

Modelos de Regressão para Dados de Contagem: Poisson e Binomial Negativo

A vida é boa somente por duas coisas: estudar matemática e ensiná-la.

Siméon-Denis Poisson

Ao final deste capítulo, você terá condições de:

- Estabelecer as circunstâncias a partir das quais os modelos de regressão para dados de contagem podem ser utilizados.
- Entender a estimativa dos parâmetros de um modelo de regressão Poisson e de um modelo de regressão binomial negativo pelo método de máxima verossimilhança.
- Avaliar os resultados dos testes estatísticos pertinentes aos modelos de regressão Poisson e binomial negativo.
- Elaborar intervalos de confiança dos parâmetros do modelo estimado para efeitos de previsão.
- Estimar modelos de regressão Poisson e binomial negativo em Microsoft Office Excel®, Stata Statistical Software® e IBM SPSS Statistics Software® e interpretar seus resultados.

14.1. INTRODUÇÃO

Os modelos de regressão Poisson e binomial negativo fazem parte do que é conhecido por modelos de regressão para dados de contagem, e têm por objetivo analisar o comportamento, em relação a variáveis preditoras, de determinada variável dependente que se apresenta na forma quantitativa, porém com valores discretos e não negativos (dados de contagem).

Nestes casos, segundo Ramalho (1996), o modelo clássico de regressão linear não é adequado para explicar como uma variável discreta, que somente pode assumir um pequeno número de valores estritamente positivos, depende de um conjunto de variáveis preditoras. Além disso, teremos também interesse em calcular, após a estimativa do modelo desejado, a probabilidade de ocorrência do fenômeno em estudo, dado o comportamento das variáveis explicativas.

Segundo o mesmo autor, é comum, quando estamos trabalhando com dados de contagem, iniciarmos a estimativa dos parâmetros por meio de um **modelo de regressão Poisson**, devido à sua simplicidade. Neste caso, a variável dependente de um modelo de regressão Poisson deve seguir uma distribuição Poisson com média igual à variância. Entretanto, de acordo com Tadano, Ugaya e Franco (2009), esta propriedade é frequentemente violada em estudos empíricos, já que é comum a existência de **superdispersão**, ou seja, é frequente que a variância da variável dependente seja maior do que a sua média. Nestes casos, trabalharemos com a estimativa de um **modelo de regressão binomial negativo**.

Ainda para Tadano, Ugaya e Franco (2009), os modelos de regressão Poisson e binomial negativo, que também se inserem no contexto dos **Modelos Lineares Generalizados (Generalized Linear Models)**, em que são utilizadas classes de modelos que oferecem alternativas para a transformação dos dados devido ao caráter não linear da variável dependente, tiveram sua origem na década de 1970, quando Wedderburn (1974) desenvolveu a teoria da quasi-verossimilhança.

Ao contrário da tradicional técnica de regressão estimada por meio de métodos de mínimos quadrados, os modelos de regressão para dados de contagem são estimados por máxima verossimilhança e a escolha da melhor estimativa depende da distribuição da variável dependente, da relação entre sua média e variância e do objetivo do estudo, com base na teoria subjacente e na experiência do pesquisador.

É comum encontrarmos exemplos de aplicação de modelos de regressão para dados de contagem em economia, finanças, demografia, ecologia e meio-ambiente, atuária, medicina e veterinária, entre outras áreas do conhecimento.

Imagine, por exemplo, que um pesquisador tenha interesse em avaliar a quantidade de vezes que um grupo de pacientes idosos vai ao médico por ano, em função da idade de cada um deles, do sexo e das características dos seus planos de saúde. Um segundo pesquisador deseja estudar a quantidade de ofertas públicas de ações que são realizadas em uma amostra de países desenvolvidos e emergentes num determinado ano, com base em seus desempenhos econômicos, como inflação, taxa de juros, produto interno bruto e taxa de investimento estrangeiro. Note que a quantidade de visitas ao médico ou a quantidade de ofertas públicas de ações são as variáveis dependentes nos dois casos, sendo representadas por dados quantitativos que assumem valores discretos e restritos a um determinado número de ocorrências, ou seja, são dados de contagem.

Entretanto, imagine que a média e a variância da variável correspondente ao número de visitas ao médico por ano sejam aproximadamente iguais. Desta forma, poderemos estimar um clássico modelo de regressão Poisson. Por outro lado, como a dispersão, entre países, da quantidade de ofertas públicas de ações é muito maior do que a média geral, estaremos lidando com o fenômeno da superdispersão e, consequentemente, poderemos estimar um modelo de regressão binomial negativo. Segundo Cameron e Trivedi (2009), a superdispersão é comumente gerada pela presença de maior heterogeneidade nos dados entre observações da amostra.

A Figura 14.1 apresenta, de maneira ilustrativa, uma variável com distribuição Poisson e outra com distribuição binomial negativa. Embora as distribuições sejam aparentemente semelhantes, nota-se que a dispersão é maior para o segundo caso (Figura 14.1b).

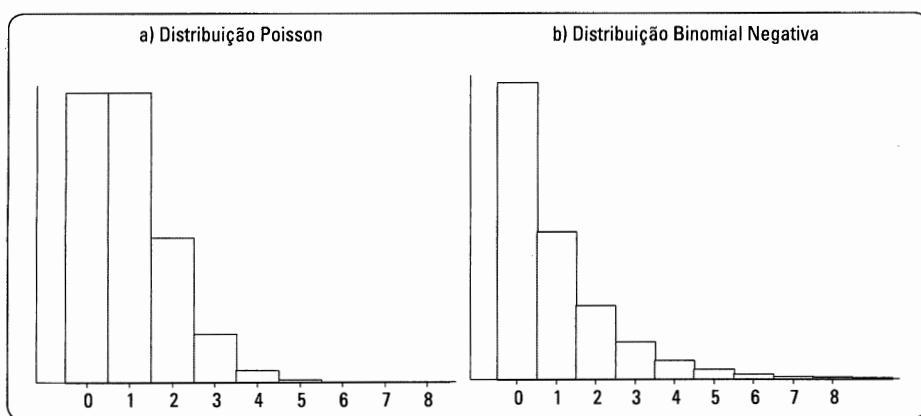


Figura 14.1 Exemplo de distribuição Poisson e de distribuição binomial negativa.

Como a variável dependente apresenta-se de maneira quantitativa, é muito comum que não seja estudada a sua distribuição e, consequentemente, é possível que um pesquisador desavisado ou iniciante estime o modelo por meio da regressão por mínimos quadrados ordinários, inclusive obtendo *outputs*. **Este procedimento está incorreto, já que poderá gerar estimadores viesados, porém infelizmente é mais comum do que parece!**

É importante mencionar que ainda fazem parte dos modelos de regressão para dados de contagem os chamados **modelos de regressão inflacionados de zeros**, cujos parâmetros podem ser estimados quando a variável dependente apresentar uma quantidade considerável de valores de contagem iguais a zero. Estudaremos especificamente os modelos inflacionados de zeros dos tipos Poisson e binomial negativo no apêndice do presente capítulo.

Conforme discutido nos capítulos anteriores, os modelos de regressão para dados de contagem também devem ser 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 para dados de contagem, com os seguintes objetivos: (1) introduzir os conceitos sobre os modelos de regressão Poisson e binomial negativo; (2) apresentar a estimativa por máxima verossimilhança em modelos de regressão para dados de contagem; (3) interpretar os resultados obtidos e elaborar previsões; e (4) apresentar a aplicação das técnicas em Excel, Stata e SPSS. Segundo a lógica dos capítulos anteriores, será inicialmente 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 em Stata e em SPSS.

14.2. O MODELO DE REGRESSÃO POISSON

Os modelos de regressão para dados de contagem têm, por objetivo principal, estudar o comportamento de uma variável dependente, definida por Y , que se apresenta com valores discretos e não negativos, com base no comportamento de variáveis explicativas. Segundo Cameron e Trivedi (2009), o ponto inicial para o estudo dos modelos de regressão para dados de contagem é a apresentação da distribuição Poisson que, para determinada observação i ($i = 1, 2, \dots, n$, em que n é o tamanho da amostra), possui, analogamente ao apresentado na expressão (5.45) do Capítulo 5, a seguinte probabilidade de ocorrência de uma contagem m em dada exposição (período, área, região, entre outros exemplos):

$$p(Y_i = m) = \frac{e^{-\lambda_i} \cdot \lambda_i^m}{m!}, \quad m = 0, 1, 2, \dots \quad (14.1)$$

em que λ é o número esperado de ocorrências ou a taxa média estimada de incidência do fenômeno em estudo para dada exposição (em inglês, **incidence rate ratio**).

A partir da expressão (14.1), podemos elaborar uma tabela com valores de p em função dos valores de m . Como m é um número inteiro e não negativo, pode variar de 0 a $+\infty$ e, dessa forma, iremos, apenas para efeitos didáticos, utilizar valores inteiros entre 0 a 20. A Tabela 14.1 traz estes valores, para três situações diferentes de λ .

Tabela 14.1 Probabilidade de ocorrência de uma contagem m para diferentes valores de λ .

	$\lambda_i = 1$	$\lambda_i = 4$	$\lambda_i = 10$
m	$p(Y_i = m) = \frac{e^{-\lambda_i} \cdot \lambda_i^m}{m!}$		
0	0,3679	0,0183	0,0000
1	0,3679	0,0733	0,0005
2	0,1839	0,1465	0,0023
3	0,0613	0,1954	0,0076
4	0,0153	0,1954	0,0189
5	0,0031	0,1563	0,0378
6	0,0005	0,1042	0,0631
7	0,0001	0,0595	0,0901
8	0,0000	0,0298	0,1126
9	0,0000	0,0132	0,1251
10	0,0000	0,0053	0,1251
11	0,0000	0,0019	0,1137
12	0,0000	0,0006	0,0948
13	0,0000	0,0002	0,0729
14	0,0000	0,0001	0,0521
15	0,0000	0,0000	0,0347
16	0,0000	0,0000	0,0217
17	0,0000	0,0000	0,0128
18	0,0000	0,0000	0,0071
19	0,0000	0,0000	0,0037
20	0,0000	0,0000	0,0019

A partir dos dados calculados na Tabela 14.1, podemos elaborar o gráfico da Figura 14.2.

Por meio da análise deste gráfico, é possível verificarmos um achatamento da curva de probabilidades e o seu deslocamento para a direita à medida que o número esperado de ocorrências (λ) aumenta, chegando ao ponto de a curva se aproximar de uma distribuição normal para valores maiores de λ .

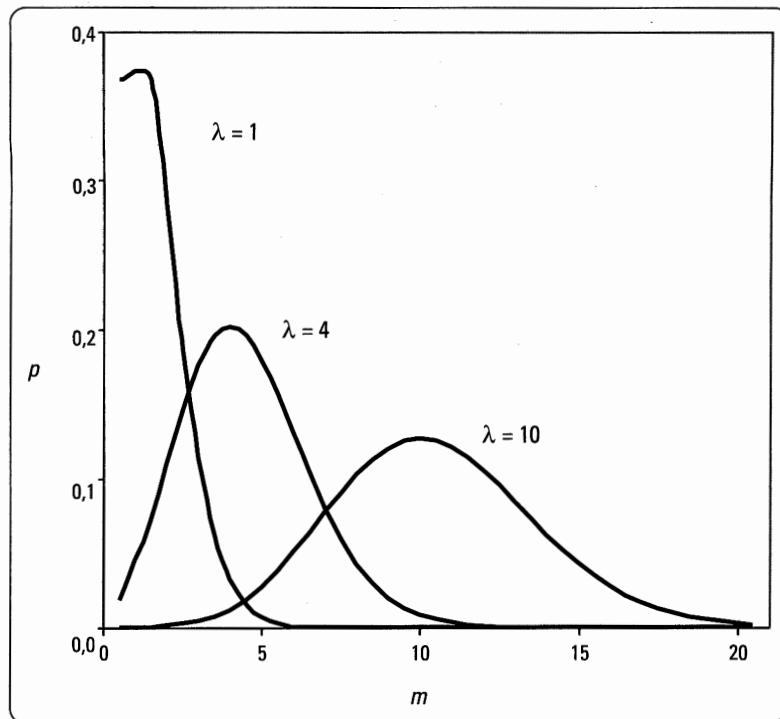


Figura 14.2 Distribuição Poisson – gráficos de probabilidade de ocorrência de uma contagem m em função do número esperado de ocorrências λ .

Na distribuição Poisson, a média e a variância da variável em estudo devem ser iguais a λ , conforme pode ser demonstrado a seguir:

- **Média:**

$$E(Y) = \sum_{m=0}^{\infty} m \cdot \frac{e^{-\lambda} \cdot \lambda^m}{m!} = \lambda \cdot \sum_{m=1}^{\infty} \frac{e^{-\lambda} \cdot \lambda^{m-1}}{(m-1)!} = \lambda \cdot 1 = \lambda \quad (14.2)$$

- **Variância:**

$$\begin{aligned} Var(Y) &= \sum_{m=0}^{\infty} \frac{e^{-\lambda} \cdot \lambda^m}{m!} \cdot (m - \lambda)^2 = \sum_{m=0}^{\infty} \frac{e^{-\lambda} \cdot \lambda^m}{m!} \cdot (m^2 - 2m\lambda + \lambda^2) \\ &= \lambda^2 \cdot \sum_{m=2}^{\infty} \frac{e^{-\lambda} \cdot \lambda^{m-2}}{(m-2)!} + \lambda \cdot \sum_{m=1}^{\infty} \frac{e^{-\lambda} \cdot \lambda^{m-1}}{(m-1)!} - \lambda^2 = \lambda \end{aligned} \quad (14.3)$$

Caso esta propriedade, conhecida por **equidispersão da distribuição Poisson**, seja atendida, poderemos estimar um modelo de regressão Poisson, definido da seguinte forma:

$$\ln(\hat{Y}_i) = \ln(\lambda_i) = \alpha + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \dots + \beta_k \cdot X_{ki} \quad (14.4)$$

que também é chamado de modelo log-linear (ou semilogarítmico à esquerda). Sendo assim, o número esperado de ocorrências em dada exposição, para determinada observação i , pode ser escrito como:

$$\lambda_i = e^{(\alpha + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \dots + \beta_k \cdot X_{ki})} \quad (14.5)$$

em que α representa a constante, β_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).

Feita esta pequena introdução sobre os modelos de regressão Poisson, partiremos, então, para a estimação propriamente dita dos seus parâmetros, por meio da apresentação de um exemplo elaborado inicialmente em Excel.

14.2.1. Estimação do modelo de regressão Poisson por máxima verossimilhança

Seguindo a lógica proposta no livro, apresentaremos agora os conceitos pertinentes à estimação por máxima verossimilhança de um modelo de regressão Poisson por meio de um exemplo similar ao desenvolvido nos capítulos anteriores. Entretanto, agora a variável dependente apresentará dados de contagem.

Imagine que o nosso mesmo professor curioso e investigativo, que já explorou consideravelmente os efeitos de determinadas variáveis explicativas sobre o tempo de deslocamento de um grupo de alunos até a escola e sobre a probabilidade de se chegar atrasado às aulas, por meio, respectivamente, das técnicas de regressão múltipla e de regressão logística binária e multinomial, tenha agora o interesse em investigar se algumas destas mesmas variáveis explicativas influenciam a quantidade de vezes que os alunos chegam atrasados durante o período de uma semana. Desta forma, o fenômeno em questão a ser estudado apresenta-se na forma quantitativa (incidência de atrasos semanalmente), porém apenas com valores não negativos e discretos (dados de contagem).

Sendo assim, o professor elaborou uma pesquisa com 100 alunos da escola onde leciona, questionando sobre a quantidade de vezes que cada um deles chegou atrasado à escola na semana anterior à pesquisa. Perguntou também sobre a distância (em quilômetros) que é percorrida ao longo do trajeto (supondo que cada aluno realize o mesmo trajeto diariamente), o número de semáforos pelos quais cada um passa e o período do dia em que cada estudante tem o hábito de se deslocar para a escola (manhã ou tarde). Parte do banco de dados elaborado encontra-se na Tabela 14.2.

Seguindo o que foi definido nos capítulos anteriores em relação à variável correspondente ao período do dia em que é realizado o trajeto, a categoria de referência 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, conforme apresentado na Tabela 14.2.

Tabela 14.2 Exemplo: quantidade de atrasos na semana x distância percorrida, quantidade de semáforos e período do dia para o trajeto até a escola.

Estudante	Quantidade de atrasos na última semana (Y_i)	Distância percorrida até a escola (quilômetros) (X_{1i})	Quantidade de semáforos (X_{2i})	Período do dia (X_{3i})
Gabriela	1	11	15	1 (manhã)
Patrícia	0	9	15	1 (manhã)
Gustavo	0	9	16	1 (manhã)
Letícia	3	10	16	0 (tarde)
Luiz Ovídio	2	12	18	1 (manhã)
Leonor	3	14	16	0 (tarde)
Dalila	1	10	15	1 (manhã)
Antônio	0	10	16	1 (manhã)
Júlia	2	10	18	1 (manhã)
Mariana	0	9	13	1 (manhã)
...				
Filomena	1	8	18	1 (manhã)
...				
Estela	0	8	13	1 (manhã)

A fim de que seja possível elaborar corretamente um modelo de regressão Poisson, devemos, inicialmente, verificar se a média da variável dependente (quantidade de atrasos) é igual à sua variância. Enquanto a Tabela 14.3 apresenta estas estatísticas, de onde se pode verificar que são muito próximas, a Figura 14.3 mostra o histograma da variável dependente do nosso exemplo.

Tabela 14.3 Média e variância da variável dependente (quantidade de atrasos na última semana).

Estatística	
Média	1,030
Variância	1,059

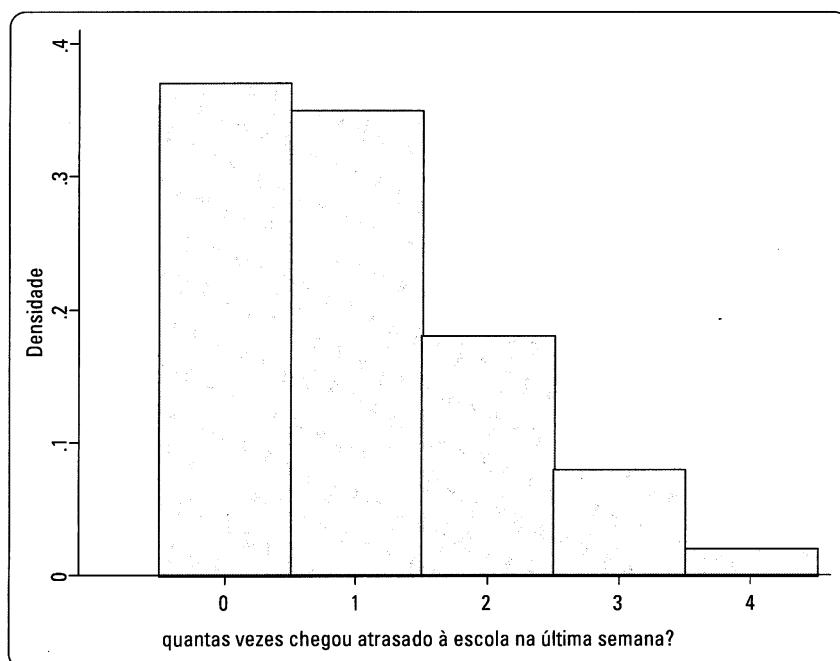


Figura 14.3 Histograma da variável dependente.

Dada a proximidade da média à variância da variável dependente, iremos optar por estimar um modelo para estudar o comportamento da incidência de atrasos à escola semanalmente, em função da distância percorrida, da quantidade de semáforos e do período do dia em que é realizado o trajeto, por meio da regressão Poisson.

Entretanto, caso a variância da variável dependente seja consideravelmente maior do que a sua média, a estimativa de um modelo Poisson poderá gerar parâmetros viesados, por conta do problema conhecido por **super-dispersão**. É sempre recomendável, portanto, que, após a estimativa de um modelo de regressão Poisson, seja elaborado um **teste para verificação da existência de superdispersão** (que será abordado na seção 14.2.4) e, caso sua presença seja detectada, será recomendada a estimativa de um modelo de regressão binomial negativo (seção 14.3).

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

Desta forma, com base na expressão (14.4), o modelo de regressão Poisson a ser estimado será:

$$\ln(\lambda_i) = \alpha + \beta_1 \cdot dist_i + \beta_2 \cdot sem_i + \beta_3 \cdot per_i$$

e a taxa média de incidência de atrasos semanalmente, para cada estudante, será dada, com base na expressão (14.5), por:

$$\lambda_i = e^{(\alpha + \beta_1 \cdot dist_i + \beta_2 \cdot sem_i + \beta_3 \cdot per_i)}$$

Assim como nos modelos de regressão logística binária e multinomial, os parâmetros de um modelo de regressão Poisson são estimados por máxima verossimilhança, em que a variável dependente segue uma distribuição Poisson. Sendo a probabilidade de ocorrência de uma específica contagem m em determinada exposição (no nosso exemplo, o período de uma semana) para uma observação i em uma amostra com n observações dada pela expressão (14.1), podemos definir a função de verossimilhança para modelos de regressão Poisson como sendo:

$$L = \prod_{i=1}^n \frac{e^{-\lambda_i} \cdot (\lambda_i)^{Y_i}}{Y_i!} \quad (14.6)$$

de onde vem que o logaritmo da função de verossimilhança (*log likelihood function*) pode ser escrito como:

$$LL = \sum_{i=1}^n [-\lambda_i + (Y_i) \cdot \ln(\lambda_i) - \ln(Y_i!)] \quad (14.7)$$

Portanto, podemos fazer a seguinte pergunta: **Quais os valores dos parâmetros do modelo proposto que fazem com que o valor de LL da expressão (14.7) 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 Poisson, 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:

$$LL = \sum_{i=1}^n [-\lambda_i + (Y_i) \cdot \ln(\lambda_i) - \ln(Y_i!)] = \text{máx} \quad (14.8)$$

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 **QuantAtrasosPoissonMáximaVerossimilhança.xls**, que servirá de auxílio para o cálculo dos parâmetros.

Neste arquivo, além da variável dependente e das variáveis explicativas, foram criadas duas novas variáveis, que correspondem, respectivamente, à taxa esperada semanal de incidência λ_i e ao logaritmo da função de verossimilhança LL_i para cada observação. A Tabela 14.4 mostra parte dos dados quando os parâmetros α, β_1, β_2 e β_3 forem iguais a 0.

Tabela 14.4 Cálculo de LL quando $\alpha = \beta_1 = \beta_2 = \beta_3 = 0$.

Estudante	Y_i	X_{1i}	X_{2i}	X_{3i}	λ_i	LL_i $-\lambda_i + (Y_i) \cdot \ln(\lambda_i) - \ln(Y_i!)$
Gabriela	1	11	15	1	1,00000	-1,00000
Patrícia	0	9	15	1	1,00000	-1,00000
Gustavo	0	9	16	1	1,00000	-1,00000
Letícia	3	10	16	0	1,00000	-2,79176
Luiz Ovídio	2	12	18	1	1,00000	-1,69315
Leonor	3	14	16	0	1,00000	-2,79176
Dalila	1	10	15	1	1,00000	-1,00000
Antônio	0	10	16	1	1,00000	-1,00000
Júlia	2	10	18	1	1,00000	-1,69315
Mariana	0	9	13	1	1,00000	-1,00000
...						
Filomena	1	8	18	1	1,00000	-1,00000
...						
Estela	0	8	13	1	1,00000	-1,00000
Somatória	$LL = \sum_{i=1}^{100} [-\lambda_i + (Y_i) \cdot \ln(\lambda_i) - \ln(Y_i!)]$					-133,16683

	A	B	C	D	E	F	G	H	I	J
1	Estudante	Atrasos (Y_i)	Distância (X_1)	Semáforos (X_2)	Período (X_3)	λ_i	LL_i			
2	Gabriela	1	11	15	1	1,00000	-1,00000			
3	Patrícia	0	9	15	1	1,00000	-1,00000	α	0,0000	
4	Gustavo	0	9	16	1	1,00000	-1,00000	β_1	0,0000	
5	Letícia	3	10	16	0	1,00000	-2,79176	β_2	0,0000	
6	Luiz Ovídio	2	12	18	1	1,00000	-1,69315	β_3	0,0000	
7	Leonor	3	14	16	0	1,00000	-2,79176			
8	Dalila	1	10	15	1	1,00000	-1,00000			
9	Antônio	0	10	16	1	1,00000	-1,00000			
10	Júlia	2	10	18	1	1,00000	-1,69315			
11	Mariana	0	9	13	1	1,00000	-1,00000			
12	Roberto	1	9	15	1	1,00000	-1,00000			
13	Renata	1	9	15	1	1,00000	-1,00000			
14	Guilherme	2	12	17	1	1,00000	-1,69315			
15	Rodrigo	1	9	12	1	1,00000	-1,00000			
16	Giulia	0	11	11	1	1,00000	-1,00000			
17	Felipe	2	9	17	1	1,00000	-1,69315			
18	Karina	1	11	14	1	1,00000	-1,00000			
19	Pietro	1	11	15	1	1,00000	-1,00000			
20	Cecília	0	11	15	1	1,00000	-1,00000			
21	Gisele	0	9	14	1	1,00000	-1,00000			
22	Elaine	1	11	13	1	1,00000	-1,00000			
23	Kamal	0	9	14	1	1,00000	-1,00000			
24	Rodolfo	0	11	15	1	1,00000	-1,00000			
25	Pilar	1	11	13	1	1,00000	-1,00000			
26	Vivian	2	13	16	1	1,00000	-1,69315			
27	Danielle	0	9	11	1	1,00000	-1,00000			
28	Juliana	0	9	16	1	1,00000	-1,00000			
101	Estela	0	8	13	1	1,00000	-1,00000			
102										
103							Somatória $LL_i = -133,16683$			

Figura 14.4 Dados do arquivo QuantAtrasosPoissonMáximaVerossimilhança.xls.

A Figura 14.4 apresenta parte dos dados presentes neste arquivo do Excel.

Como podemos verificar, quando $\alpha = \beta_1 = \beta_2 = \beta_3 = 0$, o valor da somatória do logaritmo da função de verossimilhança é igual a -133,16683. Entretanto, deve haver uma combinação ótima de valores dos parâmetros, de modo que a função-objetivo apresentada na expressão (14.8) 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 G103, que é a nossa célula de destino e que deverá ser maximizada. Além disso, os parâmetros α, β_1, β_2 e β_3 , cujos valores estão nas células J3, J5, J7 e J9, respectivamente, são as células variáveis. A janela do **Solver** ficará como mostra a Figura 14.5.

Ao clicarmos em **Resolver** e em **OK**, obteremos a solução ótima do problema de programação linear. A Tabela 14.5 apresenta parte dos resultados obtidos.

Logo, o valor máximo possível da somatória do logaritmo da função de verossimilhança é $LL_{máx} = -107,61498$. A resolução deste problema gerou as seguintes estimativas dos parâmetros:

$$\alpha = -4,3801$$

$$\beta_1 = 0,2221$$

$$\beta_2 = 0,1646$$

$$\beta_3 = -0,5731$$

e, assim, podemos escrever o nosso modelo log-linear estimado da seguinte forma:

$$\ln(\lambda_i) = -4,3801 + 0,2221.dist_i + 0,1646.sem_i - 0,5731.per_i$$

com taxa média de incidência de atrasos semanalmente dada, para cada estudante, por:

$$\lambda_i = e^{(-4,3801+0,2221.dist_i+0,1646.sem_i-0,5731.per_i)}$$

A Figura 14.6 apresenta parte dos resultados obtidos pela modelagem.

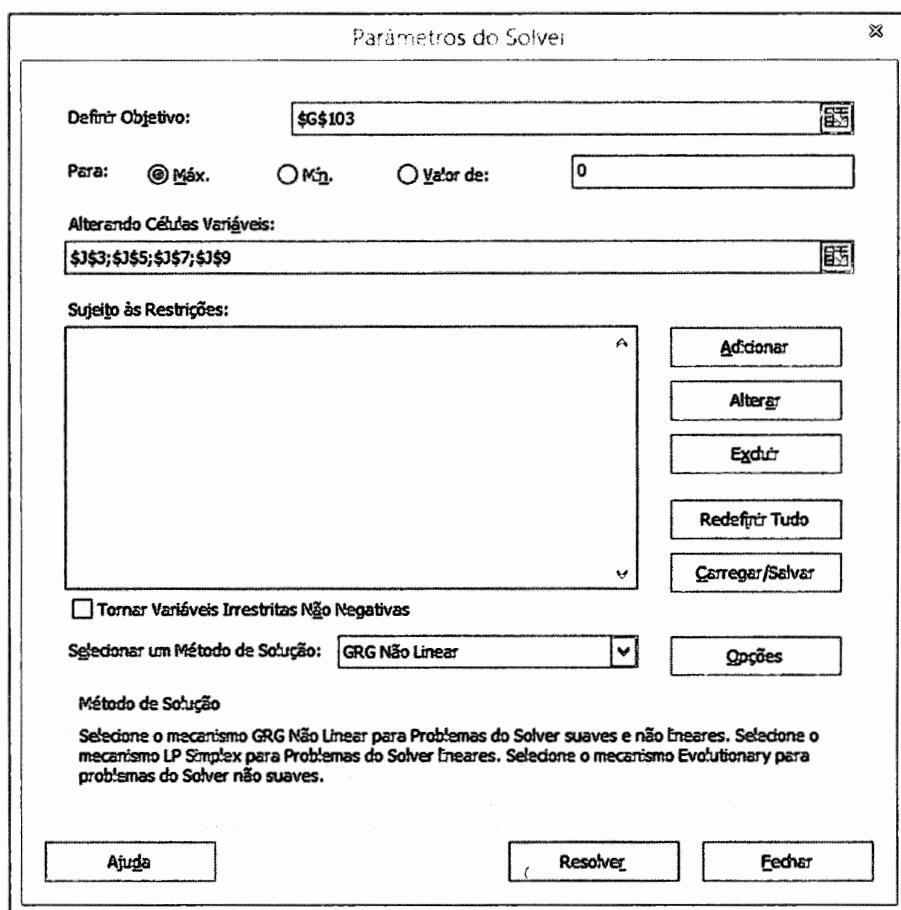


Figura 14.5 Solver – Maximização da somatória do logaritmo da função de verossimilhança.

Tabela 14.5 Valores obtidos quando da maximização de LL .

Estudante	Y_i	X_{1i}	X_{2i}	X_{3i}	λ_i	LL_i $-\lambda_i + (Y_i) \cdot \ln(\lambda_i) - \ln(Y_i!)$
Gabriela	1	11	15	1	0,96026	-1,00081
Patrícia	0	9	15	1	0,61581	-0,61581
Gustavo	0	9	16	1	0,72601	-0,72601
Letícia	3	10	16	0	1,60809	-1,97471
Luiz Ovídio	2	12	18	1	1,96485	-1,30717
Leonor	3	14	16	0	3,91008	-1,61117
Dalila	1	10	15	1	0,76899	-1,03167
Antônio	0	10	16	1	0,90659	-0,90659
Júlia	2	10	18	1	1,26006	-1,49089
Mariana	0	9	13	1	0,44306	-0,44306
...						
Filomena	1	8	18	1	0,80808	-1,02117
...						
Estela	0	8	13	1	0,35481	-0,35481
Somatória	$LL = \sum_{i=1}^{100} [-\lambda_i + (Y_i) \cdot \ln(\lambda_i) - \ln(Y_i!)]$					-107,61498

A	B	C	D	E	F	G	H	I	J
1	Estudante	Atrasos (Y)	Distância (X ₁)	Semáforos (X ₂)	Período (X ₃)	λ_i	LL _i		
2	Gabrielle	1	11	15	1	0,96026	-1,00081	α	4,3801
3	Patrícia	0	9	15	1	0,61581	-0,61581	β_1	0,2221
4	Gustavo	0	9	16	1	0,72601	-0,72601	β_2	0,1646
5	Leticia	3	10	16	0	1,60809	-1,97471	β_3	-0,5731
6	Luiz Ovídio	2	12	18	1	1,96485	-1,30717		
7	Leonor	3	14	16	0	3,91008	-1,61117		
8	Dalila	1	10	15	1	0,76899	-1,03167		
9	Antônio	0	10	16	1	0,90659	-0,90659		
10	Júlia	2	10	18	1	1,26006	-1,49089		
11	Mariana	0	9	13	1	0,44306	-0,44306		
12	Roberto	1	9	15	1	0,61581	-1,10062		
13	Renata	1	9	15	1	0,61581	-1,10062		
14	Guilherme	2	12	17	1	1,66663	-1,33817		
15	Rodrigo	1	9	12	1	0,37582	-1,35447		
16	Giulia	0	11	11	1	0,49708	-0,49708		
17	Felipe	2	9	17	1	0,85592	-1,86023		
18	Karina	1	11	14	1	0,81451	-1,01968		
19	Pietro	1	11	15	1	0,96026	-1,00081		
20	Cecília	0	11	15	1	0,96026	-0,96026		
21	Gisele	0	9	14	1	0,52235	-0,52235		
22	Elaine	1	11	13	1	0,69088	-1,06067		
23	Kamal	0	9	14	1	0,52235	-0,52235		
24	Rodolfo	0	11	15	1	0,96026	-0,96026		
25	Pilar	1	11	13	1	0,69088	-1,06067		
26	Vivian	2	13	16	1	1,76529	-1,32181		
27	Danielle	0	9	11	1	0,31878	-0,31878		
28	Juliana	0	9	16	1	0,72601	-0,72601		
101	Estela	0	8	13	1	0,35481	-0,35481		
102									
103							Somatória LL _i	-107,61498	

Figura 14.6 Obtenção dos parâmetros quando da maximização de LL pelo Solver.

Estimados os parâmetros do modelo de regressão Poisson, podemos propor quatro interessantes perguntas:

Qual é a quantidade média esperada de atrasos na semana quando se desloca 12 quilômetros e se passa por 17 semáforos diariamente, sendo o trajeto feito à tarde?

Em média, em quanto se altera a taxa de incidência semanal de atrasos ao se adotar um percurso 1 quilômetro mais longo, mantidas as demais condições constantes?

Em média, em quanto se altera a taxa de incidência semanal de atrasos ao se passar por 1 semáforo a mais no percurso até a escola, mantidas as demais condições constantes?

Em média, em quanto se altera a taxa de incidência semanal de atrasos ao se optar por ir à escola de manhã, em vez de se ir à tarde, 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 sejam apresentados apenas parâmetros estatisticamente significantes para, a partir de então, ser possível a elaboração de inferências e previsões.

Portanto, tendo sido elaborada a estimativa por máxima verossimilhança dos parâmetros da equação da taxa média de incidência de atrasos semanalmente, partiremos para o estudo da significância estatística geral do modelo obtido, bem como das significâncias estatísticas dos parâmetros, de forma análoga ao realizado nos capítulos anteriores.

14.2.2. Significância estatística geral e dos parâmetros do modelo de regressão Poisson

Assim como para os modelos de regressão logística binária e multinomial, para os modelos de regressão Poisson pode ser calculado o pseudo R^2 de McFadden, dado pela seguinte expressão:

$$\text{pseudo } R^2 = \frac{-2 \cdot LL_0 - (-2 \cdot LL_{\max})}{-2 \cdot LL_0} \quad (14.9)$$

e cuja utilidade é bastante limitada e restringe-se a casos em que o pesquisador tiver interesse em escolher um determinado modelo em detrimento de outros, prevalecendo aquele que apresentar o maior pseudo R^2 de McFadden.

Seguindo a mesma lógica proposta no capítulo anterior, iremos inicialmente calcular LL_0 , que é dado pelo valor máximo da somatória do logaritmo da função de verossimilhança para um modelo em que há apenas a constante α , conhecido por **modelo nulo**. Por meio do mesmo procedimento elaborado na seção 14.2.1,

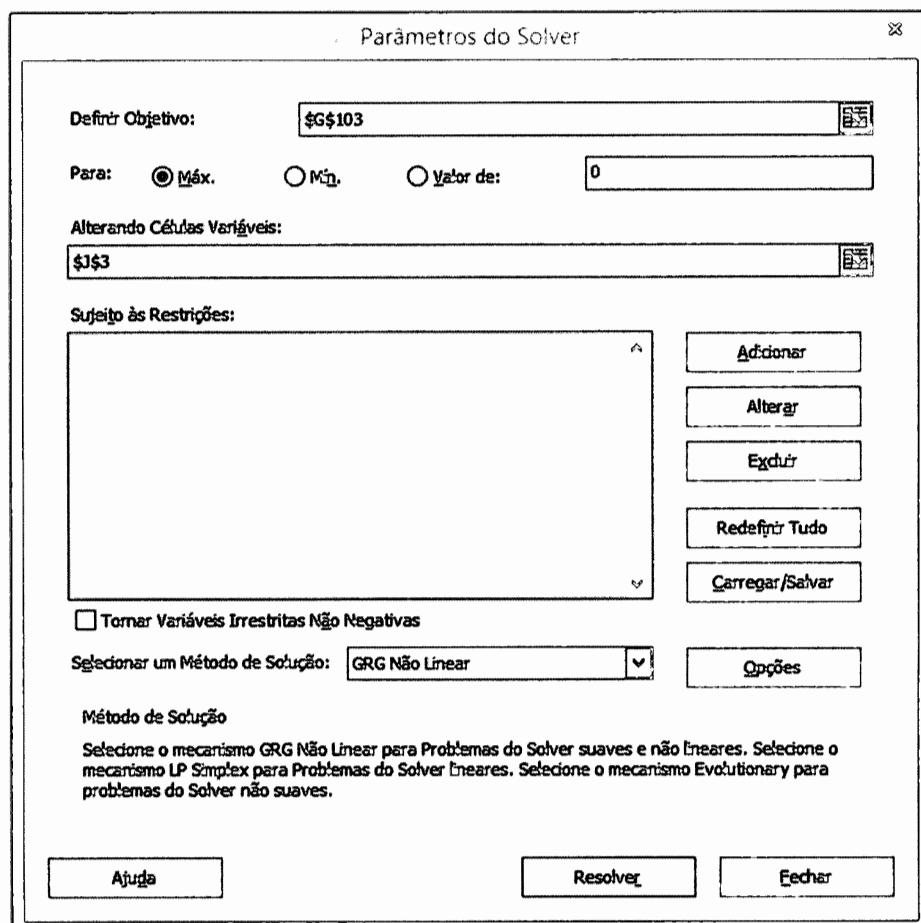


Figura 14.7 Solver – Maximização da somatória do logaritmo da função de verossimilhança para o modelo nulo.

porém agora utilizando o arquivo **QuantAtrasosPoissonMáximaVerossimilhançaModeloNulo.xls**, obtemos $LL_0 = -133,12228$. As Figuras 14.7 e 14.8 mostram, respectivamente, a janela do **Solver** e parte dos resultados obtidos pela modelagem neste arquivo.

No nosso exemplo, conforme já discutimos na seção anterior e já calculamos por meio do **Solver** do Excel, $LL_{\text{máx}}$, que é o valor máximo possível da somatória do logaritmo da função de verossimilhança, é igual a $-107,61498$.

Logo, com base na expressão (14.9), obteremos:

$$\text{pseudo } R^2 = \frac{-2.(-133,12228) - [(-2.(-107,61498))]}{-2.(-133,12228)} = 0,1916$$

Conforme discutimos, um maior pseudo R^2 de McFadden pode ser utilizado como critério para escolha de um modelo em detrimento de outro. Entretanto, não é adequado para avaliar o percentual de variância da variável dependente que é explicado pelo conjunto de variáveis explicativas consideradas no modelo.

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 14.4 e 14.5, respectivamente.

Analogamente ao procedimento apresentado nos capítulos anteriores, inicialmente iremos estudar a significância estatística geral do modelo que está sendo proposto. O teste χ^2 propicia condições à verificação da significância do modelo, uma vez que suas hipóteses nula e alternativa, para um modelo de regressão Poisson, são, respectivamente:

$$\begin{aligned} H_0: \beta_1 &= \beta_2 = \dots = \beta_k = 0 \\ H_1: \text{existe pelo menos um } \beta_j &\neq 0 \end{aligned}$$

A	B	C	D	E	F	G	H	I	J
1	Estudante	Atrasos (Y)	Distância (X ₁)	Semáforos (X ₂)	Período (X ₃)	λ_i	LL _i		
2	Gabriela	1	11	15	1	1,03000	-1,00044		
3	Patrícia	0	9	15	1	1,03000	-1,03000		
4	Gustavo	0	9	16	1	1,03000	-1,03000		
5	Letícia	3	10	16	0	1,03000	-2,73308		
6	Luiz Ovídio	2	12	18	1	1,03000	-1,66403		
7	Leonor	3	14	16	0	1,03000	-2,73308		
8	Dalila	1	10	15	1	1,03000	-1,00044		
9	Antônio	0	10	16	1	1,03000	-1,03000		
10	Julia	2	10	18	1	1,03000	-1,66403		
11	Mariana	0	9	13	1	1,03000	-1,03000		
12	Roberto	1	9	15	1	1,03000	-1,00044		
13	Renata	1	9	15	1	1,03000	-1,00044		
14	Guilherme	2	12	17	1	1,03000	-1,66403		
15	Rodrigo	1	9	12	1	1,03000	-1,00044		
16	Giulia	0	11	11	1	1,03000	-1,03000		
17	Felipe	2	9	17	1	1,03000	-1,66403		
18	Karina	1	11	14	1	1,03000	-1,00044		
19	Pietro	1	11	15	1	1,03000	-1,00044		
20	Cecília	0	11	15	1	1,03000	-1,03000		
21	Gisele	0	9	14	1	1,03000	-1,03000		
22	Elaine	1	11	13	1	1,03000	-1,00044		
23	Kamal	0	9	14	1	1,03000	-1,03000		
24	Rodolfo	0	11	15	1	1,03000	-1,03000		
25	Pilar	1	11	13	1	1,03000	-1,00044		
26	Vivian	2	13	16	1	1,03000	-1,66403		
27	Danielle	0	9	11	1	1,03000	-1,03000		
28	Juliana	0	9	16	1	1,03000	-1,03000		
101	Estela	0	8	13	1	1,03000	-1,03000		
102									
103							Somatória L	-133,12228	

Figura 14.8 Obtenção dos parâmetros quando da maximização de LL pelo Solver – modelo nulo.

Conforme já discutimos no capítulo anterior, o teste χ^2 é adequado para se avaliar a significância conjunta dos parâmetros do modelo quando este for estimado pelo método de máxima verossimilhança, como nos casos dos modelos de regressão logística binária e multinomial e de regressão para dados de contagem.

O teste χ^2 propicia ao pesquisador uma verificação inicial sobre a existência do modelo que está sendo proposto, uma vez que, se todos os parâmetros estimados β_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 taxa de incidência do fenômeno em estudo. Conforme também já apresentado no capítulo anterior, a estatística χ^2 tem a seguinte expressão:

$$\chi^2 = -2 \cdot (LL_0 - LL_{\max}) \quad (14.10)$$

Voltando ao nosso exemplo, temos que:

$$\chi^2_{3g.l.} = -2 \cdot [-133,12228 - (-107,61498)] = 51,0146$$

Para 3 graus de liberdade (número de variáveis explicativas consideradas na modelagem, ou seja, número de parâmetros β), temos, por meio da Tabela D do apêndice do livro, que o $\chi^2_c = 7,815$ (χ^2 crítico para 3 graus de liberdade e para o nível de significância de 5%). Desta forma, como o χ^2 calculado $\chi^2_{cal} = 51,0146 > \chi^2_c = 7,815$, podemos rejeitar a hipótese nula de que todos os parâmetros β_j ($j = 1, 2, 3$) sejam estatisticamente iguais a zero. Logo, pelo menos uma variável X é estatisticamente significante para explicar a incidência de atrasos à escola semanalmente e teremos um modelo de regressão Poisson estatisticamente significante para fins de previsão.

Softwares como o Stata e o SPSS não oferecem o χ^2_c para os graus de liberdade definidos e um determinado nível de significância. Todavia, oferecem o nível de significância do χ^2_{cal} para estes graus de liberdade. Desta forma, em vez de analisarmos se $\chi^2_{cal} > \chi^2_c$, devemos verificar se o nível de significância do χ^2_{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. χ^2_{cal} ou Prob. χ^2_{cal}) $< 0,05$, existe pelo menos um $\beta_j \neq 0$.

Na sequência, é preciso que o pesquisador avalie se cada um dos parâmetros do modelo de regressão Poisson é estatisticamente significante e, neste sentido, a estatística z de Wald será importante para fornecer a significância estatística de cada parâmetro a ser considerado no modelo. Conforme já discutido no capítulo anterior, a

nomenclatura z refere-se ao fato de que a distribuição desta estatística é a distribuição normal padrão, e as hipóteses do teste z de Wald para o α e para cada β_j ($j = 1, 2, \dots, k$) são, respectivamente:

$$H_0: \alpha = 0$$

$$H_1: \alpha \neq 0$$

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

As expressões para o cálculo das estatísticas z de Wald de cada parâmetro α e β_j são dadas, respectivamente, por:

$$z_\alpha = \frac{\alpha}{s.e.(\alpha)} \quad (14.11)$$

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

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 McCullagh e Nelder (1989). Os valores de $s.e.$ de cada parâmetro, para o nosso exemplo, são:

$$s.e.(\alpha) = 1,160$$

$$s.e.(\beta_1) = 0,066$$

$$s.e.(\beta_2) = 0,046$$

$$s.e.(\beta_3) = 0,262$$

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

$$z_\alpha = \frac{\alpha}{s.e.(\alpha)} = \frac{-4,3801}{1,160} = -3,776$$

$$z_{\beta_1} = \frac{\beta_1}{s.e.(\beta_1)} = \frac{0,2221}{0,066} = 3,365$$

$$z_{\beta_2} = \frac{\beta_2}{s.e.(\beta_2)} = \frac{0,1646}{0,046} = 3,580$$

$$z_{\beta_3} = \frac{\beta_3}{s.e.(\beta_3)} = \frac{-0,5731}{0,262} = -2,187$$

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.

Conforme discutimos no capítulo anterior, 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).

Assim como no caso do teste χ^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 $Sig. z_{\text{al}}$ ou $Prob. z_{\text{al}}$) $< 0,05$ para $\alpha, \alpha \neq 0$
e

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

Sendo assim, como todos os valores de $z_{\text{al}} < -1,96$ ou $> 1,96$, os *valores-P* das estatísticas z de Wald $< 0,05$ para todos os parâmetros estimados e, portanto, já chegamos ao modelo final de regressão Poisson, sem que haja a necessidade de uma eventual aplicação do procedimento *Stepwise* estudado nos capítulos anteriores. Logo, a taxa média estimada de atrasos por semana para determinado aluno i é dada por:

$$\lambda_i = e^{(-4,3801+0,2221.\text{dist}_i+0,1646.\text{sem}_i-0,5731.\text{per}_i)}$$

e, desta forma, podemos retornar às nossas quatro importantes perguntas, respondendo uma de cada vez:

Qual é a quantidade média esperada de atrasos na semana quando se desloca 12 quilômetros e se passa por 17 semáforos diariamente, sendo o trajeto feito à tarde?

Fazendo uso da expressão da taxa média estimada de atrasos em uma semana e substituindo os valores fornecidos nesta equação, teremos:

$$\lambda = e^{[-4,3801+0,2221.(12)+0,1646.(17)-0,5731.(0)]} = 2,95$$

Logo, espera-se que determinado aluno que é submetido a estas características ao se deslocar à escola apresente, em média, uma quantidade de 2,95 atrasos por semana. Como a variável *atrasos* é discreta, dificilmente existirão observações em modelos de regressão Poisson com termos de erro com valores inteiros ou até mesmo iguais a zero.

Em média, em quanto se altera a taxa de incidência semanal de atrasos ao se adotar um percurso 1 quilômetro mais longo, mantidas as demais condições constantes?

Fazendo uso da mesma expressão, temos que:

$$e^{0,2221} = 1,249$$

Logo, mantidas as demais condições constantes, a taxa de incidência semanal de atrasos ao se adotar um percurso 1 quilômetro mais longo é, em média, multiplicada por um fator de 1,249, ou seja, é, em média, 24,9% maior.

Em média, em quanto se altera a taxa de incidência semanal de atrasos ao se passar por 1 semáforo a mais no percurso até a escola, mantidas as demais condições constantes?

Neste caso, teremos:

$$e^{0,1646} = 1,179$$

Logo, mantidas as demais condições constantes, a taxa de incidência semanal de atrasos ao se adotar um percurso com 1 semáforo a mais é, em média, multiplicada por um fator de 1,179, ou seja, é, em média, 17,9% maior.

Em média, em quanto se altera a taxa de incidência semanal de atrasos ao se optar por ir à escola de manhã, em vez de se ir à tarde, mantidas as demais condições constantes?

Neste último caso, teremos:

$$e^{-0,5731} = 0,564$$

Logo, mantidas as demais condições constantes, a taxa de incidência semanal de atrasos ao se optar por ir à escola de manhã, em vez de se ir à tarde, é, em média, multiplicada por um fator de 0,564, ou seja, é, em média, 43,6% menor.

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

14.2.3. Construção dos intervalos de confiança dos parâmetros do modelo de regressão Poisson

Igualmente ao apresentado no capítulo anterior, os intervalos de confiança dos coeficientes da expressão (14.4), para os parâmetros α e β_j ($j = 1, 2, \dots, k$), ao nível de confiança de 95%, podem ser escritos, respectivamente, da seguinte forma:

$$\alpha \pm 1,96. [\text{s.e.}(\alpha)] \quad (14.12)$$

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

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

Tabela 14.6 Cálculo dos intervalos de confiança dos parâmetros.

Parâmetro	Coeficiente	Erro-Padrão (s.e.)	z	Intervalo de Confiança (95%)	
				$\alpha - 1,96 \cdot [s.e.(\alpha)]$	$\alpha + 1,96 \cdot [s.e.(\alpha)]$
				$\beta_j - 1,96 \cdot [s.e.(\beta_j)]$	$\beta_j + 1,96 \cdot [s.e.(\beta_j)]$
α (constante)	-4,3801	1,160	-3,776	-6,654	-2,106
β_1 (variável <i>dist</i>)	0,2221	0,066	3,365	0,093	0,351
β_2 (variável <i>sem</i>)	0,1646	0,046	3,580	0,074	0,254
β_3 (variável <i>per</i>)	-0,5731	0,262	-2,187	-1,086	-0,060

Assim sendo, podemos elaborar a Tabela 14.6, que traz os coeficientes estimados dos parâmetros da expressão log-linear 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 estimarmos este modelo de regressão Poisson por meio do Stata e do SPSS (seções 14.4 e 14.5, respectivamente).

Com base nos intervalos de confiança dos parâmetros, podemos escrever as expressões dos limites inferior (mínimo) e superior (máximo) do modelo log-linear de regressão Poisson, com 95% de confiança. Assim, teremos:

$$\ln(\lambda_i)_{\min} = -6,654 + 0,093.dist_i + 0,074.sem_i - 1,086.per_i$$

$$\ln(\lambda_i)_{\max} = -2,106 + 0,351.dist_i + 0,254.sem_i - 0,060.per_i$$

A partir da expressão (14.5), o intervalo de confiança da taxa estimada de incidência do fenômeno em estudo (*incidence rate ratio*, ou *irr*) correspondente à alteração em cada parâmetro β_j ($j = 1, 2, \dots, k$), ao nível de confiança de 95%, pode ser escrito da seguinte forma:

$$e^{\beta_j \pm 1,96 \cdot [s.e.(\beta_j)]} \quad (14.13)$$

Note que não apresentamos a expressão do intervalo de confiança da taxa de incidência correspondente ao parâmetro α , uma vez que só faz sentido discutirmos a mudança na taxa de incidência do fenômeno em estudo quando é alterada em uma unidade determinada variável explicativa do modelo, mantidas todas as demais condições constantes.

Para os dados do nosso exemplo e com base nos valores da Tabela 14.6, vamos, então, elaborar a Tabela 14.7, que apresenta os intervalos de confiança da taxa de incidência do fenômeno de interesse para cada parâmetro β_j .

Estes valores também poderão ser obtidos por meio do Stata e do SPSS, conforme mostraremos, respectivamente, nas seções 14.4 e 14.5.

Conforme já discutido nos capítulos anteriores, se o intervalo de confiança de um determinado parâmetro contiver o zero (ou da taxa de incidência 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 α ,

Tabela 14.7 Cálculo dos intervalos de confiança da taxa de incidência λ (*irr*) para cada parâmetro β_j .

Parâmetro	Taxa de Incidência λ (<i>irr</i>)	Intervalo de Confiança de λ (95%)	
	e^{β_j}	$e^{\beta_j - 1,96 \cdot [s.e.(\beta_j)]}$	$e^{\beta_j + 1,96 \cdot [s.e.(\beta_j)]}$
β_1 (variável <i>dist</i>)	1,249	1,097	1,421
β_2 (variável <i>sem</i>)	1,179	1,078	1,289
β_3 (variável <i>per</i>)	0,564	0,337	0,942

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 β_j contiver o zero (o que não aconteceu neste nosso exemplo), este deverá ser excluído do modelo final quando da elaboração do procedimento *Stepwise*.

Da mesma forma que para os modelos de regressão logística, a rejeição da hipótese nula para um determinado parâmetro β_j , a um especificado nível de significância, indica que a correspondente variável X é significativa para explicar a taxa de incidência do fenômeno em estudo e, consequentemente, deve permanecer no modelo final de regressão para dados de contagem. Podemos, portanto, concluir que a decisão pela exclusão de determinada variável X em um modelo de regressão para dados de contagem pode ser realizada por meio da análise direta da estatística z de Wald de seu respectivo parâmetro β_j (se $-z_{\alpha/2} < z_{cal} < z_{\alpha/2} \rightarrow \text{valor-}P > 0,05 \rightarrow$ 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 14.1 apresenta os critérios de inclusão ou exclusão de parâmetros β_j ($j = 1, 2, \dots, k$) em modelos de regressão para dados de contagem.

Quadro 14.1 Decisão de inclusão de parâmetros β_j em modelos de regressão para dados de contagem.

Parâmetro	Estatística z de Wald (para nível de significância α)	Teste z (análise do valor- P para nível de significância α)	Análise pelo Intervalo de Confiança	Decisão
β_j	$-z_{\alpha/2} < z_{cal} < z_{\alpha/2}$	$\text{valor-}P >$ nível de sig. α	O intervalo de confiança contém o zero	Excluir o parâmetro do modelo
	$z_{cal} > z_{\alpha/2}$ ou $z_{cal} < -z_{\alpha/2}$	$\text{valor-}P <$ nível de sig. α	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 $\alpha = 5\%$.

14.2.4. Teste para verificação de superdispersão em modelos de regressão Poisson

Cameron e Trivedi (1990) propõem um interessante procedimento para verificação da existência de superdispersão em modelos de regressão Poisson. Para tanto, é preciso que seja gerada uma variável Y^* , da seguinte maneira:

$$Y_i^* = \frac{[(Y_i - \lambda_i)^2 - Y_i]}{\lambda_i} \quad (14.14)$$

em que λ_i é o número esperado de ocorrências para cada observação da amostra após a estimativa do modelo de regressão Poisson e $(Y_i - \lambda_i)$ é a diferença entre o número real de ocorrências e o número previsto de ocorrências para cada observação (equivale ao termo de erro da regressão múltipla).

A Tabela 14.8 apresenta parte do banco de dados com a variável Y^* . Para fins didáticos, criamos um arquivo específico em Excel para que seja elaborado este teste, nomeado de **QuantAtrasosPoissonTesteSuperdispersão.xls**.

Após a geração de Y^* , devemos estimar o seguinte modelo auxiliar de regressão simples, sem a constante:

$$\hat{Y}_i^* = \beta \cdot \lambda_i \quad (14.15)$$

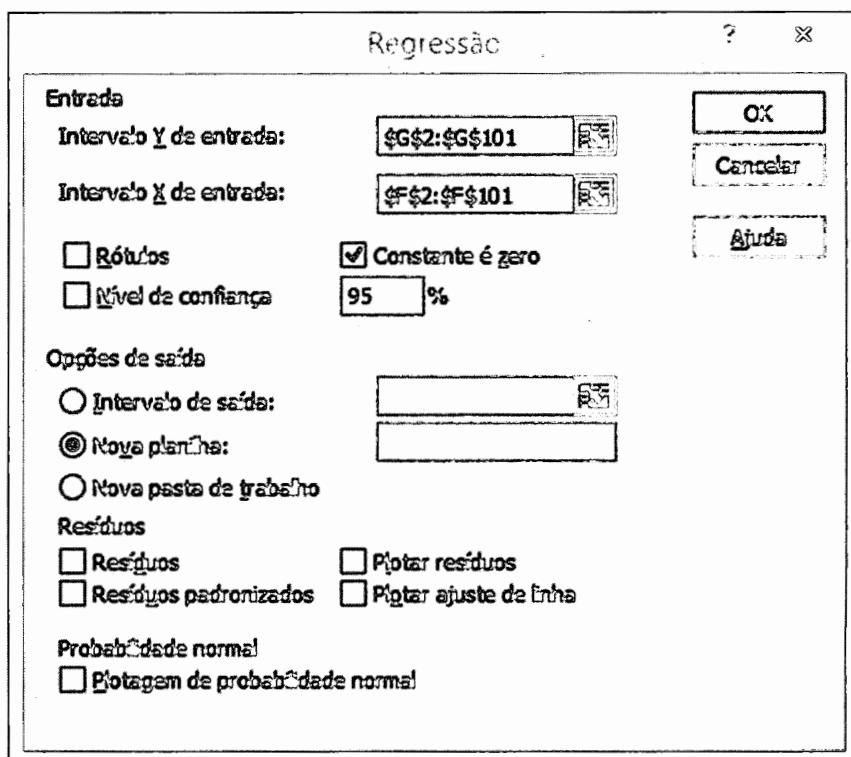
Cameron e Trivedi (1990) destacam que, se ocorrer o fenômeno da superdispersão nos dados, o parâmetro β estimado por meio do modelo representado pela expressão (14.15) será estatisticamente diferente de zero, a um determinado nível de significância.

Vamos, então, estimar a regressão auxiliar proposta, clicando em **Dados → Análise de Dados → Regressão → OK**. Na caixa de diálogo para inserção dos dados, devemos inserir as variáveis Y^* e λ , conforme mostra a Figura 14.9. Não devemos nos esquecer de marcar a opção **Constante é zero**.

Na sequência, vamos clicar em **OK**. O *output* desejado desta estimativa encontra-se na Figura 14.10.

Tabela 14.8 Cálculo da variável Y^* .

Estudante	Y_i	λ_i	$Y_i^* = \frac{[(Y_i - \lambda_i)^2 - Y_i]}{\lambda_i}$
Gabriela	1	0,96026	-1,03974
Patrícia	0	0,61581	0,61581
Gustavo	0	0,72601	0,72601
Letícia	3	1,60809	-0,66077
Luiz Ovídio	2	1,96485	-1,01726
Leonor	3	3,91008	-0,55542
Dalila	1	0,76899	-1,23101
Antônio	0	0,90659	0,90659
Júlia	2	1,26006	-1,15271
Mariana	0	0,44306	0,44306
...			
Filomena	1	0,80808	-1,19192
...			
Estela	0	0,35481	0,35481

**Figura 14.9** Caixa de diálogo para elaboração de regressão auxiliar no Excel – teste para verificação de existência de superdispersão.

	Coeficiente	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores
Variável X 1	-0,29175	0,15835	-1,84245	0,06840	-0,60596	0,02245

Figura 14.10 Resultado do teste para verificação de existência de superdispersão.

Como o *valor-P* do teste *t* correspondente ao parâmetro β da variável λ é maior do que 0,05, podemos afirmar que os dados da variável dependente **não apresentam superdispersão**, fazendo com que o modelo de regressão Poisson estimado seja adequado pela **presença de equidispersão nos dados**. Se não fosse esse o caso, deveríamos partir para a estimação de um modelo de regressão binomial negativo, a ser discutido na próxima seção.

14.3. O MODELO DE REGRESSÃO BINOMIAL NEGATIVO

Conforme discutimos, os modelos de regressão binomial negativo também são enquadrados nos chamados modelos de regressão para dados de contagem, sendo apropriados para estimação quando a variável dependente for quantitativa e com valores inteiros e não negativos (dados de contagem) e quando houver superdispersão nos dados.

Oliveira (2011) enfatiza que o interesse em se contar o número de ensaios necessários para que seja obtido o número desejado de ocorrências pode conduzir a uma distribuição binomial negativa, conforme discutimos no Capítulo 5. Segundo Lord e Park (2008), esta distribuição, primeiramente derivada por Greenwood e Yule (1920), é também conhecida por distribuição Poisson-Gama por ser uma combinação de duas distribuições que foi desenvolvida para levar em consideração o fenômeno da superdispersão que é comumente observado em dados de contagem. Ainda segundo os autores, leva este nome por aplicar o teorema binomial com um expoente negativo.

Se, por exemplo, a média do número de ocorrências de uma distribuição Poisson possuir uma parcela aleatória, a expressão (14.5) passará ser escrita da seguinte maneira:

$$\lambda_i = e^{(\alpha + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \dots + \beta_k \cdot X_{ki} + \varepsilon_i)} \quad (14.16)$$

de onde vem que:

$$\lambda_i = e^{(\alpha + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \dots + \beta_k \cdot X_{ki})} \cdot e^{(\varepsilon_i)} \quad (14.17)$$

que pode ser escrita como:

$$\lambda_i = u_i \cdot v_i \quad (14.18)$$

e que possui uma distribuição binomial negativa, em que o primeiro termo (u_i) representa o valor esperado de ocorrências e possui uma distribuição Poisson e o segundo termo (v_i) corresponde à parcela aleatória do número de ocorrências da variável dependente e possui uma distribuição Gama.

Para determinada observação i ($i = 1, 2, \dots, n$, em que n é o tamanho da amostra), a função da distribuição de probabilidade da variável v_i será dada por:

$$p(v_i) = \frac{\delta^\psi \cdot v_i^{\psi-1} \cdot e^{-\delta}}{\Gamma(\psi)}, \quad v_i = 0, 1, 2, \dots \quad (14.19)$$

em que ψ é chamado de parâmetro de forma ($\psi > 0$), δ é chamado de parâmetro de taxa ($\delta > 0$) e, para $\psi > 0$ e inteiro, $\Gamma(\psi)$ pode ser aproximado por $(\psi - 1)!$.

Com distribuição Gama, teremos, para a variável v , que:

- **Média:**

$$E(v) = \frac{\psi}{\delta} \quad (14.20)$$

- **Variância:**

$$Var(v) = \frac{\psi}{\delta^2} \quad (14.21)$$

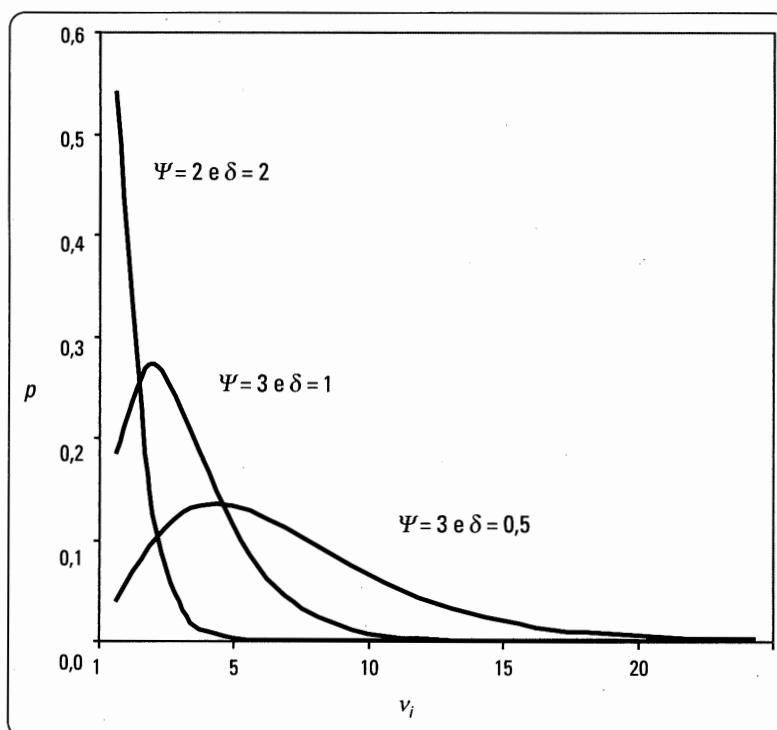
Analogamente ao realizado na seção 14.2, podemos elaborar, com base na expressão (14.19), uma tabela com valores de p em função de valores de v_i (Tabela 14.9), variando-se v_i de 1 a 20 e com três diferentes combinações de ψ e δ .

A partir dos dados calculados na Tabela 14.9, podemos elaborar o gráfico da Figura 14.11.

Apenas como curiosidade, a distribuição χ^2 é um caso particular da distribuição Gama quando $\psi = 0,5$ e $\delta = k/2$, em que k é um número inteiro e positivo.

Tabela 14.9 Distribuição Gama – funções de probabilidade de v_i para diferentes valores de ψ e δ .

v_i	$\psi = 2 \text{ e } \delta = 2$	$\psi = 3 \text{ e } \delta = 1$	$\psi = 3 \text{ e } \delta = 0,5$
	$p(v_i) = \frac{\delta^\psi \cdot v_i^{\psi-1} \cdot e^{-v_i \cdot \delta}}{\Gamma(\psi)}$		
1	0,5413	0,1839	0,0379
2	0,1465	0,2707	0,0920
3	0,0297	0,2240	0,1255
4	0,0054	0,1465	0,1353
5	0,0009	0,0842	0,1283
6	0,0001	0,0446	0,1120
7	0,0000	0,0223	0,0925
8	0,0000	0,0107	0,0733
9	0,0000	0,0050	0,0562
10	0,0000	0,0023	0,0421
11	0,0000	0,0010	0,0309
12	0,0000	0,0004	0,0223
13	0,0000	0,0002	0,0159
14	0,0000	0,0001	0,0112
15	0,0000	0,0000	0,0078
16	0,0000	0,0000	0,0054
17	0,0000	0,0000	0,0037
18	0,0000	0,0000	0,0025
19	0,0000	0,0000	0,0017
20	0,0000	0,0000	0,0011

**Figura 14.11** Distribuição Gama – gráficos das funções de probabilidade para diferentes valores de ψ e δ .

Fazendo uso da expressão (14.18), podemos transformar a função de probabilidade da distribuição Gama apresentada na expressão (14.19) como uma função do valor esperado de ocorrências da distribuição Poisson (u_i), de modo que:

$$p(u_i) = \frac{\left(\frac{\psi}{u_i}\right)^\psi \cdot \lambda_i^{\psi-1} \cdot e^{-\frac{\lambda_i \cdot \psi}{u_i}}}{\Gamma(\psi)} \quad (14.22)$$

Segundo Lord e Park (2008), podemos combinar as expressões (14.1) e (14.22), de modo a gerar a função da probabilidade de uma distribuição binomial negativa, o que nos permitirá calcular a probabilidade de ocorrência de uma contagem m , dada determinada exposição. Desta forma, teremos:

$$p(Y_i = m) = \int_0^{\infty} \frac{e^{(-\lambda_i)} \cdot \lambda_i^m}{m!} \cdot \frac{\left(\frac{\psi}{u_i}\right)^\psi \cdot \lambda_i^{\psi-1} \cdot e^{-\frac{\lambda_i \cdot \psi}{u_i}}}{\Gamma(\psi)} d\lambda_i \quad (14.23)$$

de onde vem que:

$$p(Y_i = m) = \frac{\Gamma(m + \psi)}{\Gamma(m + 1) \cdot \Gamma(\psi)} \cdot \left(\frac{\psi}{u_i + \psi} \right)^\psi \cdot \left(\frac{u_i}{u_i + \psi} \right)^m, \quad m = 0, 1, 2, \dots \quad (14.24)$$

que também pode ser escrita como:

$$p(Y_i = m) = \binom{m + \psi - 1}{\psi - 1} \cdot \left(\frac{\psi}{u_i + \psi} \right)^\psi \cdot \left(\frac{u_i}{u_i + \psi} \right)^m, \quad m = 0, 1, 2, \dots \quad (14.25)$$

que representa a função de probabilidade da distribuição binomial negativa para a ocorrência de uma contagem m , com as seguintes estatísticas:

- **Média:**

$$E(Y) = u \quad (14.26)$$

- **Variância:**

$$Var(Y) = u + \phi \cdot u^2 \quad (14.27)$$

em que $\phi = \frac{1}{\psi}$.

Desta forma, o segundo termo da expressão de variância da distribuição binomial negativa representa a superdispersão e, caso verifiquemos que $\phi \rightarrow 0$, este fenômeno não estará presente nos dados, podendo ser estimado um modelo de regressão Poisson, já que a média da variável dependente será igual à sua variância. Entretanto, caso ϕ seja estatisticamente maior do que zero, a existência de superdispersão faz com que deva ser estimado um modelo de regressão binomial negativo. Na seção 14.3.1, o parâmetro ϕ será estimado juntamente com os parâmetros do modelo de regressão binomial negativo por meio da maximização da somatória do logaritmo da função de verossimilhança, que ainda será definida, com o uso da ferramenta **Solver** do Excel. É importante ressaltarmos que softwares como o Stata e o SPSS estimam o valor de ϕ (inverso do parâmetro de forma ψ) e apresentam o seu intervalo de confiança, a partir do qual se torna possível avaliarmos se o mesmo é ou não estatisticamente igual a zero, conforme estudaremos, respectivamente, nas seções 14.4 e 14.5.

O modelo de regressão binomial negativo a ser estimado neste capítulo é também conhecido por **modelo de regressão NB2 (negative binomial 2 regression model)**, dada a especificação quadrática da variância apresentada na expressão (14.27). Entretanto, existem trabalhos que utilizam a expressão de variância como sendo apenas:

$$Var(Y) = u + \phi \cdot u \quad (14.28)$$

e, desta forma, o modelo estimado é conhecido por **modelo de regressão NB1 (negative binomial 1 regression model)**, porém, segundo Cameron e Trivedi (2009), os modelos de regressão NB2, com especificação quadrática da variância, são preferíveis aos modelos de regressão NB1 por frequentemente apresentarem melhores aproximações às funções mais gerais de variância.

Com base nas expressões (14.25), (14.26) e (14.27), iremos, a seguir, definir a expressão da somatória do logaritmo da função de verossimilhança da distribuição binomial negativa, que deverá ser maximizada. Segundo o padrão adotado, estimaremos um modelo de regressão binomial negativo (NB2) com base na elaboração de um exemplo a ser resolvido inicialmente por meio da ferramenta **Solver** do Excel.

14.3.1. Estimação do modelo de regressão binomial negativo por máxima verossimilhança

Apresentaremos, agora, os conceitos pertinentes à estimação por máxima verossimilhança de um modelo de regressão binomial negativo por meio de um exemplo similar ao desenvolvido na seção 14.2.

Imagine que o professor dê continuidade à pesquisa sobre a quantidade de atrasos dos alunos, porém agora com contagem não mais semanal e, sim, de forma mensal. Após o término do mês, o professor realizou a pesquisa com os mesmos 100 alunos da escola onde leciona, questionando agora sobre a quantidade de vezes que cada um chegou atrasado neste último mês. As variáveis X são as mesmas, ou seja, distância percorrida até a escola (em quilômetros), número de semáforos pelos quais cada um passa e o período do dia em que cada estudante tem o hábito de se deslocar para a escola (manhã ou tarde). Parte do banco de dados encontra-se na Tabela 14.10.

A Tabela 14.11 apresenta a média e a variância da variável dependente, por meio da qual podemos verificar que a variância é consideravelmente maior do que sua média, gerando indícios sobre a existência de superdispersão dos dados.

Tabela 14.10 Exemplo: quantidade de atrasos no mês x distância percorrida, quantidade de semáforos e período do dia para o trajeto até a escola.

Estudante	Quantidade de atrasos no último mês (Y_i)	Distância percorrida até a escola (quilômetros) (X_{1i})	Quantidade de semáforos (X_{2i})	Período do dia (X_{3i})
Gabriela	5	11	15	1 (manhã)
Patrícia	0	9	15	1 (manhã)
Gustavo	0	9	16	1 (manhã)
Letícia	6	10	16	0 (tarde)
Luiz Ovídio	7	12	18	1 (manhã)
Leonor	4	14	16	0 (tarde)
Dalila	5	10	15	1 (manhã)
Antônio	0	10	16	1 (manhã)
Júlia	1	10	18	1 (manhã)
Mariana	0	9	13	1 (manhã)
...				
Filomena	1	8	18	1 (manhã)
...				
Estela	0	8	13	1 (manhã)

Tabela 14.11 Média e variância da variável dependente (quantidade de atrasos no último mês).

Estatística	
Média	1,820
Variância	5,422

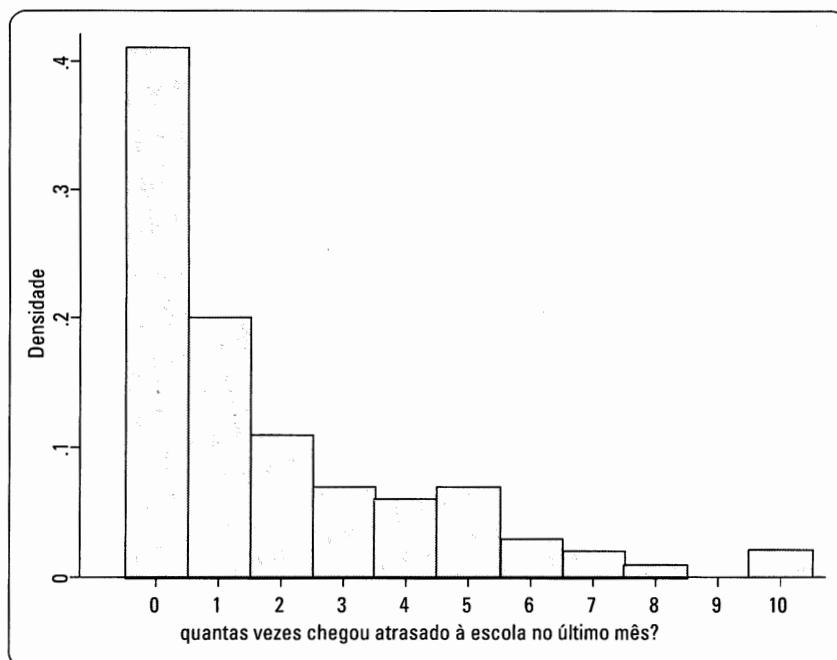


Figura 14.12 Histograma da variável dependente.

A Figura 14.12 apresenta o histograma da variável dependente para dados de contagem mensal, de onde podemos perceber que a dispersão é maior do que aquela apresentada no gráfico da Figura 14.3, elaborada para dados de contagem semanal.

Quando da estimação dos parâmetros do modelo, iremos também estimar o parâmetro ϕ da expressão (14.27), para que seja verificado se o mesmo é diferente de zero (existência de superdispersão) e, consequentemente, para que faça sentido a estimação do modelo de regressão binomial negativo.

O banco de dados completo elaborado nesta nova investigação pode ser acessado por meio do arquivo **QuantAtrasosBNeg.xls**. Estimaremos os parâmetros do modelo para avaliar a quantidade mensal esperada de atrasos de chegada à escola que, com base na expressão (14.5), será dada por:

$$u_i = e^{(\alpha + \beta_1 \cdot dist_i + \beta_2 \cdot sem_i + \beta_3 \cdot per_i)}$$

Com base na expressão (14.24), podemos escrever o logaritmo da função de verossimilhança (*log likelihood function*) de um modelo de regressão binomial negativo (NB2) como sendo:

$$LL = \sum_{i=1}^n \left[Y_i \cdot \ln \left(\frac{\phi \cdot u_i}{1 + \phi \cdot u_i} \right) - \frac{\ln(1 + \phi \cdot u_i)}{\phi} + \ln \Gamma(Y_i + \phi^{-1}) - \ln \Gamma(Y_i + 1) - \ln \Gamma(\phi^{-1}) \right] \quad (14.29)$$

Portanto, podemos fazer a seguinte pergunta: **Quais os valores dos parâmetros do modelo proposto que fazem com que o valor de LL da expressão (14.29) 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 binomial negativo, e pode ser respondida com o uso de ferramentas de programação linear, a fim de que sejam estimados os parâmetros $\phi, \alpha, \beta_1, \beta_2, \dots, \beta_k$ com base na seguinte função-objetivo:

$$LL = \sum_{i=1}^n \left[Y_i \cdot \ln \left(\frac{\phi \cdot u_i}{1 + \phi \cdot u_i} \right) - \frac{\ln(1 + \phi \cdot u_i)}{\phi} + \ln \Gamma(Y_i + \phi^{-1}) - \ln \Gamma(Y_i + 1) - \ln \Gamma(\phi^{-1}) \right] = \text{máx} \quad (14.30)$$

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 **QuantAtrasosBNegMáximaVerossimilhança.xls**, que servirá de auxílio para o cálculo dos parâmetros.

Neste arquivo, além da variável dependente e das variáveis explicativas, foram criadas duas novas variáveis, que correspondem, respectivamente, ao valor esperado de ocorrências mensais u_i , com distribuição Poisson e ao logaritmo da função de verossimilhança LL_i , proveniente da expressão (14.29) para cada observação.

Vamos, portanto, abrir a ferramenta **Solver** do Excel. A função-objetivo está na célula G103, que é a nossa célula de destino e que deverá ser maximizada. Além disso, os parâmetros $\phi, \alpha, \beta_1, \beta_2$ e β_3 , cujos valores estão nas células J2, J4, J6, J8 e J10, respectivamente, são as células variáveis. Além disso, devemos impor uma restrição de que $\phi > 0$. A janela do **Solver** ficará como mostra a Figura 14.13.

Ao clicarmos em **Resolver** e em **OK**, obteremos a solução ótima do problema de programação linear. A Tabela 14.12 apresenta parte dos resultados obtidos.

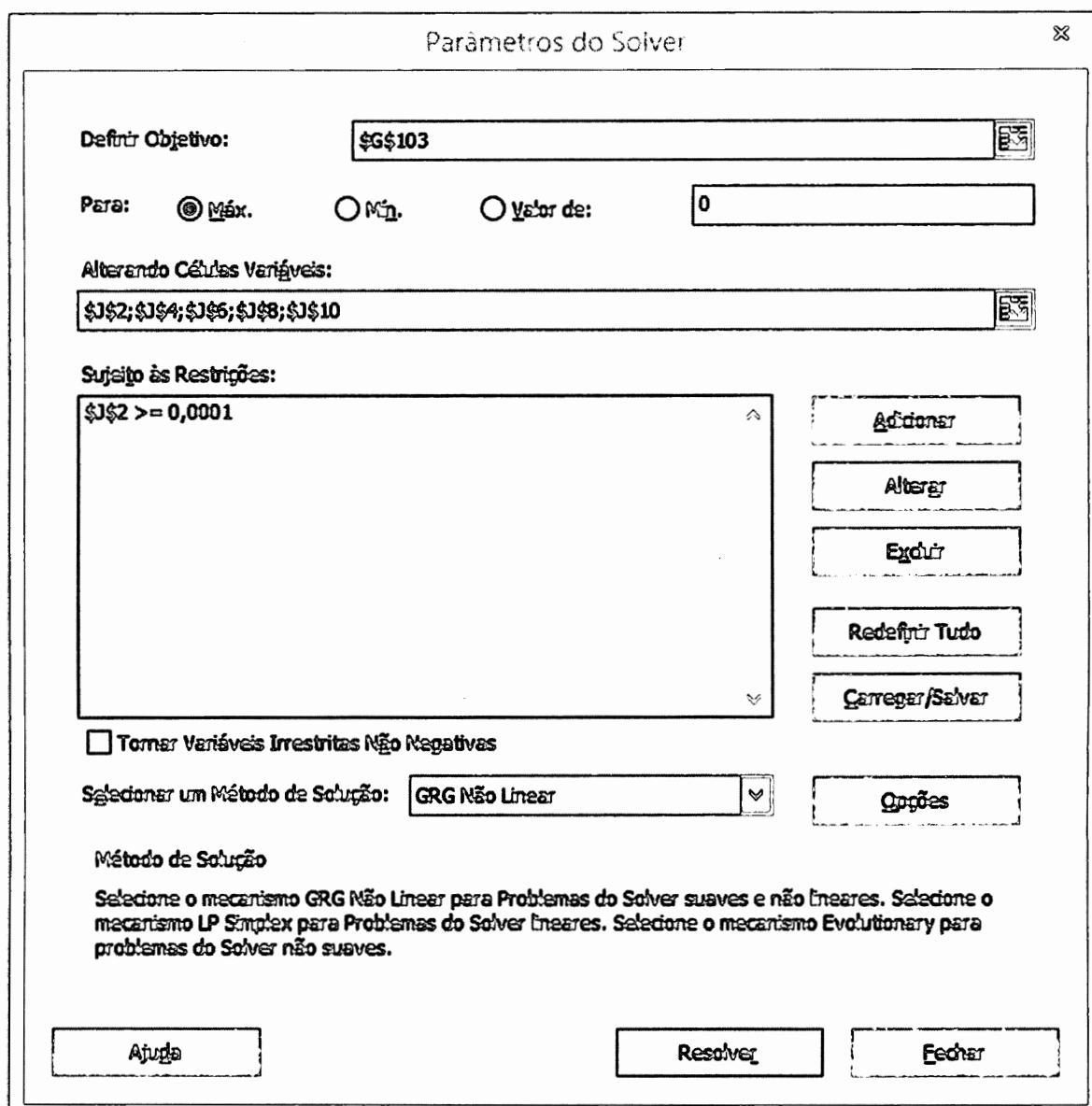


Figura 14.13 Solver – Maximização da somatória do logaritmo da função de verossimilhança.

Tabela 14.12 Valores obtidos quando da maximização de LL .

Estudante	Y_i	X_{1i}	X_{2i}	X_{3i}	u_i	LL_i
						$Y_i \cdot \ln \left(\frac{\phi \cdot u_i}{1 + \phi \cdot u_i} \right) - \frac{\ln(1 + \phi \cdot u_i)}{\phi} + \ln \Gamma(Y_i + \phi^{-1}) - \ln \Gamma(Y_i + 1) - \ln \Gamma(\phi^{-1})$
Gabriela	5	11	15	1	1,52099	-3,70395
Patrícia	0	9	15	1	0,82205	-0,74622
Gustavo	0	9	16	1	1,00138	-0,89171
Letícia	6	10	16	0	3,44343	-2,68117
Luiz Ovídio	7	12	18	1	3,73985	-2,94546
Leonor	4	14	16	0	11,78834	-3,09516
Dalila	5	10	15	1	1,111818	-4,55597
Antônio	0	10	16	1	1,36212	-1,16895
Júlia	1	10	18	1	2,02126	-1,34220
Mariana	0	9	13	1	0,55397	-0,51814
					...	
Filomena	1	8	18	1	1,09243	-1,12117
					...	
Estela	0	8	13	1	0,40726	-0,38745
Somatória	$LL = \sum_{i=1}^{100} \left[Y_i \cdot \ln \left(\frac{\phi \cdot u_i}{1 + \phi \cdot u_i} \right) - \frac{\ln(1 + \phi \cdot u_i)}{\phi} + \ln \Gamma(Y_i + \phi^{-1}) - \ln \Gamma(Y_i + 1) - \ln \Gamma(\phi^{-1}) \right]$					-151,01230

Logo, o valor máximo possível da somatória do logaritmo da função de verossimilhança é $LL_{\max} = -151,01230$. A resolução deste problema gerou as seguintes estimativas dos parâmetros:

$$\phi = 0,2553$$

$$\alpha = -4,9976$$

$$\beta_1 = 0,3077$$

$$\beta_2 = 0,1973$$

$$\beta_3 = -0,9274$$

Como $\phi \neq 0$, daremos sequência à estimação do modelo de regressão binomial negativo, porém quando estimarmos este modelo por meio dos softwares Stata e SPSS, respectivamente nas seções 14.4 e 14.5, verificaremos que ϕ é de fato estatisticamente diferente de zero. Caso um pesquisador mais curioso estimasse um modelo de regressão binomial negativo no banco de dados utilizado na seção 14.2, verificaria que a estimação de $\phi \approx 0$, como já era de se esperar, visto que o teste para verificação de existência de superdispersão não rejeitou a hipótese nula de equidispersão para aquele caso.

Logo, a expressão da quantidade mensal esperada de atrasos de chegada à escola pode ser escrita da seguinte forma:

$$u_i = e^{(-4,9976 + 0,3077 \cdot dist_i + 0,1973 \cdot sem_i - 0,9274 \cdot per_i)}$$

A Figura 14.14 apresenta parte dos resultados obtidos pela modelagem.

Estimados os parâmetros do modelo de regressão binomial negativo, podemos voltar às quatro perguntas propostas ao final da seção 14.2.1, porém agora para dados de contagem mensal:

Qual é a quantidade média esperada de atrasos no mês quando se desloca 12 quilômetros e se passa por 17 semáforos diariamente, sendo o trajeto feito à tarde?

Em média, em quanto se altera a taxa de incidência mensal de atrasos ao se adotar um percurso 1 quilômetro mais longo, mantidas as demais condições constantes?

Em média, em quanto se altera a taxa de incidência mensal de atrasos ao se passar por 1 semáforo a mais no percurso até a escola, mantidas as demais condições constantes?

Em média, em quanto se altera a taxa de incidência mensal de atrasos ao se optar por ir à escola de manhã, em vez de se ir à tarde, mantidas as demais condições constantes?

	A	B	C	D	E	F	G	H	I	J
1	Estudante	Atrasos (Y)	Distância (X_1)	Semáforos (X_2)	Período (X_3)	u_i	LL_i			
2	Gabriela	5	11	15	1	1,52099	-3,70395	ϕ	0,2553	
3	Patrícia	0	9	15	1	0,82205	-0,74622	α	-4,9976	
4	Gustavo	0	9	16	1	1,00138	-0,89171	β_1	0,3077	
5	Leície	6	10	16	0	3,44343	-2,68117	β_2	0,1973	
6	Luiz Ovídio	7	12	18	1	3,73985	-2,94546	β_3	-0,9274	
7	Leonor	4	14	16	0	11,78834	-3,09516			
8	Dália	5	10	15	1	1,11818	-4,55597			
9	Antônio	0	10	16	1	1,36212	-1,16895			
10	Júlia	1	10	18	1	2,02126	-1,34220			
11	Mariana	0	9	13	1	0,55397	-0,51814			
12	Roberto	2	9	15	1	0,82205	-1,98495			
13	Renata	0	9	15	1	0,82205	-0,74622			
14	Guilherme	4	12	17	1	3,07009	-2,06459			
15	Rodrigo	1	9	12	1	0,45476	-1,32807			
16	Giulia	0	11	11	1	0,69074	-0,63616			
17	Felipe	3	9	17	1	1,21984	-2,43101			
18	Karina	3	11	14	1	1,24860	-2,39972			
19	Pietro	1	11	15	1	1,52099	-1,19384			
20	Cecília	5	11	15	1	1,52099	-3,70395			
21	Gisele	0	9	14	1	0,67483	-0,62261			
22	Elaine	2	11	13	1	1,02499	-1,79178			
23	Kamal	0	9	14	1	0,67483	-0,62261			
24	Rodolfo	0	11	15	1	1,52099	-1,28509			
25	Pilar	0	11	13	1	1,02499	-0,91047			
26	Vivian	4	13	16	1	3,42817	-2,01900			
27	Danielle	0	9	11	1	0,37332	-0,35658			
28	Juliana	0	9	16	1	1,00138	-0,89171			
101	Estela	0	8	13	1	0,40726	-0,38745			
102										
103							Somatória LL_i	-151,01230		

Figura 14.14 Obtenção dos parâmetros quando da maximização de LL pelo Solver.

Antes de respondermos a estas importantes questões, precisamos novamente verificar se todos os parâmetros estimados são estatisticamente significantes a um determinado nível de confiança. Se não for este o caso, precisaremos reestimar 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.

Partiremos, portanto, para o estudo da significância estatística geral do modelo de regressão binomial negativo estimado, bem como das significâncias estatísticas dos parâmetros, de forma análoga ao realizado na seção 14.2.2.

14.3.2. Significância estatística geral e dos parâmetros do modelo de regressão binomial negativo

A fim de que possam ser calculados o pseudo R^2 de McFadden e a estatística χ^2 , com base, respectivamente, nas expressões (14.9) e (14.10), vamos, inicialmente, calcular LL_0 , que é dado pelo valor máximo da somatória do logaritmo da função de verossimilhança da expressão (14.29) para um modelo em que há apenas a constante α , conhecido por **modelo nulo**. Por meio do mesmo procedimento elaborado na seção 14.3.1, porém agora utilizando o arquivo **QuantAtrasosBNegMáximaVerossimilhançaModeloNulo.xls**, obteremos $LL_0 = -182,63662$. As Figuras 14.15 e 14.16 mostram, respectivamente, a janela do **Solver** e parte dos resultados obtidos pela modelagem neste arquivo.

Desta forma, temos que:

$$\text{pseudo } R^2 = \frac{-2.(-182,63662) - [(-2.(-151,01230))]}{-2.(-182,63662)} = 0,1732$$

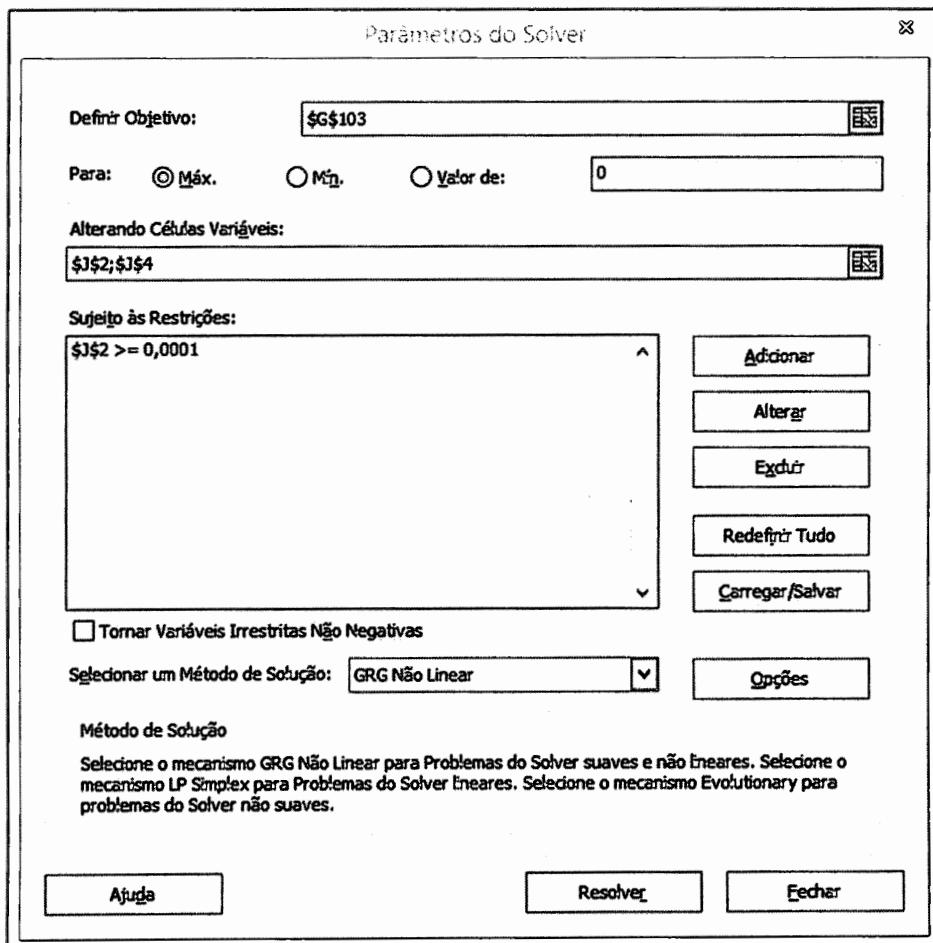


Figura 14.15 Solver – Maximização da somatória do logaritmo da função de verossimilhança para o modelo nulo.

A	B	C	D	E	F	G	H	I	J
Estudante	Atrasos (Y)	Distância (X ₁)	Semáforos (X ₂)	Período (X ₃)	$\hat{\mu}$	LL _i			
1 Gabriela	5	11	15	1	1,82000	-3,27602		ϕ	1,3521
2 Patrícia	0	9	15	1	1,82000	-0,91822			
3 Gustavo	0	9	16	1	1,82000	-0,91822		α	0,5988
4 Letícia	6	10	16	0	1,82000	-3,66141			
5 Luiz Ovídio	7	12	18	1	1,82000	-4,04033			
7 Leonor	4	14	16	0	1,82000	-2,88152			
8 Dafila	5	10	15	1	1,82000	-3,27602			
9 Antônio	0	10	16	1	1,82000	-0,91822			
10 Júlia	1	10	18	1	1,82000	-1,56086			
11 Mariana	0	9	13	1	1,82000	-0,91822			
12 Roberto	2	9	15	1	1,82000	-2,04137			
13 Renata	0	9	15	1	1,82000	-0,91822			
14 Guilherme	4	12	17	1	1,82000	-2,88152			
15 Rodrigo	1	9	12	1	1,82000	-1,56086			
16 Giulia	0	11	11	1	1,82000	-0,91822			
17 Felipe	3	9	17	1	1,82000	-2,47318			
18 Karina	3	11	14	1	1,82000	-2,47318			
19 Pietro	1	11	15	1	1,82000	-1,56086			
20 Cecília	5	11	15	1	1,82000	-3,27602			
21 Gisele	0	9	14	1	1,82000	-0,91822			
22 Elaine	2	11	13	1	1,82000	-2,04137			
23 Kamal	0	9	14	1	1,82000	-0,91822			
24 Rodolfo	0	11	15	1	1,82000	-0,91822			
25 Pilar	0	11	13	1	1,82000	-0,91822			
26 Vivian	4	13	16	1	1,82000	-2,88152			
27 Danielle	0	9	11	1	1,82000	-0,91822			
28 Juliana	0	9	16	1	1,82000	-0,91822			
101 Estela	0	8	13	1	1,82000	-0,91822			
102									
103						Somatória LL _i -182,63662			

Figura 14.16 Obtenção dos parâmetros quando da maximização de LL pelo Solver – modelo nulo.

Como sabemos, mesmo sendo bastante limitada a utilidade do pseudo R² de McFadden, softwares como o Stata e o SPSS o calculam e o apresentam em seus *outputs*, conforme veremos nas seções 14.4 e 14.5, respectivamente. A sua utilidade restringe-se à comparação de dois ou mais modelos apenas de mesma classe, ou seja, não pode ser utilizado para se comparar, por exemplo, um modelo Poisson com um modelo binomial negativo.

Além disso, temos também que:

$$\chi^2_{3g.l.} = -2 \cdot [-182,63662 - (-151,01230)] = 63,2486$$

Analogamente ao discutido na seção 14.2.2, para 3 graus de liberdade (número de variáveis explicativas consideradas na modelagem, ou seja, número de parâmetros β), temos, por meio da Tabela D do apêndice do livro, que o $\chi^2_c = 7,815$ (χ^2 crítico para 3 graus de liberdade e para o nível de significância de 5%). Desta forma, como o χ^2 calculado $\chi^2_{cal} = 63,2486 > \chi^2_c = 7,815$, podemos rejeitar a hipótese nula de que todos os parâmetros β_j ($j = 1, 2, 3$) sejam estatisticamente iguais a zero. Logo, pelo menos uma variável X é estatisticamente significante para explicar a incidência de atrasos de chegada à escola mensalmente e teremos um modelo de regressão binomial negativo estatisticamente significante para fins de previsão.

Softwares como o Stata e o SPSS não oferecem o χ^2_c para os graus de liberdade definidos e um determinado nível de significância. Todavia, oferecem o nível de significância do χ^2_{cal} para estes graus de liberdade. Desta forma, em vez de analisarmos se $\chi^2_{cal} > \chi^2_c$, devemos verificar se o nível de significância do χ^2_{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. χ^2_{cal}* ou *Prob. χ^2_{cal}*) < 0,05, existe pelo menos um $\beta_j \neq 0$.

Ainda seguindo a mesma lógica proposta na seção 14.2.2, é preciso que o avaliemos também se cada um dos parâmetros do modelo de regressão binomial negativo é estatisticamente significante, por meio também da análise da estatística *z* de Wald. Para o nosso exemplo, temos que:

$$s.e. (\alpha) = 1,249$$

$$s.e. (\beta_1) = 0,071$$

$$s.e. (\beta_2) = 0,049$$

$$s.e. (\beta_3) = 0,257$$

Logo, com base nas equações da expressão (14.11), temos que:

$$z_{\alpha} = \frac{\alpha}{s.e.(\alpha)} = \frac{-4,9976}{1,249} = -4,001$$

$$z_{\beta_1} = \frac{\beta_1}{s.e.(\beta_1)} = \frac{0,3077}{0,071} = 4,320$$

$$z_{\beta_2} = \frac{\beta_2}{s.e.(\beta_2)} = \frac{0,1973}{0,049} = 3,984$$

$$z_{\beta_3} = \frac{\beta_3}{s.e.(\beta_3)} = \frac{-0,9274}{0,257} = -3,608$$

Como todos os valores de $z_{\text{al}} < -1,96$ ou $> 1,96$, os *valores-P* das estatísticas z de Wald $< 0,05$ para todos os parâmetros estimados e, portanto, já chegamos ao modelo final de regressão binomial negativo, sem que haja necessidade de uma eventual aplicação do procedimento *Stepwise*. Sendo assim, a quantidade esperada de atrasos por mês para determinado aluno i é, de fato, dada por:

$$u_i = e^{(-4,9976 + 0,3077 \cdot \text{dist}_i + 0,1973 \cdot \text{sem}_i - 0,9274 \cdot \text{per}_i)}$$

e, desta forma, podemos retornar às perguntas propostas, respondendo uma de cada vez:

Qual é a quantidade média esperada de atrasos no mês quando se desloca 12 quilômetros e se passa por 17 semáforos diariamente, sendo o trajeto feito à tarde?

Com base na expressão da quantidade esperada de atrasos por mês e substituindo os valores propostos, teremos que:

$$u = e^{[-4,9976 + 0,3077 \cdot (12) + 0,1973 \cdot (17) - 0,9274 \cdot (0)]} = 7,76$$

Portanto, espera-se que determinado aluno que é submetido aos dados propostos ao se deslocar à escola apresente uma quantidade média de 7,76 atrasos por mês.

Em média, em quanto se altera a taxa de incidência mensal de atrasos ao se adotar um percurso 1 quilômetro mais longo, mantidas as demais condições constantes?

Fazendo uso da mesma expressão, temos que:

$$e^{0,3077} = 1,360$$

Assim, mantidas as demais condições constantes, a taxa de incidência mensal de atrasos ao se adotar um percurso 1 quilômetro mais longo é, em média, multiplicada por um fator de 1,360, ou seja, é, em média, 36,0% maior.

Em média, em quanto se altera a taxa de incidência mensal de atrasos ao se passar por 1 semáforo a mais no percurso até a escola, mantidas as demais condições constantes?

Neste caso, teremos:

$$e^{0,1973} = 1,218$$

Logo, mantidas as demais condições constantes, a taxa de incidência mensal de atrasos ao se adotar um percurso com 1 semáforo a mais é, em média, multiplicada por um fator de 1,218, ou seja, é, em média, 21,8% maior.

Em média, em quanto se altera a taxa de incidência mensal de atrasos ao se optar por ir à escola de manhã, em vez de se ir à tarde, mantidas as demais condições constantes?

Neste último caso, teremos:

$$e^{-0,9274} = 0,396$$

Logo, mantidas as demais condições constantes, a taxa de incidência mensal de atrasos ao se optar por ir à escola de manhã, em vez de se ir à tarde, é, em média, multiplicada por um fator de 0,396, ou seja, é, em média, 60,4% menor.

Como estes cálculos utilizam as estimativas médias dos parâmetros, estudaremos agora os intervalos de confiança destes parâmetros.

14.3.3. Construção dos intervalos de confiança dos parâmetros do modelo de regressão binomial negativo

Com base nos termos da expressão (14.12), podemos elaborar a Tabela 14.13, que traz os coeficientes estimados dos parâmetros do modelo 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%.

Tabela 14.13 Cálculo dos intervalos de confiança dos parâmetros.

Parâmetro	Coeficiente	Erro-Padrão (s.e.)	z	Intervalo de Confiança (95%)	
				$\alpha - 1,96 \cdot [s.e.(\alpha)]$	$\alpha + 1,96 \cdot [s.e.(\alpha)]$
α (constante)	-4,9976	1,249	-4,001	-7,446	-2,549
β_1 (variável $dist$)	0,3077	0,071	4,320	0,168	0,447
β_2 (variável sem)	0,1973	0,049	3,984	0,100	0,294
β_3 (variável per)	-0,9274	0,257	-3,608	-1,431	-0,424

Esta tabela é igual a que obteremos quando estimarmos este modelo de regressão binomial negativo por meio do Stata e do SPSS (seções 14.4 e 14.5, respectivamente).

Com base nos intervalos de confiança dos parâmetros, podemos escrever as expressões dos limites inferior (mínimo) e superior (máximo) da quantidade esperada de atrasos por mês para determinado aluno i , com 95% de confiança:

$$u_{i_{\min}} = e^{(-7,446 + 0,168 \cdot dist_i + 0,100 \cdot sem_i - 1,431 \cdot per_i)}$$

$$u_{i_{\max}} = e^{(-2,549 + 0,447 \cdot dist_i + 0,294 \cdot sem_i - 0,424 \cdot per_i)}$$

Fazendo uso da expressão (14.13), podemos elaborar a Tabela 14.14, que apresenta o intervalo de confiança da taxa mensal estimada de incidência de atrasos (**incidence rate ratio** ou **IRR**) correspondente à alteração em cada parâmetro β_j ($j = 1, 2, \dots, k$).

Estes valores também poderão ser obtidos por meio do Stata e do SPSS, conforme mostraremos, respectivamente, nas seções 14.4 e 14.5.

Como podemos verificar, os intervalos de confiança dos parâmetros estimados não contêm o zero e, consequentemente, os das taxas esperadas de incidência não contêm o 1, o que já era de se esperar, dado que, conforme

Tabela 14.14 Cálculo dos intervalos de confiança da taxa de incidência u (IRR) para cada parâmetro β_j .

Parâmetro	e^{β_j}	Intervalo de Confiança de u (95%)	
		$e^{\beta_j - 1,96 \cdot [s.e.(\beta_j)]}$	$e^{\beta_j + 1,96 \cdot [s.e.(\beta_j)]}$
β_1 (variável $dist$)	1,360	1,182	1,564
β_2 (variável sem)	1,218	1,105	1,342
β_3 (variável per)	0,396	0,239	0,655

discutimos, $z_{cal} < -1,96$ ou $> 1,96$. Logo, os parâmetros estimados são estatisticamente diferentes de zero ao nível de confiança de 95%.

Partiremos agora para a estimação dos modelos de regressão para dados de contagem por meio dos softwares Stata e SPSS.

14.4. ESTIMAÇÃO DE MODELOS DE REGRESSÃO PARA DADOS DE CONTAGEM NO SOFTWARE STATA

O objetivo desta seção não é o de discutir novamente todos os conceitos inerentes às estatísticas dos modelos de regressão Poisson e binomial negativo, 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®.

14.4.1. Modelo de regressão Poisson no software Stata

Voltando ao exemplo desenvolvido na seção 14.2, lembremos que o nosso professor tem o interesse em avaliar se a distância percorrida, a quantidade de semáforos e o período do dia em que ocorre o percurso até a escola influenciam a quantidade de atrasos semanalmente. 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 **QuantAtrasosPoisson.dta** e é exatamente igual ao apresentado parcialmente por meio da Tabela 14.2.

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 14.17 apresenta este primeiro *output* do Stata.

. desc					
obs:	100				
vars:	5				
size:	2,500 (99.9% of memory free)				
variable	name	storage	display	value	variable label
		type	format	label	
estudante		str11	%11s		
atrasos		float	%9.0g	quantas vezes chegou atrasado à escola na última semana?	
dist		byte	%8.0g	distância que percorre até a escola (km)	
sem		byte	%8.0g	quantidade de semáforos	
per		float	%9.0g	período do dia	
Sorted by:					

Figura 14.17 Descrição do banco de dados **QuantAtrasosPoisson.dta**.

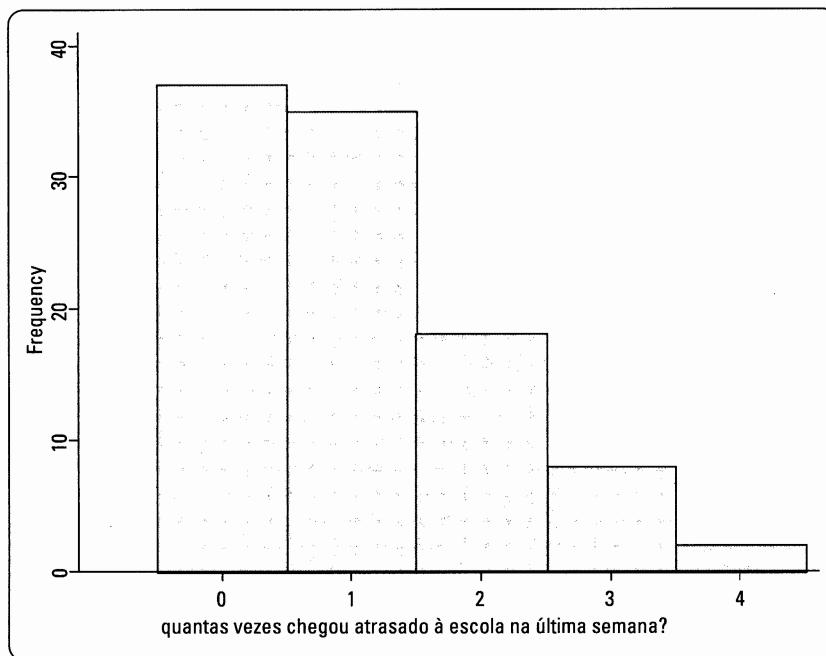
A variável dependente, que se refere à quantidade de atrasos (número de ocorrências) semanalmente ao se chegar à escola, é quantitativa, discreta e com valores não negativos. Desta forma, o comando **tab**, que frequentemente é utilizado para se obter a distribuição de frequências de uma variável qualitativa, pode ser, neste caso, utilizado, dado que a variável dependente apresenta valores inteiros e com poucas possibilidades de resposta. A Figura 14.18 apresenta a distribuição de frequências para os dados de contagem da variável dependente *atrasos*.

O comando a seguir oferece uma possibilidade de visualização do histograma da variável dependente, apresentado na Figura 14.19. O termo **discrete** informa que a variável dependente apresenta apenas valores inteiros.

hist atrasos, discrete freq

Antes da elaboração de qualquer modelo de regressão para dados de contagem, é interessante que o pesquisador avalie se a média e a variância da variável dependente são iguais ou, ao menos, próximas. Isso dará uma ideia

tab atrasos			
quantas vezes chegou atrasado à escola na última semana?	Freq.	Percent	Cum.
0	37	37.00	37.00
1	35	35.00	72.00
2	18	18.00	90.00
3	8	8.00	98.00
4	2	2.00	100.00
Total	100	100.00	

Figura 14.18 Distribuição de frequências para os dados de contagem da variável *atrasos*.**Figura 14.19** Histograma da variável dependente *atrasos*.

sobre a adequação da estimação do modelo de regressão Poisson, ou se será necessária a estimação de um modelo de regressão binomial negativo. A digitação do seguinte comando permitirá que este preliminar diagnóstico seja elaborado, cujos resultados encontram-se na Figura 14.20:

```
tabstat atrasos, stats(mean var)
```

Os *outputs* da Figura 14.20 correspondem aos apresentados na Tabela 14.3 da seção 14.2.1 e, por meio da análise da média e da variância, que são muito próximas, podemos, ainda que de forma preliminar, supor que

tabstat atrasos, stats(mean var)		
variable	mean	variance
atrasos	1.03	1.059697

Figura 14.20 Média e variância da variável dependente *atrasos*.

a estimação de um modelo de regressão Poisson seja adequada neste caso. É importante ressaltar que, quando a variável dependente apresentar dados de contagem, a estimação de um modelo de regressão Poisson deverá sempre ser elaborada inicialmente, a fim de que, a partir da mesma, possa ser aplicado um teste para verificação de existência de superdispersão. Caso ocorra superdispersão nos dados, aí sim o pesquisador poderá recorrer à estimativa de um modelo de regressão binomial negativo, em detrimento da estimativa do modelo Poisson.

Vamos, então, à estimativa do modelo de regressão Poisson. Para tanto, devemos digitar o seguinte comando:

poisson atrasos dist sem per

O comando **poisson** elabora um modelo de regressão Poisson estimado por máxima verossimilhança. Assim como para os modelos de regressão múltipla e de regressão logística binária e multinomial, 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:

poisson atrasos dist sem per, level(90)

Iremos seguir com a análise mantendo o nível padrão de confiança dos intervalos dos parâmetros, que é de 95%. Os resultados encontram-se na Figura 14.21 e são exatamente iguais aos calculados na seção 14.2.

Como os modelos de regressão Poisson fazem parte do grupo de modelos conhecidos por **Modelos Lineares Generalizados (Generalized Linear Models)**, e como estamos supondo, neste momento, que a variável dependente apresenta uma distribuição Poisson, já que o teste para verificação de existência de superdispersão nos dados ainda será elaborado, os resultados da estimativa apresentados na Figura 14.21 também podem igualmente ser obtidos por meio da digitação do seguinte comando:

glm atrasos dist sem per, family(poisson)

. poisson atrasos dist sem per						
Iteration 0:	log likelihood = -107.79072					
Iteration 1:	log likelihood = -107.61523					
Iteration 2:	log likelihood = -107.61498					
Iteration 3:	log likelihood = -107.61498					
Poisson regression				Number of obs	=	100
				LR chi2(3)	=	51.01
				Prob > chi2	=	0.0000
				Pseudo R2	=	0.1916
Log likelihood = -107.61498						
<hr/>						
atrasos	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
dist	.2221224	.0658737	3.37	0.001	.0930122	.3512325
sem	.1646107	.0458251	3.59	0.000	.0747952	.2544262
per	-.5731352	.261911	-2.19	0.029	-1.086471	-.059799
_cons	-4.379926	1.160234	-3.78	0.000	-6.653943	-2.10591

Figura 14.21 Outputs do modelo de regressão Poisson no Stata.

Inicialmente, podemos verificar que mostram, respectivamente, a janela o valor máximo do logaritmo da função de verossimilhança para o modelo completo é igual a -107,61498, que é exatamente igual ao valor calculado por meio do **Solver** do Excel (seção 14.2.1) e apresentado na Tabela 14.5 e na Figura 14.6. Caso o pesquisador queira obter o valor máximo do logaritmo da função de verossimilhança para o modelo nulo, deverá digitar o seguinte comando, cujos resultados encontram-se na Figura 14.22:

poisson atrasos

. poisson atrasos					
Iteration 0: log likelihood = -133.12228					
Iteration 1: log likelihood = -133.12228					
Poisson regression					
	Number of obs	=	100		
	LR chi2(0)	=	0.00		
	Prob > chi2	=			
	Pseudo R2	=	0.0000		
Log likelihood = -133.12228					
atrasos	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
cons	.0295588	.0985329	0.30	0.764	-.1635622 .2226798

Figura 14.22 Outputs do modelo de regressão Poisson nulo no Stata.

Logo, o valor máximo do logaritmo da função de verossimilhança para o modelo nulo é igual a -133,12228, que é exatamente igual ao valor também calculado pelo **Solver** do Excel e apresentado na Figura 14.8.

Assim, fazendo uso da expressão (14.10), temos que:

$$\chi^2_{3g.l.} = -2 \cdot [-133,12228 - (-107,61498)] = 51,01 \quad \text{com valor - } P(\text{ou Prob. } \chi^2_{cal}) = 0,000.$$

Logo, com base no teste χ^2 , podemos rejeitar a hipótese nula de que todos os parâmetros β_j ($j = 1, 2, 3$) sejam estatisticamente iguais a zero ao nível de significância de 5%, ou seja, pelo menos uma variável X é estatisticamente significante para explicar o número de atrasos que ocorre semanalmente ao se chegar 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 (14.9), exatamente como fizemos na seção 14.2.2.

$$\text{pseudo } R^2 = \frac{-2 \cdot (-133,12228) - [(-2 \cdot (-107,61498))]}{-2 \cdot (-133,12228)} = 0,1916$$

Em relação à significância estatística dos parâmetros do modelo apresentado na Figura 14.21, como todos os valores de $z_{cal} < -1,96$ ou $> 1,96$, os *valores-P* das estatísticas z de Wald $< 0,05$ para todos os parâmetros estimados e, portanto, já chegamos ao modelo final de regressão Poisson, sem que haja a necessidade de uma eventual aplicação do procedimento *Stepwise*. Se este não tivesse sido o caso, seria recomendável a estimativa do modelo final por meio do seguinte comando:

stepwise, pr(0.05): poisson atrasos dist sem per

ou do equivalente:

stepwise, pr(0.05): glm atrasos dist sem per, family(poisson)

que, para este nosso exemplo, geram exatamente os mesmos resultados apresentados na Figura 14.21.

Logo, a quantidade média estimada de atrasos por semana para determinado aluno i é dada por:

$$\lambda_i = e^{(-4,380 + 0,222.dist_i + 0,165.sem_i - 0,573.per_i)}$$

que, à exceção de pequenos arredondamentos, é exatamente o mesmo modelo estimado na seção 14.2. Além disso, também com base na Figura 14.21, as quantidades estimadas de atrasos por semana apresentam, com 95% de nível de confiança, expressões de mínimo e de máximo iguais a:

$$\lambda_{i_{\min}} = e^{(-6,654 + 0,093.dist_i + 0,075.sem_i - 1,086.per_i)}$$

$$\lambda_{i_{\max}} = e^{(-2,106 + 0,351.dist_i + 0,254.sem_i - 0,060.per_i)}$$

Após a estimação do modelo de regressão Poisson, precisamos elaborar o teste para verificação de existência de superdispersão nos dados. Para tanto, seguiremos o mesmo procedimento estudado na seção 14.2.4.

Inicialmente, devemos gerar uma variável correspondente aos valores previstos de ocorrência de atrasos semanais por aluno, que chamaremos de *lambda*. Esta variável deverá ser gerada exatamente após a estimação do modelo final, por meio da digitação do seguinte comando:

```
predict lambda
```

Na sequência, com base na expressão (14.14), reescrita a seguir, devemos criar uma nova variável no banco de dados, que chamaremos de *yasterisco*, de acordo como segue:

$$yasterisco_i = \frac{[(atrasos_i - lambda_i)^2 - atrasos_i]}{lambda_i}$$

```
gen yasterisco = ((atrasos-lambda)^2 - atrasos)/lambda
```

Por fim, devemos estimar o modelo auxiliar de regressão simples $yasterisco_i = \beta \cdot lambda_i$, de acordo com a expressão (14.15), por meio da digitação do seguinte comando:

```
reg yasterisco lambda, nocons
```

Os resultados deste procedimento encontram-se na Figura 14.23, e correspondem aos apresentados na Figura 14.10.

Cameron e Trivedi (1990) salientam que, se ocorrer o fenômeno da superdispersão nos dados, o parâmetro β estimado por meio do modelo de regressão auxiliar será estatisticamente diferente de zero, ao nível definido de significância de 5%. Como o *valor-P* do teste *t* correspondente ao parâmetro β da variável *lambda* é maior do que 0,05, podemos afirmar que os dados da variável dependente **não apresentam superdispersão**, fazendo com que o modelo de regressão Poisson estimado seja adequado pela **presença de equidispersão nos dados**. Seguiremos, portanto, com o modelo final de regressão Poisson estimado.

. predict lambda (option n assumed; predicted number of events)	
. gen yasterisco = ((atrasos-lambda)^2 - atrasos)/lambda	
. reg yasterisco lambda, nocons	
	Source SS df MS
Model 15.0749658 1 15.0749658	
Residual 439.607992 99 4.44048476	
Total 454.682957 100 4.54682957	
	Number of obs = 100
	F(1, 99) = 3.39
	Prob > F = 0.0684
	R-squared = 0.0332
	Adj R-squared = 0.0234
	Root MSE = 2.1072

	yasterisco Coef. Std. Err. t P> t [95% Conf. Interval]
	lambda -.2917561 .158346 -1.84 0.068 -.6059489 .0224366

Figura 14.23 Resultado do teste para verificação de existência de superdispersão no Stata.

O comando **prcounts**, a ser digitado após a estimação do modelo final completo elaborado por meio do comando **poisson**, permite que sejam criadas variáveis correspondentes às probabilidades de ocorrência de cada uma das possibilidades de atraso (de 0 a 9 atrasos), para cada observação. Caso o comando **prcounts** não esteja instalado no Stata, o pesquisador deverá digitar **findit prcounts** e instalá-lo no pacote estatístico.

Vamos, então, digitar o seguinte comando:

```
prcounts prpoisson, plot
```

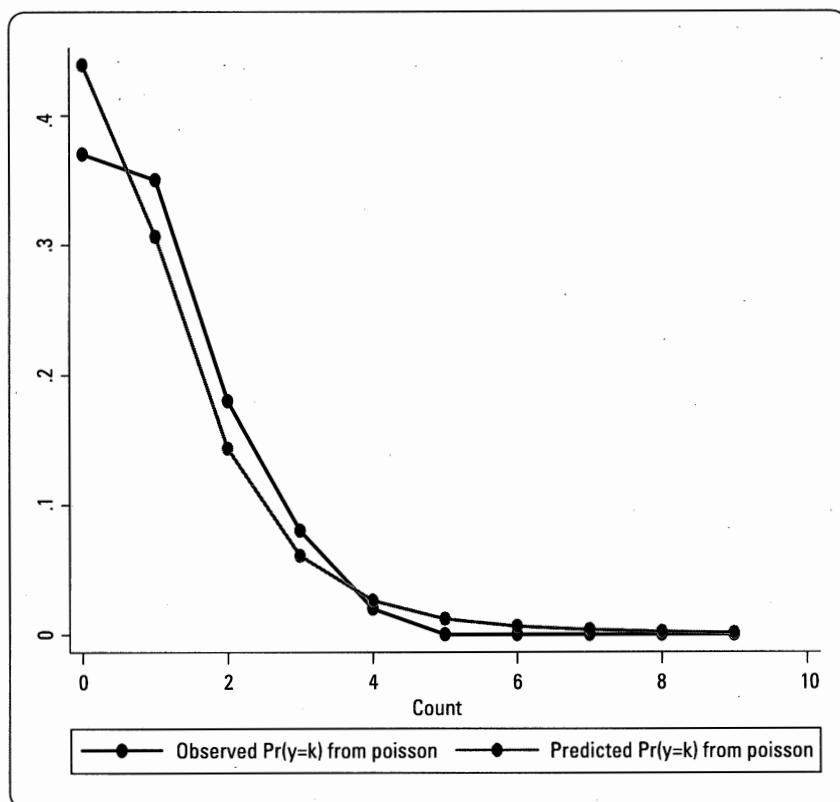


Figura 14.24 Distribuições de probabilidades observadas e previstas de ocorrência de 0 a 9 atrasos.

Além disso, são também geradas duas variáveis que correspondem, respectivamente, às probabilidades observadas e previstas de ocorrência de 0 a 9 atrasos para toda a amostra (*prpoissonobeq* e *prpoissonpreq*). Note que a variável *prpoissonobeq* apresenta, obviamente, a mesma distribuição de probabilidades apresentada na Figura 14.18. Por fim, a variável *prpoissonval* apresenta os próprios valores de 0 a 9 que serão relacionados com as probabilidades observadas e previstas. O comando a seguir permite que sejam comparadas, visualmente, as distribuições de probabilidades observadas e previstas de ocorrência de 0 a 9 atrasos:

```
graph twoway (scatter prpoissonobeq prpoissonpreq prpoissonval, connect (1 1))
```

O gráfico resultante encontra-se na Figura 14.24.

Desta forma, para que seja verificada a qualidade do ajuste do modelo final estimado, de forma análoga ao teste de Hosmer-Lemeshow utilizado quando da estimação de modelos de regressão logística binária, podemos elaborar um teste χ^2 para comparar as duas curvas apresentadas na Figura 14.24. Assim, após a estimação do modelo final, devemos digitar:

```
poisgof
```

O resultado, que se encontra na Figura 14.25, indica a existência de qualidade do ajuste do modelo final de regressão Poisson, ou seja, não existem diferenças estatisticamente significantes entre os valores previstos e observados do número de atrasos que ocorrem semanalmente.

```
. poisgof
Goodness-of-fit chi2      =   67.71699
Prob > chi2(96)           =     0.9873
```

Figura 14.25 Verificação da qualidade do ajuste do modelo de regressão Poisson estimado.

Desta forma, podemos retornar à primeira pergunta proposta ao final da seção 14.2.1:

Qual é a quantidade média esperada de atrasos na semana quando se desloca 12 quilômetros e se passa por 17 semáforos diariamente, sendo o trajeto feito à tarde?

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

```
mfx, at(dist=12 sem=17 per=0)
```

Assim como já havíamos calculado manualmente na seção 14.2.2, espera-se, portanto, que determinado aluno que é submetido a estas características ao se deslocar para a escola apresente, em média, uma quantidade de 2,95 atrasos por semana (Figura 14.26).

. mfx, at(dist=12 sem=17 per=0)							
Marginal effects after poisson							
y = Predicted number of events (predict)							
= 2.9562577							
variable	dy/dx	Std. Err.	z	P> z	[95% C.I.]	x	
dist	.6566509	.21773	3.02	0.003	.229916 1.08339	12	
sem	.4866317	.16407	2.97	0.003	.165058 .808205	17	
per*	-1.289652	.63928	-2.02	0.044	-2.54262 -.036687	0	

(* dy/dx is for discrete change of dummy variable from 0 to 1

Figura 14.26 Cálculo da quantidade esperada de atrasos semanais para valores das variáveis explicativas – comando **mfx**.

Caso haja a intenção de se obter diretamente as estimativas das taxas de incidência semanal de atrasos quando se altera em uma unidade determinada variável explicativa, mantidas as demais condições constantes, pode ser digitado o seguinte comando:

```
poisson atrasos dist sem per, irr
```

em que o termo **irr** significa *incidence rate ratio* e, para o nosso exemplo, oferece a taxa estimada de incidência de atrasos por semana correspondente à alteração em cada parâmetro β_j ($j = 1, 2, 3$). Os resultados, apresentados na Figura 14.27, também poderiam ser obtidos por meio do seguinte comando:

```
glm atrasos dist sem per, family(poisson) eform
```

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

. poisson atrasos dist sem per, irr							
Iteration 0: log likelihood = -107.79072							
Iteration 1: log likelihood = -107.61523							
Iteration 2: log likelihood = -107.61498							
Iteration 3: log likelihood = -107.61498							
Poisson regression							
Number of obs = 100							
LR chi2(3) = 51.01							
Prob > chi2 = 0.0000							
Pseudo R2 = 0.1916							
Log likelihood = -107.61498							
atrasos IRR Std. Err. z P> z [95% Conf. Interval]							
dist 1.248724	.0822581	3.37	0.001	1.097475	1.420818		
sem 1.178934	.0540247	3.59	0.000	1.077663	1.289721		
per .5637552	.1476537	-2.19	0.029	.337405	.9419538		

Figura 14.27 Outputs do modelo de regressão Poisson – *incidence rate ratios*.

Sendo assim, podemos retornar às três últimas perguntas propostas ao final da seção 14.2.1:

Em média, em quanto se altera a taxa de incidência semanal de atrasos ao se adotar um percurso 1 quilômetro mais longo, mantidas as demais condições constantes?

Em média, em quanto se altera a taxa de incidência semanal de atrasos ao se passar por 1 semáforo a mais no percurso até a escola, mantidas as demais condições constantes?

Em média, em quanto se altera a taxa de incidência semanal de atrasos ao se optar por ir à escola de manhã, em vez de se ir à tarde, mantidas as demais condições constantes?

As respostas agora podem ser dadas de maneira direta, ou seja, enquanto a taxa de incidência semanal de atrasos ao se adotar um percurso 1 quilômetro mais longo é, em média e mantidas as demais condições constantes, multiplicada por um fator de 1,249 (24,9% maior), a taxa de incidência semanal de atrasos ao se adotar um percurso com 1 semáforo a mais é, em média e também mantidas as demais condições constantes, multiplicada por um fator de 1,179 (17,9% maior). Por fim, a taxa de incidência semanal de atrasos ao se optar por ir à escola de manhã, em vez de se ir à tarde, é, em média, multiplicada por um fator de 0,564 (43,6% menor), mantidas as demais condições constantes. Estes valores são exatamente os mesmos daqueles calculados manualmente ao final da seção 14.2.2.

Um pesquisador mais curioso pode inclusive elaborar um gráfico para estudar, por exemplo, o comportamento da evolução da quantidade semanal prevista de atrasos em função da distância que é percorrida até a escola. Para tanto, pode ser digitado o seguinte comando:

```
graph twoway scatter lambda dist || mspline lambda dist
```

Por meio do gráfico elaborado e apresentado na Figura 14.28 é possível claramente perceber que distâncias maiores percorridas para se chegar à escola levam a um aumento da quantidade esperada de atrasos por semana, com taxa média de incremento de 24,9% de atrasos a cada 1 quilômetro adicional.

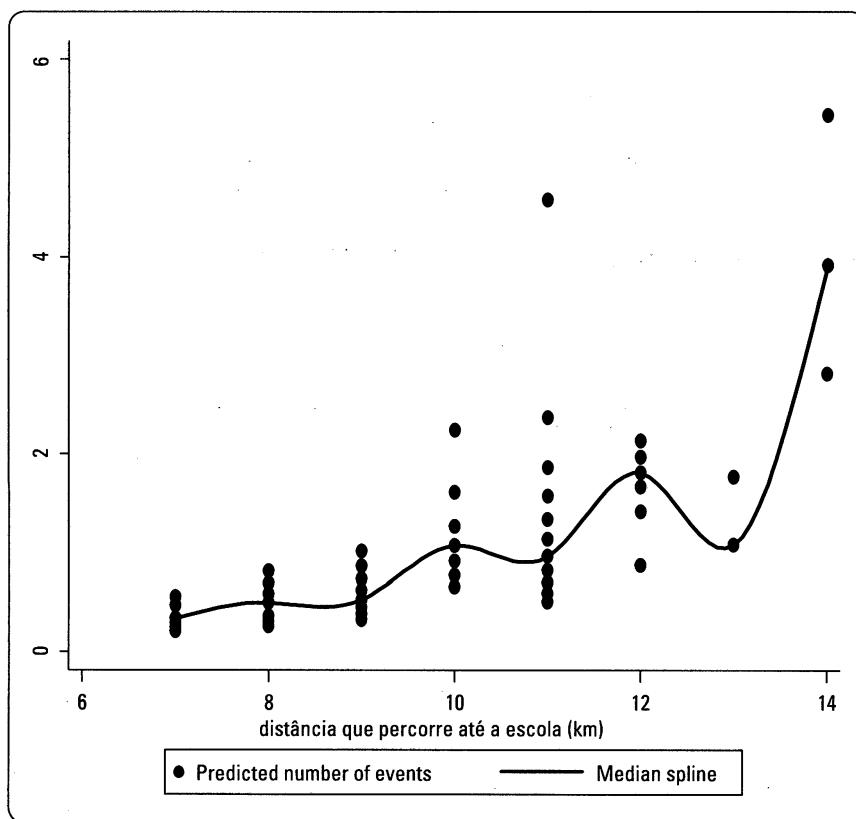


Figura 14.28 Quantidade esperada de atrasos por semana (λ) x distância percorrida ($dist$).

Entretanto, caso se deseje elaborar o mesmo gráfico, porém estratificando os comportamentos de evolução da quantidade semanal prevista de atrasos para trajetos realizados de manhã ou à tarde, deve-se digitar o seguinte comando:

```
graph twoway scatter lambda dist if per==0 || scatter lambda dist
if per==1 || mspline lambda dist if per==0 || mspline lambda dist
if per==1 ||, legend(label(3 "tarde") label(4 "manhã"))
```

O novo gráfico gerado encontra-se na Figura 14.29.

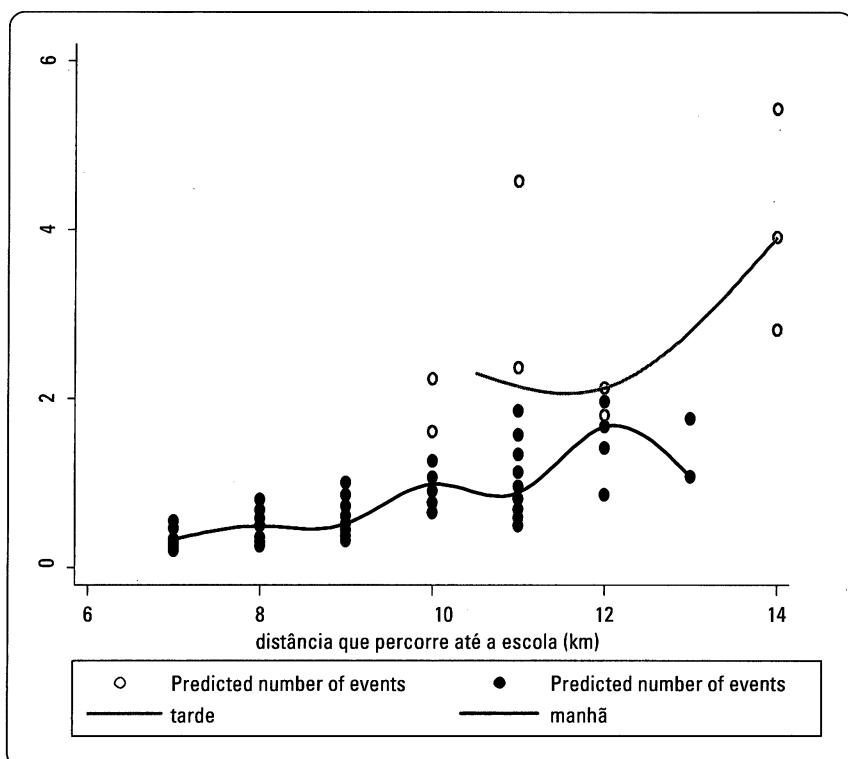


Figura 14.29 Quantidade esperada de atrasos por semana (*lambda*) x distância percorrida (*dist*) em diferentes períodos do dia (*per*).

Por meio deste gráfico é possível verificar que os trajetos para se chegar à escola realizados no período da tarde apresentam maiores distâncias, em média. Enquanto a quantidade esperada de atrasos por semana para os percursos realizados de manhã não apresenta média superior a 1 e não ultrapassa o valor de 2, a quantidade esperada de atrasos por semana para os percursos realizados à tarde e, portanto, que têm maiores distâncias, apresenta média em torno de 3, com valor mínimo ficando próximo de 2.

Por fim, podemos desejar comparar os resultados do modelo de regressão Poisson estimado por máxima verossimilhança com aqueles obtidos por um eventual modelo de regressão múltipla log-linear estimado pelo método de mínimos quadrados ordinários (*ordinary least squares*, ou *OLS*). Para tanto, vamos inicialmente gerar uma variável chamada de *lnatrasos*, que corresponde ao logaritmo natural da variável dependente *atrasos*, por meio do seguinte comando:

```
gen lnatrasos=ln(atrasos)
```

Na sequência, vamos estimar o modelo $\ln atrasos_i = \alpha + \beta_1.dist_i + \beta_2.sem_i + \beta_3.per_i$ por *OLS*, da seguinte forma:

```
quietly reg lnatrasos dist sem per
```

O termo **quietly** indica que os *outputs* não serão apresentados, porém os parâmetros serão estimados. A fim de obtermos os valores previstos da variável dependente por meio da estimação *OLS*, devemos digitar:

```
predict yhat
gen eyhat = exp(yhat)
```

em que a variável *eyhat* corresponde aos valores previstos, para cada observação, da quantidade de atrasos por semana para um modelo de regressão múltipla log-linear estimado por *OLS*.

O gráfico apresentado na Figura 14.30 oferece uma oportunidade de verificação, por meio de ajustes lineares, das diferenças dos valores previstos em função dos valores reais da variável dependente para cada uma das estimativas elaboradas (modelo de regressão Poisson estimado por máxima verossimilhança e modelo de regressão múltipla log-linear estimado por *OLS*). O comando para elaboração deste gráfico é:

```
graph twoway lfit lambda atrasos || lfit eyhat atrasos ||,
legend(label(1 "Poisson") label(2 "OLS"))
```

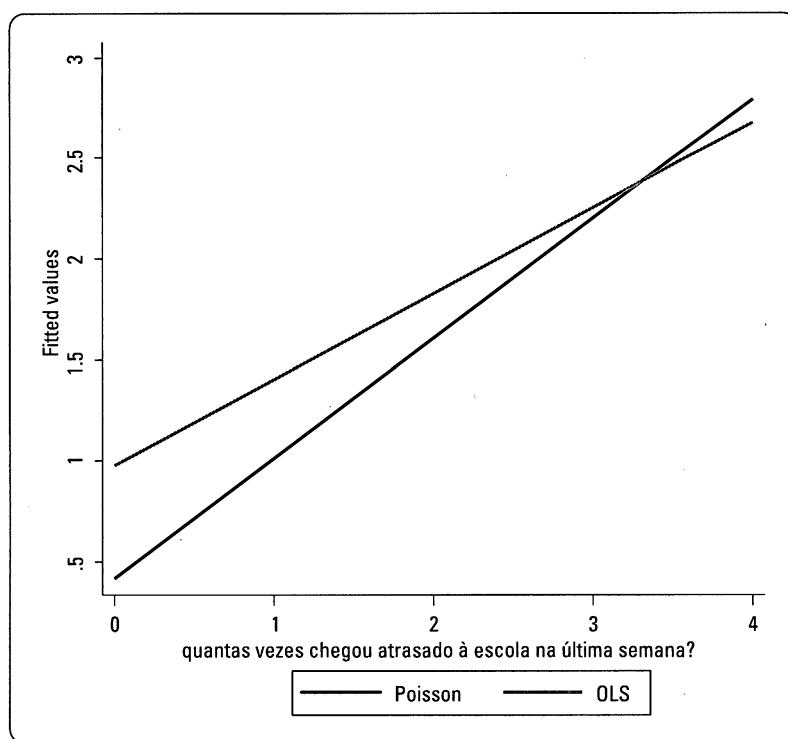


Figura 14.30 Valores previstos x valores observados para os modelos de regressão Poisson e de regressão múltipla log-linear (*OLS*).

O gráfico da Figura 14.30 nos mostra que o fato de determinada variável dependente ser quantitativa não é condição suficiente para que seja elaborado um modelo de regressão múltipla com estimação *OLS*, cujos parâmetros podem ser diferentes e viesados em relação àqueles obtidos por um modelo de regressão Poisson estimado por máxima verossimilhança. O pesquisador precisa investigar o comportamento da distribuição e a natureza da variável dependente de seu estudo, a fim de que seja estimado o modelo mais adequado e consistente para efeitos de diagnóstico da base de dados e para efeitos de previsão.

14.4.2. Modelo de regressão binomial negativo no software Stata

Voltando agora ao exemplo da seção 14.3, o professor passa a ter interesse em avaliar se a distância percorrida, a quantidade de semáforos e o período do dia em que se dá o trajeto até a escola são variáveis estatisticamente significantes para explicar a quantidade de atrasos por mês a que estão sujeitos os seus 100 alunos. O banco de dados encontra-se agora no arquivo **QuantAtrasosBNeg.dta** e é exatamente igual ao apresentado parcialmente por meio da Tabela 14.10.

Ao digitarmos o comando **desc**, podemos analisar as características do banco de dados, como o número de observações, o número de variáveis e a descrição de cada uma delas. A Figura 14.31 apresenta esta descrição.

Na sequência, seguindo a lógica apresentada na seção 14.4.1, vamos inicialmente analisar a distribuição da variável dependente neste novo exemplo, solicitando ao Stata que seja elaborada uma tabela com a distribuição de frequências e o correspondente histograma. Os comandos são:

```
tab atrasos
hist atrasos, discrete freq
```

. desc					
obs:	100	vars:	5	size:	2,500 (99.9% of memory free)
<hr/>					
variable name	storage type	display format	value label	variable	label
estudante	str11	%11s			
atrasos	float	%9.0g			quantas vezes chegou atrasado à escola no último mês?
dist	byte	%8.0g			distância que percorre até a escola (km)
sem	byte	%8.0g			quantidade de semáforos
per	float	%9.0g	per		período do dia
<hr/>					
Sorted by:					

Figura 14.31 Descrição do banco de dados QuantAtrasosBNeg.dta.

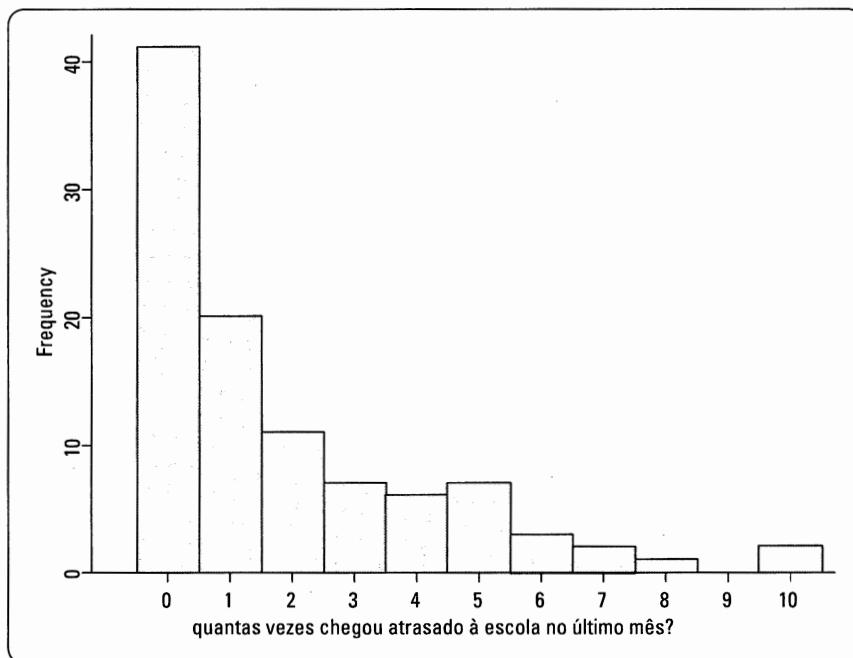
Enquanto a Figura 14.32 apresenta a tabela com a distribuição de frequências da variável dependente *atrasos*, a Figura 14.33 traz o histograma desta variável.

É importante verificar que a cauda mais longa deste histograma em comparação com aquele apresentado na Figura 14.19 é decorrente do fato de que, no presente estudo, a variável dependente contempla dados mensais de contagem, em vez de dados semanais. Esta cauda mais longa pode ser um primeiro indício de existência de superdispersão nos dados e, desta forma, faz-se necessário calcular a média e a variância desta variável dependente. Para tanto, devemos digitar o seguinte comando, cujos resultados encontram-se na Figura 14.34:

```
tabstat atrasos, stats(mean var)
```

. tab atrasos				
quantas	vezes	chegou	atrasado à	escola no
último mês?				
0	41		41.00	41.00
1	20		20.00	61.00
2	11		11.00	72.00
3	7		7.00	79.00
4	6		6.00	85.00
5	7		7.00	92.00
6	3		3.00	95.00
7	2		2.00	97.00
8	1		1.00	98.00
10	2		2.00	100.00
Total	100		100.00	

Figura 14.32 Distribuição de frequências para os dados de contagem da variável *atrasos*.

**Figura 14.33** Histograma da variável dependente *atrasos*.

. tabstat atrasos, stats(mean var)		
variable	mean	variance
atrasos	1.82	5.421818

Figura 14.34 Média e variância da variável dependente *atrasos*.

Conforme podemos verificar, a variância da variável dependente é aproximadamente 3 vezes maior do que a sua média, o que faz com que surjam indícios de existência de superdispersão.

Recomenda-se que toda modelagem em que a variável dependente contém dados de contagem seja iniciada por meio da estimação de um modelo de regressão Poisson. Desta forma, vamos digitar os seguintes comandos:

```
quietly poisson atrasos dist sem per
predict lambda
```

em que *lambda* é uma variável que corresponde aos valores previstos de ocorrência de atrasos mensalmente e é calculada com base na estimativa do modelo de regressão Poisson.

Desta forma, partiremos inicialmente para a aplicação do teste proposto por Cameron e Trivedi (1990) para verificação de existência de superdispersão nos dados da variável dependente, com base na expressão (14.14) e seguindo o procedimento já elaborado na seção 14.4.1. Assim, devemos digitar:

```
gen yasterisco = ((atrasos-lambda)^2 - atrasos)/lambda
reg yasterisco lambda, nocons
```

Os resultados deste procedimento encontram-se na Figura 14.35.

```

quietly poisson atrasos dist sem per

predict lambda
(option n assumed; predicted number of events)

gen yasterisco = ((atrasos-lambda)^2 - atrasos)/lambda

reg yasterisco lambda, nocons

Source |      SS       df      MS
-----+-----+-----+
Model | 12.8608941      1 12.8608941
Residual | 278.374591     99 2.81186456
-----+-----+
Total | 291.235486    100 2.91235486

Number of obs = 100
F( 1, 99) = 4.57
Prob > F = 0.0349
R-squared = 0.0442
Adj R-squared = 0.0345
Root MSE = 1.6769

yasterisco |   Coef.  Std. Err.      t    P>|t| [95% Conf. Interval]
-----+-----+
lambda | .1332397  .062301    2.14  0.035  .0096209  .2568584
-----+

```

Figura 14.35 Resultado do teste para verificação de existência de superdispersão no Stata.

Como o parâmetro β da variável *lambda* estimado por meio do modelo de regressão auxiliar apresentado na Figura 14.35 é, ao nível de significância de 5%, estatisticamente diferente de zero, **podemos concluir que os dados da variável dependente apresentam superdispersão**, fazendo com que o modelo de regressão Poisson estimado não seja adequado. Mais adiante teremos mais uma comprovação deste fato ao estimarmos a própria expressão da variância da variável dependente.

O teste χ^2 para comparar as distribuições de probabilidades observadas e previstas de ocorrência de atrasos mensais também indica a inexistência de qualidade do ajuste do modelo de regressão Poisson, ou seja, existem diferenças estatisticamente significantes entre os valores previstos e observados do número de atrasos que ocorrem mensalmente. O comando para a realização deste teste, que deve ser digitado após a estimativa elaborada por meio do comando **poisson**, é:

poisgof

O resultado deste teste χ^2 encontra-se na Figura 14.36.

```

poisgof

Goodness-of-fit chi2 = 145.2954
Prob > chi2(96) = 0.0009

```

Figura 14.36 Verificação da qualidade do ajuste do modelo de regressão Poisson estimado.

Portanto, partiremos para a estimativa de um modelo de regressão binomial negativo. O comando para a estimativa deste modelo, para este exemplo, é:

nbreg atrasos dist sem per

O comando **nbreg** elabora um modelo de regressão binomial negativo NB2 estimado por máxima verossimilhança (*negative binomial 2 regression model*), ou seja, considera uma especificação quadrática para a variância, conforme discutido quando da apresentação da expressão (14.27). Assim como para os modelos de regressão múltipla, de regressão logística binária e multinomial e de regressão Poisson, 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:

nbreg atrasos dist sem per, level(90)

Iremos seguir com a análise mantendo o nível padrão de confiança dos intervalos dos parâmetros, que é de 95%. Os resultados da estimação encontram-se na Figura 14.37 e são exatamente iguais aos calculados na seção 14.3.

Assim como os modelos de regressão Poisson, os modelos de regressão binomial negativo também fazem parte do grupo de modelos conhecidos por **Modelos Lineares Generalizados (Generalized Linear Models)**, e como estamos supondo que a variável dependente apresenta uma distribuição Poisson-Gama pelo fato de apresentar superdispersão nos dados, os resultados da estimação apresentados na Figura 14.37 também podem igualmente ser obtidos por meio da digitação do seguinte comando:

```
glm atrasos dist sem per, family(nbinomial ml)
```

em que o termo **ml** significa *maximum likelihood*.

```
. nbreg atrasos dist sem per

Fitting Poisson model:

Iteration 0: log likelihood = -160.97008
Iteration 1: log likelihood = -154.89761
Iteration 2: log likelihood = -154.89376
Iteration 3: log likelihood = -154.89376

Fitting constant-only model:

Iteration 0: log likelihood = -183.37156
Iteration 1: log likelihood = -182.64329
Iteration 2: log likelihood = -182.63662
Iteration 3: log likelihood = -182.63662

Fitting full model:

Iteration 0: log likelihood = -164.81888
Iteration 1: log likelihood = -163.03629
Iteration 2: log likelihood = -156.38042 (not concave)
Iteration 3: log likelihood = -155.02033
Iteration 4: log likelihood = -151.41164
Iteration 5: log likelihood = -151.31538
Iteration 6: log likelihood = -151.01444
Iteration 7: log likelihood = -151.0123
Iteration 8: log likelihood = -151.0123

Negative binomial regression                               Number of obs =      100
Dispersion = mean                                         LR chi2(3)    =     63.25
Log likelihood = -151.0123                                Prob > chi2   =     0.0000
                                                               Pseudo R2    =     0.1732

-----+
atrasos |      Coef.    Std. Err.      z    P>|z|    [95% Conf. Interval]
-----+
dist |   .3076544   .0712522    4.32    0.000    .1680026   .4473061
sem |   .1973366   .0495291    3.98    0.000    .1002612   .2944119
per |  -.9274356   .257023    -3.61    0.000   -1.431191   -.4236797
_cons |  -4.997447  1.249431    -4.00    0.000   -7.446287  -2.548607
-----+
/lnalpha |  -1.365232   .5276507          -2.399408  -.3310552
-----+
alpha |   .2553215   .1347206          .0907717   .7181655
-----+
Likelihood-ratio test of alpha=0: chibar2(01) =    7.76 Prob>=chibar2 = 0.003
```

Figura 14.37 Outputs do modelo de regressão binomial negativo no Stata.

Inicialmente, podemos verificar que o valor máximo do logaritmo da função de verossimilhança para o modelo completo é igual a -151,0123, que é exatamente igual ao valor calculado por meio do **Solver** do Excel (seção 14.3.1) e apresentado na Tabela 14.12 e na Figura 14.14. Caso o pesquisador deseje também obter o valor máximo do logaritmo da função de verossimilhança para o modelo nulo, deverá digitar o seguinte comando, cujos resultados encontram-se na Figura 14.38:

```
nbreg atrasos
```

```

. nbreg atrasos

Fitting Poisson model:

Iteration 0:  log likelihood = -223.36096
Iteration 1:  log likelihood = -223.36096

Fitting constant-only model:

Iteration 0:  log likelihood = -183.37156
Iteration 1:  log likelihood = -182.64329
Iteration 2:  log likelihood = -182.63662
Iteration 3:  log likelihood = -182.63662

Fitting full model:

Iteration 0:  log likelihood = -182.63662
Iteration 1:  log likelihood = -182.63662

Negative binomial regression                               Number of obs      =        100
Dispersion      = mean                                LR chi2(0)       =         0.00
Log likelihood   = -182.63662                          Prob > chi2     =
                                                       Pseudo R2       =        0.0000

-----+
           atrasos |   Coef.  Std. Err.      z    P>|z|      [95% Conf. Interval]
-----+
           _cons |  .5988365  .137895    4.34    0.000    .3285673  .8691057
-----+
           /lnalpha |  .3016238  .2430113          -.1746697  .7779172
-----+
           alpha |  1.352052  .3285641          .8397343  2.176933
-----+
Likelihood-ratio test of alpha=0:  chibar2(01) =  81.45  Prob>=chibar2 = 0.000

```

Figura 14.38 Outputs do modelo de regressão binomial negativo nulo no Stata.

Logo, o valor máximo do logaritmo da função de verossimilhança para o modelo nulo é igual a -182,63662, que é exatamente igual ao valor também calculado pelo **Solver** do Excel e apresentado na Figura 14.16.

Assim, fazendo uso da expressão (14.10), temos que:

$$\chi^2_{3g.l.} = -2 \cdot [-182,63662 - (-151,01230)] = 63,25 \quad \text{com valor -}P(\text{ou Prob. } \chi^2 \text{ cal}) = 0,000.$$

Logo, com base no teste χ^2 , podemos rejeitar a hipótese nula de que todos os parâmetros β_j ($j = 1, 2, 3$) sejam estatisticamente iguais a zero ao nível de significância de 5%, ou seja, pelo menos uma variável X é estatisticamente significante para explicar o número de atrasos que ocorre mensalmente ao se chegar à escola.

Também podemos calcular o pseudo R^2 de McFadden, como fizemos na seção 14.4.1, sempre lembrando, porém, que sua utilidade é bastante limitada e restringe-se à comparação de dois ou mais modelos de mesma classe, ou seja, não pode ser utilizado para se comparar, por exemplo, um modelo Poisson com um modelo binomial negativo. Assim, com base na expressão (14.9), temos que:

$$\text{pseudo } R^2 = \frac{-2 \cdot (-182,63662) - [(-2 \cdot (-151,01230))] }{-2 \cdot (-182,63662)} = 0,1732$$

Em relação à significância estatística dos parâmetros do modelo apresentado na Figura 14.37, como todos os valores de $z_{cal} < -1,96$ ou $> 1,96$, os *valores-P* das estatísticas z de Wald $< 0,05$ para todos os parâmetros estimados e, portanto, já chegamos ao modelo final de regressão binomial negativo, sem que haja necessidade de uma eventual aplicação do procedimento *Stepwise*. Se este não tivesse sido o caso, seria recomendável a estimativa do modelo final por meio de um dos seguintes comandos:

```

stepwise, pr(0.05): nbreg atrasos dist sem per
stepwise, pr(0.05): glm atrasos dist sem per, family(nbinomial ml)

```

que, para este nosso exemplo, geram exatamente os mesmos resultados apresentados na Figura 14.37.

Após a estimativa do modelo final de regressão binomial negativo, podemos gerar uma variável correspondente aos valores previstos de ocorrência de atrasos mensais por aluno, que chamaremos de u . Esta variável deverá ser gerada exatamente após a estimativa do modelo final, por meio da digitação do seguinte comando:

predict u

A expressão da quantidade média estimada de atrasos por mês para um determinado aluno i será dada, portanto, por:

$$u_i = e^{(-4,997 + 0,308 \cdot dist_i + 0,197 \cdot sem_i - 0,927 \cdot per_i)}$$

que, à exceção de pequenos arredondamentos, é exatamente o mesmo modelo estimado na seção 14.3. Além disso, também com base na Figura 14.37, as quantidades estimadas de atrasos por mês apresentam, com 95% de nível de confiança, expressões de mínimo e de máximo iguais a:

$$u_{i_{\min}} = e^{(-7,446 + 0,168 \cdot dist_i + 0,100 \cdot sem_i - 1,431 \cdot per_i)}$$

$$u_{i_{\max}} = e^{(-2,549 + 0,447 \cdot dist_i + 0,294 \cdot sem_i - 0,424 \cdot per_i)}$$

Além disso, a parte inferior da Figura 14.37 apresenta o *output* correspondente à estimativa de ϕ , que é o inverso do parâmetro de forma ψ da distribuição binomial negativa e que o Stata chama de *alpha*. Conforme podemos observar, o intervalo de confiança para ϕ (*alpha*) não contém o zero, ou seja, para o nível de confiança de 95%, podemos afirmar que ϕ é estatisticamente diferente de zero e com valor estimado igual a 0,255, conforme já calculado na seção 14.3.1 por meio do **Solver** do Excel (Figura 14.14). Os *outputs* da Figura 14.37 ainda apresentam o teste de razão de verossimilhança para o parâmetro ϕ (*alpha*), de onde se pode concluir que a hipótese nula de que este parâmetro seja estatisticamente igual a zero pode ser rejeitada ao nível de significância de 5% (*Sig.* $\chi^2 = 0,003 < 0,05$). **Isso comprova a existência de superdispersão nos dados**, ficando a variância da variável dependente, de acordo com a expressão (14.27), com a seguinte especificação:

$$Var(Y) = u + 0,255 \cdot u^2$$

O comando **glm** apresenta diretamente esta expressão de variância em seus *outputs*, conforme mostra a Figura 14.39, que equivale à Figura 14.37.

glm atrasos dist sem per, family(nbinomial ml)

Iteration 0: log likelihood = -151.49946						
Iteration 1: log likelihood = -151.01314						
Iteration 2: log likelihood = -151.0123						
Iteration 3: log likelihood = -151.0123						
 Generalized linear models						
Optimization : ML						
No. of obs = 100						
Residual df = 96						
Scale parameter = 1						
Deviance = 105.0249438						
(1/df) Deviance = 1.09401						
Pearson = 104.7027564						
(1/df) Pearson = 1.090654						
 Variance function: V(u) = u + (.2553)u^2						
Link function : g(u) = ln(u)						
[Neg. Binomial]						
[Log]						
 AIC = 3.100246						
BIC = -337.0714						
 Log likelihood = -151.0122975						
 OIM						
atrasos	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
dist	.3076544	.0680481	4.52	0.000	.1742826	.4410261
sem	.1973366	.0481042	4.10	0.000	.103054	.2916191
per	-.9274356	.2568699	-3.61	0.000	-.1.430891	-.42398
cons	-4.997447	1.17835	-4.24	0.000	-7.306971	-2.687923

Figura 14.39 Outputs do modelo de regressão binomial negativo no Stata – comando **glm**.

Se um pesquisador mais curioso estimar um modelo de regressão binomial negativo no banco de dados utilizado na seção 14.4.1 (**QuantAtrasosPoisson.dta**), verificará que ϕ (*alpha*) será estatisticamente igual a zero, o que já era de se esperar, visto que o teste para verificação de existência de superdispersão não rejeitou a hipótese nula de equidispersão para aquele caso (Figura 14.23). Em outras palavras, a estimação de um modelo de regressão Poisson para aquele banco de dados foi adequada, fato que não acontece neste nosso exemplo atual.

Desta forma, como $\phi \neq 0$, faz sentido continuarmos com a análise dos resultados obtidos pela estimação do modelo de regressão binomial negativo e, portanto, retornaremos à primeira pergunta proposta ao final da seção 14.3.1 e respondida na seção 14.3.2:

Qual é a quantidade média esperada de atrasos no mês quando se desloca 12 quilômetros e se passa por 17 semáforos diariamente, sendo o trajeto feito à tarde?

Para responder a esta pergunta, vamos novamente utilizar o comando **mfx**, digitando o seguinte:

```
mfx, at(dist=12 sem=17 per=0)
```

Com base na Figura 14.40, e conforme já calculado manualmente na seção 14.3.2, espera-se, portanto, que determinado aluno que é submetido a estas características ao se deslocar à escola apresente, em média, uma quantidade de 7,76 atrasos por mês.

Marginal effects after nbreg							
variable	dy/dx	Std. Err.	z	P> z	[95% C.I.]
dist	2.387744	.79926	2.99	0.003	.821228	3.95426	12
sem	1.531554	.54557	2.81	0.005	.462264	2.60084	17
per*	-4.691082	1.65951	-2.83	0.005	-7.94366	-1.4385	0

(* dy/dx is for discrete change of dummy variable from 0 to 1)

Figura 14.40 Cálculo da quantidade esperada de atrasos mensais para valores das variáveis explicativas – comando **mfx**.

Analogamente ao elaborado para os modelos de regressão Poisson, podemos também aqui obter diretamente as estimativas das taxas de incidência mensal de atrasos quando se altera em uma unidade determinada variável explicativa, mantidas as demais condições constantes. Desta forma, para o nosso modelo de regressão binomial negativo, podemos digitar:

```
nbreg atrasos dist sem per, irr
```

Os resultados, apresentados na Figura 14.41, também poderiam ser obtidos por meio do seguinte comando:

```
glm atrasos dist sem per, family(nbomial ml) eform
```

em que, neste caso, o termo **eform** do comando **glm** equivale ao termo **irr** do comando **nbreg**.

Desta maneira, podemos retornar às três últimas perguntas propostas ao final da seção 14.3.1:

Em média, em quanto se altera a taxa de incidência mensal de atrasos ao se adotar um percurso 1 quilômetro mais longo, mantidas as demais condições constantes?

Em média, em quanto se altera a taxa de incidência mensal de atrasos ao se passar por 1 semáforo a mais no percurso até a escola, mantidas as demais condições constantes?

Em média, em quanto se altera a taxa de incidência mensal de atrasos ao se optar por ir à escola de manhã, em vez de se ir à tarde, mantidas as demais condições constantes?

As respostas agora podem ser dadas de maneira direta, ou seja, enquanto a taxa de incidência mensal de atrasos ao se adotar um percurso 1 quilômetro mais longo é, em média e mantidas as demais condições constantes, multiplicada por um fator de 1,360 (36,0% maior), a taxa de incidência mensal de atrasos ao se adotar um

```

. nbreg atrasos dist sem per, irr

Fitting Poisson model:

Iteration 0: log likelihood = -160.97008
Iteration 1: log likelihood = -154.89761
Iteration 2: log likelihood = -154.89376
Iteration 3: log likelihood = -154.89376

Fitting constant-only model:

Iteration 0: log likelihood = -183.37156
Iteration 1: log likelihood = -182.64329
Iteration 2: log likelihood = -182.63662
Iteration 3: log likelihood = -182.63662

Fitting full model:

Iteration 0: log likelihood = -164.81888
Iteration 1: log likelihood = -163.03629
Iteration 2: log likelihood = -156.38042 (not concave)
Iteration 3: log likelihood = -155.02033
Iteration 4: log likelihood = -151.41164
Iteration 5: log likelihood = -151.31538
Iteration 6: log likelihood = -151.01444
Iteration 7: log likelihood = -151.0123
Iteration 8: log likelihood = -151.0123

Negative binomial regression
Number of obs      =       100
LR chi2(3)        =      63.25
Dispersion = mean
Prob > chi2       =     0.0000
Log likelihood    = -151.0123
Pseudo R2         =     0.1732

-----
atrasos |      IRR      Std. Err.      z     P>|z|      [95% Conf. Interval]
-----+
dist |   1.360231   .0969194     4.32     0.000     1.18294   1.564093
sem |   1.218154   .0603341     3.98     0.000     1.10546   1.342337
per |   .3955668   .1016698    -3.61     0.000     .239024   .6546335
-----+
/lnalpha |  -1.365232   .5276507          -2.399408   -.3310552
-----+
alpha |   .2553215   .1347206          .0907717   .7181655
-----+
Likelihood-ratio test of alpha=0: chibar2(01) =    7.76 Prob>chibar2 = 0.003

```

Figura 14.41 Outputs do modelo de regressão binomial negativo – incidence rate ratios.

percurso com 1 semáforo a mais é, em média e também mantidas as demais condições constantes, multiplicada por um fator de 1,218 (21,8% maior). Por fim, a taxa de incidência mensal de atrasos ao se optar por ir à escola de manhã, em vez de se ir à tarde, é, em média, multiplicada por um fator de 0,396 (60,4% menor), mantidas as demais condições constantes. Estes valores são exatamente os mesmos daqueles calculados manualmente ao final da seção 14.3.2.

Imagine, portanto, que tenhamos o interesse de, por exemplo, visualizar, por meio de um gráfico, o comportamento da evolução da quantidade mensal prevista de atrasos em função da quantidade existente de semáforos no percurso até a escola, porém separando os trajetos realizados de manhã ou à tarde. Para tanto, podemos digitar o seguinte comando:

```

graph twoway scatter u sem if per==0 || scatter u sem if per==1
|| mspline u sem if per==0 || mspline u sem if per==1 ||,
legend(label(3 "tarde") label(4 "manhã"))

```

O gráfico gerado encontra-se na Figura 14.42.

Por meio deste gráfico é possível verificar que os trajetos para se chegar à escola realizados no período da tarde possuem quantidades maiores de semáforos, em média, provavelmente porque os estudantes que se deslocam até a escola no período vespertino partem de locais mais distantes. Enquanto a quantidade esperada de atrasos por mês para os percursos realizados de manhã não apresenta média superior a 1,5 e não ultrapassa o valor de 4, a quantidade esperada de atrasos por mês para os percursos realizados à tarde e, portanto, que apresentam maiores quantidades de semáforos, apresenta média em torno de 8, com valor mínimo ficando próximo de 4.

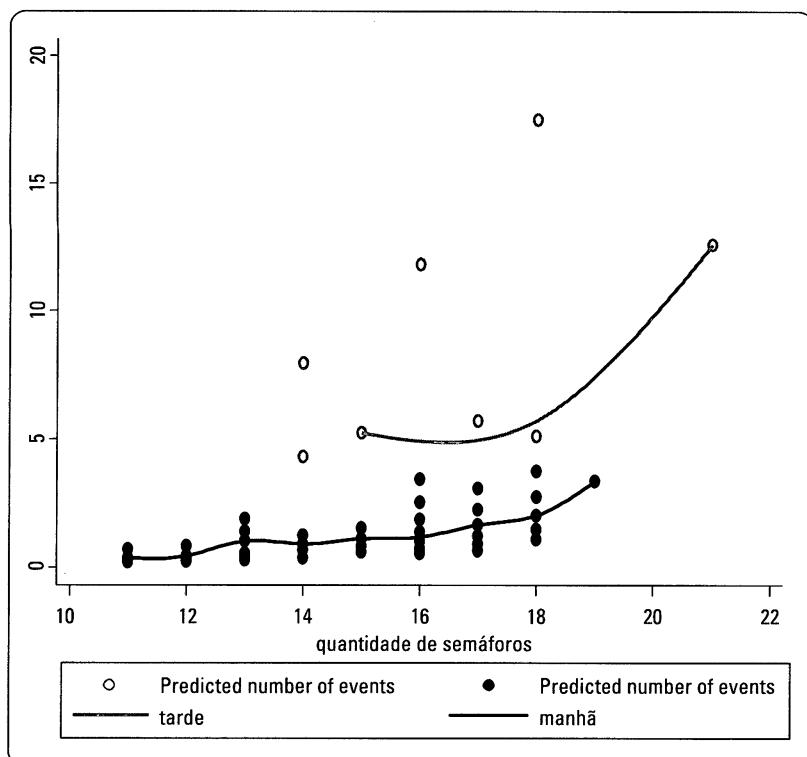


Figura 14.42 Quantidade esperada de atrasos por mês (u) x quantidade de semáforos (sem) em diferentes períodos do dia (per).

De maneira global, é possível claramente perceber que percursos com uma quantidade maior de semáforos levam a um aumento da quantidade esperada de atrasos por mês, com taxa média de incremento de 21,8% de atrasos a cada 1 semáforo adicional.

Por fim, vamos comparar as estimações dos modelos de regressão Poisson e binomial negativo elaboradas para este nosso exemplo. Primeiramente, a fim de que possamos comparar as distribuições de probabilidades observadas e previstas de ocorrência de atrasos mensais para estas duas estimativas, devemos digitar a seguinte sequência de comandos, que gerará o gráfico da Figura 14.43:

```
quietly poisson atrasos dist sem per
prcounts prpoisson, plot
quietly nbreg atrasos dist sem per
prcounts prbneg, plot
graph twoway (scatter prbnegobeq prbnegpreq prpoissonpreq
prbnegval, connect (1 1 1))
```

Por meio da análise deste gráfico, podemos verificar que a distribuição estimada (prevista) de probabilidades do modelo binomial negativo se ajusta melhor à distribuição observada (pontos mais próximos) do que a distribuição estimada de probabilidades do modelo Poisson.

Este fato também pode ser verificado quando se aplica o comando **countfit**, que oferece os valores destas probabilidades previstas para cada contagem da variável dependente. Assim, podemos digitar a seguinte sequência de comandos:

```
countfit atrasos dist sem per, prm nograph noestimates nofit
countfit atrasos dist sem per, nbreg nograph noestimates nofit
```

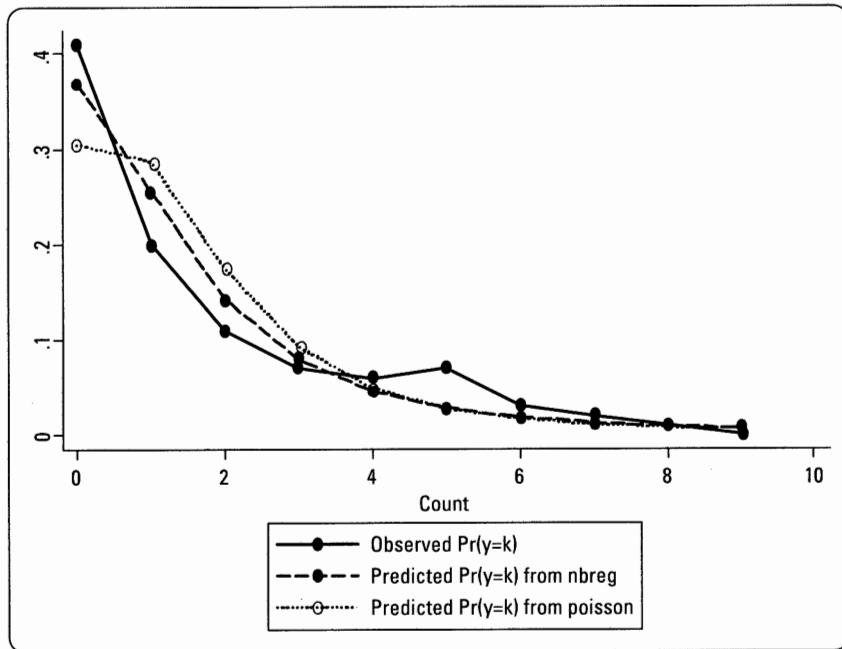


Figura 14.43 Distribuições de probabilidades observadas e previstas de ocorrência de atrasos mensais para os modelos de regressão Poisson e binomial negativo.

em que o termo **prm** refere-se ao modelo Poisson e o termo **nbreg**, ao modelo binomial negativo (NB2). Os *outputs* encontram-se na Figura 14.44.

As colunas **Actual** e **Predicted** dos *outputs* da Figura 14.44 referem-se, respectivamente, às probabilidades observadas e previstas para cada um dos modelos estimados e, por meio delas, também poderia ter sido obtido o gráfico da Figura 14.43.

Podemos verificar que o ajuste do modelo de regressão binomial negativo é melhor do que o ajuste do modelo de regressão Poisson. Isso pode inicialmente já ser percebido pela análise da diferença máxima entre as probabilidades observadas e previstas que, enquanto para o modelo Poisson, é de 0,105, para o modelo binomial negativo é, em módulo, igual a 0,056. Além disso, a média destas diferenças é de 0,036 para o modelo Poisson e de 0,022 para o modelo binomial negativo. Enquanto os valores da coluna **|Diff|** correspondem a estas diferenças em módulo para cada contagem da variável dependente (de 0 a 9), os valores da coluna **Pearson**, segundo Cameron e Trivedi (2009), representam um bom indicador do ajuste do modelo e são calculados com base na seguinte expressão:

$$\text{Pearson} = N \cdot \frac{(\text{Diff})^2}{\text{Predicted}} \quad (14.31)$$

em que N é o tamanho da amostra. Conforme também podemos verificar por meio da análise destes mesmos *outputs* (Figura 14.44), o valor total de **Pearson** é mais baixo para o modelo de regressão binomial negativo, indicando o seu melhor ajuste em relação ao modelo de regressão Poisson.

Além disso, podemos elaborar um gráfico que relaciona as quantidades previstas com as quantidades observadas de atrasos mensais para cada observação da amostra, para os modelos de regressão Poisson e binomial negativo estimados para o banco de dados deste exemplo. É importante lembrarmos que, enquanto a variável u corresponde aos valores previstos de ocorrência de atrasos mensais por aluno obtidos pelo modelo binomial negativo, a variável $lambda$ corresponde a estes valores previstos pelo modelo Poisson. Assim, devemos digitar o seguinte comando, a fim de que seja gerado o gráfico da Figura 14.45:

```
graph twoway mspline u atrasos || mspline lambda atrasos ||,
legend(label(1 "Binomial Negativo") label(2 "Poisson"))
```

Comparison of Mean Observed and Predicted Count				
Model	Maximum Difference	At Value	Mean Diff	
PRM	0.105	0	0.036	
PRM: Predicted and actual probabilities				
Count	Actual	Predicted	Diff	Pearson
0	0.410	0.305	0.105	3.632
1	0.200	0.287	0.087	2.651
2	0.110	0.175	0.065	2.410
3	0.070	0.093	0.023	0.564
4	0.060	0.049	0.011	0.242
5	0.070	0.028	0.042	6.516
6	0.030	0.017	0.013	1.028
7	0.020	0.011	0.009	0.706
8	0.010	0.008	0.002	0.054
9	0.000	0.006	0.006	0.604
Sum	0.980	0.979	0.364	18.408
. countfit atrasos dist sem per, nbreg nograph noestimates nofit				
Comparison of Mean Observed and Predicted Count				
Model	Maximum Difference	At Value	Mean Diff	
NBRM	-0.056	1	0.022	
NBRM: Predicted and actual probabilities				
Count	Actual	Predicted	Diff	Pearson
0	0.410	0.369	0.041	0.451
1	0.200	0.256	0.056	1.234
2	0.110	0.143	0.033	0.756
3	0.070	0.079	0.009	0.105
4	0.060	0.046	0.014	0.426
5	0.070	0.028	0.042	6.085
6	0.030	0.019	0.011	0.704
7	0.020	0.013	0.007	0.416
8	0.010	0.009	0.001	0.009
9	0.000	0.007	0.007	0.671
Sum	0.980	0.969	0.221	10.858

Figura 14.44 Probabilidades observadas e previstas para cada contagem da variável dependente e respectivos termos de erro.

Esta figura mostra que a variância da quantidade prevista de atrasos mensais é bem superior para o caso do modelo de regressão binomial negativo, cuja estimação consegue capturar a existência de superdispersão nos dados. Para o exemplo utilizado na seção 14.4.1, caso tivéssemos elaborado este mesmo gráfico, resultante das estimações do modelo de regressão Poisson e do modelo de regressão binomial negativo, as duas curvas seriam exatamente iguais (superpostas), o que demonstra, mais uma vez, que a estimação do modelo de regressão Poisson, naquele caso, foi adequada, ao contrário da presente situação, em que prevalece a estimação do modelo de regressão binomial negativo.

Por fim, assim como fizemos ao final da seção 14.4.1, podemos desejar comparar os resultados do modelo de regressão binomial negativo estimado por máxima verossimilhança com os resultados obtidos por outras estimações como, no caso, aqueles obtidos pelo modelo de regressão Poisson também estimado por máxima verossimilhança e os obtidos por um eventual modelo de regressão múltipla log-linear estimado por mínimos quadrados ordinários (*ordinary least squares*, ou *OLS*). Para tanto, vamos inicialmente gerar uma variável chamada de *lnatrasos*, que corresponde ao logaritmo natural da variável dependente *atrasos*, por meio da digitação do seguinte comando:

```
gen lnatrasos=ln(atrasos)
```

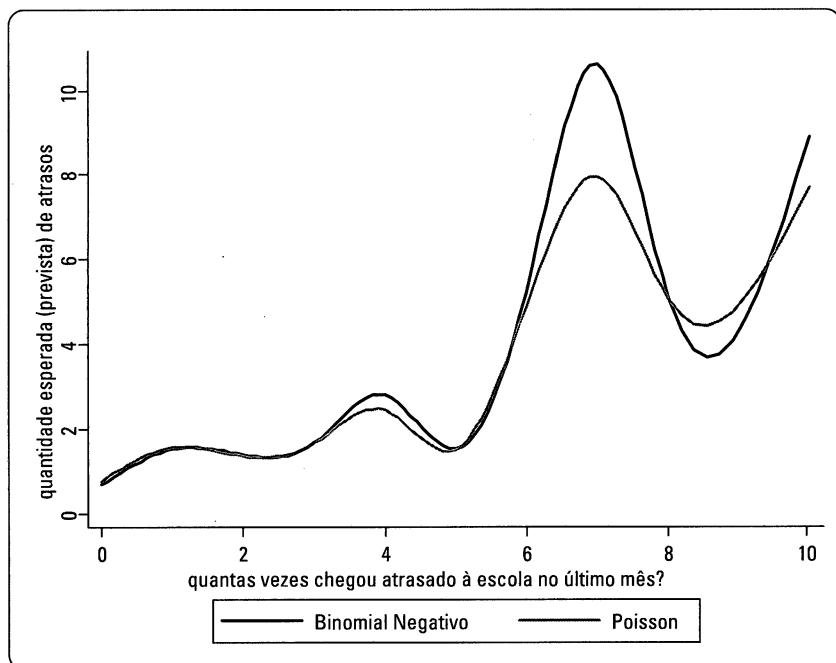


Figura 14.45 Quantidade prevista x quantidade real de atrasos mensais para os modelos binomial negativo e Poisson.

Na sequência, vamos estimar o modelo $\ln atrasos_i = \alpha + \beta_1 \cdot dist_i + \beta_2 \cdot sem_i + \beta_3 \cdot per_i$ por OLS, gerando no banco de dados uma variável correspondente aos valores previstos, para cada observação, da quantidade de atrasos mensais (variável *eyhat*), por meio da digitação da seguinte sequência de comandos:

```
quietly reg lnatrasos dist sem per
predict yhat
gen eyhat = exp(yhat)
```

O gráfico apresentado na Figura 14.46 oferece uma oportunidade de verificação, por meio de ajustes lineares, das diferenças dos valores previstos em função dos valores reais da variável dependente entre as estimações elaboradas (modelos de regressão binomial negativo e Poisson estimados por máxima verossimilhança e modelo de regressão múltipla log-linear estimado por OLS). O comando para elaboração deste gráfico é:

```
graph twoway lfit u atrasos || lfit lambda atrasos ||
lfit eyhat atrasos ||, legend(label(1 "Binomial Negativo")
label(2 "Poisson") label(3 "OLS"))
```

Este gráfico nos mostra que o modelo binomial negativo estimado acabou por gerar valores previstos mais similares aos valores reais da variável dependente, visto que seu ajuste linear é consistentemente mais próximo de uma reta imaginária com inclinação de 45°, principalmente para valores mais elevados de *Y*. Os modelos de regressão Poisson e log-linear, por outro lado, geraram estimativas viesadas dos parâmetros em relação ao modelo de regressão binomial negativo, o que demonstra que é fundamental que o pesquisador elabore diagnósticos preliminares sobre o comportamento da distribuição e a natureza da variável dependente antes da estimação de determinado modelo de regressão. Enquanto a presença de uma variável dependente quantitativa não garante a qualidade do ajuste de um modelo de regressão múltipla estimado por OLS, uma variável dependente quantitativa que contém dados de contagem também não garante a qualidade do ajuste de um modelo de regressão Poisson.

A capacidade do Stata para a elaboração dos mais diversos tipos de modelos é enorme, porém acreditamos que o que foi exposto aqui é considerado obrigatório para pesquisadores que tenham a intenção de estimar, de forma correta, os modelos de regressão para dados de contagem.

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

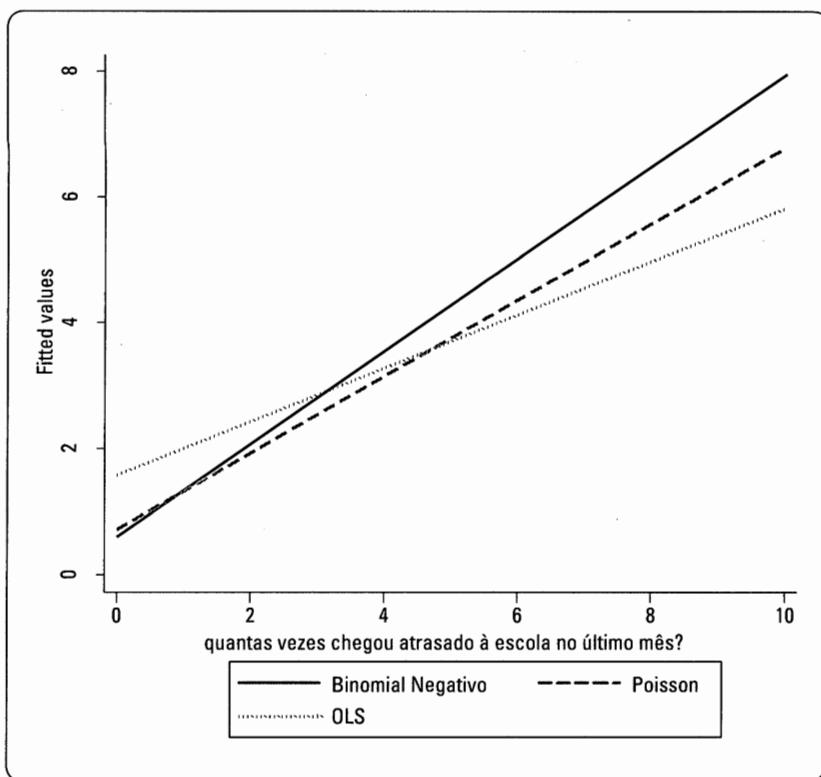


Figura 14.46 Valores previstos x valores observados para os modelos de regressão binomial negativo, Poisson e log-linear (OLS).

14.5. ESTIMAÇÃO DE MODELOS DE REGRESSÃO PARA DADOS DE CONTAGEM 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®.

Assim como realizado nos capítulos anteriores, nosso objetivo não é apresentar 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 estimar os modelos de regressão para dados de contagem 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, possa decidir qual software utilizar, em função das características de cada um e da própria acessibilidade para uso.

14.5.1. Modelo de regressão Poisson 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 **QuantAtrasosPoisson.sav** e, após o abrirmos, vamos inicialmente clicar em **Analyze → Descriptive Statistics → Frequencies...**, a fim de elaborarmos o primeiro diagnóstico sobre a distribuição da variável dependente. A caixa de diálogo da Figura 14.47 será aberta.

Conforme mostra esta figura, devemos inserir a variável dependente *atrasos* (quantas vezes chegou atrasado à escola na última semana?) em **Variable(s)**. No botão **Statistics...**, devemos marcar as opções **Mean** e **Variance**, conforme mostra a Figura 14.48.

Ao clicarmos em **Continue**, voltaremos à caixa de diálogo anterior. No botão **Charts...**, marcaremos a opção **Histograms**, conforme mostra a Figura 14.49.

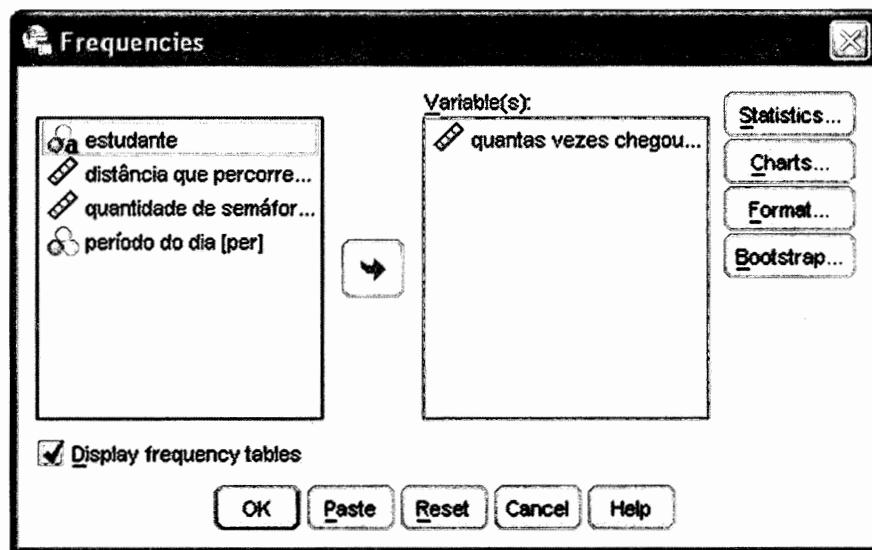


Figura 14.47 Caixa de diálogo para elaboração da tabela de frequências da variável dependente.

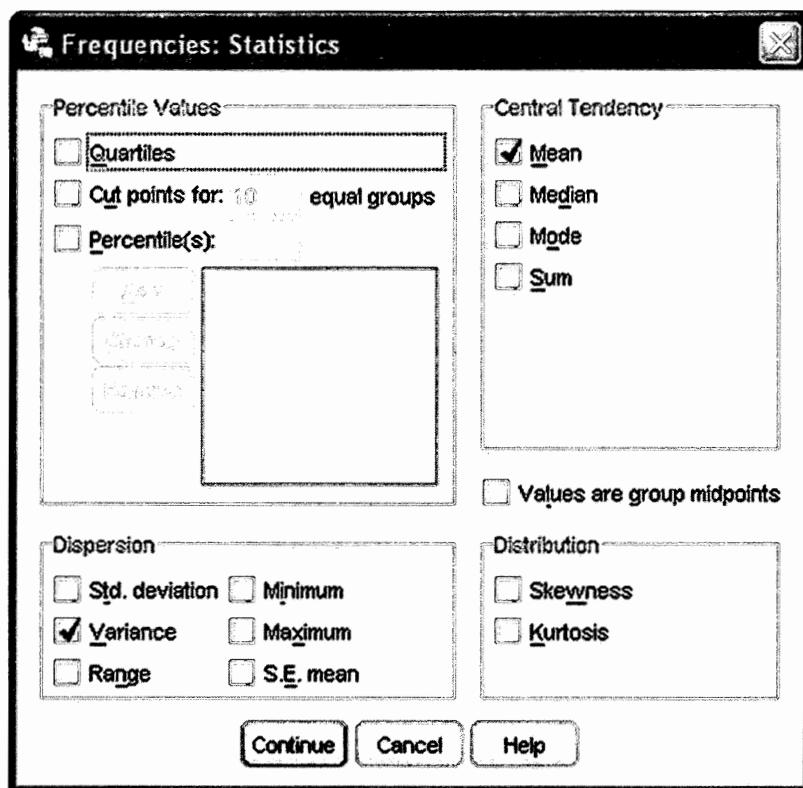


Figura 14.48 Seleção para cálculo da média e da variância da variável dependente.

Na sequência, devemos clicar em **Continue** e em **OK**. Os *outputs* encontram-se na Figura 14.50.

Estes *outputs* são os mesmos daqueles apresentados na Tabela 14.3 e na Figura 14.3 da seção 14.2.1 e também nas Figuras 14.18, 14.19 e 14.20 da seção 14.4.1 e, por meio deles, podemos verificar, ainda que de forma preliminar, que há indícios de inexistência de superdispersão nos dados, uma vez que a média e a variância são muito próximas. Partiremos, portanto, para a estimativa de um modelo de regressão Poisson, e, a partir de seus resultados, iremos elaborar o teste para verificação de existência de superdispersão.

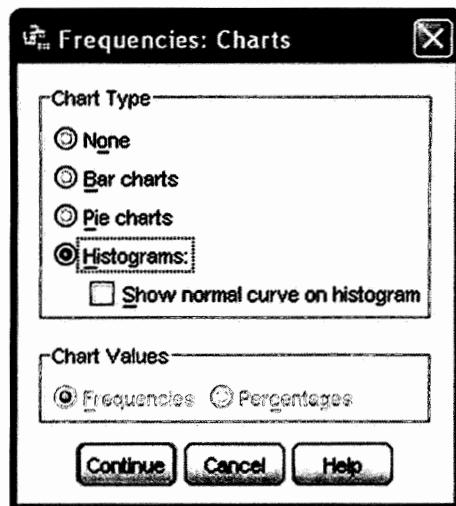


Figura 14.49 Caixa de diálogo para elaboração do histograma da variável dependente.

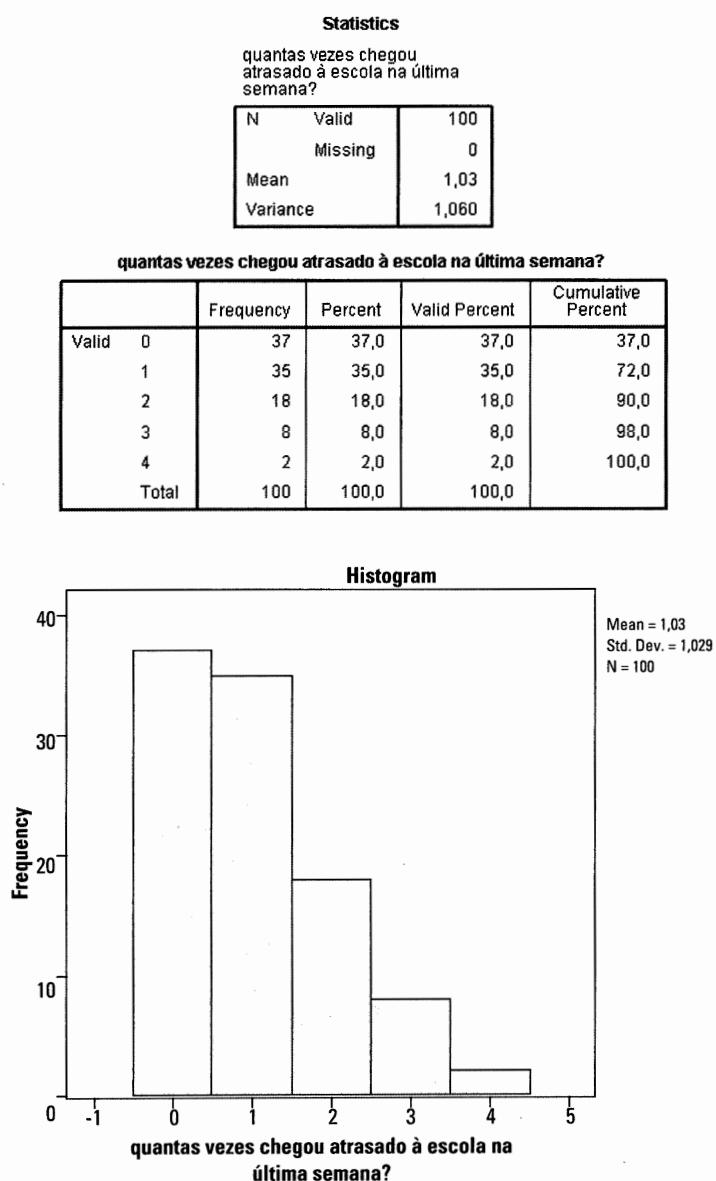


Figura 14.50 Média, variância, tabela de frequências e histograma da variável dependente.

Assim sendo, vamos clicar em **Analyze → Generalized Linear Models → Generalized Linear Models....**. Uma caixa de diálogo será aberta e devemos marcar, na pasta **Type of Model**, a opção **Poisson loglinear** (em **Counts**), conforme mostra a Figura 14.51.

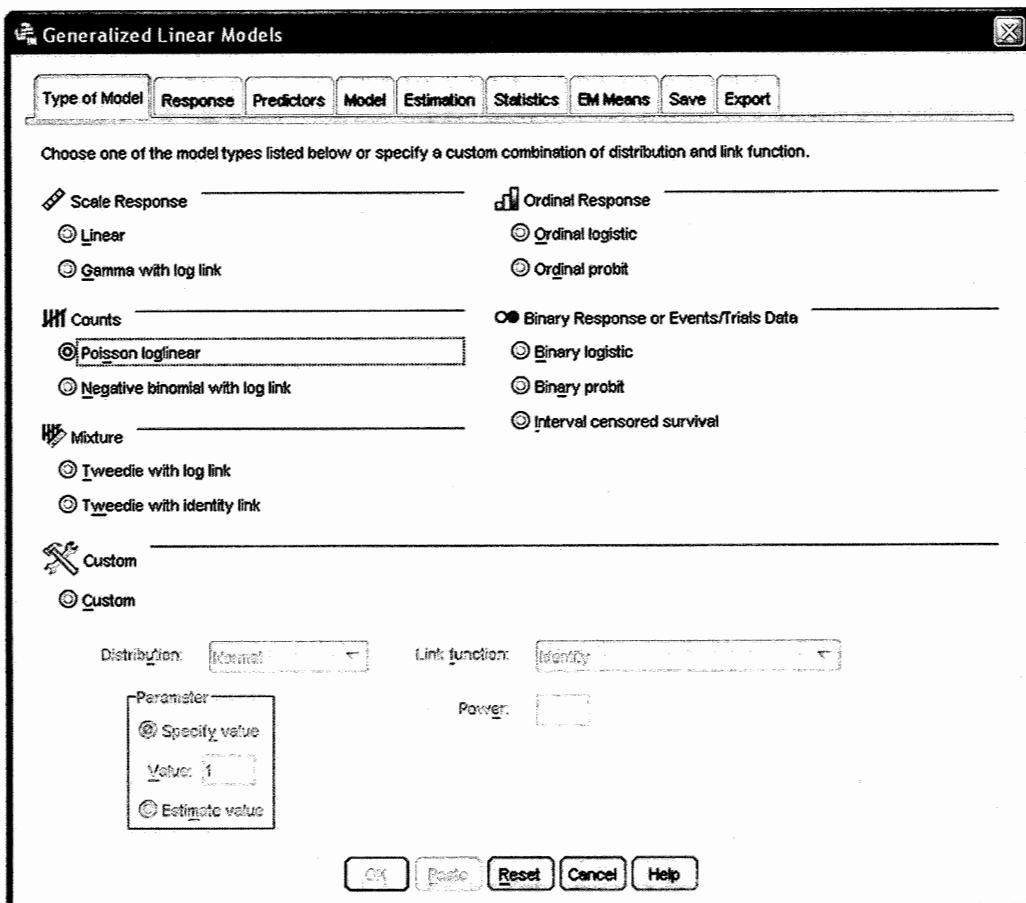


Figura 14.51 Caixa de diálogo inicial para estimação do modelo Poisson no SPSS.

É importante ressaltar que o pesquisador pode fazer uso desta mesma caixa de diálogo caso deseje estimar, por exemplo, um modelo de regressão múltipla ou um modelo de regressão logística, visto que estes também compõem os chamados **Modelos Lineares Generalizados**.

Na pasta **Response**, devemos incluir a variável *atrasos* na caixa **Dependent Variable**, conforme mostra a Figura 14.52.

Enquanto na pasta **Predictors** devemos incluir as variáveis *dist*, *sem* e *per* na caixa **Covariates**, na pasta **Model** devemos inserir estas mesmas três variáveis na caixa **Model**, conforme mostram, respectivamente, as Figuras 14.53 e 14.54.

Na pasta **Statistics**, além das opções já selecionadas de forma padrão pelo SPSS, devemos marcar também a opção **Include exponential parameter estimates**, conforme mostra a Figura 14.55.

Por fim, conforme mostra a Figura 14.56, marcaremos, na pasta **Save**, apenas a primeira opção, ou seja, **Predicted value of mean response**, que criará no banco de dados uma variável correspondente a λ_i (quantidade prevista de atrasos semanais por aluno).

Na sequência, devemos clicar em **OK**. A Figura 14.57 apresenta os principais *outputs* da estimação.

O primeiro *output* da estimação (**Goodness of Fit**) apresenta o valor da somatória do logaritmo da função de máxima verossimilhança da estimação proposta (*Log Likelihood*), que é de -107,615 e é exatamente igual ao valor obtido quando da modelagem no Excel (Tabela 14.5 e Figura 14.6) e no Stata (Figuras 14.21 e 14.27). Por meio do mesmo *output* podemos também verificar que a qualidade do ajuste do modelo estimado é adequada, visto que, para um $\chi^2_{cal} = 67,717$ (o SPSS chama de *Deviance*), temos, para 96 graus de liberdade, que $Sig. \chi^2 > 0,05$, ou seja,

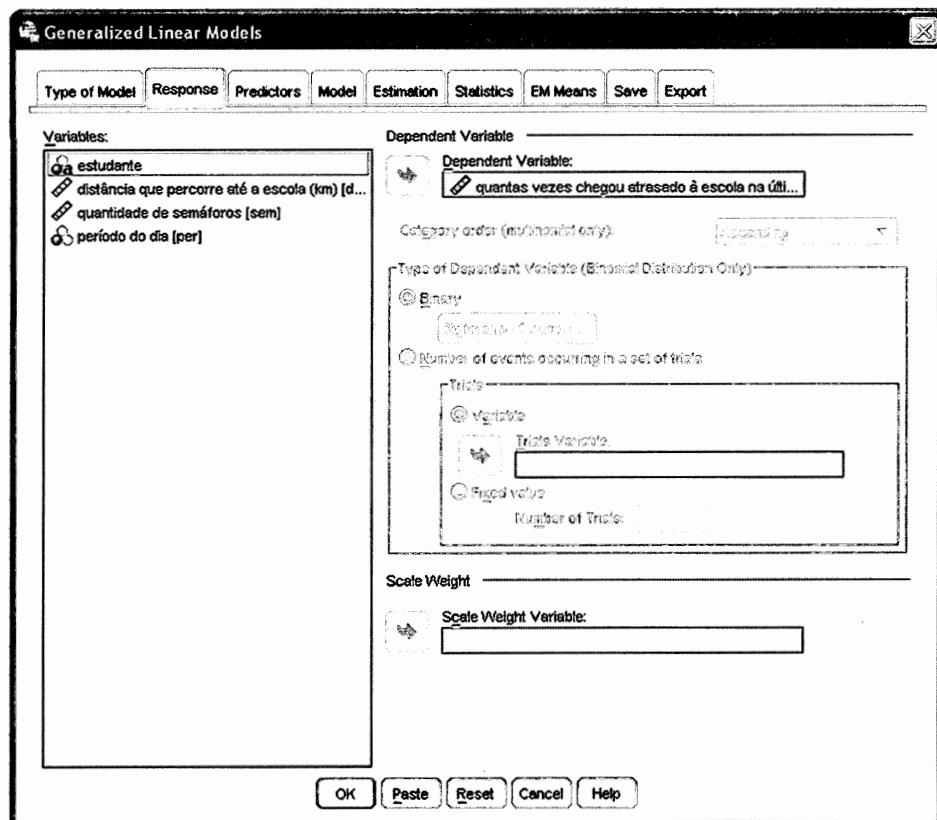


Figura 14.52 Caixa de diálogo para seleção da variável dependente.

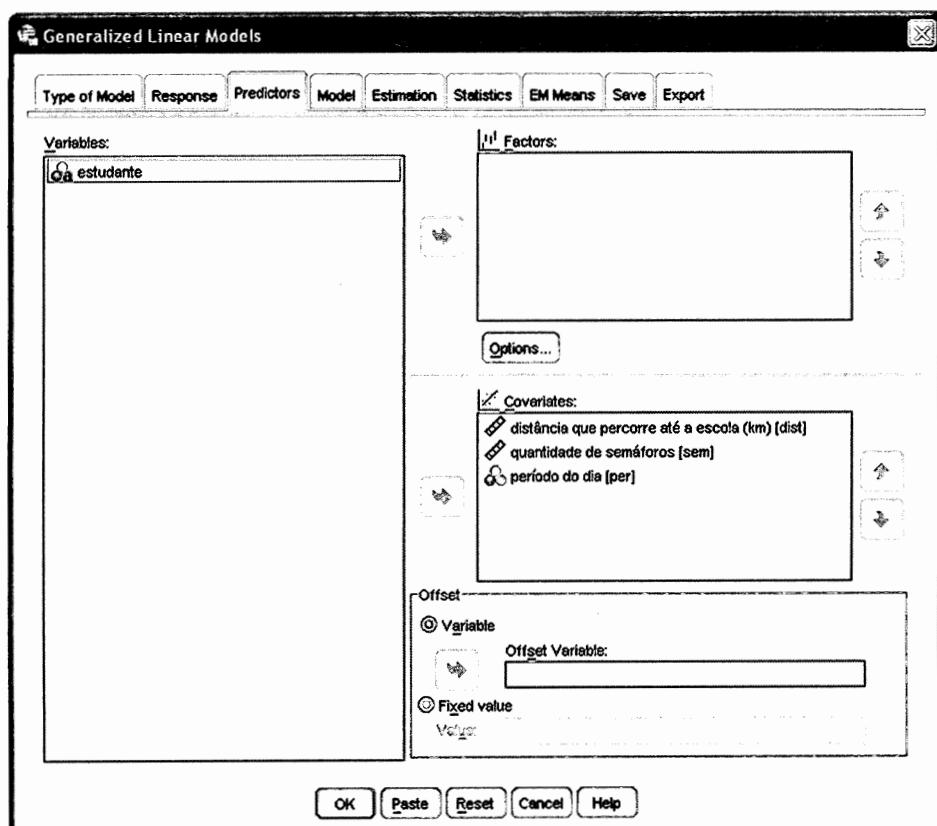


Figura 14.53 Caixa de diálogo para seleção das variáveis explicativas.

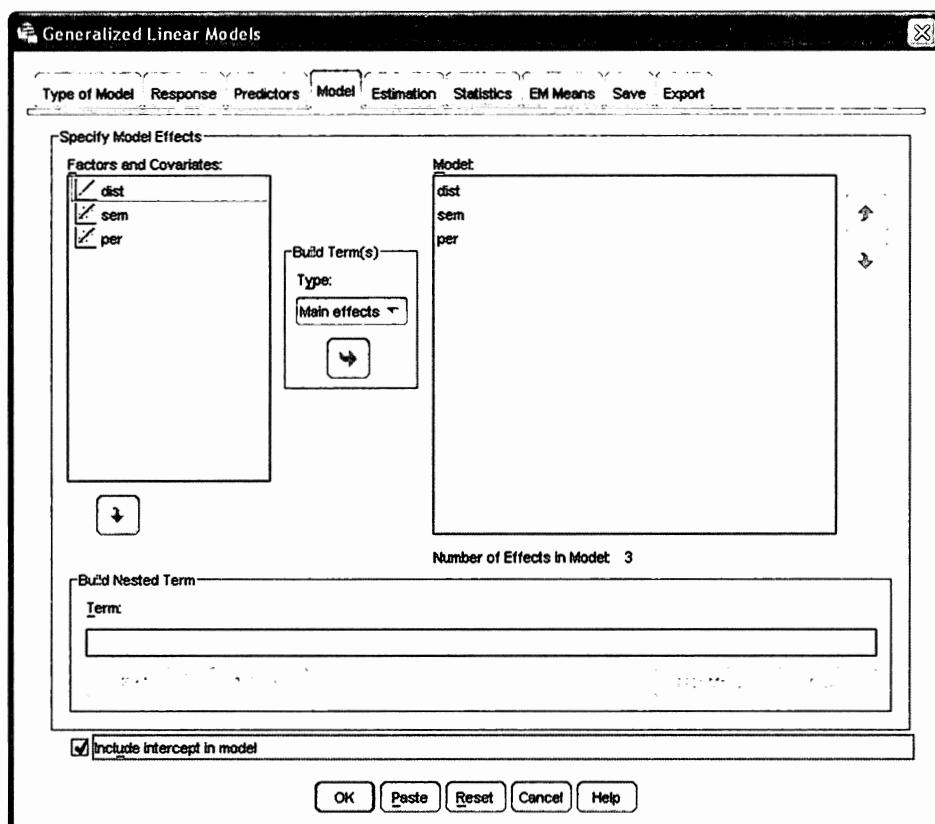


Figura 14.54 Caixa de diálogo para inclusão das variáveis explicativas na estimação do modelo.

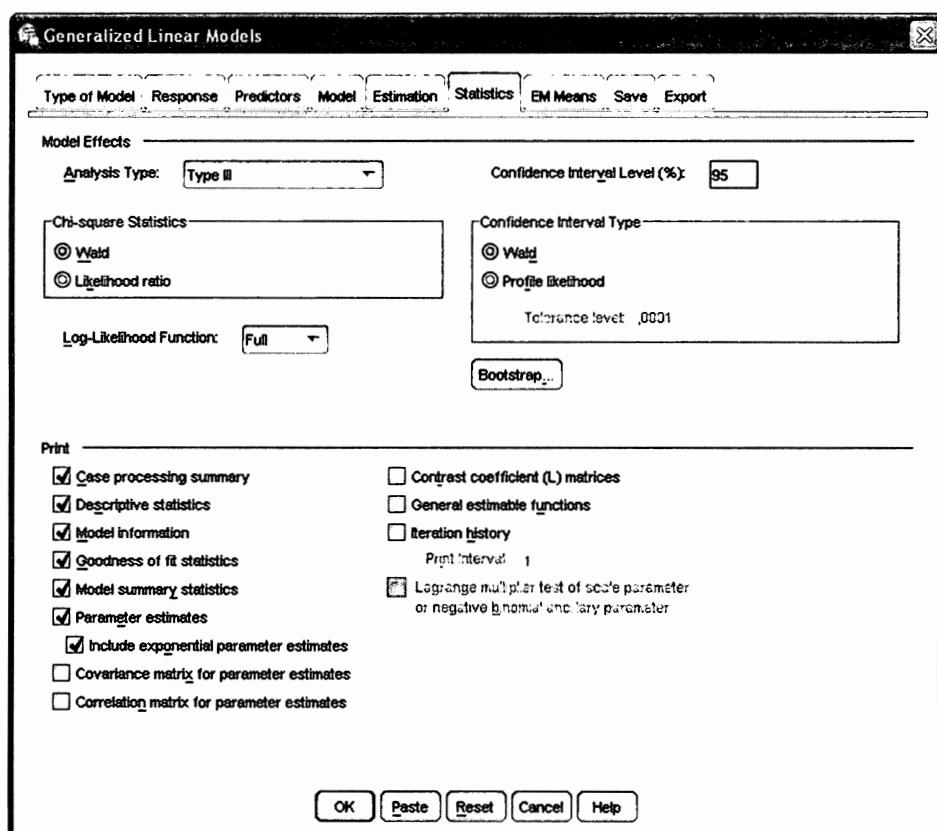


Figura 14.55 Caixa de diálogo para seleção das estatísticas do modelo de regressão Poisson.

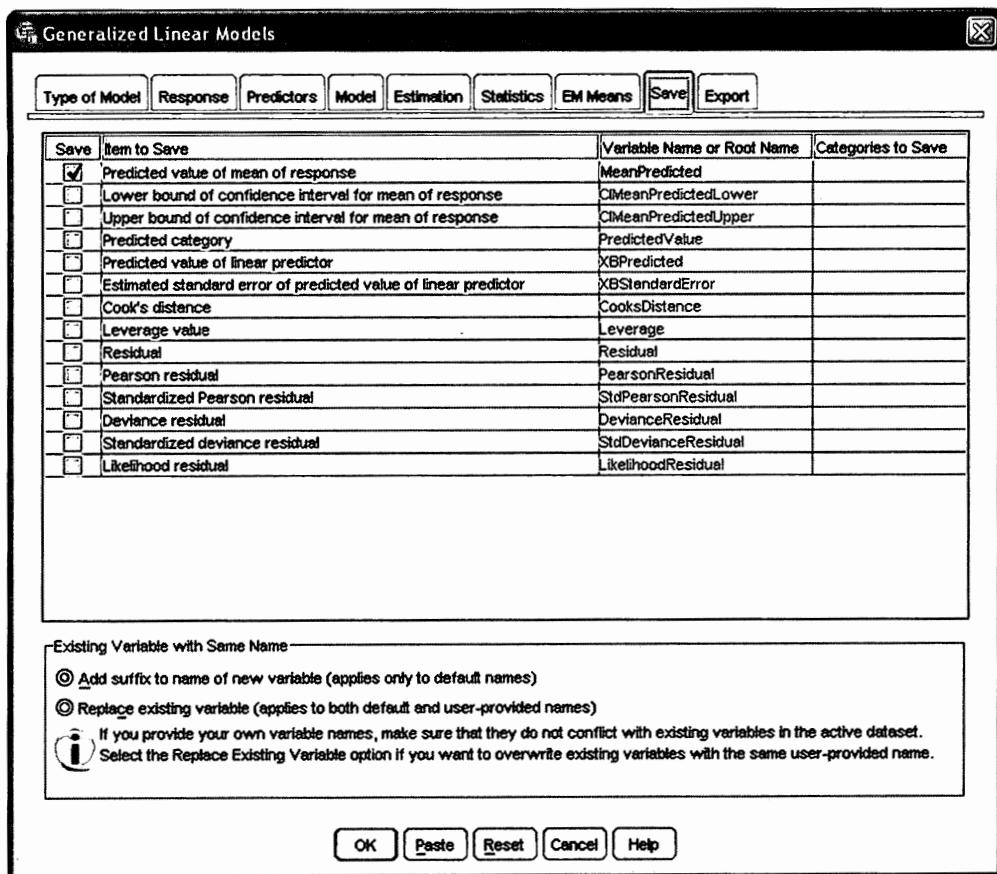


Figura 14.56 Caixa de diálogo para criação da variável λ_i , referente ao número previsto de atrasos semanais por aluno.

não existem diferenças estatisticamente significantes entre os valores previstos e observados do número de atrasos que ocorrem semanalmente. Esta parte do *output* corresponde ao apresentado na Figura 14.25 quando da estimação do modelo pelo Stata.

Podemos também verificar, com base no teste χ^2 (*Likelihood Ratio Chi-Square* = 51,015, *Sig. χ^2* = 0,000 < 0,05 apresentado no *output Omnibus Test*), que a hipótese nula de que todos os parâmetros β_j ($j = 1, 2, 3$) sejam estatisticamente iguais a zero pode ser rejeitada ao nível de significância de 5%, ou seja, pelo menos uma variável X é estatisticamente significante para explicar a ocorrência de atrasos por semana.

Os parâmetros estimados encontram-se no *output Parameter Estimates* e são exatamente iguais aos calculados manualmente e apresentados na Figura 14.6 (Excel) e também obtidos por meio do comando **poisson** do Stata (Figura 14.21). Este mesmo *output* também apresenta as *incidence rate ratios* (ou *IRR*) de cada variável explicativa, que o SPSS chama de *Exp(B)*, conforme também já apresentado por meio da Figura 14.27. Como todos os intervalos de confiança dos parâmetros estimados (95% *Wald Confidence Interval*) não contêm o zero e, consequentemente, os de *Exp(B)* não contêm o 1, já chegamos ao modelo final de regressão Poisson (todos os *Sig. Wald Chi-Square* < 0,05).

Portanto, a expressão da quantidade média estimada de atrasos por semana para um determinado aluno i pode ser escrita como:

$$\lambda_i = e^{(-4,380 + 0,222 \cdot dist_i + 0,165 \cdot sem_i - 0,573 \cdot per_i)}$$

com expressões de mínimo e máximo, a 95% de nível de confiança, iguais a:

$$\lambda_{i_{\min}} = e^{(-6,654 + 0,093 \cdot dist_i + 0,075 \cdot sem_i - 1,086 \cdot per_i)}$$

$$\lambda_{i_{\max}} = e^{(-2,106 + 0,351 \cdot dist_i + 0,254 \cdot sem_i - 0,060 \cdot per_i)}$$

Goodness of Fit ^b			
	Value	df	Value/df
Deviance	67,717	96	,705
Scaled Deviance	67,717	96	
Pearson Chi-Square	73,043	96	,761
Scaled Pearson Chi-Square	73,043	96	
Log Likelihood ^a	-107,615		
Akaike's Information Criterion (AIC)	223,230		
Finite Sample Corrected AIC (AICC)	223,651		
Bayesian Information Criterion (BIC)	233,651		
Consistent AIC (CAIC)	237,651		

Dependent Variable: quantas vezes chegou atrasado à escola na última semana?

Model: (Intercept), dist, sem, per

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

Omnibus Test^a

Likelihood Ratio Chi-Square	df	Sig.
51,015	3	,000

Dependent Variable: quantas vezes chegou atrasado à escola na última semana?

Model: (Intercept), dist, sem, per

a. Compares the fitted model against the intercept-only model.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test			Exp(B)	95% Wald Confidence Interval for Exp(B)	
			Lower	Upper	Wald Chi-Square	df	Sig.		Lower	Upper
(Intercept)	-4,380	1,1602	-6,654	-2,106	14,251	1	,000	,013	,001	,122
dist	,222	,0659	,093	,351	11,370	1	,001	1,249	1,097	1,421
sem	,165	,0458	,075	,254	12,904	1	,000	1,179	1,078	1,290
per	-,573	,2619	-1,086	-,060	4,789	1	,029	,564	,337	,942
(Scale)	1 ^a									

Dependent Variable: quantas vezes chegou atrasado à escola na última semana?

Model: (Intercept), dist, sem, per

a. Fixed at the displayed value.

Figura 14.57 Outputs do modelo de regressão Poisson no SPSS.

Após a estimação do modelo de regressão Poisson, precisamos elaborar o teste para verificação de existência de superdispersão nos dados. Para tanto, seguiremos o mesmo procedimento estudado nas seções 14.2.4 e 14.4.1. Assim, vamos inicialmente criar uma nova variável, que chamaremos de *yasterisco*. Para tanto, em **Transform → Compute Variable...**, devemos proceder como mostra a Figura 14.58. Note que a expressão a ser digitada na caixa **Numeric Expression** refere-se à expressão (14.14) e, no SPSS, o duplo asterisco corresponde ao operador expoente. A variável *MeanPredicted*, gerada no banco de dados após a estimação do modelo, refere-se à quantidade prevista de atrasos semanais para cada aluno (λ).

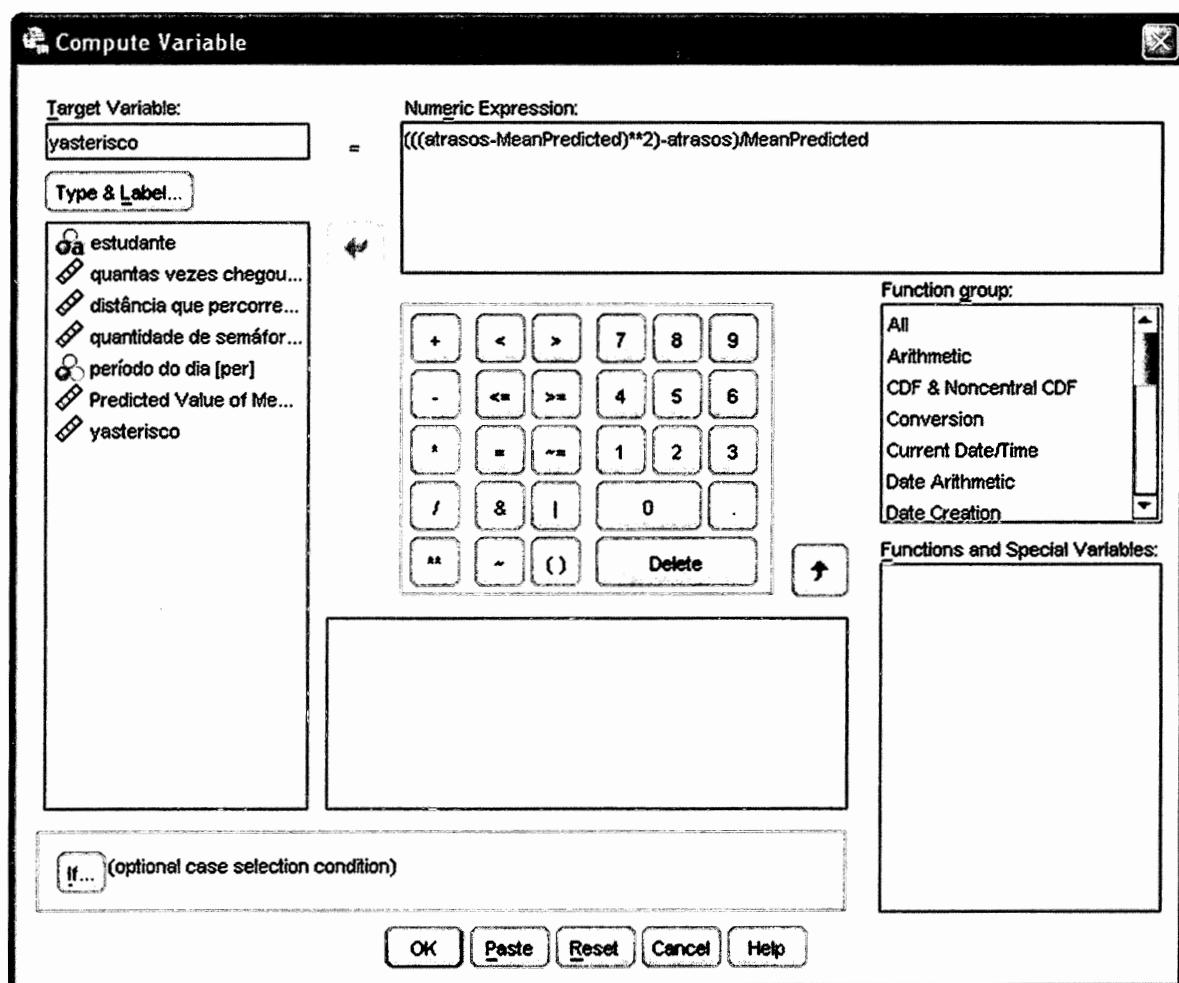


Figura 14.58 Criação da variável *yasterisco* para elaboração do teste para verificação de existência de superdispersão nos dados.

Após clicarmos em **OK**, a nova variável *yasterisco* surgirá na base de dados. Devemos agora regredi-la em função da variável *MeanPredicted*, de acordo com a expressão (14.15). Para tanto, vamos clicar em **Analyze → Regression → Linear...**, e inserir a variável *yasterisco* na caixa **Dependent** e a variável *MeanPredicted* em **Independent(s)**, conforme mostra a Figura 14.59.

No botão **Options...**, devemos desmarcar a opção **Include constant in equation**, conforme mostra a Figura 14.60. Na sequência, podemos clicar em **Continue** e em **OK**.

O *output* que nos interessa encontra-se na Figura 14.61.

Como o *valor-P (Sig.)* do teste *t* correspondente ao parâmetro β da variável *MeanPredicted* (*Predicted Value of Mean of Response*) é maior do que 0,05, podemos afirmar que os dados da variável dependente **não apresentam superdispersão** ao nível de significância de 5%, fazendo com que o modelo de regressão Poisson estimado seja adequado pela **presença de equidispersão nos dados**. O *output* da Figura 14.61 equivale aos *outputs* das Figuras 14.10 (Excel) e 14.23 (Stata).

Na sequência, assim como realizado na seção 14.4.1, vamos comparar os resultados do modelo de regressão Poisson estimado por máxima verossimilhança com aqueles obtidos por um modelo de regressão múltipla log-linear estimado pelo método de mínimos quadrados ordinários (*ordinary least squares*, ou *OLS*). Para tanto, vamos inicialmente gerar a variável *Inatrasos*, que corresponde ao logaritmo natural da variável dependente *atrasos*, clicando em **Transform → Compute Variable...**, conforme mostra a Figura 14.62.

Desta forma, o modelo $\ln atrasos_i = \alpha + \beta_1 \cdot dist_i + \beta_2 \cdot sem_i + \beta_3 \cdot per_i$ pode ser estimado por *OLS*. Para tanto, vamos clicar em **Analyze → Regression → Linear...**, e inserir a variável *Inatrasos* na caixa **Dependent** e as variáveis *dist*, *sem* e *per* na caixa **Independent(s)**, conforme mostra a Figura 14.63.

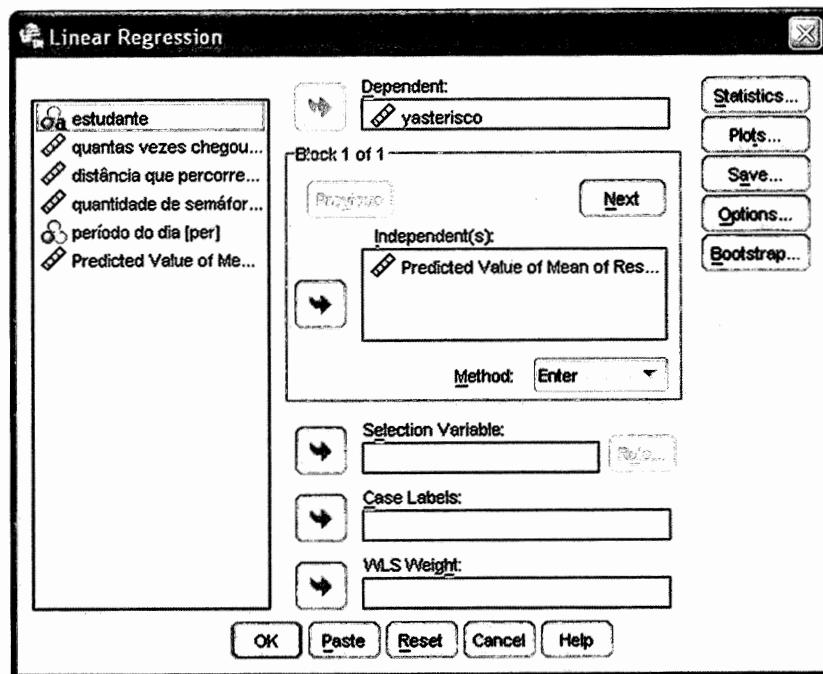


Figura 14.59 Regressão auxiliar para elaboração do teste para verificação de existência de superdispersão nos dados.

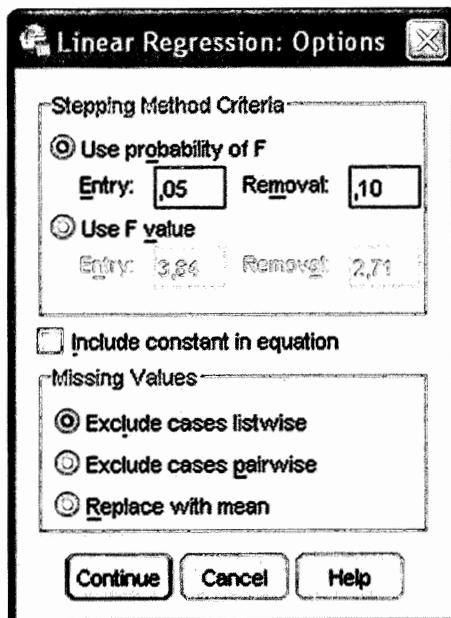


Figura 14.60 Exclusão da constante para a elaboração da regressão auxiliar.

Model	Coefficients ^{a,b}			t	Sig.
	B	Unstandardized Coefficients	Standardized Coefficients		
		Beta			
1 Predicted Value of Mean of Response	-.292	.158	-.182	-1,843	,068

a. Dependent Variable: yasterisco
 b. Linear Regression through the Origin

Figura 14.61 Resultado do teste para verificação de existência de superdispersão no SPSS.

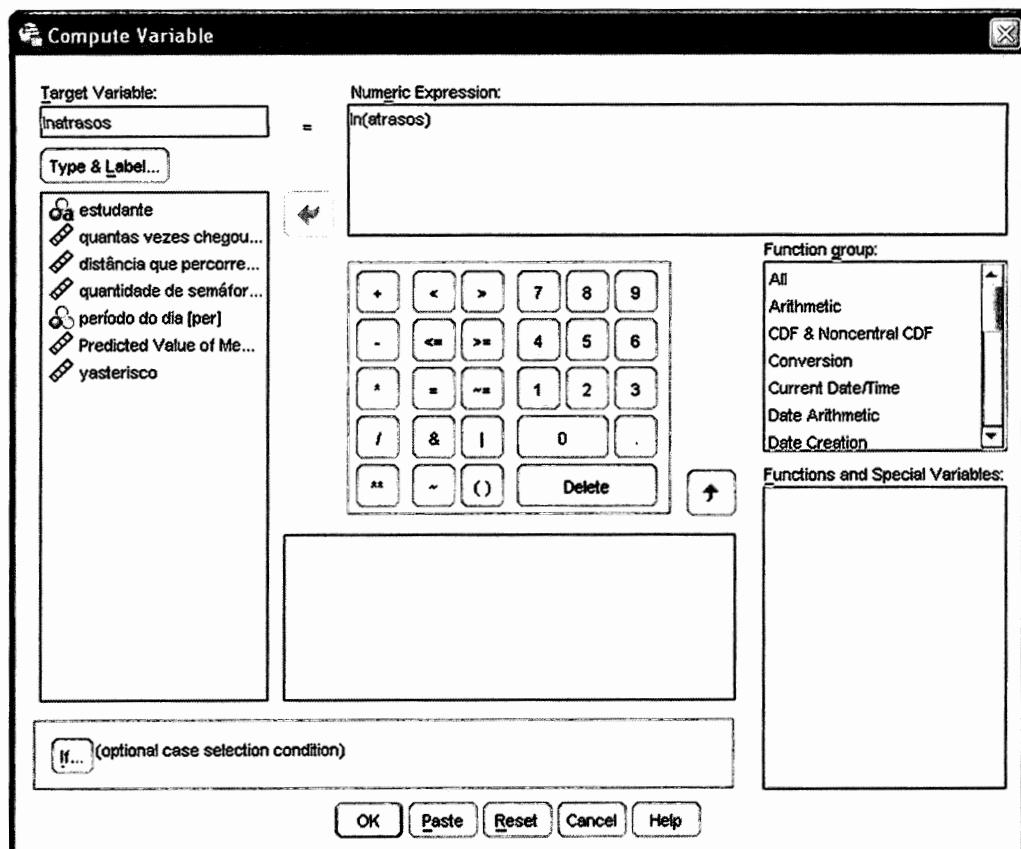


Figura 14.62 Criação da variável *Inatrasos* para estimação de um modelo de regressão log-linear.

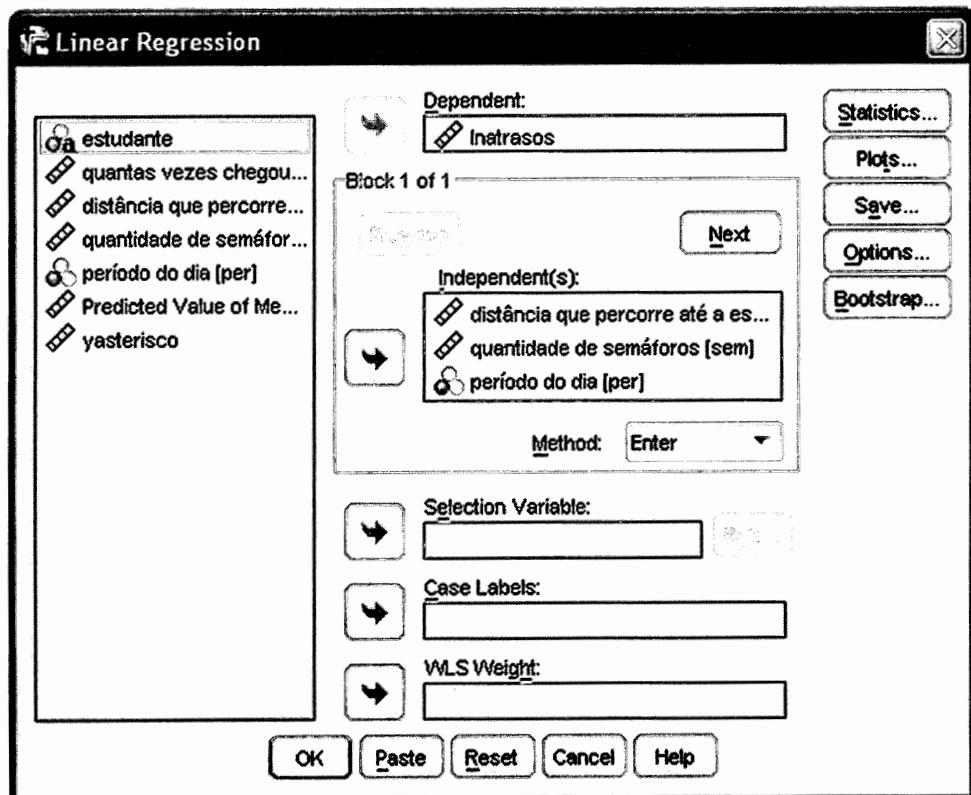


Figura 14.63 Caixa de diálogo para estimação da regressão log-linear.

No botão **Save...**, devemos marcar a opção **Unstandardized**, em **Predicted Values**, conforme mostra a Figura 14.64. Na sequência, podemos clicar em **Continue** e em **OK**. Este procedimento criará no banco de dados uma nova variável, chamada pelo SPSS de *PRE_1*, que corresponde à variável *yhat* gerada quando da estimação pelo Stata (valores previstos do logaritmo natural do número de atrasos semanais por aluno).

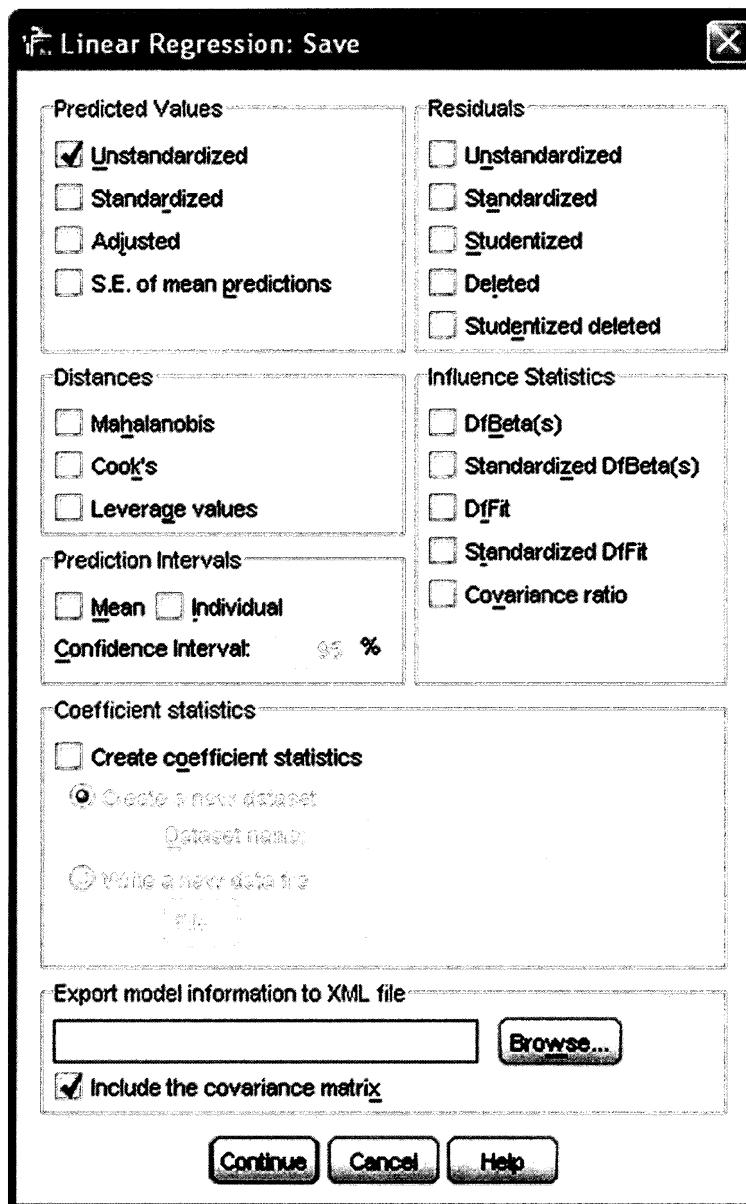


Figura 14.64 Procedimento para criação da variável *PRE_1*.

Não apresentaremos os resultados desta regressão múltipla estimada pelo SPSS, uma vez que nos interessa, neste momento, apenas gerar outra variável, a partir da variável *PRE_1*, que representará os valores previstos do número de atrasos semanais propriamente ditos por aluno. Esta variável, que chamaremos de *eyhat*, poderá ser criada clicando-se novamente em **Transform → Compute Variable...**, conforme mostra a Figura 14.65.

A fim de elaborarmos um gráfico similar ao apresentado na Figura 14.30, ou seja, um gráfico que permite que sejam comparados, para cada uma das estimações, os valores previstos e os valores reais do número de atrasos por semana, vamos agora clicar em **Graphs → Legacy Dialogs → Line...** e, na sequência, nas opções **Multiple** e **Summaries of separate variables**, como apresentado na Figura 14.66.

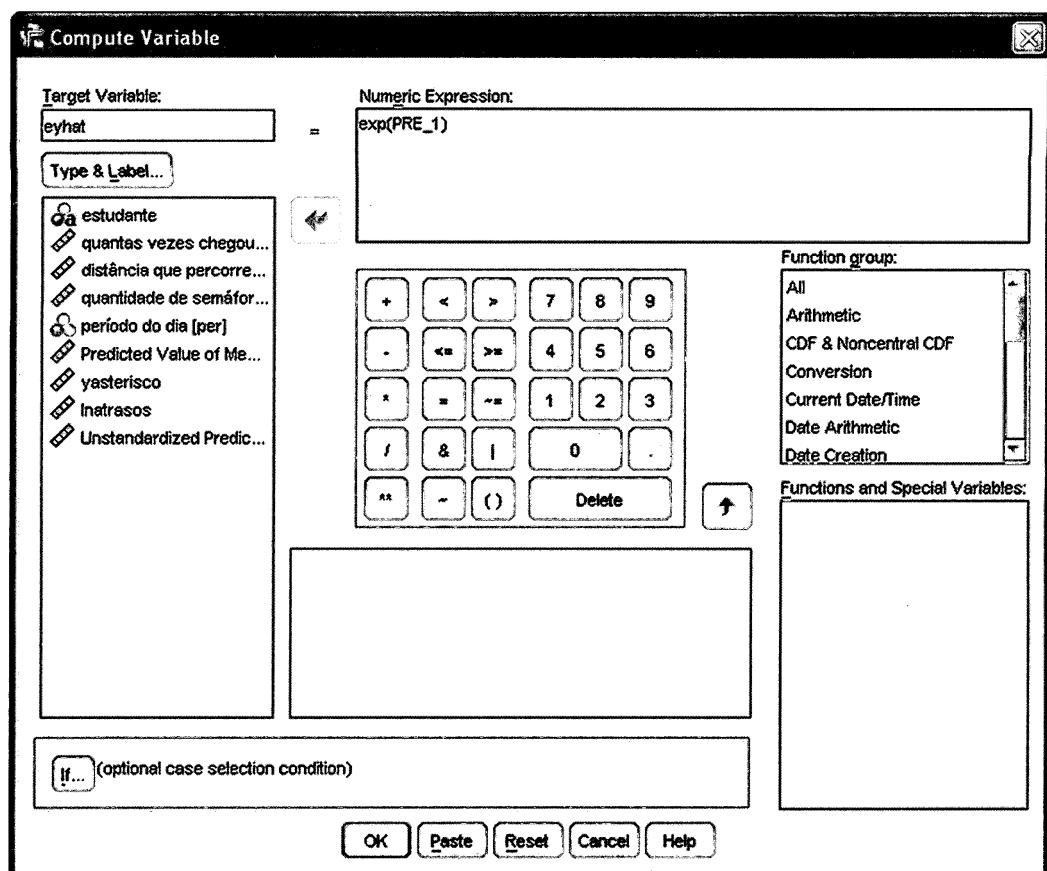


Figura 14.65 Criação da variável *eyhat* a partir da variável *PRE_1*.

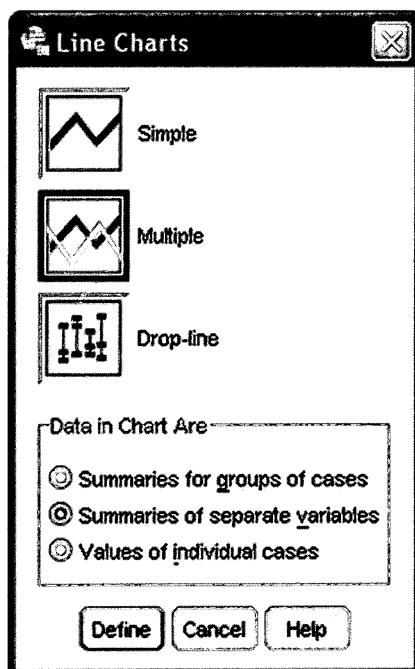


Figura 14.66 Caixa de diálogo para elaboração de gráfico para comparação das estimações.

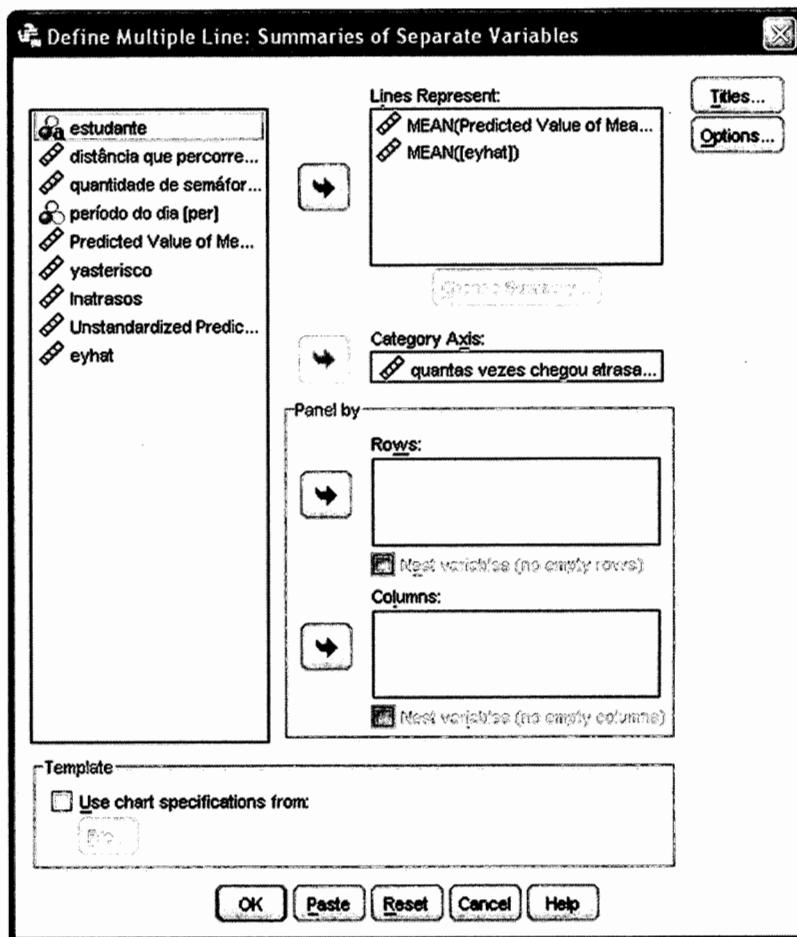


Figura 14.67 Seleção das variáveis a serem inseridas no gráfico.

Ao clicarmos em **Define**, surgirá uma caixa de diálogo como a apresentada na Figura 14.67. Devemos inserir as variáveis *MeanPredicted* (quantidade prevista de atrasos semanais para cada aluno estimada por máxima verosimilhança para o modelo de regressão Poisson) e *eyhat* (quantidade prevista de atrasos semanais para cada aluno estimada por *OLS* para o modelo de regressão múltipla log-linear) na caixa **Lines Represent** e a variável *atrasos* em **Category Axis**. Na sequência, podemos clicar em **OK**.

O gráfico da Figura 14.68 oferece uma oportunidade de comparação dos comportamentos dos valores previstos com os valores reais da variável dependente para cada uma das estimativas elaboradas, de onde se pode verificar que são diferentes. Conforme discutido, o fato de determinada variável dependente ser quantitativa não é condição suficiente para que seja elaborado um modelo de regressão múltipla com estimativa *OLS*. Dados de contagem apresentam distribuições particulares e o pesquisador sempre precisa estar atento a este fato, a fim de que sejam estimados modelos adequados e consistentes para efeitos de diagnóstico e de previsão.

14.5.2. Modelo de regressão binomial negativo no software SPSS

Seguindo a mesma lógica proposta na seção anterior, vamos agora abrir o arquivo **QuantAtrasosBNeg.sav**, que traz dados sobre a quantidade mensal de atrasos dos 100 alunos, a distância percorrida no trajeto (em quilômetros), o número de semáforos pelos quais cada um passa e o período do dia em que cada estudante tem o hábito de se deslocar para a escola (manhã ou tarde).

Clicando em **Analyze → Descriptive Statistics → Frequencies...**, podemos inicialmente elaborar o diagnóstico sobre a distribuição da variável dependente. Nesta caixa de diálogo, não apresentada novamente

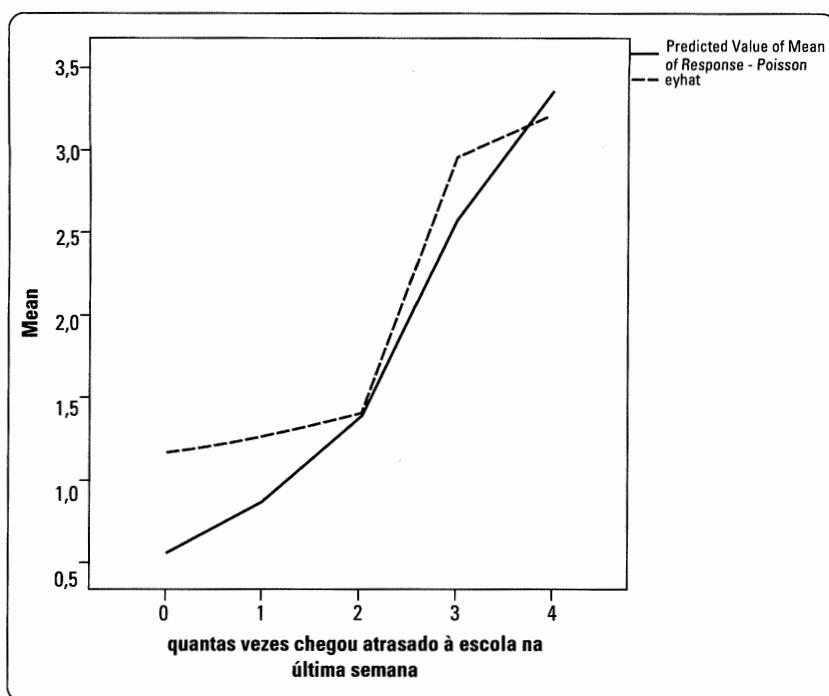


Figura 14.68 Valores previstos x valores observados para os modelos de regressão Poisson e de regressão múltipla log-linear (OLS).

aqui, devemos inserir a variável dependente *atrasos* (quantas vezes chegou atrasado à escola no último mês?) em **Variable(s)** e, no botão **Statistics...**, devemos marcar as opções **Mean** e **Variance**. Já no botão **Charts...**, marcaremos a opção **Histograms** para, então, clicarmos em **Continue** e em **OK**. Os *outputs* encontram-se na Figura 14.69.

Estes *outputs* são os mesmos daqueles apresentados na Tabela 14.11 e na Figura 14.12 da seção 14.3.1 e também nas Figuras 14.32, 14.33 e 14.34 da seção 14.4.2 e, por meio deles, podemos verificar, ainda que de forma preliminar, que há indícios de existência de superdispersão nos dados, uma vez que a variância é superior à média da variável dependente.

Recomenda-se, portanto, que seja inicialmente estimado um modelo de regressão Poisson, para, a partir de seus resultados, ser elaborado o teste para verificação de existência de superdispersão nos dados. Não iremos mostrar novamente as janelas para estimação deste modelo no SPSS, assim como foi feito na seção anterior, porém serão descritos os passos para a sua elaboração.

Assim sendo, vamos inicialmente clicar em **Analyze → Generalized Linear Models → Generalized Linear Models...**. Na caixa de diálogo que será aberta, devemos selecionar, na pasta **Type of Model**, a opção **Poisson loglinear** (em **Counts**). Já na pasta **Response**, devemos incluir a variável *atrasos* na caixa **Dependent Variable**. Enquanto na pasta **Predictors**, devemos incluir as variáveis *dist*, *sem* e *per* na caixa **Covariates**, na pasta **Model** devemos inserir estas mesmas três variáveis na caixa **Model**. Na pasta **Statistics**, além das opções já selecionadas de forma padrão pelo SPSS, devemos selecionar também a opção **Include exponential parameter estimates** e, por fim, na pasta **Save**, selecionaremos apenas a opção **Predicted value of mean response**. Ao clicarmos em **OK**, serão gerados os *outputs* da estimação do modelo de regressão Poisson, que não serão, em sua totalidade, apresentados aqui.

A Figura 14.70 apresenta apenas o *output* que nos interessa neste momento (**Goodness of Fit**) e, por meio dele, podemos verificar que a qualidade do ajuste do modelo estimado não é adequada, visto que, para um $\chi^2_{cal} = 145,295$ (*Deviance*), temos, para 96 graus de liberdade, que $Sig. \chi^2 < 0,05$, ou seja, existem diferenças estatisticamente significantes entre os valores previstos pelo modelo Poisson e os valores observados do número de atrasos que ocorrem por mês. Esta parte muito importante do *output* corresponde ao apresentado na Figura 14.36 quando da estimação do modelo pelo Stata.

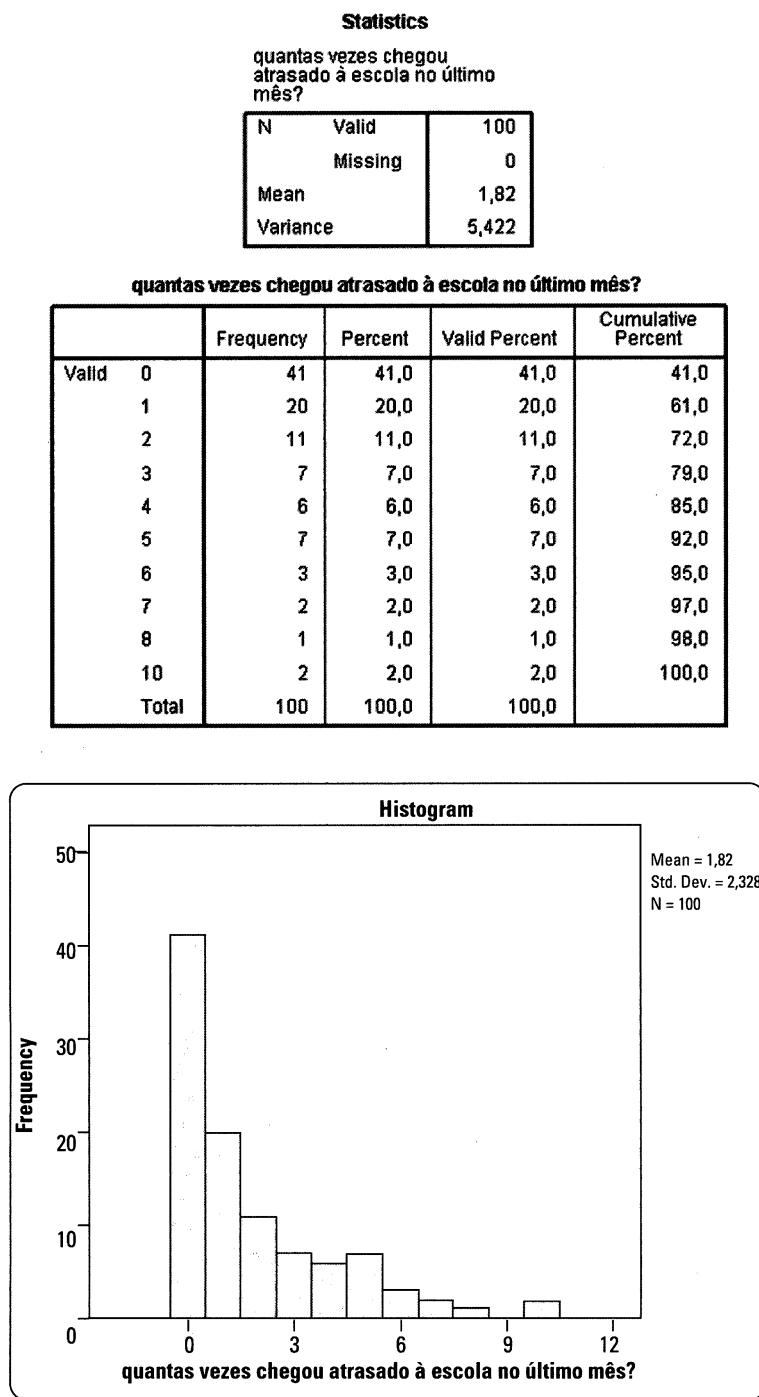


Figura 14.69 Média, variância, tabela de frequências e histograma da variável dependente.

A qualidade do ajuste do modelo de regressão Poisson estimado pode não ter sido adequada pela presença de superdispersão nos dados da variável dependente e, portanto, vamos agora elaborar o teste para verificação da existência deste fenômeno. Seguindo o que foi exposto na seção anterior, precisamos criar uma nova variável, que também chamaremos aqui de *yasterisco* e, para tanto, vamos clicar em **Transform → Compute Variable**.... A expressão que deve ser digitada na caixa **Numeric Expression** refere-se à expressão (14.14) e, no SPSS, será a mesma daquela apresentada na Figura 14.58, ou seja, $((\text{atrasos}-\text{MeanPredicted})^{**2})-\text{atrasos})/\text{MeanPredicted}$, em que a variável *MeanPredicted*, gerada no banco de dados após a estimação do modelo de regressão Poisson, refere-se à quantidade prevista de atrasos mensais para cada aluno. Também não apresentaremos aqui as figuras dispostas na seção anterior.

Goodness of Fit ^b			
	Value	df	Value/df
Deviance	145,295	96	1,513
Scaled Deviance	145,295	96	
Pearson Chi-Square	142,235	96	1,482
Scaled Pearson Chi-Square	142,235	96	
Log Likelihood ^a	-154,894		
Akaike's Information Criterion (AIC)	317,788		
Finite Sample Corrected AIC (AICC)	318,209		
Bayesian Information Criterion (BIC)	328,208		
Consistent AIC (CAIC)	332,208		

Dependent Variable: quantas vezes chegou atrasado à escola no último mês?

Model: (Intercept), dist, sem, per

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

Figura 14.70 Qualidade do ajuste do modelo de regressão Poisson inicialmente estimado.

Após clicarmos em **OK**, a nova variável *yasterisco* surgirá na base de dados. Vamos, portanto, regredi-la em função da variável *MeanPredicted*, seguindo a expressão (14.15). Para tanto, devemos clicar em **Analyze → Regression → Linear...**, e inserir a variável *yasterisco* na caixa **Dependent** e a variável *MeanPredicted* em **Independent(s)**. Por fim, no botão **Options...**, devemos desmarcar a opção **Include constant in equation** e, na sequência, devemos clicar em **Continue** e em **OK**. O *output* que nos interessa encontra-se na Figura 14.71.

Model	Coefficients ^{a,b}					
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	
	B	Std. Error	Beta			
1 Predicted Value of Mean of Response	,133	,062	,210	2,139	,035	

a. Dependent Variable: *yasterisco*

b. Linear Regression through the Origin

Figura 14.71 Resultado do teste para verificação de existência de superdispersão no SPSS.

Como o *valor-P* (*Sig.*) do teste *t* correspondente ao parâmetro β da variável *MeanPredicted* (*Predicted Value of Mean of Response*) é menor do que 0,05, podemos afirmar que os dados da variável dependente **apresentam superdispersão** ao nível de significância de 5%, fazendo com que o modelo de regressão Poisson estimado não seja adequado. O *output* da Figura 14.71 equivale ao *output* da Figura 14.35 (estimação pelo Stata).

Vamos então à estimação do modelo de regressão binomial negativo. Para tanto, devemos clicar em **Analyze → Generalized Linear Models → Generalized Linear Models...** e, na caixa de diálogo que será aberta, devemos marcar, na pasta **Type of Model**, a opção **Custom**. Nesta mesma pasta, devemos ainda selecionar as opções **Negative binomial** (em **Distribution**), **Log** (em **Link function**) e **Estimate value** (em **Parameter**). Esta última opção refere-se à estimação do parâmetro ϕ e, portanto, será estimado um modelo de regressão NB2. A Figura 14.72 mostra como ficará esta pasta após a seleção das opções.

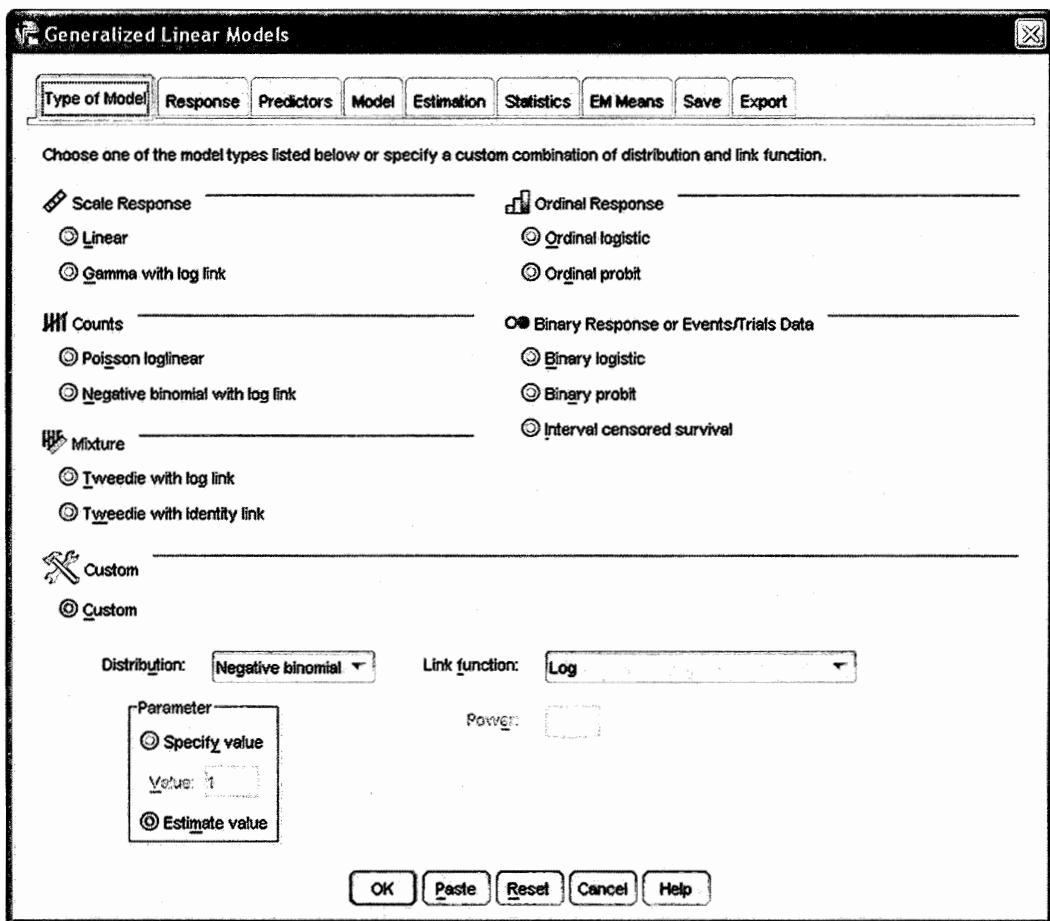


Figura 14.72 Caixa de diálogo inicial para estimação do modelo NB2 no SPSS.

Para as demais pastas, o pesquisador pode optar por manter as mesmas opções que já foram selecionadas quando da estimação inicial do modelo de regressão Poisson. Os *outputs* gerados por meio da estimação do presente modelo de regressão binomial negativo encontram-se na Figura 14.73.

O primeiro *output* desta figura (**Goodness of Fit**) apresenta o valor da somatória do logaritmo da função de máxima verossimilhança da estimação do modelo NB2 (*Log Likelihood*), que é de -151,012 e é exatamente igual ao valor obtido quando da modelagem no Excel (Tabela 14.12 e Figura 14.14) e no Stata (Figuras 14.37, 14.39 e 14.41). Por meio do mesmo *output*, podemos também verificar que a qualidade do ajuste do modelo estimado é agora adequada, visto que, para um $\chi^2_{cal} = 105,025$ (*Deviance*), temos, para 96 graus de liberdade, que $Sig. \chi^2 > 0,05$ (já que $\chi^2_c = 119,871$ para 96 graus de liberdade e nível de significância de 5%), ou seja, não existem diferenças estatisticamente significantes entre os valores previstos e os observados da quantidade de atrasos que ocorrem por mês ao se chegar à escola. Esta parte do *output* corresponde ao *Deviance* que é apresentado pelo Stata quando da estimação do modelo de regressão binomial negativo obtida pelo comando `glm..., family(nbinomial ml)` (Figura 14.39).

Podemos também verificar, com base no teste χ^2 (*Likelihood Ratio Chi-Square* = 63,249, *Sig. χ^2* = 0,000 < 0,05 apresentado no *output* **Omnibus Test**), que a hipótese nula de que todos os parâmetros β_j ($j = 1, 2, 3$) sejam estatisticamente iguais a zero pode ser rejeitada ao nível de significância de 5%, ou seja, pelo menos uma variável X é estatisticamente significante para explicar a ocorrência de atrasos por mês.

Os parâmetros estimados encontram-se no *output* **Parameter Estimates** e são exatamente iguais aos calculados manualmente e apresentados na Figura 14.14 (Excel) e também obtidos por meio dos comandos `nbreg` ou `glm..., family(nbinomial ml)` do Stata (Figuras 14.37 e 14.39, respectivamente). Este mesmo *output* também apresenta as **incidence rate ratios** (ou *IRR*) de cada variável explicativa, que o SPSS chama de *Exp(B)*, conforme também já apresentado por meio da Figura 14.41. Como todos os intervalos de confiança dos parâmetros estimados

Goodness of Fit^b

	Value	df	Value/df
Deviance	105,025	95	1,106
Scaled Deviance	105,025	95	
Pearson Chi-Square	104,703	95	1,102
Scaled Pearson Chi-Square	104,703	95	
Log Likelihood ^a	-151,012		
Akaike's Information Criterion (AIC)	312,025		
Finite Sample Corrected AIC (AICC)	312,663		
Bayesian Information Criterion (BIC)	325,050		
Consistent AIC (CAIC)	330,050		

Dependent Variable: quantas vezes chegou atrasado à escola no último mês?

Model: (Intercept), dist, sem, per

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

Omnibus Test^a

Likelihood Ratio Chi-Square	df	Sig.
63,249	3	,000

Dependent Variable: quantas vezes chegou atrasado à escola no último mês?

Model: (Intercept), dist, sem, per

a. Compares the fitted model against the Intercept-only model.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test			Exp(B)	95% Wald Confidence Interval for Exp(B)	
			Lower	Upper	Wald Chi-Square	df	Sig.		Lower	Upper
(Intercept)	-4,997	1,2494	-7,446	-2,549	15,998	1	,000	,007	,001	,078
dist	,308	,0713	,168	,447	18,644	1	,000	1,380	1,183	1,564
sem	,197	,0495	,100	,294	15,874	1	,000	1,218	1,105	1,342
per	-,927	,2570	-1,431	-,424	13,020	1	,000	,396	,239	,665
(Scale)	1 [#]									
(Negative binomial)	,255	,1248	,098	,666						

Dependent Variable: quantas vezes chegou atrasado à escola no último mês?
Model: (Intercept), dist, sem, per

a. Fixed at the displayed value.

Figura 14.73 Outputs do modelo de regressão binomial negativo (NB2) no SPSS.

(95% Wald Confidence Interval) não contêm o zero e, consequentemente, os de $Exp(B)$ não contêm o 1, já chegamos ao modelo final de regressão binomial negativo (todos os $Sig. Wald Chi-Square < 0,05$).

Logo, a expressão da quantidade média estimada de atrasos por mês para um determinado aluno i pode ser escrita como:

$$\mu_i = e^{(-4,997 + 0,308 \cdot dist_i + 0,197 \cdot sem_i - 0,927 \cdot per_i)}$$

Além disso, também com base no output final da Figura 14.73, as quantidades estimadas de atrasos por mês apresentam, com 95% de nível de confiança, expressões de mínimo e de máximo iguais a:

$$u_{i_{\min}} = e^{(-7,446 + 0,168 \cdot dist_i + 0,100 \cdot sem_i - 1,431 \cdot per_i)}$$

$$u_{i_{\max}} = e^{(-2,549 + 0,447 \cdot dist_i + 0,294 \cdot sem_i - 0,424 \cdot per_i)}$$

Por fim, a parte inferior do *output* final da Figura 14.73 apresenta a estimativa de ϕ (*Negative binomial*). Conforme podemos observar, o intervalo de confiança para ϕ não contém o zero, ou seja, para o nível de confiança de 95%, podemos afirmar que ϕ é estatisticamente diferente de zero e com valor estimado igual a 0,255, conforme já calculado na seção 14.3.1 por meio do **Solver** do Excel (Figura 14.14) e na seção 14.4.2 por meio do Stata (Figuras 14.37, 14.39 e 14.41). **Isso comprova a existência de superdispersão nos dados**, com a variância da variável dependente apresentando a seguinte expressão:

$$Var(Y) = u + 0,255 \cdot u^2$$

Por fim, vamos agora elaborar um gráfico similar ao apresentado na Figura 14.45, porém com a inclusão também dos valores estimados por *OLS* de um modelo de regressão múltipla log-linear. Em outras palavras, elaboraremos um gráfico que permite que sejam comparados, para cada um dos modelos estimados (binomial negativo, Poisson e regressão log-linear por *OLS*), os valores previstos e os valores reais do número de atrasos por mês.

Como os valores previstos das estimações dos modelos Poisson e binomial negativo já se encontram no banco de dados (variáveis *MeanPredicted* e *MeanPredicted_1*, respectivamente), precisamos, neste momento, estimar o modelo de regressão múltipla log-linear por *OLS*, cujos resultados não serão aqui apresentados, porém os procedimentos serão descritos.

Desta forma, vamos gerar uma variável chamada de *lnatrasos*, que corresponde ao logaritmo natural da variável dependente *atrasos*, clicando em **Transform → Compute Variable...**. A expressão que deve ser digitada na caixa **Numeric Expression** é *ln(atrasos)* para que, desta forma, o modelo $\ln atrasos_i = \alpha + \beta_1 \cdot dist_i + \beta_2 \cdot sem_i + \beta_3 \cdot per_i$ possa ser estimado por *OLS*.

Na sequência, vamos clicar em **Analyze → Regression → Linear...**, e inserir a variável *lnatrasos* na caixa **Dependent** e as variáveis *dist*, *sem* e *per* na caixa **Independent(s)**. No botão **Save...**, devemos marcar a opção **Unstandardized**, em **Predicted Values** e, por fim, podemos clicar em **Continue** e em **OK**. Este procedimento criará no banco de dados uma nova variável, chamada pelo SPSS de *PRE_1* (valores previstos do logaritmo natural do número de atrasos por mês).

Entretanto, a variável que desejamos criar refere-se aos valores previstos do número de atrasos mensais, e não aos valores previstos do logaritmo natural do número de atrasos mensais. Portanto, precisamos clicar novamente em **Transform → Compute Variable...** e criar uma variável chamada de *eyhat*, cuja expressão a ser digitada na caixa **Numeric Expression** é *exp(PRE_1)*.

Desta forma, podemos elaborar o gráfico desejado, clicando em **Graphs → Legacy Dialogs → Line...** e, na sequência, nas opções **Multiple** e **Summaries of separate variables**. Ao clicarmos em **Define**, surgirá uma caixa de diálogo em que deveremos inserir as variáveis *MeanPredicted* (valores previstos pelo modelo Poisson), *MeanPredicted_1* (valores previstos pelo modelo binomial negativo) e *eyhat* (valores previstos pelo modelo de regressão log-linear estimado por *OLS*) na caixa **Lines Represent** e a variável *atrasos* em **Category Axis**. Na sequência, podemos clicar em **OK**.

O gráfico gerado pode ser editado por meio de um duplo clique, e aqui se optou pela apresentação de uma interpolação do tipo **Spline**, conforme mostra a Figura 14.74. O gráfico final encontra-se na Figura 14.75.

Por meio da análise do gráfico da Figura 14.75 podemos verificar que a variância da quantidade prevista de atrasos mensais é bem superior para o caso do modelo de regressão binomial negativo, cuja estimativa consegue de fato capturar a existência de superdispersão nos dados, principalmente para valores maiores de atrasos por mês.

Isso confirma o fato de que distribuições de dados de contagem com amplitudes maiores de seus valores observados podem aumentar a variância da variável em estudo numa proporção maior do que a sua média, o que pode acarretar em uma superdispersão nos dados. Enquanto não se verificou a existência de superdispersão para os dados de contagem semanal, com menos possibilidades de ocorrência, este fenômeno tornou-se presente quando os dados de contagem passaram a se apresentar de forma mensal, ou seja, com mais amplas possibilidades de ocorrência. Conforme estudamos neste capítulo, enquanto o primeiro caso foi abordado por meio da estimação de

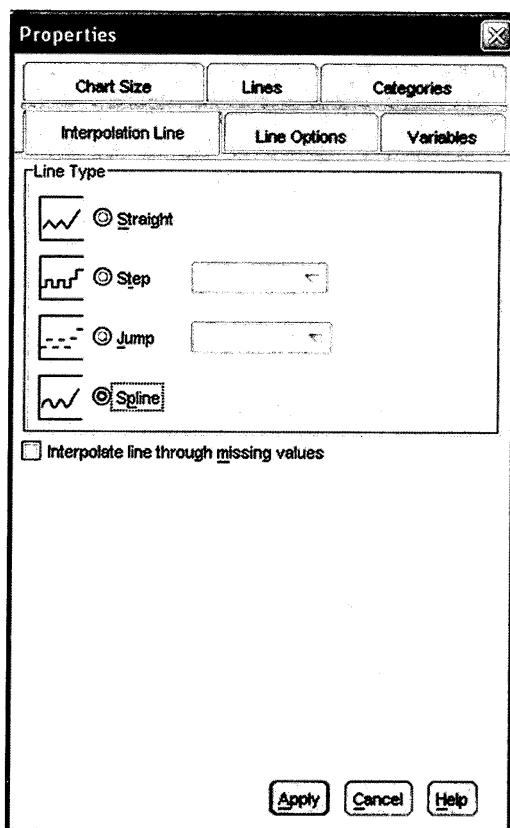


Figura 14.74 Definição da interpolação do tipo *Spline* para elaboração de gráficos.

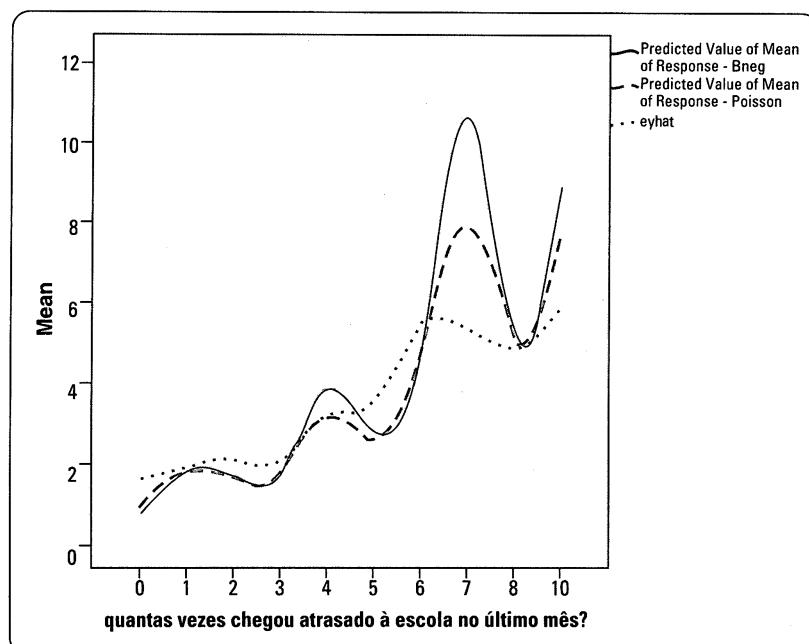


Figura 14.75 Valores previstos x valores observados de atrasos mensais para os modelos de regressão binomial negativo, Poisson e log-linear (*OLS*).

um modelo de regressão Poisson, os dados do segundo caso acabaram por apresentar um melhor ajuste quando se estimou um modelo de regressão binomial negativo.

14.6. CONSIDERAÇÕES FINAIS

A estimação de modelos de regressão em que a variável dependente é composta por dados de contagem apresenta inúmeras aplicações, porém ainda é pouco explorada, seja pelo desconhecimento dos modelos existentes, seja pelo senso comum, ainda que incorreto, de que se a variável dependente for quantitativa, cabe a estimação OLS, independentemente da sua distribuição.

Os modelos de regressão Poisson e binomial negativo são modelos log-lineares (ou semilogarítmicos à esquerda) e representam os modelos para dados de contagem mais conhecidos, sendo estimados por máxima verossimilhança. Enquanto a estimação correta de um modelo de regressão Poisson exige que não ocorra o fenômeno da superdispersão nos dados da variável dependente, a estimação de um modelo de regressão binomial negativo permite que a variância da variável dependente seja estatisticamente superior à sua média.¹

Recomenda-se que, antes que seja definido o mais adequado e consistente modelo de regressão quando houver dados de contagem, seja elaborado um diagnóstico sobre a distribuição da variável dependente e estimado um modelo de regressão Poisson para, a partir de então, ser elaborado um teste para verificação de existência de superdispersão nos dados. Caso isso se comprove, deve ser estimado um modelo de regressão binomial negativo, sendo recomendável o modelo do tipo NB2.

Os modelos de regressão Poisson e binomial negativo devem ser estimados por meio do uso correto do software escolhido, e a inclusão inicial de potenciais variáveis explicativas do fenômeno em estudo deve ser sempre feita com base na teoria subjacente e na intuição do pesquisador.

14.7. EXERCÍCIOS

1. Uma financeira de um grande estabelecimento varejista de eletroeletrônicos deseja saber se a renda e a idade dos consumidores explicam a incidência do uso de financiamento, via crédito direto ao consumidor (CDC), quando da compra de bens como telefones celulares, tablets, laptops, televisões, videogames, aparelhos de DVD, entre outros, a fim de que seja possível elaborar uma campanha de promoção dessa forma de financiamento segmentada pelo perfil dos clientes. Para tanto, a área de marketing da financeira selecionou, aleatoriamente, uma amostra de 200 consumidores provenientes de sua base total de clientes, com as seguintes variáveis:

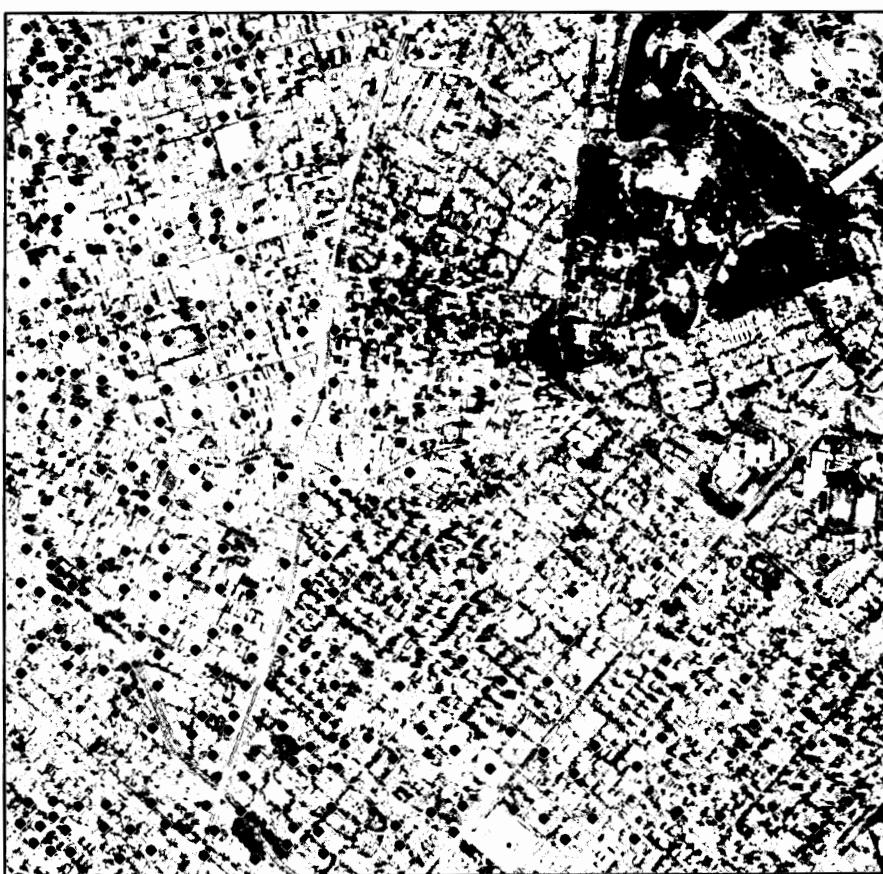
Variável	Descrição
<i>id</i>	Variável <i>string</i> que varia de 001 a 200 e que identifica o consumidor.
<i>quantcompras</i>	Variável dependente correspondente à quantidade de compras de bens duráveis realizadas por meio de CDC no último ano por consumidor (dados de contagem).
<i>renda</i>	Renda mensal do consumidor (R\$).
<i>idade</i>	Idade do consumidor (anos).

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

- Elabore um diagnóstico preliminar sobre a existência de superdispersão nos dados da variável *quantcompras*. Apresente a sua média e a sua variância, e elabore o seu histograma.
- Estime um modelo de regressão Poisson e, com base em seus resultados, elabore o teste para verificação de existência de superdispersão nos dados. Qual a conclusão deste teste, ao nível de significância de 5%?
- Elabore um teste χ^2 para comparar as distribuições de probabilidades observadas e previstas de incidência anual de uso do CDC. O resultado do teste, ao nível de significância de 5%, indica a existência de qualidade do ajuste do modelo de regressão Poisson?

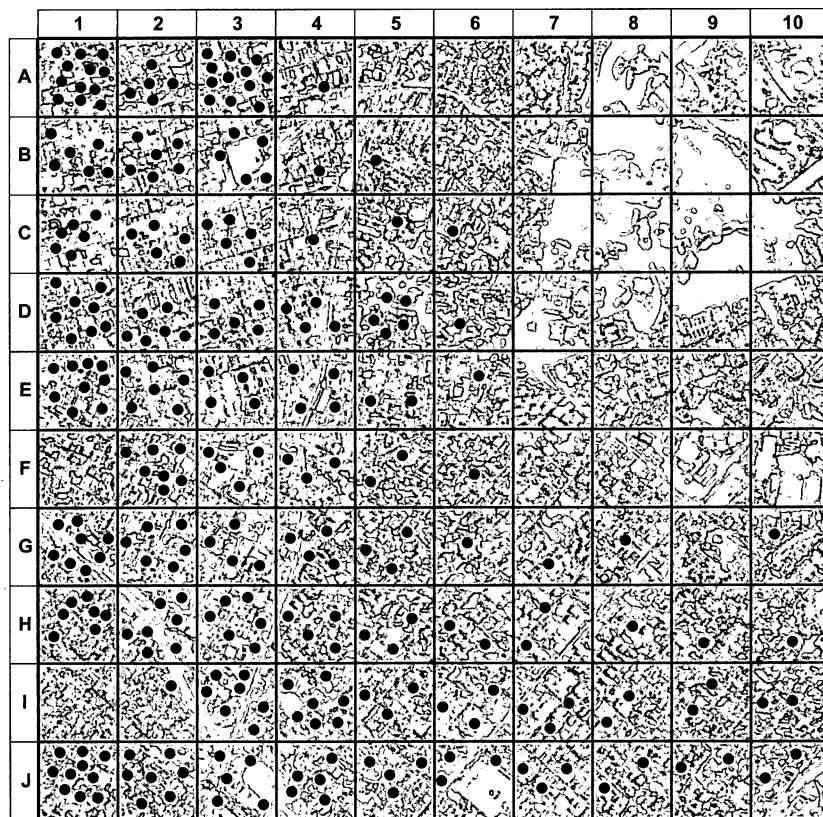
¹ Embora não seja escopo deste livro, muitos autores compararam estimações por máxima verossimilhança de modelos de regressão Poisson e binomial negativo com estimativas por máxima verossimilhança de modelos que consideram a variável dependente censurada, com base no desenvolvimento de modelos conhecidos por Tobit. Para maiores informações, recomendamos o estudo de Cameron e Trivedi (2009).

- d. Se a resposta do item anterior for sim, apresente a expressão final para a quantidade média estimada de uso anual de financiamento por meio de CDC quando da compra de bens duráveis, em função das variáveis explicativas que se mostraram estatisticamente significantes, ao nível de confiança de 95%.
- e. Qual a quantidade média esperada de uso do CDC por ano para um consumidor com renda mensal de R\$2.600,00 e 47 anos de idade?
- f. Em média, em quanto se altera a taxa de incidência anual de uso do financiamento por CDC ao se aumentar em R\$100,00 a renda mensal do consumidor, mantidas as demais condições constantes?
- g. Em média, em quanto se altera a taxa de incidência anual de uso do financiamento por CDC quando se aumenta a idade média do consumidor em 1 ano, mantidas as demais condições constantes?
- h. Elabore um gráfico (**mspline** no Stata ou **Spline** no SPSS) que mostra o valor previsto de incidência anual de uso do CDC em função da renda mensal do consumidor. Faça uma breve discussão.
- i. Estime um modelo de regressão múltipla log-linear por OLS e compare os resultados previstos deste modelo com aqueles estimados pelo modelo Poisson.
- j. Caso haja o interesse em aumentar o financiamento por meio de CDC, qual público-alvo precisa ser abordado nesta campanha de marketing da financeira?
2. Com o intuito de estudar se a proximidade de parques e áreas verdes e de shoppings e centros de consumo faz com que seja reduzida a intenção de se vender um apartamento, uma empresa do setor imobiliário residencial resolveu marcar a localização de cada um dos 276 imóveis à venda num determinado município, conforme mostra a figura a seguir:



Fonte do Mapa: Google Maps.

A fim de facilitar a elaboração do estudo, a imobiliária criou uma malha quadrangular sobre o mapa do município, com a intenção de identificar as características de cada microrregião. Foram criadas, por meio deste usual procedimento, 100 quadrículas (10 x 10) com dimensões iguais e identificadas de acordo com a figura a seguir:



Fonte do Mapa: Google Maps.

Para uma melhor visualização da quantidade de imóveis à venda em cada microrregião, na próxima figura optou-se por ocultar o mapa do município.

Foram, portanto, levantadas as seguintes variáveis em cada uma das microrregiões do município, aqui definidas pelas quadrículas:

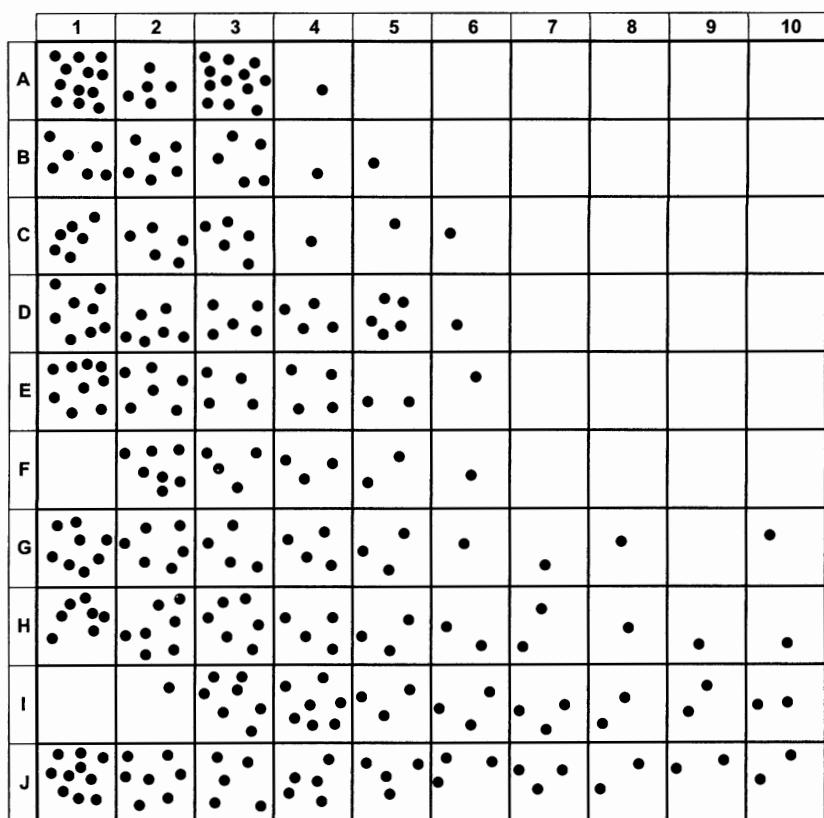
Variável	Descrição
<i>quadrícula</i>	Variável <i>string</i> que identifica a microrregião (quadrícula). É nomeada com um número <i>i</i> seguido de uma letra <i>j</i> , em que o número <i>i</i> varia de 1 a 10 e a letra <i>j</i> , de A a J.
<i>quantimóveis</i>	Variável dependente correspondente à quantidade de imóveis residenciais à venda por quadrícula (dados de contagem).
<i>distparque</i>	Distância da quadrícula ao principal parque do município (em metros).
<i>shopping</i>	Variável binária que indica se há shoppings ou centros de consumo na quadrícula (Não = 0; Sim = 1).

Os dados encontram-se nos arquivos **Imobiliária.sav** e **Imobiliária.dta**. Pede-se:

- Elabore um diagnóstico preliminar sobre a existência de superdispersão nos dados da variável *quantimóveis*. Apresente sua média, sua variância e seu histograma.
- Estime o modelo de regressão Poisson a seguir e, com base em seus resultados, elabore o teste para verificação de existência de superdispersão nos dados. Qual a conclusão deste teste, ao nível de significância de 5%? Elabore também um teste χ^2 para comparar as distribuições de probabilidades observadas e previstas para a quantidade de imóveis à venda por quadrícula. O resultado do teste, ao nível de significância de 5%, indica a existência de qualidade do ajuste do modelo de regressão Poisson? Justifique.

$$\text{quantimóveis}_{ij} = e^{(\alpha + \beta_1 \cdot \text{parque}_{ij} + \beta_2 \cdot \text{shopping}_{ij})}$$

- Estime um modelo de regressão binomial negativo do tipo NB2.



- d. Pode-se dizer, ao nível de confiança de 95%, que o parâmetro ϕ (inverso do parâmetro de forma da distribuição Gama) é estatisticamente diferente de zero? Se sim, deve-se optar pela estimação do modelo binomial negativo?

Os próximos sete itens referem-se à estimação do modelo de regressão binomial negativo do tipo NB2:

- e. Qual a expressão da quantidade média estimada de imóveis à venda para determinada quadrícula ij ?
- f. Qual é a quantidade média esperada de imóveis à venda para uma microrregião (quadrícula) que se encontra a 820 metros de distância do parque e não possui centros de consumo?
- g. Em média, em quanto se altera a taxa de incidência de imóveis à venda por quadrícula quando há uma aproximação média de 100 metros do parque, mantidas as demais condições constantes?
- h. Em média, em quanto se altera a taxa de incidência de imóveis à venda quando passa a existir um centro de consumo ou um shopping na microrregião (quadrícula), mantidas as demais condições constantes?
- i. Elabore um gráfico (`mspline` no Stata ou `Spline` no SPSS) que mostra o comportamento da quantidade prevista de imóveis à venda por quadrícula em função da distância até o parque.
- j. Elabore o mesmo gráfico, porém agora estratificando as quadrículas que têm centros de consumo das que não têm.
- k. Pode-se dizer que a proximidade de parques e áreas verdes e de shoppings e centros de consumo inibe a intenção de se colocar à venda um imóvel residencial?

Além disso, pede-se:

- l. Compare as estimações dos modelos de regressão Poisson e binomial negativo por meio de um gráfico que apresenta as distribuições de probabilidades observadas e previstas de incidência de imóveis à venda por quadrícula.
- m. Compare também a qualidade do ajuste dos dois modelos (Poisson e binomial negativo) por meio da análise das diferenças máximas entre as distribuições de probabilidades observadas e previstas que ocorrem em ambos os casos. Além disso, elabore esta análise comparando os valores totais de Pearson das duas estimativas.
- n. Estime um modelo de regressão múltipla log-linear por OLS e compare os resultados previstos deste modelo com aqueles estimados pelos modelos de regressão Poisson e binomial negativo.

APÊNDICE

Modelos de regressão inflacionados de zeros

A) Breve Introdução

Como parte dos **Modelos Lineares Generalizados**, os modelos de regressão para dados de contagem são utilizados para os casos em que o fenômeno que se deseja estudar apresenta-se na forma de uma variável quantitativa, porém apenas com valores discretos e não negativos, conforme estudamos ao longo do capítulo. Entretanto, é comum que algumas variáveis com dados de contagem apresentem uma **quantidade excessiva de zeros**, o que pode fazer com que parâmetros estimados quando da elaboração de modelos tradicionais de regressão dos tipos Poisson ou binomial negativo sejam viesados por não conseguirem capturar a presença exacerbada de contagens nulas. Nessas situações, podem ser utilizados os **modelos de regressão inflacionados de zeros**, e neste apêndice estudaremos tais modelos também com foco nos tipos Poisson e binomial negativo.¹

Os modelos de regressão inflacionados de zeros, de acordo com Lambert (1992), são considerados uma combinação entre um modelo para dados de contagem e um modelo para dados binários, já que são utilizados para investigar as razões que levam a determinada quantidade de ocorrências (contagens) de um fenômeno, bem como as razões que levam (ou não) à ocorrência propriamente dita desse fenômeno, independentemente da quantidade de contagens observadas.

Neste sentido, enquanto um modelo Poisson inflacionado de zeros é estimado a partir da **combinação de uma distribuição de Bernoulli com uma distribuição Poisson**, determinado modelo binomial negativo inflacionado de zeros é estimado por meio da **combinação de uma distribuição de Bernoulli com uma distribuição Poisson-Gama**, e a escolha de um ou de outro obedece ao que estudamos ao longo do capítulo, ou seja, passa pela existência de superdispersão nos dados, ou seja, pela análise do inverso do parâmetro de forma da distribuição Gama e do correspondente teste de razão de verossimilhança para o referido parâmetro. Voltaremos a discutir essa questão mais adiante, quando da elaboração de um exemplo em Stata.

A própria definição sobre a existência ou não de uma quantidade excessiva de zeros na variável dependente Y é elaborada por meio de um teste específico, conhecido por **teste de Vuong** (1989), que representará o primeiro *output* a ser analisado na estimação de modelos de regressão inflacionados de zeros.

Em relação específica aos **modelos de regressão Poisson inflacionados de zeros**, podemos definir que, enquanto a **probabilidade p de ocorrência de nenhuma contagem** para dada observação i ($i = 1, 2, \dots, n$, em que n é o tamanho da amostra), ou seja, $p(Y_i = 0)$, é calculada levando-se em consideração a soma de um componente dicotômico com um componente de contagem e, portanto, deve-se definir a probabilidade P_{logit_i} de não ocorrer nenhuma contagem devido exclusivamente ao componente dicotômico, a **probabilidade p de ocorrência de determinada contagem m** ($m = 1, 2, \dots$), ou seja, $p(Y_i = m)$, segue a própria expressão da probabilidade da distribuição Poisson, multiplicada por $(1 - P_{logit_i})$.

Portanto, fazendo uso das expressões (13.10) e (14.1), temos que:

$$\begin{cases} p(Y_i = 0) = P_{logit_i} + (1 - P_{logit_i}) \cdot e^{-\lambda_i} \\ p(Y_i = m) = (1 - P_{logit_i}) \cdot \frac{e^{-\lambda_i} \cdot \lambda_i^m}{m!}, \quad m = 1, 2, \dots \end{cases} \quad (14.32)$$

sendo $Y \sim ZIP(\lambda, P_{logit_i})$, em que ZIP significa **zero inflated Poisson**, e sabendo-se que:

¹ É importante mencionar que, alternativamente aos modelos de regressão inflacionados de zeros dos tipos Poisson e binomial negativo, o pesquisador também pode optar pela estimação de modelos *hurdle* quando do estudo do comportamento de determinada variável dependente com dados de contagem e quantidade excessiva de zeros. Os modelos *hurdle*, embora não contemplados na presente edição deste livro, podem ser estudados em Cameron e Trivedi (2009).

$$p_{logit_i} = \frac{1}{1 + e^{-(\gamma + \delta_1 W_{1i} + \delta_2 W_{2i} + \dots + \delta_q W_{qi})}} \quad (14.33)$$

e

$$\lambda_i = e^{(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki})} \quad (14.34)$$

Podemos verificar que, se $p_{logit_i} = 0$, claramente a distribuição de probabilidades da expressão (14.32) se resume à distribuição Poisson, inclusive para casos em que $Y_i = 0$. Em outras palavras, os modelos de regressão Poisson inflacionados de zeros apresentam dois processos geradores de zeros, sendo um devido à distribuição binária (neste caso, são gerados os chamados **zeros estruturais**) e outro devido à distribuição Poisson (nesta situação, são gerados dados de contagem, entre os quais os chamados **zeros amostrais**).²

Com base nas expressões (14.33) e (14.34), podemos, portanto, definir que, enquanto a ocorrência de zeros estruturais é influenciada por um vetor de variáveis explicativas W_1, W_2, \dots, W_q , a ocorrência de determinada contagem m é influenciada por um vetor de variáveis X_1, X_2, \dots, X_k . Em alguns casos, o pesquisador pode inserir a mesma variável nos dois vetores, caso deseje investigar se essa variável influencia, concomitantemente, a ocorrência do evento e, em caso afirmativo, a quantidade de ocorrências (contagens) do referido fenômeno.

A partir da expressão (14.32), e seguindo a lógica para a definição do **logaritmo da função de verossimilhança (log likelihood function)** apresentado na expressão 14.7, podemos chegar à seguinte função-objetivo, que tem por intuito estimar os parâmetros $\alpha, \beta_1, \beta_2, \dots, \beta_k$ e $\gamma, \delta_1, \delta_2, \dots, \delta_k$ de determinado modelo de regressão Poisson inflacionado de zeros:

$$\begin{aligned} LL = & \sum_{Y_i=0} \ln \left[p_{logit_i} + (1 - p_{logit_i}) \cdot e^{-\lambda_i} \right] + \\ & \sum_{Y_i>0} \left[\ln(1 - p_{logit_i}) - \lambda_i + (Y_i) \cdot \ln(\lambda_i) - \ln(Y_i!) \right] = \text{máx} \end{aligned} \quad (14.35)$$

cuja solução, assim como apresentado ao longo do capítulo, pode ser obtida por meio de ferramentas de programação linear.

Já em relação aos **modelos de regressão do tipo binomial negativo inflacionados de zeros**, podemos definir que, enquanto a **probabilidade p de ocorrência de nenhuma contagem** para dada observação i , ou seja, $p(Y_i = 0)$, é também calculada levando-se em consideração a soma de um componente dicotômico com um componente de contagem, a **probabilidade p de ocorrência de determinada contagem m** ($m = 1, 2, \dots$), ou seja, $p(Y_i = m)$, segue agora a expressão da probabilidade da distribuição Poisson-Gama. Nesse sentido, fazendo uso das expressões (13.10) e (14.25), temos que:

$$\begin{cases} p(Y_i = 0) = p_{logit_i} + (1 - p_{logit_i}) \cdot \left(\frac{1}{1 + \phi u_i} \right)^{\frac{1}{\phi}} \\ p(Y_i = m) = (1 - p_{logit_i}) \cdot \left[\binom{m + \phi^{-1} - 1}{\phi^{-1} - 1} \cdot \left(\frac{1}{1 + \phi u_i} \right)^{\frac{1}{\phi}} \cdot \left(\frac{\phi u_i}{\phi u_i + 1} \right)^m \right], \quad m = 1, 2, \dots \end{cases} \quad (14.36)$$

sendo $Y \sim \text{ZINB}(\phi, u, P_{logit})$, em que ZINB significa *zero inflated negative binomial* e ϕ representa o inverso do parâmetro de forma de determinada distribuição Gama, e sabendo-se, de forma análoga ao apresentado para os modelos de regressão Poisson inflacionados de zeros, que:

$$p_{logit_i} = \frac{1}{1 + e^{-(\gamma + \delta_1 W_{1i} + \delta_2 W_{2i} + \dots + \delta_q W_{qi})}} \quad (14.37)$$

² Note que a expressão (14.33) refere-se ao modelo logit estudado no Capítulo 13. O pesquisador pode, entretanto, optar por utilizar a expressão de probabilidades do modelo probit, estudada no apêndice do mesmo capítulo, para investigar a existência de zeros estruturais referentes à distribuição de Bernoulli.

e

$$u_i = e^{(\alpha + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \dots + \beta_k \cdot X_{ki})} \quad (14.38)$$

Podemos novamente verificar que, se $p_{logit_i} = 0$, a distribuição de probabilidades da expressão (14.36) se resume à distribuição Poisson-Gama, inclusive para casos em que $Y_i = 0$. Logo, os modelos de regressão do tipo binomial negativo inflacionados de zeros também apresentam dois processos geradores de zeros, oriundos da distribuição binária e da distribuição Poisson-Gama.

Portanto, com base na expressão (14.36), e a partir do logaritmo da função de verossimilhança (*log likelihood function*) definido na expressão 14.29, chegamos à seguinte função-objetivo, que tem por intuito estimar os parâmetros, ϕ , α , β_1 , β_2 , ..., β_k e γ , δ_1 , δ_2 , ..., δ_k de determinado modelo de regressão binomial negativo inflacionado de zeros:

$$\begin{aligned} LL = \sum_{Y_i=0} \ln \left[p_{logit_i} + (1 - p_{logit_i}) \cdot \left(\frac{1}{1 + \phi \cdot u_i} \right)^{\frac{1}{\phi}} \right] + \\ \sum_{Y_i>0} \left[\ln(1 - p_{logit_i}) + Y_i \cdot \ln \left(\frac{\phi \cdot u_i}{1 + \phi \cdot u_i} \right) - \frac{\ln(1 + \phi \cdot u_i)}{\phi} \right. \\ \left. + \ln \Gamma(Y_i + \phi^{-1}) - \ln \Gamma(Y_i + 1) - \ln \Gamma(\phi^{-1}) \right] = \text{máx} \end{aligned} \quad (14.39)$$

cuja solução também pode ser obtida por meio de ferramentas de programação linear.

Na sequência, apresentaremos um exemplo elaborado em Stata, em que são estimados os parâmetros de um modelo de regressão Poisson e de um modelo de regressão binomial negativo, ambos inflacionados de zeros. Inicialmente, será estudada a significância da quantidade excessiva de zeros na variável dependente Y (teste de Vuong) para, posteriormente, ser avaliada a significância do inverso parâmetro de forma ϕ da distribuição Gama (teste de razão de verossimilhança para o parâmetro ϕ), ou seja, a existência de superdispersão nos dados. O Quadro 14.2 apresenta a relação entre os modelos de regressão para dados de contagem e a existência de superdispersão e de excesso de zeros nos dados da variável dependente.

Quadro 14.2 Modelos de regressão para dados de contagem, superdispersão e excesso de zeros nos dados da variável dependente.

Verificação	Modelo de Regressão para Dados de Contagem			
	Poisson	Binomial Negativo	Poisson Inflacionado de Zeros (ZIP)	Binomial Negativo Inflacionado de Zeros (ZINB)
Superdispersão nos Dados da Variável Dependente	Não	Sim	Não	Sim
Quantidade Excessiva de Zeros na Variável Dependente	Não	Não	Sim	Sim

Desta forma, enquanto os modelos inflacionados de zeros dos tipos Poisson e binomial negativo são mais apropriados quando houver uma quantidade excessiva de zeros na variável dependente, o uso desses últimos é ainda mais recomendável quando houver superdispersão nos dados.

B) Exemplo: Modelo de Regressão Poisson Inflacionado de Zeros no Stata

A fim de elaborarmos modelos de regressão inflacionados de zeros, faremos uso do banco de dados **Acidentes.dta**. Para a elaboração dessa base, foi investigada a quantidade de acidentes de trânsito que ocorreram em uma semana em 100 cidades de determinado país, que representa a variável dependente com dados de contagem. Além disso, inseriu-se na base a população urbana, a idade média dos habitantes com carteira de habilitação em vigência e o fato de o município adotar lei seca após as 22:00h. O comando **desc** permite que estudemos as características do banco de dados, conforme mostra a Figura 14.76.

. desc				
obs: 100				
vars: 4				
size: 1,700 (99.9% of memory free)				
variable name	storage type	display format	value label	variable label
acidentes	byte	%8.0g		quantidade de acidentes de trânsito na última semana
pop	float	%9.5f		população urbana (x milhão)
idade	float	%9.2f		idade média dos habitantes com carteira de habilitação em vigência
leiseca	float	%9.0g	leiseca	o município adota seia seca após as 22:00h?
Sorted by:				

Figura 14.76 Descrição do Banco de Dados **Acidentes.dta**.

Neste exemplo, vamos definir a variável *pop* como variável X , e as variáveis *idade* e *leiseca* como variáveis W_1 e W_2 . Em outras palavras, nosso intuito é verificar se a probabilidade de não ocorrência de acidentes, ou seja, de ocorrência de zeros estruturais, é influenciada pela idade média dos motoristas e pelo fato de haver lei seca após as 22:00h nos municípios e, além disso, se a ocorrência de determinada contagem de acidentes na semana em estudo é influenciada pela população de cada município i ($i = 1, \dots, 100$). Portanto, para o modelo de regressão Poisson inflacionado de zeros, devem ser estimados os parâmetros das seguintes expressões:

$$plogit_i = \frac{1}{1 + e^{-(\gamma + \delta_1 \cdot \text{idade}_i + \delta_2 \cdot \text{leiseca}_i)}}$$

e

$$\lambda_i = e^{\alpha + \beta \cdot \text{pop}_i}$$

Inicialmente, vamos analisar a distribuição de frequências da variável *acidentes*, digitando os seguintes comandos:

```
tab acidentes
hist acidentes, discrete freq
```

As Figuras 14.77 e 14.78 apresentam, respectivamente, a tabela de frequências e o histograma e, por meio deles, é possível verificarmos, para o país em estudo, que 58% dos municípios analisados não apresentaram nenhum acidente de trânsito na semana pesquisada, o que indica, ainda que de forma preliminar, a existência de uma quantidade excessiva de zeros na variável dependente.

. tab acidentes			
quantidade de acidentes de trânsito na última semana	Freq.	Percent	Cum.
0	58	58.00	58.00
1	8	8.00	66.00
2	6	6.00	72.00
3	6	6.00	78.00
4	4	4.00	82.00
5	3	3.00	85.00
6	2	2.00	87.00
7	1	1.00	88.00
8	2	2.00	90.00
9	2	2.00	92.00
10	1	1.00	93.00
14	1	1.00	94.00
16	1	1.00	95.00
20	1	1.00	96.00
25	1	1.00	97.00
30	1	1.00	98.00
31	1	1.00	99.00
33	1	1.00	100.00
Total	100	100.00	

Figura 14.77 Tabela de frequências da variável dependente **acidentes**.

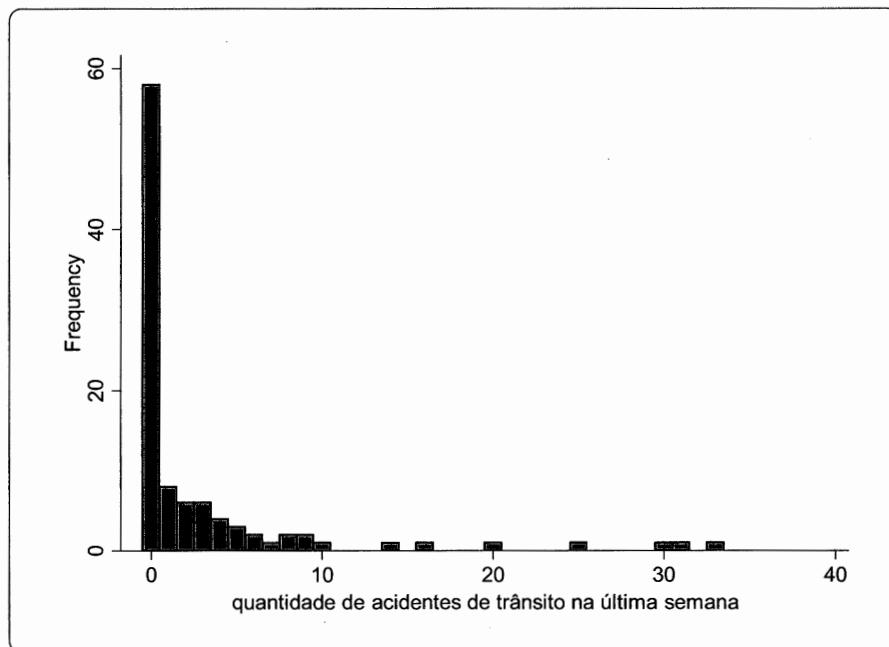


Figura 14.78 Histograma da variável dependente *acidentes*.

Para a elaboração do modelo de regressão Poisson inflacionado de zeros, devemos digitar o seguinte comando:

```
zip acidentes pop, inf(idade leiseca) vuong nolog
```

em que a variável explicativa X (*pop*) deve vir logo após a variável dependente (*acidentes*) e as variáveis W_1 e W_2 (*idade* e *leiseca*) devem vir entre parêntesis, logo após o termo **inf**, que significa **inflate** e corresponde à inflação de zeros estruturais. O termo **vuong** faz com que seja elaborado o teste de Vuong (1989), destinado à verificação da adequação do modelo inflacionado de zeros em relação ao modelo tradicional especificado (neste caso, Poisson), ou seja, tem por finalidade verificar a existência de uma quantidade excessiva de zeros na variável dependente. O termo **nolog** faz com que sejam omitidos os *outputs* referentes às iterações da modelagem para que já seja apresentado o valor máximo do logaritmo da função de verossimilhança.

Além disso, é importante mencionar que o comando apresentado oferece implicitamente, como padrão, a **expressão de probabilidades do modelo logit** para a verificação de existência de zeros estruturais referentes à distribuição de Bernoulli. Entretanto, caso o pesquisador opte por trabalhar com a **expressão de probabilidades do modelo probit**, estudada no apêndice do Capítulo 13, deverá adicionar o termo **probit** ao final do comando.

Os *outputs* encontram-se na Figura 14.79.

Zero-inflated Poisson regression						
				Number of obs	=	100
				Nonzero obs	=	42
				Zero obs	=	58
Inflation model = logit				LR chi2(1)	=	37.72
Log likelihood = -256.0484				Prob > chi2	=	0.0000
<hr/>						
acidentes	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
acidentes						
pop	.5039652	.0863993	5.83	0.000	.3346256	.6733047
_cons	.9329778	.1987482	4.69	0.000	.5434386	1.322517
inflate						
idade	.2252293	.0584096	3.86	0.000	.1107485	.3397101
leiseca	1.725743	.5531873	3.12	0.002	.6415157	2.80997
_cons	-11.72936	3.030402	-3.87	0.000	-17.66884	-5.789881
<hr/>						
Vuong test of zip vs. standard Poisson:				z =	4.19	Pr>z = 0.0000

Figura 14.79 Outputs do modelo de regressão Poisson inflacionado de zeros no Stata.

O primeiro resultado que deve ser analisado refere-se ao teste de Vuong, cuja estatística é normalmente distribuída, com valores positivos e significantes indicando a adequação do modelo Poisson inflacionado de zeros, e com valores negativos e significantes indicando a adequação do modelo tradicional Poisson. Para os dados do nosso exemplo, podemos verificar que o teste de Vuong indica a melhor adequação do modelo inflacionado de zeros sobre o modelo tradicional, visto que $z = 4,19$ e $Pr > z = 0,000$.

Antes de analisarmos os demais *outputs*, é importante mencionar que Desmarais e Harden (2013) propõem uma correção ao teste de Vuong, que se baseia nas estatísticas **Akaike information criterion (AIC)** e **Bayesian (Schwarz) information criterion (BIC)** e que deve ser elaborada para que se eliminem eventuais vieses que podem prejudicar a decisão sobre a escolha do modelo mais adequado. Para tanto, basta que seja substituído o termo **zip** pelo termo **zipcv** (que significa *zero inflated Poisson with corrected Vuong*), e o novo comando ficará conforme segue:

```
zipcv accidentes pop, inf(idade leiseca) vuong nolog
```

porém antes de sua elaboração no Stata, devemos instalar o comando **zipcv**, digitando **findit zipcv** e clicando no link [st0319 from http://www.stata-journal.com/software/sj13-4](http://www.stata-journal.com/software/sj13-4). Na sequência, devemos clicar em **click here to install**.

Os novos *outputs* estão na Figura 14.80.

. zipcv accidentes pop, inf(idade leiseca) vuong nolog						
Zero-inflated Poisson regression			Number of obs	=	100	
			Nonzero obs	=	42	
			Zero obs	=	58	
Inflation model = logit			LR chi2(1)	=	37.72	
Log likelihood = -256.0484			Prob > chi2	=	0.0000	
<hr/>						
accidentes	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
accidentes						
pop	.5039652	.0863993	5.83	0.000	.3346256	.6733047
_cons	.9329778	.1987482	4.69	0.000	.5434386	1.322517
inflated						
idade	.2252293	.0584096	3.86	0.000	.1107485	.3397101
leiseca	1.725743	.5531873	3.12	0.002	.6415157	2.80997
_cons	-11.72936	3.030402	-3.87	0.000	-17.66884	-5.789881
<hr/>						
Vuong test of zip vs. standard Poisson: z = 4.19 Pr>z = 0.0000						
Pr<z = 1.0000						
with AIC (Akaike) correction: z = 4.13 Pr>z = 0.0000						
Pr<z = 1.0000						
with BIC (Schwarz) correction: z = 4.04 Pr>z = 0.0000						
Pr<z = 1.0000						

Figura 14.80 Outputs do modelo de regressão Poisson inflacionado de zeros com correção no teste de Vuong.

Para os dados do nosso exemplo, enquanto a estatística do teste de Vuong é $z = 4,19$, as estatísticas com correção AIC e BIC são $z = 4,13$ e $z = 4,04$, respectivamente, ou seja, todas apresentam $Pr > z = 0,000$. Em outras palavras, os resultados do teste de Vuong com correção AIC e BIC continuam permitindo, neste caso, que afirmemos que o modelo inflacionado de zeros é mais apropriado.

Note que os demais *outputs* apresentados nas Figuras 14.79 e 14.80 são exatamente os mesmos. Com base nesses *outputs*, podemos verificar que os parâmetros estimados são estatisticamente diferentes de zero, a 95% de confiança, e as expressões finais de P_{logit_i} e de λ_i são dadas por:

$$P_{logit_i} = \frac{1}{1 + e^{-(11,729 + 0,225 \cdot idade_i + 1,726 \cdot leiseca_i)}}$$

$$\lambda_i = e^{(0,933 + 0,504 \cdot pop_i)}$$

Um pesquisador mais curioso poderá obter esses mesmos *outputs* por meio do arquivo **Acidentes ZIP Máxima Verossimilhança.xls**, usando a ferramenta **Solver** do Excel, conforme padrão também adotado ao longo do capítulo e do livro. Nesse arquivo, os critérios do **Solver** já estão previamente definidos.

Portanto, fazendo uso da expressão (14.32) e dos parâmetros estimados, podemos calcular algebraicamente, da seguinte forma, a quantidade média esperada de acidentes de trânsito na semana para um município com 700.000 habitantes, com idade média de seus motoristas igual a 40 anos e que não adota a lei seca após as 22:00h:

$$\lambda_{inflate} = \left\{ 1 - \frac{1}{1 + e^{-[-11,729 + 0,225.(40) + 1,726.(0)]}} \right\} \cdot \left\{ e^{[0,933 + 0,504.(0,700)]} \right\} = 3,39$$

O mesmo resultado pode ser encontrado pelo pesquisador caso seja digitado o seguinte comando, cujo *output* encontra-se na Figura 14.81:

```
mfx, at(pop=0.7 idade=40 leiseca=0)
```

Marginal effects after zip							
	Y = Predicted number of events (predict)						
	= 3.3938647						
variable	dy/dx	Std. Err.	z	P> z	[95% C.I.]
pop	1.71039	.14686	11.65	0.000	1.42256	1.99822	.7
idade	-.0472341	.02209	-2.14	0.032	-.090529	-.003939	40
leiseca*	-.7532942	.43112	-1.75	0.081	-1.59827	.091684	0

(* dy/dx is for discrete change of dummy variable from 0 to 1)

Figura 14.81 Cálculo da quantidade esperada de acidentes semanais para valores das variáveis explicativas – comando **mfx**.

Por fim, podemos, por meio de um gráfico, comparar os valores previstos da quantidade média de acidentes de trânsito na semana obtidos pelo modelo de regressão Poisson inflacionado de zeros com aqueles que seriam obtidos por um modelo tradicional de regressão Poisson, sem considerar, portanto, as variáveis que influenciam a ocorrência de zeros estruturais, ou seja, o componente dicotômico (variáveis *idade* e *leiseca*). Para tanto, podemos digitar a seguinte sequência de comandos:

```
quietly zipcv acidentes pop, inf(idade leiseca) vuong nolog
predict lambda_inf

quietly poisson acidentes pop
predict lambda

graph twoway scatter acidentes pop || mspline lambda_inf pop || mspline
lambda pop ||, legend(label(2 "ZIP") label(3 "Poisson"))
```

O gráfico gerado é apresentado na Figura 14.82 e, por meio dele, podemos verificar que os valores previstos pelo modelo de regressão Poisson inflacionado de zeros (ZIP) ajustam-se de forma mais adequada à quantidade excessiva de zeros na variável dependente.

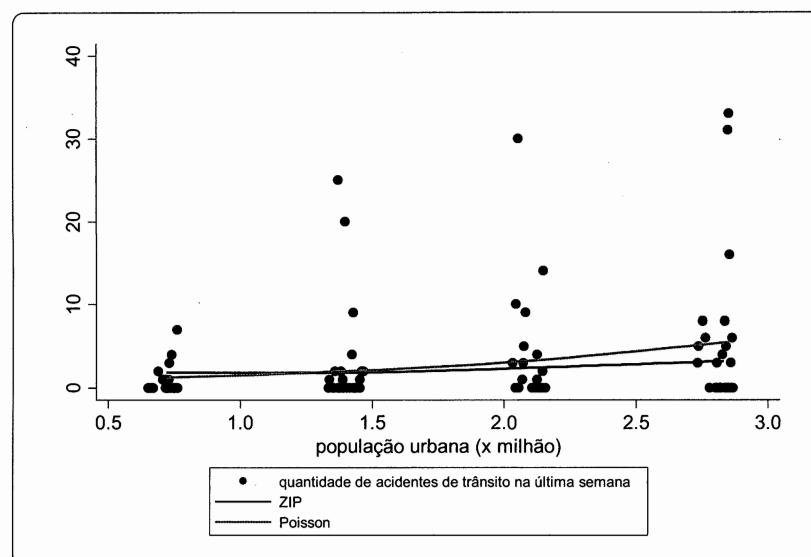


Figura 14.82 Quantidade esperada de acidentes de trânsito por semana x população do município (*pop*) para os modelos ZIP e Poisson.

Na sequência, vamos analisar, com base no mesmo banco de dados, os resultados obtidos por meio da estimação de um modelo de regressão binomial negativo inflacionado de zeros.

C) Exemplo: Modelo de Regressão Binomial Negativo Inflacionado de Zeros no Stata

Seguindo a mesma lógica, vamos fazer uso novamente do banco de dados **Acidentes.dta**, porém agora com foco na estimação de um modelo de regressão binomial negativo inflacionado de zeros. Portanto, serão estimados os parâmetros das seguintes expressões:

$$p_{logit_i} = \frac{1}{1 + e^{-(\gamma + \delta_1.idade_i + \delta_2.leiseca_i)}}$$

e

$$u_i = e^{(\alpha + \beta.pop_i)}$$

Assim como discutido ao longo do capítulo, vamos inicialmente analisar a média e a variância da variável *acidentes*, digitando o seguinte comando:

```
tabstat acidentes, stats(mean var)
```

A Figura 14.83 apresenta o resultado gerado.

. tabstat acidentes, stats(mean var)		
variable	mean	variance
acidentes	3.01	42.9999

Figura 14.83 Média e variância da variável dependente *acidentes*.

Conforme podemos verificar, a variância da variável dependente é aproximadamente 14 vezes maior do que a sua média, o que representa um forte indício da existência de superdispersão nos dados. Vamos, portanto, partir para a estimação do modelo de regressão binomial negativo inflacionado de zeros e, para tanto, devemos digitar o seguinte comando:

```
zinbcv acidentes pop, inf(idade leiseca) vuong nolog zip
```

que possui a mesma lógica do comando utilizado para a estimação do modelo ZIP. Note que optamos por utilizar o termo **zinbcv** (*zero inflated negative binomial with corrected Vuong*) em vez do termo **zinb**, visto que, embora os parâmetros estimados sejam exatamente iguais, o primeiro apresenta os resultados do teste de Vuong com correção AIC e BIC. Além disso, o termo **zip** ao final do comando faz com que seja elaborado o teste de razão de verossimilhança para o parâmetro ϕ (*alpha* no Stata), ou seja, propicia uma comparação da adequação do modelo ZINB em relação ao modelo ZIP. Os *outputs* são apresentados na Figura 14.84.

Inicialmente, podemos verificar que o intervalo de confiança do parâmetro ϕ , que é o inverso do parâmetro de forma ψ da distribuição binomial negativa e que o Stata cita como **alpha**, não contém o zero, ou seja, para o nível de confiança de 95%, podemos afirmar que ϕ é estatisticamente diferente de zero e com valor estimado igual a 1,271. Por meio do teste de razão de verossimilhança para o parâmetro ϕ , pode-se concluir que a hipótese nula de que este parâmetro seja estatisticamente igual a zero pode ser rejeitada ao nível de significância de 5% (*Sig. χ^2* = 0,000 < 0,05), o que comprova a existência de superdispersão nos dados e indica que o modelo ZINB é preferível ao modelo ZIP.

Além disso, o teste de Vuong com correção AIC e BIC, por apresentar significantes estatísticas z a 95% de confiança, indica que o modelo binomial negativo inflacionado de zeros (ZINB) seja preferível ao modelo tradicional binomial negativo, pois comprova a existência de uma quantidade excessiva de zeros.

Também podemos verificar que o parâmetro estimado da variável *pop* é estatisticamente diferente de zero a 95% de confiança, ou seja, esta variável é significante para explicar o comportamento da quantidade de acidentes

de trânsito na semana (componente de contagem). Da mesma forma, as variáveis *idade* e *leiseca* são estatisticamente significantes para explicar a quantidade excessiva de zeros (zeros estruturais) na variável *acidentes* (componente dicotômico).

<code>. zinbcv acidentes pop, inf(idade leiseca) vuong nolog zip</code>						
Zero-inflated negative binomial regression			Number of obs	=	100	
			Nonzero obs	=	42	
			Zero obs	=	58	
Inflation model = logit						
Log likelihood = -164.4035						
acidentes	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
acidentes						
pop	.8661751	.2621428	3.30	0.001	.3523847	1.379966
_cons	.0253062	.5403137	0.05	0.963	-1.033689	1.084301
inflate						
idade	.2882047	.0998951	2.89	0.004	.0924139	.4839954
leiseca	2.85907	1.076625	2.66	0.008	.7489239	4.969217
_cons	-16.23734	5.726858	-2.84	0.005	-27.46178	-5.012905
/lnalpha						
	.2399887	.3137446	0.76	0.444	-.3749393	.8549167
alpha						
	1.271235	.398843			.687331	2.351179
Likelihood-ratio test of alpha=0: chibar2(01) = 183.29 Pr>=chibar2 = 0.0000						
Vuong test of zinb vs. standard negative binomial: z = 3.88 Pr>z = 0.0001						
Pr<z = 0.9999						
with AIC (Akaike) correction: z = 3.31 Pr>z = 0.0005						
Pr<z = 0.9995						
with BIC (Schwarz) correction: z = 2.57 Pr>z = 0.0051						
Pr<z = 0.9949						

Figura 14.84 Outputs do modelo de regressão inflacionado de zeros no Stata.

Com base nesses outputs, podemos chegar às expressões finais de P_{logit_i} e de u_i , dadas por:

$$P_{logit_i} = \frac{1}{1 + e^{-(16.237 + 0.288 \cdot idade_i + 2.859 \cdot leiseca_i)}}$$

e

$$u_i = e^{(0.025 + 0.866 \cdot pop_i)}$$

Assim, um pesquisador curioso poderá obter esses mesmos outputs por meio do arquivo **Acidentes ZINB Máxima Verossimilhança.xls**, fazendo uso da ferramenta **Solver** do Excel, conforme padrão também adotado ao longo do capítulo e do livro. Nesse arquivo, os critérios do **Solver** já estão previamente definidos.

Fazendo uso da expressão (14.36) e dos parâmetros estimados, podemos novamente calcular, de forma algébrica, a quantidade média esperada de acidentes de trânsito na semana para um município com 700.000 habitantes, com idade média de seus motoristas igual a 40 anos e que não adota a lei seca após as 22:00h, conforme segue:

$$u_{inflated} = \left\{ 1 - \frac{1}{1 + e^{-[-16.237 + 0.288 \cdot (40) + 2.859 \cdot (0)]}}} \right\} \cdot \left\{ e^{[0.025 + 0.866 \cdot (0.700)]} \right\} = 1.86$$

O mesmo resultado também pode ser encontrado pelo pesquisador se digitado o seguinte comando, cujo output é apresentado na Figura 14.85:

mfx, at(pop=0.7 idade=40 leiseca=0)

```

. mfx, at( pop=0.7 idade=40 leiseca=0)
Marginal effects after zinb
Y = Predicted number of events (predict)
= 1.8638732
-----+
variable | dy/dx   Std. Err.      z    P>|z| [ 95% C.I. ]   x
-----+
pop | 1.614441 .29961  5.39  0.000  1.02722  2.20166 .7
idade | -.004798 .00811 -0.59  0.554 -.020686 .01109 40
leiseca | -.2387158 .26031 -0.92  0.359 -.74891 .271479 0
-----+
(*) dy/dx is for discrete change of dummy variable from 0 to 1

```

Figura 14.85 Cálculo da quantidade esperada de acidentes semanais para valores das variáveis explicativas – comando **mfx**.

Em tese, a modelagem poderia ser, neste momento, finalizada. Entretanto, se houver também o interesse em estimar os parâmetros de um modelo ZIP, a fim apenas de compará-los com os obtidos pelo modelo ZINB, poderemos digitar a seguinte sequência de comandos:

```

eststo: quietly zip acidentes pop, inf(idade leiseca) vuong
prcounts lambda_inflate, plot

eststo: quietly zinb acidentes pop, inf(idade leiseca) vuong
prcounts u_inflate, plot

esttab, scalars(l1) se

```

que gera os *outputs* apresentados na Figura 14.86.

```

. eststo: quietly zip acidentes pop, inf(idade leiseca) vuong
(est1 stored)
. prcounts lambda_inflate, plot

. eststo: quietly zinb acidentes pop, inf(idade leiseca) vuong
(est2 stored)
. prcounts u_inflate, plot

. esttab, scalars(l1) se

-----+
(1)           (2)
acidentes     acidentes
-----+
acidentes
pop          0.504***       0.866***  

              (0.0864)        (0.262)

_cons         0.933***       0.0253  

              (0.199)         (0.540)
-----+
inflate
idade        0.225***       0.288**  

              (0.0584)        (0.0999)

leiseca       1.726**        2.859**  

              (0.553)         (1.077)

_cons         -11.73***      -16.24**  

              (3.030)         (5.727)
-----+
lnalpha
_cons         0.240  

              (0.314)
-----+
N             100          100
11            -256.0       -164.4
-----+
Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

```

Figura 14.86 Principais resultados obtidos nas estimações ZIP e ZINB.

Esses *outputs* consolidados permitem que verifiquemos, além das diferenças entre as estimativas dos parâmetros nos dois modelos, que o valor obtido do **logaritmo da função de verossimilhança** (11, ou **log likelihood**) é consideravelmente maior para o modelo ZINB (modelo 2 na Figura 14.86), o que é mais um indício de melhor adequação deste sobre o modelo ZIP para os dados do nosso exemplo.

Outra maneira de comparar as estimativas dos modelos ZINB e ZIP é por meio da análise das distribuições de probabilidades observadas e previstas da ocorrência de acidentes semanais para essas duas estimativas, analogamente ao que discutimos ao longo do capítulo, fazendo uso das variáveis geradas na elaboração dos comandos **prcounts**. Para tanto, devemos digitar o seguinte comando, que gerará o gráfico da Figura 14.87:

```
graph twoway (scatter u_inflateobeq u_inflatetreq lambda_inflatetreq
u_inflateval, connect (1 1 1))
```

em que as variáveis *u_inflatetreq* e *lambda_inflatetreq* correspondem às probabilidades previstas de ocorrência de 0 a 9 acidentes obtidas, respectivamente, pelos modelos ZINB e ZIP. Além disso, enquanto a variável *u_inflateobeq* corresponde às probabilidades observadas da variável dependente *e*, portanto, apresenta a mesma distribuição de probabilidades apresentada na Figura 14.77 para até 9 acidentes de trânsito, a variável *u_inflateval* apresenta os próprios valores de 0 a 9 que serão relacionados com as probabilidades observadas.

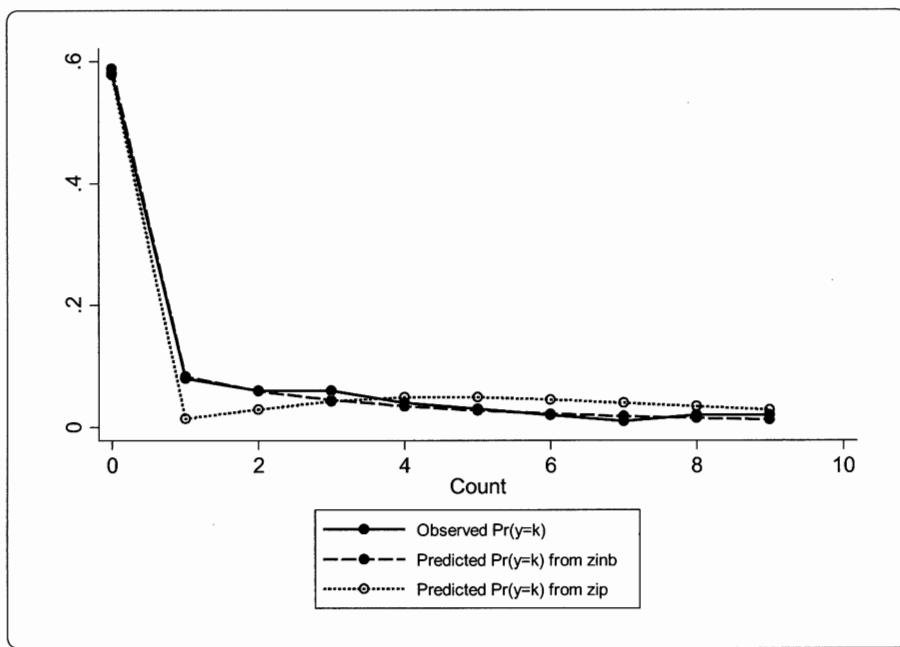


Figura 14.87 Distribuições de probabilidades observadas e previstas de ocorrência de acidentes de trânsito semanais para os modelos ZINB e ZIP.

Por meio da análise do gráfico da Figura 14.87, podemos verificar que a distribuição estimada (prevista) de probabilidades do modelo ZINB se ajusta bem melhor à distribuição observada do que a distribuição estimada de probabilidades do modelo ZIP, para uma contagem de até 9 acidentes de trânsito por semana.

Alternativamente, assim como discutimos ao longo do capítulo, esse fato também pode ser verificado na aplicação do comando **countfit**, que oferece, além dos valores das probabilidades observadas e previstas para cada contagem (de 0 a 9) da variável dependente, os termos de erro resultantes da diferença entre as probabilidades obtidas pelos modelos ZINB e ZIP. Dessa forma, podemos digitar o seguinte comando:

```
countfit acidentes pop, zip zinb noestimates
```

que gera os *outputs* da Figura 14.88 e o gráfico da Figura 14.89.

Comparison of Mean Observed and Predicted Count				
Model	Maximum Difference	At Value	Mean Diff	
ZIP	0.070	1	0.024	
ZINB	0.016	3	0.006	
ZIP: Predicted and actual probabilities				
Count	Actual	Predicted	Diff	Pearson
0	0.580	0.580	0.000	0.000
1	0.080	0.010	0.070	47.385
2	0.060	0.023	0.037	6.248
3	0.060	0.035	0.025	1.839
4	0.040	0.043	0.003	0.021
5	0.030	0.046	0.016	0.566
6	0.020	0.045	0.025	1.412
7	0.010	0.042	0.032	2.441
8	0.020	0.038	0.018	0.826
9	0.020	0.033	0.013	0.495
Sum	0.920	0.894	0.239	61.233
ZINB: Predicted and actual probabilities				
Count	Actual	Predicted	Diff	Pearson
0	0.580	0.580	0.000	0.000
1	0.080	0.090	0.010	0.108
2	0.060	0.059	0.001	0.001
3	0.060	0.044	0.016	0.607
4	0.040	0.034	0.006	0.113
5	0.030	0.027	0.003	0.034
6	0.020	0.022	0.002	0.018
7	0.010	0.018	0.008	0.368
8	0.020	0.015	0.005	0.149
9	0.020	0.013	0.007	0.391
Sum	0.920	0.902	0.058	1.789
Tests and Fit Statistics				
ZIP	BIC=	570.596	AIC=	560.176
vs ZINB	BIC=	391.416	dif=	179.180
	AIC=	378.390	dif=	181.786
			ZINB	ZINB
			ZIP	ZIP
				Very strong

Figura 14.88 Probabilidades observadas e previstas para cada contagem da variável dependente e respectivos termos de erro.

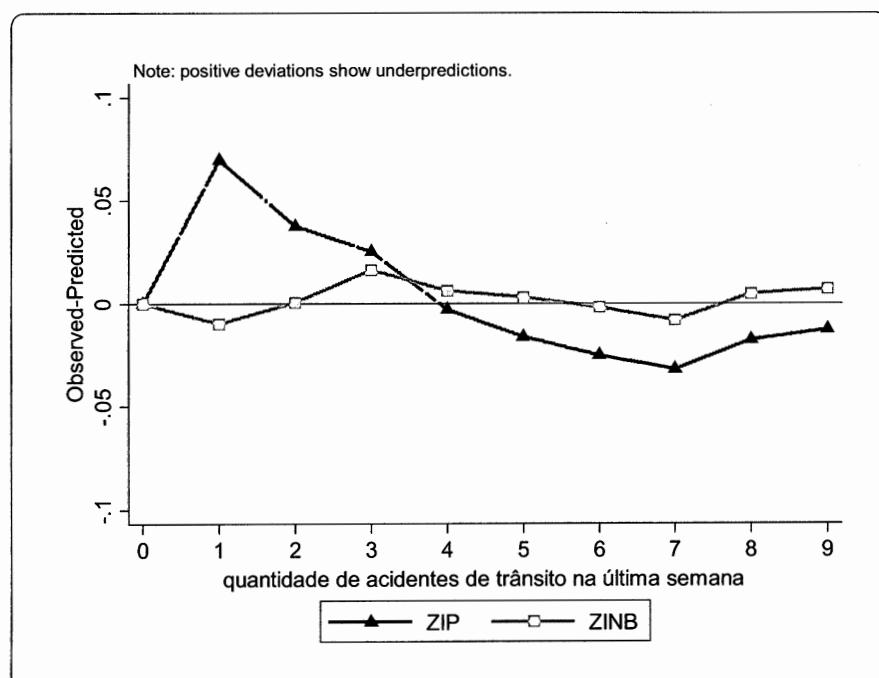


Figura 14.89 Termos de erro resultantes da diferença entre as probabilidades observadas e previstas (modelos ZINB e ZIP).

As Figuras 14.88 e 14.89 nos mostram, mais uma vez, que o ajuste do modelo ZINB é melhor do que o ajuste do modelo ZIP, pelas seguintes razões:

- enquanto a diferença máxima entre as probabilidades observadas e previstas para o modelo ZIP é, em módulo, igual a 0,070, para o modelo ZINB é, em módulo, igual a 0,016.
- a média dessas diferenças é de 0,024 para o modelo ZIP e de 0,006 para o modelo ZINB.
- o valor total de Pearson é mais baixo no modelo ZINB (1,789) do que no modelo ZIP (61,233).

O gráfico da Figura 14.89 permite que a análise comparativa entre os termos de erro gerados nos dois modelos seja elaborada de maneira visual, merecendo destaque o ajuste do modelo ZINB, em que a curva de erros é consistentemente mais próxima de zero.

Assim como realizado anteriormente, podemos também comparar, graficamente, os valores previstos da quantidade média de acidentes de trânsito na semana obtidos pelos modelos ZIP e ZINB com aqueles que seriam obtidos pelos correspondentes modelos tradicionais de regressão dos tipos Poisson e binomial negativo (comando **nbreg**), sem a consideração das variáveis que influenciam apenas ocorrência de zeros estruturais (variáveis *idade* e *leiseca*). Para tanto, podemos digitar a seguinte sequência de comandos:

```
quietly poisson acidentes pop
predict lambda
```

```
quietly nbreg acidentes pop
predict u
```

```
graph twoway mspline lambda_inflaterate pop || mspline u_inflaterate
pop || mspline lambda pop || mspline u pop||, legend(label(1 "ZIP") label(2
"ZINB") label(3 "Poisson") label(4 "Binomial Negativo"))
```

O gráfico gerado é apresentado na Figura 14.90.

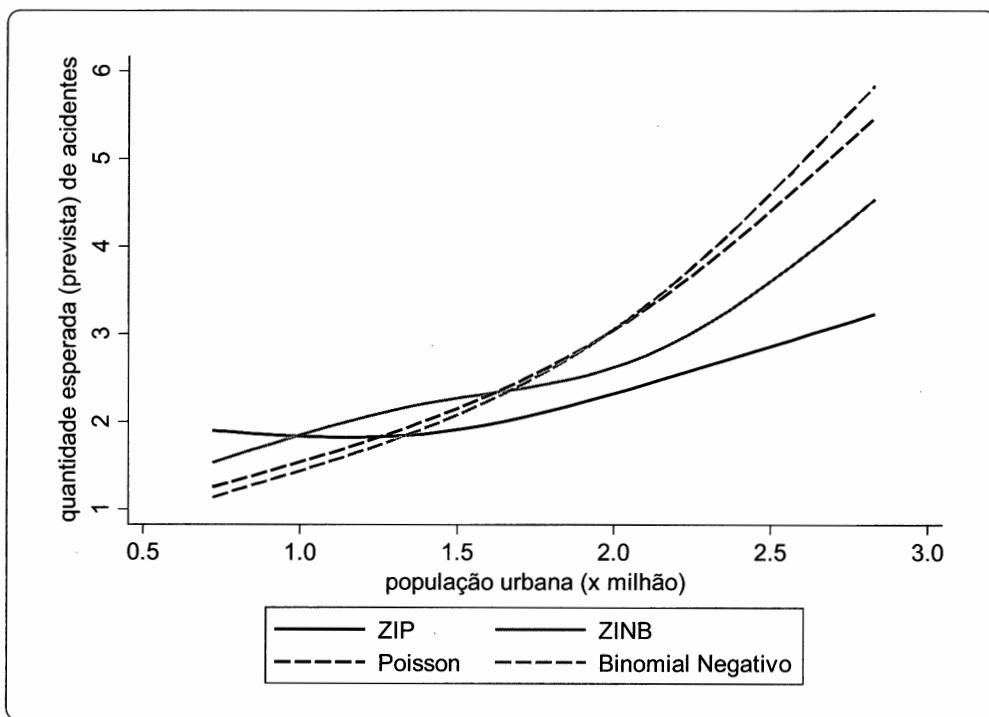


Figura 14.90 Quantidade esperada de acidentes de trânsito por semana x população do município (*pop*) para os modelos ZIP, ZINB, Poisson e binomial negativo.

Duas considerações podem ser feitas em relação a esse gráfico. A primeira diz respeito à variância da quantidade prevista de acidentes semanais, que faz com que as curvas dos modelos ZINB e binomial negativo sejam mais alongadas à parte superior direita do gráfico do que aquelas geradas pelos correspondentes modelos ZIP e Poisson, que não conseguem capturar a existência de superdispersão nos dados. Além disso, podemos também observar que os valores previstos gerados pelos modelos ZINB e ZIP ajustam-se de forma mais adequada à quantidade excessiva de zeros do que os valores previstos gerados pelos modelos Poisson e binomial negativo, visto que apresentam inclinações menores, principalmente para valores mais baixos da quantidade esperada de acidentes.

Neste sentido, é importante que o pesquisador possua uma visão completa dos modelos de regressão para dados de contagem, a fim de que possa estimar, da maneira mais adequada possível, os parâmetros de seu modelo, considerando sempre a natureza e o comportamento da variável dependente que representa o fenômeno em estudo.

MODELOS DE REGRESSÃO PARA DADOS EM PAINEL

Os modelos de regressão para dados em painel são muito úteis quando se deseja estudar o comportamento de determinado fenômeno, representado pela variável dependente, na presença de estruturas de **dados agrupados**, com **medidas repetidas ou longitudinais**.

Enquanto nas **estruturas de dados agrupados** determinadas variáveis explicativas não apresentam variação entre as observações (que representam um nível de análise) provenientes de determinado grupo (que representa outro nível de análise), nas **estruturas de dados com medidas repetidas** existe, além disso, a evolução temporal, fato que permite ao pesquisador investigar as razões individuais que possam levar cada uma das observações a apresentar comportamentos diferentes da variável dependente, para um mesmo grupo ou para grupos distintos, ao longo do tempo. Por exemplo, determinados dados de uma escola que não variam entre seus estudantes, como localização e porte, podem ser comparados com dados de outras escolas; e determinados dados de um estudante, como sexo e religião, que não variam ao longo do tempo, podem ser comparados com dados de outros estudantes, o que permite que sejam analisadas as diferentes influências sobre o comportamento da variável dependente. Em todas essas situações (dados agrupados ou dados com medidas repetidas), os bancos de dados oferecem **estruturas aninhadas**, a partir das quais podem ser estimados **modelos hierárquicos**, também conhecidos por **modelos multinível de regressão para dados em painel**, a serem estudados no Capítulo 16.

No entanto, antes disso, estudaremos, no Capítulo 15, os **modelos longitudinais de regressão para dados em painel**, que podem ser estimados a partir da existência de bancos de dados cujas estruturas (longitudinais) oferecem uma lógica dentro da qual as observações apresentam dados que se alteram ao longo do tempo, tanto para a variável dependente, quanto para as variáveis explicativas, o que permite que o pesquisador estude o comportamento de diversas **cross-sections ao longo do tempo**. Em determinadas áreas, o uso dos bancos de dados com estrutura longitudinal é mais frequente do que os com estrutura aninhada, razão pela qual os modelos longitudinais de regressão para dados em painel são comumente chamados apenas de **modelos de regressão para dados em painel**, mesmo sabendo-se que esses englobam também os modelos de regressão multinível.

A Figura III.2.1 apresenta, para os modelos de regressão para dados em painel, as estruturas de dados agrupados, com medidas repetidas e longitudinais e a relação entre elas, o aninhamento nos dados e a evolução temporal, como foco para o que será estudado nos Capítulos 15 e 16.

Nos três capítulos anteriores, que compõem o que chamamos de **Modelos Lineares Generalizados**, estudamos os modelos de regressão simples e múltipla, os modelos de regressão logística e os modelos de regressão para dados de contagem, com uma abordagem prioritariamente de *cross-section*, ou seja, com exemplos de bancos de dados que reproduzem, de certa forma, uma fotografia do momento em que são coletados os dados. Em outras palavras, **para modelos em cross-section, os indivíduos variam, porém o tempo é fixo**. Além disso, quando estudamos o fenômeno da autocorrelação dos resíduos no Capítulo 12, os exemplos passam a trazer bancos de dados que reproduzem, de certa forma, um filme da evolução temporal de determinadas variáveis, porém para um único indivíduo. Portanto, **para modelos em série temporal, os períodos de tempo variam, porém para um único indivíduo**.

Mantendo essa lógica, no Capítulo 15 estudaremos, por meio de estruturas de dados longitudinais, os **modelos longitudinais lineares de regressão para dados em painel**, que correspondem aos modelos estudados no Capítulo 12, e os **modelos longitudinais não lineares de regressão para dados em painel**, como os modelos logísticos e os modelos Poisson e binomial negativo, que correspondem, respectivamente, aos modelos estudados nos Capítulos 13 e 14.

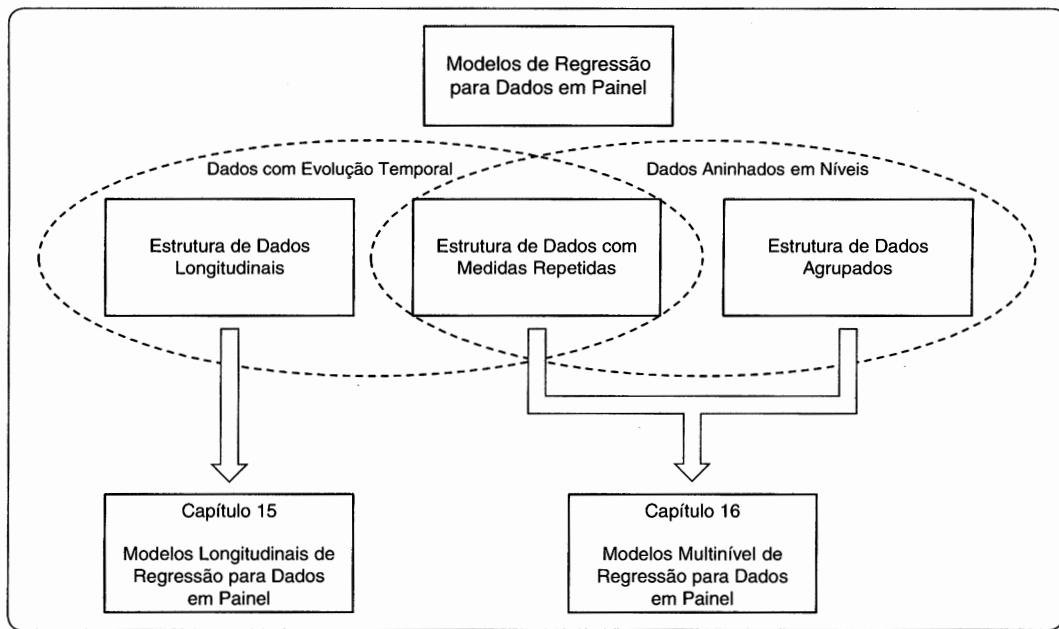


Figura III.2.1 Relação entre Estruturas de Dados, Aninhamento e Evolução Temporal em Modelos de Regressão para Dados em Painel.

Além disso, fazendo uso dos conceitos estudados no Capítulo 12 em relação aos modelos de regressão simples e múltipla e dos conceitos estudados no Capítulo 15 sobre dados com evolução temporal, teremos condições, no Capítulo 16, de estudar, a partir de estruturas de dados agrupados, os **modelos hierárquicos lineares de dois níveis**, e a partir de estruturas de dados com medidas repetidas, os **modelos hierárquicos lineares de três níveis com medidas repetidas**. No apêndice do Capítulo 16 apresentaremos exemplos de **modelos hierárquicos não lineares dos tipos logístico, Poisson e binomial negativo**.

Portanto, a estrutura adotada nos três capítulos anteriores e a correspondência com as seções dos Capítulos 15 e 16 encontram-se na Figura III.2.2.

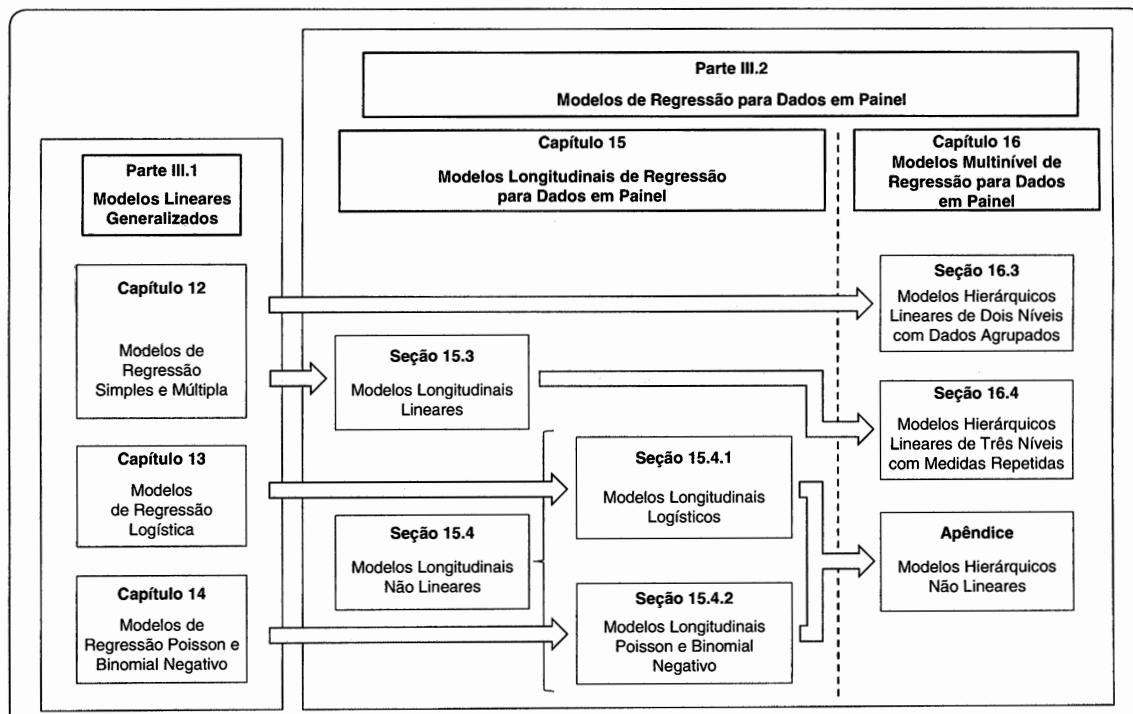


Figura III.2.2 Estrutura dos Capítulos 12, 13 e 14 e Correspondência com os Capítulos 15 e 16.

Em relação especificamente aos modelos longitudinais lineares, a serem estudados no Capítulo 15, faremos distinção entre as estimações que podem ser utilizadas quando o banco de dados oferecer um painel curto, ou seja, apresentar uma quantidade de indivíduos superior à quantidade de períodos, ou um painel longo, que é definido quando a quantidade de períodos exceder o número de indivíduos na amostra.

Seguindo a lógica apresentada no estudo dos três capítulos anteriores, podemos escrever a expressão geral de um modelo longitudinal de regressão para dados em painel da seguinte forma:

$$\eta_{it} = \alpha_i + \beta_1 \cdot X_{1it} + \beta_2 \cdot X_{2it} + \dots + \beta_k \cdot X_{kit} \quad (\text{III.2.1})$$

em que η é conhecido por função de ligação canônica, α representa os termos do intercepto, β_j ($j = 1, 2, \dots, k$) são os coeficientes de cada variável explicativa e correspondem aos parâmetros a serem estimados e X_j são as variáveis explicativas (métricas ou *dummies*), que variam entre indivíduos e ao longo do tempo. Os subscritos i representam cada um dos indivíduos da amostra ($i = 1, 2, \dots, n$, em que n é o tamanho da amostra) e t , os períodos em que são coletados os dados.

O Quadro III.2.1 relaciona cada caso particular dos modelos longitudinais de regressão para dados em painel com a característica da variável dependente, a sua distribuição e a respectiva função de ligação canônica.

Quadro III.2.1 Modelos longitudinais de regressão para dados em painel, características da variável dependente e funções de ligação canônica.

Modelo Longitudinal de Regressão para Dados em Painel	Característica da Variável Dependente	Distribuição	Função de Ligação Canônica (η)
Linear	Quantitativa	Normal	\hat{Y}
Não Linear Logístico	Qualitativa com 2 Categorias (<i>dummy</i>)	Bernoulli	$\ln\left(\frac{p}{1-p}\right)$
Não Linear Poisson	Quantitativa com Valores Inteiros e Não Negativos (Dados de Contagem)	Poisson	$\ln(\lambda)$
Não Linear Binomial Negativo	Quantitativa com Valores Inteiros e Não Negativos (Dados de Contagem)	Poisson-Gama	$\ln(u)$

Logo, para uma dada variável dependente Y , que representa o fenômeno em estudo e que varia entre indivíduos e ao longo do tempo, podemos especificar cada um dos modelos apresentados no Quadro III.2.1 da seguinte maneira:

Modelo Longitudinal Linear:

$$\hat{Y}_{it} = \alpha_i + \beta_1 \cdot X_{1it} + \beta_2 \cdot X_{2it} + \dots + \beta_k \cdot X_{kit} \quad (\text{III.2.2})$$

em que \hat{Y} é o valor esperado da variável dependente Y .

Modelo Longitudinal Não Linear Logístico:

$$\ln\left(\frac{p_{it}}{1-p_{it}}\right) = \alpha_i + \beta_1 \cdot X_{1it} + \beta_2 \cdot X_{2it} + \dots + \beta_k \cdot X_{kit} \quad (\text{III.2.3})$$

em que p é a probabilidade de ocorrência do evento de interesse no instante t para dado indivíduo i .

Modelo Longitudinal Não Linear Poisson:

$$\ln(\lambda_{it}) = \alpha_i + \beta_1 \cdot X_{1it} + \beta_2 \cdot X_{2it} + \dots + \beta_k \cdot X_{kit} \quad (\text{III.2.4})$$

em que λ é o valor esperado da quantidade de ocorrências do fenômeno em estudo (que apresenta distribuição Poisson) no instante t para dado indivíduo i .

Modelo Longitudinal Não Linear Binomial Negativo:

$$\ln(u_{it}) = \alpha_i + \beta_1 \cdot X_{1it} + \beta_2 \cdot X_{2it} + \dots + \beta_k \cdot X_{kit} \quad (\text{III.2.5})$$

em que u é o valor esperado da quantidade de ocorrências do fenômeno em estudo (que apresenta distribuição Poisson-Gama) no instante t para dado indivíduo i .

As estimações tradicionais elaboradas nos capítulos anteriores serão novamente utilizadas no Capítulo 15, e tais métodos, de forma análoga aos Modelos Lineares Generalizados, são conhecidos, para os casos em que há dados longitudinais, como **GEE (Generalized Estimating Equations)**. Além disso, em função das características dos dados, também serão estimados parâmetros de modelos que podem levar em consideração a existência de **efeitos fixos** ou de **efeitos aleatórios** nos termos do intercepto, conforme discutiremos ao longo do mesmo capítulo. Logo, para cada um dos modelos propostos, serão estimados parâmetros por meio dos métodos *GEE*, por efeitos fixos ou por efeitos aleatórios. A Figura III.2.3 apresenta essa lógica, a ser utilizada no Capítulo 15.

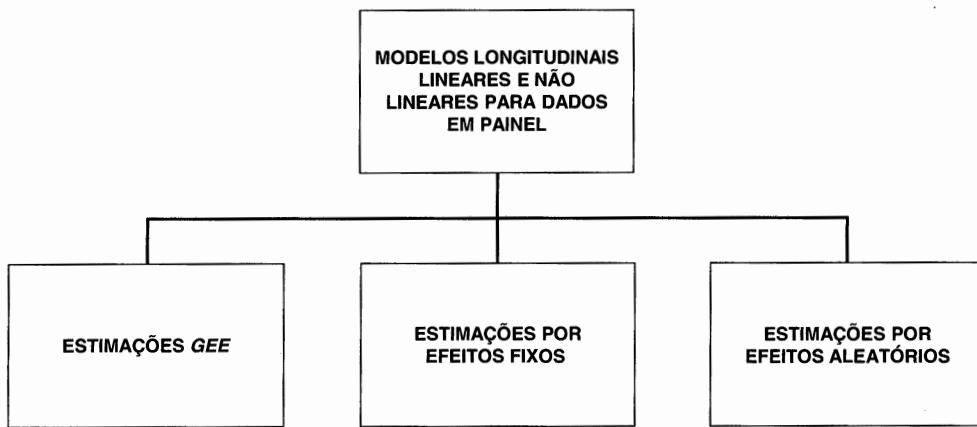


Figura III.2.3 Estimativas de Parâmetros em Modelos Longitudinais de Regressão para Dados em Painel.

Já em relação aos modelos de regressão multinível, podemos especificar cada um dos modelos que serão estudados no Capítulo 16 da seguinte maneira:

Modelo Hierárquico Linear de Dois Níveis (Dados Agrupados):

Nível 1:

$$Y_{ij} = b_{0j} + \sum_{q=1}^Q b_{qj} \cdot X_{qij} + r_{ij} \quad (\text{III.2.6})$$

Nível 2:

$$b_{qj} = \gamma_{q0} + \sum_{s=1}^{S_q} \gamma_{qs} \cdot W_{sj} + u_{qj} \quad (\text{III.2.7})$$

em que os coeficientes b representam os coeficientes do nível 1, X_q ($q = 0, 1, \dots, Q$) é uma q -ésima variável explicativa de nível 1 com dados para os indivíduos $i = 1, \dots, n$ pertencentes aos grupos $j = 1, \dots, J$, os coeficientes γ representam os parâmetros do nível 2, W_s ($s = 1, \dots, S_q$) é uma s -ésima variável explicativa de nível 2 com dados para os grupos (porém invariante em i para determinado grupo j), r_{ij} representa os termos de erro do nível 1 e u_{qj} os termos de erro do nível 2.

Modelo Hierárquico Linear de Três Níveis com Medidas Repetidas:

Nível 1:

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk} \cdot \text{período}_{jk} + e_{ijk} \quad (\text{III.2.8})$$

Nível 2:

$$\pi_{pjk} = b_{p0k} + \sum_{q=1}^{Q_p} b_{pqk} \cdot X_{qjk} + r_{pjk} \quad (\text{III.2.9})$$

Nível 3:

$$b_{pqk} = \gamma_{pq0} + \sum_{s=1}^{S_{pq}} \gamma_{pqs} \cdot W_{sk} + u_{pqk} \quad (\text{III.2.10})$$

em que a variável explicativa *período* do nível 1 representa a medida repetida (variável temporal em que $t = 1, \dots, T$ períodos), os coeficientes π_p ($p = 0$ para intercepto e $p = 1$ para inclinação) representam os parâmetros do nível 1, os coeficientes b representam os parâmetros do nível 2, X_q ($q = 0, 1, \dots, Q_p$) é uma q -ésima variável explicativa de nível 2 com dados para os indivíduos pertencentes aos grupos (porém invariante em t para determinado indivíduo j), os coeficientes γ representam os parâmetros do nível 3, W_s ($s = 1, \dots, S_{pq}$) é uma s -ésima variável explicativa de nível 3 com dados para os grupos (porém invariante em t e em j para determinado grupo k), e_{ijk} representa os termos de erro do nível 1, r_{jk} os termos de erro do nível 2 e u_k os termos de erro do nível 3.

Note, para ambos os casos, que existem variáveis explicativas distintas em cada nível em decorrência de não haver alterações em seus dados em níveis inferiores, o que caracteriza o aninhamento. **Esse fato representa a principal diferença entre os modelos com estruturas aninhadas e os modelos com estruturas longitudinais.**

Também podem ser definidos modelos hierárquicos não lineares caso a variável dependente seja categórica ou apresentar dados de contagem, conforme estudaremos no apêndice do Capítulo 16. Nessas situações, as funções de ligação canônica referentes à variável dependente serão as mesmas daquelas apresentadas no Quadro III.2.1 para os modelos longitudinais.

Os Capítulos 15 e 16 estão estruturados dentro de uma mesma lógica de apresentação em que, inicialmente, são introduzidos os conceitos pertinentes a cada modelo. Dada a complexidade computacional, no Capítulo 15 os parâmetros dos modelos são estimados por meio do uso do software Stata. Entretanto, no Capítulo 16 optamos por elaborar as modelagens multinível em Stata e em SPSS, fato que torna o pesquisador apto a comparar os *outputs* gerados por ambos os softwares, visto que é consideravelmente escassa a literatura que permite esta análise, principalmente com base em modelagens elaboradas em SPSS. Ao término dos capítulos, são propostos exercícios complementares, cujas respostas estão disponibilizadas no final do livro.