- *Variação Total* é igual ao quadrado dos desvios das observações y com relação ao valor da média $\overline{y}$ das mesmas observações y, isto é, $\sum_{i=1}^{n}(y_i - \overline{y})^2$.
- *Variação não-explicada* é igual ao quadrado dos desvios das observações y com relação aos valores estimados pelo modelo de regressão $\hat{y}$, isto é, $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.
- *Variação explicada* é igual ao quadrado dos desvios dos valores estimados pelo modelo de regressão $\hat{y}$ com relação ao valor da média das observações y, isto é, $\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$.

Da Figura 16.3 pode-se ver que:

Variação total = Variação não-explicada + Variação explicada

Se demonstra que $\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y})^2 + \sum_{i=1}^{n}(\hat{y} - \overline{y})^2$

Definimos como *coeficiente de determinação* r^2 à relação:

$$r^2 = \frac{\textit{Variação explicada}}{\textit{Variação total}}$$

Substituindo as expressões matemáticas na expressão anterior temos:

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

Na planilha **Coeficiente de Determinação** incluída na pasta CAP_16 foi obtido o *coeficiente de determinação* do Exemplo 16.1, Figura 16.4. Os resultados parciais são:

- *Variação Total*, conhecida também como $SST = \sum_{i=1}^{n}(y_i - \overline{y})^2 = 129.951$
- *Variação não-explicada*, conhecida também como $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 33.980,61$
- *Variação explicada*, conhecida também como SSR[10] $= \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = 95.969,39$

[10] Em inglês, *SSR* é *Sum of Squares for Regression*

Substituindo na expressão do *coeficiente de determinação* temos:

$$r^2 = \frac{95.969,39}{129.950} = 0,7385$$

	A	B	C	D	E	F	G	H
1	DETERMINAÇÃO DO COEFICIENTE DE DETERMINAÇÃO							
2								
3						Variação		
4		x	y		Projeção	Explicada	Não explicada	Total
5		30	430		409,21	853,48	432,04	2.500
6		21	335		321,57	3.413,93	180,33	2.025
7		35	520		457,91	6.069,21	3.855,77	19.600
8		42	490		526,07	21.337,07	1.301,20	12.100
9		37	470		477,38	9.483,14	54,49	8.100
10		20	210		311,83	4.646,74	10.369,96	28.900
11		8	195		194,98	34.234,14	0,00	34.225
12		17	270		282,62	9.483,14	159,23	12.100
13		35	400		457,91	6.069,21	3.353,01	400
14		25	480		360,52	379,33	14.274,58	10.000
15		Média y	380		Soma	95.969,39	33.980,61	129.950
16		b	9,74					
17		a	117,07			Coeficiente de Determinação		0,7385
18								

Figura 16.4 • Cálculo do coeficiente de determinação do Exemplo 16.1

O *coeficiente de determinação* r^2, cujo valor é sempre positivo, deve ser interpretado como a proporção da variação total na variável dependente y que é explicada pela variação da variável independente x. No caso do Exemplo 16.1, podemos dizer que 73,85% das variações das vendas podem ser explicadas pela variabilidade do investimento em propaganda. Se demonstra, também, que *o coeficiente de determinação é igual ao quadrado do coeficiente de correlação*, e vice versa. Ou seja, a partir do coeficiente de correlação r obtemos o valor do coeficiente de determinação: $r^2 = (r)^2$. No caso do Exemplo 16.1, como o r=0,859366 temos que $r^2 = (0,859366)^2 = 0,7385$. O coeficiente de determinação r^2 é sempre positivo e pode ser analisado de forma equivalente como foi analisado o coeficiente de correlação r. O coeficiente de correlação é mais indicado para ser usado como medida da força da relação entre as variáveis, e o coeficiente de determinação é mais apropriado para definir quanto a reta de regressão explica o ajuste da reta.

USANDO O EXCEL

As funções estatísticas do Excel EPADYX e RQUAD calculam, respectivamente, o *erro padrão da estimativa* e o *coeficiente de determinação*.

A partir da linha 17 da planilha **Funções** incluída na pasta CAP_16, as funções EPADYX e RQUAD são aplicadas ao Exemplo 16.1, cujas sintaxes são as seguintes:

EPADYX(*série_y; série_x*)
A função estatística EPADYX[11] dá como resultado o valor do erro padrão da estimativa S_e da reta de regressão linear $\hat{y} = a + bx$, quando são conhecidas as duas séries de dados ou observações, *série_y* e *série_x*. Ao usar a função EPADYX deve-se tomar cuidado de fornecer os dados na ordem correta, o primeiro argumento *série_y* se refere à série de valores da variável dependente y, e o argumento *série_x* se refere à série de valores da variável independente x. Os argumentos *série_y* e *série_x* devem ser números ou nomes, matrizes ou referências que contenham números. No caso de informar como matriz, devemos registrar a fórmula:
=EPADYX({430;335;520;490;470;210;195;270;400;480};
{30;21;35;42;37;20;8;17;35;25}) → 65,17

O resultado do erro padrão da estimativa significa que o valor real de vendas é diferente do valor estimado no valor igual a $65,17 milhões. Embora a reta de regressão possa ajudar a estimar valores de vendas, não podemos esperar uma diferença menor que $65,17 milhões com relação aos dados amostrais.

RQUAD(*série_y; série_x*)
A função estatística RQUAD[12] dá como resultado o valor do *coeficiente de determinação* r^2 da reta de regressão linear $\hat{y} = a + bx$, quando são conhecidas as duas séries de dados ou observações, *série_y* e *série_x*. Pela definição dos argumentos da função RQUAD, deve-se tomar cuidado de fornecer os dados na ordem correta, primeiro o argumento *série_y* e depois o argumento *série_x*. Entretanto, essa restrição é desnecessária, da mesma forma como foi visto ao apresentar as funções CORREL e PEARSON, pois não é necessário manter a ordem das variáveis. Os argumentos *série_y* e *série_x* devem ser números ou nomes, matrizes ou referências que contenham números. No caso de informar como matriz, devemos registrar a fórmula:
=RQUAD({430;335;520;490;470;210;195;270;400;480};
{30;21;35;42;37;20;8;17;35;25}) → 0,7385

A partir da função RQUAD pode ser calculado o valor do coeficiente de correlação, realizando o caminho oposto, isto é, $\sqrt{r^2} = \pm r$. O sinal de r deve ser o mesmo que o sinal do coeficiente b da reta de regressão, ou da função INCLINAÇÃO. No caso do Exemplo 16.1, o coeficiente de determinação é igual ao quadrado do coeficiente de correlação (0,8594) cujo quadrado é igual a 0,7385.
O resultado do coeficiente de determinação significa que 73,85% da variação das vendas são explicadas pelo variável investimento em propaganda, ficando 26,15% sem explicação.

[11] Em inglês, a função estatística EPADYX é *STEYX*.
[12] Em inglês, a função estatística RQUAD é *RSQ*.

AS PREMISSAS DO MODELO DE REGRESSÃO LINEAR

A amostragem aleatória realizada para obter a reta de regressão representa alguns pontos da população, que é bem maior. A regressão realizada é, na realidade, uma estimativa da relação entre as variáveis, relação essa que é desconhecida. Portanto, os coeficientes da regressão, a e b, são estimativas pontuais dos dois parâmetros populacionais correspondentes, denominados como α e β.

$$\hat{y} = a + bx$$
$$\hat{y} = \alpha + \beta x + e$$

O valor e[13] representa a *dispersão* na população, devido ao fato de não existir um relacionamento perfeito entre as duas variáveis na população. De uma outra maneira, existem outras variáveis que não foram consideradas na regressão e que também tem uma certa influência, minoritária, nos resultados, pois a regressão foi realizada com as duas variáveis mais importantes do experimento.

Devido à variabilidade amostral, deve-se aceitar que cada amostra aleatória gerará uma equação de regressão diferente. Portanto, o coeficiente a é um estimador de α e b é um estimador de β. Se fosse amostrada toda a população, então o valor de a seria igual a α e o valor de b igual a β. A dispersão na população significa que para cada valor de x se terão diversos valores de y. Portanto, para cada valor de x existe uma distribuição de freqüências de y, que o modelo de regressão linear supõe que é uma distribuição normal. Esta distribuição é denominada como *distribuição condicional*, pois depende da *condição* x, e todas as distribuições condicionais tem o mesmo desvio padrão, denominado como *desvio padrão condicional*. As premissas do modelo de regressão linear são:

1. Para cada valor de x existe um grupo de valores de y. Todos os grupos de y têm distribuição normal com o mesmo desvio padrão.
2. As médias das distribuições normais de y pertencem à reta de regressão.
3. O valor esperado dos desvios ou erros é nulo; ou seja, $E[d_i]=0$; pois a variância é mínima.
4. A variância dos desvios, $Var(d_i)$, é constante e igual à variância da população; isto é, todos os desvios tem a mesma variância.
5. Os desvios são variáveis aleatórias independentes e têm distribuição normal. Portanto, a covariância entre os desvios, tomados dois a dois, é nula e os desvios e a variável independente x não tem nenhuma correlação.

Tenha presente o leitor que se os dados amostrais disponíveis não são apropriados para ser aplicado o método de regressão linear, então as inferências obtidas da regressão poderão ser incorretas.

[13] Denominado também como *resíduo*.

ANÁLISE DOS COEFICIENTES

Devido à variabilidade amostral, o modelo de regressão obtido é um dos modelos possíveis da população. Vamos supor que o modelo baseado na população seja $\hat{y} = \alpha + \beta x + e$, onde, e representa o erro cometido na projeção. Devemos ter presente que embora possamos usar a população, continuaremos a ter uma diferença entre as observações reais e os valores projetados. Essa diferença é devida às limitações do modelo representar a realidade usando apenas uma variável aleatória. Ao mesmo tempo, os valores dos coeficientes a e b obtidos de uma amostragem aleatória não serão iguais, em geral, aos coeficientes α e β da população. Entretanto, se demonstra que a e b são os melhores estimadores não tendenciosos de α e β, respectivamente. Portanto, os coeficientes a e b terão o menor valor de variância comparados com qualquer outro par de valores.

ERRO PADRÃO DO COEFICIENTE *b*

O *erro padrão do coeficiente b*, denominado como S_b indica aproximadamente quão distante o coeficiente b está do coeficiente da população β devido à variabilidade amostral. A fórmula usada, com n-2 graus de liberdade, é a seguinte:

$$S_b = \sqrt{S_b^2} = \sqrt{\frac{S_e^2}{(n-1)\times Var(x)}} = \frac{S_e}{\sqrt{(n-1)\times Var(x)}}$$

Da fórmula pode-se deduzir que o *erro padrão do coeficiente b* é diretamente proporcional ao erro padrão da estimativa S_e, e inversamente proporcional ao valor do desvio padrão de x e o tamanho da amostra menos 1. O *erro padrão do coeficiente b* do Exemplo 16.1 é igual a 2,05, obtido da seguinte fórmula[14]:

$$S_b = \frac{65{,}17}{\sqrt{(10-1)\times\frac{10}{9}\times 101{,}20}} = 2{,}05$$

ERRO PADRÃO DO COEFICIENTE *a*

O *erro padrão do coeficiente a*, denominado como S_a, indica aproximadamente quão distante o coeficiente a está do coeficiente da população α devido à variabilidade amostral. A fórmula usada, com n-2 graus de liberdade, é a seguinte:

[14] Como a variância conhecida se refere à população, devemos obter seu valor equivalente como amostra multiplicando aquele valor por 10 e depois dividindo por 9.

$$S_a = S_e\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{(n-1)\times S_x^2}}$$

Da fórmula pode-se deduzir que erro padrão do coeficiente a é proporcional, também, ao erro padrão da estimativa S_e, diminuindo seu valor com o valor do desvio padrão de x e o tamanho da amostra menos 1, e aumentando com o valor da média da amostra x. O *erro padrão do coeficiente a* no Exemplo 16.1 é 59,03, obtido de:

$$S_a = 65{,}17\sqrt{\frac{1}{10} + \frac{27^2}{(10-1)\times\frac{10}{9}\times 101{,}20}} = 59{,}03$$

INTERVALO DE CONFIANÇA DO COEFICIENTE *b*

Se a partir de uma equação de regressão obtemos o valor previsto da variável dependente y, para um dado valor da variável independente x, interessa conhecer o valor do intervalo, por exemplo, para 95% de probabilidade de conter o verdadeiro valor de y. Apesar de ser possível realizar uma regressão linear dentro das premissas estabelecidas, existem fontes de erro provenientes de:

1. A reta da população é desconhecida, porém é estimada a partir das observações. Aceitando que b=β, pode acontecer que a média da variável x seja a mesma para a população como para a amostra, entretanto os valores das médias de y para a população e para a amostra podem ser diferentes. Em outras palavras, como as retas são paralelas, nas projeções existirá sempre uma diferença constante igual à diferença dos valores das médias de y.
2. Aceitando que $b \neq \beta$, as retas da população e da amostra têm apenas um único valor coincidente, passam pelos mesmos valores de média de x e y, que são os valores das médias das observações. Nesse caso, o erro da projeção aumenta, ou diminui, com a diferença de declividade; como não existe um erro constante, cada valor projetado tem um erro diferente.
3. Existem erros de outras fontes que se refletem em valores de observações que não representam fielmente a população.

O coeficiente b de uma amostra é um estimador do coeficiente da população β. Interessa conhecer o intervalo de variação do coeficiente b para um determinado coeficiente de confiança, por exemplo 0,95. Usando a distribuição t, o intervalo de b é obtido com a expressão:

$$\beta = b \pm t_{\alpha/2}\times S_b = b \pm t_{\alpha/2}\times\frac{S_e}{\sqrt{(n-1)\times Var(x)}}$$

No Exemplo 16.1, para um coeficiente de confiança de 0,95 e (10-2)=8 graus de liberdade, o valor de *t* é obtido com =INVT(0,05;8) é 2,306. Portanto,

$$\beta = b \pm t_{\alpha/2} \times S_b = 9{,}7381 \pm 2{,}306 \times 2{,}05$$
$$\beta = 9{,}7381 \pm 4{,}7273$$

Então, podemos dizer que temos 95% de probabilidade de que o coeficiente β da população se encontre entre os valores 5,0108 e 14,4654. Como os valores são positivos, pode-se dizer que temos 95% de confiança que o coeficiente β seja positivo. De uma outra maneira, por cada unidade de aumento da variável *x*, a variável dependente *y* aumentará entre os valores do intervalo 5,0108 e 14,4654.

TESTE DE HIPÓTESE DO COEFICIENTE *b*

Será que o modelo de regressão linear obtido é útil para projetar valores de *y*? Supondo que as variáveis *x* e *y* não são relacionadas, o que podemos dizer dos valores dos coeficientes da reta? *A hipótese nula estabelece que as variáveis x e y não são relacionadas*, isto é, β=0[15]. O teste de hipóteses que deve ser realizado é o seguinte:

$H_0 : \beta=0$
$H_1 : \beta\neq0$

Quando o tamanho da amostra é pequeno devemos usar a distribuição *t*. O valor do *t observado* é $t = \frac{b-\beta}{S_b}$. Como o valor do coeficiente β é zero, o valor de *t observado* será $t = \frac{b-0}{S_b}$, onde S_b é o erro padrão do coeficiente *b*. Aplicando este teste no Exemplo 16.1, o valor do *t observado* é igual a 4,75, obtido de $t = \frac{9{,}74-0}{2{,}05} = 4{,}75$.

O valor do *t observado* indica quantos desvios padrões o coeficiente de regressão está distanciado do coeficiente da população, considerando que a hipótese β=0 seja verdadeira. Quanto maior for o valor de *t*, o resultado sugere que o coeficiente da população não é nulo.

Adotando um nível de significância de 0,05 nas duas caudas e 8 graus de liberdade, o valor do *t crítico* é obtido com a função estatística INVT. No Exemplo 16.1, o valor do *t crítico* é igual a 2,306 obtido com a seguinte fórmula =INVT(0,05;8). Como o valor de *t observado* (4,75) é maior que o valor do *t crítico* (2,306), deve-se rejeitar a hipótese nula e aceitar que β≠0; isto é, podemos dizer que o *coeficiente da população não é nulo*, e portanto, *o coeficiente b é um bom estimador*.

[15] Na ausência de melhores informações, a melhor estimativa de uma variável aleatória é sua própria média.

TESTE DE HIPÓTESES COM A DISTRIBUIÇÃO *F*

A distribuição *t* é usada para realizar testes de hipóteses dos coeficientes da reta de regressão. A distribuição *F* é usada para realizar testes de hipóteses da equação da reta de regressão como um todo. A distribuição *F* testa a hipótese de que nenhum dos coeficientes de regressão tenha significado. O *F observado* é igual a:

$$F = \frac{Variância\ \ Explicada}{Variância\ \ Não\ Explicada}$$

Para obter o valor de cada variância devemos dividir a soma dos quadrados dos desvios pelo número de graus de liberdade correspondente. Para a *variância explicada*, o número de graus de liberdade é igual ao número de amostras *k* menos 1, isto é, *k*–1. Para a *variância não explicada*, o número de graus de liberdade é igual ao número de observações da amostra *n* menos o número de amostras *k*; isto é, *n*–*k*. Portanto, o *F observado* é igual a:

$$F = \frac{\frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{k-1}}{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-k}}$$

Aproveitando os resultados das variações, na célula H19 da planilha **Coeficiente de Determinação** foi calculado o valor *F observado* do Exemplo 16.1 igual a 22,5939. Aplicando a definição do coeficiente de determinação, se obtém a expressão:

$$F = \frac{\frac{r^2}{k-1}}{\frac{1-r^2}{n-k}}$$

O valor do *F observado* do Exemplo 16.1 é 22,59, considerando o *coeficiente de determinação* igual a 0,7385, obtido da seguinte forma:

$$F = \frac{\frac{0{,}7385}{2-1}}{\frac{1-0{,}7385}{10-2}} = 22{,}59$$

Este valor deve ser comparado com o *F crítico* para um nível de significância definido. O procedimento formal realizado é o teste de hipóteses:

$H_0 : \beta=0$
$H_1 : \beta\neq0$

Adotando um nível de significância igual a 0,01, para $k-1=2-1=1$ graus de liberdade do numerador e $n-2=8$ graus de liberdade do denominador, com a função estatística =INVF(0,01;1;8) obtemos valor crítico da distribuição F igual a 11,2586. Como o valor observado é maior que o valor crítico, devemos rejeitar a hipótese nula e aceitar a regressão, isto é, $\beta \neq 0$. Verifique o leitor que ao aplicar o teste F em regressão linear simples:

- Os graus de liberdade do numerador ficam definidos e constantes de valor igual a 1.
- A distribuição F é igual à distribuição t ao quadrado, isto é $F = t^2$.

USANDO O EXCEL

Na planilha **Funções** incluída na pasta CAP_16 obtivemos a maioria dos resultados de uma regressão linear. Agora será apresentada a função estatística PROJ.LIN que fornece todos os resultados numa única célula, de onde, o usuário pode extrair os resultados do seu interesse. A sintaxe da função PROJ.LIN é a seguinte:

PROJ.LIN(*série_y; série_x; constante; estatísticas*)
A função estatística PROJ.LIN[16] dá como resultado uma *matriz* com os resultados da reta de regressão linear múltipla $\hat{y} = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n$, aplicando o método dos quadrados mínimos, onde os argumentos têm os seguintes significados:

- É conhecida a série de dados ou observações, *série_y*.
- São conhecidas uma ou mais séries de dados ou observações, *série_x*, tendo presente que:
  - Quando existe apenas uma variável independente, a forma dos intervalos das duas variáveis y e x podem ter qualquer forma.
  - Quando existe mais de uma variável independente, deve ser informado o intervalo abrangendo todas as variáveis independentes juntas.
  - Quando a informação da variável independente é omitida, a função assume que x é uma *matriz* de números {1, 2, 3, ...} com o mesmo cumprimento que a variável y.
- Definindo o argumento *constante* como:
  - VERDADEIRO (ou omitido) a função fornece todos os coeficientes a e $b's$ da reta de regressão linear múltipla completa.
  - FALSO a função fornece apenas os coeficientes $b's$ da reta de regressão $\hat{y} = b_1x_1 + b_2x_2 + \cdots + b_nx_n$, isto é $a=0$;
- Definindo o argumento *estatísticas* como:
  - FALSO a função fornece somente os coeficientes a e $b's$.
  - VERDADEIRO (ou omitido) a função fornece os coeficientes a e $b's$ e as seguintes estatísticas: erros padrões dos coeficientes a e $b's$; o coeficiente de determinação r^2; o erro padrão da estimativa S_e; o valor da estatística F; o número de graus de liberdade gl da regressão; a soma dos quadrados dos desvios explicados *SSR*; a soma dos quadrados dos desvios não-explicados *SSE*.

[16] Em inglês, a função estatística PROJ.LIN é *LINEST*.

Para compreender o uso da função PROJ.LIN, realizaremos aplicações gradativas para o caso de regressão linear simples $\hat{y} = a + bx$, acompanhando o Exemplo 16.1, reproduzido na planilha **Função PROJ.LIN** incluída na pasta CAP_16, conforme apresentado na Figura 16.5.

	A	B	C	D	E	F	G
1	Função PROJ.LIN						
2							
3		y	x			Constante	Estatísticas
4		430	30			VERDADEIRO	FALSO
5		335	21			Função PROJ.LIN	9,7381
6		520	35			b =	9,7381
7		490	42			a =	117,0702
8		470	37				
9		210	20			Constante	Estatísticas
10		195	8			VERDADEIRO	VERDADEIRO
11		270	17			Função PROJ.LIN	9,7381
12		400	35			1	2
13		480	25		1	9,7381	117,0702
14					2	2,0487	59,0299
15					3	0,7385	65,1734
16					4	22,5939	8
17					5	95.969,3923	33.980,6077
18							

Figura 16.5 • A função PROJ.LIN resolvendo o Exemplo 16.1

PROJ.LIN(*série_y; série_x; VERDADEIRO; FALSO*).
Informando o argumento *constante* como VERDADEIRO definimos que a equação de regressão é $\hat{y} = a + bx$. Quando o argumento *estatísticas* é FALSO, a função PROJ.LIN fornece somente os coeficientes da reta de regressão. No intervalo de células F3:G7 da planilha **Função PROJ.LIN** tratamos deste caso. Na célula G5 digitamos a fórmula: =PROJ.LIN(B4:B13;C4:C13;F4;G4) obtendo como resultado o valor 9,7381.

Na realidade, a função PROJ.LIN registra na célula G5 uma *matriz* com dois resultados, apresentando na célula apenas um dos dois resultados. Para ver a *matriz* com os dois resultados registrados na célula G5 procedemos da seguinte maneira:

- Selecionamos com o cursor a célula G5.
- Depois, pressionando primeiro a tecla de função F2 e depois a tecla de função F9 aparecerá a *matriz*: ={9,73814229249012.117,070158102767}.

O primeiro valor, já apresentado na célula G5, corresponde ao valor do coeficiente b e o segundo valor, separado por um ponto, símbolo (.), corresponde ao valor do

coeficiente *a*. Para extrair os resultados da *matriz* devemos usamos a função ÍNDICE, como realizado nas células G6 e G7. As fórmulas são as seguintes:

- Célula G6[17]: =ÍNDICE(PROJ.LIN(B4:B13;C4:C13;F4;G4);1;1) obtendo o valor do coeficiente *b* igual a 9,73814229249012.
- Célula G7: =ÍNDICE(PROJ.LIN(B4:B13;C4:C13;F4;G4);1;2) obtendo o valor do coeficiente *a* igual a 117,070158102767. Finalmente, a equação da reta de regressão linear simples é $\hat{y} = 117{,}070158 + 9{,}738142x$

PROJ.LIN(*série_y; série_x; VERDADEIRO; VERDADEIRO*).
Se o argumento *constante* for VERDADEIRO a equação de regressão usada pela função PROJ.LIN é $\hat{y} = a + bx$. Se o argumento *estatísticas* for VERDADEIRO, a função PROJ.LIN fornecerá uma *matriz* com 10 resultados, os dois coeficientes da reta de regressão mais 8 resultados de interesse na regressão linear.

No intervalo de células F9:G17 da planilha **Função PROJ.LIN** tratamos deste caso. Na célula G11 digitamos a fórmula: =PROJ.LIN(B4:B13;C4:C13;F10;G10) obtendo como resultado o valor 9,7381. Na realidade, a função PROJ.LIN registrou na célula G11 uma *matriz* de 5 linhas e 2 colunas, apresentando, apenas, um dos dez resultados disponíveis na célula. Para ver os dez resultados da *matriz* registrada na célula G11 procedemos da seguinte maneira:

- Selecionamos com o cursor a célula G11.
- Depois, pressionando primeiro a tecla de função F2 e depois a tecla de função F9 aparecerá na célula e/ou na barra de fórmulas a seguinte expressão:
  ={9,73814229249012.117,070158102767;
  2,04870921558875.59,0298507995893;
  0,738510136917969.65,1734298885576;
  22,593920183783.8;
  95969,3922924901.33980,6077075099}

Nesta expressão existem 5 grupos de resultados separados pelo ponto e vírgula, símbolo (;). Cada um dos 5 grupos está composto por dois resultados separados pelo ponto, símbolo (.). O valor apresentado pela função PROJ.LIN na célula G11, é o primeiro resultado do primeiro grupo de resultados, que corresponde ao valor do coeficiente *b*, sendo que o segundo, separado pelo símbolo (.), corresponde ao valor do coeficiente *a*. Para extrair cada um desses valores de forma separada devemos usar a função ÍNDICE, formando a tabela limitada pelas células E12 e G17. A primeira coluna da tabela (com valores 1, 2, 3, 4, 5) identifica cada grupo de resultados, e a primeira linha da tabela (com valores 1 e 2) identifica o primeiro e o segundo resultado de cada grupo de resultados. Para extrair todos os resultados, registramos a fórmula na:

- Célula F13 =ÍNDICE(PROJ.LIN($B$4:$B$13;$C$4:$C$13;$F$10;$G$10);$E13;F$12)

[17] Em inglês, a função de procura e referência ÍNDICE é *INDEX*. A sintaxe desta função é ÍNDICE(*matriz; linha; coluna*).

Depois copiamos esta fórmula até a célula G17, obtendo a tabela do intervalo F13:G17 da planilha **Função PROJ.LIN**, Figura 16.5. Cada uma das 10 células da tabela, separadas em 5 grupos (as 5 linhas da tabela) tem o seguinte significado:

Coeficiente *b*=9,738142	Coeficiente *a*=117,0702
Erro padrão do coeficiente *b* S_b=2,048709	Erro padrão do coeficiente *a* S_a=59,02985
Coeficiente de determinação r^2=0,7385	Erro padrão da estimativa S_e=65,17343
F observado =22,59392	Graus de liberdade da regressão 8
Soma dos quadrados dos desvios explicados *SSR*=95.969,39	Soma dos quadrados dos desvios não-explicados *SSE*=33.980,61

Análise da Condição *constante* = FALSO
Quando o argumento *constante* da função PROJ.LIN é FALSO, a reta de regressão usada, aplicando o método dos quadrados mínimos, é $\hat{y} = b_1x_1 + b_2x_2 + \cdots + b_nx_n$, isto é, a=0 e a reta de regressão passa pela origem dos eixos. No caso de regressão linear simples $\hat{y} = bx$, se demonstra que:

$$b = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sum_{i=1}^{n} x_i^2}$$

Mudando o valor dos argumentos *constante* nas células F4 e F10 da planilha **Função PROJ.LIN** o leitor pode ver os resultados para os dois valores desse argumento. A partir da coluna J da planilha **Função PROJ.LIN** foram repetidos os mesmos cálculos informando as séries de informações como matrizes.

OUTRAS FUNÇÕES ESTATÍSTICAS DO EXCEL

Apresentaremos outras funções estatísticas disponíveis no Excel.

PROJ.LOG(*série_y; série_x; constante; estatísticas*)
A função estatística PROJ.LOG[18] dá como resultado uma *matriz* de resultados de uma curva exponencial de regressão do tipo $\hat{y} = b \times m_1^{x_1} \times m_2^{x_2} \times \cdots \times m_n^{x_n}$, aplicando o método dos quadrados mínimos, onde os argumentos têm os seguintes significados:

- É conhecida a série de dados ou observações, *série_y*.

[18] Em inglês, a função estatística PROJ.LOG é *LOGEST*.