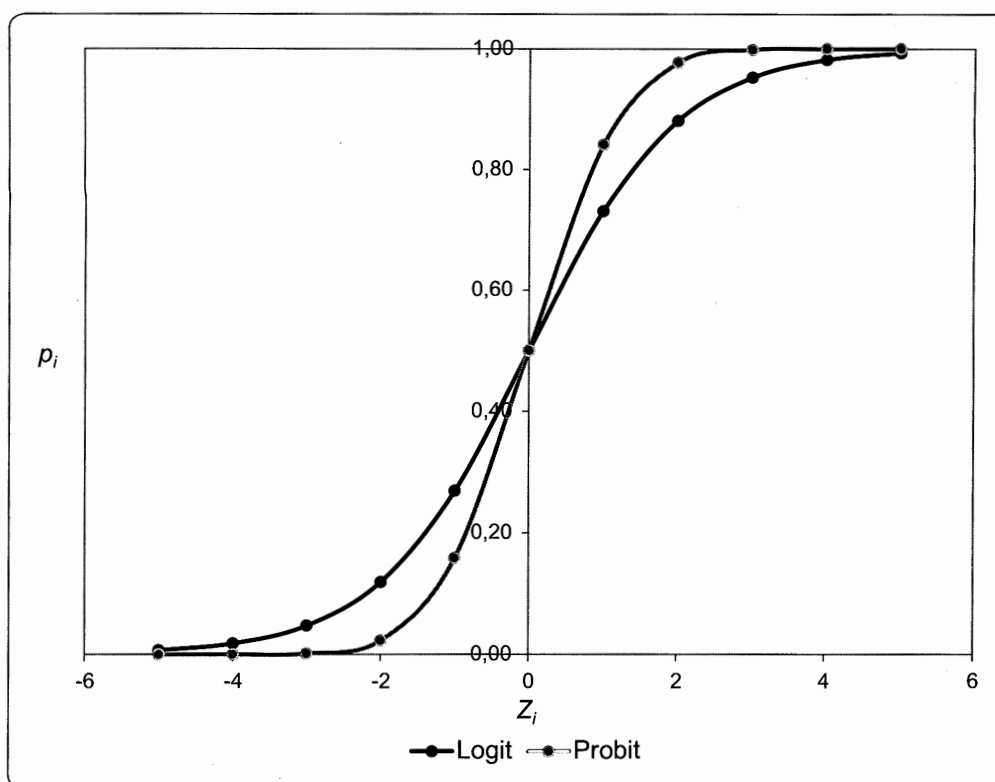


Figura 13.67 Gráfico de  $p = f(Z)$  para os modelos logit e probit.

Na sequência, apresentaremos um exemplo em que é estimado um modelo de regressão probit, cujos resultados são comparados com os obtidos por um modelo de regressão logística binária.

### B) Exemplo: Modelo de Regressão Probit no Stata

Faremos uso do banco de dados **Thriatlon.dta**, que apresenta dados levantados por meio de uma pesquisa realizada com 200 atletas amadores que participaram de determinada prova de triathlon do tipo *sprint*. O levantamento consistiu em verificar se determinado atleta completou ou não a prova, com o intuito de avaliar se tal fato relaciona-se com a quantidade de carboidratos, em gramas, por quilo de peso corporal ingerida no dia anterior. Para a variável dependente, como o evento de interesse refere-se a *Sim* (prova finalizada), essa categoria apresenta valores iguais a 1 no banco de dados, ficando a categoria *Não* (prova não finalizada) com valores iguais a 0. Nosso intuito, portanto, é estimar os parâmetros de  $Z$ , que é dado, para cada atleta  $i$ , por:

$$Z_i = \alpha + \beta_1 \cdot \text{carboidratos}_i$$

a partir da maximização do logaritmo da função de verossimilhança apresentada na expressão (13.47), em que:

$$p_i = \Phi(Z_i) = \Phi(\alpha + \beta_1 \cdot \text{carboidratos}_i)$$

O modelo proposto para este exemplo pode ser considerado de relação dose-resposta, visto que a quantidade, ou dose, de carboidratos ingeridos no dia anterior à prova de triathlon pode se relacionar com a finalização da mesma.

No Stata, podemos estimar os parâmetros do nosso modelo de regressão probit por meio da digitação do seguinte comando:

```
probit thriatlon carboidratos
```

cujos *outputs* encontram-se na Figura 13.68.



Alternativamente a esse comando, poderíamos ter digitado o seguinte comando:

**glm thriatlon carboidratos, family(binomial) link(probit)**

que gera exatamente os mesmos estimadores dos parâmetros, já que os modelos de regressão probit também se inserem dentro do grupo de **Modelos Lineares Generalizados** (*Generalized Linear Models*).

. probit thriatlon carboidratos						
Iteration 0: log likelihood = -121.31362						
Iteration 1: log likelihood = -97.527113						
Iteration 2: log likelihood = -97.429774						
Iteration 3: log likelihood = -97.429732						
Iteration 4: log likelihood = -97.429732						
Probit regression			Number of obs = 200			
			LR chi2(1) = 47.77			
			Prob > chi2 = 0.0000			
Log likelihood = -97.429732			Pseudo R2 = 0.1969			
thriatlon	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
carboidratos	.379623	.0600936	6.32	0.000	.2618417	.4974042
_cons	-1.64247	.2058876	-7.98	0.000	-2.046002	-1.238937

Figura 13.68 Outputs da regressão probit no Stata.

É importante mencionar que um pesquisador mais curioso poderá obter esses mesmos *outputs* por meio do arquivo **Thriatlon Probit Máxima Verossimilhança.xls**, fazendo uso da ferramenta **Solver** do Excel, conforme padrão também adotado ao longo do capítulo e do livro. Neste arquivo, os critérios do **Solver** já estão previamente definidos.

Com base nos *outputs* da Figura 13.68, podemos verificar que os parâmetros estimados são estatisticamente diferentes de zero, a 95% de confiança, e a expressão final de probabilidade estimada de que um atleta  $i$  complete a prova é dada por:

$$p_i = \Phi(-1,642 + 0,379 \cdot \text{carboidratos}_i)$$

Nesse sentido, a probabilidade média estimada de finalização da prova de triathlon para, por exemplo, um participante que tenha ingerido no dia anterior 10 gramas de carboidratos por quilo de peso corporal, pode ser obtida por meio da digitação do seguinte comando:

**mfx, at(carboidratos = 10)**

O *output* é apresentado na Figura 13.69 e, por meio do qual, podemos chegar à resposta de 0,984 (98,4%). Essa resposta também pode ser obtida a partir da seguinte expressão:

$$p_i = \Phi[-1,642 + 0,379 \cdot (10)] = \Phi(2,148)$$

em que o valor 2,148 representa a abscissa ( $Z_{score}$ ) da distribuição normal padrão acumulada, que resulta em um valor de probabilidade de 0,984. Para fins de verificação, o pesquisador pode digitar o comando **display normal(2.148)** no Stata ou até mesmo a função **=DIST.NORMPN(2,148; 1)** em qualquer célula do Excel.

. mfx, at(carboidratos = 10)						
Marginal effects after probit						
y = Pr(thriatlon) (predict)						
= .9843705						
variable	dy/dx	Std. Err.	z	P> z	[ 95% C.I. ]	X
carboidratos	.0148931	.01167	1.28	0.202	-.007981 .037767	10

Figura 13.69 Cálculo da probabilidade estimada quando *carboidratos* = 10 – comando **mfx**.

Além disso, podemos verificar, assim como para a estimação dos modelos de regressão logística binária, que o Stata também apresenta, em seus *outputs*, o valor do pseudo  $R^2$  de McFadden na estimação de modelos de regressão probit, cujo cálculo também é feito com base na expressão (13.16) e cuja utilidade restringe-se apenas a casos em que o pesquisador tiver interesse em comparar dois ou mais modelos distintos (critério de maior pseudo  $R^2$  de McFadden).

Caso o pesquisador também deseje estimar os parâmetros do modelo correspondente de regressão logística binária, a fim de compará-los com os obtidos pela modelagem de regressão probit, poderá digitar a seguinte sequência de comandos:

```
eststo: quietly logit thriatlon carboidratos
predict prob1
```

```
eststo: quietly probit thriatlon carboidratos
predict prob2
```

```
esttab, scalars(11) se pr2
```

A Figura 13.70 apresenta os principais resultados obtidos em cada estimação.

```
. eststo: quietly logit thriatlon carboidratos
(est1 stored)
. predict prob1
(option pr assumed; Pr(thriatlon))

. eststo: quietly probit thriatlon carboidratos
(est2 stored)
. predict prob2
(option pr assumed; Pr(thriatlon))

. esttab, scalars(11) se pr2
```

	(1) thriatlon	(2) thriatlon
carboidratos	0.642*** (0.109)	0.380*** (0.0601)
_cons	-2.767*** (0.382)	-1.642*** (0.206)
N	200	200
pseudo R-sq	0.196	0.197
ll	-97.52	-97.43

Standard errors in parentheses  
\* p<0.05, \*\* p<0.01, \*\*\* p<0.001

Figura 13.70 Principais resultados obtidos nas estimações logit e probit.

A partir dos *outputs* consolidados, é possível verificarmos que, embora existam diferenças entre as estimações dos parâmetros em cada caso, os valores obtidos do **logaritmo da função de verossimilhança (11, ou *log likelihood*)** e do **pseudo  $R^2$  de McFadden** são ligeiramente maiores para o modelo probit (modelo 2 na Figura 13.70), o que o torna preferível ao modelo logit para os dados do nosso exemplo.

Em relação aos parâmetros estimados propriamente ditos, podemos inclusive chegar às seguintes relações:

$$\frac{\alpha_{\text{logit}}}{\alpha_{\text{probit}}} = \frac{-2,767}{-1,642} = 1,69$$

$$\frac{\beta_{\text{logit}}}{\beta_{\text{probit}}} = \frac{0,642}{0,380} = 1,69$$

que estão de acordo com o discutido por Amemiya (1981).

Para efeitos de interpretação, podemos afirmar que, enquanto a ingestão de 1 grama a mais de carboidratos por quilo de peso corporal incrementa o logaritmo natural da chance de finalização da prova de triathlon, em média, em 0,642 (modelo logit), o mesmo fato faz com que o *Zscore* da distribuição normal padrão acumulada seja incrementado, em média, em 0,380 (modelo probit).

Na sequência, podemos estudar e comparar os níveis de significância do teste de Hosmer-Lemeshow e as áreas abaixo da curva ROC dos dois modelos. Para tanto, devemos digitar os seguintes comandos:

```
quietly logit thriathlon carboidratos
estat gof, group(10)
lroc, nograph
```

```
quietly probit thriathlon carboidratos
estat gof, group(10)
lroc, nograph
```

Os novos *outputs* encontram-se na Figura 13.71.

```
. quietly logit thriathlon carboidratos
. estat gof, group(10)

Logistic model for thriathlon, goodness-of-fit test

(Table collapsed on quantiles of estimated probabilities)

      number of observations =      200
      number of groups      =       10
Hosmer-Lemeshow chi2(8)    =       9.14
      Prob > chi2          =     0.3305

. lroc, nograph

Logistic model for thriathlon

number of observations =      200
area under ROC curve   =     0.7892

. quietly probit thriathlon carboidratos
. estat gof, group(10)

Probit model for thriathlon, goodness-of-fit test

(Table collapsed on quantiles of estimated probabilities)

      number of observations =      200
      number of groups      =       10
Hosmer-Lemeshow chi2(8)    =       8.93
      Prob > chi2          =     0.3479

. lroc, nograph

Probit model for thriathlon

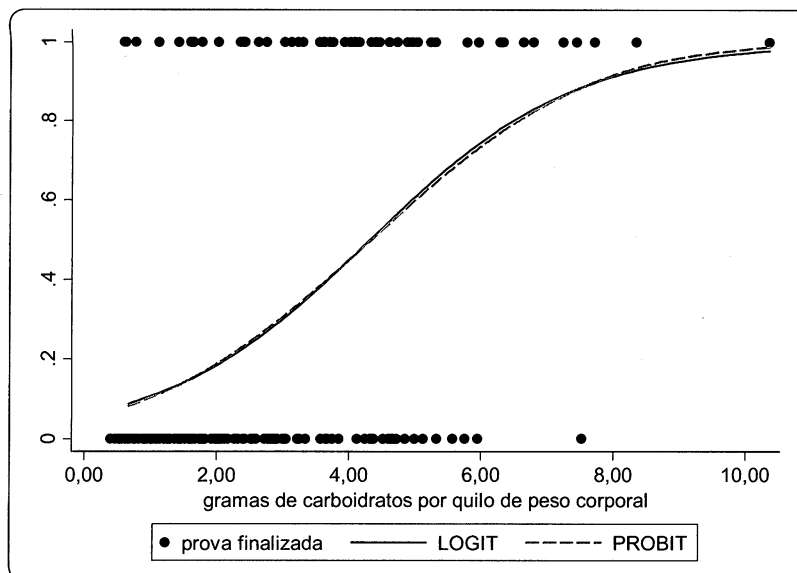
number of observations =      200
area under ROC curve   =     0.7892
```

**Figura 13.71** Testes de Hosmer-Lemeshow e áreas abaixo da curva ROC obtidos nas estimações logit e probit.

A partir desses *outputs*, podemos verificar que as áreas abaixo da curva ROC são iguais nos dois modelos. Entretanto, embora as estimações não apresentem problemas em relação à qualidade do ajuste proposto, visto que não há rejeição da hipótese nula de que as frequências esperadas e observadas sejam iguais, ao nível de confiança de 95%, o nível de significância do teste de Hosmer-Lemeshow do modelo probit ( $\chi^2 = 8,93$ , Sig.  $\chi^2 = 0,3479$ ) é levemente superior ao do modelo logit ( $\chi^2 = 9,14$ , Sig.  $\chi^2 = 0,3305$ ), fato que sugere que o primeiro (probit) apresenta uma qualidade um pouco melhor do ajuste proposto.

Por fim, podemos elaborar um gráfico que relaciona os valores esperados (previstos) de probabilidade de finalização da prova de triathlon para cada atleta (variáveis já geradas *prob1* e *prob2* para, respectivamente, os modelos logit e probit) com a variável *carboidratos*. Este gráfico é apresentado na Figura 13.72, e o comando para a sua geração é:

```
graph twoway scatter thriatlon carboidratos || mspline prob1
carboidratos || mspline prob2 carboidratos ||, legend(label(2 "LOGIT")
label(3 "PROBIT"))
```



**Figura 13.72** Probabilidades de ocorrência do evento (finalizar o triathlon) em função da variável *carboidratos*, com ajustes logit e probit.

Embora este gráfico mostre, para os dados deste exemplo, que não existem diferenças consideráveis entre os ajustes logit e probit, os critérios discutidos favorecem a adoção do último.

É recomendável, para modelos em que a variável dependente for binária, que o pesquisador justifique a adoção de determinado critério de estimação, ou ao menos investigue se há certa aderência da curva de probabilidades de ocorrência do evento em análise à distribuição normal padrão acumulada. Se esse for o caso, os modelos de regressão probit podem ser mais adequados para a geração de probabilidades previstas condizentes com a realidade estudada.