

# Modelos de Regressão Simples e Múltipla

*A política serve a um momento no presente, mas uma equação é eterna.*

Albert Einstein

Ao final deste capítulo, você terá condições de:

- Estabelecer as circunstâncias a partir das quais os modelos de regressão simples e múltipla podem ser utilizados.
- Estimar os parâmetros dos modelos de regressão simples e múltipla.
- Avaliar os resultados dos testes estatísticos pertinentes aos modelos de regressão.
- Elaborar intervalos de confiança dos parâmetros do modelo para efeitos de previsão.
- Entender os pressupostos dos modelos de regressão pelo método de mínimos quadrados ordinários.
- Especificar modelos de regressão não lineares e compreender a transformação de Box-Cox.
- Estimar modelos de regressão em Microsoft Office Excel®, Stata Statistical Software® e IBM SPSS Statistics Software® e interpretar seus resultados.

## 12.1. INTRODUÇÃO

Das técnicas estudadas neste livro, sem dúvida nenhuma, aquelas conhecidas por **modelos de regressão simples e múltipla** são as mais utilizadas em diversos campos do conhecimento.

Imagine que um grupo de pesquisadores tenha o interesse em estudar como as taxas de retorno de um ativo financeiro comportam-se em relação ao mercado, ou como o custo de uma empresa varia quando o parque fabril aumenta a sua capacidade produtiva ou incrementa o número de horas trabalhadas, ou, ainda, como o número de dormitórios e a área útil de uma amostra de imóveis residenciais podem influenciar a formação dos preços de venda.

Note, em todos estes exemplos, que os fenômenos principais sobre os quais há o interesse de estudo são representados, em cada caso, por uma **variável métrica**, ou **quantitativa**, e, portanto, podem ser estudados por meio da estimação de modelos de regressão, que têm por finalidade principal analisar como se comportam as relações entre um conjunto de variáveis explicativas, métricas ou **dummies**, e uma variável dependente métrica (**fenômeno em estudo**), desde que respeitadas algumas condições e atendidos alguns pressupostos, conforme veremos ao longo deste capítulo.

É importante enfatizar que todo e qualquer modelo de regressão deve ser definido 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 simples e múltipla, com os seguintes objetivos: (1) introduzir os conceitos sobre regressão simples e múltipla; (2) interpretar os resultados obtidos e elaborar previsões; (3) discutir os pressupostos da técnica; e (4) apresentar a aplicação da técnica em Excel, Stata e SPSS. Inicialmente, será elaborada a solução em Excel de um exemplo concomitantemente à apresentação dos conceitos e à resolução manual deste mesmo exemplo. Somente após a introdução dos conceitos serão apresentados os procedimentos para a elaboração da técnica de regressão no Stata e no SPSS.

## 12.2. MODELOS LINEARES DE REGRESSÃO

Inicialmente, abordaremos os modelos lineares de regressão e seus pressupostos, ficando a análise das regressões não lineares destinada à seção 12.4.

Segundo Fávero *et al.* (2009), a técnica de **regressão linear** oferece, prioritariamente, a possibilidade de que seja estudada a relação entre uma ou mais variáveis explicativas, que se apresentam na forma linear, e uma variável dependente quantitativa. Assim, um modelo geral de regressão linear pode ser definido da seguinte maneira:

$$Y_i = a + b_1 \cdot X_{1i} + b_2 \cdot X_{2i} + \dots + b_k \cdot X_{ki} + u_i \quad (12.1)$$

em que  $Y$  representa o fenômeno em estudo (**variável dependente quantitativa**),  $a$  representa o **intercepto (constante ou coeficiente linear)**,  $b_j$  ( $j = 1, 2, \dots, k$ ) são os coeficientes de cada variável (**coeficientes angulares**),  $X_j$  são as **variáveis explicativas** (métricas ou *dummies*) e  $u$  é o **termo de erro** (diferença entre o valor real de  $Y$  e o valor previsto de  $Y$  por meio do modelo para cada observação). Os subscritos  $i$  representam cada uma das observações da amostra em análise ( $i = 1, 2, \dots, n$ , em que  $n$  é o tamanho da amostra).

A equação apresentada por meio da expressão (12.1) representa um **modelo de regressão múltipla**, uma vez que considera a inclusão de diversas variáveis explicativas para o estudo do comportamento do fenômeno em questão. Por outro lado, caso seja inserida apenas uma variável  $X$ , estaremos diante de um **modelo de regressão simples**. Para efeitos didáticos, introduziremos os conceitos e apresentaremos o passo a passo da estimativa dos parâmetros por meio de um modelo de regressão simples. Na sequência, ampliaremos a discussão por meio da estimativa de modelos de regressão múltipla, inclusive com a consideração de variáveis *dummy* do lado direito da equação.

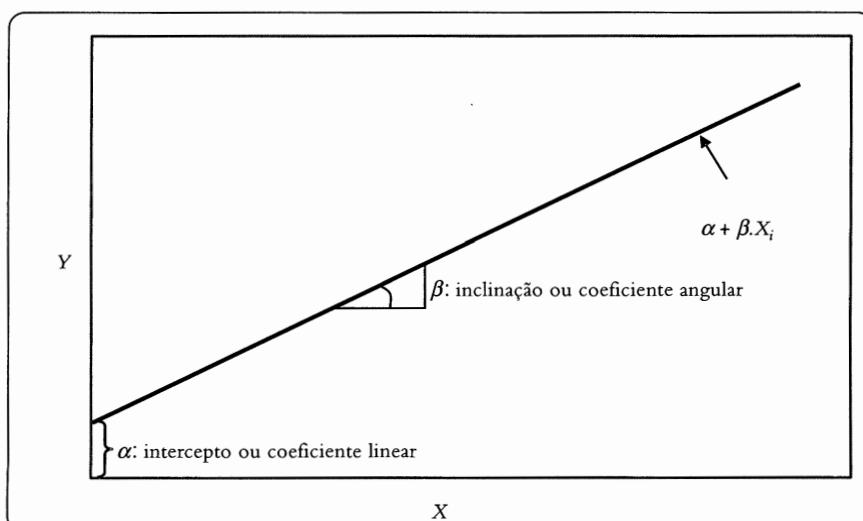
É importante enfatizar, portanto, que o modelo de regressão linear simples a ser estimado apresenta a seguinte expressão:

$$\hat{Y}_i = \alpha + \beta \cdot X_i \quad (12.2)$$

em que  $\hat{Y}_i$  representa o **valor previsto** da variável dependente que será obtido por meio do modelo estimado para cada observação  $i$ , e  $\alpha$  e  $\beta$  representam, respectivamente, os **parâmetros estimados** do intercepto e da inclinação do modelo proposto. A Figura 12.1 apresenta, graficamente, a configuração geral de um modelo estimado de regressão linear simples.

Podemos, portanto, verificar que, enquanto o parâmetro estimado  $\alpha$  mostra o ponto da reta de regressão em que  $X = 0$ , o parâmetro estimado  $\beta$  representa a inclinação da reta, ou seja, o incremento (ou decréscimo) de  $Y$  para cada unidade adicional de  $X$ , em média.

Logo, a inclusão do termo de erro  $u$  na expressão (12.1), também conhecido por **resíduo**, é justificada pelo fato de que qualquer relação que seja proposta dificilmente se apresentará de maneira perfeita. Em outras palavras, muito provavelmente o fenômeno que se deseja estudar, representado pela variável  $Y$ , apresentará relação com



**Figura 12.1** Modelo estimado de regressão linear simples.

alguma outra variável  $X$  não incluída no modelo proposto e que, portanto, precisará ser representada pelo termo de erro  $u$ . Sendo assim, o termo de erro  $u$ , para cada observação  $i$ , pode ser escrito como:

$$u_i = Y_i - \hat{Y}_i \quad (12.3)$$

De acordo com Kennedy (2008), Fávero *et al.* (2009) e Wooldridge (2012), os termos de erro ocorrem em função de algumas razões que precisam ser conhecidas e consideradas pelos pesquisadores, como:

- Existência de variáveis agregadas e/ou não aleatórias.
- Incidência de falhas quanto da especificação do modelo (formas funcionais não lineares e omissão de variáveis explicativas relevantes).
- Ocorrência de erros quando do levantamento dos dados.

Mais considerações sobre os termos de erro serão feitas quando do estudo dos pressupostos dos modelos de regressão, na seção 12.3.

Discutidos estes conceitos preliminares, vamos partir para o estudo propriamente dito da estimação de um modelo de regressão linear.

### 12.2.1. Estimação do modelo de regressão linear por mínimos quadrados ordinários

Frequentemente vislumbramos, de forma racional ou intuitiva, a relação entre comportamentos de variáveis que se apresentam de forma direta ou indireta. Será que se eu frequentar mais as piscinas do meu clube aumentarei a minha massa muscular? Será que se eu mudar de emprego terei mais tempo para ficar com meus filhos? Será que se eu poupar maior parcela de meu salário poderei me aposentar mais jovem? Estas questões oferecem nitidamente relações entre determinada variável dependente, que representa o fenômeno que se deseja estudar, e, no caso, uma única variável explicativa.

O objetivo principal da análise de regressão é, portanto, propiciar ao pesquisador condições de avaliar como se comporta uma variável  $Y$  com base no comportamento de uma ou mais variáveis  $X$ , sem que, necessariamente, ocorra uma relação de causa e efeito.

Introduziremos os conceitos de regressão por meio de um exemplo que considera apenas uma variável explicativa (regressão linear simples). Imagine que, em determinado dia de aula, um professor tenha o interesse em saber, para uma turma de 10 estudantes de uma mesma classe, qual a relação entre a distância percorrida para se chegar à escola e o tempo de percurso. Sendo assim, o professor elaborou um questionamento com cada um dos seus 10 alunos e montou um banco de dados, que se encontra na Tabela 12.1.

Na verdade, o professor deseja saber a equação que regula o fenômeno “tempo de percurso até a escola” em função da “distância percorrida pelos alunos”. É sabido que outras variáveis influenciam o tempo de determinado percurso, como o trajeto adotado, o tipo de transporte ou o horário em que o aluno partiu para a escola naquele dia. Entretanto, o professor tem conhecimento de que tais variáveis não entrarão no modelo, já que nem mesmo as coletou para a formação da base de dados.

**Tabela 12.1** Exemplo: tempo de percurso x distância percorrida.

Estudante	Tempo para chegar à escola (minutos)	Distância percorrida até a escola (quilômetros)
Gabriela	15	8
Dalila	20	6
Gustavo	20	15
Letícia	40	20
Luiz Ovídio	50	25
Leonor	25	11
Ana	10	5
Antônio	55	32
Júlia	35	28
Mariana	30	20

Pode-se, portanto, modelar o problema da seguinte maneira:

$$\text{tempo} = f(\text{dist})$$

Assim sendo, a equação, ou modelo de regressão simples, será:

$$\text{tempo}_i = a + b \cdot \text{dist}_i + u_i$$

e, dessa forma, o valor esperado (estimativa) da variável dependente, para cada observação  $i$ , será dado por:

$$\hat{\text{tempo}}_i = \alpha + \beta \cdot \text{dist}_i$$

em que  $\alpha$  e  $\beta$  são, respectivamente, as estimativas dos parâmetros  $a$  e  $b$ .

Esta última equação mostra que o **valor esperado** da variável  $\text{tempo}$  ( $\hat{Y}$ ), também conhecido por **média condicional**, é calculado para cada observação da amostra, em função do comportamento da variável  $\text{dist}$ , sendo que o subscrito  $i$  representa, para os dados do nosso exemplo, os próprios alunos da escola ( $i = 1, 2, \dots, 10$ ). O nosso objetivo aqui é, portanto, estudar se o comportamento da variável dependente  $\text{tempo}$  apresenta relação com a variação da distância, em quilômetros, a que cada um dos alunos se submete para chegar à escola em determinado dia de aula. No apêndice deste capítulo, faremos uma breve apresentação dos **modelos de regressão quantílica**, cujo objetivo é estimar a **mediana** (e outros percentis) da variável dependente, ao contrário da média, também condicional aos valores das variáveis explicativas.

No nosso exemplo, não faz muito sentido discutirmos qual seria o tempo percorrido no caso de a distância até a escola ser zero (parâmetro  $\alpha$ ). O parâmetro  $\beta$ , por outro lado, nos informará qual é o incremento no tempo para se chegar à escola ao se aumentar a distância percorrida em um quilômetro, em média.

Vamos, desta forma, elaborar um gráfico (Figura 12.2) que relaciona o tempo de percurso ( $Y$ ) com a distância percorrida ( $X$ ), em que cada ponto representa um dos alunos.

Como comentado anteriormente, não é somente a distância percorrida que afeta o tempo para se chegar à escola, uma vez que este pode também ser afetado por outras variáveis relacionadas ao tráfego, ao meio de transporte ou ao próprio indivíduo e, desta maneira, o termo de erro  $u$  deverá capturar o efeito das demais variáveis não incluídas no modelo. Logo, para que estimemos a equação que melhor se ajusta a esta nuvem de pontos, devemos estabelecer duas condições fundamentais relacionadas aos resíduos.

### 1. A somatória dos resíduos deve ser zero: $\sum_{i=1}^n u_i = 0$ , em que $n$ é o tamanho da amostra.

Com apenas esta primeira condição, podem ser encontradas diversas retas de regressão em que a somatória dos resíduos seja zero, como mostra a Figura 12.3.

Nota-se, para o mesmo banco de dados, que diversas retas podem respeitar a condição de que a somatória dos resíduos seja igual a zero. Portanto, faz-se necessário o estabelecimento de uma segunda condição.

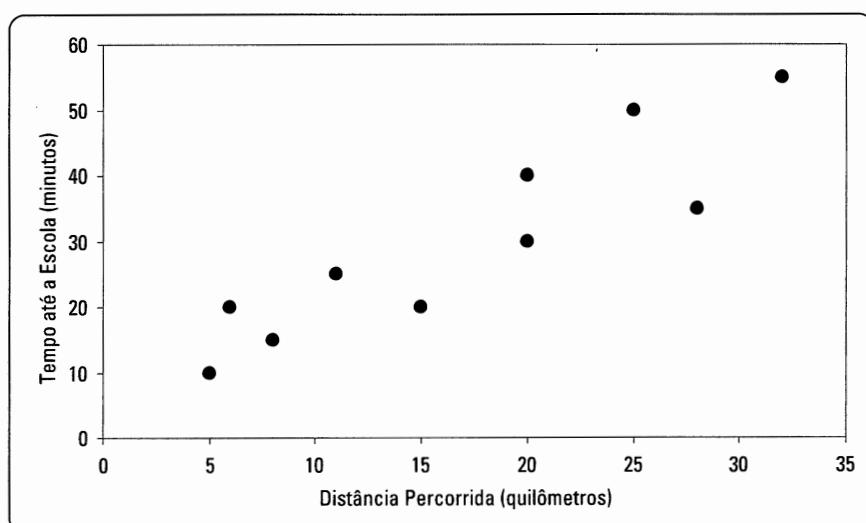
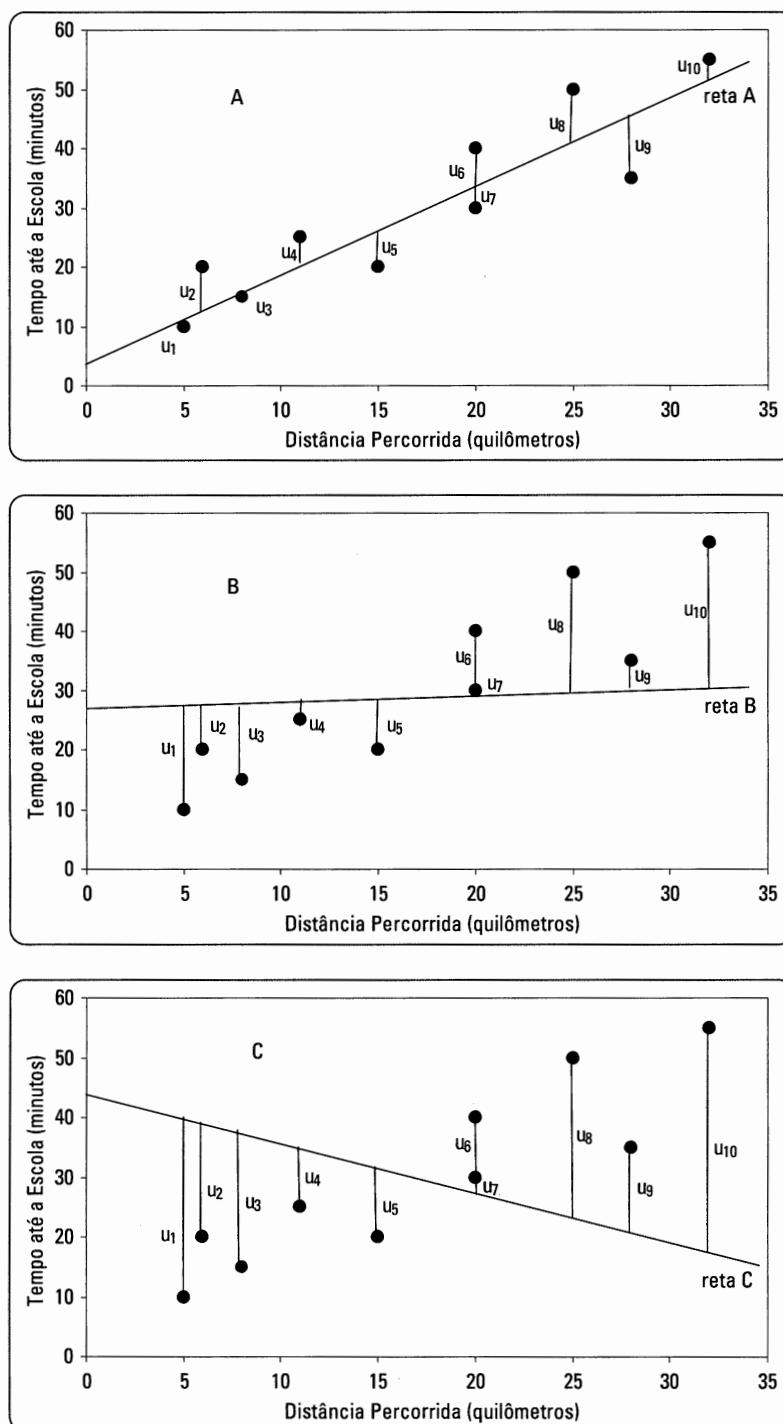


Figura 12.2 Tempo de percurso x distância percorrida para cada aluno.



**Figura 12.3** Exemplos de retas de regressão em que a somatória dos resíduos é zero.

## 2. A somatória dos resíduos ao quadrado é a mínima possível: $\sum_{i=1}^n u_i^2 = \text{mín.}$

Com esta condição, escolhe-se a reta que apresenta o melhor ajuste à nuvem de pontos, partindo-se, portanto, da definição de **mínimos quadrados**, ou seja, deve-se determinar  $\alpha$  e  $\beta$  de modo que a somatória dos quadrados dos resíduos seja a menor possível (**método de Mínimos Quadrados Ordinários - MQO**, ou, em inglês, *Ordinary Least Squares - OLS*). Assim:

$$\sum_{i=1}^n (Y_i - \beta \cdot X_i - \alpha)^2 = \text{mín} \quad (12.4)$$

A minimização ocorre ao se derivar a expressão (12.4) em  $\alpha$  e  $\beta$  e igualar a zero as expressões resultantes. Assim:

$$\frac{\partial \left[ \sum_{i=1}^n (Y_i - \beta \cdot X_i - \alpha)^2 \right]}{\partial \alpha} = -2 \sum_{i=1}^n (Y_i - \beta \cdot X_i - \alpha) = 0 \quad (12.5)$$

$$\frac{\partial \left[ \sum_{i=1}^n (Y_i - \beta \cdot X_i - \alpha)^2 \right]}{\partial \beta} = -2 \sum_{i=1}^n X_i \cdot (Y_i - \beta \cdot X_i - \alpha) = 0 \quad (12.6)$$

Ao se distribuir e dividir a expressão (12.5) por  $2 \cdot n$ , em que  $n$  é o tamanho da amostra, tem-se que:

$$\frac{-2 \sum_{i=1}^n Y_i}{2n} + \frac{2 \sum_{i=1}^n \beta \cdot X_i}{2n} + \frac{2 \sum_{i=1}^n \alpha}{2n} = \frac{0}{2n} \quad (12.7)$$

de onde vem que:

$$-\bar{Y} + \beta \cdot \bar{X} + \alpha = 0 \quad (12.8)$$

e, portanto:

$$\alpha = \bar{Y} - \beta \cdot \bar{X} \quad (12.9)$$

em que  $\bar{Y}$  e  $\bar{X}$  representam, respectivamente, a média amostral de  $Y$  e de  $X$ .

Ao se substituir este resultado na expressão (12.6), tem-se que:

$$-2 \sum_{i=1}^n X_i \cdot (Y_i - \beta \cdot X_i - \bar{Y} + \beta \cdot \bar{X}) = 0 \quad (12.10)$$

que, ao se desenvolver:

$$\sum_{i=1}^n X_i \cdot (Y_i - \bar{Y}) + \beta \sum_{i=1}^n X_i \cdot (\bar{X} - X_i) = 0 \quad (12.11)$$

e que gera, portanto:

$$\beta = \frac{\sum_{i=1}^n (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (12.12)$$

Retornando ao nosso exemplo, o professor então elaborou uma planilha de cálculo a fim de obter a reta de regressão linear, conforme mostra a Tabela 12.2.

**Tabela 12.2** Planilha de cálculo para a determinação de  $\alpha$  e  $\beta$ .

Observação (i)	Tempo ( $Y_i$ )	Distância ( $X_i$ )	$Y_i - \bar{Y}$	$X_i - \bar{X}$	$(X_i - \bar{X}).(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$
1	15	8	-15	-9	135	81
2	20	6	-10	-11	110	121
3	20	15	-10	-2	20	4
4	40	20	10	3	30	9
5	50	25	20	8	160	64
6	25	11	-5	-6	30	36
7	10	5	-20	-12	240	144
8	55	32	25	15	375	225
9	35	28	5	11	55	121
10	30	20	0	3	0	9
<b>Soma</b>	<b>300</b>	<b>170</b>			<b>1155</b>	<b>814</b>
<b>Média</b>	<b>30</b>	<b>17</b>				

Por meio da planilha apresentada na Tabela 12.2 podemos calcular os estimadores  $\alpha$  e  $\beta$ , de acordo como segue:

$$\beta = \frac{\sum_{i=1}^{10} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{10} (X_i - \bar{X})^2} = \frac{1155}{814} = 1,4189$$

$$\alpha = \bar{Y} - \beta \cdot \bar{X} = 30 - 1,4189 \cdot 17 = 5,8784$$

E a equação de regressão linear simples pode ser escrita como:

$$\hat{tempo}_i = 5,8784 + 1,4189 \cdot dist_i$$

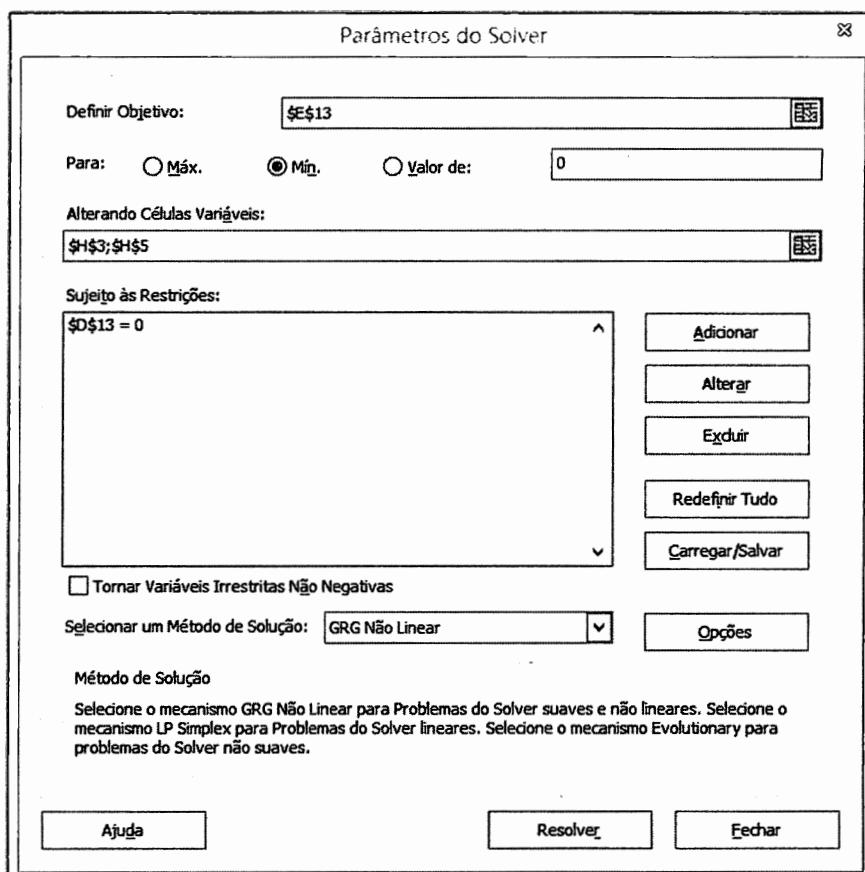
A estimação dos parâmetros do modelo do nosso exemplo também pode ser efetuada por meio da ferramenta **Solver** do Excel, respeitando-se as condições de que  $\sum_{i=1}^{10} u_i = 0$  e  $\sum_{i=1}^{10} u_i^2 = \text{mín}$ . Desta forma, vamos inicialmente abrir o arquivo **TempoMínimosQuadrados.xls** que contém os dados do nosso exemplo, além das colunas referentes ao  $\hat{Y}$ , ao  $u$  e ao  $u^2$  de cada observação. A Figura 12.4 apresenta este arquivo, antes da elaboração do procedimento **Solver**.

Seguindo a lógica proposta por Belfiore e Fávero (2012), vamos então abrir a ferramenta **Solver** do Excel. A função-objetivo está na célula E13, que é a nossa célula de destino e que deverá ser minimizada (somatória dos

	A	B	C	D	E	F	G	H
1	Tempo (Y)	Distância ( $X_i$ )	$\hat{Y}_i$	$u_i$	$u_i^2$			
2	15	8	0	15,00000	225,00000			
3	20	6	0	20,00000	400,00000			
4	20	15	0	20,00000	400,00000			
5	40	20	0	40,00000	1600,00000			
6	50	25	0	50,00000	2500,00000			
7	25	11	0	25,00000	625,00000			
8	10	5	0	10,00000	100,00000			
9	55	32	0	55,00000	3025,00000			
10	35	28	0	35,00000	1225,00000			
11	30	20	0	30,00000	900,00000			
12								
13				Somatória	300,00000	11000,00000		

$\alpha$    
 $\beta$

**Figura 12.4** Dados do arquivo **TempoMínimosQuadrados.xls**.



**Figura 12.5** Solver – Minimização da somatória dos resíduos ao quadrado.

quadrados dos resíduos). Além disso, os parâmetros  $\alpha$  e  $\beta$ , cujos valores estão nas células H3 e H5, respectivamente, são as células variáveis. Por fim, devemos impor que o valor da célula D13 seja igual a zero (restrição de que a soma dos resíduos seja igual a zero). A janela do **Solver** ficará como mostra a Figura 12.5.

Ao clicarmos em **Resolver** e em **OK**, obteremos a solução ótima do problema de minimização dos resíduos ao quadrado. A Figura 12.6 apresenta os resultados obtidos pela modelagem.

Logo, o intercepto  $\alpha$  é 5,8784 e o coeficiente angular  $\beta$  é 1,4189, conforme havíamos estimado por meio da solução analítica. De forma elementar, o tempo médio para se chegar à escola por parte dos alunos que não percorrem distância alguma, ou seja, que já se encontram na escola, é de 5,8784 minutos, o que não faz muito sentido do ponto de vista físico. Em alguns casos, este tipo de situação pode ocorrer com frequência, em que valores de  $\alpha$  não são condizentes com a realidade. Do ponto de vista matemático, isto não está errado, porém o pesquisador deve sempre analisar o sentido físico ou econômico da situação em estudo, bem como a teoria subjacente utilizada. Ao analisarmos o gráfico da Figura 12.2 iremos perceber que não há nenhum estudante com distância percorrida próxima de zero, e o intercepto reflete apenas o prolongamento, projeção ou extrapolação da reta de regressão até o eixo Y. É comum, inclusive, que alguns modelos apresentem  $\alpha$  negativo quando do estudo de fenômenos que não podem oferecer valores negativos. O pesquisador deve, portanto, ficar sempre atento a este fato, já que um modelo de regressão pode ser bastante útil para que sejam elaboradas inferências sobre o comportamento de uma variável Y dentro dos limites de variação de X, ou seja, para a elaboração de **interpolações**. Já as **extrapolações** podem oferecer inconsistências por eventuais mudanças de comportamento da variável Y fora dos limites de variação de X na amostra em estudo.

Dando sequência à análise, cada quilômetro adicional de distância entre o local de partida de cada aluno e a escola incrementa o tempo de percurso em 1,4189 minutos, em média. Assim, um estudante que mora 10 quilômetros mais longe da escola do que outro tenderá a gastar, em média, pouco mais de 14 minutos ( $1,4189 \times 10$ ) a mais para chegar à escola do que seu colega que mora mais perto. A Figura 12.7 apresenta a reta de regressão linear simples do nosso exemplo.

	A	B	C	D	E	F	G	H
1	Tempo (Y)	Distância (X <sub>i</sub> )	$\hat{Y}_i$	$u_i$	$u_i^2$			
2	15	8	17	-2,22973	4,97169			
3	20	6	14	5,60811	31,45088			
4	20	15	27	-7,16216	51,29657			
5	40	20	34	5,74324	32,98484			
6	50	25	41	8,64865	74,79912			
7	25	11	21	3,51351	12,34478			
8	10	5	13	-2,97297	8,83857			
9	55	32	51	3,71622	13,81026			
10	35	28	46	-10,60811	112,53196			
11	30	20	34	-4,25676	18,11998			
12								
13			Somatória	0,00000	361,14865			

**Figura 12.6** Obtenção dos parâmetros quando da minimização da somatória de  $u^2$  pelo Solver.

Concomitantemente à discussão de cada um dos conceitos e à resolução do exemplo proposto de forma analítica e pelo **Solver**, iremos também apresentar a solução por meio da ferramenta **Regressão** do Excel, passo a passo. Nas seções 12.5 e 12.6 partiremos para a solução final por meio dos softwares Stata e SPSS, respectivamente. Desta maneira, vamos agora abrir o arquivo **Tempodist.xls** que contém os dados do nosso exemplo, ou seja, dados fictícios de tempo de percurso e distância percorrida por um grupo de 10 alunos até o local da escola.

Ao clicarmos em **Dados → Análise de Dados**, aparecerá a caixa de diálogo da Figura 12.8.

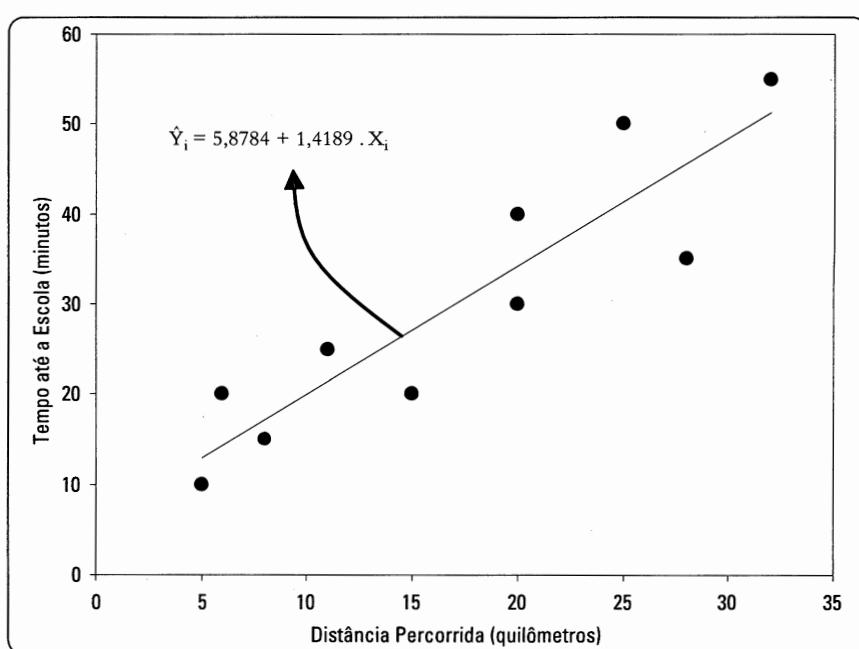
Vamos clicar em **Regressão** e, em seguida, em **OK**. A caixa de diálogo para inserção dos dados a serem considerados na regressão aparecerá na sequência (Figura 12.9).

Para o nosso exemplo, a variável *tempo* (min) é a dependente (Y) e a variável *dist* (km) é a explicativa (X). Portanto, devemos inserir seus dados nos respectivos intervalos de entrada, conforme mostra a Figura 12.10.

Além da inserção dos dados, vamos também marcar a opção **Resíduos**, conforme mostra a Figura 12.10. Na sequência, vamos clicar em **OK**. Uma nova planilha será gerada, com os *outputs* da regressão. Iremos analisar cada um deles à medida que formos introduzindo os conceitos e elaborando também os cálculos manualmente.

Conforme podemos observar por meio da Figura 12.11, 4 grupos de *outputs* são gerados: estatísticas da regressão, tabela de análise de variância (*analysis of variance*, ou ANOVA), tabela de coeficientes da regressão e tabela de resíduos. Iremos discutir cada um deles.

Como calculado anteriormente, podemos verificar os coeficientes da equação de regressão nos *outputs* (Figura 12.12).



**Figura 12.7** Reta de regressão linear simples entre tempo e distância percorrida.

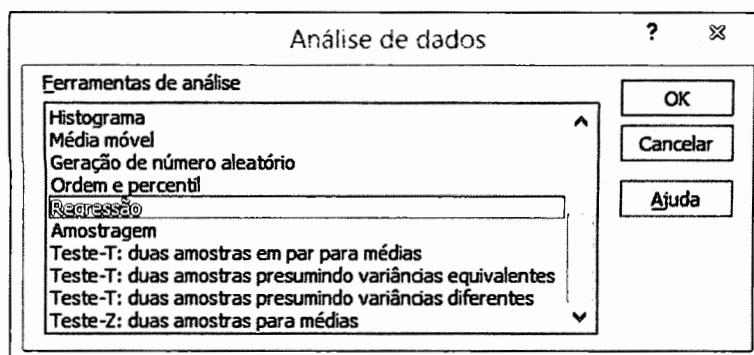


Figura 12.8 Caixa de diálogo para análise de dados no Excel.

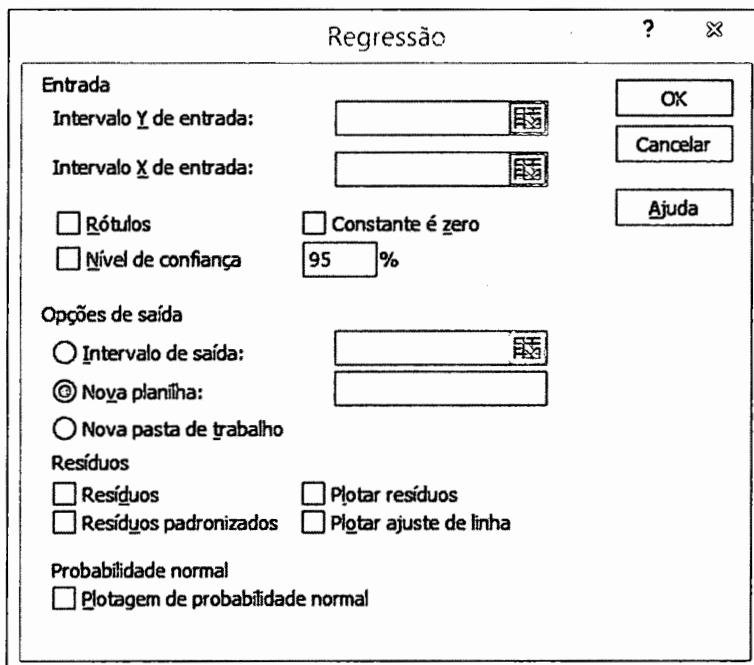


Figura 12.9 Caixa de diálogo para elaboração de regressão linear no Excel.

	A	B	C	D	E	F	G
	Tempo (min) (Y)	Distância (km) (X <sub>1</sub> )					
1	15	8					
2	20	6					
3	20	15					
4	40	20					
5	50	25					
6	25	11					
7	10	5					
8	55	32					
9	35	28					
10	30	20					

Figura 12.10 Inserção dos dados para elaboração de regressão linear no Excel.

	A	B	C	D	E	F	G	H	I
1	RESUMO DOS RESULTADOS								
2									
3	<i>Estatística de regressão</i>								
4	R múltiplo	0,90522134							
5	R-Quadrado	0,81942568							
6	R-quadrado ajustado	0,79685389							
7	Erro padrão	6,71889731							
8	Observações	10							
9									
10	ANOVA								
11		gl	SQ	MQ	F	F de significação			
12	Regressão	1	1638,851351	1638,85135	36,303087	0,000314449			
13	Resíduo	8	361,1486486	45,1435811					
14	Total	9	2000						
15									
16		Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores	Inferior 95,0%	Superior 95,0%
17	Interseção	5,87837838	4,532327565	1,29698886	0,23078848	-4,573187721	16,32994448	-4,573187721	16,32994448
18	Variável X 1	1,41891892	0,235497229	6,02520431	0,00031445	0,875861336	1,961976502	0,875861336	1,961976502
19									
20									
21									
22	RESULTADOS DE RESÍDUOS								
23									
24	Observação	Y previsto	Resíduos						
25	1	17,2297297	-2,22972973						
26	2	14,3918919	5,608108108						
27	3	27,1621622	-7,16216216						
28	4	34,2567568	5,743243243						
29	5	41,3513514	8,648648649						
30	6	21,4864865	3,513513514						
31	7	12,972973	-2,97297297						
32	8	51,2837838	3,716216216						
33	9	45,6081081	-10,6081081						
34	10	34,2567568	-4,25675676						

**Figura 12.11** Outputs da regressão linear simples no Excel.

	A	B	C	D	E	F	G	H	I
1	RESUMO DOS RESULTADOS								
2									
3	<i>Estatística de regressão</i>								
4	R múltiplo	0,90522134							
5	R-Quadrado	0,81942568							
6	R-quadrado ajustado	0,79685389							
7	Erro padrão	6,71889731							
8	Observações	10							
9									
10	ANOVA								
11		gl	SQ	MQ	F	F de significação			
12	Regressão	1	1638,851351	1638,85135	36,303087	0,000314449			
13	Resíduo	8	361,1486486	45,1435811					
14	Total	9	2000						
15									
16		Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores	Inferior 95,0%	Superior 95,0%
17	Interseção	5,87837838	4,532327565	1,29698886	0,23078848	-4,573187721	16,32994448	-4,573187721	16,32994448
18	Variável X 1	1,41891892	0,235497229	6,02520431	0,00031445	0,875861336	1,961976502	0,875861336	1,961976502
19									
20									
21									
22	RESULTADOS DE RESÍDUOS								
23									
24	Observação	Y previsto	Resíduos						
25	1	17,2297297	-2,22972973						
26	2	14,3918919	5,608108108						
27	3	27,1621622	-7,16216216						
28	4	34,2567568	5,743243243						
29	5	41,3513514	8,648648649						
30	6	21,4864865	3,513513514						
31	7	12,972973	-2,97297297						
32	8	51,2837838	3,716216216						
33	9	45,6081081	-10,6081081						
34	10	34,2567568	-4,25675676						

Equação de Regressão Linear

$$\text{tempo}_i = 5,8784 + 1,4189 \cdot \text{dist}_i$$

**Figura 12.12** Coeficientes da equação de regressão linear.

### 12.2.2. Poder explicativo do modelo de regressão: $R^2$

Segundo Fávero *et al.* (2009), para mensurarmos o poder explicativo de determinado modelo de regressão, ou o percentual de variabilidade da variável  $Y$  que é explicado pelo comportamento de variação das variáveis explicativas, precisamos entender alguns importantes conceitos. Enquanto a **soma total dos quadrados (SQT)** mostra a variação em  $Y$  em torno da própria média, a **soma dos quadrados da regressão (SQR)** oferece a variação de  $Y$  considerando as variáveis  $X$  utilizadas no modelo. Além disso, a **soma dos quadrados dos resíduos (SQU)** apresenta a variação de  $Y$  que não é explicada pelo modelo elaborado. Logo, podemos definir que:

$$SQT = SQR + SQU \quad (12.13)$$

sendo:

$$Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) \quad (12.14)$$

em que  $Y_i$  equivale ao valor de  $Y$  de cada observação  $i$  da amostra,  $\bar{Y}$  é a média de  $Y$  e  $\hat{Y}_i$  representa o valor ajustado da reta da regressão para cada observação  $i$ . Assim, temos que:

$Y_i - \bar{Y}$ : desvio total dos valores de cada observação em relação à média,

$(\hat{Y}_i - \bar{Y})$ : desvio dos valores da reta de regressão para cada observação em relação à média,

$(Y_i - \hat{Y}_i)$ : desvio dos valores de cada observação em relação à reta de regressão,

que resulta em:

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (12.15)$$

ou:

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (u_i)^2 \quad (12.16)$$

que é a própria expressão (12.13).

A Figura 12.13 mostra graficamente esta relação.

Feitas estas considerações e definida a equação de regressão, partiremos para o estudo do poder explicativo do modelo de regressão, também conhecido por **coeficiente de ajuste  $R^2$** . Stock e Watson (2004) definem o  $R^2$  como a fração da variância da amostra de  $Y_i$  explicada (ou prevista) pelas variáveis explicativas. Da mesma

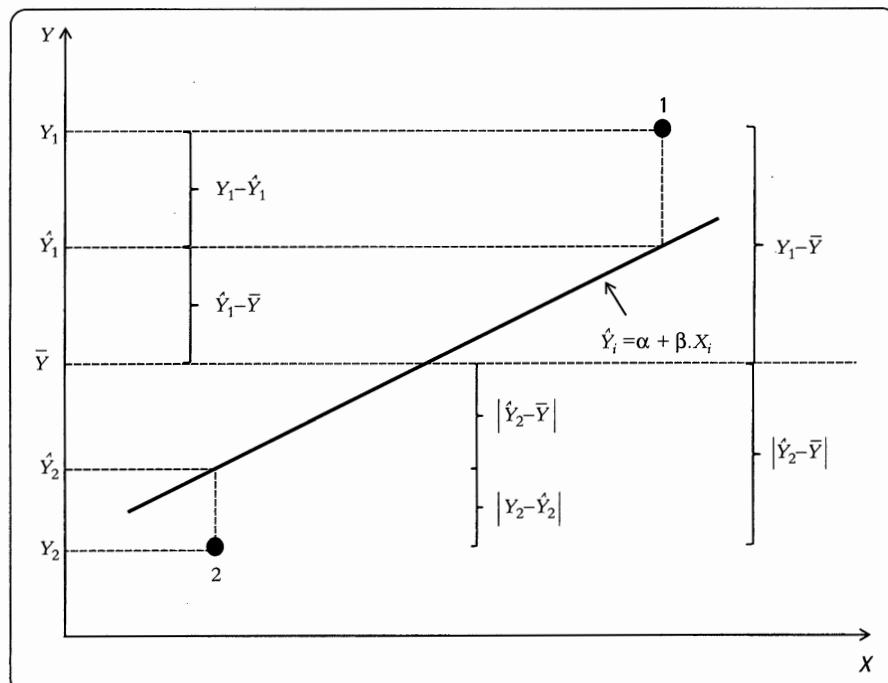


Figura 12.13 Desvios de  $Y$  para duas observações.

forma, Wooldridge (2012) considera o  $R^2$  como a proporção da variação amostral da variável dependente explicada pelo conjunto de variáveis explicativas, podendo ser utilizado como uma medida do grau de ajuste do modelo proposto.

Segundo Fávero *et al.* (2009), a capacidade explicativa do modelo é analisada pelo  $R^2$  da regressão, conhecido também por **coeficiente de ajuste** ou **de explicação**. Para um modelo de regressão simples, esta medida mostra quanto do comportamento da variável  $Y$  é explicado pelo comportamento de variação da variável  $X$ , sempre lembrando que não existe, necessariamente, uma relação de causa e efeito entre as variáveis  $X$  e  $Y$ . Para um modelo de regressão múltipla, esta medida mostra quanto do comportamento da variável  $Y$  é explicado pela variação conjunta das variáveis  $X$  consideradas no modelo.

O  $R^2$  é obtido da seguinte forma:

$$R^2 = \frac{SQR}{SQR + SQU} = \frac{SQR}{SQT} \quad (12.17)$$

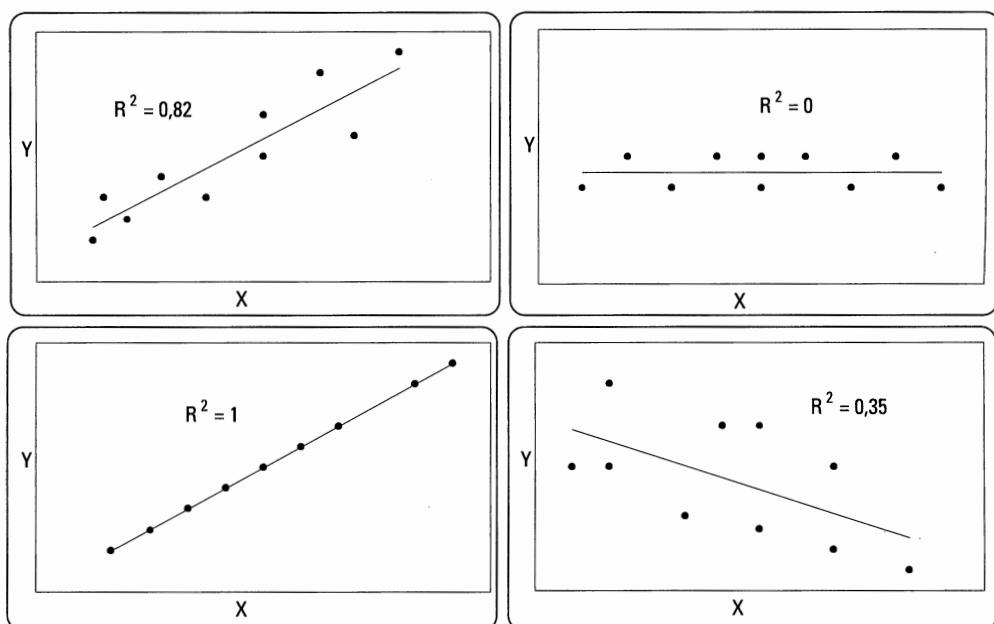
ou

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (u_i)^2} \quad (12.18)$$

Ainda de acordo com Fávero *et al.* (2009), o  $R^2$  pode variar entre 0 e 1 (0% a 100%), porém é praticamente impossível a obtenção de um  $R^2$  igual a 1, uma vez que dificilmente todos os pontos situar-se-ão em cima de uma reta. Em outras palavras, se o  $R^2$  for 1, não haverá resíduos para cada uma das observações da amostra em estudo, e a variabilidade da variável  $Y$  estará sendo totalmente explicada pelo vetor de variáveis  $X$  consideradas no modelo de regressão. É importante enfatizar que, em diversos campos do conhecimento humano, como em ciências sociais aplicadas, este fato é realmente muito pouco provável de acontecer.

Quanto mais dispersa for a nuvem de pontos, menos as variáveis  $X$  e  $Y$  se relacionarão, maiores serão os resíduos e mais próximo de zero será o  $R^2$ . Em um caso extremo, se a variação de  $X$  não corresponder a nenhuma variação em  $Y$ , o  $R^2$  será zero. A Figura 12.14 apresenta, de forma ilustrativa, o comportamento do  $R^2$  para diferentes casos.

Voltando ao nosso exemplo em que o professor tem intenção de estudar o comportamento do tempo que os alunos levam para chegar à escola e se este fenômeno é influenciado pela distância percorrida pelos estudantes, apresentamos uma planilha (Tabela 12.3) que nos auxiliará no cálculo do  $R^2$ .



**Figura 12.14** Comportamento do  $R^2$  para diferentes regressões lineares simples.

**Tabela 12.3** Planilha para o cálculo do coeficiente de ajuste do modelo de regressão R<sup>2</sup>.

Observação (i)	Tempo (Y <sub>i</sub> )	Distância (X <sub>i</sub> )	$\hat{Y}_i$	$u_i$ (Y <sub>i</sub> - $\hat{Y}_i$ )	$(\hat{Y}_i - \bar{Y})^2$	$(u_i)^2$
1	15	8	17,23	-2,23	163,08	4,97
2	20	6	14,39	5,61	243,61	31,45
3	20	15	27,16	-7,16	8,05	51,30
4	40	20	34,26	5,74	18,12	32,98
5	50	25	41,35	8,65	128,85	74,80
6	25	11	21,49	3,51	72,48	12,34
7	10	5	12,97	-2,97	289,92	8,84
8	55	32	51,28	3,72	453,00	13,81
9	35	28	45,61	-10,61	243,61	112,53
10	30	20	34,26	-4,26	18,12	18,12
<b>Soma</b>	<b>300</b>	<b>170</b>			<b>1638,85</b>	<b>361,15</b>
<b>Média</b>	<b>30</b>	<b>17</b>				

Obs.: Em que  $\hat{Y}_i = \text{tempo}_i = 5,8784 + 1,4189 \cdot \text{dist}_i$ .

Esta planilha permite que calculemos o R<sup>2</sup> do modelo de regressão linear simples do nosso exemplo. Assim:

$$R^2 = \frac{\sum_{i=1}^{10} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{10} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{10} (u_i)^2} = \frac{1638,85}{1638,85 + 361,15} = 0,8194$$

Dessa forma, podemos agora afirmar que, para a mostra estudada, 81,94% da variabilidade do tempo para se chegar à escola é devido à variável referente à distância percorrida durante o percurso elaborado por cada um dos alunos. E, portanto, pouco mais de 18% desta variabilidade é devido a outras variáveis não incluídas no modelo e que, portanto, foram decorrentes da variação dos resíduos.

Os *outputs* gerados no Excel também trazem esta informação, conforme pode ser observado na Figura 12.15.

Note que estes *outputs* também fornecem os valores de  $\hat{Y}$  e dos resíduos para cada observação, bem como o valor mínimo da somatória dos resíduos ao quadrado, que são exatamente iguais aos obtidos quando da estimativa dos parâmetros por meio da ferramenta **Solver** do Excel (Figura 12.6) e também calculados e apresentados na Tabela 12.3. Por meio desses valores, temos condições de calcular o R<sup>2</sup>.

Segundo Stock e Watson (2004) e Fávero *et al.* (2009), o **coeficiente de ajuste R<sup>2</sup> não diz** aos pesquisadores se determinada variável explicativa é estatisticamente significante e se esta variável é a causa verdadeira da alteração de comportamento da variável dependente. Mais do que isso, o R<sup>2</sup> também não oferece condições de se avaliar a existência de um eventual viés de omissão de variáveis explicativas e se a escolha daquelas que foram inseridas no modelo proposto foi adequada.

A importância dada à dimensão do R<sup>2</sup> é frequentemente demasiada e, em diversas situações, os pesquisadores destacam a adequabilidade de seus modelos pela obtenção de altos valores de R<sup>2</sup>, dando ênfase inclusive à relação de causa e efeito entre as variáveis explicativas e a variável dependente, mesmo que isso seja bastante equivocado, uma vez que esta medida apenas captura a relação entre as variáveis utilizadas no modelo. Wooldridge (2012) é ainda mais enfático, destacando que é fundamental não dar importância considerável ao valor do R<sup>2</sup> na avaliação de modelos de regressão.

Segundo Fávero *et al.* (2009), se conseguirmos, por exemplo, encontrar uma variável que explique 40% do retorno das ações, num primeiro momento pode parecer uma capacidade explicativa baixa. Porém, se uma única variável conseguir capturar toda esta relação numa situação de existência de inúmeros outros fatores econômicos, financeiros, perceptuais e sociais, o modelo poderá ser bastante satisfatório.

A significância estatística geral do modelo e de seus parâmetros estimados não é dada pelo R<sup>2</sup>, mas por meio de testes estatísticos apropriados que passaremos a estudar na próxima seção.

	A	B	C	D	E	F	G	H	I
1	RESUMO DOS RESULTADOS								
2									
3	Estatística de regressão								
4	R múltiplo	0,90522144							
5	R-Quadrado	0,81942568							
6	R-quadrado ajustado	0,79065389							
7	Erro padrão	6,71889731							
8	Observações	10							
9									
10	ANOVA								
11	g/	SQ	MQ	F	F de significação				
12	Regressão	1 1638,851351	1638,85135	36,303087	0,000314449				
13	Resíduo	361,1486486	45,1435811						
14	Total	9 2000							
15									
16	Coefficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores	Inferior 95,0%	Superior 95,0%	
17	Interseção	5,87837838	4,532327565	1,29698886	0,23078848	-4,573187721	16,32994448	-4,573187721	16,32994448
18	Variável X 1	1,41891892	0,235497229	6,025020431	0,00031445	0,875861336	1,961976502	0,875861336	1,961976502
19									
20									
21									
22	RESULTADOS DE RESÍDUOS								
23									
24	Observação	Y previsto	Resíduos						
25	1	17,2297297	-2,22972973						
26	2	14,3918919	5,608108108						
27	3	27,1621622	-7,16216216						
28	4	34,2567568	5,743243243						
29	5	41,3513514	8,648648649						
30	6	21,4864865	3,513513514						
31	7	12,972973	-2,97297297						
32	8	51,2837838	3,716216216						
33	9	45,6081081	-10,6081081						
34	10	34,2567568	-4,25675676						

Figura 12.15 Coeficiente de ajuste da regressão.

### 12.2.3. A significância geral do modelo e de cada um dos parâmetros

Inicialmente, é de fundamental importância que estudemos a significância estatística geral do modelo estimado. Com tal finalidade, devemos fazer uso do **teste F**, cujas hipóteses nula e alternativa, para um modelo geral de regressão, são, respectivamente:

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

E, para um modelo de regressão simples, portanto, estas hipóteses passam a ser:

$$\begin{aligned} H_0: \beta = 0 \\ H_1: \beta \neq 0 \end{aligned}$$

Este teste possibilita ao pesquisador verificar se o modelo que está sendo estimado de fato existe, uma vez que, se todos os  $\beta_j$  ( $j = 1, 2, \dots, k$ ) forem estatisticamente iguais a zero, o comportamento de alteração de cada uma das variáveis explicativas não influenciará em absolutamente nada o comportamento de variação da variável dependente. A **estatística F** apresenta a seguinte expressão:

$$F = \frac{\frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{(k-1)}}{\frac{\sum_{i=1}^n (u_i)^2}{(n-k)}} = \frac{\frac{SQR}{(k-1)}}{\frac{SQU}{(n-k)}} \quad (12.19)$$

em que  $k$  representa o número de parâmetros do modelo estimado (inclusive o intercepto) e  $n$ , o tamanho da amostra.

Podemos, portanto, obter a expressão da estatística  $F$  com base na expressão do  $R^2$  apresentada em (12.17). Sendo assim, temos que:

$$F = \frac{\frac{SQR}{(k-1)}}{\frac{SQU}{(n-k)}} = \frac{\frac{R^2}{(k-1)}}{\frac{(1-R^2)}{(n-k)}} \quad (12.20)$$

Logo, voltando ao nosso exemplo inicial, obtemos:

$$F = \frac{\frac{1638,85}{(2-1)}}{\frac{361,15}{(10-2)}} = 36,30$$

que, para 1 grau de liberdade da regressão ( $k - 1 = 1$ ) e 8 graus de liberdade para os resíduos ( $n - k = 10 - 2 = 8$ ), temos, por meio da Tabela A do apêndice do livro, que o  $F_c = 5,32$  ( $F$  crítico ao nível de significância de 5%). Desta forma, como o  $F$  calculado  $F_{cal} = 36,30 > F_c = F_{1,8,5\%} = 5,32$ , podemos rejeitar a hipótese nula de que todos os parâmetros  $\beta_j$  ( $j = 1$ ) sejam estatisticamente iguais a zero. Logo, pelo menos uma variável  $X$  é estatisticamente significante para explicar a variabilidade de  $Y$  e teremos um modelo de regressão estatisticamente significante para fins de previsão. Como, neste caso, temos apenas uma única variável  $X$  (regressão simples), esta será estatisticamente significante, ao nível de significância de 5%, para explicar o comportamento de variação de  $Y$ .

Os *outputs* oferecem, por meio da análise de variância (ANOVA), a estatística  $F$ , conforme estudado no Capítulo 7, e o seu correspondente nível de significância (Figura 12.16).

Softwares como o Stata e o SPSS não oferecem diretamente o  $F_c$  para os graus de liberdade definidos e determinado nível de significância. Todavia, oferecem o nível de significância do  $F_{cal}$  para estes graus de liberdade. Desta forma, em vez de analisarmos se  $F_{cal} > F_c$ , devemos verificar se o nível de significância do  $F_{cal}$  é menor do que 0,05 (5%) a fim de darmos continuidade à análise de regressão. O Excel chama este nível de significância de  $F$  de *significação*. Assim:

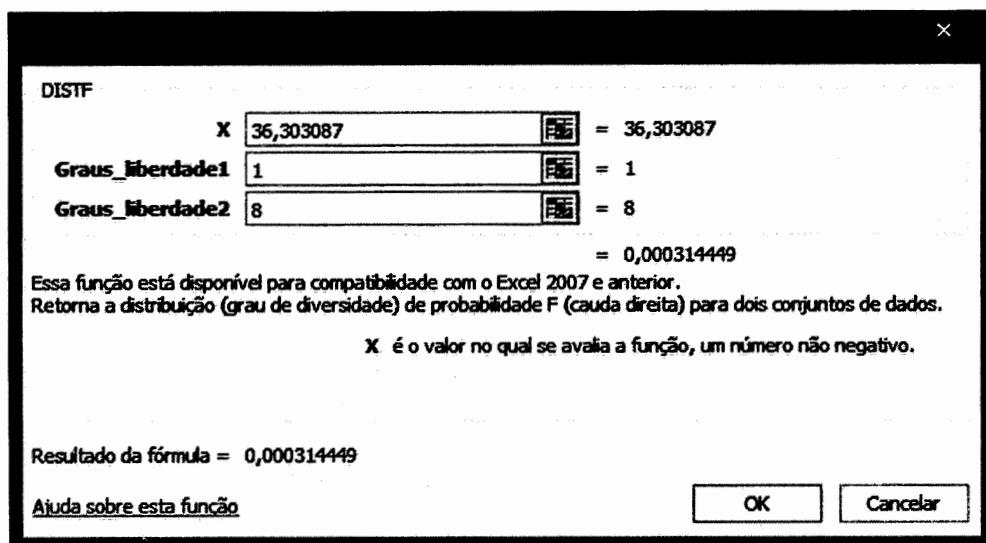
Se  $F$  de *significação*  $< 0,05$ , existe pelo menos um  $\beta_j \neq 0$ .

	A	B	C	D	E	F	G	H	I
1	RESUMO DOS RESULTADOS								
2									
3	<i>Estatística de regressão</i>					$\sum_{i=1}^n (Y_i - \bar{Y})^2$	SQR	1638,85	
4	R múltiplo	0,90522134				$(k-1)$	$(k-1)$		
5	R-Quadrado	0,81942568					$(2-1)$		
6	R-quadrado ajustado	0,79685389					361,15	36,30	
7	Erro padrão	6,71889731				$\sum_{i=1}^n (u_i)^2$	SQU		
8	Observações	10				$(n-k)$	$(n-k)$	$(10-2)$	
9						$(n-k)$			
10	ANOVA								
11		gl	SQ	MQ	F	<i>F</i> de <i>significação</i>			
12	Regressão	1	1638,851351	1638,85135	36,303087	0,000314449			Nível de significância do $F_{cal} < 0,05$
13	Resíduo	8	361,1486486	45,1435811					
14	Total	9	2000						
15									
16		Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores	Inferior 95,0%	Superior 95,0%
17	Interseção	5,87837838	4,532327565	1,29698886	0,23078848	-4,573187721	16,32994448	-4,573187721	16,32994448
18	Variável X 1	1,41891892	0,235497229	6,02520431	0,00031445	0,875861336	1,961976502	0,875861336	1,961976502
19									
20									
21									
22	RESULTADOS DE RESÍDUOS								
23									
24	Observação	Y previsto	Resíduos						
25	1	17,2297297	-2,22972973						
26	2	14,3918919	5,608108108						
27	3	27,1621622	-7,16216216						
28	4	34,2567568	5,743243243						
29	5	41,3513514	8,648648649						
30	6	21,4864865	3,513513514						
31	7	12,972973	-2,97297297						
32	8	51,2837838	3,716216216						
33	9	45,6081081	-10,6081081						
34	10	34,2567568	-4,25675676						

Figura 12.16 Output da ANOVA – Teste  $F$  para avaliação conjunta de significância dos parâmetros.

O nível de significância do  $F_{cal}$  pode ser obtido no Excel por meio do comando **Fórmulas → Inserir Função → DISTF**, que abrirá uma caixa de diálogo conforme mostra a Figura 12.17.

Muitos modelos apresentam mais de uma variável explicativa  $X$  (regressões múltiplas) e, como o teste  $F$  avalia a significância conjunta das variáveis explicativas, acaba por não se definir qual ou quais destas variáveis consideradas no modelo apresentam parâmetros estimados estatisticamente diferentes de zero, a determinado nível de significância. Desta maneira, é preciso que o pesquisador avalie se cada um dos parâmetros do modelo de regressão é estatisticamente diferente de zero, a fim de que a sua respectiva variável  $X$  seja, de fato, incluída no modelo final proposto.



**Figura 12.17** Obtenção do nível de significância de  $F$  (comando **Inserir Função**).

A **estatística  $t$** , também estudada no Capítulo 7, é importante para fornecer ao pesquisador a significância estatística de cada parâmetro a ser considerado no modelo de regressão, e as hipóteses do teste correspondente (**teste  $t$** ) para o intercepto e para cada  $\beta_j$  ( $j = 1, 2, \dots, k$ ) são, respectivamente:

$$H_0: \alpha = 0$$

$$H_1: \alpha \neq 0$$

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

Este teste propicia ao pesquisador uma verificação sobre a significância estatística de cada parâmetro estimado,  $\alpha$  e  $\beta_j$ , e sua expressão é dada por:

$$t_\alpha = \frac{\alpha}{s.e.(\alpha)} \quad (12.21)$$

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

em que  $s.e.$  corresponde ao **erro-padrão (standard error)** de cada parâmetro em análise e será discutido adiante. Após a obtenção das estatísticas  $t$ , o pesquisador pode utilizar as respectivas tabelas de distribuição para obtenção dos valores críticos a um dado nível de significância e verificar se tais testes rejeitam ou não a hipótese nula. Entretanto, como no caso do teste  $F$ , os pacotes estatísticos também oferecem os valores dos níveis de significância dos testes  $t$ , chamados de *valor-P* (ou *P-value*), o que facilita a decisão, já que, com 95% de nível de confiança (5% de nível de significância), teremos:

Se  $valor-P t < 0,05$  para o intercepto,  $\alpha \neq 0$

e

Se  $valor-P t < 0,05$  para determinada variável  $X$ ,  $\beta \neq 0$ .

Utilizando os dados do nosso exemplo inicial, temos que o erro-padrão da regressão é:

$$s.e. = \sqrt{\frac{\sum_{i=1}^n (u_i)^2}{(n-k)}} = \sqrt{\frac{361,15}{(10-2)}} = 6,7189$$

que também é fornecido pelos *outputs* do Excel (Figura 12.18).

	A	B	C	D	E	F	G	H	I
1	RESUMO DOS RESULTADOS								
2									
3	<i>Estatística de regressão</i>								
4	R múltiplo	0,90522134	<b>Erro-Padrão</b>						
5	R-Quadrado	0,81942568							
6	R-quadrado ajustado	0,79585389							
7	Erro padrão	6,7189731							
8	Observações	10							
9									
10	ANOVA								
11	g/f	SQ	MQ	F	F de significação				
12	Regressão	1	1638,851351	1638,85135	36,303087	0,000314449			
13	Resíduo	8	361,1486486	45,1435811					
14	Total	9	2000						
15									
16		Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores	Inferior 95,0%	Superior 95,0%
17	Interseção	5,87837838	4,532327565	1,29698886	0,23078848	-4,573187721	16,32994448	-4,573187721	16,32994448
18	Variável X 1	1,41891892	0,235497229	6,02520431	0,00031445	0,875861336	1,961976502	0,875861336	1,961976502
19									
20									
21									
22	RESULTADOS DE RESÍDUOS								
23									
24	Observação	Y previsto	Resíduos						
25	1	17,2297297	-2,22972973						
26	2	14,3918919	5,608108108						
27	3	27,1621622	-7,16216216						
28	4	34,2567568	5,743243243						
29	5	41,3513514	8,648648649						
30	6	21,4864865	3,513513514						
31	7	12,972973	-2,97297297						
32	8	51,2837838	3,716216216						
33	9	45,6081081	-10,6081081						
34	10	34,2567568	-4,25675676						

**Figura 12.18** Cálculo do erro-padrão.

A partir da expressão (12.21), podemos calcular, para o nosso exemplo:

$$t_\alpha = \frac{\alpha}{s.e.(\alpha)} = \frac{5,8784}{6,7189 \cdot \sqrt{a_{jj}}}$$

$$t_\beta = \frac{\beta}{s.e.(\beta)} = \frac{1,4189}{6,7189 \cdot \sqrt{a_{jj}}}$$

em que  $a_{jj}$  é o j-ésimo elemento da diagonal principal resultante do seguinte cálculo matricial:

$$\left[ \begin{pmatrix} 1 & 1 & 1 & \dots \\ 8 & 6 & 15 & \dots \end{pmatrix} \begin{pmatrix} 1 & 8 \\ 1 & 6 \\ 1 & 15 \\ \dots & \dots \end{pmatrix} \right]^{-1} = \begin{pmatrix} 0,4550 & -0,0209 \\ -0,0209 & 0,0012 \end{pmatrix}$$

que resulta, portanto, em:

$$t_{\alpha} = \frac{\alpha}{s.e.(\alpha)} = \frac{5,8784}{6,7189 \cdot \sqrt{0,4550}} = \frac{5,8784}{4,532} = 1,2969$$

$$t_{\beta} = \frac{\beta}{s.e.(\beta)} = \frac{1,4189}{6,7189 \cdot \sqrt{0,0012}} = \frac{1,4189}{0,2354} = 6,0252$$

que, para 8 graus de liberdade ( $n - k = 10 - 2 = 8$ ), temos, por meio da Tabela B do apêndice do livro, que o  $t_c = 2,306$  para o nível de significância de 5% (probabilidade na cauda superior de 0,025 para a distribuição bicaudal). Desta forma, como o  $t_{cal} = 1,2969 < t_c = t_{8, 2,5\%} = 2,306$ , não podemos rejeitar a hipótese nula de que o parâmetro  $\alpha$  seja estatisticamente igual a zero a este nível de significância para a amostra em questão.

O mesmo, todavia, não ocorre para o parâmetro  $\beta$ , já que  $t_{cal} = 6,0252 > t_c = t_{8, 2,5\%} = 2,306$ . Podemos, portanto, rejeitar a hipótese nula neste caso, ou seja, ao nível de significância de 5% não podemos afirmar que este parâmetro seja estatisticamente igual a zero.

	A	B	C	D	E	F	G	H	I
1	RESUMO DOS RESULTADOS								
2									
3	Estatística de regressão								
4	R múltiplo	0,90522134							
5	R-Quadrado	0,81942568							
6	R-quadrado ajustado	0,79685389							
7	Erro padrão	6,71889731							
8	Observações	10							
9									
10	ANOVA								
11		gl	SQ	MS	F	F de significação			
12	Regressão	1	1638,851351	1638,85135	36,303087	0,000314449			
13	Resíduo	8	361,1486486	45,1435891					
14	Total	9	2000						
15									
16		Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores	Inferior 95,0%	Superior 95,0%
17	Interseção	5,87837838	4,532327565	1,29698886	0,23078848	-4,573187721	16,32994448	-4,573187721	16,32994448
18	Variável X 1	1,41891892	0,235497229	6,02520431	0,00031445	0,875861336	1,961976502	0,875861336	1,961976502
19									
20									
21									
22	RESULTADOS DE RESÍDUOS								
23									
24	Observação	Y previsto	Resíduos						
25	1	17,2297297	-2,22972973						
26	2	14,3918919	5,608108108						
27	3	27,1621622	-7,16216216						
28	4	34,2567568	5,743243243						
29	5	41,3513514	8,648648649						
30	6	21,4864865	3,513513514						
31	7	12,972973	-2,97297297						
32	8	51,2837838	3,716216216						
33	9	45,6081081	-10,6081081						
34	10	34,2567568	-4,25675676						

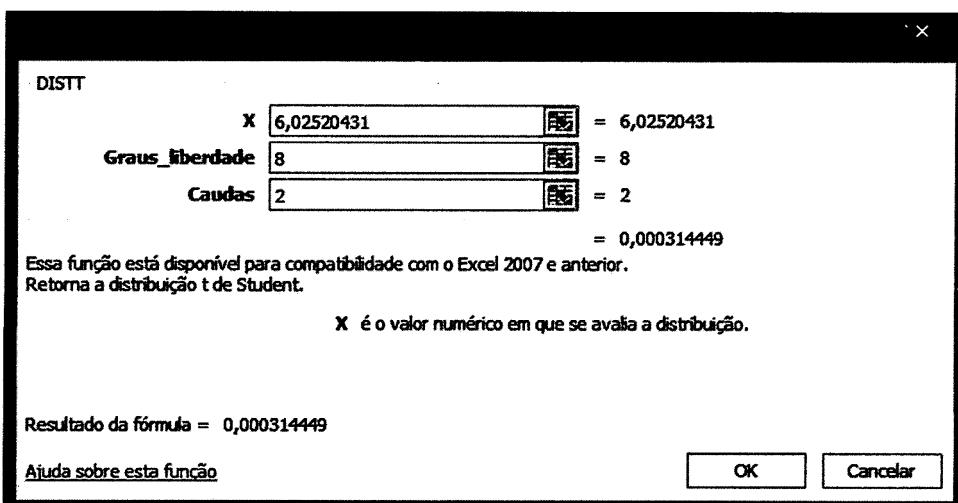
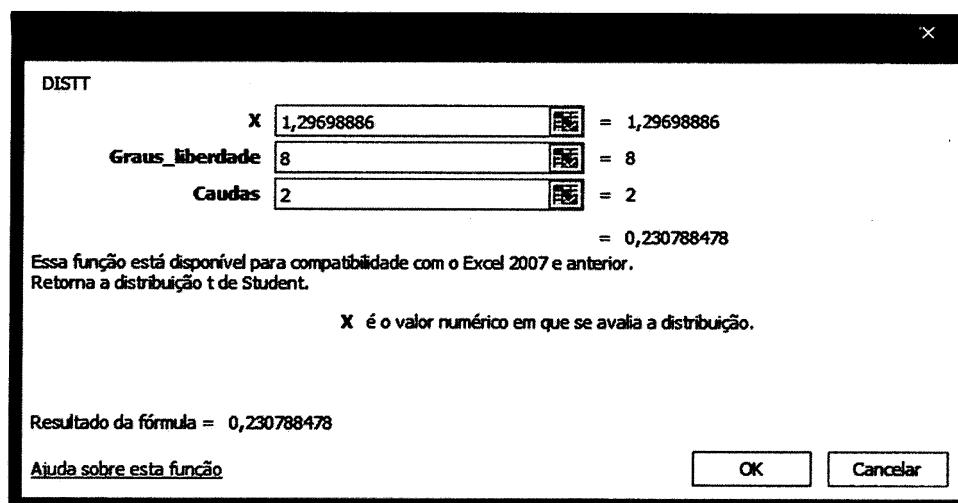
**Figura 12.19** Cálculo dos coeficientes e teste t de significância dos parâmetros.

Analogamente ao teste F, em vez de analisarmos se  $t_{cal} > t_c$  para cada parâmetro, podemos diretamente verificar se o nível de significância (valor-P) de cada  $t_{cal}$  é menor do que 0,05 (5%), a fim de mantermos o parâmetro no modelo final. O valor-P de cada  $t_{cal}$  pode ser obtido no Excel por meio do comando **Fórmulas → Inserir Função → DISTT**, que abrirá uma caixa de diálogo conforme mostra a Figura 12.20. Nesta figura, já estão apresentadas as caixas de diálogo correspondentes aos parâmetros  $\alpha$  e  $\beta$ .

É importante mencionar que, para regressões simples, a estatística  $F = t^2$  do parâmetro  $\beta$ , conforme demonstram Fávero *et al.* (2009). No nosso exemplo, portanto, podemos verificar que:

$$t_{\beta}^2 = F$$

$$t_{\beta}^2 = (6,0252)^2 = 36,30 = F$$



**Figura 12.20** Obtenção dos níveis de significância de  $t$  para os parâmetros  $\alpha$  e  $\beta$  (comando **Inserir Função**).

Como a hipótese  $H_1$  do teste  $F$  nos diz que pelo menos um parâmetro  $\beta$  é estatisticamente diferente de zero para determinado nível de significância, e visto que uma regressão simples apresenta apenas um único parâmetro  $\beta$ , se  $H_0$  for rejeitada para o teste  $F$ ,  $H_0$  também o será para o teste  $t$  deste parâmetro  $\beta$ .

Já para o parâmetro  $\alpha$ , como  $t_{cal} < t_c$  (*valor-P* de  $t_{cal}$  para o parâmetro  $\alpha > 0,05$ ) no nosso exemplo, poderíamos pensar na elaboração de uma nova regressão forçando que o intercepto seja igual a zero. Isso poderia ser elaborado por meio da caixa de diálogo de Regressão do Excel, com a seleção da opção **Constante é zero**.

Todavia, não iremos elaborar tal procedimento, uma vez que a não rejeição da hipótese nula de que o parâmetro  $\alpha$  seja estatisticamente igual a zero é decorrência da pequena amostra utilizada, porém não impede que um pesquisador faça previsões por meio da utilização do modelo obtido. A imposição de que o  $\alpha$  seja zero poderá gerar vieses de previsão pela geração de outra reta que possivelmente não será a mais adequada para se elaborarem interpolações nos dados. A Figura 12.21 ilustra este fato.

Desta forma, o fato de não podermos rejeitar que o parâmetro  $\alpha$  seja estatisticamente igual a zero a determinado nível de significância não implica que, necessariamente, devemos forçar a sua exclusão do modelo. Todavia, se esta for a decisão do pesquisador, é importante que se tenha ao menos a consciência de que apenas será gerado um modelo diferente daquele obtido inicialmente, com consequências para a elaboração de previsões.

A não rejeição da hipótese nula para o parâmetro  $\beta$  a determinado nível de significância, por outro lado, deve indicar que a correspondente variável  $X$  não se correlaciona com a variável  $Y$  e, portanto, deve ser excluída do modelo final.

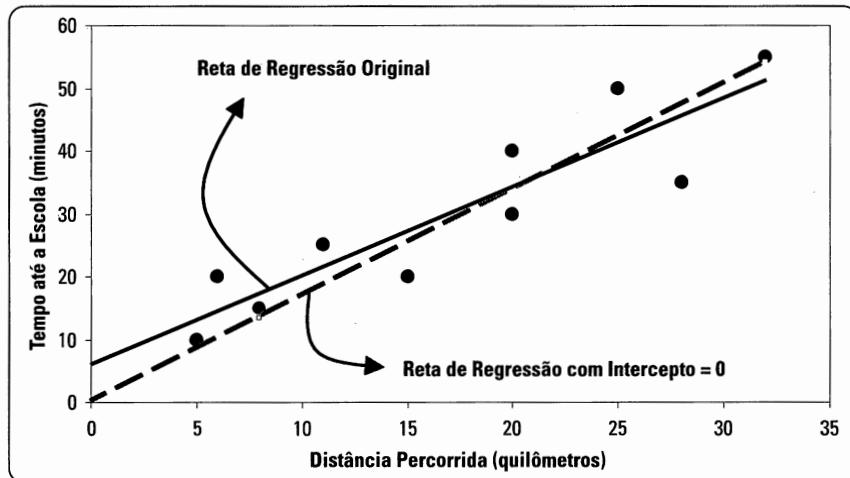


Figura 12.21 Retas de regressão original e com intercepto igual a zero.

Quando apresentarmos, mais adiante neste capítulo, a análise de regressão por meio dos softwares Stata (seção 12.5) e SPSS (seção 12.6), será introduzido o **procedimento Stepwise**, que tem a propriedade de automaticamente excluir ou manter os parâmetros  $\beta$  no modelo em função dos critérios apresentados e oferecer o modelo final apenas com parâmetros  $\beta$  estatisticamente diferentes de zero para determinado nível de significância.

#### 12.2.4. Construção dos intervalos de confiança dos parâmetros do modelo e elaboração de previsões

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

$$\begin{aligned}
 P\left[\alpha - t_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sum_{i=1}^n (u_i)^2}{(n-k)} \cdot \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)} \leq \alpha \leq \alpha + t_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sum_{i=1}^n (u_i)^2}{(n-k)} \cdot \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)}\right] = 95\% \quad (12.22) \\
 P\left[\beta_j - t_{\frac{\alpha}{2}} \cdot \frac{s.e.}{\sqrt{\left(\sum_{i=1}^n X_i^2\right) - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}} \leq \beta_j \leq \beta_j + t_{\frac{\alpha}{2}} \cdot \frac{s.e.}{\sqrt{\left(\sum_{i=1}^n X_i^2\right) - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}}\right] = 95\%
 \end{aligned}$$

Portanto, para o nosso exemplo, temos que:

**Parâmetro  $\alpha$ :**

$$P\left[5,8784 - 2,306 \cdot \sqrt{\frac{361,1486}{(8)} \cdot \left(\frac{1}{10} + \frac{289}{814}\right)} \leq \alpha \leq 5,8784 + 2,306 \cdot \sqrt{\frac{361,1486}{(8)} \cdot \left(\frac{1}{10} + \frac{289}{814}\right)}\right] = 95\%$$

$$P[-4,5731 \leq \alpha \leq 16,3299] = 95\%$$

Como o intervalo de confiança para o parâmetro  $\alpha$  contém o zero, não podemos rejeitar, ao nível de confiança de 95%, que este parâmetro seja estatisticamente igual a zero, conforme já verificado quando do cálculo da estatística  $t$ .

**Parâmetro  $\beta$ :**

$$P\left[1,4189 - 2,306 \cdot \frac{6,7189}{\sqrt{(3704) - \frac{(170)^2}{10}}} \leq \beta \leq 1,4189 + 2,306 \cdot \frac{6,7189}{\sqrt{(3704) - \frac{(170)^2}{10}}}\right] = 95\%$$

$$P[0,8758 \leq \beta \leq 1,9619] = 95\%$$

Como o intervalo de confiança para o parâmetro  $\beta$  não contém o zero, podemos rejeitar, ao nível de confiança de 95%, que este parâmetro seja estatisticamente igual a zero, conforme também já verificado quando do cálculo da estatística  $t$ .

Estes intervalos também são gerados nos *outputs* do Excel. Como o padrão do software é utilizar um nível de confiança de 95%, estes intervalos são mostrados duas vezes, a fim de permitir que o pesquisador altere manualmente o nível de confiança desejado, selecionando a opção **Nível de confiança** na caixa de diálogo de Regressão do Excel, e ainda tenha condições de analisar os intervalos para o nível de confiança mais comumente utilizado (95%). Em outras palavras, os intervalos para o nível de confiança de 95% no Excel serão sempre apresentados, dando ao pesquisador a possibilidade de analisar paralelamente intervalos com outro nível de confiança.

Iremos, desta forma, alterar a caixa de diálogo da regressão (Figura 12.22), a fim de permitir que o software também calcule os intervalos dos parâmetros ao nível de confiança de, por exemplo, 90%. Estes *outputs* estão apresentados na Figura 12.23.

Percebe-se que os valores das bandas inferior e superior são simétricos em relação ao parâmetro médio estimado e oferecem ao pesquisador uma possibilidade de serem elaboradas previsões com determinado nível de confiança. No caso do parâmetro  $\beta$  do nosso exemplo, como os extremos das bandas inferior e superior são positivos, podemos dizer que este parâmetro é positivo, com 95% de confiança. Além disso, podemos também dizer que o intervalo [0,8758; 1,9619] contém  $\beta$  com 95% de confiança.

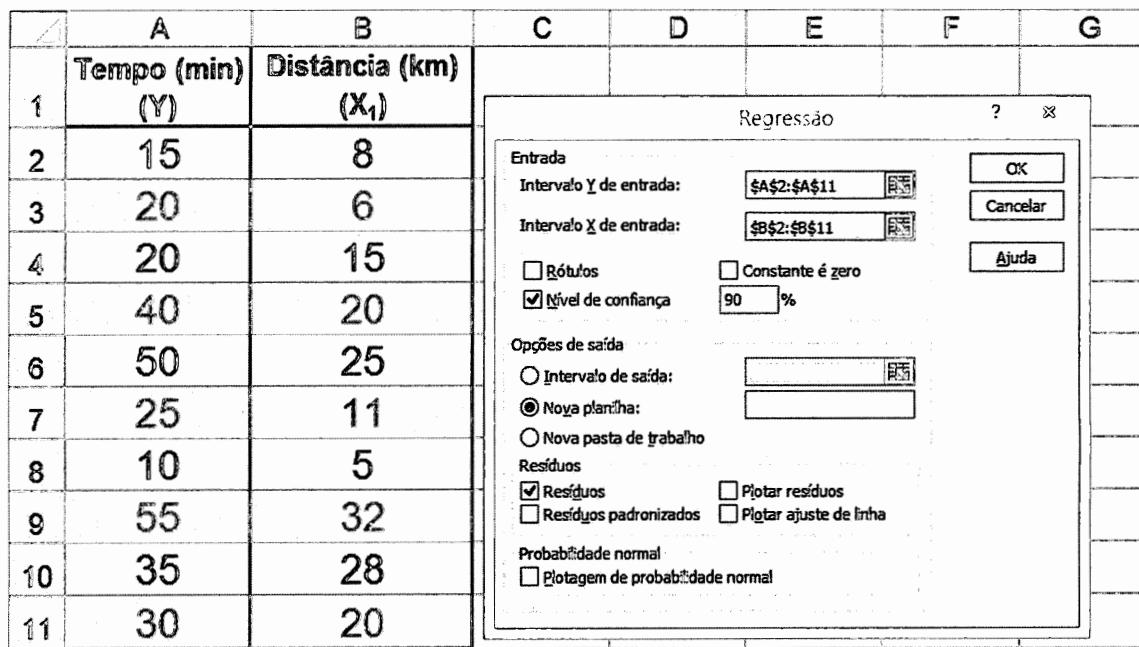


Figura 12.22 Alteração do nível de confiança dos intervalos dos parâmetros para 90%.

Diferentemente do que fizemos para o nível de confiança de 95%, não iremos calcular manualmente os intervalos dos parâmetros para o nível de confiança de 90%. Porém a análise dos *outputs* do Excel nos permite afirmar que o intervalo [0,9810; 1,8568] contém  $\beta$  com 90% de confiança. Desta maneira, podemos dizer que, quanto menores os níveis de confiança, mais estreitos (menor amplitude) serão os intervalos para conter determinado parâmetro. Por outro lado, quanto maiores forem os níveis de confiança, maior amplitude terão os intervalos para conter este parâmetro.

A Figura 12.24 ilustra o que acontece quando temos uma nuvem dispersa de pontos em torno de uma reta de regressão.

Podemos notar que, por mais que o parâmetro  $\alpha$  seja positivo e matematicamente igual a 5,8784, não podemos afirmar que ele seja estatisticamente diferente de zero para esta pequena amostra, uma vez que o intervalo de confiança contém o intercepto igual a zero (origem). Uma amostra maior poderia resolver este problema.

Já para o parâmetro  $\beta$ , podemos notar que a inclinação tem sido sempre positiva, com valor médio calculado matematicamente e igual a 1,4189. Podemos visualmente notar que seu intervalo de confiança não contém a inclinação igual a zero.

	A	B	C	D	E	F	G	H	I
1	RESUMO DOS RESULTADOS								
2									
3	<i>Estatística de regressão</i>								
4	R múltiplo	0,90522134							
5	R-Quadrado	0,81942568							
6	R-quadrado ajustado	0,79685389							
7	Erro padrão	6,71889731							
8	Observações	10							
9									
10	ANOVA								
11		g/	SO	MQ	F	F de significação			
12	Regressão	1	1638,851351	1638,85135	36,303087	0,000314449			
13	Resíduo	8	361,1486486	45,1435811					
14	Total	9	2000						
15									
16		Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores	Inferior 90,0%	Superior 90,0%
17	Interseção	5,87837838	4,532327565	1,29698886	0,23078848	-4,573187721	16,32994448	-2,549702432	14,30645919
18	Variável X 1	1,41891892	0,235497229	6,02520431	0,00031445	0,875861336	1,961976502	0,98100051	1,856837328
19									
20									
21									
22	RESULTADOS DE RESÍDUOS								
23									
24	Observação	Y previsto	Resíduos			Intervalos dos Parâmetros com Nível de Confiança de 95%	Intervalos dos Parâmetros com Nível de Confiança de 90%		
25	1	17,2297297	-2,22972973						
26	2	14,3918919	5,608108108						
27	3	27,1621622	-7,16216216						
28	4	34,2567568	5,743243243						
29	5	41,3513514	8,648648649						
30	6	21,4864865	3,513513514						
31	7	12,972973	-2,97297297						
32	8	51,2837838	3,716216216						
33	9	45,6081081	-10,6081081						
34	10	34,2567568	-4,25675676						

Figura 12.23 Intervalos com níveis de confiança de 95% e 90% para cada um dos parâmetros.

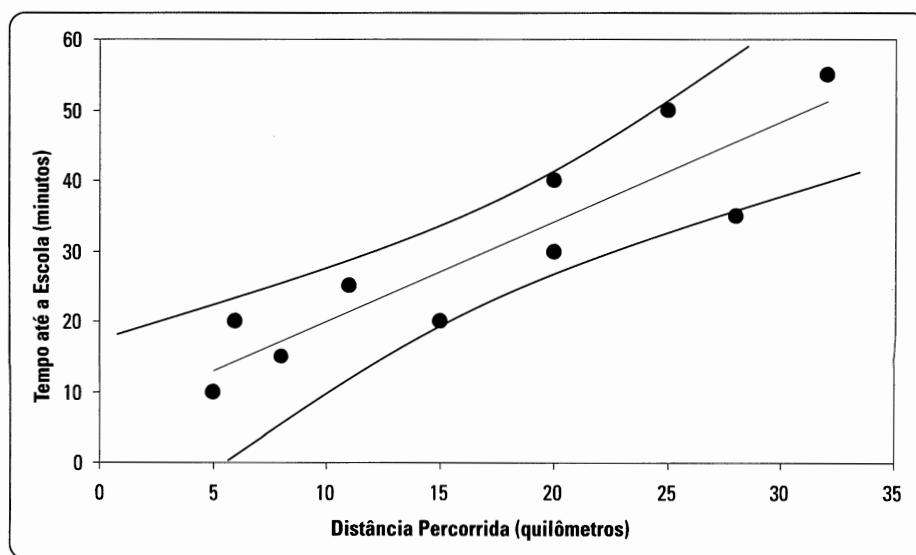


Figura 12.24 Intervalos de confiança para a dispersão de pontos em torno da reta de regressão.

Conforme já discutido, a rejeição da hipótese nula para o parâmetro  $\beta$ , a determinado nível de significância, indica que a correspondente variável  $X$  correlaciona-se com a variável  $Y$  e, consequentemente, deve permanecer no modelo final. Podemos, portanto, concluir que a decisão pela exclusão de uma variável  $X$  em determinado modelo de regressão pode ser realizada por meio da análise direta da estatística  $t$  de seu respectivo parâmetro

$\beta$  (se  $t_{cal} < t_c \rightarrow valor-P > 0,05 \rightarrow$  não podemos rejeitar que o parâmetro seja estatisticamente igual a zero) ou por meio da análise do intervalo de confiança (se o mesmo contém o zero). O Quadro 12.1 apresenta os critérios de inclusão ou exclusão de parâmetros  $\beta_j$  ( $j = 1, 2, \dots, k$ ) em modelos de regressão.

**Quadro 12.1** Decisão de inclusão de parâmetros  $\beta_j$  em modelos de regressão.

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

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

Após a discussão desses conceitos, o professor propôs o seguinte exercício à turma de estudantes: **Qual a previsão do tempo médio de percurso ( $Y$  estimado, ou  $\hat{Y}$ ) de um aluno que percorre 17 quilômetros para chegar à escola? Quais seriam os valores mínimo e máximo que este tempo de percurso poderia assumir, com 95% de confiança?**

A primeira parte do exercício pode ser resolvida pela simples substituição do valor de  $X_i = 17$  na equação inicialmente obtida. Assim:

$$\hat{tempo}_i = 5,8784 + 1,4189.dist_i = 5,8784 + 1,4189.(17) = 29,9997 \text{ min}$$

A segunda parte do exercício nos remete aos *outputs* da Figura 12.23, já que os parâmetros  $\alpha$  e  $\beta$  assumem intervalos de  $[-4,5731; 16,3299]$  e  $[0,8758; 1,9619]$ , respectivamente, ao nível de confiança de 95%. Sendo assim, as equações que determinam os valores mínimo e máximo do tempo de percurso para este nível de confiança são:

Tempo mínimo:

$$\hat{tempo}_{\min} = -4,5731 + 0,8758.dist_i = -4,5731 + 0,8758.(17) = 10,3155 \text{ min}$$

Tempo máximo:

$$\hat{tempo}_{\max} = 16,3299 + 1,9619.dist_i = 16,3299 + 1,9619.(17) = 49,6822 \text{ min}$$

Logo, podemos dizer que há 95% de confiança de que um aluno que percorre 17 quilômetros para chegar à escola leve entre 10,3155 min e 49,6822 min, com tempo médio estimado de 29,9997 min.

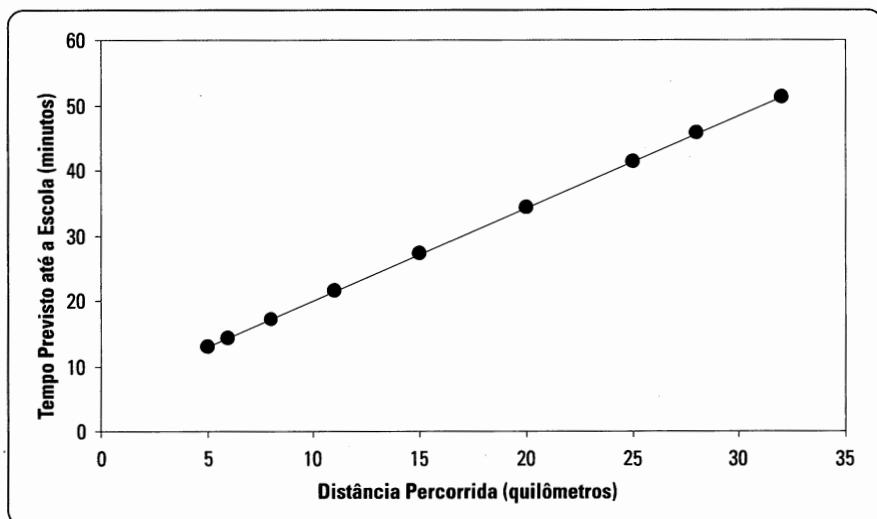
Obviamente que a amplitude destes valores não é pequena, por conta do intervalo de confiança do parâmetro  $\alpha$  ser bastante amplo. Este fato poderia ser corrigido pelo incremento do tamanho da amostra ou pela inclusão de novas variáveis  $X$  estatisticamente significantes no modelo (que passaria a ser um modelo de regressão múltipla), já que, neste último caso, aumentar-se-ia o valor do  $R^2$ .

Após o professor apresentar os resultados de seu modelo aos estudantes, um curioso aluno levantou-se e perguntou: **Mas então, professor, existe alguma influência do coeficiente de ajuste  $R^2$  dos modelos de regressão sobre a amplitude dos intervalos de confiança? Se elaborássemos esta regressão linear substituindo  $Y$  por  $\hat{Y}$ , como seriam os resultados? A equação seria alterada? E o  $R^2$ ? E os intervalos de confiança?**

E o professor substituiu  $Y$  por  $\hat{Y}$  e elaborou novamente a regressão por meio do banco de dados apresentado na Tabela 12.4.

**Tabela 12.4** Banco de dados para a elaboração da nova regressão.

Observação ( $i$ )	Tempo previsto ( $\hat{Y}_i$ )	Distância ( $X_i$ )
1	17,23	8
2	14,39	6
3	27,16	15
4	34,26	20
5	41,35	25
6	21,49	11
7	12,97	5
8	51,28	32
9	45,61	28
10	34,26	20



**Figura 12.25** Gráfico de dispersão e reta de regressão linear entre tempo previsto ( $\hat{Y}$ ) e distância percorrida ( $X$ ).

O primeiro passo adotado pelo professor foi elaborar o novo gráfico de dispersão, já com a reta estimada de regressão. Este gráfico está apresentado na Figura 12.25.

Como podemos observar, obviamente todos os pontos agora se situam sobre a reta de regressão, uma vez que tal procedimento forçou esta situação pelo fato de o cálculo de cada  $\hat{Y}_i$  ter utilizado a própria reta de regressão obtida anteriormente. Vamos aos novos *outputs* (Figura 12.26).

Como já esperávamos, o  $R^2$  é 1. E a equação do modelo é exatamente aquela já calculada anteriormente, uma vez que é a mesma reta.

Porém, podemos observar que os testes  $F$  e  $t$  fazem com que rejeitemos fortemente as suas respectivas hipóteses nulas. Mesmo para o parâmetro  $\alpha$ , que anteriormente não podia ser considerado estatisticamente diferente de zero, agora apresenta seu teste  $t$  nos dizendo que podemos rejeitar, ao nível de confiança de 95% (ou até maior), que este parâmetro é estatisticamente igual a zero. Isso ocorre porque anteriormente a pequena amostra utilizada ( $n = 10$  observações) não nos permitia afirmar que o intercepto era diferente de zero, já que a dispersão de pontos gerava um intervalo de confiança que continha o intercepto igual a zero (Figura 12.24).

Por outro lado, quando todos os pontos estão sobre a reta, cada um dos termos do resíduo passa a ser zero, o que faz com que o  $R^2$  se torne 1. Além disso, a equação obtida não é mais uma reta ajustada a uma dispersão de pontos, mas a própria reta que passa por todos os pontos e explica completamente o comportamento da amostra. Assim, não temos dispersão em torno da reta de regressão e os intervalos de confiança passam a apresentar amplitude nula, como também podemos observar por meio da Figura 12.26. Neste caso, para qualquer nível de confiança, não são mais alterados os valores de cada intervalo dos parâmetros, o que nos faz afirmar que o intervalo  $[5,8784; 5,8784]$  contém  $\alpha$  e o intervalo  $[1,4189; 1,4189]$  contém  $\beta$  com 100% de confiança. Em outras palavras, neste caso extremo  $\alpha$  é matematicamente igual a 5,8784 e  $\beta$  é matematicamente igual a 1,4189.

Assim sendo, o  $R^2$  é um indicador de quão amplos serão os intervalos de confiança dos parâmetros. Portanto, modelos com  $R^2$  mais elevados propiciam ao pesquisador a elaboração de previsões com maior acurácia, dado que a nuvem de pontos será menos dispersa em torno da reta de regressão, o que reduzirá a amplitude dos intervalos de confiança dos parâmetros.

Por outro lado, modelos com baixos valores de  $R^2$  podem prejudicar a elaboração de previsões em razão da maior amplitude dos intervalos de confiança dos parâmetros, mas não invalidam a existência do modelo propriamente dito. Conforme já discutimos, muitos pesquisadores dão importância demasiada ao  $R^2$ , porém será o teste  $F$  que permitirá ao mesmo afirmar que existe um modelo de regressão (pelo menos uma variável  $X$  considerada é estatisticamente significante para explicar  $Y$ ). Assim, não é raro encontrarmos em Administração, em Contabilidade ou em Economia modelos com baixíssimos valores de  $R^2$  e com valores de  $F$  estatisticamente significantes, o que demonstra que o fenômeno estudado  $Y$  sofreu mudanças em seu comportamento em decorrência de algumas variáveis  $X$  adequadamente incluídas no modelo, porém baixa será a acurácia de previsão pela impossibilidade de se monitorarem todas as variáveis que efetivamente explicam a variação daquele fenômeno  $Y$ . Dentro das mencionadas áreas do conhecimento, tal fato é facilmente encontrado em trabalhos sobre Finanças e Mercado de Ações.

	A	B	C	D	E	F	G
1	RESUMO DOS RESULTADOS						
2							
3	<i>Estatística de regressão</i>						
4	R múltiplo	1					
5	R-Quadrado	1					
6	R-quadrado ajustado	1					
7	Erro padrão	0					
8	Observações	10					
9							
10	ANOVA						
11		g/	SQ	MQ	F	<i>F de significação</i>	
12	Regressão	1	1638,851351	1638,851351	4,03541E+32	0,00	
13	Resíduo	8	0	0			
14	Total	9	1638,851351				
15							
16		Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores
17	Interseção	5,878378378	0,00	4,32423E+15	0,00	5,878378378	5,878378378
18	Variável X 1	1,418918919	0,00	2,00883E+16	0,00	1,418918919	1,418918919
19							
20							
21							
22	RESULTADOS DE RESÍDUOS						
23							
24	Observação	Y previsto	Resíduos				
25	1	17,22972973	0				
26	2	14,39189189	0				
27	3	27,16216216	0				
28	4	34,25675676	0				
29	5	41,35135135	0				
30	6	21,48648649	0				
31	7	12,97297297	0				
32	8	51,28378378	0				
33	9	45,60810811	0				
34	10	34,25675676	0				

**Figura 12.26 Outputs da regressão linear entre tempo previsto ( $\hat{Y}$ ) e distância percorrida ( $X$ ).**

### 12.2.5. Estimação de modelos lineares de regressão múltipla

Segundo Fávero *et al.* (2009), a regressão linear múltipla apresenta a mesma lógica apresentada para a regressão linear simples, porém agora com a inclusão de mais de uma variável explicativa  $X$  no modelo. A utilização de muitas variáveis explicativas dependerá da teoria subjacente e de estudos predecessores, bem como da experiência e do bom senso do pesquisador, a fim de que seja possível fundamentar a decisão.

Inicialmente, o conceito *ceteris paribus* (mantidas as demais condições constantes) deve ser utilizado na análise da regressão múltipla, uma vez que a interpretação do parâmetro de cada variável será feita isoladamente. Assim, em um modelo que possui duas variáveis explicativas,  $X_1$  e  $X_2$ , os respectivos coeficientes serão analisados de forma a considerar todos os outros fatores constantes.

Para exemplificarmos a análise de regressão linear múltipla, utilizaremos o mesmo exemplo até agora abordado neste capítulo. Porém, neste momento, imaginemos que o professor tenha tomado a decisão de coletar mais uma variável de cada um dos alunos. Esta variável será referente ao número de semáforos pelos quais cada aluno é obrigado a passar, e a chamaremos de variável *sem*. Desta forma, o modelo teórico passará a ser:

$$\text{tempo}_i = a + b_1 \cdot \text{dist}_i + b_2 \cdot \text{sem}_i + u_i$$

que, analogamente ao apresentado para a regressão simples, temos que:

$$\hat{\text{tempo}}_i = \alpha + \beta_1 \cdot \text{dist}_i + \beta_2 \cdot \text{sem}_i$$

em que  $\alpha$ ,  $\beta_1$  e  $\beta_2$  são, respectivamente, as estimativas dos parâmetros  $a$ ,  $b_1$  e  $b_2$ .

O novo banco de dados encontra-se na Tabela 12.5, bem como no arquivo **Tempodistsem.xls**.

**Tabela 12.5** Exemplo: tempo de percurso x distância percorrida e quantidade de semáforos.

Estudante	Tempo para chegar à escola (minutos) ( $Y_i$ )	Distância percorrida até a escola (quilômetros) ( $X_{1i}$ )	Quantidade de semáforos ( $X_{2i}$ )
Gabriela	15	8	0
Dalila	20	6	1
Gustavo	20	15	0
Letícia	40	20	1
Luiz Ovídio	50	25	2
Leonor	25	11	1
Ana	10	5	0
Antônio	55	32	3
Júlia	35	28	1
Mariana	30	20	1

Iremos agora desenvolver algebricamente os procedimentos para o cálculo dos parâmetros do modelo, assim como fizemos para o modelo de regressão simples. Por meio da seguinte expressão:

$$Y_i = \alpha + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + u_i$$

podemos também definir que a somatória dos quadrados dos resíduos seja mínima. Assim:

$$\sum_{i=1}^n (Y_i - \beta_1 \cdot X_{1i} - \beta_2 \cdot X_{2i} - \alpha)^2 = \text{mín}$$

A minimização ocorre ao se derivar a expressão anterior em  $\alpha$ ,  $\beta_1$  e  $\beta_2$  e igualar as expressões resultantes a zero. Assim:

$$\frac{\partial}{\partial \alpha} \left[ \sum_{i=1}^n (Y_i - \beta_1 \cdot X_{1i} - \beta_2 \cdot X_{2i} - \alpha)^2 \right] = -2 \sum_{i=1}^n (Y_i - \beta_1 \cdot X_{1i} - \beta_2 \cdot X_{2i} - \alpha) = 0 \quad (12.23)$$

$$\frac{\partial}{\partial \beta_1} \left[ \sum_{i=1}^n (Y_i - \beta_1 \cdot X_{1i} - \beta_2 \cdot X_{2i} - \alpha)^2 \right] = -2 \sum_{i=1}^n X_{1i} \cdot (Y_i - \beta_1 \cdot X_{1i} - \beta_2 \cdot X_{2i} - \alpha) = 0 \quad (12.24)$$

$$\frac{\partial}{\partial \beta_2} \left[ \sum_{i=1}^n (Y_i - \beta_1 \cdot X_{1i} - \beta_2 \cdot X_{2i} - \alpha)^2 \right] = -2 \sum_{i=1}^n X_{2i} \cdot (Y_i - \beta_1 \cdot X_{1i} - \beta_2 \cdot X_{2i} - \alpha) = 0 \quad (12.25)$$

que gera o seguinte sistema de três equações e três incógnitas:

$$\begin{cases} \sum_{i=1}^n Y_i = n \cdot \alpha + \beta_1 \cdot \sum_{i=1}^n X_{1i} + \beta_2 \cdot \sum_{i=1}^n X_{2i} \\ \sum_{i=1}^n Y_i \cdot X_{1i} = \alpha \cdot \sum_{i=1}^n X_{1i} + \beta_1 \cdot \sum_{i=1}^n X_{1i}^2 + \beta_2 \cdot \sum_{i=1}^n X_{1i} \cdot X_{2i} \\ \sum_{i=1}^n Y_i \cdot X_{2i} = \alpha \cdot \sum_{i=1}^n X_{2i} + \beta_1 \cdot \sum_{i=1}^n X_{1i} \cdot X_{2i} + \beta_2 \cdot \sum_{i=1}^n X_{2i}^2 \end{cases} \quad (12.26)$$

Dividindo-se a primeira equação da expressão (12.26) por  $n$ , chegamos a:

$$\alpha = \bar{Y} - \beta_1 \cdot \bar{X}_1 - \beta_2 \cdot \bar{X}_2 \quad (12.27)$$

Por meio da substituição da expressão (12.27) nas duas últimas equações da expressão (12.26), chegaremos ao seguinte sistema de duas equações e duas incógnitas:

$$\begin{cases} \sum_{i=1}^n Y_i \cdot X_{1i} - \frac{\sum_{i=1}^n Y_i \cdot \sum_{i=1}^n X_{1i}}{n} = \beta_1 \cdot \left[ \sum_{i=1}^n X_{1i}^2 - \frac{\left( \sum_{i=1}^n X_{1i} \right)^2}{n} \right] + \beta_2 \cdot \left[ \sum_{i=1}^n X_{1i} \cdot X_{2i} - \frac{\left( \sum_{i=1}^n X_{1i} \right) \left( \sum_{i=1}^n X_{2i} \right)}{n} \right] \\ \sum_{i=1}^n Y_i \cdot X_{2i} - \frac{\sum_{i=1}^n Y_i \cdot \sum_{i=1}^n X_{2i}}{n} = \beta_1 \cdot \left[ \sum_{i=1}^n X_{1i} \cdot X_{2i} - \frac{\left( \sum_{i=1}^n X_{1i} \right) \left( \sum_{i=1}^n X_{2i} \right)}{n} \right] + \beta_2 \cdot \left[ \sum_{i=1}^n X_{2i}^2 - \frac{\left( \sum_{i=1}^n X_{2i} \right)^2}{n} \right] \end{cases} \quad (12.28)$$

Vamos agora calcular manualmente os parâmetros do modelo do nosso exemplo. Para tanto, iremos utilizar a planilha apresentada na Tabela 12.6.

Vamos agora substituir os valores no sistema representado pela expressão (12.28). Assim:

$$\begin{cases} 6255 - \frac{300 \cdot 170}{10} = \beta_1 \cdot \left[ 3704 - \frac{(170)^2}{10} \right] + \beta_2 \cdot \left[ 231 - \frac{(170) \cdot (10)}{10} \right] \\ 415 - \frac{300 \cdot 10}{10} = \beta_1 \cdot \left[ 231 - \frac{(170) \cdot (10)}{10} \right] + \beta_2 \cdot \left[ 18 - \frac{(10)^2}{10} \right] \end{cases}$$

que resulta em:

$$\begin{cases} 1155 = 814 \cdot \beta_1 + 61 \cdot \beta_2 \\ 115 = 61 \cdot \beta_1 + 8 \cdot \beta_2 \end{cases}$$

Resolvendo o sistema, chegamos a:

$$\beta_1 = 0,7972 \text{ e } \beta_2 = 8,2963$$

**Tabela 12.6** Planilha para o cálculo dos parâmetros da regressão linear múltipla.

Obs. ( <i>i</i> )	$Y_i$	$X_{1i}$	$X_{2i}$	$Y_i \cdot X_{1i}$	$Y_i \cdot X_{2i}$	$X_{1i} \cdot X_{2i}$	$(Y_i)^2$	$(X_{1i})^2$	$(X_{2i})^2$
1	15	8	0	120	0	0	225	64	0
2	20	6	1	120	20	6	400	36	1
3	20	15	0	300	0	0	400	225	0
4	40	20	1	800	40	20	1600	400	1
5	50	25	2	1250	100	50	2500	625	4
6	25	11	1	275	25	11	625	121	1
7	10	5	0	50	0	0	100	25	0
8	55	32	3	1760	165	96	3025	1024	9
9	35	28	1	980	35	28	1225	784	1
10	30	20	1	600	30	20	225	400	1
<b>Soma</b>	<b>300</b>	<b>170</b>	<b>10</b>	<b>6255</b>	<b>415</b>	<b>231</b>	<b>11000</b>	<b>3704</b>	<b>18</b>
<b>Média</b>	<b>30</b>	<b>17</b>	<b>1</b>						

Assim, temos que:

$$\alpha = \bar{Y} - \beta_1 \bar{X}_1 - \beta_2 \bar{X}_2 = 30 - 0,7972.(17) - 8,2963.(1) = 8,1512$$

Portanto, a equação do tempo estimado para se chegar à escola agora passa a ser:

$$\hat{\text{tempo}}_i = 8,1512 + 0,7972.\text{dist}_i + 8,2963.\text{sem}_i$$

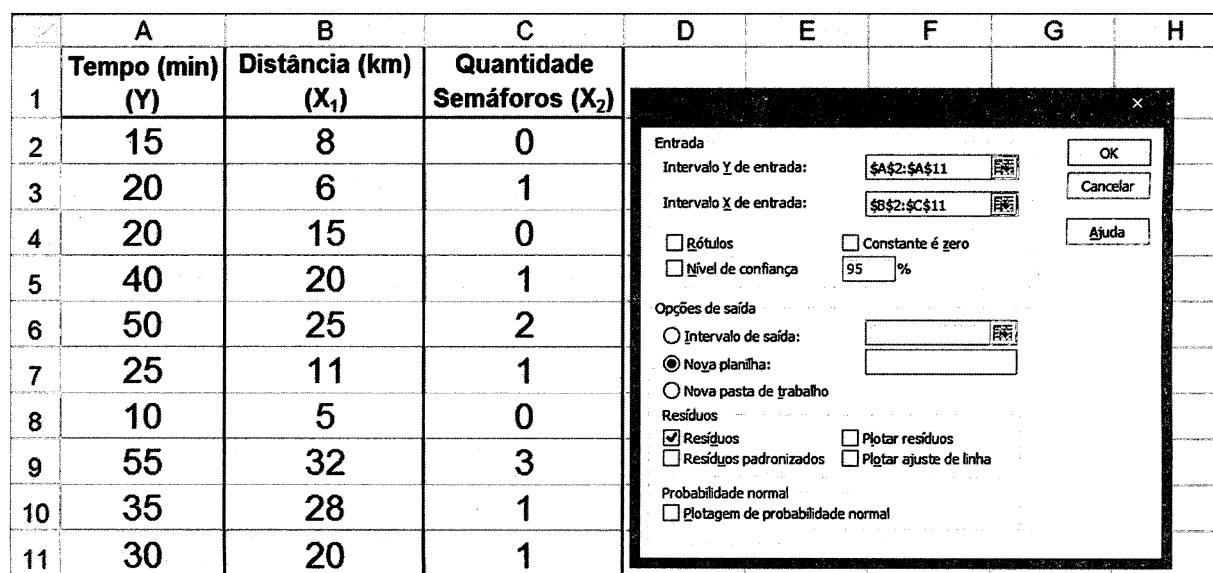
Ressalta-se que as estimativas destes parâmetros também poderiam ter sido obtidas por meio do procedimento **Solver** do Excel, como elaborado na seção 12.2.1.

Os cálculos do coeficiente de ajuste  $R^2$ , das estatísticas  $F$  e  $t$  e dos valores extremos dos intervalos de confiança não serão novamente elaborados de forma manual, dado que seguem exatamente o mesmo procedimento já executado nas seções 12.2.2, 12.2.3 e 12.2.4 e podem ser realizados por meio das respectivas expressões apresentadas até o presente momento. A Tabela 12.7 poderá auxiliar neste sentido.

Vamos diretamente para a elaboração desta regressão linear múltipla no Excel (arquivo **Tempodistsem.xls**). Na caixa de diálogo da regressão, devemos selecionar conjuntamente as variáveis referentes à distância percorrida e à quantidade de semáforos, como mostra a Figura 12.27.

**Tabela 12.7** Planilha para o cálculo das demais estatísticas.

Observação ( <i>i</i> )	Tempo ( $Y_i$ )	Distância ( $X_{1i}$ )	Semáforos ( $X_{2i}$ )	$\hat{Y}_i$	$u_i$ ( $Y_i - \hat{Y}_i$ )	$(\hat{Y}_i - \bar{Y})^2$	$(u_i)^2$
1	15	8	8	14,53	0,47	239,36	0,22
2	20	6	6	21,23	-1,23	76,90	1,51
3	20	15	15	20,11	-0,11	97,83	0,01
4	40	20	20	32,39	7,61	5,72	57,89
5	50	25	25	44,67	5,33	215,32	28,37
6	25	11	11	25,22	-0,22	22,88	0,05
7	10	5	5	12,14	-2,14	319,08	4,57
8	55	32	32	58,55	-3,55	815,14	12,61
9	35	28	28	38,77	-3,77	76,90	14,21
10	30	20	20	32,39	-2,39	5,72	5,72
<b>Soma</b>	<b>300</b>	<b>170</b>	<b>10</b>			<b>1874,85</b>	<b>125,15</b>
<b>Média</b>	<b>30</b>	<b>17</b>	<b>1</b>				



**Figura 12.27** Regressão linear múltipla – seleção conjunta das variáveis explicativas.

A Figura 12.28 apresenta os *outputs* gerados.

Nestes *outputs* podemos encontrar as estimativas dos parâmetros do nosso modelo de regressão linear múltipla determinadas algebricamente.

Neste momento é importante introduzirmos o conceito de **R<sup>2</sup> ajustado**. Segundo Fávero *et al.* (2009), quando há o intuito de comparar o coeficiente de ajuste (R<sup>2</sup>) entre dois modelos com tamanhos de amostra diferentes ou com quantidades distintas de parâmetros, faz-se necessário o uso do R<sup>2</sup> ajustado, que é uma medida do R<sup>2</sup> da regressão estimada pelo método de mínimos quadrados ordinários ajustada pelo número de graus de liberdade, uma vez que a estimativa amostral de R<sup>2</sup> tende a superestimar o parâmetro populacional. A expressão do R<sup>2</sup> ajustado é:

$$R_{ajust}^2 = 1 - \frac{n-1}{n-k}(1-R^2) \quad (12.29)$$

em que *n* é o tamanho da amostra e *k* é o número de parâmetros do modelo de regressão (número de variáveis explicativas mais o intercepto). Quando o número de observações é muito grande, o ajuste pelos graus de liberdade torna-se desprezível, porém quando há um número significativamente diferente de variáveis *X* para duas amostras, deve-se utilizar o R<sup>2</sup> ajustado para a elaboração de comparações entre os modelos e optar pelo modelo com maior R<sup>2</sup> ajustado.

O R<sup>2</sup> aumenta quando uma nova variável é adicionada ao modelo, entretanto o R<sup>2</sup> ajustado nem sempre aumentará, bem como poderá diminuir ou até ficar negativo. Para este último caso, Stock e Watson (2004) explicam que o R<sup>2</sup> ajustado pode ficar negativo quando as variáveis explicativas, tomadas em conjunto, reduzirem a soma dos quadrados dos resíduos em um montante tão pequeno que esta redução não consiga compensar o fator (n-1)/(n-k).

Para o nosso exemplo, temos que:

$$R_{ajust}^2 = 1 - \frac{10-1}{10-3}(1-0,9374) = 0,9195$$

Portanto, até o presente momento, em detrimento da regressão simples aplicada inicialmente, devemos optar por esta regressão múltipla como sendo um melhor modelo para se estudar o comportamento do tempo de percurso para se chegar até a escola, uma vez que o R<sup>2</sup> ajustado é maior para este caso.

	A	B	C	D	E	F	G	H	I
1	RESUMO DOS RESULTADOS								
2									
3	<i>Estatística de regressão</i>								
4	R múltiplo	0,96820652							
5	R-Quadrado	0,93742386							
6	R-quadrado ajustado	0,91954497							
7	Erro padrão	4,22834441							
8	Observações	10							
9									
10	ANOVA								
11		gl	SQ	MQ	F	<i>F de significação</i>			
12	Regressão	2	1874,847725	937,423862	52,4318637	6,12958E-05			
13	Resíduo	7	125,1522752	17,8788965					
14	Total	9	2000						
15									
16		<i>Coeficientes</i>	<i>Erro padrão</i>	<i>Stat t</i>	<i>valor-P</i>	<i>95% inferiores</i>	<i>95% superiores</i>	<i>Inferior 95,0%</i>	<i>Superior 95,0%</i>
17	Interseção	8,15120029	2,920086914	2,79142386	0,02685329	1,246291955	15,05610862	1,246291955	15,05610862
18	Variável X 1	0,7972053	0,226378631	3,52155722	0,00970731	0,261904901	1,332505704	0,261904901	1,332505704
19	Variável X 2	8,29630957	2,283508533	3,63314148	0,00836288	2,896669913	13,69594922	2,896669913	13,69594922
20									
21									
22									
23	RESULTADOS DE RESÍDUOS								
24									
25	Observação	<i>Y previsto</i>	<i>Resíduos</i>						
26	1	14,5288427	0,471157291						
27	2	21,2307417	-1,23074167						
28	3	20,1092798	-0,10927983						
29	4	32,3916159	7,608384092						
30	5	44,673952	5,326048011						
31	6	25,2167682	-0,21676818						
32	7	12,1372268	-2,1372268						
33	8	58,5506987	-3,55069867						
34	9	38,7692583	-3,76925833						
35	10	32,3916159	-2,39161591						

Figura 12.28 Outputs da regressão linear múltipla no Excel.

Vamos dar sequência à análise dos demais *outputs*. Inicialmente, o teste *F* já nos informa que pelo menos uma das variáveis *X* relaciona-se significativamente com *Y*. Além disso, podemos também verificar, ao nível de significância de 5%, que todos os parâmetros ( $\alpha$ ,  $\beta_1$  e  $\beta_2$ ) são estatisticamente diferentes de zero (*valor-P* < 0,05 → intervalo de confiança não contém o zero). Conforme já discutido, a não rejeição da hipótese nula de que o intercepto seja estatisticamente igual a zero pode ser alterada ao se incluir uma variável explicativa significante no modelo. Notamos também que houve um perceptivo aumento no valor do  $R^2$ , o que fez também com que os intervalos de confiança dos parâmetros se tornassem mais estreitos.

Dessa forma, podemos concluir, para este caso, que o aumento de um semáforo ao longo do trajeto até a escola incrementa o tempo médio de percurso em 8,2963 minutos, *ceteris paribus*. Por outro lado, um incremento de um quilômetro na distância a ser percorrida aumenta agora apenas 0,7972 minutos no tempo médio de percurso, *ceteris paribus*. A redução no valor estimado de  $\beta$  da variável *dist* ocorreu porque parte do comportamento desta variável está contemplada na própria variável *sem*. Em outras palavras, distâncias maiores são mais suscetíveis a uma quantidade maior de semáforos e, portanto, há uma correlação alta entre elas.

Segundo Kennedy (2008), Gujarati (2011) e Wooldridge (2012), a existência de altas correlações entre variáveis explicativas, conhecida por **multicolinearidade**, não afeta a intenção de elaboração de previsões. Gujarati (2011) ainda destaca que a existência de altas correlações entre variáveis explicativas não gera necessariamente estimadores ruins ou fracos e que a presença de multicolinearidade não significa que o modelo possua problemas. Discutiremos mais sobre a multicolinearidade na seção 12.3.2.

As equações que determinam os valores mínimo e máximo para o tempo de percurso, ao nível de confiança de 95%, são:

Tempo mínimo:

$$\hat{\text{tempo}}_{\min} = 1,2463 + 0,2619 \cdot \text{dist}_i + 2,8967 \cdot \text{sem}_i$$

Tempo máximo:

$$\hat{\text{tempo}}_{\max} = 15,0561 + 1,3325 \cdot \text{dist}_i + 13,6959 \cdot \text{sem}_i$$

### 12.2.6. Variáveis dummy em modelos de regressão

De acordo com Sharma (1996) e Fávero *et al.* (2009), a determinação do número de variáveis necessárias para a investigação de um fenômeno é direta e simplesmente igual ao número de variáveis utilizadas para mensurar as respectivas características. Entretanto, o procedimento para determinar o número de variáveis explicativas cujos dados estejam em escalas qualitativas é diferente.

Imagine, por exemplo, que desejamos estudar como se altera o comportamento de determinado fenômeno organizacional, como a lucratividade total, quando são consideradas, no mesmo banco de dados, empresas provenientes de diferentes setores. Ou, em outra situação, desejamos verificar se o tíquete médio de compras realizadas em supermercados apresenta diferenças significativas ao compararmos consumidores provenientes de diferentes sexos e faixas de idade. Num terceira situação, desejamos estudar como se comportam as taxas de crescimento do PIB de diferentes países considerados emergentes e desenvolvidos. Em todas estas hipotéticas situações, as variáveis dependentes são quantitativas (lucratividade total, tíquete médio ou taxa de crescimento do PIB), porém desejamos saber como estas se comportam em função de variáveis explicativas qualitativas (setor, sexo, faixa de idade, classificação do país) que serão incluídas do lado direito dos respectivos modelos de regressão a serem estimados.

Não podemos simplesmente atribuir valores a cada uma das categorias da variável qualitativa, pois isso seria um erro grave, denominado de **ponderação arbitrária**, uma vez que estariam supondo que as diferenças na variável dependente seriam previamente conhecidas e de magnitudes iguais às diferenças dos valores atribuídos a cada uma das categorias da variável explicativa qualitativa. Nestas situações, a fim de que este problema seja completamente eliminado, devemos recorrer ao artifício das **variáveis dummy**, ou **binárias**, que assumem valores iguais a 0 ou 1, de forma a estratificar a amostra da maneira que for definido determinado critério, evento ou atributo, para, aí assim, serem incluídas no modelo em análise. Até mesmo um determinado período (dia, mês ou ano) em que ocorre um importante evento pode ser objeto de análise.

As variáveis *dummy* devem, portanto, ser utilizadas quando desejarmos estudar a relação entre o comportamento de determinada variável explicativa qualitativa e o fenômeno em questão, representado pela variável dependente.

Voltando ao nosso exemplo, imagine agora que o professor também tenha perguntado aos estudantes em que período do dia vieram à escola, ou seja, se cada um deles veio de manhã, a fim de ficar estudando na biblioteca, ou

se veio apenas no final da tarde para a aula noturna. A intenção do professor agora é saber se o tempo de percurso até a escola sofre variação em função da distância percorrida, da quantidade de semáforos e também do período do dia em que os estudantes se deslocam para chegar até a escola. Portanto, uma nova variável foi acrescentada ao banco de dados, conforme mostra a Tabela 12.8.

Devemos, portanto, definir qual das categorias da variável qualitativa será a referência (*dummy* = 0). Como, neste caso, temos somente duas categorias (manhã ou tarde), apenas uma única variável *dummy* deverá ser criada, em que a categoria de referência assumirá valor 0 e a outra categoria, valor 1. Este procedimento permitirá ao pesquisador estudar as diferenças que acontecem na variável *Y* ao se alterar a categoria da variável qualitativa, uma vez que o  $\beta$  desta *dummy* representará exatamente a diferença que ocorre no comportamento da variável *Y* quando se passa da categoria de referência da variável qualitativa para a outra categoria, estando o comportamento da categoria de referência representado pelo intercepto  $\alpha$ . Portanto, a decisão de escolha sobre qual será a categoria de referência é do próprio pesquisador e os parâmetros do modelo serão obtidos com base no critério adotado.

Desta forma, o professor decidiu que a categoria de referência será o período da tarde, ou seja, as células do banco de dados com esta categoria assumirão valores iguais a 0. Logo, as células com a categoria *manhã* assumirão valores iguais a 1. Isso porque o professor deseja avaliar se a ida à escola no período da manhã traz algum benefício ou prejuízo de tempo em relação ao período da tarde, que é imediatamente anterior à aula. Chamaremos esta *dummy* de variável *per*. Assim sendo, o banco de dados passa a ficar de acordo com o apresentado na Tabela 12.9.

Portanto, o novo modelo passa a ser:

$$\text{tempo}_i = \alpha + b_1 \cdot \text{dist}_i + b_2 \cdot \text{sem}_i + b_3 \cdot \text{per}_i + u_i$$

**Tabela 12.8** Exemplo: tempo de percurso x distância percorrida, quantidade de semáforos e período do dia para o trajeto até a escola.

Estudante	Tempo para chegar à escola (minutos) ( $Y_i$ )	Distância percorrida até a escola (quilômetros) ( $X_{1i}$ )	Quantidade de semáforos ( $X_{2i}$ )	Período do dia ( $X_{3i}$ )
Gabriela	15	8	0	Manhã
Dalila	20	6	1	Manhã
Gustavo	20	15	0	Manhã
Letícia	40	20	1	Tarde
Luiz Ovídio	50	25	2	Tarde
Leonor	25	11	1	Manhã
Ana	10	5	0	Manhã
Antônio	55	32	3	Tarde
Júlia	35	28	1	Manhã
Mariana	30	20	1	Manhã

**Tabela 12.9** Substituição das categorias da variável qualitativa pela *dummy*.

Estudante	Tempo para chegar à escola (minutos) ( $Y_i$ )	Distância percorrida até a escola (quilômetros) ( $X_{1i}$ )	Quantidade de semáforos ( $X_{2i}$ )	Período do dia <i>dummy per</i> ( $X_{3i}$ )
Gabriela	15	8	0	1
Dalila	20	6	1	1
Gustavo	20	15	0	1
Letícia	40	20	1	0
Luiz Ovídio	50	25	2	0
Leonor	25	11	1	1
Ana	10	5	0	1
Antônio	55	32	3	0
Júlia	35	28	1	1
Mariana	30	20	1	1

Analogamente ao apresentado para a regressão simples, temos, portanto, que:

$$\hat{tempo}_i = \alpha + \beta_1 \cdot dist_i + \beta_2 \cdot sem_i + \beta_3 \cdot per_i$$

em que  $\alpha, \beta_1, \beta_2$  e  $\beta_3$  são, respectivamente, as estimativas dos parâmetros  $a, b_1, b_2$  e  $b_3$ .

Resolvendo novamente pelo Excel, devemos agora incluir a variável *dummy per* no vetor de variáveis explicativas, conforme mostra a Figura 12.29 (arquivo **Tempodistsemper.xls**).

Os *outputs* são apresentados na Figura 12.30.

Por meio destes *outputs*, podemos, inicialmente, verificar que o coeficiente de ajuste  $R^2$  subiu para 0,9839, o que nos permite dizer que mais de 98% do comportamento de variação do tempo para se chegar à escola é explicado pela variação conjunta das três variáveis  $X$  (*dist, sem* e *per*). Além disso, este modelo é preferível em relação aos anteriormente estudados, uma vez que apresenta maior  $R^2$  ajustado.

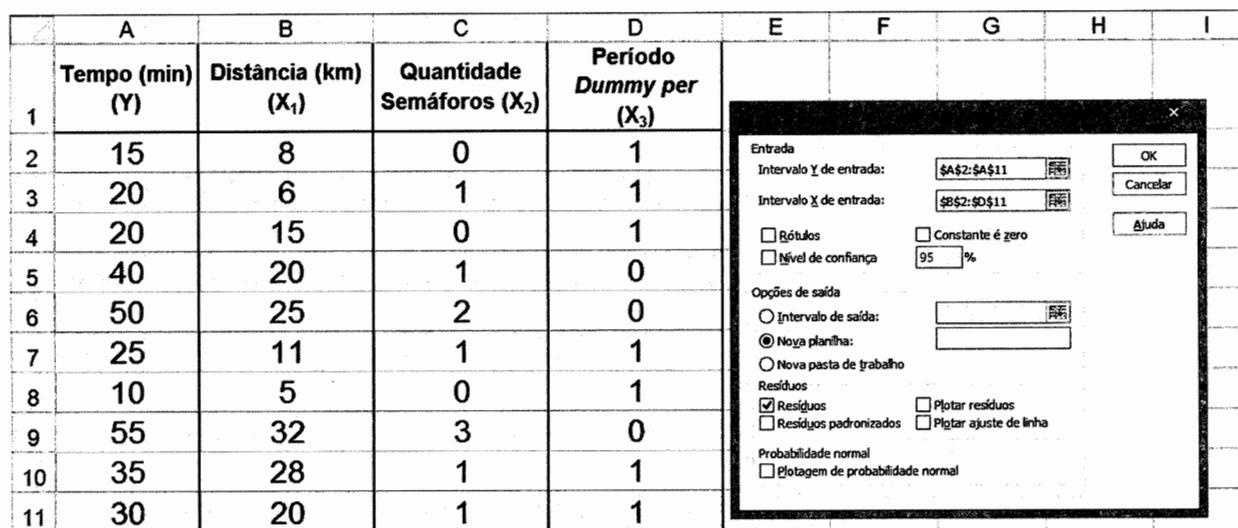


Figura 12.29 Regressão linear múltipla – seleção conjunta das variáveis explicativas com *dummy*.

A figura mostra uma tabela de resultados da regressão linear múltipla no Excel, com 24 linhas de dados:

	A	B	C	D	E	F	G	H	I
1	RESUMO DOS RESULTADOS								
2									
3	<i>Estatística de regressão</i>								
4	R múltiplo	0,99192476							
5	R-Quadrado	0,98391472							
6	R-quadrado ajustado	0,97587208							
7	Erro padrão	2,31554735							
8	Observações	10							
9									
10	ANOVA								
11		gl	SQ	MQ	F	F de significação			
12	Regressão	3	1967,82944	655,943148	122,337293	9,04894E-06			
13	Resíduo	6	32,1705573	5,36175955					
14	Total	9	2000						
15									
16		Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores	Inferior 95,0%	Superior 95,0%
17	Interseção	19,6352768	3,18782198	6,15946465	0,00084014	11,8349574	27,43559611	11,8349574	27,43559611
18	Variável X 1	0,70844841	0,12578944	5,63201834	0,00134086	0,400652751	1,016244075	0,400652751	1,016244075
19	Variável X 2	5,25727366	1,4478751	3,63102706	0,0109517	1,714450931	8,800096381	1,714450931	8,800096381
20	Variável X 3	-9,9088192	2,37945112	-4,16432979	0,00591578	-15,73112633	-4,086512054	-15,73112633	-4,086512054
21									
22									
23									
24	RESULTADOS DE RESÍDUOS								
25									
26	Observação	Y prevista	Resíduos						
27	1	15,3940449	-0,39404487						
28	2	19,2344217	0,7655783						
29	3	20,3531838	-0,35318376						
30	4	39,0615187	0,93848133						
31	5	47,8610344	2,13896561						
32	6	22,7766638	2,22333623						
33	7	13,2686996	-3,26869963						
34	8	58,0774469	-3,07744694						
35	9	34,8202868	0,17971322						
36	10	29,1526995	0,84730052						

Figura 12.30 Outputs da regressão linear múltipla com *dummy* no Excel.

Enquanto o teste  $F$  nos permite afirmar que pelo menos um parâmetro estimado  $\beta$  é estatisticamente diferente de zero ao nível de significância de 5%, os testes  $t$  de cada parâmetro mostram que todos eles ( $\beta_1, \beta_2, \beta_3$  e o próprio  $\alpha$ ) são estatisticamente diferentes de zero a este nível de significância, pois cada  $valor-P < 0,05$ . Assim, nenhuma variável  $X$  precisa ser excluída da modelagem e a equação final que estima o tempo para se chegar à escola apresenta-se da seguinte forma:

$$\hat{tempo}_i = 19,6353 + 0,7084.dist_i + 5,2573.sem_i - 9,9088.per_i_{\substack{\{tarde=0 \\ \{manh\aa=1\}}}}$$

Desta forma, podemos afirmar, para o nosso exemplo, que o tempo médio previsto para se chegar à escola é de 9,9088 minutos a menos para os alunos que optarem por ir no período da manhã em relação àqueles que optarem por ir à tarde, *ceteris paribus*. Isso provavelmente deve ter acontecido por motivos associados ao trânsito, porém estudos mais aprofundados poderiam ser elaborados neste momento. Assim, o professor propôs mais um exercício: **qual o tempo estimado para se chegar à escola por parte de um aluno que se desloca 17 quilômetros, passa por dois semáforos e vem à escola pouco antes do início da aula noturna, ou seja, no período da tarde?** A solução encontra-se a seguir:

$$\hat{tempo} = 19,6353 + 0,7084.(17) + 5,2573.(2) - 9,9088.(0) = 42,1934 \text{ min}$$

Ressalta-se que eventuais diferenças a partir da terceira casa decimal podem ocorrem por problemas de arredondamento. Utilizamos aqui os próprios valores obtidos nos *outputs* do Excel.

**E qual seria o tempo estimado para outro aluno que também se desloca 17 quilômetros, passa também por dois semáforos, porém decide ir à escola de manhã?**

$$\hat{tempo} = 19,6353 + 0,7084.(17) + 5,2573.(2) - 9,9088.(1) = 32,2846 \text{ min}$$

Conforme já discutimos, a diferença entre estas duas situações é capturada pelo  $\beta_3$  da variável *dummy*. A condição *ceteris paribus* impõe que nenhuma outra alteração seja considerada, exatamente como mostrado neste último exercício.

Imagine agora que o professor, ainda não satisfeito, tenha realizado um último questionamento aos estudantes, referente ao estilo de direção. Assim, perguntou como cada um se considera em termos de **perfil ao volante: calmo, moderado ou agressivo**. Ao obter as respostas, montou o último banco de dados, apresentado na Tabela 12.10.

Para elaborar a regressão, o professor precisa transformar a variável *perfil ao volante* em *dummies*. Para a situação em que houver um número de categorias maior do que 2 para determinada variável qualitativa (por exemplo, estado civil, time de futebol, religião, setor de atuação, entre outros exemplos), é necessário que o pesquisador utilize um número maior de variáveis *dummy* e, de maneira geral, para uma variável qualitativa com  $n$  categorias serão necessárias  $(n - 1)$  *dummies*, uma vez que determinada categoria deverá ser escolhida como referência e seu comportamento será capturado pelo parâmetro estimado  $\alpha$ .

**Tabela 12.10** Exemplo: tempo de percurso x distância percorrida, quantidade de semáforos, período do dia para o trajeto até a escola e perfil ao volante.

Estudante	Tempo para chegar à escola (minutos) ( $Y_i$ )	Distância percorrida até a escola (quilômetros) ( $X_{1i}$ )	Quantidade de semáforos ( $X_{2i}$ )	Período do dia ( $X_{3i}$ )	Perfil ao volante ( $X_{4i}$ )
Gabriela	15	8	0	manhã	calmo
Dalila	20	6	1	manhã	moderado
Gustavo	20	15	0	manhã	moderado
Letícia	40	20	1	tarde	agressivo
Luiz Ovídio	50	25	2	tarde	agressivo
Leonor	25	11	1	manhã	moderado
Ana	10	5	0	manhã	calmo
Antônio	55	32	3	tarde	calmo
Júlia	35	28	1	manhã	moderado
Mariana	30	20	1	manhã	moderado

Conforme discutimos, infelizmente é bastante comum que encontremos na prática procedimentos que substituam arbitrariamente as categorias de variáveis qualitativas por valores como 1 e 2, quando houver duas categorias, 1, 2 e 3, quando houver três categorias e assim sucessivamente. **Isso é um erro grave**, uma vez que, desta forma, partirmos do pressuposto de que as diferenças que ocorrem no comportamento da variável Y ao alterarmos a categoria da variável qualitativa seriam sempre de mesma magnitude, o que não necessariamente é verdade. Em outras palavras, não podemos presumir que a diferença média no tempo de percurso entre os indivíduos calmos e moderados será a mesma que entre os moderados e os agressivos.

No nosso exemplo, portanto, a variável *perfil ao volante* deverá ser transformada em duas *dummies* (variáveis *perfil2* e *perfil3*), já que definiremos a categoria *calmo* como sendo a referência (comportamento presente no intercepto). Enquanto a Tabela 12.11 apresenta os critérios para a criação das duas *dummies*, a Tabela 12.12 mostra o banco de dados final a ser utilizado na regressão.

E, desta forma, o modelo terá a seguinte equação:

$$\text{tempo}_i = a + b_1 \cdot \text{dist}_i + b_2 \cdot \text{sem}_i + b_3 \cdot \text{per}_i + b_4 \cdot \text{perfil2}_i + b_5 \cdot \text{perfil3}_i + u_i$$

e, analogamente ao apresentado para os modelos anteriores, temos que:

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

em que  $\alpha, \beta_1, \beta_2, \beta_3, \beta_4$  e  $\beta_5$  são, respectivamente, as estimativas dos parâmetros  $a, b_1, b_2, b_3, b_4$  e  $b_5$ .

Desta forma, analisando os parâmetros das variáveis *perfil2* e *perfil3*, temos que:

$\beta_4$  = diferença média no tempo de percurso entre um indivíduo considerado moderado e um indivíduo considerado calmo.

$\beta_5$  = diferença média no tempo de percurso entre um indivíduo considerado agressivo e um indivíduo considerado calmo.

$(\beta_5 - \beta_4)$  = diferença média no tempo de percurso entre um indivíduo considerado agressivo e um indivíduo considerado moderado.

**Tabela 12.11** Critérios para a criação das duas variáveis *dummy* a partir da variável qualitativa *perfil ao volante*.

Categoria da variável qualitativa <i>perfil ao volante</i>	Variável <i>dummy perfil2</i>	Variável <i>dummy perfil3</i>
Calm	0	0
Moderado	1	0
Agressivo	0	1

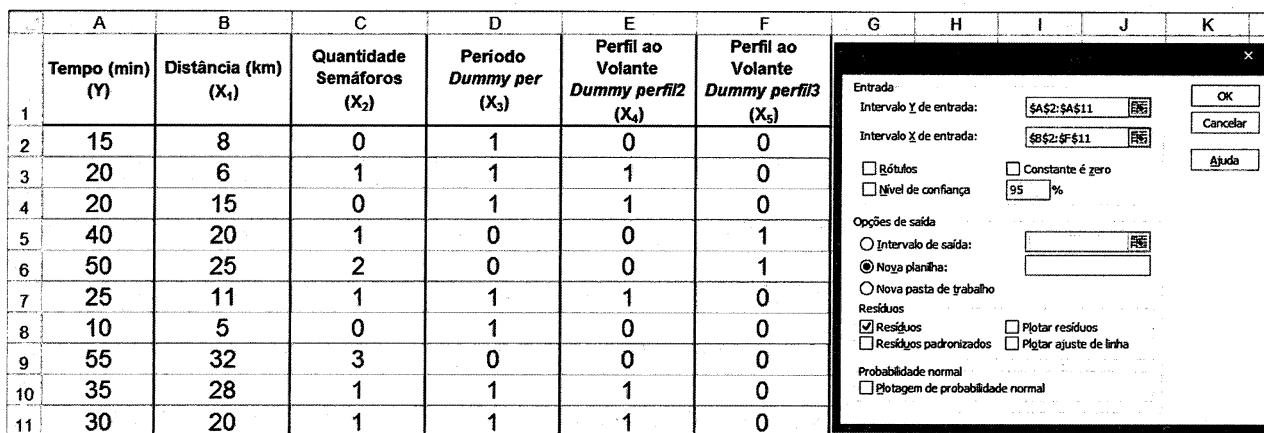
**Tabela 12.12** Substituição das categorias das variáveis qualitativas pelas respectivas variáveis *dummy*.

Estudante	Tempo para chegar à escola (minutos) ( $Y_i$ )	Distância percorrida até a escola (quilômetros) ( $X_{1i}$ )	Quantidade de semáforos ( $X_{2i}$ )	Período do dia <i>Dummy per</i> ( $X_{3i}$ )	Perfil ao Volante <i>Dummy perfil2</i> ( $X_{4i}$ )	Perfil ao Volante <i>Dummy perfil3</i> ( $X_{5i}$ )
Gabriela	15	8	0	1	0	0
Dalila	20	6	1	1	1	0
Gustavo	20	15	0	1	1	0
Letícia	40	20	1	0	0	1
Luiz Ovídio	50	25	2	0	0	1
Leonor	25	11	1	1	1	0
Ana	10	5	0	1	0	0
Antônio	55	32	3	0	0	0
Júlia	35	28	1	1	1	0
Mariana	30	20	1	1	1	0

Resolvendo novamente pelo Excel, devemos agora incluir as variáveis *dummy perfil2* e *perfil3* no vetor de variáveis explicativas. A Figura 12.31 mostra este procedimento, elaborado por meio do arquivo **Tempodistsempperfil.xls**. Os *outputs* são apresentados na Figura 12.32.

Podemos agora notar que, embora o coeficiente de ajuste do modelo  $R^2$  tenha sido muito elevado ( $R^2 = 0,9969$ ), os parâmetros das variáveis referentes ao período em que o trajeto foi efetuado ( $X_3$ ) e à categoria *moderado* da variável *perfil ao volante* ( $X_4$ ) não se mostraram estatisticamente diferentes de zero ao nível de significância de 5%. Desta forma, tais variáveis serão retiradas da análise e o modelo será elaborado novamente.

Entretanto, é importante analisarmos que, na presença das demais variáveis, o tempo do percurso até a escola passa a não apresentar mais diferenças se o percurso for realizado de manhã ou à tarde. O mesmo vale em relação ao perfil ao volante, já que se percebe que não há diferenças estatisticamente significantes no tempo de percurso para estudantes com perfil moderado em relação àqueles que se julgam calmos. Ressalta-se, numa regressão



**Figura 12.31** Regressão linear múltipla – seleção conjunta das variáveis explicativas com todas as *dummies*.

A	B	C	D	E	F	G	H	I
1	RESUMO DOS RESULTADOS							
2								
3	<i>Estatística de regressão</i>							
4	R múltiplo	0,998488967						
5	R-Quadrado	0,998980217						
6	R-quadrado ajustado	0,993205489						
7	Erro padrão	1,228776327						
8	Observações	10						
9								
10	ANOVA							
11	<i>gl</i>	<i>SQ</i>	<i>MQ</i>	<i>F</i>	<i>F de significação</i>			
12	Regressão	5	1993,960435	398,792087	264,11974	3,97756E-05		
13	Resíduo	4	6,039566047	1,50989126				
14	Total	9	2000					
15								
16	<i>Coeficientes</i>	<i>Erro padrão</i>	<i>Staf t</i>	<i>valor-P</i>	<i>95% inferiores</i>	<i>95% superiores</i>	<i>Inferior 95,0%</i>	<i>Superior 95,0%</i>
17	Interseção	13,49010874	3,8600886236	3,49404461	0,02503087	2,770570048	24,20964743	2,770570048
18	Variável X 1	0,874048902	0,07171531	9,39892617	0,00071409	0,474933281	0,873160522	0,474933281
19	Variável X 2	6,846796803	1,09498739	6,07086934	0,00371883	3,8068957826	9,686635981	3,8068957826
20	Variável X 3	-5,3714136	3,778780741	-1,42146739	0,22823382	-15,86299089	5,120163692	-15,86299089
21	Variável X 4	1,779116992	1,441459897	1,23424662	0,28466664	-2,223017254	5,781251238	-2,223017254
22	Variável X 5	6,37364077	2,243105048	2,84143659	0,04679951	0,145782739	12,6014988	0,145782739
23								
24								
25								
26	RESULTADOS DE RESÍDUOS							
27								
28	<i>Observação</i>	<i>Y prevista</i>	<i>Resíduos</i>					
29	1	13,51107035	1,488929648					
30	2	20,58889034	-0,58889034					
31	3	20,00851566	-0,00851566					
32	4	39,99148434	0,00851566					
33	5	50,00851566	-0,00851566					
34	6	23,85912485	1,040875147					
35	7	11,48892965	-1,48892965					
36	8	55	0					
37	9	35,41792218	-0,41792218					
38	10	30,02554697	-0,02554697					

**Figura 12.32** Outputs da regressão linear múltipla com diversas *dummies* no Excel.

múltipla, que tão importante quanto a análise dos parâmetros estatisticamente significantes é a análise dos parâmetros que não se mostraram estatisticamente diferentes de zero.

O procedimento *Stepwise*, disponível no Stata, no SPSS e em diversos outros softwares de modelagem, apresenta a propriedade de automaticamente excluir as variáveis explicativas cujos parâmetros não se mostrarem estatisticamente diferentes de zero. Como o software Excel não possui esse procedimento, iremos manualmente excluir as variáveis *per* e *perfil2* e elaborar novamente a regressão. Os novos *outputs* estão apresentados na Figura 12.33. Recomenda-se, todavia, que o pesquisador sempre tome bastante cuidado com a exclusão manual simultânea de variáveis cujos parâmetros, num primeiro momento, não se mostrarem estatisticamente diferentes de zero, uma vez que determinado parâmetro  $\beta$  pode tornar-se estatisticamente diferente de zero, mesmo que inicialmente não o fosse, ao se eliminar da análise outra variável cujo parâmetro  $\beta$  também não se mostre estatisticamente diferente de zero. Felizmente isso não ocorre neste exemplo e, assim, optamos por excluir as duas variáveis simultaneamente. Isto será comprovado quando elaborarmos esta regressão por meio do procedimento *Stepwise* nos softwares Stata (seção 12.5) e SPSS (seção 12.6).

E, dessa forma, o modelo final, com todos os parâmetros estatisticamente diferentes de zero ao nível de significância de 5%, com  $R^2 = 0,9954$  e com maior  $R^2$  ajustado entre todos aqueles discutidos ao longo do capítulo, passa a ser:

$$\hat{\text{tempo}}_i = 8,2919 + 0,7105 \cdot \text{dist}_i + 7,8368 \cdot \text{sem}_i + 8,9676 \cdot \text{perfil3}_{\begin{cases} i_{\text{calmo}}=0 \\ i_{\text{agressivo}}=1 \end{cases}}$$

É importante também verificarmos que houve uma redução das amplitudes dos intervalos de confiança para cada um dos parâmetros. Dessa forma, podemos perguntar:

**Qual seria o tempo estimado para outro aluno que também se desloca 17 quilômetros, passa também por dois semáforos, também decide ir à escola de manhã, porém tem um perfil considerado agressivo ao volante?**

$$\hat{\text{tempo}} = 8,2919 + 0,7105 \cdot (17) + 7,8368 \cdot (2) + 8,9676 \cdot (1) = 45,0109 \text{ min}$$

Por fim, podemos afirmar, *ceteris paribus*, que um estudante considerado agressivo ao volante leva, em média, 8,9676 minutos a mais para chegar à escola em relação a outro considerado calmo. Isso demonstra, entre outras coisas, que agressividade no trânsito realmente não leva a nada!

	A	B	C	D	E	F	G	H	I
1	RESUMO DOS RESULTADOS								
2									
3	<i>Estatística de regressão</i>								
4	R múltiplo	0,99770703							
5	R-Quadrado	0,99541932							
6	R-quadrado ajustado	0,99312897							
7	Erro padrão	1,23567574							
8	Observações	10							
9									
10	ANOVA								
11		gl	SQ	MQ	F	F de significação			
12	Regressão	3	1990,83863	663,612878	434,616052	2,0989E-07			
13	Resíduo	6	9,16136725	1,52689454					
14	Total	9	2000						
15									
16		Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores	Inferior 95,0%	Superior 95,0%
17	Interseção	8,29193164	0,85350815	9,71511709	6,8281E-05	10,38039085	6,203472428	10,38039085	
18	Variável X 1	0,7104531	0,06690063	10,6195276	4,1071E-05	0,546753153	0,874153047	0,546753153	0,874153047
19	Variável X 2	7,8368442	0,66940306	11,7072129	2,3427E-05	6,198873914	9,474814481	6,198873914	9,474814481
20	Variável X 5	8,96760731	1,02889043	8,71580395	0,0001261	6,450003119	11,48521151	6,450003119	11,48521151
21									
22									
23									
24	RESULTADOS DE RESÍDUOS								
25									
26	Observação	Y previsto	Resíduos						
27	1	13,9755564	1,02444356						
28	2	20,3914944	-0,39149444						
29	3	18,9487281	1,05127186						
30	4	39,3054452	0,69455485						
31	5	50,6945548	-0,69455485						
32	6	23,9437599	1,05624006						
33	7	11,8441971	-1,84419714						
34	8	54,5369634	0,46303657						
35	9	36,0214626	-1,02146264						
36	10	30,3378378	-0,33783784						

**Figura 12.33** Outputs da regressão linear múltipla após a exclusão de variáveis.

## 12.3. PRESSUPOSTOS DOS MODELOS DE REGRESSÃO POR MÍNIMOS QUADRADOS ORDINÁRIOS (MQO OU OLS)

Após a apresentação do modelo de regressão múltipla estimado pelo método de mínimos quadrados ordinários, o Quadro 12.2 traz os seus pressupostos, as consequências de suas violações e os procedimentos para a verificação de cada um deles.

Na sequência, iremos apresentar e discutir cada um dos pressupostos.

**Quadro 12.2** Pressupostos do modelo de regressão.

Pressuposto	Violão	Verificação do Pressuposto
Os resíduos apresentam distribuição normal.	Valor- <i>P</i> dos testes <i>t</i> e do teste <i>F</i> não são válidos.	Teste de Shapiro-Wilk. Teste de Shapiro-Francia.
Não existem correlações elevadas entre as variáveis explicativas e existem mais observações do que variáveis explicativas.	Multicolinearidade.	Matriz de Correlação Simples. Determinante da matriz ( $\mathbf{X}'\mathbf{X}$ ). <i>VIF</i> ( <i>Variance Inflation Factor</i> ) e <i>Tolerance</i> .
Os resíduos não apresentam correlação com qualquer variável <i>X</i> .	Heterocedasticidade.	Teste de Breusch-Pagan/ Cook-Weisberg.
Os resíduos são aleatórios e independentes.	Autocorrelação dos resíduos para modelos temporais.	Teste de Durbin-Watson. Teste de Breusch-Godfrey.

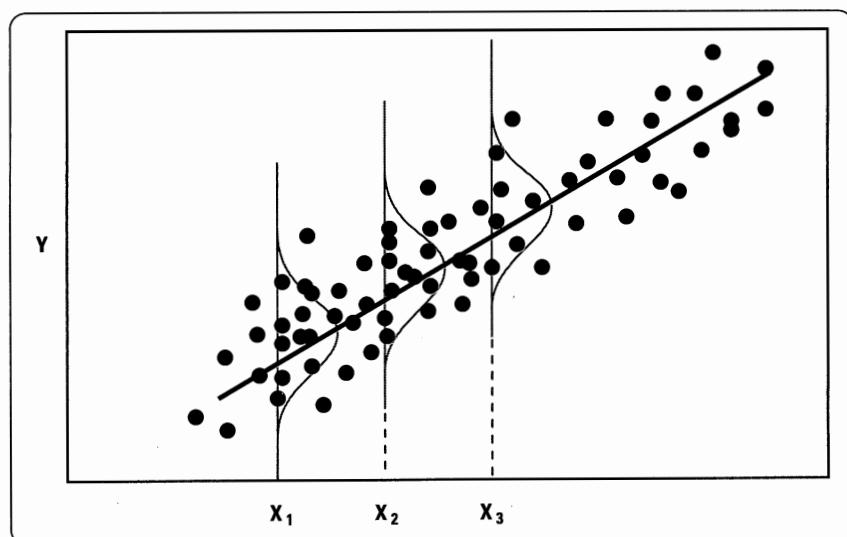
Fonte: Kennedy (2008).

### 12.3.1. Normalidade dos resíduos

A **normalidade dos resíduos** é requerida apenas e tão somente para que sejam validados os testes de hipótese dos modelos de regressão, ou seja, o pressuposto da normalidade assegura que o *valor-P* dos testes *t* e do teste *F* sejam válidos. Entretanto, Wooldridge (2012) argumenta que a violação deste pressuposto pode ser minimizada quando da utilização de grandes amostras, devido às propriedades assintóticas dos estimadores obtidos por mínimos quadrados ordinários.

É bastante comum que este pressuposto seja violado por pesquisadores quando da estimação de modelos de regressão pelo método de mínimos quadrados ordinários, porém é importante que esta hipótese possa ser atendida para a obtenção de uma série de resultados estatísticos voltados para a definição da melhor forma funcional do modelo e para a determinação dos intervalos de confiança para previsão (Figura 12.34), que são definidos, como já estudamos, com base na estimação dos parâmetros do modelo.

Ressalta-se que a aderência à distribuição normal da variável dependente, em modelos de regressão por mínimos quadrados ordinários, pode fazer com que sejam gerados termos de erro também normais e, consequentemente, estimados parâmetros mais adequados à determinação dos intervalos de confiança para efeitos de previsão.



**Figura 12.34** Distribuição normal dos resíduos.

Assim sendo, recomenda-se que seja aplicado, dependendo do tamanho da amostra, o **teste de Shapiro-Wilk** ou o **teste de Shapiro-Francia** aos termos de erro, a fim de que seja verificado o pressuposto da normalidade dos resíduos. Segundo Maroco (2014), enquanto o teste de Shapiro-Wilk é mais indicado para pequenas amostras (aqueles com até 30 observações), o teste de Shapiro-Francia é mais recomendado para grandes amostras, conforme discutimos no Capítulo 7.

Na seção 12.5 iremos apresentar a aplicação destes testes, bem como seus resultados, por meio da utilização do Stata.

A não aderência à normalidade dos termos de erro pode indicar que o modelo foi especificado incorretamente quanto à forma funcional e que houve a omissão de variáveis explicativas relevantes. A fim de que seja corrigido este problema, pode-se alterar a formulação matemática, bem como incluir novas variáveis explicativas no modelo.

Na seção 12.3.5 apresentaremos o *linktest* e o teste *RESET*, para identificação de problemas de especificação na forma funcional e de omissão de variáveis relevantes, respectivamente, e na seção 12.4 iremos discorrer sobre as especificações não lineares, com destaque para determinadas formas funcionais. Nesta mesma seção, discutiremos as transformações de Box-Cox, que têm por intuito maximizar a aderência à normalidade da distribuição de determinada variável gerada a partir de uma variável original com distribuição não normal. É muito comum que este procedimento seja aplicado à variável dependente de um modelo cuja estimação gerou termos de erro não aderentes à normalidade.

Vale a pena comentar que é comum que se discuta sobre a necessidade de que as variáveis explicativas apresentem distribuições aderentes à normalidade, o que é um grande erro. Se este fosse o caso, não seria possível utilizarmos variáveis *dummy* em nossos modelos.

### 12.3.2. O problema da multicolinearidade

O problema da **multicolinearidade** ocorre quando há correlações muito elevadas entre variáveis explicativas e, em casos extremos, tais correlações podem ser perfeitas, indicando uma relação linear entre as variáveis.

Inicialmente, apresentaremos o modelo geral de regressão linear múltipla na forma matricial. Partindo-se de:

$$Y_i = a + b_1 \cdot X_{1i} + b_2 \cdot X_{2i} + \dots + b_k \cdot X_{ki} + u_i \quad (12.30)$$

podemos escrever que:

$$\mathbf{Y} = \mathbf{X} \cdot \mathbf{b} + \mathbf{U} \quad (12.31)$$

ou:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix}_{nx1} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ 1 & X_{31} & X_{32} & \dots & X_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix}_{nxk+1} \cdot \begin{bmatrix} a \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix}_{k+1x1} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{bmatrix}_{nx1} \quad (12.32)$$

de onde se pode demonstrar que as estimativas dos parâmetros são dadas pelo seguinte vetor:

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) \quad (12.33)$$

Imaginemos um modelo específico com apenas duas variáveis explicativas, como segue:

$$Y_i = a + b_1 \cdot X_{1i} + b_2 \cdot X_{2i} + u_i \quad (12.34)$$

Se, por exemplo,  $X_{2i} = 4 \cdot X_{1i}$ , não seria possível separar as variações ocorridas na variável dependente em decorrência de alterações em  $X_1$  advindas da influência de  $X_2$ . Portanto, segundo Vasconcellos e Alves (2000), seria impossível, para esta situação, que fossem estimados todos os parâmetros da equação da expressão (12.34), já que ficaria impossibilitada a inversão da matriz  $(\mathbf{X}'\mathbf{X})$  e, consequentemente, o cálculo do vetor de parâmetros  $\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$ . Entretanto, poderia ser estimado o seguinte modelo:

$$Y_i = a + (b_1 + 4 \cdot b_2) \cdot X_{1i} + u_i \quad (12.35)$$

cujo parâmetro estimado seria uma combinação linear entre  $b_1$  e  $b_2$ .

Problemas maiores, entretanto, ocorrerão quando a correlação entre as variáveis explicativas for muito alta, porém não perfeita, conforme será discutido mais adiante por meio da apresentação de exemplos numéricos e de aplicações em bancos de dados.

### 12.3.2.1. Causas da multicolinearidade

Uma das principais causas da multicolinearidade é a existência de variáveis que apresentam a mesma tendência durante alguns períodos. Imaginemos, por exemplo, que se deseja estudar se a rentabilidade, ao longo do tempo, de um determinado fundo de renda fixa atrelado a índices de preços varia em função de índices de inflação com defasagem de três meses. Ou seja, há o intuito de se criar um modelo em que a rentabilidade do fundo em um período  $t$  seja função de determinados índices de inflação em  $t - 3$ . Para tanto, o pesquisador inclui, como variáveis explicativas, os índices IPCA e IGP-m (ambos em  $t - 3$ ). Como tais índices apresentam correlação ao longo do tempo, muito provavelmente o modelo gerado apresentará multicolinearidade.

Tal fenômeno não é restrito a bases de dados em que há a evolução temporal. Imaginemos outra situação em que um pesquisador deseja estudar se o faturamento de uma amostra de lojas de supermercados em um mês é função da área de vendas (em  $m^2$ ) e do número de funcionários alocados em cada uma das lojas. Como é sabido que, para este tipo de operação varejista, há certa correlação entre área de vendas e número de funcionários, problemas de multicolinearidade nesta *cross-section* também poderão acontecer.

Outra causa bastante comum da multicolinearidade é a utilização de bancos de dados com um número insuficiente de observações.

### 12.3.2.2. Consequências da multicolinearidade

Segundo Vasconcellos e Alves (2000), a existência de multicolinearidade tem impacto direto no cálculo da matriz  $(\mathbf{X}'\mathbf{X})$ . Para tratar deste problema, apresentaremos, por meio de exemplos numéricos, os cálculos das matrizes  $(\mathbf{X}'\mathbf{X})$  e  $(\mathbf{X}'\mathbf{X})^{-1}$  em três casos distintos, nos quais existe correlação entre as duas variáveis explicativas: (a) correlação perfeita; (b) correlação muito alta, porém não perfeita; (c) correlação baixa.

#### (a) Correlação perfeita

Imagine uma matriz  $\mathbf{X}$  com apenas duas variáveis explicativas e duas observações:

$$\mathbf{X} = \begin{bmatrix} 1 & 4 \\ 2 & 8 \end{bmatrix}$$

Logo:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 5 & 20 \\ 20 & 80 \end{bmatrix}$$

e, portanto,  $\det(\mathbf{X}'\mathbf{X}) = 0$ , ou seja,  $(\mathbf{X}'\mathbf{X})^{-1}$  não pode ser calculada.

#### (b) Correlação muito alta, porém não perfeita

Imagine agora que a matriz  $\mathbf{X}$  apresente os seguintes valores:

$$\mathbf{X} = \begin{bmatrix} 1 & 4 \\ 2 & 7,9 \end{bmatrix}$$

Logo:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 5 & 19,8 \\ 19,8 & 78,41 \end{bmatrix}$$

de onde vem que  $\det(\mathbf{X}'\mathbf{X}) = 0,01$  e, portanto:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 7,841 & -1,980 \\ -1,980 & 500 \end{bmatrix}$$

Segundo Vasconcellos e Alves (2000), como a matriz de variância e covariância dos parâmetros do modelo é dada por  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ , e como os elementos da diagonal principal desta matriz aparecem no denominador das estatísticas  $t$ , conforme estudado na seção 12.2.3 (expressão 12.21), estas tendem, neste caso, a apresentar valores subestimados pela existência de valores elevados na matriz  $(\mathbf{X}'\mathbf{X})^{-1}$ , o que pode eventualmente fazer com que um pesquisador considere não significantes os efeitos de algumas das variáveis explicativas. Porém, como os cálculos da estatística  $F$  e do  $R^2$  não são afetados por este fenômeno, é comum que se encontrem modelos em que os coeficientes das variáveis explicativas não sejam estatisticamente significantes, com o teste  $F$  rejeitando a hipótese nula ao mesmo nível de significância, ou seja, indicando que pelo menos um parâmetro seja estatisticamente diferente de zero. Em muitos casos, esta inconsistência ainda vem acompanhada de um alto valor de  $R^2$ .

### (c) Correlação baixa

Imagine, por fim, que a matriz  $\mathbf{X}$  passe a apresentar os seguintes valores:

$$\mathbf{X} = \begin{bmatrix} 1 & 4 \\ 2 & 3 \end{bmatrix}$$

Logo:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 5 & 10 \\ 10 & 25 \end{bmatrix}$$

de onde vem que  $\det(\mathbf{X}'\mathbf{X}) = 25$  e, portanto:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 1 & -0,4 \\ -0,4 & 0,2 \end{bmatrix}$$

Podemos agora verificar que, dada a baixa correlação entre  $X_1$  e  $X_2$ , os valores presentes na matriz  $(\mathbf{X}'\mathbf{X})^{-1}$  são baixos, o que gerará pouca influência para a redução da estatística  $t$  quando do seu cálculo.

Na seção 12.3.2.3, a seguir, serão elaborados modelos com o uso de bancos de dados que propiciam o estudo destas três situações.

### 12.3.2.3. Aplicação de exemplos com multicolinearidade no Excel

Voltando ao exemplo utilizado ao longo do capítulo, imaginemos agora que o professor deseja avaliar a influência da distância percorrida (*dist*) e da quantidade de cruzamentos (*cruz*) ao longo do trajeto sobre o tempo para se chegar à escola (*tempo*). Para tanto, fez uma pesquisa com alunos de três turmas diferentes (A, B e C), de modo que seja obtido, para cada turma, o seguinte modelo:

$$\text{tempo}_i = a + b_1 \cdot \text{dist}_i + b_2 \cdot \text{cruz}_i + u_i$$

Os três casos apresentados a seguir referem-se, respectivamente, aos dados obtidos em cada uma das três turmas de alunos.

#### (a) Turma A: O caso da correlação perfeita

A turma A tem alunos que moram apenas no centro da cidade, ou seja, coincidentemente existe uma relação perfeita entre a distância percorrida e a quantidade de cruzamentos, uma vez que os trajetos possuem as mesmas características e são sempre realizados em zona urbana. O banco de dados coletado na turma A está apresentado na Tabela 12.13.

Por meio do arquivo **Tempodistcruz\_turma\_A.xls**, podemos elaborar a regressão múltipla, conforme mostra a Figura 12.35. Os *outputs* são apresentados na Figura 12.36.

Conforme podemos verificar, a estimativa do parâmetro da variável  $X_1$  (*dist*) não foi calculada visto que a correlação entre *dist* e *cruz* é perfeita e, portanto, fica impossível a inversão da matriz  $(\mathbf{X}'\mathbf{X})$  que, neste caso, é dada por:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 3.704 & 7.408 \\ 7.408 & 14.816 \end{bmatrix}, \text{ de onde vem que } \det(\mathbf{X}'\mathbf{X}) = 0.$$

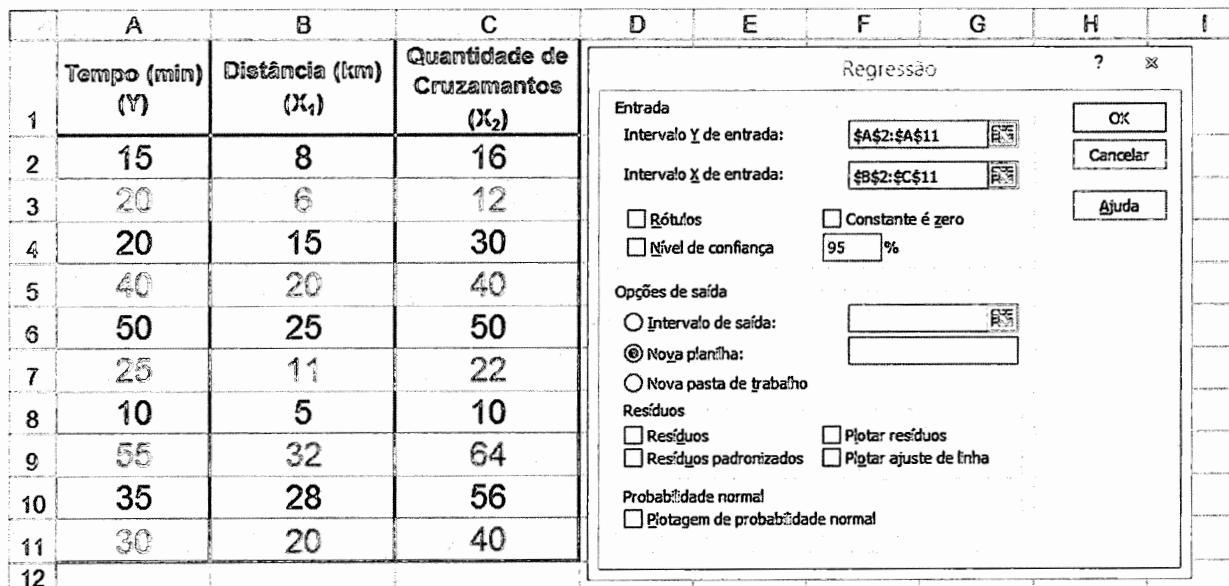
De qualquer modo, como sabemos que  $\text{cruz}_i = 2 \cdot \text{dist}_i$ , poderemos estimar o seguinte modelo:

$$\text{tempo}_i = a + (b_1 + 2 \cdot b_2) \cdot \text{dist}_i + u_i$$

em que o parâmetro estimado será uma combinação linear entre  $b_1$  e  $b_2$ .

**Tabela 12.13** Turma A e o exemplo de correlação perfeita entre as variáveis explicativas (distância percorrida e quantidade de cruzamentos).

Estudante	Tempo para chegar à escola (minutos) (Y <sub>i</sub> )	Distância percorrida até a escola (quilômetros) (X <sub>1i</sub> )	Quantidade de cruzamentos (X <sub>2i</sub> )
Gabriela	15	8	16
Dalila	20	6	12
Gustavo	20	15	30
Letícia	40	20	40
Luiz Ovídio	50	25	50
Leonor	25	11	22
Ana	10	5	10
Antônio	55	32	64
Júlia	35	28	56
Mariana	30	20	40



**Figura 12.35** Regressão linear múltipla para a turma A.

A	B	C	D	E	F	G	H	I
1	RESUMO DOS RESULTADOS							
2								
3	Estatística de regressão							
4	R múltiplo	0,905221341						
5	R-Quadrado	0,819425676						
6	R-quadrado ajustado	0,671853885						
7	Erro padrão	6,718897311						
8	Observações	10						
9								
10	ANOVA							
11	g/	SQ	MQ	F	F de significação			
12	Regressão	2	1638,851351	819,4256757	36,303087	0,000201618		
13	Resíduo	8	361,1486486	45,14358108				
14	Total	10	2000					
15								
16	Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores	Interior 95,0%	Superior 95,0%
17	Interseção	5,87837878	4,532327565	1,296988864	0,230788477	-4,573187721	16,32994448	-4,573187721
18	Variável X 1	0	0	65535	#NÚMI	0	0	0
19	Variável X 2	0,709459459	0,117748614	6,025204312	0,000314449	0,437930668	0,980988251	0,980988251

**Figura 12.36** Outputs da regressão linear múltipla para a turma A.

**(b) Turma B: O caso da correlação muito alta, porém não perfeita**

A turma B, muito parecida com a turma A em termos de características dos deslocamentos, possui apenas um estudante (Américo) que, por se deslocar por uma via expressa, passa por um cruzamento a menos, proporcionalmente, em relação aos demais, conforme pode ser observado na Tabela 12.14. Desta forma, a correlação entre *dist* e *cruz* passa a não ser mais perfeita, mesmo que ainda seja extremamente elevada (no caso deste exemplo, igual a 0,9998).

Por meio do arquivo **Tempodistcruz\_turma\_B.xls**, podemos elaborar a mesma regressão múltipla, cujos outputs são apresentados na Figura 12.37.

Neste caso, conforme já discutimos, é possível verificar que há uma inconsistência entre o resultado do teste *F* e os resultados dos testes *t*, já que estes últimos apresentam valores subestimados de suas estatísticas pelo fato de haver valores mais elevados na matriz  $(\mathbf{X}'\mathbf{X})^{-1}$ , ou seja, pelo fato de  $\det(\mathbf{X}'\mathbf{X})$  ser mais baixo. Neste caso, temos:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 3.704 & 7.388 \\ 7.388 & 14.737 \end{bmatrix}, \text{ de onde vem que } \det(\mathbf{X}'\mathbf{X}) = 3.304, \text{ que aparentemente é um valor alto, porém é}$$

consideravelmente mais baixo do que o calculado para o caso da turma C a seguir. Além disso, neste caso, temos que:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 4,460 & -2,236 \\ -2,236 & 1,121 \end{bmatrix}$$

Em decorrência disso, os outputs (Figura 12.37) podem fazer com que um pesquisador, erroneamente, afirme que nenhum parâmetro do modelo em questão seja estatisticamente significante, mesmo que o teste *F* tenha indicado que pelo menos um deles seja estatisticamente diferente de zero, ao nível de significância de, por exemplo, 5%, e que o próprio  $R^2$  tenha se mostrado relativamente alto ( $R^2 = 0,8379$ ). Este fenômeno representa o maior erro que se pode cometer em modelos com alta multicolinearidade entre variáveis explicativas.

**Tabela 12.14** Turma B e o exemplo de correlação muito alta entre as variáveis explicativas (distância percorrida e quantidade de cruzamentos).

Estudante	Tempo para chegar à escola (minutos) ( $Y_i$ )	Distância percorrida até a escola (quilômetros) ( $X_{1i}$ )	Quantidade de cruzamentos ( $X_{2i}$ )
Giulia	15	8	16
Luiz Felipe	20	6	12
Antonietta	20	15	30
Américo	40	20	39
Ferruccio	50	25	50
Filomena	25	11	22
Camilo	10	5	10
Guilherme	55	32	64
Maria Paula	35	28	56
Mateus	30	20	40

	A	B	C	D	E	F	G	H	I
1	RESUMO DOS RESULTADOS								
2									
3	<i>Estatística de regressão</i>								
4	R múltiplo	0,915411662							
5	R-Quadrado	0,83797851							
6	R-quadrado ajustado	0,791686656							
7	Erro padrão	6,803811742							
8	Observações	10							
9									
10	ANOVA								
11		g/	SQ	MQ	F	F de significação			
12	Regressão	2	1675,95702	837,9785102	18,10207269	0,001712002			
13	Resíduo	7	324,0429795	46,2918542					
14	Total	9	2000						
15									
16		Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores	Inferior 95,0%	Superior 95,0%
17	Interseção	5,637092316	4,597513664	1,226117578	0,259796941	-5,234299987	16,50848462	-5,234299987	16,50848462
18	Variável X 1	14,31661139	14,40799895	0,993657165	0,353488584	-19,75289233	48,3861151	-19,75289233	48,3861151
19	Variável X 2	-6,4607518	7,216310503	-0,89529848	0,40036477	-23,52461462	10,60311102	-23,52461462	10,60311102
20									
21									

**Figura 12.37** Outputs da regressão linear múltipla para a turma B.

**(c) Turma C: O caso da correlação mais baixa**

A turma C é mais heterogênea em termos de características dos deslocamentos, já que é formada por estudantes que também vêm de outros municípios e, portanto, utilizam estradas com uma quantidade proporcionalmente menor de cruzamentos ao longo do trajeto. A correlação entre *dist* e *cruz*, neste caso, passa a ser de 0,6505. A Tabela 12.15 apresenta o banco de dados coletado na turma C.

O arquivo **Tempodistcruz\_turma\_C.xls** traz os dados no formato do Excel, pelo qual podemos elaborar a mesma regressão múltipla, cujos *outputs* são apresentados na Figura 12.38.

Podemos agora verificar que, dada uma correlação mais baixa entre *dist* e *cruz*, os valores presentes na matriz  $(\mathbf{X}'\mathbf{X})^{-1}$  são bem mais baixos do que aqueles calculados para a turma B, o que gerará pouca influência para a redução das estatísticas *t* quando dos seus cálculos e, consequentemente, não ocorrerão inconsistências entre os resultados dos testes *t* e do teste *F*. Neste caso, temos:

$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 3.704 & 4.959 \\ 4.959 & 7.965 \end{bmatrix}$ , de onde vem que  $\det(\mathbf{X}'\mathbf{X}) = 4.910.679$ , que é um valor bem mais alto do que aquele calculado para o caso anterior. Além disso, temos que:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0,0016 & -0,0010 \\ -0,0010 & 0,0008 \end{bmatrix}$$

**Tabela 12.15** Turma C e o exemplo de correlação mais baixa entre as variáveis explicativas (distância percorrida e quantidade de cruzamentos).

Estudante	Tempo para chegar à escola (minutos) (Y <sub>i</sub> )	Distância percorrida até a escola (quilômetros) (X <sub>1i</sub> )	Quantidade de cruzamentos (X <sub>2i</sub> )
Juliana	15	8	12
Raquel	20	6	20
Larissa	20	15	25
Rogério	40	20	37
Isabel	50	25	32
Wilson	25	11	17
Luciana	10	5	9
Sandra	55	32	60
Oswaldo	35	28	12
Lucas	30	20	17

	A	B	C	D	E	F	G	H	I
1	RESUMO DOS RESULTADOS								
2									
3	<i>Estatística de regressão</i>								
4	R múltiplo	0,949472258							
5	R-Quadrado	0,901497568							
6	R-quadrado ajustado	0,873354016							
7	Erro padrão	5,305049665							
8	Observações	10							
9									
10	ANOVA								
11		gl	SQ	MQ	F	F de significação			
12	Regressão	2	1802,995136	901,4975682	32,03211768	0,000299961			
13	Resíduo	7	197,0048637	28,14355195					
14	Total	9	2000						
15									
16		Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores	Inferior 95,0%	Superior 95,0%
17	Interseção	3,663526099	3,694245028	0,991684653	0,354385408	-5,071975283	12,39902748	-5,071975283	12,39902748
18	Variável X 1	1,034262702	0,244832986	4,224360118	0,003915388	0,455324686	1,613200719	0,455324686	1,613200719
19	Variável X 2	0,363236845	0,150406681	2,415031322	0,046429664	0,00758156	0,718892129	0,00758156	0,718892129
20									
21									

**Figura 12.38** Outputs da regressão linear múltipla para a turma C.

#### 12.3.2.4. Diagnósticos de multicolinearidade

O primeiro e mais simples método para diagnóstico de multicolinearidade refere-se à identificação de altas correlações entre variáveis explicativas por meio da análise da matriz de correlação simples. Se, por um lado, este método apresenta uma grande facilidade de aplicação, por outro não consegue identificar eventuais relações existentes entre mais de duas variáveis simultaneamente.

O segundo método, menos utilizado, diz respeito ao estudo do determinante da matriz  $(\mathbf{X}'\mathbf{X})$ . Conforme estudamos nas duas seções anteriores, valores de  $\det(\mathbf{X}'\mathbf{X})$  muito baixos podem indicar a presença de altas correlações entre as variáveis explicativas, o que prejudica a análise das estatísticas  $t$ .

Por fim, mas não menos importante, é o diagnóstico de multicolinearidade elaborado por meio da estimação de regressões auxiliares. Segundo Vasconcellos e Alves (2000), a partir da expressão (12.30) podem ser estimadas regressões, de modo que:

$$\begin{aligned} X_{1i} &= a + b_1 \cdot X_{2i} + b_2 \cdot X_{3i} + \dots + b_{k-1} \cdot X_{ki} + u_i \\ X_{2i} &= a + b_1 \cdot X_{1i} + b_2 \cdot X_{3i} + \dots + b_{k-1} \cdot X_{ki} + u_i \\ &\vdots \\ X_{ki} &= a + b_1 \cdot X_{1i} + b_2 \cdot X_{2i} + \dots + b_{k-1} \cdot X_{ki-1} + u_i \end{aligned} \quad (12.36)$$

e, para cada uma delas, haverá um  $R_k^2$ . Se um ou mais destes  $R_k^2$  auxiliares for elevado, poderemos considerar a existência de multicolinearidade. Desta forma, podemos definir, a partir dos mesmos, as estatísticas **Tolerance** e **VIF** (*Variance Inflation Factor*), como segue:

$$Tolerance = 1 - R_k^2 \quad (12.37)$$

$$VIF = \frac{1}{Tolerance} \quad (12.38)$$

Assim sendo, se a *Tolerance* for muito baixa e, consequentemente, a estatística *VIF* alta, teremos um indício de que há problemas de multicolinearidade. Em outras palavras, se a *Tolerance* for baixa para determinada regressão auxiliar, significa que a variável explicativa que faz o papel de dependente nesta regressão auxiliar compartilha um percentual elevado de variância com as demais variáveis explicativas.

Enquanto muitos autores afirmam que problemas de multicolinearidade surgem com valores de *VIF* acima de 10, podemos perceber que um valor de *VIF* igual a 4 resulta em uma *Tolerance* de 0,25, ou seja, em um  $R_k^2$  de 0,75 para aquela determinada regressão auxiliar, o que representa um percentual relativamente elevado de variância compartilhada entre determinada variável explicativa e as demais.

#### 12.3.2.5. Possíveis soluções para o problema da multicolinearidade

A multicolinearidade representa um dos problemas mais difíceis de serem tratados em modelagem de dados. Enquanto alguns apenas aplicam o procedimento *Stepwise*, para que sejam eliminadas as variáveis explicativas que estão correlacionadas, o que de fato pode corrigir a multicolinearidade, tal solução pode criar um problema de especificação pela omissão de variável relevante, conforme discutiremos na seção 12.3.5.

A criação de fatores ortogonais a partir das variáveis explicativas, por meio da aplicação da técnica de análise fatorial, pode corrigir problemas de multicolinearidade. Para efeitos de previsão, entretanto, é sabido que os valores correspondentes aos fatores para novas observações não serão conhecidos, o que gera um problema para o pesquisador. Além disso, a criação de fatores sempre acarreta perda de uma parcela de variância das variáveis explicativas originais.

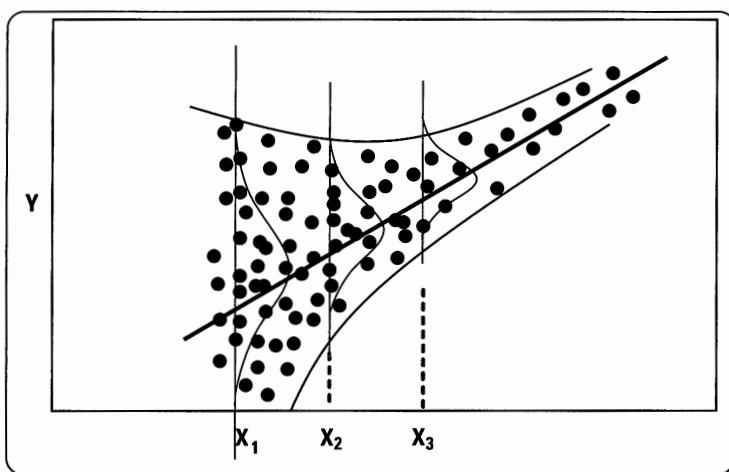
A boa notícia, conforme também discutem Vasconcellos e Alves (2000), é que a existência de multicolinearidade não afeta a intenção de elaboração de previsões, desde que as mesmas condições que geraram os resultados se mantenham para a previsão. Desta forma, as previsões incorporarão o mesmo padrão de relação entre as variáveis explicativas, o que não representa problema algum. Gujarati (2011) ainda destaca que a existência de altas correlações entre variáveis explicativas não gera necessariamente estimadores ruins ou fracos e que a presença de multicolinearidade não significa que o modelo possui problemas. Em outras palavras, alguns autores argumentam que uma solução para a multicolinearidade é identificá-la, reconhecê-la e não fazer nada.

### 12.3.3. O problema da heterocedasticidade

Além dos pressupostos discutidos anteriormente, a distribuição de probabilidades de cada termo aleatório de  $Y_i = a + b_1 \cdot X_{1i} + b_2 \cdot X_{2i} + \dots + b_k \cdot X_{ki} + u_i$  ( $i = 1, 2, \dots, n$ ) é tal que todas as distribuições devem apresentar a mesma variância, ou seja, devem ser homocedásticas. Assim:

$$\text{Var}(u_i) = E(u_i)^2 = \sigma_u^2 \quad (12.39)$$

A Figura 12.39 propicia, para um modelo de regressão linear simples, uma visualização do problema da **heterocedasticidade**, ou seja, a não constância da variância dos resíduos ao longo da variável explicativa. Em outras palavras, deve estar ocorrendo uma correlação entre os termos do erro e a variável  $X$ , percebida pela formação de um “cone” que se estreita à medida que  $X$  aumenta. Obviamente, o problema de heterocedasticidade também ocorreria se este “cone” se apresentasse de forma espelhada, ou seja, se o estreitamento (redução dos valores dos termos de erro) ocorresse com a redução dos valores da variável  $X$ .



**Figura 12.39** O problema da heterocedasticidade.

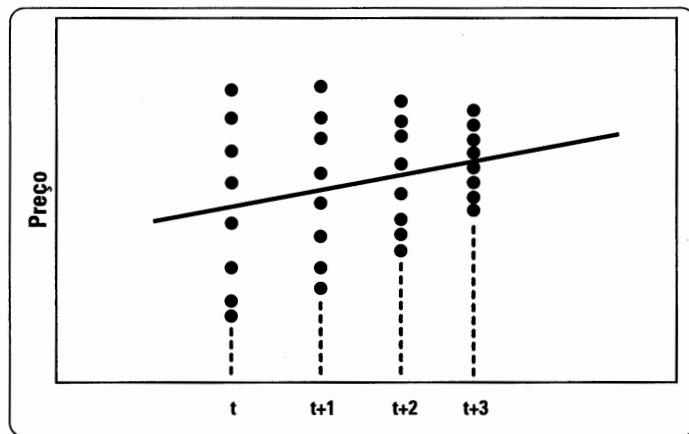
#### 12.3.3.1. Causas da heterocedasticidade

Segundo Vasconcellos e Alves (2000) e Greene (2012), erros de especificação quanto à forma funcional ou quanto à omissão de variável relevante podem gerar termos de erro heterocedásticos no modelo.

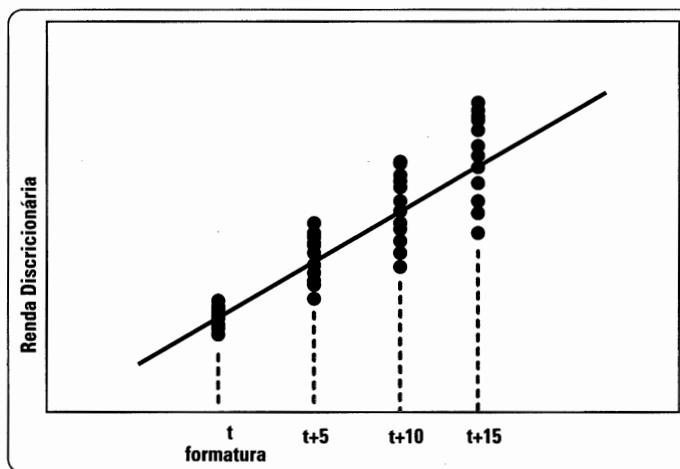
Este fenômeno também pode ser gerado por modelos de aprendizagem e erro. Neste caso, imaginemos que um grupo de analistas deseje elaborar previsões a respeito do preço futuro da soja no mercado de derivativos. Os mesmos analistas fazem suas previsões em  $t, t+1, t+2$  e  $t+3$  meses, a fim de que seja avaliada a curva de aprendizagem de cada um deles sobre o fenômeno em questão (precificação correta da *commodity*). O gráfico da Figura 12.40 é elaborado após o experimento e, por meio de sua análise, podemos verificar que os analistas passam a prever de forma mais apurada o preço da soja com o passar do tempo, muito provavelmente por conta do processo de aprendizagem a que são submetidos.

Analogamente, o incremento da renda discricionária (parcela da renda total de um indivíduo que não está comprometida, ou seja, que permite que o indivíduo possa exercer algum grau de discrição quanto ao seu destino) também pode fazer com que sejam gerados problemas de heterocedasticidade em modelos de regressão. Imaginemos uma pesquisa realizada com estudantes formados em um curso de Direito. De tempos em tempos, digamos de 5 em 5 anos, os mesmos estudantes são questionados sobre a sua renda discricionária naquele exato momento. O gráfico da Figura 12.41 é, então, elaborado e, por meio dele, verificamos que a renda discricionária dos estudantes passa a apresentar diferenças maiores ao longo do tempo, se comparadas àquelas dos tempos de recém-formados.

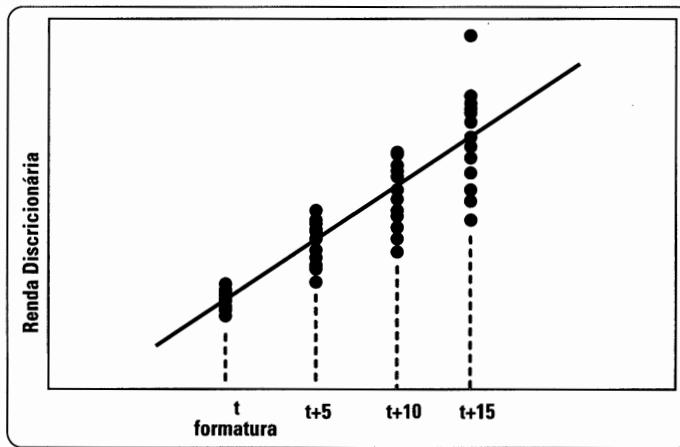
Ainda com base no mesmo exemplo da renda discricionária, imaginemos agora que outra amostra tenha a mesma configuração, porém com apenas um indivíduo apresentando valor discrepante de sua renda discricionária em  $t+15$ , conforme mostra a Figura 12.42. Este *outlier* aumentará ainda mais, neste caso, a intensidade da heterocedasticidade no modelo proposto.



**Figura 12.40** Modelos de aprendizagem e erro como causa da heterocedasticidade.



**Figura 12.41** Incremento da renda discricionária como causa da heterocedasticidade.



**Figura 12.42** Existência de *outlier* como causa da heterocedasticidade.

### 12.3.3.2. Consequências da heterocedasticidade

Todas as causas aqui discutidas (erros de especificação do modelo, modelos de aprendizagem e erro, aumento da renda discricionária e presença de *outliers*) podem levar à heterocedasticidade, que gera estimadores dos parâmetros não viesados, porém inefficientes, e erros-padrão dos parâmetros viesados, o que acarreta problemas com os testes de hipótese das estatísticas  $t$ .

A fim de que seja detectada a presença de heterocedasticidade, apresentaremos, na sequência, o teste de Breusch-Pagan/Cook-Weisberg. Alguns procedimentos para eventual correção da heterocedasticidade também serão discutidos, como a estimativa pelo método de mínimos quadrados ponderados e o método de Huber-White para erros-padrão robustos.

### 12.3.3.3. Diagnóstico de heterocedasticidade: teste de Breusch-Pagan/Cook-Weisberg

O teste de Breusch-Pagan/Cook-Weisberg, que se baseia no **multiplicador de Lagrange (LM)**, apresenta, como hipótese nula, o fato de a variância dos termos de erro ser constante (erros homocedásticos) e, como hipótese alternativa, o fato de a variância dos termos de erro não ser constante, ou seja, os termos de erro serem uma função de uma ou mais variáveis explicativas (erros heterocedásticos). É importante mencionar que este teste é indicado para os casos em que a suposição de normalidade dos resíduos for verificada.

Para obter o resultado do teste, podemos, inicialmente, elaborar um determinado modelo de regressão, a partir do qual vamos obter o vetor de resíduos ( $u_i$ ) e o vetor de valores previstos da variável dependente ( $\hat{Y}_i$ ). Na sequência, podemos padronizar os resíduos ao quadrado, obrigando que a média desta nova variável seja igual a 1. Ou seja, cada resíduo padronizado será obtido por meio da seguinte expressão:

$$up_i = \frac{u_i^2}{\left( \sum_{i=1}^n u_i^2 \right) / n} \quad (12.40)$$

em que  $n$  é o número de observações.

Em seguida, podemos elaborar a regressão  $up_i = a + b.\hat{Y}_i + \xi_i$ , a partir da qual se calcula a soma dos quadrados da regressão (SQR) que, dividindo-se por dois, chega-se à estatística  $\chi^2_{BP/CW}$ .

Assim sendo, o teste de Breusch-Pagan/Cook-Weisberg apresenta, como hipótese nula, o fato de a estatística calculada  $\chi^2_{BP/CW}$  possuir distribuição qui-quadrado com 1 grau de liberdade, ou seja, que  $\chi^2_{BP/CW} < \chi^2_{1 \text{ g.l.}}$  para determinado nível de significância. Em outras palavras, se os termos do erro forem homocedásticos, os resíduos ao quadrado não aumentam ou diminuem com o aumento de  $\hat{Y}$ .

Na seção 12.5, iremos apresentar a aplicação deste teste, bem como seus resultados, por meio da utilização do Stata.

### 12.3.3.4. Método de mínimos quadrados ponderados: uma possível solução

Conforme mencionamos, falhas na especificação do modelo podem gerar termos de erro heterocedásticos e, como sabemos e discutiremos na seção 12.4, as relações entre variáveis são complexas e nem sempre seguem uma linearidade. E não havendo determinada teoria subjacente que indique a relação entre duas ou mais variáveis, cabe ao pesquisador, por meio, por exemplo, da elaboração de gráficos dos resíduos em função da variável dependente ou das variáveis explicativas, tentar inferir sobre um eventual ajuste não linear a ser aplicado ao modelo em estudo, como o logarítmico, o quadrático ou o inverso.

Neste sentido, o **método de mínimos quadrados ponderados**, que é um caso particular do método de mínimos quadrados generalizados, pode ser aplicado quando se diagnostica que a variância dos termos de erro depende da variável explicativa, ou seja, quando a expressão (12.39) sofre alguma alteração, de modo que:

$$Var(u_i) = \sigma_u^2 \cdot X_i$$

ou

$$Var(u_i) = \sigma_u^2 \cdot X_i^2$$

ou

$$Var(u_i) = \sigma_u^2 \cdot \sqrt{X_i}$$

ou qualquer outra relação entre  $Var(u_i)$  e  $X_i$ .

Assim sendo, o modelo poderá ser transformado de maneira que os termos de erro passem a apresentar variância constante. Imagine, por exemplo, que a relação entre  $u_i$  e  $X_i$  seja linear, ou seja, que  $|u_i| = c \cdot X_i$ , e, desta forma,  $E(u_i)^2 = E(c \cdot X_i)^2 = c^2 \cdot X_i^2$ , em que  $c$  é uma constante. Isto posto, podemos propor um novo modelo, da seguinte forma:

$$\frac{Y_i}{X_i} = \frac{a}{X_i} + \frac{b \cdot X_i}{X_i} + \frac{u_i}{X_i} \quad (12.41)$$

A partir da expressão (12.41), temos que os novos termos de erro apresentam a seguinte variância:

$$E\left(\frac{u_i}{X_i}\right)^2 = \frac{1}{X_i^2} \cdot E(u_i)^2 = \frac{1}{X_i^2} \cdot c^2 \cdot X_i^2 = c^2, \text{ que é constante.}$$

Portanto, o modelo proposto por meio da expressão (12.41) pode ser estimado por mínimos quadrados ordinários.

#### 12.3.3.5. Método de Huber-White para erros-padrão robustos

Para termos uma sucinta ideia do procedimento proposto em seminal artigo escrito por White (1980), que segue o trabalho de Huber (1967), vamos novamente utilizar a expressão:

$$Y_i = a + b \cdot X_i + u_i, \text{ com } Var(u_i) = E(u_i)^2 = \sigma_u^2 \quad (12.42)$$

e

$$Var(\hat{b}) = \frac{\sum X_i^2 \cdot \sigma_u^2}{(\sum X_i^2)} \quad (12.43)$$

Porém, como  $\sigma_u^2$  não é diretamente observável, White (1980) propõe que se adote  $\hat{u}_i^2$ , em vez de  $\sigma_u^2$ , para a estimação de  $Var(\hat{b})$ , da seguinte maneira:

$$Var(\hat{b}) = \frac{\sum X_i^2 \cdot \hat{u}_i^2}{(\sum X_i^2)} \quad (12.44)$$

White (1980) demonstra que a  $Var(\hat{b})$  apresentada por meio da expressão (12.44) é um estimador consistente da variância apresentada por meio da expressão (12.43), ou seja, à medida que o tamanho da amostra aumenta indefinidamente, a segunda converge para a primeira.

Este procedimento pode ser generalizado para o modelo de regressão múltipla:

$$Y_i = a + b_1 \cdot X_{1i} + b_2 \cdot X_{2i} + \dots + b_k \cdot X_{ki} + u_i \quad (12.45)$$

de onde vem que:

$$Var(\hat{b}_j) = \frac{\sum \hat{w}_{ji}^2 \cdot \hat{u}_i^2}{(\sum \hat{w}_{ji}^2)^2} \quad (12.46)$$

em que  $j = 1, 2, \dots, k$ ,  $\hat{u}_i$  são os resíduos obtidos por meio da elaboração da regressão original e  $\hat{w}_{ji}$  representam os resíduos obtidos por meio da elaboração de cada regressão auxiliar do regressor  $X_j$  contra todos os demais regressores.

Dada a facilidade computacional de se aplicar este método, atualmente é muito frequente que os pesquisadores utilizem os erros-padrão robustos à heterocedasticidade em seus trabalhos acadêmicos, a tal ponto de nem mais se preocuparem em verificar a existência da própria heterocedasticidade. Entretanto, esta decisão, que acaba

por tentar eliminar uma incerteza correspondente à fonte da heterocedasticidade e que eventualmente gera uma eventual confiança em resultados mais robustos, não representa uma verdadeira solução na grande maioria das vezes. É importante salientar que este procedimento, que gera estimativas dos erros-padrão dos parâmetros diferentes daquelas que seriam obtidas com a aplicação direta do método de mínimos quadrados ordinários (afetando as estatísticas  $t$ ), não altera as estimativas dos parâmetros do modelo de regressão propriamente ditos.

Desta forma, a adoção deste procedimento pode apenas fazer com que o pesquisador finja que o problema não existe, ao invés de tentar identificar as razões por meio das quais ele surge.

#### 12.3.4. O problema da autocorrelação dos resíduos

A hipótese de aleatoriedade e independência dos termos de erro apenas faz sentido de ser estudada em modelos em que há a **evolução temporal dos dados**. Em outras palavras, se estivermos trabalhando com uma base de dados em *cross-section*, este pressuposto não se justifica, já que a mudança da sequência em que as observações estão dispostas numa *cross-section* não altera em nada o banco de dados, porém modifica a correlação entre os termos de erro de uma observação para a seguinte. Por outro lado, como devemos obrigatoriamente respeitar a sequência das observações em bancos de dados com evolução temporal ( $t, t+1, t+2$  etc.), a correlação ( $\rho$ ) dos termos de erro entre observações passa a fazer sentido. Dessa forma, podemos propor o seguinte modelo, agora com subscritos  $t$  em vez de  $i$ :

$$Y_t = a + b_1 \cdot X_{1t} + b_2 \cdot X_{2t} + \dots + b_k \cdot X_{kt} + \varepsilon_t \quad (12.47)$$

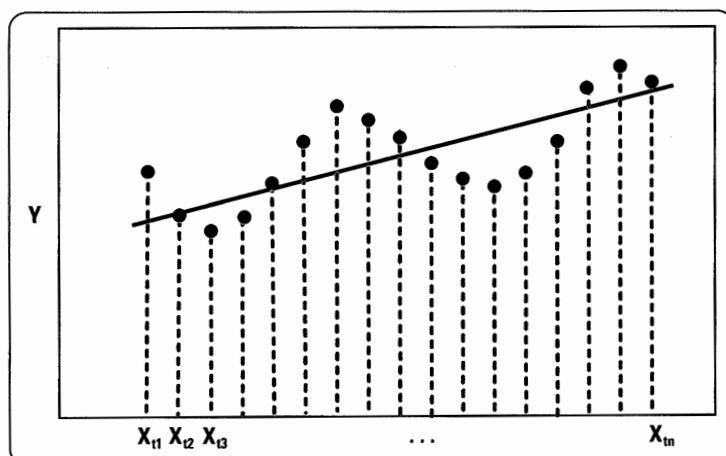
em que:

$$\varepsilon_t = \rho \cdot \varepsilon_{t-1} + u_t, \text{ com } -1 \leq \rho \leq 1 \quad (12.48)$$

Ou seja, os termos de erro  $\varepsilon_t$  não são independentes e, de acordo com a expressão (12.48), apresentam **autocorrelação de primeira ordem**, ou seja, cada valor de  $\varepsilon$  depende do valor de  $\varepsilon$  do período anterior e de um termo aleatório e independente  $u$ , com distribuição normal, média zero e variância constante. Neste caso, portanto, temos que:

$$\begin{aligned} \varepsilon_{t-1} &= \rho \cdot \varepsilon_{t-2} + u_{t-1} \\ \varepsilon_{t-2} &= \rho \cdot \varepsilon_{t-3} + u_{t-2} \\ &\vdots \\ \varepsilon_{t-p} &= \rho \cdot \varepsilon_{t-p-1} + u_{t-p} \end{aligned} \quad (12.49)$$

A Figura 12.43 propicia, para um modelo de regressão linear simples, uma visualização do problema da autocorrelação dos resíduos, ou seja, nitidamente os termos de erro não apresentam aleatoriedade e correlacionam-se temporalmente.



**Figura 12.43** O problema da autocorrelação dos resíduos.

### 12.3.4.1. Causas da autocorrelação dos resíduos

Segundo Vasconcellos e Alves (2000) e Greene (2012), erros de especificação quanto à forma funcional ou quanto à omissão de variável explicativa relevante podem gerar termos de erro autocorrelacionados. Além disso, a autocorrelação dos resíduos também pode ser causada por fenômenos sazonais e, consequentemente, pela desazonalização destas séries.

Imaginemos que um pesquisador deseja investigar a relação existente entre consumo de sorvete (em toneladas) em determinada cidade e o crescimento da população ao longo dos trimestres. Para tanto, coletou dados por 2 anos (8 trimestres) e elaborou o gráfico apresentado na Figura 12.44. Por meio deste gráfico, podemos perceber que o crescimento da população da cidade ao longo do tempo faz com que o consumo de sorvete aumente. Entretanto, por conta da sazonalidade que existe, já que o consumo de sorvete é maior em períodos de primavera e verão e menor em períodos de outono e inverno, a forma funcional linear (modelo dessazonalizado) faz com que sejam gerados termos de erro autocorrelacionados ao longo do tempo.

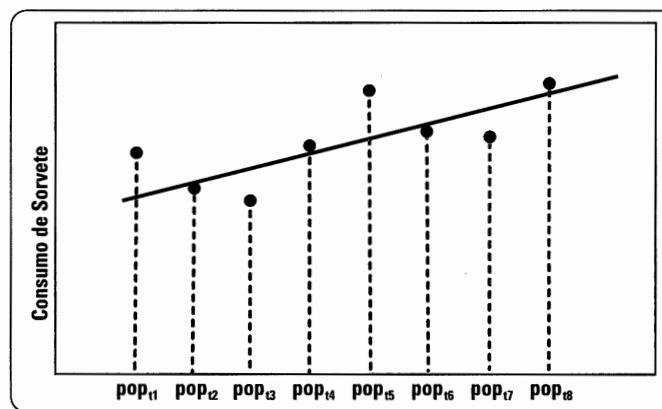


Figura 12.44 Sazonalidade como causa da autocorrelação dos resíduos.

### 12.3.4.2. Consequências da autocorrelação dos resíduos

Todas as causas aqui apresentadas (erros de especificação do modelo quanto à forma funcional, omissão de variável explicativa relevante e dessazonalização de séries) podem levar à autocorrelação dos resíduos, que gera estimadores dos parâmetros não viesados, porém ineficientes, e erros-padrão dos parâmetros subestimados, o que acarreta problemas com os testes de hipótese das estatísticas  $t$ .

A fim de que seja detectada a presença de autocorrelação dos resíduos, apresentaremos, a seguir, os testes de Durbin-Watson e de Breusch-Godfrey.

### 12.3.4.3. Diagnóstico de autocorrelação dos resíduos: teste de Durbin-Watson

O teste de Durbin-Watson é o mais utilizado por pesquisadores que têm a intenção de verificar a existência de autocorrelação dos resíduos, embora sua aplicação só seja válida para se testar a existência de autocorrelação de primeira ordem. A estatística do teste é dada por:

$$DW = \frac{\sum_{t=2}^n (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^n \varepsilon_t^2} \quad (12.50)$$

em que  $\varepsilon_t$  representa os termos de erro estimados para o modelo da expressão (12.47). Como sabemos que a correlação entre  $\varepsilon_t$  e  $\varepsilon_{t-1}$  é dada por:

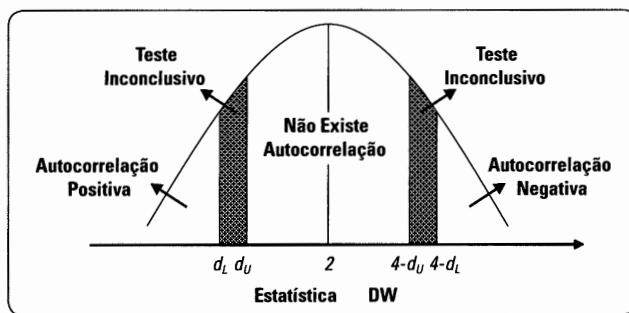
$$\rho = \frac{\sum_{t=2}^n \varepsilon_t \cdot \varepsilon_{t-1}}{\sum_{t=2}^n \varepsilon_{t-1}^2} \quad (12.51)$$

para valores de  $n$  suficientemente grandes, podemos deduzir que:

$$DW \equiv 2.(1 - \hat{\rho}) \quad (12.52)$$

e é por este motivo que muitos pesquisadores afirmam que um teste de Durbin-Watson com estatística  $DW$  aproximadamente igual a 2 resulta em inexistência de autocorrelação dos resíduos ( $\hat{\rho} \approx 0$ ). Embora isso seja verdade para processos autorregressivos de primeira ordem, uma tabela com valores críticos  $d_U$  e  $d_L$  da distribuição de  $DW$  pode oferecer ao pesquisador uma possibilidade mais concreta sobre a real existência de autocorrelação, já que oferece os valores de  $d_U$  e  $d_L$  em função do número de observações da amostra, do número de parâmetros do modelo e do nível de significância estatística que deseja o pesquisador. Enquanto a Tabela C do apêndice do livro traz estes valores críticos, a Figura 12.45 apresenta a distribuição de  $DW$  e os critérios para existência ou não de autocorrelação.

Embora bastante utilizado, o teste de Durbin-Watson, conforme já discutido, só é válido para verificação de existência de autocorrelação de primeira ordem dos termos de erro. Além disso, não é apropriado para modelos em que a variável dependente defasada é incluída como uma das variáveis explicativas. E é neste sentido que o teste de Breusch-Godfrey passa a ser uma alternativa bastante interessante.



**Figura 12.45** Distribuição de  $DW$  e critérios para existência de autocorrelação.

#### 12.3.4.4. Diagnóstico de autocorrelação dos resíduos: teste de Breusch-Godfrey

O teste de Breusch-Godfrey, originado por dois importantes artigos publicados individualmente em 1978 (Breusch, 1978; Godfrey, 1978) permite que se teste a existência de autocorrelação dos resíduos em um modelo que apresenta a variável dependente defasada como uma de suas variáveis explicativas. Além disso, também permite que o pesquisador verifique se a autocorrelação é de ordem 1, de ordem 2 ou de ordem  $p$ , sendo, portanto, mais geral do que o teste de Durbin-Watson.

Dado novamente o mesmo modelo de regressão linear múltipla:

$$Y_t = a + b_1 \cdot X_{1t} + b_2 \cdot X_{2t} + \dots + b_k \cdot X_{kt} + \varepsilon_t \quad (12.53)$$

podemos definir que os termos de erro sofrem um processo autorregressivo de ordem  $p$ , de modo que:

$$\varepsilon_t = \rho_1 \cdot \varepsilon_{t-1} + \rho_2 \cdot \varepsilon_{t-2} + \dots + \rho_p \cdot \varepsilon_{t-p} + u_t \quad (12.54)$$

em que  $u$  possui distribuição normal, média zero e variância constante.

Assim, por meio da estimação por mínimos quadrados ordinários do modelo representado pela expressão (12.53), podemos obter  $\hat{\varepsilon}_t$  e elaborar a seguinte regressão:

$$\hat{\varepsilon}_t = d_1 \cdot X_{1t} + d_2 \cdot X_{2t} + \dots + d_k \cdot X_{kt} + \hat{\rho}_1 \cdot \hat{\varepsilon}_{t-1} + \hat{\rho}_2 \cdot \hat{\varepsilon}_{t-2} + \dots + \hat{\rho}_p \cdot \hat{\varepsilon}_{t-p} + v_t \quad (12.55)$$

Breusch e Godfrey provam que a estatística do teste é dada por:

$$BG = (n - p) \cdot R^2 \sim \chi^2_p \quad (12.56)$$

em que  $n$  é o tamanho da amostra,  $p$  é a dimensão do processo autorregressivo e  $R^2$  é o coeficiente de ajuste obtido por meio da estimação do modelo da expressão (12.55). Desta forma, se  $(n - p) \cdot R^2$  for maior do que o

valor crítico da distribuição qui-quadrado com  $p$  graus de liberdade, rejeitamos a hipótese nula de inexistência de autocorrelação dos resíduos, ou seja, pelo menos um parâmetro  $\hat{\rho}$  na expressão (12.55) é estatisticamente diferente de zero.

A principal desvantagem do teste de Breusch-Godfrey é não permitir que se defina, *a priori*, o número de defasagens  $p$  na expressão (12.54), fazendo com que o pesquisador tenha que testar diversas possibilidades de  $p$ .

#### 12.3.4.5. Possíveis soluções para o problema da autocorrelação dos resíduos

A autocorrelação dos resíduos pode ser tratada pela alteração da forma funcional do modelo ou pela inclusão de variável relevante que havia sido omitida. Os testes para identificação destes problemas de especificação encontram-se na seção 12.3.5.

Entretanto, caso se chegue à conclusão de que a autocorrelação é considerada “pura”, ou seja, não advinda de problemas de especificação pela inadequada forma funcional ou pela omissão de variável relevante, pode-se tratar o problema por meio do método de mínimos quadrados generalizados, que tem por objetivo encontrar a melhor transformação do modelo original de modo a gerar termos de erro não autocorrelacionados.

Imaginemos novamente o nosso modelo original, porém com apenas uma variável explicativa. Assim:

$$Y_t = a + b \cdot X_t + \varepsilon_t \quad (12.57)$$

sendo:

$$\varepsilon_t = \rho \cdot \varepsilon_{t-1} + u_t \quad (12.58)$$

em que  $u$  possui distribuição normal, média zero e variância constante.

Como o nosso intuito é modificar o modelo da expressão (12.57), de modo que os termos de erro passem a ser  $u$ , e não mais  $\varepsilon$ , podemos multiplicar os termos desta expressão por  $\rho$  e defasá-los em 1 período. Assim, temos:

$$\rho \cdot Y_{t-1} = \rho \cdot a + \rho \cdot b \cdot X_{t-1} + \rho \cdot \varepsilon_{t-1} \quad (12.59)$$

Ao subtrairmos a expressão (12.59) da expressão (12.57), passamos a ter:

$$Y_t - \rho \cdot Y_{t-1} = a \cdot (1 - \rho) + b \cdot (X_t - \rho \cdot X_{t-1}) + u_t \quad (12.60)$$

que passa a ser um modelo com termos de erro não correlacionados. Para que seja feita esta transformação, é necessário, todavia, que o pesquisador conheça  $\rho$ .

Na seção 12.5, que traz a aplicação dos modelos de regressão múltipla por meio do software Stata, serão apresentados os procedimentos para verificação de cada um dos pressupostos, com os respectivos testes e resultados.

#### 12.3.5. Detecção de problemas de especificação: o *linktest* e o teste *RESET*

Como podemos perceber, grande parte das violações dos pressupostos em regressão é gerada por falhas de especificação do modelo, ou seja, por problemas na definição da forma funcional e por omissão de variáveis explicativas relevantes. Existem muitos métodos de detecção de problemas de especificação, porém os mais utilizados referem-se ao *linktest* e ao teste *RESET*.

O *linktest* nada mais é do que um procedimento que cria duas novas variáveis a partir da elaboração de um modelo de regressão, que nada mais são do que as variáveis  $\hat{Y}$  e  $\hat{Y}^2$ . Assim, a partir da estimativa de um modelo original:

$$Y_i = a + b_1 \cdot X_{1i} + b_2 \cdot X_{2i} + \dots + b_k \cdot X_{ki} + u_i \quad (12.61)$$

podemos estimar o seguinte modelo:

$$Y_i = a + d_1 \cdot \hat{Y}_i + d_2 \cdot (\hat{Y}_i)^2 + v_i \quad (12.62)$$

de onde se espera que  $\hat{Y}$  seja estatisticamente significante e  $\hat{Y}^2$  não seja, uma vez que, se o modelo original for especificado corretamente em termos de forma funcional, o quadrado dos valores previstos da variável dependente não deverá apresentar um poder explicativo sobre a variável dependente original. O *linktest* aplicado diretamente

no Stata apresenta exatamente esta configuração, porém um pesquisador que tiver interesse em avaliar a significância estatística da variável  $\hat{Y}$  com outros expoentes poderá fazê-lo manualmente.

Já o teste **RESET (Regression Specification Error Test)** avalia a existência de erros de especificação do modelo pela omissão de variáveis relevantes. Similarmente ao *linktest*, o teste *RESET* também cria novas variáveis com base nos valores de  $\hat{Y}$  gerados a partir da estimativa do modelo original representado pela expressão (12.61). Assim, podemos estimar o seguinte modelo:

$$Y_i = a + b_1 \cdot X_{1i} + b_2 \cdot X_{2i} + \dots + b_k \cdot X_{ki} + d_1 \cdot (\hat{Y}_i)^2 + d_2 \cdot (\hat{Y}_i)^3 + d_3 \cdot (\hat{Y}_i)^4 + v_i \quad (12.63)$$

A partir da estimativa do modelo representado pela expressão (12.63), podemos calcular a estatística  $F$  da seguinte forma:

$$F = \frac{\frac{\left( \sum_{i=1}^n u_i^2 - \sum_{i=1}^n v_i^2 \right)}{3}}{\frac{\left( \sum_{i=1}^n v_i^2 \right)}{(n - k - 4)}} \quad (12.64)$$

em que  $n$  é o número de observações e  $k$  é o número de variáveis explicativas do modelo original.

Desta forma, se a estatística  $F$  calculada para  $(3, n - k - 4)$  graus de liberdade for menor do que o correspondente  $F$  crítico ( $H_0$  do teste *RESET*), podemos afirmar que o modelo original não apresenta omissão de variáveis explicativas relevantes.

Da mesma forma que para o *linktest*, na seção 12.5 elaboraremos o teste *RESET* a partir da estimativa de um modelo no Stata.

## 12.4. MODELOS NÃO LINEARES DE REGRESSÃO

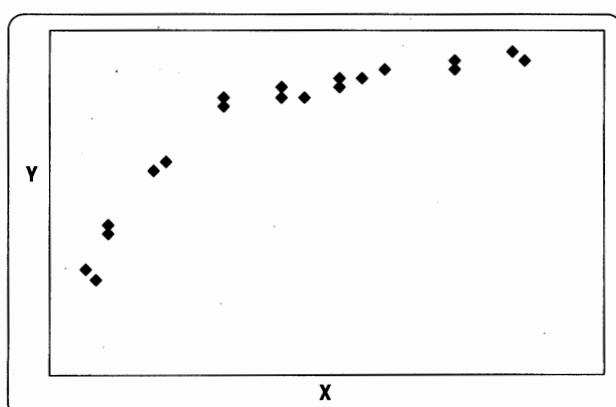
Conforme já estudamos, um modelo de regressão linear com uma única variável  $X$  pode ser representado por:

$$Y_i = a + b \cdot X_i + u_i \quad (12.65)$$

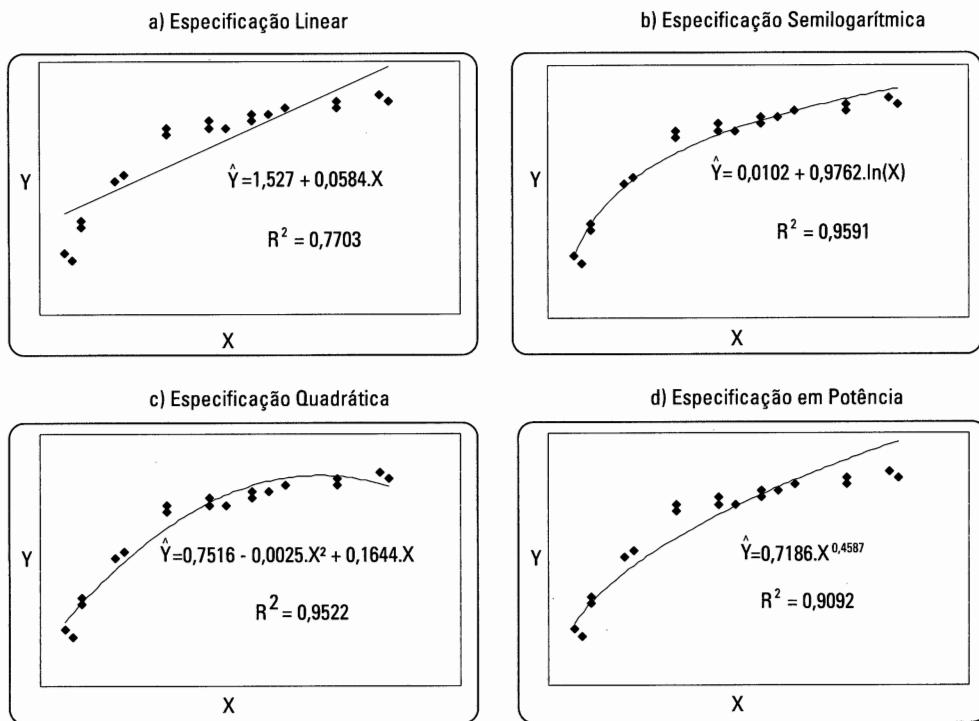
Porém, imagine uma situação em que a variável  $Y$  seja mais bem explicada por um comportamento não linear da variável  $X$ . Desta forma, a adoção, por parte do pesquisador, de uma forma funcional linear poderá gerar um modelo com menor  $R^2$  e, consequentemente, com pior poder preditivo.

Imagine uma situação hipotética apresentada por meio da Figura 12.46. Nitidamente,  $Y$  e  $X$  se relacionam de maneira não linear.

Um pesquisador, bastante curioso, elaborou quatro modelos de regressão, com o intuito de escolher o mais apropriado para efeitos de previsão. As formas funcionais escolhidas foram a linear, a semilogarítmica, a quadrática e a conhecida por potência. A Figura 12.47 apresenta os resultados destes quatro modelos.



**Figura 12.46** Exemplo de comportamento não linear entre uma variável  $Y$  e uma variável  $X$ .



**Figura 12.47** Resultados da aplicação de quatro diferentes formas funcionais em regressão.

Ao analisar os resultados, o pesquisador verificou que a forma funcional semilogarítmica apresentou maior  $R^2$ , o que vai propiciar melhor poder preditivo do modelo e, portanto, será o modelo a ser escolhido. Além disso, percebeu também, neste caso, que a forma funcional linear foi a que apresentou  $R^2$  mais baixo.

As relações entre variáveis podem se dar por meio de inúmeras formas funcionais não lineares que eventualmente devem ser consideradas quando da estimação de modelos de regressão, para que seja, de maneira mais adequada, compreendido o comportamento dos diferentes fenômenos. Neste sentido, o Quadro 12.3 apresenta as principais formas funcionais utilizadas.

Segundo Linneman (1980) e Aguirre e Macedo (1996), a definição da melhor forma funcional é uma questão empírica a ser decidida a favor do melhor ajuste dos dados. Ressaltamos, todavia, que o pesquisador tem liberdade de aplicar as formas funcionais que melhor lhe convier com base na teoria subjacente, na análise preliminar dos dados e também em sua experiência, porém a decisão a favor de determinada forma funcional, respeitando-se os pressupostos da técnica, tem como base o maior  $R^2$  (para as mesmas amostras e com a mesma quantidade de parâmetros; caso contrário, deve-se optar pela escolha da forma funcional cujo modelo apresentar o maior  $R^2$  ajustado, conforme já discutimos).

**Quadro 12.3** Principais formas funcionais em modelos de regressão.

Forma Funcional	Modelo
Linear	$Y_i = a + b.X_i + u_i$
Semilogarítmica à Direita	$Y_i = a + b.\ln(X_i) + u_i$
Semilogarítmica à Esquerda	$\ln(Y_i) = a + b.X_i + u_i$
Logarítmica (ou Log-Log)	$\ln(Y_i) = a + b.\ln(X_i) + u_i$
Inversa	$Y_i = a + b.\left(\frac{1}{X_i}\right) + u_i$
Quadrática	$Y_i = a + b.(X_i)^2 + u_i$
Cúbica	$Y_i = a + b.(X_i)^3 + u_i$
Potência	$Y_i = a.(X_i)^b + u_i$

Fonte: Fouto (2004) e Fávero (2005).

Segundo Fouto (2004) e Fávero (2005), enquanto na forma funcional linear o parâmetro  $b$  indica o efeito marginal da variação de  $X$  sobre a variável  $Y$ , na forma funcional semilogarítmica à direita o parâmetro  $b$  representa o efeito marginal da variação de  $\ln(X)$  sobre a variável  $Y$ .

Já os parâmetros dos modelos com formas funcionais inversa, quadrática e cúbica representam, respectivamente, o efeito marginal, sobre a variável  $Y$ , da variação do inverso, do quadrado e do cubo de  $X$ .

Por fim, nas formas funcionais semilogarítmica à esquerda e logarítmica (log-log), o coeficiente da variável  $X$  pode ser interpretado como uma elasticidade parcial. É importante mencionar que os modelos de regressão logística binária e multinomial, os modelos de regressão para dados de contagem do tipo Poisson e binomial negativo e os modelos de regressão para dados de sobrevivência são casos particulares dos modelos semilogarítmicos à esquerda, também conhecidos por modelos log-lineares ou exponenciais não lineares, e serão estudados, respectivamente, nos Capítulos 13, 14 e 17.

#### 12.4.1. Transformação de Box-Cox: o modelo geral de regressão

Box e Cox (1964), em seminal artigo, apresentam um modelo geral de regressão a partir do qual todas as formas funcionais apresentadas derivam, ou seja, são casos particulares. Segundo os autores, e conforme discutem Fávero (2005) e Fávero *et al.* (2009), a partir do modelo de regressão linear com uma única variável  $X$ , representado por meio da expressão (12.65), pode-se obter um modelo transformado a partir da substituição de  $Y$  por  $(Y^\lambda - 1) / \lambda$  e de  $X$  por  $(X^\theta - 1) / \theta$ , em que  $\lambda$  e  $\theta$  são os parâmetros da transformação. Assim, o modelo passa a ser:

$$\frac{Y_i^\lambda - 1}{\lambda} = a + b \left( \frac{X_i^\theta - 1}{\theta} \right) + u_i \quad (12.66)$$

A partir da expressão (12.66), podemos atribuir, conforme mostra a Quadro 12.4, valores para  $\lambda$  e  $\theta$  de modo a obtermos casos particulares para algumas das principais formas funcionais definidas no Quadro 12.3.

Box e Cox (1964) demonstram, por expansão de Taylor, que um logaritmo natural ( $\ln$ ) é obtido quando determinado parâmetro ( $\lambda$  ou  $\theta$ ) for igual a zero.

Uma nova variável obtida por meio de uma transformação de Box-Cox aplicada a uma variável original passa a apresentar uma nova distribuição (novo histograma). Por esta razão, é muito comum que pesquisadores obtenham novas variáveis transformadas a partir de variáveis originais, nos casos em que estas últimas apresentarem grandes amplitudes e valores muito discrepantes. Por exemplo, imagine uma base de dados com preços por metro quadrado de aluguel de lojas, que podem variar de R\$100/m<sup>2</sup> a R\$10.000/m<sup>2</sup>. Neste caso, a aplicação do logaritmo natural diminuiria consideravelmente a amplitude e a discrepância dos valores ( $\ln(100) = 4,6$  e  $\ln(10.000) = 9,2$ ). Em finanças e contabilidade, por exemplo, porte empresarial é uma variável que já é tradicionalmente conhecida como sendo o logaritmo natural dos ativos da empresa.

Para variáveis *dummy*, obviamente qualquer transformação de Box-Cox não faz o menor sentido, já que, como estas assumem valores iguais a 0 ou 1, qualquer expoente não alterará o valor original da variável.

Conforme estudamos na seção 12.3, os pressupostos relacionados aos resíduos (normalidade, homocedasticidade e ausência de autocorrelação) em modelos de regressão podem ser violados por falhas de especificação na forma funcional. Desta maneira, uma transformação de Box-Cox pode auxiliar o pesquisador na definição de outras formas funcionais, que não a linear, propiciando inclusive que se responda a seguinte pergunta: **Qual parâmetro de Box-Cox ( $\lambda$  para a variável dependente e  $\theta$  para uma variável explicativa) que maximiza**

**Quadro 12.4** Transformações de Box-Cox e valores de  $\lambda$  e  $\theta$  para cada forma funcional.

Parâmetro $\lambda$	Parâmetro $\theta$	Forma Funcional
1	1	Linear
1	0	Semilogarítmica à direita
0	1	Semilogarítmica à esquerda
0	0	Logarítmica (ou Log-Log)
1	-1	Inversa
1	2	Quadrática
1	3	Cúbica

a aderência à normalidade da distribuição de uma nova variável transformada gerada a partir de uma variável original? Como os parâmetros de Box-Cox variam de  $-\infty$  a  $+\infty$ , qualquer valor pode ser obtido. Faremos uso do software Stata, na seção 12.5, para responder a esta importante questão.

## 12.5. ESTIMAÇÃO DE MODELOS DE REGRESSÃO NO SOFTWARE STATA

O objetivo desta seção não é o de discutir novamente todos os conceitos inerentes às estatísticas e aos pressupostos da técnica de regressão, porém propiciar ao pesquisador que se conheçam os comandos do Stata, bem como mostrar as suas vantagens em relação a outros softwares, no que diz respeito aos modelos de dependência. O mesmo exemplo da seção 12.2 será aqui utilizado, sendo este critério adotado ao longo de todo o livro. A reprodução das imagens do Stata Statistical Software® nesta seção tem autorização da StataCorp LP®.

Voltando então ao exemplo, lembremos que um professor tinha o interesse em avaliar se o tempo de deslocamento de seus estudantes até a escola, independentemente de onde estariam partindo, era influenciado por variáveis como distância, quantidade de semáforos, período do dia em que se dava o trajeto e perfil do condutor ao volante. Já partiremos para o banco de dados final construído pelo professor por meio dos questionamentos elaborados ao seu grupo de 10 estudantes. O banco de dados encontra-se no arquivo **Tempodistsemperperfil.dta** e é exatamente igual ao apresentado na Tabela 12.10.

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 12.48 apresenta este primeiro *output* do Stata.

Embora a variável *per* seja qualitativa, possui apenas duas categorias que, no banco de dados, já estão rotuladas como *dummy* (manhã = 1; tarde = 0). Por outro lado, a variável *perfil* possui três categorias e, portanto, será preciso que criemos ( $n - 1 = 2$ ) *dummies*, conforme discutido na seção 12.2.6. O comando **tab** oferece a distribuição de frequências de uma variável qualitativa, com destaque para a quantidade de categorias. Se o pesquisador tiver dúvidas sobre o número de categorias, poderá recorrer facilmente a este comando (Figura 12.49).

<b>. desc</b>				
<b>obs:</b>	10			
<b>vars:</b>	6			
<b>size:</b>	200 (99.9% of memory free)			
<hr/>				
variable	storage	display	value	variable label
name	type	format	label	
estudante	str11	%11s		
tempo	byte	%8.0g		tempo para se chegar à escola (minutos)
dist	byte	%8.0g		distância percorrida até a escola (km)
sem	byte	%8.0g		quantidade de semáforos
per	byte	%8.0g	per	periódio do dia
perfil	byte	%9.0g	perfil	perfil ao volante
<hr/>				
Sorted by:				

Figura 12.48 Descrição do banco de dados **Tempodistsemperperfil.dta**.

<b>. tab perfil</b>				
perfil ao	Freq.	Percent	Cum.	
calmo	3	30.00	30.00	
moderado	5	50.00	80.00	
agressivo	2	20.00	100.00	
Total	10	100.00		

Figura 12.49 Distribuição de frequências da variável *perfil*.

```
. xi i.perfil
i.perfil      Iperfil 1-3      (naturally coded; Iperfil 1 omitted)
```

Figura 12.50 Criação das duas *dummies* a partir da variável *perfil*.

O comando `xi i.perfil` nos fornecerá estas duas *dummies*, aqui nomeadas pelo Stata de `_Iperfil_2` e `_Iperfil_3`, mantendo exatamente o critério apresentado na Tabela 12.11 (Figura 12.50).

Antes de elaborarmos o modelo de regressão múltipla propriamente dito, podemos gerar um gráfico que mostra as inter-relações entre as variáveis, duas a duas. Este gráfico, conhecido por `matrix`, pode propiciar ao pesquisador um melhor entendimento de como as variáveis se relacionam, oferecendo inclusive uma eventual sugestão sobre formas funcionais não lineares. Vamos, neste caso, elaborar o gráfico apenas com as variáveis quantitativas do modelo (Figura 12.51), a fim de facilitar a visualização. Assim, devemos digitar o seguinte comando:

```
graph matrix tempo dist sem
```

Por meio deste gráfico, podemos verificar que as relações entre a variável `tempo` e as variáveis `dist` e `sem` são positivas a aparentemente lineares. É possível verificar também que talvez exista certa multicolinearidade entre as variáveis explicativas. Uma matriz de correlações simples também pode ser gerada antes da elaboração da regressão, a fim de municiar o pesquisador com informações nesta fase de diagnóstico do banco de dados. Para tanto, devemos digitar o seguinte comando:

```
pwcorr tempo dist sem per _Iperfil_2 _Iperfil_3, sig
```

A Figura 12.52 apresenta a matriz de correlações simples.

Por meio desta matriz, podemos verificar realmente que as correlações entre as variáveis `tempo` e `dist` e entre `tempo` e `sem` são altas e estatisticamente significantes, ao nível de significância de 5%. É importante mencionar que os valores apresentados embaixo de cada correlação referem-se aos respectivos níveis de significância. Por meio da mesma matriz, por outro lado, é possível perceber que podem surgir eventuais problemas de multicolinearidade entre algumas variáveis explicativas, como, por exemplo, entre `per` e `_Iperfil_3`. Conforme veremos adiante, embora a correlação entre `tempo` e `per` seja maior, em módulo, do que entre `tempo` e `_Iperfil_3`, a variável `per` será excluída do modelo final pelo **procedimento Stepwise**, diferentemente da variável `_Iperfil_3`.

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

```
reg tempo dist sem per _Iperfil_2 _Iperfil_3
```

O comando `reg` elabora uma regressão por meio do método de mínimos quadrados ordinários. Se o pesquisador não informar o nível de confiança desejado para a definição dos intervalos dos parâmetros estimados, o

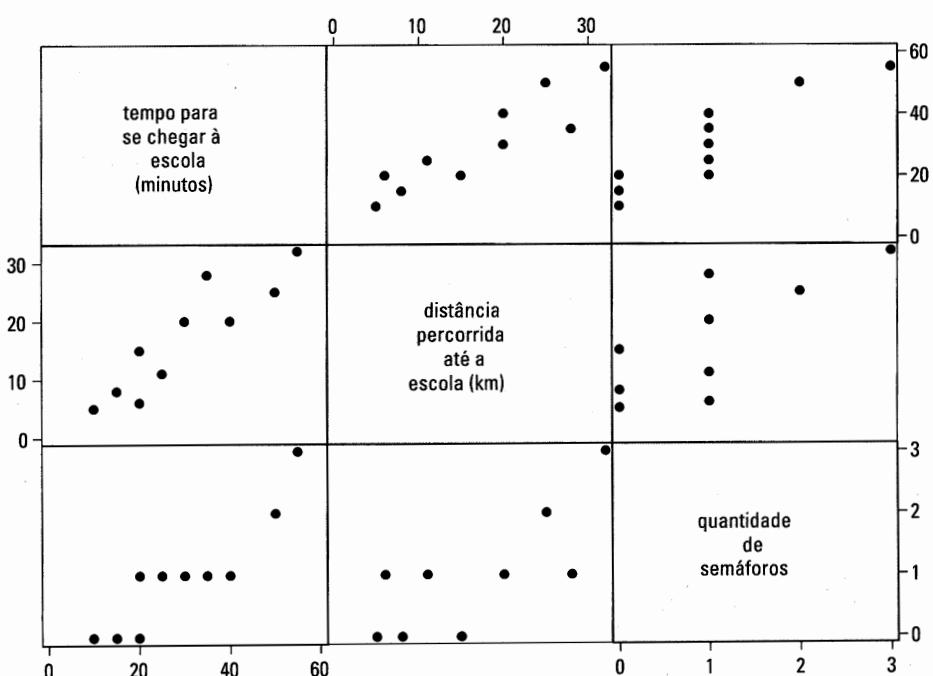


Figura 12.51 Inter-relação entre variáveis – gráfico `matrix`.

. pwcorr tempo dist sem per _Iperfil_2 _Iperfil_3, sig						
	tempo	dist	sem	per	_Iperf~2	_Iperf~3
tempo	1.0000					
dist	0.9052 0.0003	1.0000				
sem	0.9092 0.0003	0.7559 0.0114	1.0000			
per	-0.8487 0.0019	-0.6289 0.0515	-0.7319 0.0161	1.0000		
_Iperfil_2	-0.2828 0.4284	-0.1108 0.7605	-0.2236 0.5346	0.6547 0.0400	1.0000	
_Iperfil_3	0.5303 0.1148	0.3048 0.3918	0.2795 0.4341	-0.7638 0.0101	-0.5000 0.1411	1.0000

**Figura 12.52** Matriz de correlações simples.

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:

```
reg tempo dist sem per _Iperfil_2 _Iperfil_3, level(90)
```

Iremos seguir com a análise mantendo o nível de confiança dos intervalos dos parâmetros em 95%. Os *outputs* encontram-se na Figura 12.53 e são exatamente iguais aos apresentados na Figura 12.32.

Como a técnica de regressão faz parte do grupo de modelos conhecidos por **Modelos Lineares Generalizados (Generalized Linear Models)**, e como a variável dependente apresenta distribuição normal (também conhecida por distribuição de Gauss ou distribuição gaussiana), os parâmetros estimados por mínimos quadrados ordinários (comando **reg**) e apresentados na Figura 12.53 também poderiam ser igualmente obtidos por meio da estimação por máxima verossimilhança, a ser estudada no próximo capítulo. Para tanto, poderia ter sido digitado o seguinte comando:

```
glm tempo dist sem per _Iperfil_2 _Iperfil_3, family(gaussian)
```

Conforme já discutimos, os parâmetros das variáveis *per* e *\_Iperfil\_2* não se mostraram estatisticamente significantes neste modelo na presença das demais variáveis, ao nível de significância de 5%. Partiremos, então, para a aplicação do procedimento *Stepwise*, que exclui as variáveis cujos parâmetros não se mostrem estatisticamente significantes, embora isso possa criar um problema de especificação pela omissão de determinada variável que seria relevante para explicar o comportamento da variável dependente, caso não houvesse outras variáveis explicativas no modelo final. Mais adiante, aplicaremos o teste *RESET* para a verificação de eventual existência de erros de especificação do modelo pela omissão de variáveis relevantes.

. reg tempo dist sem per _Iperfil_2 _Iperfil_3						
Source	SS	df	MS	Number of obs = 10 F( 5, 4) = 264.12 Prob > F = 0.0000 R-squared = 0.9970 Adj R-squared = 0.9932 Root MSE = 1.2288		
Model	1993.96043	5	398.792087			
Residual	6.03956505	4	1.50989126			
Total	2000	9	222.222222			
	tempo	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
	dist	.6740469	.0717153	9.40	0.001	.4749333 .8731605
	sem	6.646797	1.094867	6.07	0.004	3.606958 9.686636
	per	-5.371414	3.778781	-1.42	0.228	-15.86299 5.120164
	_Iperfil_2	1.779117	1.44146	1.23	0.285	-2.223017 5.781251
	_Iperfil_3	6.373641	2.243105	2.84	0.047	.1457827 12.6015
	_cons	13.49011	3.860886	3.49	0.025	2.77057 24.20965

**Figura 12.53** Outputs da regressão linear múltipla no Stata.

Vamos, então, digitar o seguinte comando:

```
stepwise, pr(0.05) : reg tempo dist sem per _Iperfil_2 _Iperfil_3
```

Para a elaboração do comando **stepwise**, o pesquisador precisa definir o nível de significância do teste  $t$  a partir do qual as variáveis explicativas são excluídas do modelo. Os *outputs* encontram-se na Figura 12.54 e são exatamente iguais aos apresentados na Figura 12.33.

Analogamente, os parâmetros estimados e apresentados na Figura 12.54 também poderiam ser obtidos por meio do seguinte comando:

```
stepwise, pr(0.05) : glm tempo dist sem per _Iperfil_2 _Iperfil_3,
family(gaussian)
```

<b>. stepwise, pr(0.05) : reg tempo dist sem per _Iperfil_2 _Iperfil_3</b>					
<b>begin with full model</b>					
<b>p = 0.2847 &gt;= 0.0500 removing _Iperfil_2</b>					
<b>p = 0.5141 &gt;= 0.0500 removing per</b>					
Source	SS	df	MS	Number of obs	
Model	1990.83863	3	663.612878	F( 3, 6) =	434.62
Residual	9.16136725	6	1.52689454	Prob > F =	0.0000
Total	2000	9	222.222222	R-squared =	0.9954
				Adj R-squared =	0.9931
				Root MSE =	1.2357
tempo	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
dist	.7104531	.0669006	10.62	0.000	.5467532 .874153
sem	7.836844	.6694031	11.71	0.000	6.198874 9.474814
_Iperfil_3	8.967607	1.02889	8.72	0.000	6.450003 11.48521
cons	8.291932	.8535082	9.72	0.000	6.203472 10.38039

**Figura 12.54** Outputs da regressão linear múltipla com procedimento Stepwise no Stata.

Conforme já estudado na seção 12.2.6, chegamos ao seguinte modelo de regressão linear múltipla:

$$\hat{tempo}_i = 8,2919 + 0,7105.dist_i + 7,8368.sem_i + 8,9676._Iperfil\_3_{i \begin{smallmatrix} \text{calmo}=0 \\ \text{agressivo}=1 \end{smallmatrix}}$$

O comando **predict yhat** faz com que seja gerada uma nova variável (*yhat*) no banco de dados, que oferece os valores previstos ( $\hat{Y}$ ) para cada observação do último modelo elaborado.

Entretanto, podemos também desejar saber o valor previsto para determinada observação que não se encontra na base de dados. Ou seja, podemos novamente elaborar a pergunta feita ao final da seção 12.2.6 e respondida, naquele momento, de forma manual: **Qual é o tempo estimado para um aluno que se desloca 17 quilômetros, passa por dois semáforos, decide ir à escola de manhã e tem um perfil considerado agressivo ao volante?**

Por meio do comando **mfx**, o Stata permite que o pesquisador responda esta pergunta diretamente. Assim, devemos digitar o seguinte comando:

```
mfx, at(dist=17 sem=2 _Iperfil_3=1)
```

Obviamente, o termo **per = 1** não precisa ser incluído no comando **mfx**, já que a variável *per* não está presente no modelo final. O *output* é apresentado na Figura 12.55 e, por meio dele, podemos chegar à resposta de 45,0109 minutos, que é exatamente igual àquela calculada manualmente na seção 12.2.6.

Definido o modelo, partiremos para a verificação dos pressupostos da técnica, conforme estudado na seção 12.3. Anteriormente, entretanto, é sempre interessante que o pesquisador, ao estimar determinado modelo, elabore uma análise acerca de eventuais observações que sejam discrepantes na base de dados e estejam influenciando de maneira considerável as estimativas dos parâmetros do modelo, e, como sabemos, esta influência, assim como a presença de *outliers*, pode ser uma das causas da heterocedasticidade.

Para tanto, introduziremos o conceito de distância *leverage* que, para cada observação *i*, corresponde ao valor da *i*-ésima posição da diagonal principal da matriz  $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ . Uma observação pode ser considerada como

```
. mfx, at(dist=17 sem=2 _Iperfil_3=1)

Marginal effects after regress
y = Fitted values (predict)
= 45.01093

variable | dy/dx Std. Err. z P>|z| [ 95% C.I. ] x
-----+-----+-----+-----+-----+-----+-----+
dist | .7104531 .0669 10.62 0.000 .57933 .841576 17
sem | 7.836844 .6694 11.71 0.000 6.52484 9.14885 2
_Iperf~3*| 8.967607 1.02889 8.72 0.000 6.95102 10.9842 1
-----+-----+-----+-----+-----+-----+-----+
(* dy/dx is for discrete change of dummy variable from 0 to 1
```

**Figura 12.55** Cálculo da estimação de Y para valores das variáveis explicativas – comando **mfx**.

grande influente da estimativa dos parâmetros de um modelo se a sua distância *leverage* for maior que  $(2.k / n)$ , em que  $k$  é o número de variáveis explicativas e  $n$  é o tamanho da amostra. As distâncias *leverage* são geradas no Stata por meio do comando:

**predict lev, leverage**

No nosso exemplo, solicitaremos que o Stata gere as distâncias *leverage* para o modelo final estimado com o procedimento *Stepwise*. Estas distâncias estão apresentadas na Tabela 12.16.

No modelo final, como  $(2.k / n) = (2.3 / 10) = 0,6$ , a observação 8 (Antônio) é aquela com maior potencial para influenciar a estimativa dos parâmetros e, consequentemente, deve-se dispensar atenção especial a ela, já que eventuais problemas de heterocedasticidade podem surgir em decorrência desse fato. Um gráfico das distâncias *leverage* em função dos termos de erro padronizados ao quadrado (Figura 12.56) pode propiciar ao pesquisador uma fácil análise das observações com maior influência sobre os parâmetros do modelo (altas distâncias *leverage*) e, ao mesmo tempo, uma análise das observações consideradas *outliers* (elevados resíduos padronizados ao quadrado). Como sabemos, ambas podem gerar problemas de estimação. O comando para elaboração deste gráfico no nosso exemplo é:

**lvr2plot, mlabel(estudante)**

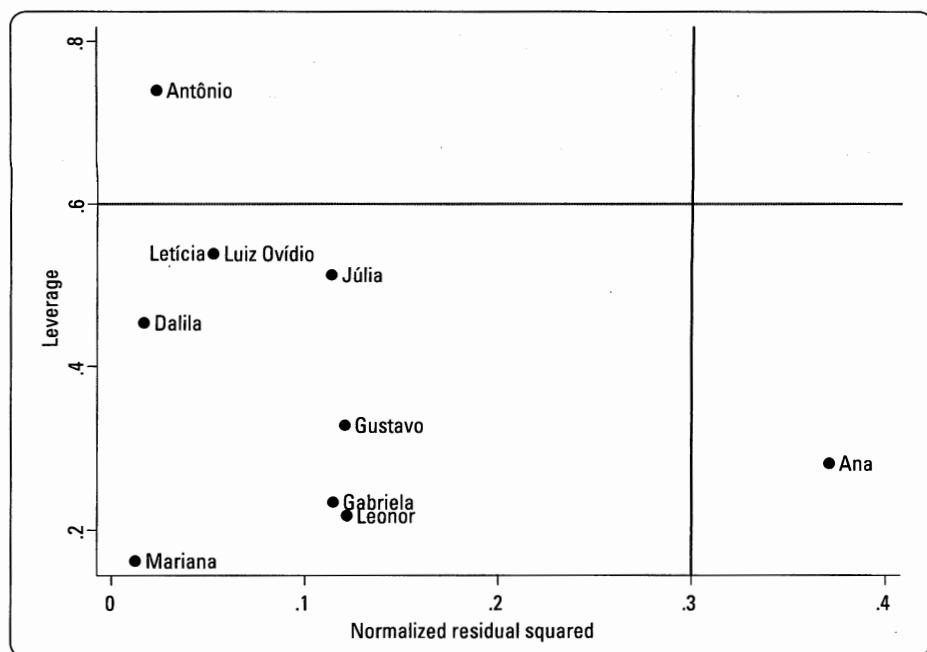
Por meio do gráfico da Figura 12.56, podemos perceber que, enquanto Antônio tem maior influência sobre os parâmetros do modelo, Ana tem propensão a ser um *outlier* na amostra por pelo fato de apresentar maior termo de erro em módulo (e, consequentemente, maior termo de erro padronizado ao quadrado). O grau de influência destas observações sobre o surgimento da heterocedasticidade no modelo deverá ser investigado quando da elaboração dos testes de verificação dos pressupostos. Vamos então a eles!

O primeiro pressuposto, conforme mostra o Quadro 12.2, refere-se à normalidade dos resíduos. Vamos, dessa forma, gerar uma variável que corresponde aos termos de erro do modelo final. Para tanto, devemos digitar o seguinte comando:

**predict res, res**

**Tabela 12.16** Distâncias *leverage* para o modelo final.

Observação (i)	$lev_i$ (Modelo Final)
Gabriela	0,23
Dalila	0,45
Gustavo	0,33
Letícia	0,54
Luiz Ovídio	0,54
Leonor	0,22
Ana	0,28
Antônio	0,74
Júlia	0,51
Mariana	0,16

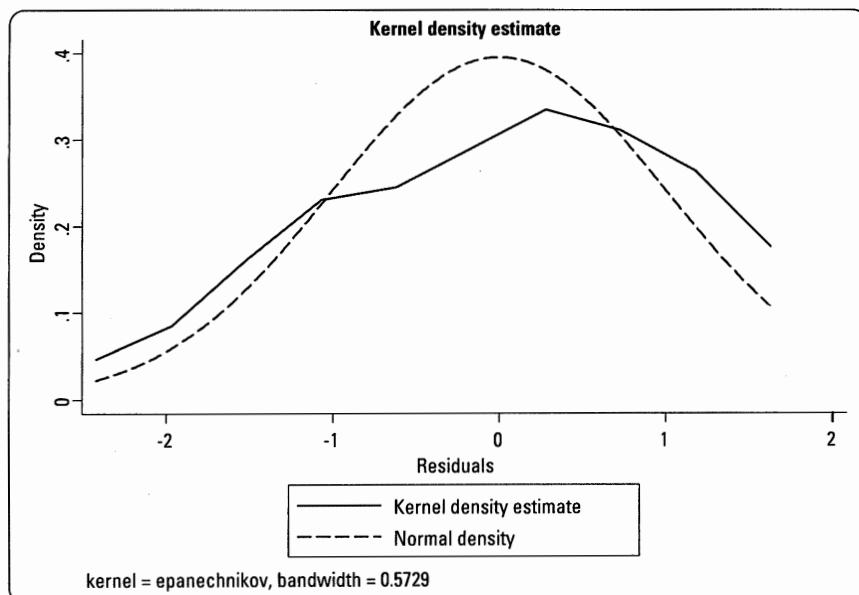


**Figura 12.56** Distâncias leverage em função dos resíduos padronizados ao quadrado.

Após gerarmos a variável *res*, que oferece os valores dos termos de erro de cada observação para o modelo final estimado com o procedimento *Stepwise*, podemos elaborar um gráfico que permite a comparação visual da distribuição dos termos de erro gerados pelo modelo com a distribuição normal padrão. Assim, devemos digitar o seguinte comando:

**kdensity res, normal**

O gráfico gerado encontra-se na Figura 12.57 e, por meio do mesmo, podemos ter uma ideia do quanto a distribuição dos resíduos gerados (*Kernel density estimate*) se aproxima da distribuição normal padrão.



**Figura 12.57** Gráfico de aderência entre a distribuição dos resíduos e a distribuição normal.

Como a amostra deste exemplo é de apenas 10 observações, aplicaremos o teste de Shapiro-Wilk, recomendado para amostras com até 30 observações (conforme discutimos no Capítulo 7), para que possamos efetivamente corroborar a hipótese de que a distribuição dos resíduos é aderente à distribuição normal. Para tanto, utilizaremos o seguinte comando:

**swilk res**

O *output* do teste encontra-se na Figura 12.58 e, por meio de sua análise, podemos verificar que os termos de erro apresentam distribuição normal ao nível de significância de 5%, não havendo rejeição de sua hipótese nula.

Shapiro-Wilk W test for normal data					
Variable	Obs	W	V	z	Prob>z
res	10	0.90525	1.460	0.675	0.24995

**Figura 12.58** Resultado do teste de normalidade de Shapiro-Wilk para os resíduos.

Para amostras maiores, conforme discutimos, recomenda-se a aplicação do teste de Shapiro-Francia, cujo comando é:

**sfrancia res**

O segundo pressuposto a ser verificado diz respeito à inexistência de multicolinearidade das variáveis explicativas. Após a elaboração do modelo completo (ainda sem o procedimento *Stepwise*), podemos digitar o seguinte comando:

**estat vif**

Os *outputs* são apresentados na Figura 12.59 e, por meio deles, podemos verificar que a estatística *VIF* da variável *per* é a mais elevada de todas ( $VIF_{per} = 19,86$ ), o que indica que o  $R^2$  resultante de uma regressão com esta variável como dependente de todas as outras seria de aproximadamente 95% ( $Tolerance_{per} = 0,05$ ). A própria Figura 12.52 nos mostra que as correlações simples entre a variável *per* e as demais variáveis explicativas são bastante elevadas, o que já dá inicialmente a entender que há existência de multicolinearidade. Entretanto, como sabemos, o modelo final não inclui esta variável, e tampouco a variável *Iperfil\_2*. A Figura 12.60 mostra os *outputs* gerados por meio do comando **estat vif** aplicado após a elaboração do procedimento *Stepwise*.

Como o modelo final obtido após o procedimento *Stepwise* não apresenta estatísticas *VIF* muito elevadas para nenhuma variável explicativa, podemos considerar que a multicolinearidade existente no modelo completo foi bastante reduzida. A própria variável *sem*, presente no modelo final, teve sua estatística *VIF* reduzida de 6,35 para 2,35 com a exclusão principalmente da variável *per*. É importante apenas que verifiquemos, por meio do teste *RESET*, se a exclusão destas variáveis criará algum problema de especificação por omissão de variável relevante. Isso será elaborado mais adiante.

estat vif		
Variable	VIF	1/VIF
per	19.86	0.050353
sem	6.35	0.157446
<i>Iperfil_3</i>	5.33	0.187554
<i>Iperfil_2</i>	3.44	0.290670
dist	2.77	0.360660
Mean VIF	7.55	

**Figura 12.59** Estatísticas *VIF* e *Tolerance* das variáveis explicativas para o modelo completo.

estat vif		
Variable	VIF	1/VIF
dist	2.39	0.419106
sem	2.35	0.425935
<i>Iperfil_3</i>	1.11	0.901469
Mean VIF	1.95	

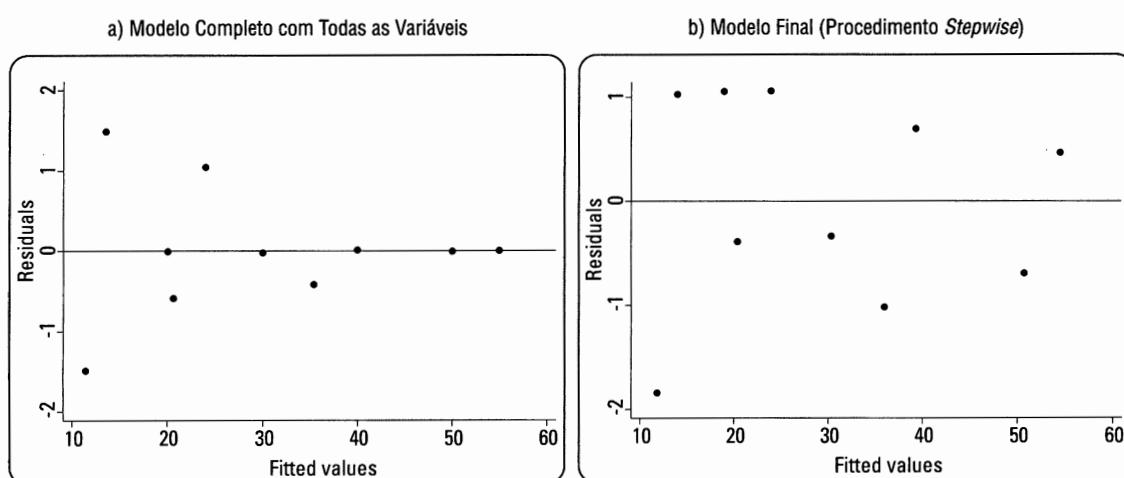
**Figura 12.60** Estatísticas *VIF* e *Tolerance* das variáveis explicativas para o modelo final.

O terceiro pressuposto refere-se à ausência de heterocedasticidade. Inicialmente, apenas para efeitos de diagnóstico, vamos elaborar um gráfico dos valores dos termos de erro em função dos valores previstos ( $\hat{Y}$ ) do modelo estimado. A Figura 12.61 apresenta os gráficos gerados após as estimativas do modelo completo e do modelo final, em que são plotados os valores dos resíduos padronizados em função dos valores estimados da variável dependente. O comando para a elaboração destes gráficos, que deve ser digitado após a estimativa de cada um dos modelos, é:

```
rvfplot, yline(0)
```

Enquanto a Figura 12.61a mostra a formação de um “cone” nitidamente visível, o mesmo já não pode ser afirmado em relação à Figura 12.61b. De fato, como veremos adiante, o modelo completo, com a inclusão de todas as variáveis explicativas, apresenta heterocedasticidade, enquanto o modelo final obtido por meio do procedimento *Stepwise* gera termos de erro homocedásticos.

Para a verificação da existência de heterocedasticidade, aplicaremos o teste de Breusch-Pagan/Cook-Weisberg que, conforme já discutimos, apresenta, como hipótese nula, o fato de a variância dos termos de erro ser constante (erros homocedásticos) e, como hipótese alternativa, o fato de a variância dos termos de erro não ser constante, ou seja, os termos de erro serem uma função de uma ou mais variáveis explicativas (erros heterocedásticos). Este teste é indicado para os casos em que a suposição de normalidade dos resíduos for verificada, como no presente exemplo.



**Figura 12.61** Método gráfico para identificação de heterocedasticidade.

A seção 12.3.3.3, conforme vimos, descreve o teste e oferece uma possibilidade de que o mesmo seja elaborado de forma manual, passo a passo. Faremos isso inicialmente, a fim de que o pesquisador possa analisar os *outputs* e confrontá-los com os resultados gerados pelo Stata.

Para tanto, precisamos desenvolver uma tabela que permita o cálculo da estatística de Breusch-Pagan, a partir da estimativa do modelo final:

$$\text{tempo}_i = 8,2919 + 0,7105 \cdot \text{dist}_i + 7,8368 \cdot \text{sem}_i + 8,9676 \cdot \text{Iperfil\_3}_i + u_i$$

Com base na estimativa de  $u_i$  para cada observação, podemos calcular os valores de  $u_i^2$  e, por meio da expressão (12.40), os valores de  $up_i$ . A Tabela 12.17 traz estes valores.

Para a obtenção do resultado do teste, o procedimento é que se elabore a regressão  $up_i = a + b \cdot \hat{Y}_i + \xi_i$ , de onde se calcula a soma dos quadrados da regressão (SQR) que, dividindo-se por 2, chega-se à estatística  $\chi^2_{BP/CW}$ . No nosso exemplo,  $SQR = 3,18$ , de onde vem que  $\chi^2_{BP/CW} = 1,59 < \chi^2_{1\ g.l.} = 3,84$  para o nível de significância de 5%, ou seja, a hipótese nula do teste (**termos de erro homocedásticos**) não pode ser rejeitada.

O comando para a aplicação direta do teste no Stata é dado por:

```
estat hettest
```

**Tabela 12.17** Elaboração do teste de Breusch-Pagan/Cook-Weisberg.

Observação (i)	$u_i$ $(Y_i - \hat{Y}_i)$	$u_i^2$	$up_i = \frac{u_i^2}{\left( \sum_{i=1}^n u_i^2 \right) / n}$	$\hat{Y}_i$
Gabriela	1,02444	1,04948	1,14555	13,97556
Dalila	-0,39149	0,15327	0,16730	20,39149
Gustavo	1,05127	1,10517	1,20634	18,94873
Letícia	0,69455	0,48241	0,52657	39,30545
Luiz Ovídio	-0,69455	0,48241	0,52657	50,69455
Leonor	1,05624	1,11564	1,21777	23,94376
Ana	-1,84420	3,40106	3,71240	11,84420
Antônio	0,46304	0,21440	0,23403	54,53696
Júlia	-1,02146	1,04339	1,13890	36,02146
Mariana	-0,33784	0,11413	0,12458	30,33784
<b>Soma</b>		<b>9,16137</b>		
<b>Média</b>		<b>0,91614</b>		

que avalia a existência de heterocedasticidade do último modelo gerado. O resultado deste teste para o modelo completo com a inclusão de todas as variáveis explicativas, embora não apresentado aqui, mostra que há existência de heterocedasticidade, como inclusive já esperávamos quando da análise da Figura 12.61a. Por outro lado, a Figura 12.62 apresenta o resultado do teste para o modelo final resultante do procedimento *Stepwise*, que é exatamente o mesmo daquele obtido manualmente, e, por meio de sua análise, podemos afirmar que este modelo final não apresenta problemas de heterocedasticidade (*valor-P*  $\chi^2 = 0,2069 > 0,05$ ).

```
. estat hettest
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of tempo

chi2(1)      =      1.59
Prob > chi2  =  0.2069
```

**Figura 12.62** Teste de Breusch-Pagan/Cook-Weisberg para heterocedasticidade.

Analogamente ao teste de Breusch-Pagan/Cook-Weisberg, o teste de White também avalia a rejeição ou não da hipótese nula de que os termos de erro sejam homocedásticos, a um determinado nível de significância. O comando para a realização deste teste é:

```
estat imtest, white
```

O *output* é apresentado na Figura 12.63 e oferece a mesma conclusão sobre a inexistência de heterocedasticidade dos resíduos no modelo final.

Como não verificamos a existência de heterocedasticidade no modelo final proposto, não elaboraremos a estimativa pelo método de mínimos quadrados ponderados. Entretanto, caso um pesquisador queira, por alguma razão, estimar um modelo com ponderação pela variável *per*, poderá propor a seguinte estimativa:

$$\frac{tempo_i}{per_i} = \frac{a}{per_i} + b_1 \cdot \frac{dist_i}{per_i} + b_2 \cdot \frac{sem_i}{per_i} + b_3 \cdot \frac{per_i}{per_i} + b_4 \cdot \frac{-Iperfil\_2_i}{per_i} + b_5 \cdot \frac{-Iperfil\_3_i}{per_i} + \frac{u_i}{per_i}$$

O comando para a estimativa do modelo por mínimos quadrados ponderados pela variável *per* seria:

```
wls0 tempo dist sem per _Iperfil_2 _Iperfil_3, wvar(per) type(abse)
```

```
. estat imtest, white

White's test for Ho: homoskedasticity
against Ha: unrestricted heteroskedasticity

chi2(7)      =      7.09
Prob > chi2  =  0.4201

Cameron & Trivedi's decomposition of IM-test

-----+
   Source |     chi2      df      p
-----+
   Heteroskedasticity | 7.09      7  0.4201
   Skewness | 1.90      3  0.5935
   Kurtosis | 1.42      1  0.2341
-----+
   Total | 10.40     11  0.4947
```

**Figura 12.63** Teste de White para heterocedasticidade.

Também não apresentaremos os *outputs* da estimação com erros-padrão robustos de Huber-White, dada a inexistência de heterocedasticidade neste exemplo. Entretanto, caso um pesquisador interessado deseje estudar a técnica, o comando para a elaboração desta estimação seria:

```
reg tempo dist sem per _Iperfil_2 _Iperfil_3, rob
```

Como o banco de dados do nosso exemplo é uma *cross-section*, não verificaremos o pressuposto de autocorrelação dos resíduos neste caso. Entretanto, mais adiante, por meio de outro banco de dados, estudaremos a aplicação dos testes voltados à verificação de tal pressuposto no Stata.

Sendo assim, partiremos para a aplicação do *linktest* que, conforme discutido na seção 12.3.5, se refere a um procedimento que cria duas novas variáveis a partir da elaboração de um modelo de regressão, que nada mais são do que as variáveis  $\hat{Y}$  e  $\hat{Y}^2$ , de onde se espera, ao regredirmos  $Y$  em função destas duas variáveis, que  $\hat{Y}$  seja estatisticamente significante e  $\hat{Y}^2$  não seja, uma vez que, se o modelo original for especificado corretamente em termos de forma funcional, o quadrado dos valores previstos da variável dependente não deverá apresentar um poder explicativo sobre a variável dependente original. O comando para aplicação deste teste no Stata é:

**linktest**

que deve ser digitado após a elaboração do modelo final. Os *outputs* do teste encontram-se na Figura 12.64.

Por meio da análise destes *outputs*, mais especificamente em relação ao *valor-P* da estatística *t* da variável *\_hatsq* (que se refere a  $\hat{Y}^2$ , ou seja, ao valor estimado ao quadrado da variável *tempo*), podemos afirmar que o *linktest* não rejeita a hipótese nula de que o modelo foi especificado corretamente em termos de forma funcional, ou seja, a forma funcional linear neste caso é adequada.

O teste *RESET*, também discutido na seção 12.3.5, avalia a existência de erros de especificação do modelo pela omissão de variáveis relevantes e, analogamente ao *linktest*, cria novas variáveis com base nos valores de  $\hat{Y}$  gerados a partir da estimação do modelo original. Desta forma, após a elaboração do modelo final por meio do

<b>. linktest</b>						Number of obs = 10 F( 2, 7) = 773.68 Prob > F = 0.0000 R-squared = 0.9955 Adj R-squared = 0.9942 Root MSE = 1.1343
Source	SS	df	MS	t	P> t	
Model	1990.99304	2	995.496519			
Residual	9.00696205	7	1.28670886			
Total	2000	9	222.222222			
tempo	Coeff.	Std. Err.		t	P> t	[95% Conf. Interval]
_hat	1.048706	.142885		7.34	0.000	.7108366 1.386575
_hatsq	-.0007371	.0021279		-0.35	0.739	-.0057687 .0042945
_cons	-.6510503	2.059793		-0.32	0.761	-5.521687 4.219586

**Figura 12.64** Linktest para verificação da adequação da forma funcional do modelo.

procedimento *Stepwise* e seguindo a expressão (12.63), iremos estimar o seguinte modelo, a partir do qual calcularemos manualmente a estatística  $F$  apresentada na expressão (12.64):

$$\text{tempo}_i = a + b_1 \cdot \text{dist}_i + b_2 \cdot \text{sem}_i + b_3 \cdot \text{Iperfil\_3}_i + d_1 \cdot (\hat{\text{tempo}}_i)^2 + d_2 \cdot (\hat{\text{tempo}}_i)^3 + d_3 \cdot (\hat{\text{tempo}}_i)^4 + \nu_i$$

Com base na estimativa do modelo final gerado pelo procedimento *Stepwise* (que possui termos de erro  $u_i$ ) e neste último modelo desenvolvido a partir da expressão (12.63) para se aplicar o teste *RESET* (que possui termos de erro  $v_i$ ), podemos criar a Tabela 12.18.

**Tabela 12.18** Construção da estatística  $F$  do teste *RESET*.

Observação (i)	$u_i$	$u_i^2$	$v_i$	$v_i^2$
Gabriela	1,02444	1,04948	1,27097	1,61537
Dalila	-0,39149	0,15327	-0,31770	0,10093
Gustavo	1,05127	1,10517	-0,49256	0,24261
Letícia	0,69455	0,48241	0,48498	0,23521
Luiz Ovídio	-0,69455	0,48241	-0,48498	0,23521
Leonor	1,05624	1,11564	0,51232	0,26247
Ana	-1,84420	3,40106	-0,75292	0,56689
Antônio	0,46304	0,21440	0,25524	0,06515
Júlia	-1,02146	1,04339	0,12753	0,01626
Mariana	-0,33784	0,11413	-0,60288	0,36346
<b>Soma</b>		<b>9,16137</b>		<b>3,70356</b>

E, a partir da Tabela 12.18, podemos calcular a estatística  $F$  do teste *RESET*, como segue:

$$F = \frac{\frac{\left( \sum_{i=1}^n u_i^2 - \sum_{i=1}^n v_i^2 \right)}{(9,16137 - 3,70356)}}{\frac{3}{\frac{\left( \sum_{i=1}^n v_i^2 \right)}{(10 - 3 - 4)}}} = \frac{3}{\frac{(3,70356)}{(10 - 3 - 4)}} = 1,47$$

Como a estatística  $F$  calculada para  $(3, 3)$  graus de liberdade é menor do que o correspondente  $F$  crítico ( $F_{(3,3)} = 9,28$  para o nível de significância de 5%), podemos afirmar que o modelo original não apresenta omissão de variáveis explicativas relevantes.

Para que seja elaborado o teste *RESET* no Stata, devemos digitar o seguinte comando após a estimativa do modelo final gerado por meio do procedimento *Stepwise*:

**ovtest**

O *output* encontra-se na Figura 12.65.

Desta forma, o *linktest* e o teste *RESET* nos indicam que não temos erros de especificação no modelo final gerado por meio do procedimento *Stepwise*. Se não fosse esse o caso, precisaríamos reespecificar o modelo por meio da mudança de sua forma funcional ou por meio da inclusão de variáveis explicativas relevantes que foram excluídas quando da estimativa.

```
. ovtest
Ramsey RESET test using powers of the fitted values of tempo
Ho: model has no omitted variables
      F(3, 3) =      1.47
      Prob > F = 0.3788
```

**Figura 12.65** Teste *RESET* para verificação de omissão de variáveis relevantes no modelo.

Portanto, o modelo proposto estimado com o procedimento *Stepwise* não apresentou problemas em relação a nenhum dos pressupostos e nem tampouco há a presença de erros de especificação.

A fim de que seja possível estudarmos uma eventual inexistência de linearidade em modelos de regressão, iremos agora trabalhar com outro banco de dados.

Imaginemos agora que o nosso professor tenha sido convidado para fazer uma palestra para 50 profissionais do setor público a respeito de mobilidade urbana, visto que ele tem pesquisado bastante sobre o tempo de locomoção das pessoas no município em função da distância percorrida e de outras variáveis, como a quantidade de semáforos por que passam diariamente. Ao término de sua palestra, muito aplaudida, o professor não pôde perder a oportunidade de coletar mais dados para suas investigações e, por conta disso, questionou cada um dos 50 presentes sobre o tempo de locomoção até o prédio em que estavam, a distância percorrida no trajeto e a quantidade de semáforos por que cada um havia passado naquela manhã. Assim, montou o banco de dados que se encontra no arquivo **Palestratempodistsem.dta**.

Seguindo os passos do professor, devemos inicialmente elaborar uma regressão linear múltipla para avaliar a influência das variáveis *dist* e *sem* sobre a variável *tempo*. Assim, devemos digitar o seguinte comando:

```
reg tempo dist sem
```

Os resultados encontram-se na Figura 12.66.

<b>. reg tempo dist sem</b>						
Source	SS	df	MS	Number of obs = 50		
Model	6185.00996	2	3092.50498	F( 2, 47) = 53.86		
Residual	2698.61004	47	57.4172349	Prob > F = 0.0000		
Total	8883.62	49	181.298367	R-squared = 0.6962		
				Adj R-squared = 0.6833		
				Root MSE = 7.5774		
tempo	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dist	.7728111	.1850909	4.18	0.000	.4004562	1.145166
sem	1.154891	.2750456	4.20	0.000	.601571	1.708212
_cons	13.06767	5.007771	2.61	0.012	2.993332	23.142

**Figura 12.66** Resultados da regressão linear múltipla.

Embora a análise preliminar dos resultados mostre uma estimativa satisfatória, o modelo apresentado na Figura 12.66 apresenta termos de erro com distribuição não aderente à normalidade, conforme podemos verificar por meio do teste de Shapiro-Francia (amostra com mais de 30 observações), obtido por meio da digitação do seguinte comando:

```
predict res, res  
sfrancia res
```

O resultado do teste encontra-se na Figura 12.67.

<b>. predict res, res</b>					
<b>. sfrancia res</b>					
Shapiro-Francia W' test for normal data					
Variable	Obs	W'	V'	z	Prob>z
res	50	0.93155	3.549	2.378	0.00869

**Figura 12.67** Resultado do teste de Shapiro-Francia para verificação de normalidade dos resíduos.

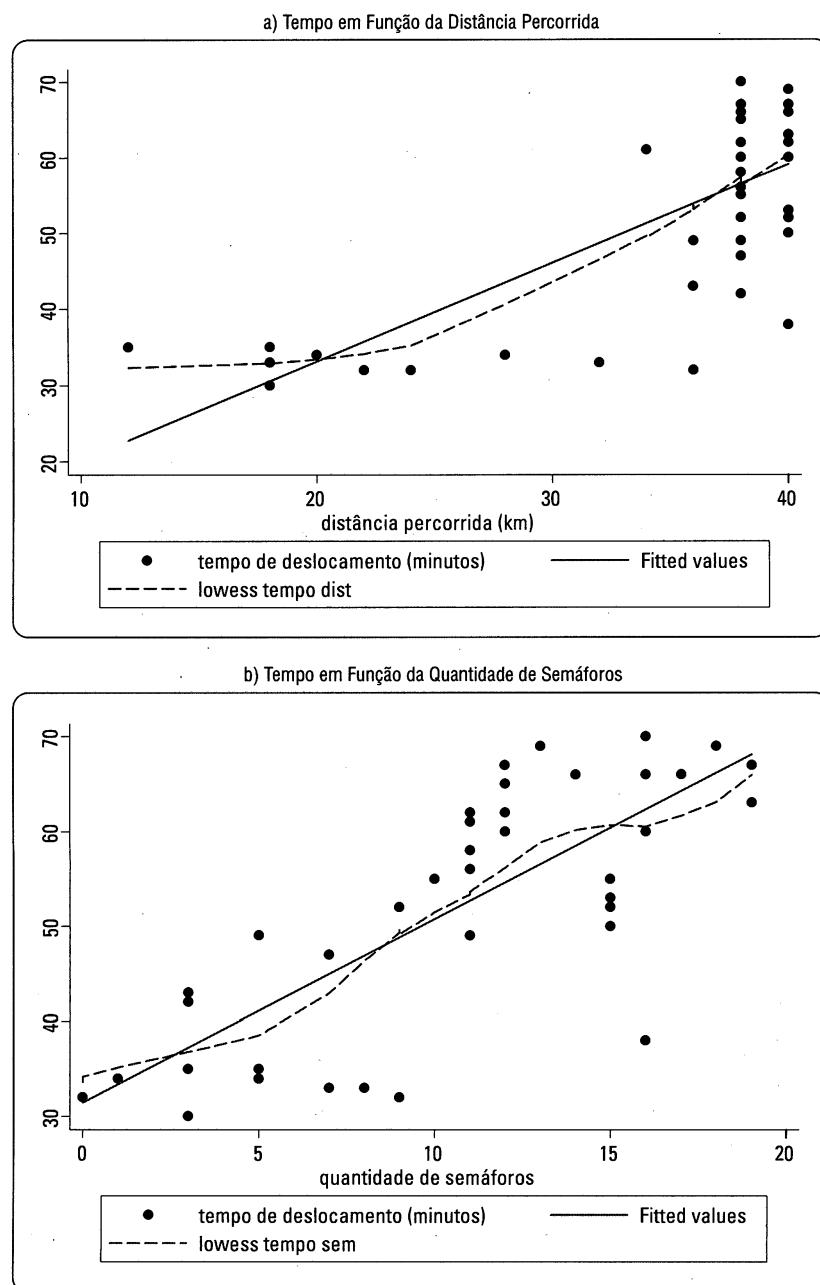
Como discutimos na seção 12.3.1, o pressuposto da normalidade assegura que o *valor-P* dos testes *t* e do teste *F* sejam válidos. Entretanto, a violação de tal pressuposto pode ser resultante de erros de especificação quanto à forma funcional do modelo.

Desta maneira, precisaremos elaborar gráficos da variável dependente em função de cada uma das variáveis explicativas individualmente e, nestes gráficos, apresentaremos o ajuste linear (valores previstos) e o ajuste conhecido por *lowess* (*locally weighted scatterplot smoothing*), que se refere a um método não paramétrico que utiliza múltiplas regressões para identificar o padrão de comportamento dos dados e, por alisamento, ajustar uma curva não necessariamente linear. Desta forma, devemos digitar os seguintes comandos:

```
graph twoway scatter tempo dist || lfit tempo dist || lowess tempo dist
graph twoway scatter tempo sem || lfit tempo sem || lowess tempo sem
```

A Figura 12.68 apresenta os dois gráficos gerados.

Nitidamente podemos perceber, por meio destes gráficos, que há diferenças entre os ajustes linear e *lowess*, principalmente para a variável *dist* (Figura 12.68a). Outra forma usual e similar de detectar a não linearidade do modelo é por meio de gráficos que apresentam a relação entre os **resíduos parciais aumentados** (*augmented residuals partials*)



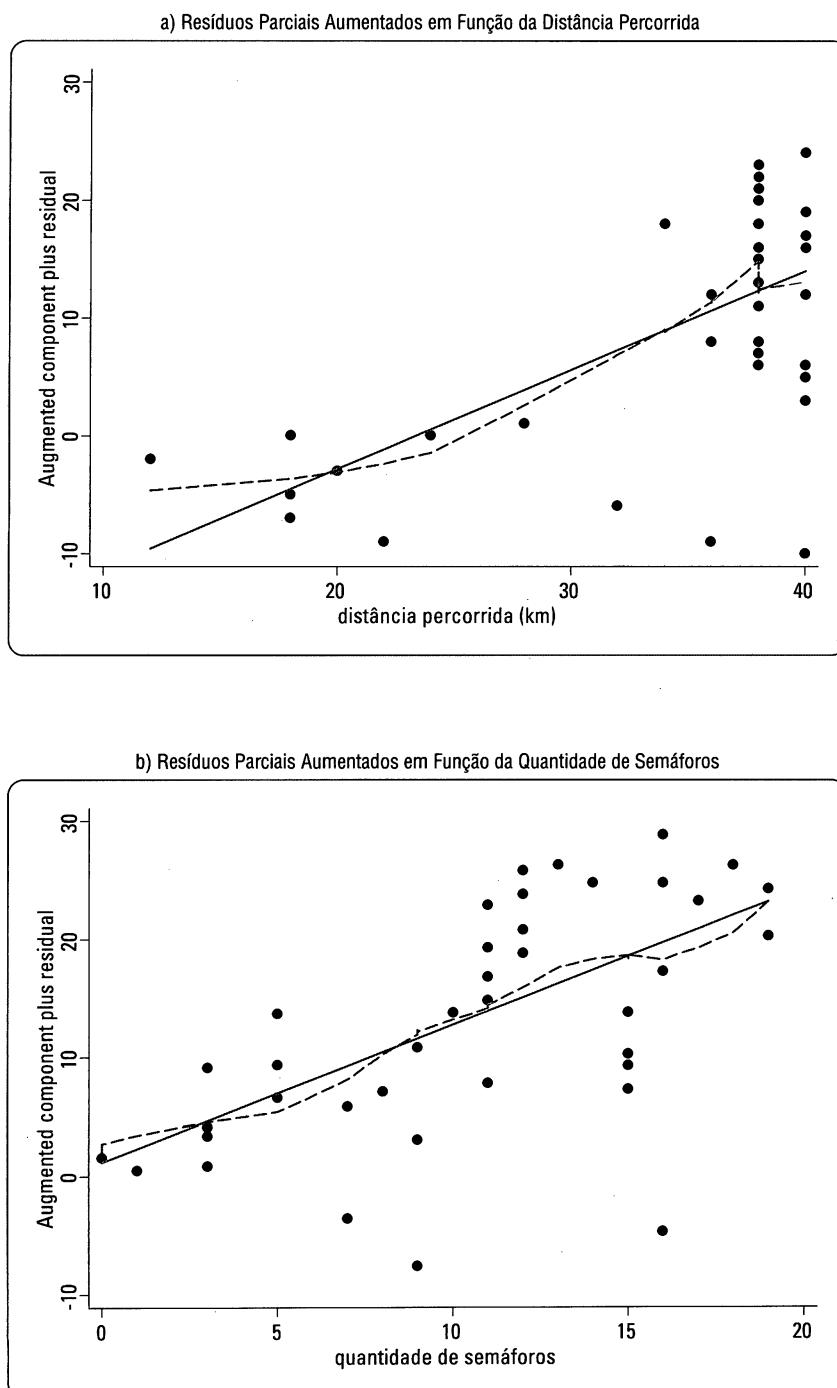
**Figura 12.68** Gráficos com ajuste linear e ajuste *lowess*.

(*component-plus-residuals*) e cada uma das variáveis explicativas. Para a obtenção destes gráficos, devemos digitar os seguintes comandos:

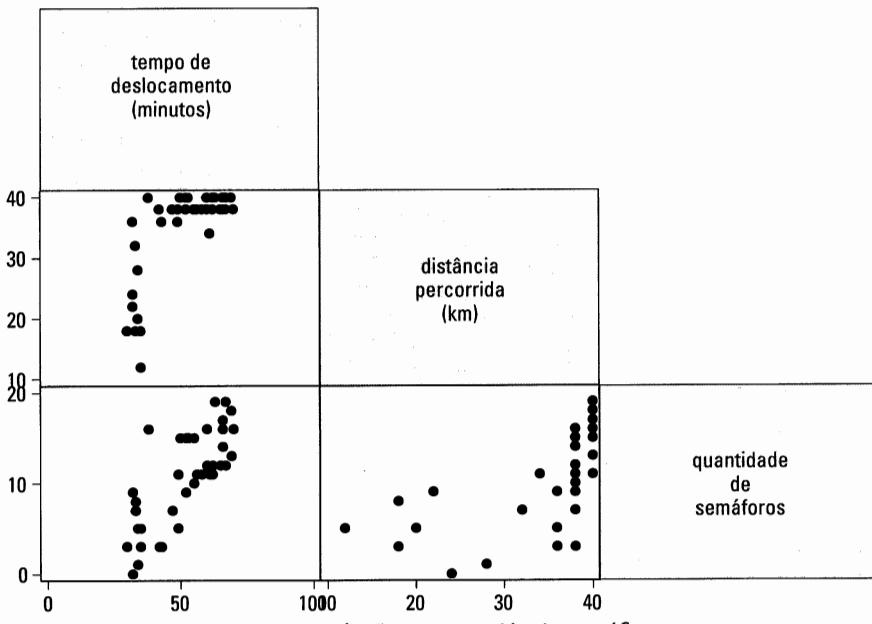
```
acprplot dist, lowess
acprplot sem, lowess
```

A Figura 12.69 apresenta os dois gráficos gerados.

Analogamente à Figura 12.68, o gráfico da Figura 12.69a também mostra que o ajuste *lowess* não se aproxima do ajuste linear, ao contrário do gráfico da Figura 12.69b, o que pode indicar problemas quanto à forma funcional linear da variável *dist* no modelo de regressão. Podemos perceber, para esta variável, que há uma quantidade



**Figura 12.69** Gráficos com ajuste linear e ajuste *lowess* para os resíduos parciais aumentados.



**Figura 12.70** Inter-relação entre variáveis – gráfico **matrix**.

considerável de pontos que potencialmente influenciam o comportamento do modelo. O gráfico **matrix** apresenta claramente este fenômeno, conforme mostra a Figura 12.70, gerada pela digitação do seguinte comando:

```
graph matrix tempo dist sem, half
```

Por meio deste gráfico, verificamos que a relação entre as variáveis *tempo* e *sem* é aparentemente linear, porém a relação entre *tempo* e *dist* é claramente não linear, conforme já discutido. Iremos, desta forma, nos focar na variável *dist*.

Inicialmente, faremos uma transformação logarítmica na variável *dist*, de modo a criarmos a variável *lndist*, como segue:

```
gen lndist=ln(dist)
```

E, desta forma, podemos estimar um novo modelo de regressão, com a seguinte forma funcional:

$$\text{tempo}_i = a + b_1 \cdot \ln \text{dist}_i + b_2 \cdot \text{sem}_i + u_i$$

cujos parâmetros e resultado do teste de Shapiro-Francia para os resíduos podem ser obtidos no Stata pela digitação dos comandos:

```
reg tempo lndist sem
predict res1, res
sfrancia res1
```

e cujos resultados encontram-se na Figura 12.71.

Isto mostra que, embora a transformação logarítmica em variáveis explicativas possa, em alguns casos, melhorar a qualidade do ajuste do modelo, o que não é verdade neste caso, isto ainda não garante que o pressuposto da normalidade dos resíduos seja atendido. O próprio gráfico da Figura 12.72, obtido por meio do comando a seguir, nos mostra que a forma funcional logarítmica da variável *dist* não se ajusta adequadamente à variável *tempo*.

```
acprplot lndist, lowess
```

Desta forma, conforme estudamos na seção 12.4.1, vamos elaborar uma transformação de Box-Cox à variável dependente, de modo que a nova variável criada apresente distribuição com maior aproximação possível da distribuição normal, mesmo que não haja garantia alguma de que esta transformação vá efetivamente gerar uma

```

. gen lndist=ln(dist)
. reg tempo lndist sem

      Source |       SS           df          MS
      Model |  6082.22904        2   3041.11452
      Residual |  2801.39096      47   59.6040629
      Total |    8883.62        49   181.298367

      Number of obs =      50
      F(  2,     47) =   51.02
      Prob > F = 0.0000
      R-squared = 0.6847
      Adj R-squared = 0.6712
      Root MSE = 7.7204

      tempo |     Coef.   Std. Err.      t    P>|t|   [95% Conf. Interval]
      lndist |  18.73429   4.826059     3.88   0.000    9.025515   28.44307
      sem |  1.277542   .2664751     4.79   0.000   .741463   1.81362
      cons | -27.26546   15.31812    -1.78   0.082  -58.08154   3.550618

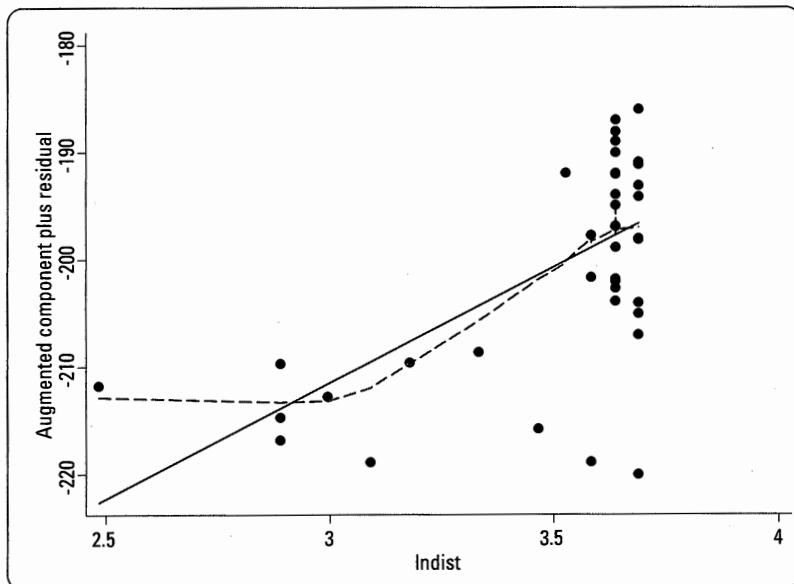
. predict res1, res
. sfrancia res1

      Shapiro-Francia W' test for normal data

      Variable |   Obs      W'      V'      z    Prob>z
      res1 |    50  0.93561   3.339   2.267   0.01168

```

**Figura 12.71** Resultados da estimativa do modelo não linear e do teste de Shapiro-Francia.



**Figura 12.72** Gráfico com ajuste linear e ajuste lowess para os resíduos parciais aumentados em função do logaritmo natural da distância percorrida.

variável com distribuição normal. Para tanto, vamos criar uma variável chamada de *bctempo*, a partir da variável *tempo* e por meio da transformação de Box-Cox. Para tanto, devemos digitar o seguinte comando:

**bcskew0 bctempo = tempo**

A Figura 12.73 apresenta o resultado da transformação de Box-Cox, com ênfase para o parâmetro  $\lambda$  apresentado na expressão (12.66) (parâmetro **L** no *output* do Stata).

Logo, temos que:

$$bctempo_i = \left( \frac{tempo_i^\lambda - 1}{\lambda} \right) = \left( \frac{tempo_i^{2,6486} - 1}{2,6486} \right)$$

. bcskew0 bctempo = tempo			
Transform	L	[95% Conf. Interval]	Skewness
(tempo^L-1)/L	2.648597	(not calculated)	-1.88e-06

Figura 12.73 Transformação de Box-Cox na variável dependente.

O gráfico que mostra o quanto a distribuição da variável *bctempo* (*Kernel density estimate*) se aproxima da distribuição normal padrão pode ser gerado e comparado com o gráfico que considera a variável *tempo* original. Estes gráficos podem ser obtidos por meio dos comandos:

```
kdensity tempo, normal
kdensity bctempo, normal
```

e são apresentados na Figura 12.74.

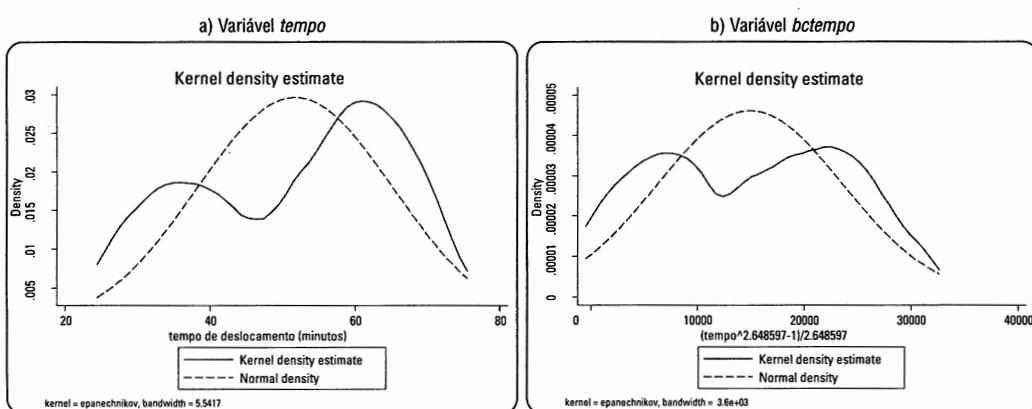


Figura 12.74 Gráfico de aderência entre a distribuição da variável Y e a distribuição normal.

Embora as duas variáveis não apresentem uma aderência muito próxima à normalidade, percebe-se claramente que a maior proximidade se dá com a variável *bctempo*. Vamos, então, estimar o seguinte modelo:

$$bctempo_i = a + b_1 \cdot dist_i + b_2 \cdot sem_i + u_i$$

cujos parâmetros e resultado do teste de Shapiro-Francia para os resíduos podem ser obtidos no Stata pela digitação dos comandos:

```
reg bctempo dist sem
predict res2, res
sfrancia res2
```

e cujos resultados encontram-se na Figura 12.75.

Isto mostra que a aderência da distribuição da variável dependente à normalidade, em modelos de regressão, pode fazer com que sejam estimados, por meio do método de mínimos quadrados ordinários, parâmetros mais adequados à determinação dos intervalos de confiança para efeitos de previsão, já que podem ser gerados termos de erro normais. No apêndice deste capítulo, faremos uma breve apresentação dos modelos de regressão quantílica, que podem ser utilizados alternativamente aos modelos estimados pelo método de mínimos quadrados ordinários para os casos em que nem mesmo a transformação de Box-Cox na variável dependente garante a determinação de resíduos com distribuição aderente à normalidade. Situações como essa podem ocorrer, entre outras razões, quando a variável dependente apresentar considerável assimetria em sua distribuição.

Logo, chegamos ao seguinte modelo:

$$\left( \frac{tempo_i^{2,6486} - 1}{2,6486} \right) = -7.193,16 + 386,6511 \cdot dist_i + 840,903 \cdot sem_i + u_i$$

que apresenta baixo problema de heterocedasticidade (na verdade, apresenta termos de erro homocedásticos ao nível de significância de 1%) e estatísticas *VIF* de 1,83. O próprio gráfico da Figura 12.76 mostra que a

```
. reg bctempo dist sem
      Source |       SS          df        MS
      Model |  2.3519e+09        2  1.1760e+09
      Residual |  1.3171e+09      47  28024387.8
      Total |  3.6691e+09      49  74878715.3
      Number of obs =      50
      F( 2, 47) =    41.96
      Prob > F =    0.0000
      R-squared =   0.6410
      Adj R-squared =  0.6257
      Root MSE =    5293.8

      bctempo |     Coef.    Std. Err.      t    P>|t|    [95% Conf. Interval]
      dist |  386.6511    129.31     2.99    0.004    126.513    646.7892
      sem |   840.903    192.155     4.38    0.000    454.3371   1227.469
      cons | -7193.16   3498.576    -2.06    0.045   -14231.39  -154.9323

      . predict res2, res
      . sfrancia res2
      Shapiro-Francia W' test for normal data
      Variable |   Obs      W'        V'        z    Prob>z
      res2 |   50  0.97217    1.443    0.706    0.24018
```

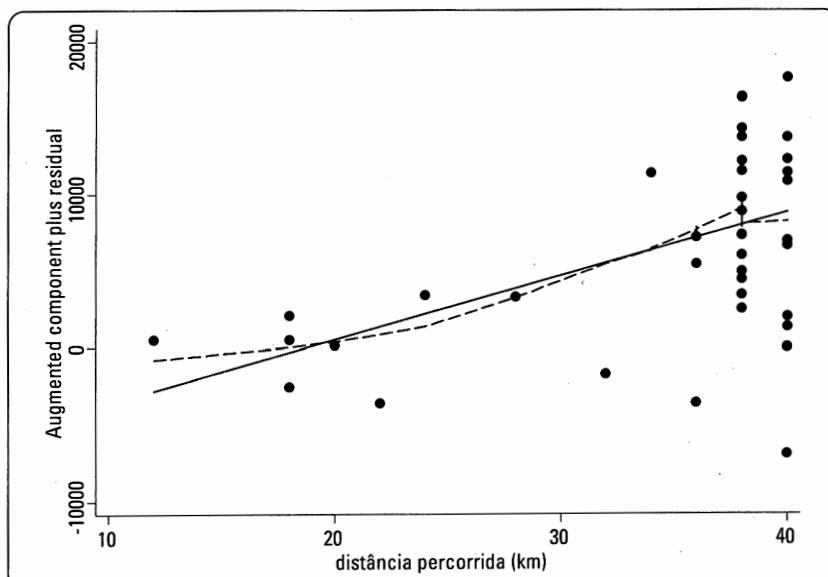
**Figura 12.75** Resultados da estimação do modelo com transformação de Box-Cox na variável dependente e do teste de Shapiro-Francia.

transformação de Box-Cox na variável dependente aproxima consideravelmente o ajuste estimado ao ajuste *lowess*. Tal gráfico pode ser obtido por meio do comando:

**acprplot dist, lowess**

Logo, caberá ao pesquisador, em função do diagnóstico dos dados que sempre precisará ser feito, em função da sua experiência e com base na teoria subjacente, definir uma adequada forma funcional quando da estimação de modelos de regressão, a fim de que se atendam os pressupostos e que sejam obtidos estimadores mais eficientes para a elaboração de previsões.

Por fim, iremos agora estudar o problema da autocorrelação dos resíduos por meio do Stata. Imaginemos que o professor, ao terminar a palestra e voltar para a escola, tenha tido a ideia de acompanhar o tempo de percurso dos alunos ao longo de um período de 30 dias. Para tanto, dia após dia ele coletou os dados dos alunos referentes ao tempo de deslocamento, à distância percorrida e à quantidade de semáforos. Só que, ao invés de elaborar o banco de dados por aluno e por dia, o que resultaria num painel de dados longitudinais (que estudaremos no Capítulo 15), o professor tabulou os



**Figura 12.76** Gráfico com ajuste linear e ajuste *lowess* para os resíduos parciais aumentados em função da distância percorrida para o modelo com transformação de Box-Cox.

dados médios de cada variável por dia, ou seja, o tempo médio de trajeto percorrido por dia, a distância média percorrida pelos alunos em cada dia e a quantidade média de semáforos. O objetivo do professor agora (e o nosso também) é estimar o seguinte modelo:

$$\text{tempo}_t = a + b_1 \cdot \text{dist}_t + b_2 \cdot \text{sem}_t + \varepsilon_t \quad (t = 1, 2, \dots, 30)$$

e o banco de dados encontra-se no arquivo **Análisetemporaltempodistsem.dta**.

Antes de estimarmos o modelo proposto, é preciso que seja definida a variável correspondente à evolução temporal (no caso, a variável *dia*). Para tanto, devemos digitar, logo ao abrir o arquivo, o seguinte comando:

**tsset dia**

Uma informação como a que aparece na Figura 12.77 surgirá na tela.

Caso o pesquisador se esqueça de definir a variável referente à evolução temporal, o que é muito comum, o Stata não permitirá que sejam elaborados os testes de Durbin-Watson e de Breusch-Godfrey, e uma mensagem de erro aparecerá na janela de *outputs* do software, informando ao pesquisador que a variável temporal precisa ser definida. Por outro lado, diversos pacotes estatísticos, como o SPSS, propiciam o cálculo das estatísticas de Durbin-Watson, por exemplo, mesmo que o banco de dados esteja em *cross-section*, o que é um erro grave.

```
. tsset dia
      time variable: dia, 1 to 30
      delta: 1 unit
```

**Figura 12.77** Definição da variável temporal.

Após a elaboração da regressão propriamente dita, por meio do comando a seguir, poderemos então elaborar os testes voltados à verificação de existência de autocorrelação dos resíduos.

**reg tempo dist sem**

Os resultados da estimação encontram-se na Figura 12.78.

Embora o modelo estimado apresente problemas, ao nível de significância de 5%, em relação à normalidade dos resíduos (teste de Shapiro-Wilk) e à heterocedasticidade (teste de Breusch-Pagan/Cook-Weisberg), restrin- giremos a análise, neste momento, à autocorrelação dos resíduos. Para tanto, iremos inicialmente elaborar o teste de Durbin-Watson, por meio do seguinte comando:

**estat dwatson**

<b>. reg tempo dist sem</b>					
Source	SS	df	MS	Number of obs = 30	
Model	3642.45366	2	1821.22683	F( 2, 27) = 34.17	
Residual	1438.91301	27	53.2930744	Prob > F = 0.0000	
Total	5081.36667	29	175.21954	R-squared = 0.7168	
				Adj R-squared = 0.6958	
				Root MSE = 7.3002	
<b>tempo   Coef. Std. Err. t P&gt; t  [95% Conf. Interval]</b>					
dist   .7816866 .2019979 3.87 0.001 .3672211 1.196152					
sem   1.040915 .3335171 3.12 0.004 .3565945 1.725236					
cons   14.32001 5.508772 2.60 0.015 3.016943 25.62308					

**Figura 12.78** Resultados da estimação do modelo temporal.

O resultado do teste encontra-se na Figura 12.79.

Por meio da Tabela C do apêndice do livro, e de acordo com a Figura 12.45 da seção 12.3.4.3, temos, ao nível de significância de 5% e para um modelo com 3 parâmetros e 30 observações, que  $d_U = 1,567 < 1,779 < 2,433 = 4 - d_U$ , ou seja, a estatística *DW* aproximadamente igual a 2 resulta em inexistência de autocorrelação de primeira ordem dos resíduos.

Conforme discutido na seção 12.3.4.4, como o teste de Durbin-Watson só é válido para a verificação da existência de autocorrelação de primeira ordem dos termos de erro, o teste de Breusch-Godfrey passa a ser mais geral

```
. estat dwatson
Durbin-Watson d-statistic( 3, 30) = 1.779404
```

**Figura 12.79** Resultado do teste de Durbin-Watson.

na medida em que também é adequado para avaliar a existência de autocorrelação dos resíduos com defasagens maiores. Numa base com dados diários, por exemplo, talvez seja interessante que o pesquisador estude eventuais autocorrelações de ordem 7, a fim de que sejam capturados fenômenos com sazonalidade semanal. Seguindo a mesma lógica, para dados mensais, talvez seja interessante que o pesquisador avalie a existência de eventuais autocorrelações de ordem 12, a fim de tentar capturar sazonalidades anuais.

Para fins didáticos, no nosso exemplo vamos elaborar o teste de Breusch-Godfrey com todas as defasagens possíveis para este banco de dados, ou seja, com ordens que variam de 1 a 28 ( $t - 1, t - 2, t - 3, \dots, t - 28$ ). O comando a ser digitado é:

```
estat bgodfrey, lags(1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
21 22 23 24 25 26 27 28)
```

Os resultados encontram-se na Figura 12.80.

Por meio da Figura 12.80, podemos perceber que não há problemas de autocorrelação dos resíduos para qualquer que seja a defasagem proposta.

A capacidade do Stata para a estimação de modelos e a elaboração de testes estatísticos é enorme, porém acreditamos que o que foi exposto aqui é considerado obrigatório para pesquisadores que desejam utilizar de forma correta as técnicas de regressão simples e múltipla.

Partiremos agora para a resolução dos mesmos exemplos por meio do SPSS, ressaltando que, embora a sua capacidade de processamento e geração de *outputs* seja considerada por muitos como sendo mais limitada do que a do Stata, é tido por vezes como um software mais amigável e mais fácil de ser utilizado.

Breusch-Godfrey LM test for autocorrelation			
lags(p)	chi2	df	Prob > chi2
1	0.213	1	0.6447
2	1.478	2	0.4775
3	2.292	3	0.5140
4	3.137	4	0.5352
5	3.138	5	0.6787
6	3.658	6	0.7228
7	4.382	7	0.7349
8	4.423	8	0.8171
9	4.765	9	0.8543
10	5.176	10	0.8791
11	5.181	11	0.9221
12	15.487	12	0.2159
13	17.025	13	0.1982
14	17.644	14	0.2235
15	18.444	15	0.2400
16	18.623	16	0.2887
17	19.119	17	0.3217
18	19.157	18	0.3822
19	20.730	19	0.3519
20	20.831	20	0.4072
21	22.068	21	0.3956
22	22.186	22	0.4488
23	26.104	23	0.2960
24	26.155	24	0.3453
25	26.169	25	0.3986
26	28.427	26	0.3378
27	30.000	27	0.3142
28	30.000	28	0.3632

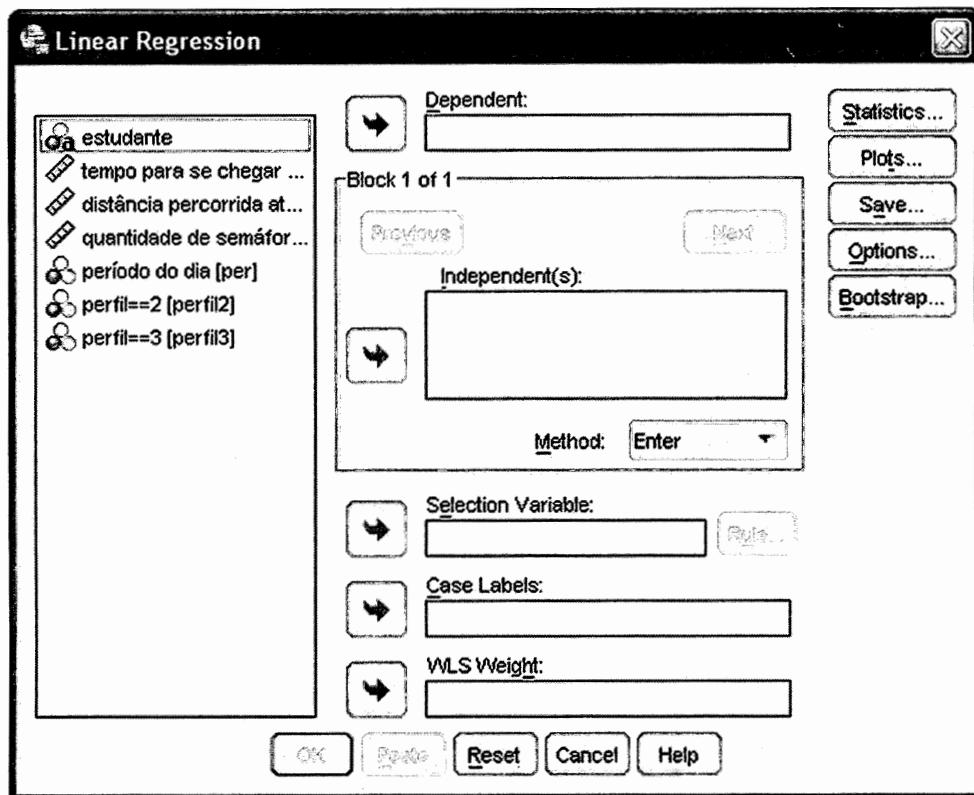
H0: no serial correlation

**Figura 12.80** Resultados do teste de Breusch-Godfrey.

## 12.6. ESTIMAÇÃO DE MODELOS DE REGRESSÃO NO SOFTWARE SPSS

Apresentaremos agora o passo a passo para a elaboração do nosso exemplo por meio do IBM SPSS Statistics Software®, e a reprodução de suas imagens nesta seção tem autorização da International Business Machines Corporation®.

Seguindo a mesma lógica proposta quando da aplicação dos modelos por meio do software Stata, já partiremos para o banco de dados final construído pelo professor a partir dos questionamentos feitos a cada um de seus 10 estudantes. Os dados encontram-se no arquivo **Tempodistsempperfil.sav** e, após o abrirmos, vamos inicialmente clicar em **Analyze → Regression → Linear....** A caixa de diálogo da Figura 12.81 será aberta.



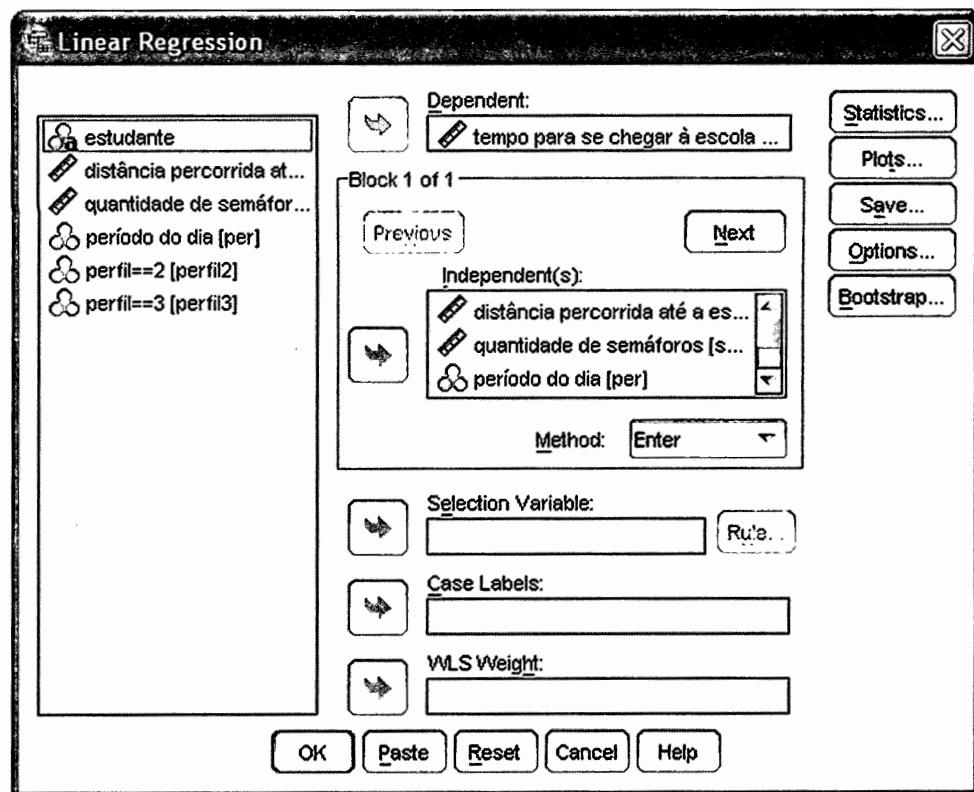
**Figura 12.81** Caixa de diálogo para elaboração da regressão linear no SPSS.

Devemos selecionar a variável *tempo* e incluí-la na caixa **Dependent**. As demais variáveis devem ser simultaneamente selecionadas e inseridas na caixa **Independent(s)**. Manteremos, neste primeiro momento, a opção pelo **Method: Enter**, conforme podemos observar por meio da Figura 12.82. O procedimento *Enter*, ao contrário do procedimento *Stepwise*, inclui todas as variáveis na estimação, mesmo aquelas cujos parâmetros sejam estatisticamente iguais a zero, e corresponde exatamente ao procedimento padrão elaborado pelo Excel e também pelo Stata quando se aplica o comando **reg**.

O botão **Statistics...** permite que selezionemos a opção que fornecerá os parâmetros e os respectivos intervalos de confiança nos *outputs*. A caixa de diálogo que é aberta, ao clicarmos nesta opção, está apresentada na Figura 12.83, em que foram selecionadas as opções **Estimates** (para que sejam apresentados os parâmetros propriamente ditos com as respectivas estatísticas *t*) e **Confidence intervals** (para que sejam calculados os intervalos de confiança destes parâmetros).

Voltaremos à caixa de diálogo principal da regressão linear ao clicarmos em **Continue**.

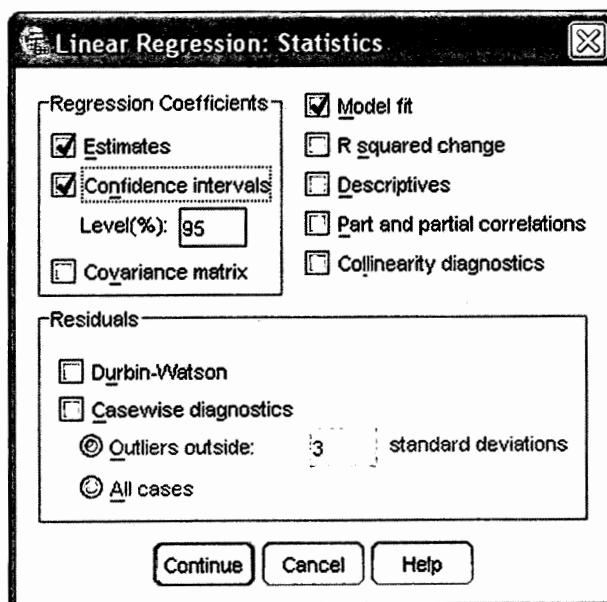
O botão **Options...** permite que alteremos os níveis de significância para rejeição da hipótese nula do teste *F* e, consequentemente, das hipóteses nulas dos testes *t*. O padrão do SPSS, conforme pode ser observado por meio da caixa de diálogo que é aberta ao clicarmos nesta opção, é de 5% para o nível de significância. Nesta mesma caixa de diálogo, podemos impor que o parâmetro  $\alpha$  seja igual a zero (ao desabilitarmos a opção **Include constant in equation**). Manteremos o padrão de 5% para os níveis de significância e deixaremos o intercepto no modelo (opção **Include constant in equation** selecionada). Esta caixa de diálogo é apresentada na Figura 12.84.



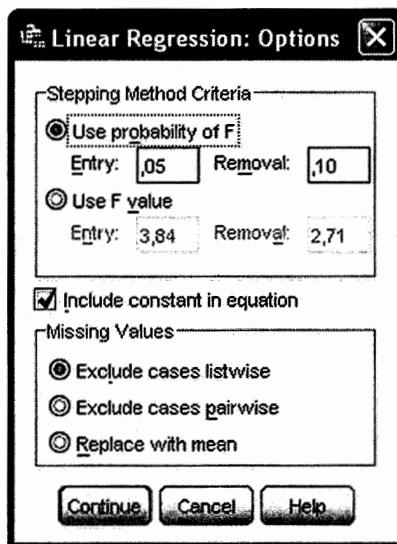
**Figura 12.82** Caixa de diálogo para elaboração da regressão linear no SPSS com inclusão da variável dependente e das variáveis explicativas e seleção do procedimento *Enter*.

Vamos agora selecionar **Continue** e **OK**. Os *outputs* gerados estão apresentados na Figura 12.85.

Não iremos novamente analisar *outputs* gerados, uma vez que podemos verificar que são exatamente iguais àqueles obtidos quando da elaboração da regressão linear múltipla no Excel (Figura 12.32) e no Stata (Figura 12.53). Vale a pena comentar que o *F de significação* do Excel é chamado de *Sig. F* e o *valor-P* é chamado de *Sig. t* no SPSS.



**Figura 12.83** Caixa de diálogo para seleção dos parâmetros e dos intervalos de confiança.



**Figura 12.84** Caixa de diálogo para eventual alteração dos níveis de significância e exclusão do intercepto em modelos de regressão linear.

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,998 <sup>a</sup>	,997	,993	1,229

a. Predictors: (Constant), perfil==3, quantidade de semáforos, perfil==2, distância percorrida até a escola (km), período do dia

ANOVA <sup>b</sup>					
Model	Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1993,960	5	398,792	264,120
	Residual	6,040	4	1,510	
	Total	2000,000	9		

a. Predictors: (Constant), perfil==3, quantidade de semáforos, perfil==2, distância percorrida até a escola (km), período do dia

b. Dependent Variable: tempo para se chegar à escola (minutos)

Model	Coefficients <sup>a</sup>						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
1	B	Std. Error	Beta			Lower Bound	Upper Bound
	(Constant)	13,490	3,861	3,494	,025	2,771	24,210
	distância percorrida até a escola (km)	,874	,072	,430	9,399	,001	,475
	quantidade de semáforos	6,647	1,095	,420	6,071	,004	3,607
	período do dia	-5,371	3,779	-,174	-1,421	,228	-15,863
	perfil==2	1,779	1,441	,063	1,234	,285	-2,223
	perfil==3	6,374	2,243	,180	2,841	,047	,146

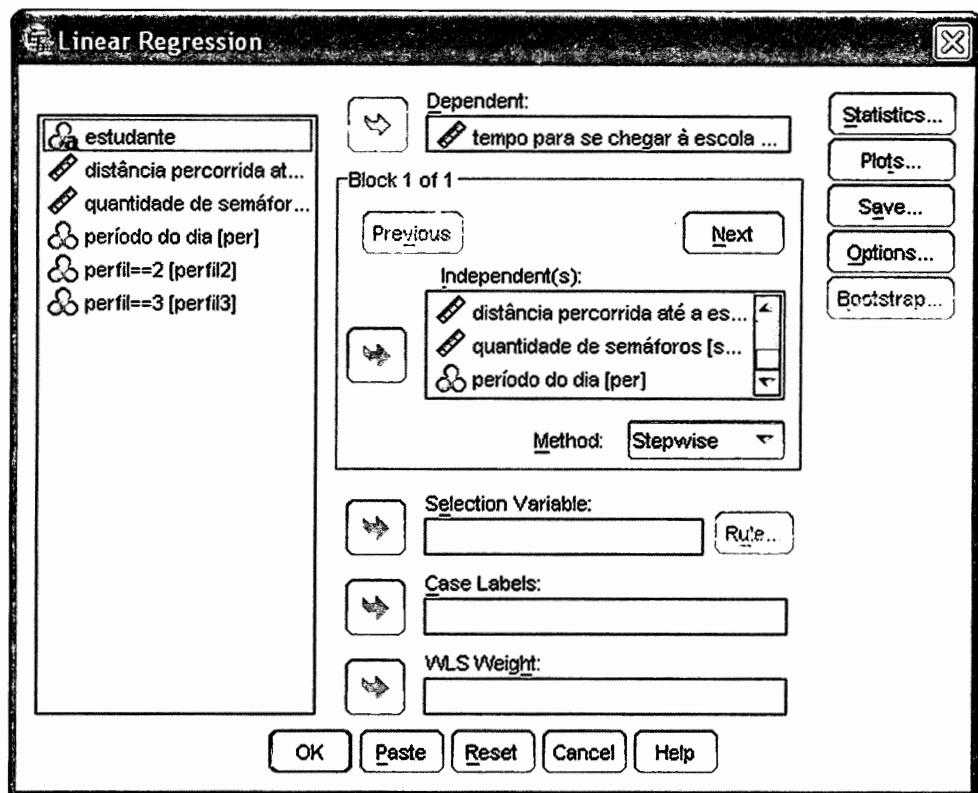
a. Dependent Variable: tempo para se chegar à escola (minutos)

**Figura 12.85** Outputs da regressão linear múltipla no SPSS – procedimento Enter.

Vamos agora, enfim, elaborar a regressão linear múltipla por meio do procedimento *Stepwise*. Para elaborarmos este procedimento, devemos selecionar a opção **Method: Stepwise** na caixa de diálogo principal da regressão linear no SPSS, conforme mostra a Figura 12.86.

Voltaremos novamente à caixa de diálogo principal da regressão linear ao clicarmos em **Continue**.

O botão **Save...** permite que sejam criadas, no próprio banco de dados original, as variáveis referentes ao  $\hat{Y}$  e aos resíduos do modelo final gerado pelo procedimento *Stepwise*. Sendo assim, ao clicarmos nesta opção, será aberta uma caixa de diálogo, conforme mostra a Figura 12.87. Com esta finalidade, devemos marcar as opções **Unstandardized** (em **Predicted Values**) e **Unstandardized** (em **Residuals**).



**Figura 12.86** Caixa de diálogo com seleção do procedimento *Stepwise*.

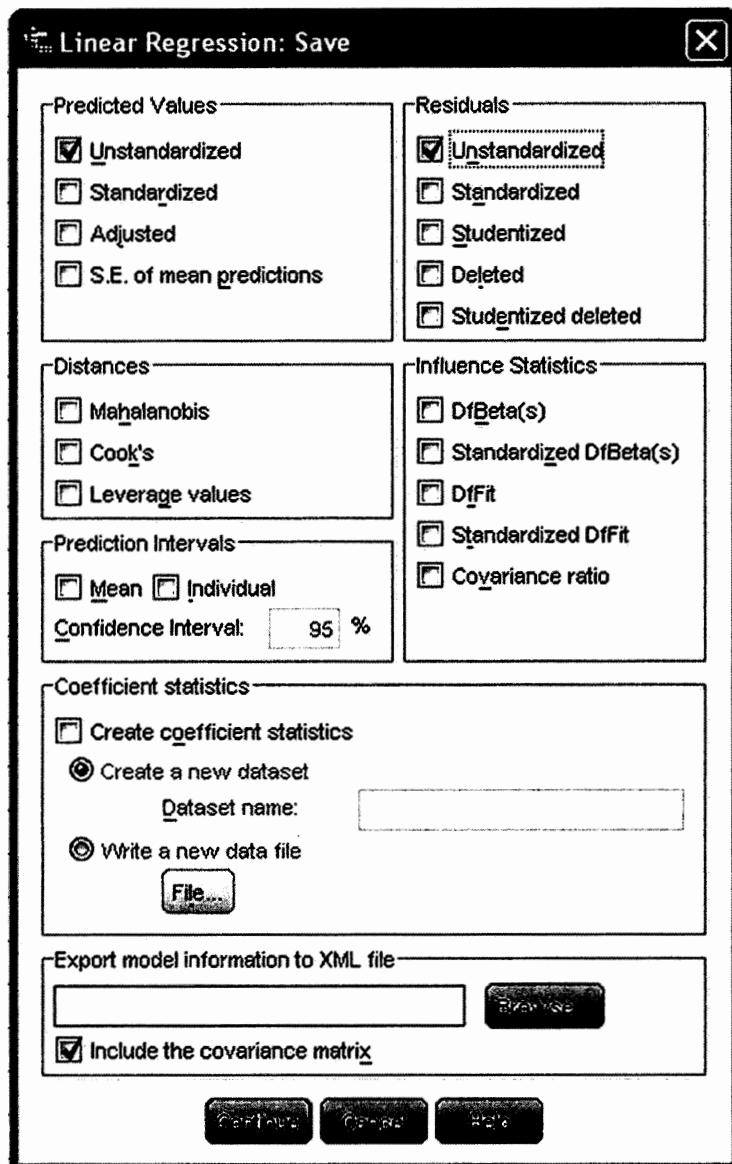
Ao clicarmos em **Continue** e, na sequência, em **OK**, novos *outputs* são gerados, conforme mostra a Figura 12.88. Note que, além dos *outputs*, são criadas duas novas variáveis no banco de dados original, chamadas de *PРЕ\_1* e *RES\_1*, que correspondem, respectivamente, aos valores de  $\hat{Y}$  e aos valores estimados dos resíduos (exatamente aqueles já mostrados na Figura 12.33).

O procedimento *Stepwise* elaborado pelo SPSS mostra o passo a passo dos modelos que foram elaborados, partindo da inclusão da variável mais significativa (maior estatística  $t$  em módulo entre todas as explicativas) até a inclusão daquela com menor estatística  $t$ , porém ainda com  $Sig. t < 0,05$ . Tão importante quanto a análise das variáveis incluídas no modelo final é a análise da lista de variáveis excluídas (**Excluded Variables**). Assim, podemos verificar que, ao se incluir no modelo 1 apenas a variável explicativa *sem*, a lista de variáveis excluídas apresenta todas as demais. Se, para o primeiro passo, houver alguma variável explicativa que tenha sido excluída, porém apresenta-se de forma significativa ( $Sig. t < 0,05$ ), como ocorre para a variável *dist*, esta será incluída no modelo no passo seguinte (modelo 2). E assim sucessivamente, até que a lista de variáveis excluídas não apresente mais nenhuma variável com  $Sig. t < 0,05$ . As variáveis remanescentes nesta lista, para o nosso exemplo, são *per* e *perfil2*, conforme já discutimos quando da elaboração da regressão no Excel e no Stata; o modelo final (modelo 3 do procedimento *Stepwise*), que é exatamente aquele já apresentado por meio das Figuras 12.33 e 12.54, conta apenas com as variáveis explicativas *dist*, *sem* e *perfil3*, e com  $R^2 = 0,995$ . Assim, conforme já vimos, o modelo linear final estimado é:

$$\hat{\text{tempo}}_i = 8,292 + 0,710 \cdot \text{dist}_i + 7,837 \cdot \text{sem}_i + 8,968 \cdot \text{perfil3}_i \begin{cases} \text{calmo}=0 \\ \text{agressivo}=1 \end{cases}$$

Partiremos agora para a verificação dos pressupostos do modelo. Inicialmente, vamos elaborar o teste de Shapiro-Wilk para verificação de normalidade dos resíduos. Para tanto, devemos clicar em **Analyze** → **Descriptive Statistics** → **Explore**.... Na caixa de diálogo que é aberta, devemos inserir a variável *RES\_1* (*Unstandardized Residual*) em **Dependent List** e clicar em **Plots...**. Nesta janela, devemos selecionar a opção **Normality plots with tests**, clicar em **Continue** e em **OK**. A Figura 12.89 mostra este passo a passo.

O teste de Shapiro-Wilk indica que os termos de erro apresentam distribuição aderente à normalidade, já que seu resultado (Figura 12.90) não indica a rejeição de sua hipótese nula. Podemos verificar que o resultado é exatamente igual ao obtido pelo Stata e apresentado por meio da Figura 12.58.



**Figura 12.87** Caixa de diálogo para inserção dos valores previstos ( $\hat{Y}$ ) e dos resíduos no próprio banco de dados.

Na sequência, vamos elaborar o diagnóstico de multicolinearidade das variáveis explicativas. Para tanto, devemos solicitar ao software que gere as estatísticas *VIF* e *Tolerance* quando for feita a estimativa do modelo. Assim, em **Analyze → Regression → Linear...**, no botão **Statistics...** devemos marcar a opção **Collinearity diagnostics**, conforme mostra a Figura 12.91.

Os *outputs* gerados são os mesmos dos apresentados na Figura 12.88, porém agora as estatísticas *VIF* e *Tolerance* são calculadas para cada variável explicativa, conforme mostra o modelo 3 da Figura 12.92. Conforme já discutido quando da apresentação da Figura 12.60, como o modelo final obtido após o procedimento *Stepwise* não apresenta estatísticas *VIF* muito elevadas para nenhuma variável explicativa, podemos considerar que não há problemas de multicolinearidade.

Com relação ao problema da heterocedasticidade, o mais comum é que se elabore inicialmente um gráfico para se avaliar o comportamento dos resíduos em função da variável dependente. Assim, devemos novamente clicar em **Analyze → Regression → Linear....**. O botão **Plots...** permite que sejam elaborados gráficos de diagnóstico do comportamento dos resíduos em função dos valores estimados da variável dependente e, ao clicarmos neste botão, será aberta uma caixa de diálogo, conforme mostra a Figura 12.93. Vamos solicitar que seja gerado o gráfico dos valores estimados dos termos de erro padronizados em função dos valores estimados padronizados da variável dependente. Este procedimento é análogo ao que gerou o gráfico da Figura 12.61b.

**Model Summary<sup>d</sup>**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,909 <sup>a</sup>	,827	,805	6,585
2	,968 <sup>b</sup>	,937	,920	4,228
3	,998 <sup>c</sup>	,995	,993	1,236

- a. Predictors: (Constant), quantidade de semáforos  
 b. Predictors: (Constant), quantidade de semáforos, distância percorrida até a escola (km)  
 c. Predictors: (Constant), quantidade de semáforos, distância percorrida até a escola (km), perfil=3  
 d. Dependent Variable: tempo para se chegar à escola (minutos)

**ANOVA<sup>d</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1653,125	1	1653,125	38,126	,000 <sup>a</sup>
	Residual	346,875	8	43,356		
	Total	2000,000	9			
2	Regression	1874,848	2	937,424	52,432	,000 <sup>b</sup>
	Residual	125,152	7	17,879		
	Total	2000,000	9			
3	Regression	1990,839	3	663,613	434,616	,000 <sup>c</sup>
	Residual	9,161	6	1,527		
	Total	2000,000	9			

- a. Predictors: (Constant), quantidade de semáforos  
 b. Predictors: (Constant), quantidade de semáforos, distância percorrida até a escola (km)  
 c. Predictors: (Constant), quantidade de semáforos, distância percorrida até a escola (km), perfil=3  
 d. Dependent Variable: tempo para se chegar à escola (minutos)

**Coefficients<sup>a</sup>**

Model	Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.	95,0% Confidence Interval for B	
	B	Std. Error				Lower Bound	Upper Bound
1	(Constant)	15,625	3,123	5,003	,001	8,422	22,828
	quantidade de semáforos	14,375	2,328			9,006	19,744
2	(Constant)	8,151	2,920	2,791	,027	1,246	15,056
	quantidade de semáforos	8,296	2,284			2,897	13,696
	distância percorrida até a escola (km)	,797	,226			,262	1,333
3	(Constant)	8,292	,854	9,715	,000	6,203	10,380
	quantidade de semáforos	7,837	,669			6,199	9,475
	distância percorrida até a escola (km)	,710	,067			,547	,874
	perfil=3	8,968	1,029			6,450	11,485

- a. Dependent Variable: tempo para se chegar à escola (minutos)

**Excluded Variables<sup>d</sup>**

Model	Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
					Tolerance
1	distância percorrida até a escola (km)	,509 <sup>a</sup>	3,522	,010	,800 ,429
	período do dia	-,395 <sup>a</sup>	-2,237	,060	-,646 ,464
	perfil=2	-,084 <sup>a</sup>	-,529	,813	-,196 ,950
	perfil=3	,300 <sup>a</sup>	2,528	,039	,681 ,922
2	período do dia	-,321 <sup>b</sup>	-4,164	,006	-,882 ,451
	perfil=2	-,116 <sup>b</sup>	-1,233	,264	-,450 ,942
	perfil=3	,254 <sup>b</sup>	8,716	,000	,963 ,901
3	período do dia	-,050 <sup>c</sup>	-,702	,514	-,299 ,124
	perfil=2	,007 <sup>c</sup>	,198	,851	,088 ,717

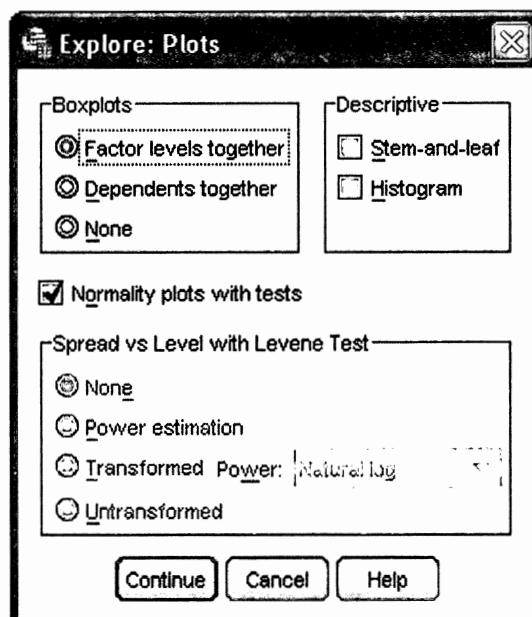
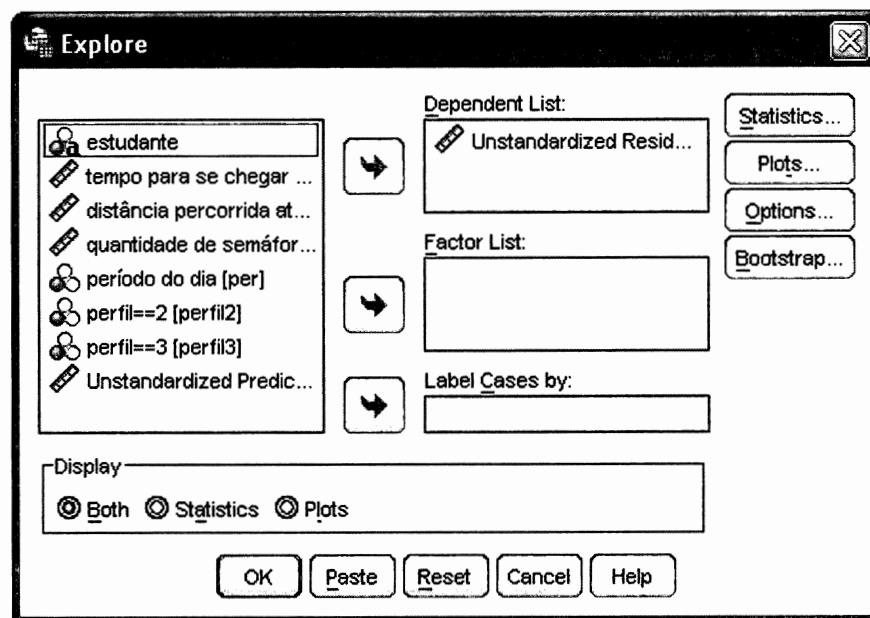
- a. Predictors in the Model: (Constant), quantidade de semáforos  
 b. Predictors in the Model: (Constant), quantidade de semáforos, distância percorrida até a escola (km)  
 c. Predictors in the Model: (Constant), quantidade de semáforos, distância percorrida até a escola (km), perfil=3  
 d. Dependent Variable: tempo para se chegar à escola (minutos)

**Residuals Statistics<sup>a</sup>**

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	11,84	54,54	30,00	14,873	10
Residual	-1,844	1,056	,000	1,009	10
Std. Predicted Value	-1,221	1,650	,000	1,000	10
Std. Residual	-1,492	,855	,000	,816	10

- a. Dependent Variable: tempo para se chegar à escola (minutos)

**Figura 12.88** Outputs da regressão linear múltipla no SPSS – procedimento Stepwise.



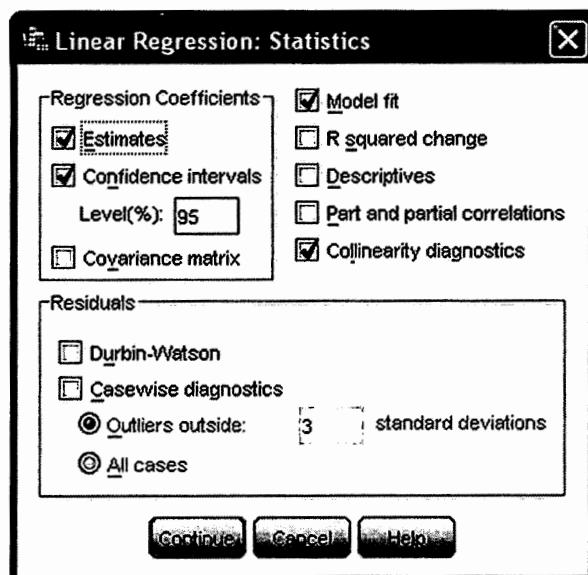
**Figura 12.89** Procedimento para elaboração do teste de Shapiro-Wilk para a variável *RES\_1*.

	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	,177	10	,200*	,905	10	,250

a. Lilliefors Significance Correction

\*. This is a lower bound of the true significance.

**Figura 12.90** Resultado do teste de normalidade de Shapiro-Wilk para os resíduos.

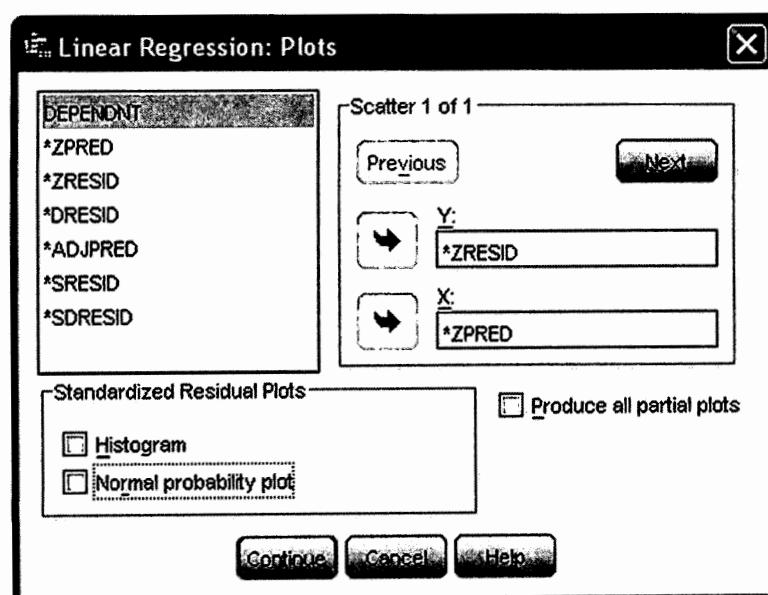


**Figura 12.91** Caixa de diálogo para elaboração do diagnóstico de multicolinearidade.

Model	Coefficients <sup>a</sup>								
	Unstandardized Coefficients			t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
	B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1 (Constant)	15,625	3,123		5,003	,001	8,422	22,828	1,000	1,000
quantidade de semáforos	14,375	2,328	,909	6,175	,000	9,006	19,744		
2 (Constant)	8,151	2,920		2,791	,027	1,246	15,056	,429	2,333
quantidade de semáforos	8,296	2,284	,525	3,633	,008	2,897	13,696		
distância percorrida até a escola (km)	,797	,226	,509	3,522	,010	,262	1,333		
3 (Constant)	8,292	,854		9,715	,000	6,203	10,380	,426	2,348
quantidade de semáforos	7,837	,669	,496	11,707	,000	6,199	9,475		
distância percorrida até a escola (km)	,710	,067	,453	10,620	,000	,547	,874		
perfil==3	8,968	1,029	,254	8,716	,000	6,450	11,485		

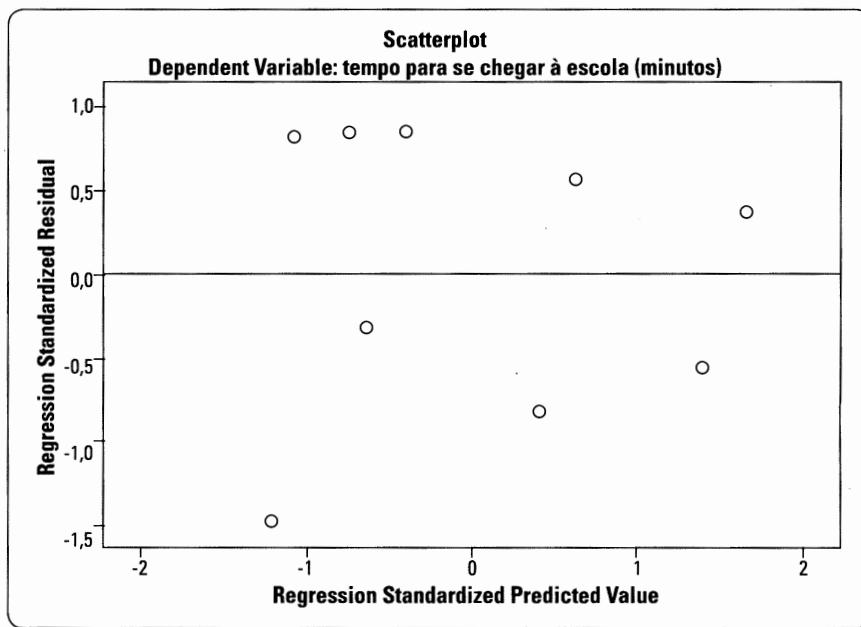
a. Dependent Variable: tempo para se chegar à escola (minutos)

**Figura 12.92** Estatísticas VIF e Tolerance das variáveis explicativas.



**Figura 12.93** Caixa de diálogo para elaboração do gráfico de diagnóstico do comportamento dos resíduos em função da variável dependente.

O gráfico gerado, apresentado na Figura 12.94, mostra que não há indícios de existência de heterocedasticidade, conforme já discutimos quando da análise da Figura 12.61b.



**Figura 12.94** Gráfico de diagnóstico do comportamento dos resíduos em função da variável dependente.

Embora o SPSS não possua uma opção direta para realização do teste de Breusch-Pagan/Cook-Weisberg, iremos construir o procedimento para a sua elaboração no SPSS. Assim, vamos inicialmente criar uma nova variável, que chamaremos de *RES\_1SQ* e que se refere ao quadrado dos resíduos. Para tanto, em **Transform → Compute Variable...**, devemos proceder como mostra a Figura 12.95. No SPSS, o duplo asterisco corresponde ao operador expoente.

Feito isso, vamos calcular a soma dos resíduos ao quadrado, clicando em **Analyze → Descriptive Statistics → Descriptives...** e marcando a opção **Sum** no botão **Options...**, conforme mostra a Figura 12.96.

A soma dos termos da variável *RES\_1SQ* é 9,16137, o que está de acordo com o apresentado na Tabela 12.17. Vamos agora criar uma nova variável, chamada de *RESUP*, em que:

$$RESUP_i = \frac{RES\_1SQ_i}{\left( \sum_{i=1}^n RES\_1SQ \right) / n} = \frac{RES\_1SQ_i}{(9,16137) / 10}$$

segundo a expressão (12.40). Logo, em **Transform → Compute Variable...** devemos proceder de acordo com o apresentado na Figura 12.97.

Na sequência, devemos elaborar a regressão de *RESUP* em função dos valores estimados da variável dependente, ou seja, em função da variável *PRED\_1*. Não iremos mostrar todos os *outputs* desta estimação, porém a Figura 12.98 apresenta a tabela ANOVA resultante.

Por meio da tabela ANOVA, verificamos que a soma dos quadrados da regressão (*SQR*) é 3,185 que, dividindo-se por 2, chega-se à estatística  $\chi^2_{BP/CW} = 1,59 < \chi^2_{1 g.l.} = 3,84$  para o nível de significância de 5%, ou seja, a hipótese nula do teste (termos de erro homocedásticos) não pode ser rejeitada, conforme também já foi analisado por meio da Figura 12.62.

Seguindo a lógica apresentada na seção 12.5, vamos, neste momento, abrir o arquivo **Palestratempodistsem.sav** e estimar o seguinte modelo de regressão não linear:

$$tempo_i = a + b_1 \cdot \ln dist_i + b_2 \cdot sem_i + u_i$$

Para tanto, precisamos criar a variável *ln dist* (Figura 12.99), clicando em **Transform → Compute Variable....**

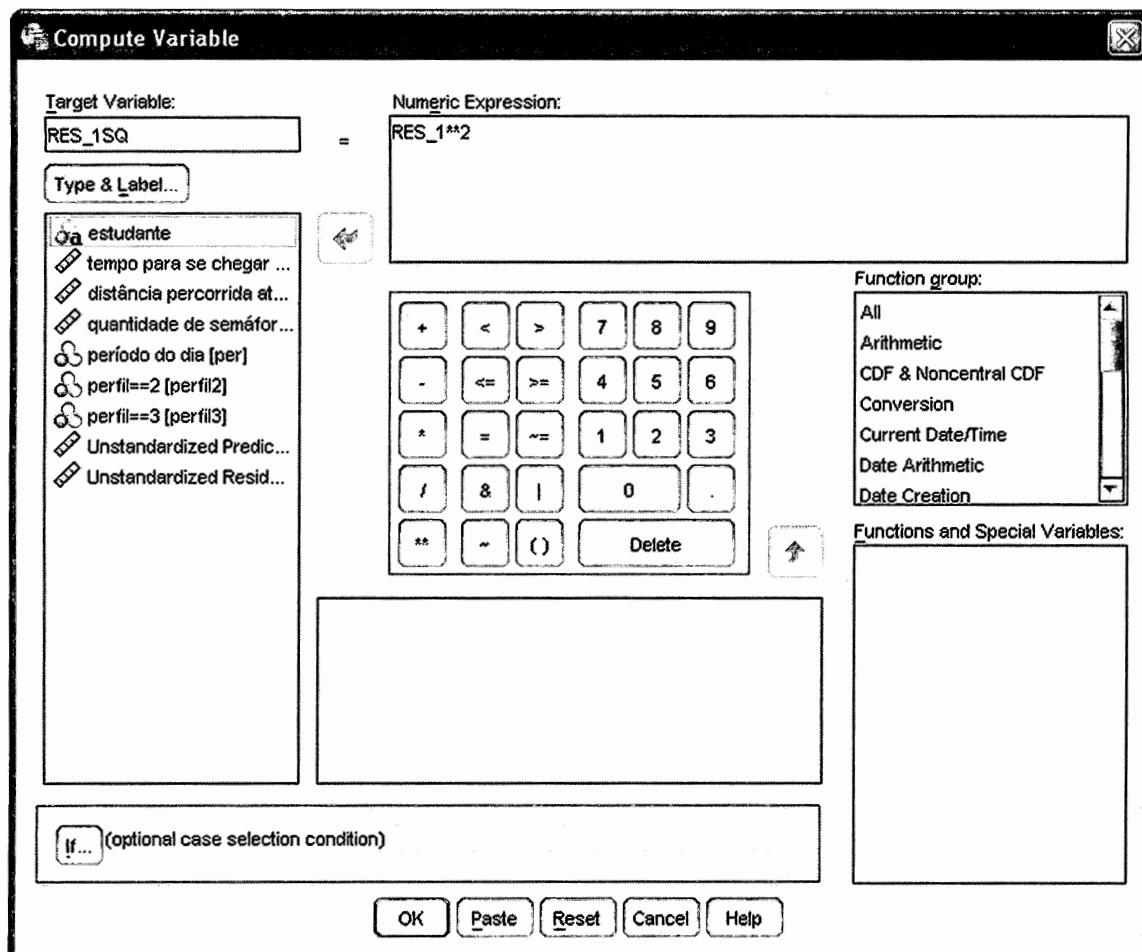
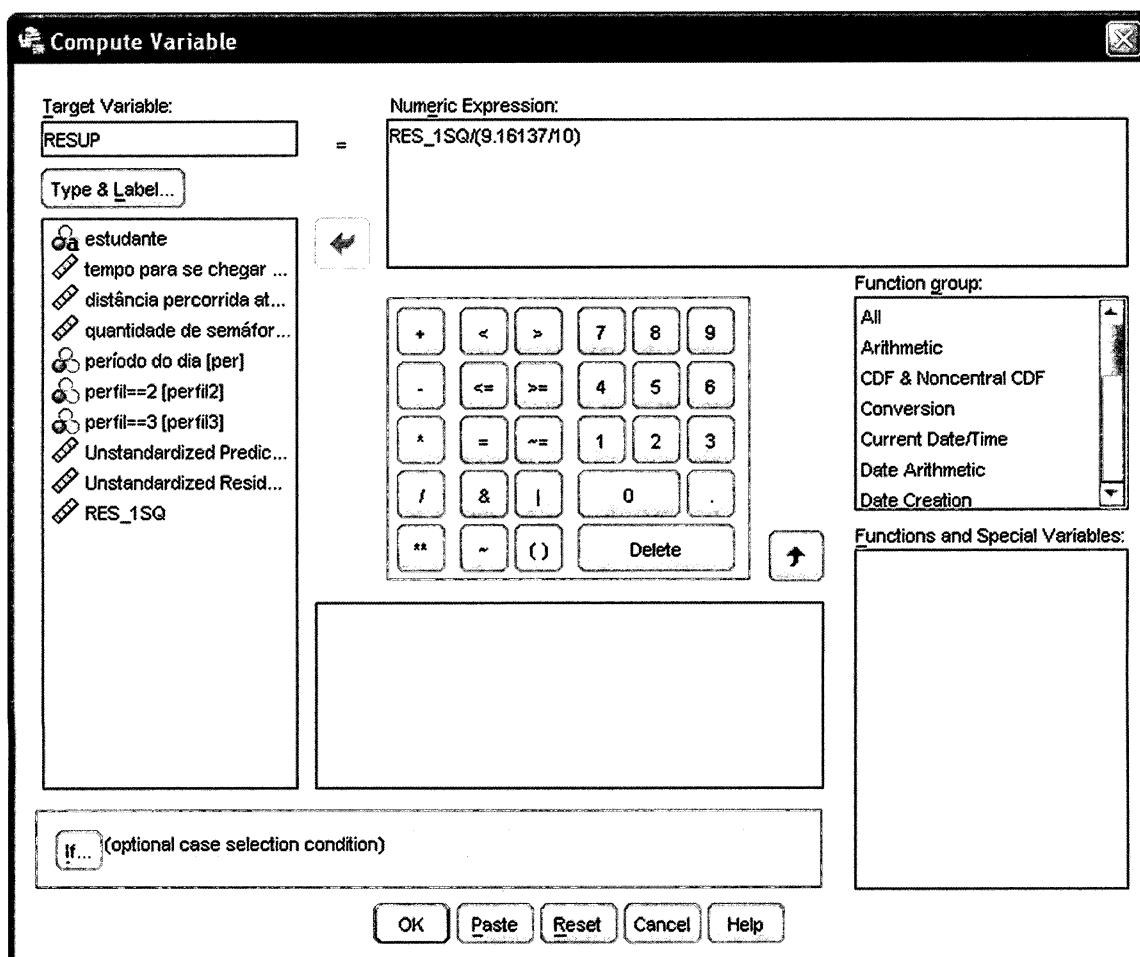


Figura 12.95 Criação da variável referente aos resíduos ao quadrado (RES\_1SQ).

The screenshot shows the 'Descriptives' dialog box and its 'Options' sub-dialog. In the main dialog, 'Variable(s)' is set to 'RES\_1SQ'. In the 'Options' dialog, 'Sum' is checked under 'Dispersion' and 'Variable list' is selected under 'Display Order'. Both dialogs have OK, Paste, Reset, Cancel, and Help buttons.

Figura 12.96 Cálculo da soma dos resíduos ao quadrado.



**Figura 12.97** Criação da variável RESUP.

ANOVA <sup>b</sup>					
Model		Sum of Squares	df	Mean Square	F
1	Regression	3,185	1	3,185	3,749
	Residual	6,797	8	,850	
	Total	9,982	9		

a. Predictors: (Constant), Unstandardized Predicted Value

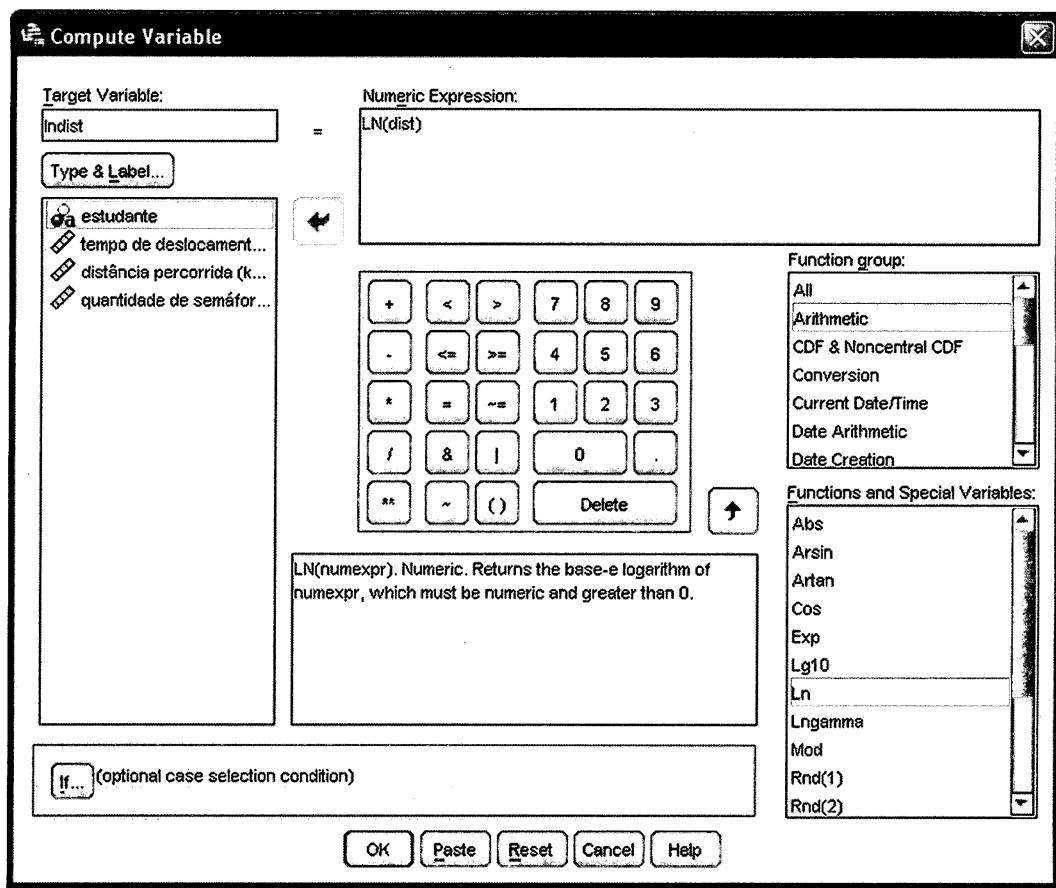
b. Dependent Variable: RESUP

**Figura 12.98** Tabela ANOVA da regressão de RESUP em função de PRE\_1.

A partir de então, podemos estimar o modelo não linear proposto. Os *outputs* não serão aqui apresentados, porém são os mesmos da Figura 12.71.

Diferentemente do Stata, o SPSS não oferece uma opção direta para elaboração de transformações de Box-Cox, de modo que não estimaremos o modelo cujos resultados são apresentados na Figura 12.75. Caso um pesquisador deseje elaborar aquela estimação, deverá criar manualmente, em **Transform → Compute Variable...**, uma nova variável dependente transformada. Entretanto, como não se conhece, *a priori*, o parâmetro da transformação de Box-Cox que maximiza a aproximação da distribuição da nova variável à distribuição normal, recomendamos fortemente que ao menos a obtenção do parâmetro  $\lambda$  seja feita por meio do Stata, com o procedimento elaborado para se chegar aos resultados da Figura 12.73.

Por fim, mas não menos importante, vamos apresentar o procedimento para verificação de existência de autocorrelação dos resíduos no SPSS. Como este software não dispõe de procedimento direto para elaboração do teste de Breusch-Godfrey, iremos nos ater à aplicação do teste de Durbin-Watson. Para tanto, devemos abrir o arquivo **Análisetemporaltempodistsem.sav**.

Figura 12.99 Criação da variável *Indist*.

Quando da elaboração da regressão propriamente dita, em **Analyze → Regression → Linear...**, o botão **Statistics...** oferece a opção para a realização do teste de Durbin-Watson. Devemos marcar esta opção, conforme mostra a Figura 12.100. Note que não há qualquer menção ao fato de que o banco de dados apresenta uma variável correspondente à evolução temporal, o que quer dizer que uma modelagem numa base em *cross-section* também permitiria a elaboração do referido teste, o que, conforme já discutimos, é um erro grave.

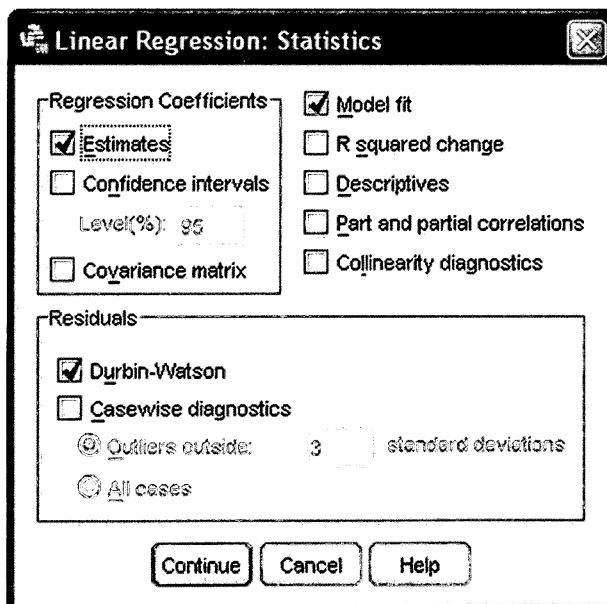


Figura 12.100 Caixa de diálogo para a elaboração do teste de Durbin-Watson.

O resultado do teste está na Figura 12.101, e é exatamente igual ao que já foi apresentado por meio da Figura 12.79.

<b>Model Summary<sup>b</sup></b>					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,847 <sup>a</sup>	,717	,696	7,30021	1,779

a. Predictors: (Constant), quantidade média de semáforos, distância média percorrida (km)

b. Dependent Variable: tempo médio de deslocamento até a escola (minutos)

**Figura 12.101** Resultado do teste de Durbin-Watson.

Conforme já discutido, a estatística  $DW = 1,779$  indica a inexistência de autocorrelação de primeira ordem dos termos de erro, ao nível de significância de 5% e para um modelo com 3 parâmetros e 30 observações.

## 12.7. CONSIDERAÇÕES FINAIS

Os modelos de regressão simples e múltipla estimados pelo método de mínimos quadrados ordinários (MQO, ou OLS) representam o grupo de técnicas de regressão mais utilizadas em ambientes acadêmicos e organizacionais, dada a facilidade de aplicação e de interpretação dos resultados obtidos, além do fato de estarem disponíveis na grande maioria dos softwares, mesmo naqueles em que não haja especificamente um foco voltado à análise estatística de dados. É importante também ressaltar a praticidade das técnicas estudadas neste capítulo para fins de elaboração de diagnósticos e previsões.

É de fundamental importância que o pesquisador sempre avalie e discuta o atendimento aos pressupostos da técnica e, mais do que isso, sempre reflita sobre a possibilidade de que sejam estimados modelos não necessariamente com formas funcionais lineares.

Explicitamos, por fim, que o pesquisador não precisa restringir a análise do comportamento de determinado fenômeno apenas e tão somente com base na teoria subjacente. A aplicação de modelagens de regressão pede, por vezes, que sejam incluídas variáveis com base na experiência e intuição do pesquisador, a fim de que possam ser gerados modelos cada vez mais interessantes e diferentes do que tradicionalmente vem sendo proposto. Assim, novas óticas e perspectivas para o estudo do comportamento de fenômenos sempre poderão surgir, o que contribui para o desenvolvimento científico e para o surgimento de trabalhos empíricos cada vez mais inovadores.

## 12.8. EXERCÍCIOS

1. A tabela a seguir traz os dados de crescimento do PIB e investimento em educação de determinada nação, ao longo de 15 anos:

Ano	Taxa de Crescimento do PIB (%)	Investimento em Educação (bilhões de US\$)
01	-1,50	7,00
02	-0,90	9,00
03	1,30	15,00
04	0,80	12,00
05	0,30	10,00
06	2,00	15,00
07	4,00	20,00
08	3,70	17,00
09	0,20	8,00
10	-2,00	5,00
11	1,00	13,00
12	1,10	13,00
13	4,00	19,00
14	2,70	19,00
15	2,50	17,00

Pergunta-se:

- Qual a equação que avalia o comportamento da taxa de crescimento do PIB ( $Y$ ) em função do investimento em educação ( $X$ )?
- Qual percentual da variância da taxa de crescimento do PIB é explicado pelo investimento em educação ( $R^2$ )?
- A variável referente o investimento em educação é estatisticamente significante, a 5% de nível de significância, para explicar o comportamento da taxa de crescimento do PIB?
- Qual o investimento em educação que, em média, resulta numa taxa esperada de crescimento do PIB igual a zero?
- Qual seria a taxa esperada de crescimento do PIB se o governo desta nação optasse por não investir em educação num determinado ano?
- Se o investimento em educação num determinado ano for de US\$11 bilhões, qual será a taxa esperada de crescimento do PIB? E quais serão os valores mínimo e máximo de previsão para a taxa de crescimento do PIB, ao nível de confiança de 95%?

2. Os arquivos **Corrupção.sav** e **Corrupção.dta** trazem dados sobre 52 países em determinado ano, a saber:

Variável	Descrição
<i>país</i>	Variável string que identifica o país $i$ .
<i>cpi</i>	<i>Corruption Perception Index</i> , que corresponde à percepção dos cidadãos em relação ao abuso do setor público sobre os benefícios privados de uma nação, cobrindo aspectos administrativos e políticos. Quanto menor o índice, maior a percepção de corrupção no país (Fonte: Transparência Internacional).
<i>idade</i>	Idade média dos bilionários do país (Fonte: Forbes).
<i>horas</i>	Quantidade média de horas trabalhadas por semana no país, ou seja, o total anual de horas trabalhadas dividido por 52 semanas (Fonte: Organização Internacional do Trabalho).

Deseja-se investigar se a percepção de corrupção de um país é função da idade média de seus bilionários e da quantidade média de horas trabalhadas semanalmente e, para tanto, será estimado o seguinte modelo:

$$cpi_i = a + b_1 \cdot idade_i + b_2 \cdot horas_i + u_i$$

Pede-se:

- Analise o nível de significância do teste  $F$ . Pelo menos uma das variáveis (*idade* e *horas*) é estatisticamente significante para explicar o comportamento da variável *cpi*, ao nível de significância de 5%?
- Se a resposta do item anterior for sim, analise o nível de significância de cada variável explicativa (testes  $t$ ). Ambas são estatisticamente significantes para explicar o comportamento de *cpi*, ao nível de significância de 5%?
- Qual a equação final estimada para o modelo de regressão linear múltipla?
- Qual o  $R^2$ ?
- Discuta os resultados em termos de sinal dos coeficientes das variáveis explicativas.
- Salve os resíduos do modelo final e verifique a existência de normalidade nestes termos de erro.
- Por meio do teste de Breusch-Pagan/Cook-Weisberg, verifique se há indícios de existência de heterocedasticidade no modelo final proposto.
- Apresente as estatísticas *VIF* e *Tolerance* e discuta os resultados.

3. Os arquivos **Corrupçãoemer.sav** e **Corrupçãoemer.dta** trazem os mesmos dados do exercício anterior, porém agora com a inclusão de mais uma variável, a saber:

Variável	Descrição
<i>emercente</i>	Variável dummy correspondente ao fato de o país ser considerado desenvolvido ou emergente, segundo o critério da Compustat Global. Neste caso, se o país for desenvolvido, a variável <i>emercente</i> = 0; caso contrário, a variável <i>emercente</i> = 1.

Deseja-se inicialmente investigar se, de fato, os países considerados emergentes apresentam menores índices *cpi*. Sendo assim, pede-se:

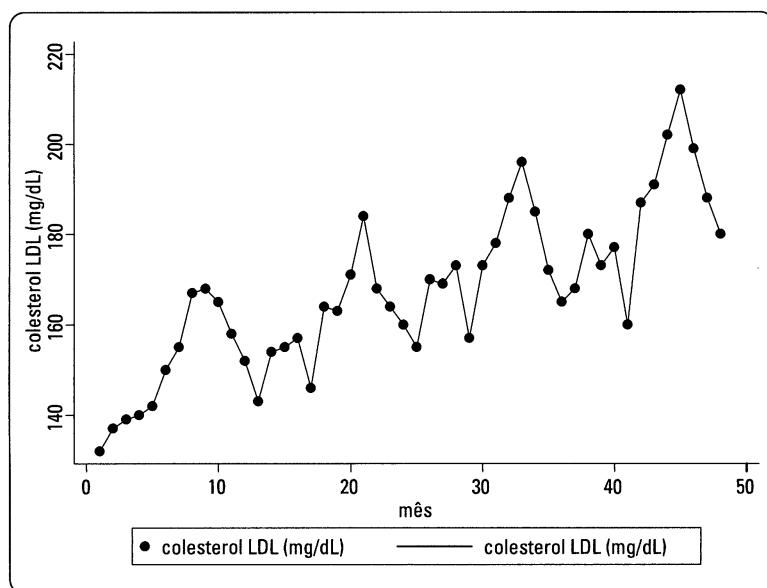
- a. Qual a diferença entre o valor médio do índice *cpi* dos países emergentes e o dos países desenvolvidos? Esta diferença é estatisticamente significante, ao nível de significância de 5%?
- b. Elabore, por meio do procedimento *Stepwise* com nível de significância de 10% para rejeição da hipótese nula dos testes *t*, a estimativa do modelo com a forma funcional linear a seguir. Escreva a equação do modelo final estimado.

$$cpi_i = a + b_1 \cdot idade_i + b_2 \cdot horas_i + b_3 \cdot emergente_i + u_i$$

- c. A partir desta estimativa, pergunta-se: qual seria a previsão, em média, do índice *cpi* para um país considerado emergente, com idade média de seus bilionários de 51 anos e com uma quantidade média de 37 horas trabalhadas semanalmente?
- d. Quais os valores mínimo e máximo do intervalo de confiança para a previsão do item anterior, ao nível de confiança de 90%?
- e. Imagine que um pesquisador proponha, para o problema em questão, que seja estimado o seguinte modelo com forma funcional não linear. Escreva a equação do modelo final estimado por meio do procedimento *Stepwise* e com nível de significância também de 10% para rejeição da hipótese nula dos testes *t*.

$$cpi_i = a + b_1 \cdot idade_i + b_2 \cdot \ln(horas_i) + b_3 \cdot emergente_i + u_i$$

- f. Dado que não foram identificados problemas referentes aos pressupostos dos modelos de regressão em ambos os casos, qual seria a forma funcional escolhida para efeitos de previsão?
4. Um cardiologista tem monitorado, ao longo dos últimos 48 meses, o índice de colesterol LDL (mg/dL), o índice de massa corpórea (kg/m<sup>2</sup>) e a frequência semanal de realização de atividades físicas de um dos principais executivos brasileiros. Seu intuito é orientá-lo sobre a importância da manutenção ou perda de peso e da realização periódica de atividades físicas. A evolução do índice de colesterol LDL (mg/dL) deste executivo, ao longo do período analisado, encontra-se no gráfico a seguir:



Os dados encontram-se nos arquivos **Colesterol.sav** e **Colesterol.dta**, compostos pelas seguintes variáveis:

Variável	Descrição
<i>mês</i>	Mês <i>t</i> da análise.
<i>colesterol</i>	Índice de colesterol LDL (mg/dL).
<i>imc</i>	Índice de massa corpórea (kg/m <sup>2</sup> ).
<i>esporte</i>	Número de vezes em que pratica atividades físicas na semana (média no mês).

Deseja-se investigar se o comportamento, ao longo tempo, do índice de colesterol LDL é influenciado pelo índice de massa corpórea do executivo e pela quantidade de vezes em que ele pratica atividades físicas semanalmente. Para tanto, será estimado o seguinte modelo:

$$\text{colesterol}_t = a + b_1 \cdot \text{imc}_t + b_2 \cdot \text{esporte}_t + \varepsilon_t$$

Desta forma, pede-se:

- a. Qual a equação final estimada para o modelo de regressão linear múltipla?
- b. Discuta os resultados em termos de sinal dos coeficientes das variáveis explicativas.
- c. Embora o modelo final estimado não apresente problemas em relação à normalidade dos resíduos, à heterocedasticidade e à multicolinearidade, o mesmo não pode ser dito em relação à autocorrelação dos resíduos. Elabore o teste de Durbin-Watson, apresente e discuta o resultado.
- d. Elabore o teste de Breusch-Godfrey (não disponível no SPSS) com defasagens de ordem 1, 3, 4 e 12 e discuta os resultados.

## APÊNDICE

# Modelos de regressão quantílica

### A) Breve Introdução

Os modelos de **regressão quantílica**, em geral, e os modelos de **regressão à mediana**, em particular, têm por objetivo principal estimar os percentis da variável dependente, condicionais aos valores das variáveis explicativas. Enquanto a regressão à mediana expressa a mediana (percentil 50%) da distribuição condicional da variável dependente como uma função linear das variáveis explicativas, as demais regressões quantílicas estimam os parâmetros de um modelo com base em qualquer outro percentil desta distribuição condicional (25% ou 75%, por exemplo). Se, para exemplificar, o pesquisador especificar um modelo de regressão quantílica a 25%, os parâmetros estimados descreverão o comportamento do 25º percentil da distribuição condicional da variável dependente.

Esses modelos permitem que seja **caracterizada toda a distribuição condicional da variável dependente**, com base em determinadas variáveis explicativas, já que são obtidas **diferentes estimativas de parâmetros para percentis distintos**, que podem ser interpretados como diferenças no comportamento da variável dependente frente a alterações nas variáveis explicativas nos mais diversos pontos de distribuição condicional da primeira. Esse fato representa uma importante vantagem desses modelos sobre os modelos de **regressão à média** estimados pelo método de mínimos quadrados ordinários (MQO) estudado ao longo do capítulo.

A estimativa dos modelos de regressão quantílica é similar à estimativa por mínimos quadrados ordinários, porém, enquanto esta última minimiza a soma dos quadrados dos resíduos, a primeira minimiza a **soma ponderada dos resíduos absolutos**.

Como a mediana, que é medida de tendência central, não é afetada pela presença de **outliers**, ao contrário da média, muitos pesquisadores fazem uso de modelos de regressão à mediana quando da presença de observações extremas ou discrepantes, visto que são estimados parâmetros não sensíveis à existência de perturbações nos dados. Entretanto, vale a pena comentar, conforme discutem Rousseeuw e Leroy (1987), que mesmo os estimadores de modelos de regressão quantílica podem ser sensíveis à existência de outliers se a **distância leverage** dessas observações forem consideravelmente elevadas.

Esta técnica foi inicialmente proposta por Koenker e Bassett (1978) com o objetivo de estimar os parâmetros do seguinte modelo de regressão:

$$Y_i = a + b_{\theta 1} \cdot X_{1i} + b_{\theta 2} \cdot X_{2i} + \dots + b_{\theta k} \cdot X_{ki} + u_{\theta i} = X_i' \cdot b_{\theta} + u_{\theta i} \quad (12.67)$$

sendo:

$$\text{Perc}_{\theta}(Y_i | X_i) = X_i' \cdot b_{\theta} \quad (12.68)$$

em que  $\text{Perc}_{\theta}(Y_i | X_i)$  representa o percentil  $\theta$  ( $0 < \theta < 1$ ) da variável dependente  $Y$ , condicional ao vetor de variáveis explicativas  $X'$ . A estimativa dos parâmetros da expressão (12.67) pode ser obtida pela solução de um problema de programação linear, cuja função-objetivo é dada pela seguinte expressão:

$$\left[ \sum_{i: Y_i \geq X_i' \cdot b} \theta \cdot |Y_i - X_i' \cdot b| + \sum_{i: Y_i < X_i' \cdot b} (1-\theta) \cdot |Y_i - X_i' \cdot b| \right] = \min \quad (12.69)$$

A estimativa de modelos de regressão quantílica não tem como pressuposto a existência de **normalidade dos resíduos**, o que faz com que possam ser utilizados alternativamente aos modelos estimados pelo método de mínimos quadrados ordinários para os casos em que nem mesmo a **transformação de Box-Cox** na variável dependente

garante a determinação de resíduos com distribuição aderente à normalidade. Situações como essa podem ocorrer, entre outras razões, quando a variável dependente apresentar considerável assimetria em sua distribuição.

Desta forma, esses modelos fazem parte do grupo de estimações que podem ser utilizadas em estudos que apresentam **variáveis dependentes com distribuições assimétricas**, e deseja-se investigar os **diferentes comportamentos das variáveis explicativas para distintos percentis da distribuição**.

De maneira resumida, e seguindo Buchinsky (1998), os modelos de regressão quantílica apresentam as seguintes características e vantagens:

- permitem que os efeitos de cada variável explicativa sobre o comportamento da variável dependente variem entre os percentis;
- a função-objetivo (função de verossimilhança) da regressão quantílica representa a minimização da soma ponderada dos resíduos absolutos, o que faz com que os parâmetros estimados não sejam sensíveis a observações extremas ou discrepantes;
- oferecem estimativas mais eficientes dos parâmetros do que aquelas obtidas pelo método de mínimos quadrados ordinários quando os termos de erro não apresentarem distribuição normal;
- podem ser utilizados quando a variável dependente apresentar distribuição assimétrica.

Como, por exemplo, a distribuição de renda é **intrinsecamente assimétrica** para diferentes populações e ocorrem **variações ao longo dos percentis**, os modelos de regressão quantílica podem ser bastante úteis para o estudo do comportamento de rendimentos, condicional a determinadas variáveis explicativas. Para esses casos, os modelos tradicionais de regressão à média podem ser insatisfatórios, pelo fato de levarem, eventualmente, o pesquisador a conclusões incompletas.

Na sequência, apresentaremos um exemplo em que é estimado um modelo de regressão quantílica, tendo como variável dependente a renda média familiar de determinados indivíduos.

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

Faremos uso do banco de dados **Renda Quantílica.dta**, dada a existência de *outliers* multivariados na amostra, que podem ser identificados por meio da aplicação do algoritmo **bacon** estudado no apêndice do Capítulo 9. Esta base apresenta dados referentes à renda média familiar (R\$) e ao tempo de formado (anos) de 400 profissionais que concluíram o curso de economia em determinada faculdade. Partiremos, portanto, para a estimação dos parâmetros do seguinte modelo:

$$\hat{renda}_i = \alpha + \beta_1 \cdot tformado_i$$

Inicialmente, vamos analisar o histograma da variável dependente *renda*, digitando o seguinte comando:

```
hist renda, freq
```

O gráfico gerado encontra-se na Figura 12.102.

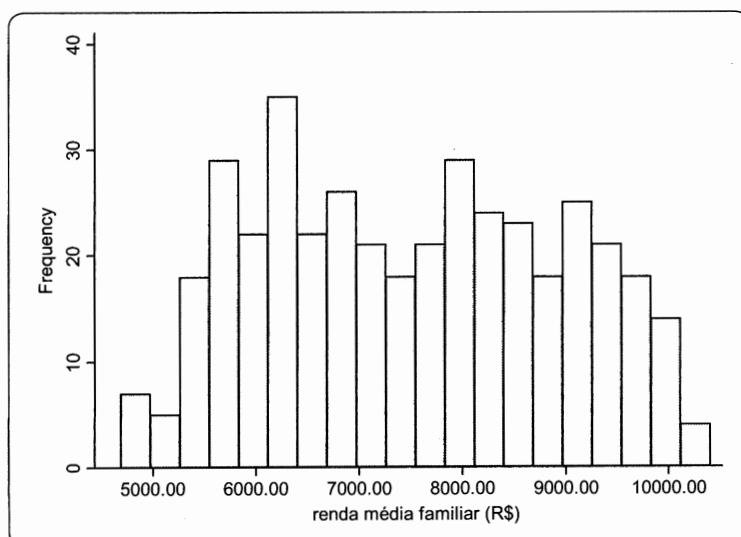
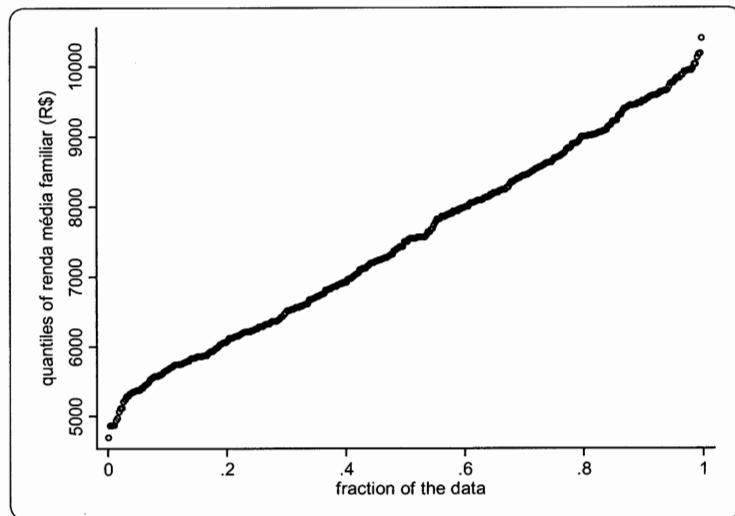


Figura 12.102 Histograma da variável dependente.

A partir desse histograma, podemos perceber a existência de certa assimetria, que representa um primeiro indício favorável à estimação de um modelo de regressão quantílica.

Na sequência, podemos digitar o seguinte comando, que irá gerar o gráfico da Figura 12.103.

```
qplot renda
```



**Figura 12.103** Gráfico de percentis da variável dependente.

Este gráfico mostra os valores de cada percentil da variável dependente *renda*. Por meio do comando **sum renda, detail**, cujos *outputs* não são apresentados aqui, podemos verificar que os valores dos quartis da variável *renda* são iguais a R\$ 6.250,00 (percentil 25%), R\$ 7.500,00 (mediana) e R\$ 8.670,00 (percentil 75%).

Embora também não apresentado aqui, é importante mencionar que os termos de erro gerados a partir da estimação de um modelo de regressão por mínimos quadrados ordinários não apresentam aderência à normalidade, e tal fato tampouco acontece na estimação deste mesmo modelo fazendo-se uso da transformação de Box-Cox na variável dependente, o que novamente favorece a estimação de um modelo de regressão quantílica para os dados do nosso exemplo. Um pesquisador mais curioso poderá comprovar esses fatos, com base nos conceitos estudados ao longo do capítulo.

Incialmente, vamos estimar os parâmetros de um modelo de regressão quantílica com percentil 50% (regressão à mediana), digitando o seguinte comando:

```
qreg renda tformado, quantile(0.50)
```

em que o comando **qreg** estima um modelo de regressão quantílica, sendo o termo **quantile(0.50)** referente a um modelo de regressão à mediana, que poderia ter sido omitido neste caso por ser o próprio padrão do comando **qreg** no Stata. Os *outputs* gerados encontram-se na Figura 12.104.

. qreg renda tformado, quantile(0.50)
Iteration 1: WLS sum of weighted deviations = 466946.48
Iteration 1: sum of abs. weighted deviations = 467240
Iteration 2: sum of abs. weighted deviations = 464146
Iteration 3: sum of abs. weighted deviations = 464040
Median regression
Raw sum of deviations 491360 (about 7500)
Min sum of deviations 464040
Number of obs = 400
Pseudo R2 = 0.0556
-----
renda   Coef. Std. Err. t P> t  [95% Conf. Interval]
tformado   273.3333 48.54141 5.63 0.000 177.9037 368.7629
_cons   5243.333 395.699 13.25 0.000 4465.412 6021.255

**Figura 12.104** Outputs da regressão à mediana no Stata.

É importante mencionar que um pesquisador ainda mais curioso poderá obter esses mesmos *outputs* por meio do arquivo **Renda Quantílica Mínimos Resíduos Absolutos.xls**, fazendo uso da ferramenta **Solver** do Excel, conforme padrão também adotado ao longo do capítulo. Embora não exposto aqui, neste arquivo o pesquisador também terá a opção de determinar o percentil desejado para a estimativa dos parâmetros de qualquer modelo de regressão quantílica.

Podemos verificar (Figura 12.104) que todos os parâmetros estimados são estatisticamente diferentes de zero, a 95% de confiança, e o modelo obtido pode ser escrito da seguinte forma:

$$\hat{renda}_{(mediana)i} = 5.243,333 + 273,333 \cdot tformado_i$$

Neste sentido, a mediana esperada da renda média familiar de determinado economista com 7 anos de formado pode ser obtida da seguinte forma:

$$\hat{renda}_{(mediana)i} = 5.243,333 + 273,333 \cdot (7) = R\$ 7.156,667$$

Desta forma, os parâmetros de um modelo de regressão quantílica podem ser interpretados por meio da derivada parcial do percentil condicional em função de determinada variável explicativa.

Os *outputs* também mostram que a soma absoluta das diferenças entre os valores reais da renda média familiar e o valor de sua mediana não condicional (R\$ 7.500,00) é igual a 491.360. Em outras palavras, temos que:

$$\sum_{i=1}^{400} |renda_i - 7.500,00| = 491.360$$

Já a soma ponderada dos resíduos absolutos para a expressão geral obtida (distribuição condicional da variável *renda* como função linear da variável *tformado*) é igual a 464.040, conforme também podemos verificar pelo mesmo arquivo em Excel.

Sendo assim, o pseudo  $R^2$  apresentado nos *outputs* pode ser calculado da seguinte forma:

$$pseudo R^2 = 1 - \frac{464.040}{491.360} = 0,0556$$

cuja utilidade é bastante limitada e restringe-se a casos em que o pesquisador tiver interesse em comparar dois ou mais modelos distintos.

Se o pesquisador também desejar estimar os parâmetros dos modelos de regressão quantílica, por exemplo, com percentis 25% e 75%, a fim de compará-los com os obtidos pela modelagem de regressão à mediana e também com aqueles obtidos por uma estimativa por mínimos quadrados ordinários, poderá digitar a seguinte sequência de comandos:

```
* REGRESSÃO POR MÍNIMOS QUADRADOS ORDINÁRIOS
quietly reg renda tformado
estimates store MQO

* REGRESSÃO QUANTÍLICA (PERCENTIL 25%)
quietly qreg renda tformado, quantile(0.25)
estimates store QREG25

* REGRESSÃO À MEDIANA (PERCENTIL 50%)
quietly qreg renda tformado, quantile(0.50)
estimates store QREG50

* REGRESSÃO QUANTÍLICA (PERCENTIL 75%)
quietly qreg renda tformado, quantile(0.75)
estimates store QREG75

estimates table MQO QREG25 QREG50 QREG75, se
```

A Figura 12.105 apresenta os parâmetros estimados em cada modelo.

Variable	MQO	QREG25	QREG50	QREG75
tformado	197.58258   35.529997	250 27.482074	273.33333 48.541413	80 70.509666
_cons	5932.1141   289.87448	4360 223.97567	5243.3334 395.69901	7960 576.43629

legend: b/se

**Figura 12.105** Parâmetros estimados em cada modelo e respectivos erros-padrão.

A partir dos *outputs* consolidados na Figura 12.105, é possível percebermos que existem discrepâncias entre os parâmetros estimados por mínimos quadrados ordinários e os obtidos pelas regressões quantílicas. Podemos inclusive verificar que os erros-padrão dos parâmetros (valores situados abaixo dos respectivos parâmetros) são menores para a regressão quantílica com percentil 25%, o que reflete maior precisão da estimação em torno desse percentil para a distribuição condicional da variável dependente.

A sequência de comandos a seguir permite inclusive que visualizemos, por meio de gráficos, as diferenças entre os estimadores obtidos pelas regressões quantílicas e os obtidos por mínimos quadrados ordinários:

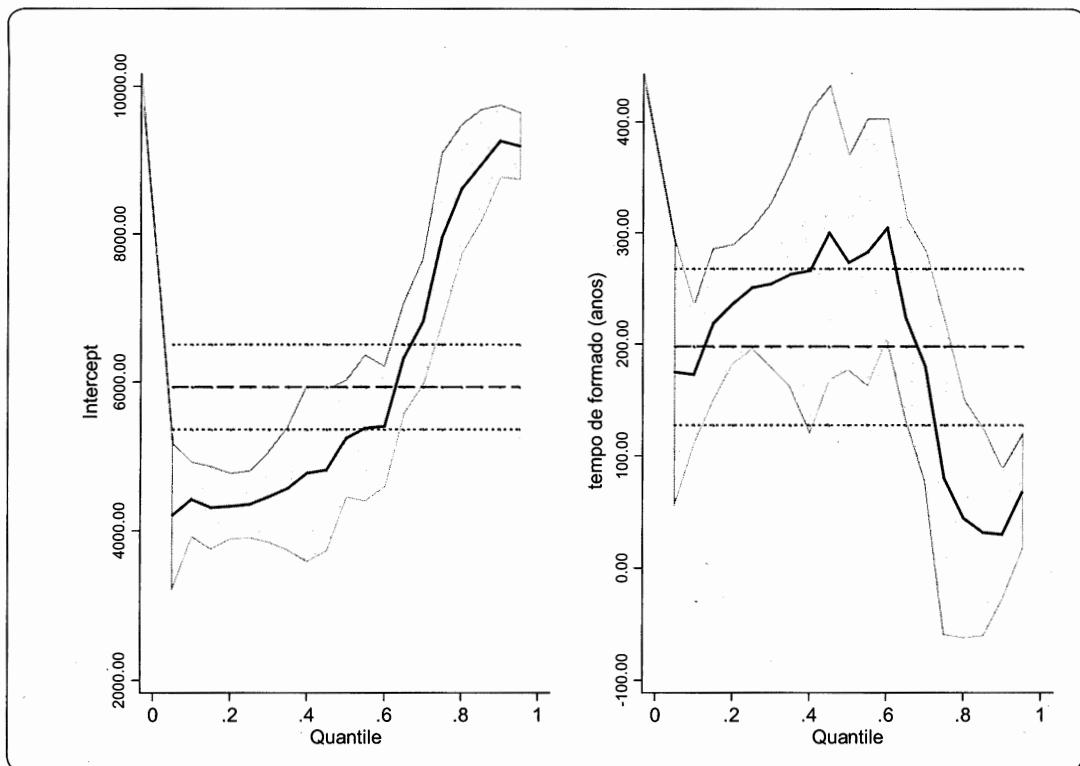
```
quietly qreg renda tformado
grqreg, cons ci ols olsci
```

Os gráficos gerados, que se encontram na Figura 12.106, apresentam os parâmetros  $\alpha$  e  $\beta$  estimados, não restritos apenas aos percentis 25%, 50% e 75%, com respectivos intervalos de confiança a 95% (termo *ci*). Além disso, enquanto o termo *cons* permite que seja elaborado o gráfico do intercepto, os termos *ols* e *olsci* incluem nos gráficos os parâmetros estimados por mínimos quadrados ordinários e os respectivos intervalos de confiança, também a 95%.

Por meio desses gráficos, comprovamos que os parâmetros estimados por mínimos quadrados ordinários e os respectivos intervalos de confiança não variam com os percentis, ao contrário daqueles estimados pelos modelos de regressão quantílica, e, conforme discutimos, esse fato representa uma das principais vantagens desses modelos sobre os modelos de regressão à média, visto que permite que seja caracterizada toda a distribuição condicional da variável dependente em função de determinada variável explicativa, fornecendo uma visão mais ampla da relação entre elas e não restringindo a análise à média condicional.

Para os dados do nosso exemplo, podemos inclusive verificar que o parâmetro  $\beta$  correspondente à variável *tformado* deixa de ser estatisticamente diferente de zero, ao nível de confiança de 95%, para percentis mais elevados, visto que seu intervalo de confiança passa a conter o zero. Para a verificação desse fato, basta que o pesquisador digite, por exemplo, o comando *qreg renda tformado, quantile(0.80)* e analise a estatística *t* do referido parâmetro.

É importante mencionar que, em outros casos, podem inclusive ocorrer alterações de sinal de determinado parâmetro  $\beta$  à medida que variam os percentis, o que propicia ao pesquisador uma análise mais completa acerca das diferenças no comportamento da variável dependente frente a alterações em cada variável explicativa nos mais diversos pontos da distribuição condicional da primeira.



**Figura 12.106** Parâmetros estimados para regressões quantílicas e por mínimos quadrados ordinários, com respectivos intervalos de confiança.

Para efeitos didáticos, vamos elaborar um gráfico que apresenta os ajustes lineares entre os valores previstos da variável dependente, gerados pelos modelos de regressão por mínimos quadrados ordinários e quantílicos com percentis 25%, 50% e 75%, e a variável explicativa. O intuito é comparar esses ajustes lineares. Para tanto, podemos digitar a seguinte sequência de comandos:

```
* REGRESSÃO POR MÍNIMOS QUADRADOS ORDINÁRIOS
quietly reg renda tformado
predict ymqo

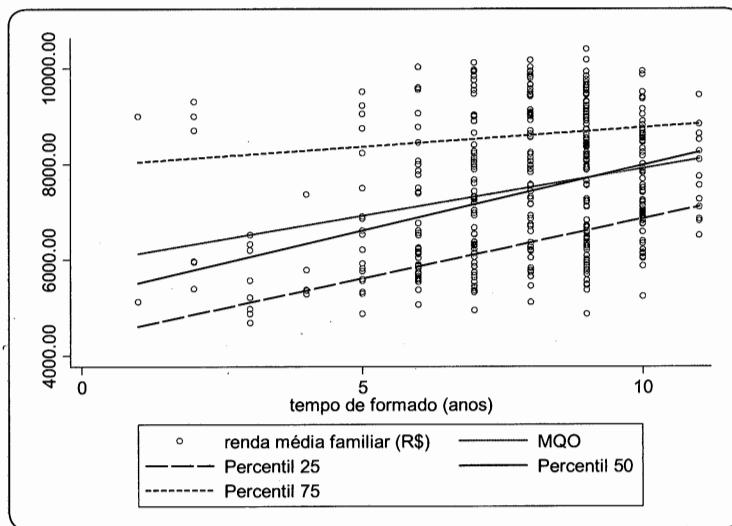
* REGRESSÃO QUANTÍLICA (PERCENTIL 25%)
quietly qreg renda tformado, quantile(0.25)
predict yqreg25

* REGRESSÃO À MEDIANA (PERCENTIL 50%)
quietly qreg renda tformado, quantile(0.50)
predict yqreg50

* REGRESSÃO QUANTÍLICA (PERCENTIL 75%)
quietly qreg renda tformado, quantile(0.75)
predict yqreg75

graph twoway scatter renda tformado || lfit ymqo tformado || lfit yqreg25
tformado || lfit yqreg50 tformado || lfit yqreg75 tformado ||, legend(label(2
"MQO") label(3 "Percentil 25") label(4 "Percentil 50") label(5 "Percentil 75"))
```

O gráfico gerado está na Figura 12.107.



**Figura 12.107** Comportamento da variável dependente em função da variável explicativa *tformado*, com destaque para as estimações MQO e quantílicas.

Esse gráfico apresenta a renda média familiar ajustada por sua média e para os percentis 25%, 50% e 75%, em função do tempo de formado do indivíduo. Embora seja possível evidenciar, por meio deste exemplo, o crescimento da renda média familiar em todos os percentis à medida que o tempo de formado aumenta, podemos verificar a existência de diferenças entre o ajuste à média (MQO) e o ajuste à mediana (percentil 50%), fato que ocorre em razão da existência de *outliers* e da influência que esses exercem sobre a estimativa dos parâmetros por mínimos quadrados ordinários. Nesse sentido, o pesquisador precisa estar sempre atento à sensibilidade dos parâmetros e existência de observações extremas ou discrepantes na base de dados, que podem fazer com que determinado método de estimação seja preferível.

Em resumo, e conforme discutimos inicialmente, os modelos de regressão quantílica são mais adequados para o estudo da relação entre as variáveis apresentadas neste exemplo, visto que tornam possível a análise, para os diversos percentis, dos efeitos da variável *tformado* sobre o comportamento da variável *renda*, propiciam a estimativa de parâmetros não sensíveis à existência de *outliers* e distribuição assimétrica da variável dependente, e possibilitam a determinação de um modelo sem que haja a necessidade de que os resíduos apresentem distribuição normal.