R:

ANOVA

Fonte da variação	*SQ*	*gl*	*MQ*	*F*	*valor-P*	*F crítico*
Entre grupos	2,3409	2	1,1704	4,2235	0,0324	3,5915
Dentro dos grupos	4,7111	17	0,2771			
Total	7,0520	19				

Problema 3

É afirmado que o número de carros roubados por dia não depende da região da cidade. Para verificar essa afirmação, a cidade foi dividida em quatro zonas e durante 10 dias foram registrados os carros roubados nas quatro zonas, conforme registrado na tabela seguinte. Pede-se verificar essa afirmação, adotando um nível de significância de 0,05.

Zona 1	Zona 2	Zona 3	Zona 4
12	12	10	13
15	11	12	15
14	13	14	14
12	18	12	15
15	15	11	17
18	14	13	14
12	13	10	13
14	12	12	14
12	11	13	15
11	10	11	16

R: Rejeitar a hipótese nula. Existem diferenças entre as quatro zonas.

ANOVA

Fonte da variação	*SQ*	*gl*	*MQ*	*F*	*valor-P*	*F crítico*
Entre grupos	41	3	13,6666667	4,12060302	0,01304178	2,86626545
Dentro dos grupos	119,4	36	3,31666667			
Total	160,4	39				

(2)

Capítulo 16

REGRESSÃO LINEAR SIMPLES

Neste capítulo iniciaremos o estudo da relação linear entre variáveis aleatórias, dando destaque à análise entre duas variáveis aleatórias, denominada como *análise bidimensional*. Os procedimentos que operam com dados bidimensionais são a *análise de correlação* estudada no Capítulo 7, e a *análise de regressão simples* cujo objetivo é encontrar uma equação matemática que permita:

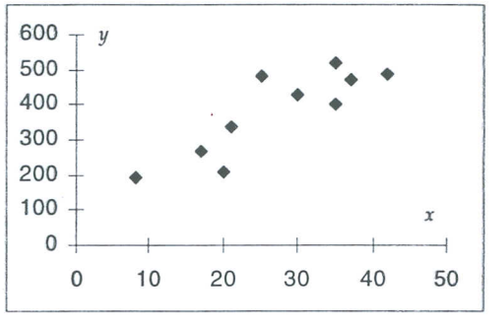
- *Descrever e compreender a relação entre duas variáveis aleatórias.*
- *Projetar ou estimar uma nova observação* ou *ajustar e controlar processos.* Conhecida a relação entre duas variáveis aleatórias; por exemplo, as vendas de um produto em função de diferentes valores de investimento na campanha de divulgação,
  - Pode-se usar uma das observações para prever a outra.
  - Podem-se realizar ajustes na procura de melhores resultados.

Exemplo 16.1

O diretor de vendas de uma rede de varejo com vendas a nível nacional, está querendo analisar a relação que existe entre o investimento em propaganda e o valor das vendas da empresa. O objetivo é ter uma equação matemática que permita realizar projeções e estimativas de vendas a partir do investimento em propaganda; isto é, encontrar a relação entre a variável dependente *vendas* e a variável independente *investimento em propaganda*. O departamento de vendas da rede de varejo relacionou as vendas anuais em milhões, denominada como *variável dependente y*, com o investimento anual em propaganda em milhões denominada como *variável independente x*, cujos valores estão registrados na tabela seguinte. Pede-se analisar a possibilidade de definir um modelo que represente a relação entre as variáveis da amostra.

x	30	21	35	42	37	20	8	17	35	25
y	430	335	520	490	470	210	195	270	400	480

Solução. O primeiro passo é representar os pares de observações num gráfico, como pode-se ver na figura seguinte. O registro dos 10 pares de dados no gráfico de dispersão *x,y* mostram uma *tendência de crescimento* positivo, isto é, na medida que o investimento em propaganda aumenta, as vendas também aumentam, e vice versa.



Aplicando o conhecimento de correlação, podemos dizer que as duas variáveis estão correlacionadas de forma positiva com um coeficiente de correlação menor que +1.

Pode-se ver que a tendência de crescimento positivo dos dados do Exemplo 16.1 sugere um modelo representado por uma linha reta. Portanto, o *modelo da reta*[1] é um modelo interessante para encontrar uma lei matemática entre a *variável independente* e *variável dependente*. Em outras palavras, o objetivo é *ajustar* uma reta a partir dos dados amostrais.

A RETA DE REGRESSÃO

A *reta de regressão* é representada pela equação $\hat{y} = a + bx$, onde, $\hat{y}$ é a variável dependente e x é a variável independente. Se os n pares de valores amostrais formassem uma reta, então a equação da reta ajustada conteria os n pontos amostrais. Em geral, os n pares de valores não estarão contidos numa reta, eles estarão distribuídos ao redor da reta ajustada. Nesse modelo se verifica que:

- Para um valor x_i podem existir um ou mais valores de y_i amostrados.
- Para esse mesmo valor x_i se terá um valor projetado $\hat{y}_i$.

[1] Da mesma forma como usamos a *média* para resumir uma variável aleatória, a *reta de regressão* é usada para resumir a estimativa linear entre duas variáveis aleatórias. Essa semelhança pode ser estendida para a variabilidade de médias entre amostras de uma mesma população, existindo variabilidade de retas.

- Para cada valor de x_i existirá um *desvio* d_i dos valores de $\hat{y}_i$ conforme indicado na Figura 16.1. Sempre teremos observações que não são pontos da reta.

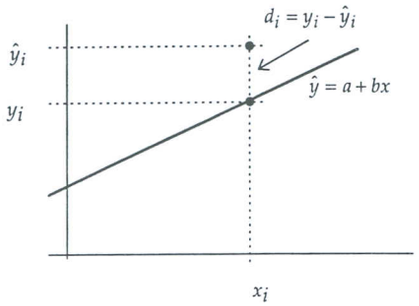


Figura 16.1 • Desvio do valor projetado

O gráfico da Figura 16.1 mostra que, em geral, para cada valor de x_i o valor observado e o valor projetado serão diferentes. Isto é, existirá um *desvio* d_i medido como:

$$d_i = y_i - \hat{y}_i .$$

Para encontrar a equação da reta, começamos por incluir a equação da reta de regressão na fórmula do desvio, isto é,

$$d_i = y_i - (a + bx_i)$$
$$d_i = y_i - a - bx_i$$

O objetivo é obter os valores dos coeficientes a e b da reta $\hat{y} = a + bx$, a partir dos n dados amostrais. Os coeficientes a e b são denominados como *coeficientes de regressão*, ou simplesmente *coeficientes* ou *constantes*, com os seguintes significados:

- O coeficiente b é a *declividade* da reta e define o aumento ou diminuição da variável y por unidade de variação da variável x.
- A constante a é denominada como *intercepto y*, sendo igual ao valor de $\hat{y}$ quando x é igual a zero[2].

Que critério devemos aplicar para obter os valores dos coeficientes a e b? A definição do critério para o ajuste de uma reta sobre o gráfico dos n pontos observados pode ser feito de diversas formas. Podemos entender que, quanto menor for a soma dos desvios de todos os pares de observações, melhor será o *poder de explicação* do modelo. A partir dessa idéia, temos os dois seguintes critérios:

- O primeiro critério é ajustar uma reta horizontal de valor igual à média dos valores de y, isto é, $\overline{y}$; pois, a média é uma reta de regressão com b=0. Este crité-

[2] Em alguns casos o valor de x=0 não tem significado prático.

rio não necessita de regressão, entretanto, é uma referência usada para medir o grau de explicação[3] da reta de regressão.

- O segundo critério é ajustar uma reta que divida os pontos observados de forma que a soma dos desvios seja nula. Entretanto, sabemos que a simples soma dos desvios[4] não oferece uma boa resposta devido às compensações dos valores dos desvios positivos e negativos.

O procedimento usado para obter os coeficientes da equação da reta de regressão, parte da soma dos quadrados dos desvios de todos os pontos observados, isto é,

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

O critério é encontrar os coeficientes a e b da reta de regressão que minimizam a soma dos quadrados dos desvios, denominado como *método dos quadrados mínimos*. Portanto, o objetivo é encontrar a e b de forma que a soma dos quadrados dos desvios seja um valor mínimo, isto é,

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \Rightarrow \textit{deve ser um valor mínimo}$$

Ao estabelecer que a soma dos quadrados dos desvios seja um valor mínimo, devemos aplicar conceitos de cálculo diferencial com derivadas parciais. Ao mesmo tempo, como as incógnitas do problema são duas, os coeficientes a e b, para poder resolver necessitamos formar um sistema com duas equações. Aplicando esses conceitos se obtém as equações dos coeficientes a e b seguintes:

$$\hat{y} = a + bx \quad \text{sendo,} \quad \begin{cases} a = \dfrac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} \\ b = \dfrac{n\sum_{i=1}^{n} x_i \times y_i - \sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \end{cases}$$

Finalmente, minimizar a soma dos quadrados dos desvios não garante que se tenha obtido a *melhor* reta ajustada, é apenas uma propriedade desejada de ajuste da reta.

[3] Mais adiante será usada para definir o *coeficiente de determinação*.

[4] A soma dos desvios é sempre igual a zero. Como existem muitas retas que cumprem com essa condição, não é usado este critério para ajustar uma reta.

Exemplo 16.2

Com os dados do Exemplo 16.1, pede-se:

1. Obter a reta de regressão linear, com o método dos quadrados mínimos.
2. Desenhar os dados e a reta de regressão.

Solução. A tabela seguinte apresenta os resultados intermediários necessários para calcular os coeficientes de regressão.

	x	y	x.y	x^2
	30	430	12.900	900
	21	335	7.035	441
	35	520	18.200	1.225
	42	490	20.580	1.764
	37	470	17.390	1.369
	20	210	4.200	400
	8	195	1.560	64
	17	270	4.590	289
	35	400	14.000	1.225
	25	480	12.000	625
Somas	**270**	**3.800**	**112.455**	**8.302**

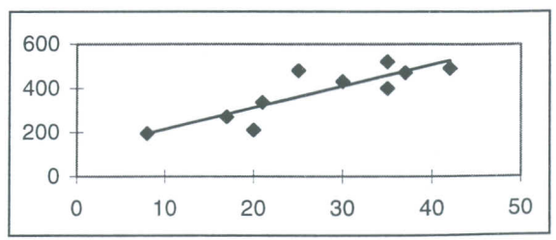
Aplicando as fórmulas calculamos os coeficientes da reta de regressão.

$$b = \frac{10 \times 112.455 - 270 \times 3.800}{10 \times 8.302 - 270^2} = 9{,}7381$$

$$a = \frac{3.800 - 9{,}7381 \times 270}{10} = 117{,}07$$

A equação da reta de regressão que cumpre com as premissas estabelecidas é a seguinte:

$$\hat{y} = 117{,}07 + 9{,}7381 \times x$$



Com os dados e resultados obtidos construímos o gráfico onde representamos os dados e a reta de regressão.

Uma das aplicações da regressão linear é projetar valores da variável dependente y para valores definidos da variável independente x.

Exemplo 16.3
Com os dados do Exemplo 16.1, pede-se projetar os valores de vendas para valores de investimentos em propaganda iguais de 20, 30 e 45 milhões.

Solução. Aplicando a equação da reta de regressão, obtemos a tabela com os resultados das projeções.

x	$\hat{y}$
20	311,83
30	409,21
45	555,29

O método de ajuste pelo método dos quadrados mínimos é preferível, pois:
1. Obtém as melhores estimações, isto é, as estimativas não terão tendenciosidade.
2. Onera os desvios maiores, fato desejável que evita grandes desvios.
3. Permite realizar testes de significância na equação de regressão.
4. A reta de regressão passa pelo ponto formado pelos valores das médias das duas séries de observações.

OUTRA FORMA DE APRESENTAÇÃO DA EQUAÇÃO DA RETA

As expressões matemáticas do cálculo dos valores dos coeficientes a e b não mostram que estão sendo aplicadas medidas estatísticas. Entretanto, esses conceitos estão presentes facilitando a compreensão e os seus cálculos. Realizando transformações algébricas nas expressões de a e b obtidas acima, se obtém outra forma de calcular a e b, mudando somente a simbologia, pois o conceito é o mesmo.

$$\begin{cases} a = \overline{y} - b\overline{x} \\ b = \dfrac{Cov(x,y)}{Var(x)} \end{cases}$$

Tendo presente que $Cov(x,y) = r_{xy}\sigma_x\sigma_y$, o coeficiente b será igual a:

$$b = \frac{r_{xy}\sigma_x\sigma_y}{\sigma_x^2}$$

Finalmente, as expressões dos coeficientes de regressão são:

$$\begin{cases} a = \overline{y} - b\overline{x} \\ b = r_{xy}\dfrac{\sigma_y}{\sigma_x} \end{cases}$$

Uma vantagem adicional desta forma de cálculo é que com os mesmos dados é possível calcular as duas possíveis retas de regressão linear, permutando as variáveis. Como regra geral, recomendamos ter presente que:
- O valor do coeficiente b é obtido como resultado da divisão da covariância das duas variáveis aleatórias pela variância da variável aleatória independente.
- O valor do coeficiente a é obtido como resultado da subtração da média da variável dependente menos o produto do coeficiente b pela média da variável independente.

Por exemplo, se y é a variável independente e x a variável dependente, o valor dos coeficientes da reta de regressão $\hat{x} = f(y)$, serão calculados com as fórmulas:

$$\begin{cases} a = \overline{x} - b\overline{y} \\ b = \dfrac{Cov(x,y)}{Var(y)} = r_{xy}\dfrac{\sigma_x}{\sigma_y} \end{cases}$$

Exemplo 16.4
Calcular o valor dos coeficientes da reta de regressão linear do Exemplo 16.1 usando as novas expressões dos coeficientes de regressão.

Solução. Resolvendo diretamente com o Excel na planilha **Exemplo 16.4** incluída na pasta CAP_16 obtemos a planilha da figura seguinte.

	A	B	C	D	E	F	G
1	Exemplo 16.4						
2							
3			x	y		Retas de Regressão	
4			30	430		y=f(x)	
5			21	335		b	9,7381
6			35	520		a	117,0702
7			42	490		x=f(y)	
8			37	470		b	0,0758
9			20	210		a	-1,8180
10			8	195			
11			17	270			
12			35	400			
13			25	480			
14		Média	27,00	380,00			
15		S	10,60	120,16			
16		r	0,86				
17							

...plo 16.1, a reta de regressão
...(y) tem os coeficientes a=–1,8180 e b=0,0758

USANDO O EXCEL

Na resolução dos exemplos anteriores foram obtidos os valores dos coeficientes de regressão e projetados valores de vendas usando a planilha Excel, seguindo o roteiro do desenvolvimento apresentado. Para cálculos de regressão linear, o Excel dispõe de muitas funções estatísticas. As funções estatísticas diretas que correspondem aos conceitos vistos até este momento são: INTERCEPÇÃO, INCLINAÇÃO, PREVISÃO e TENDÊNCIA, apresentadas na planilha **Funções** incluída na pasta CAP_16 e registradas na Figura 16.2.

INTERCEPÇÃO(*série_y; série_x*)
A função estatística INTERCEPÇÃO[5]dá como resultado o valor do coeficiente de regressão b denominado como *intercepto* da reta de regressão linear $\hat{y} = a + bx$, quando são conhecidas as duas séries de dados, *série_y* e *série_x*. Em outras palavras, a função INTERCEPÇÃO dá o valor de $\hat{y}$ quando x=0, resultando em $\hat{y} = a$. Ao usar a função INTERCEPÇÃO 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, numa célula do Excel devemos registrar a fórmula:
=INTERCEPÇÃO({430;335;520;490;470;210;195;270;400;480};
{30;21;35;42;37;20;8;17;35;25})

NCLINAÇÃO(*série_y; série_x*)
. função estatística INCLINAÇÃO[6]dá como resultado o valor do coeficiente b da ...ta de regressão linear $\hat{y} = a + bx$, quando são conhecidas as duas séries de dados, ...rie_y e *série_x*. Ao usar esta função deve-se tomar cuidado de fornecer os dados ... ordem correta, o primeiro argumento *Série_y* se refere à série de valores da vari...el dependente y, e o argumento *Série_x* se refere à série de valores da variável ...dependente x. Os argumentos *série_y* e *série_x* devem ser números ou nomes, ...trizes ou referências que contenham números. No caso de informar como ma...z, numa célula do Excel devemos registrar a fórmula:
...NCLINAÇÃO({430;335;520;490;470;210;195;270;400;480};
{30;21;35;42;37;20;8;17;35;25})

[5] ...n inglês, a função estatística INTERCEPÇÃO é *INTERCEPT*.
[6] ...n inglês, a função estatística INCLINAÇÃO é *SLOPE*.

PREVISÃO(*x; série_y; série_x*)
A função estatística PREVISÃO[7]dá como resultado o valor projetado $\hat{y}$ da reta de regressão linear simples $\hat{y} = a + bx$, para um único valor x, quando são conhecidas as duas séries de dados ou observações, *série_y* e *série_x*. Ao usar esta função 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, para projetar as vendas para um investimento igual a 20 milhões, numa célula do Excel devemos registrar a fórmula:
=PREVISÃO(20;{430;335;520;490;470;210;195;270;400;480};
{30;21;35;42;37;20;8;17;35;25}) → 311,83

Para projetar valores de $\hat{y}$, deve-se tomar o cuidado de escolher valores de x dentro da faixa de valores observados da variável independente x. A Figura 16.2 apresenta a resolução dos exemplos anteriores usando somente as funções estatísticas apresentadas.

	A	B	C	D	E	F	G
1	FUNÇÕES ESTATÍSTICAS						
2	Funções INCLINAÇÃO, INTERCEPÇÃO e PREVISÃO						
3							
4		x	y		Informando Intervalos		
5		30	430		b =	9,7381	
6		21	335		a =	117,0702	
7		35	520		Informando como Matriz		
8		42	490		b =	9,7381	
9		37	470		a =	117,0702	
10		20	210				
11		8	195		x	Projeção	Matriz
12		17	270		20	311,83	311,83
13		35	400		30	409,21	409,21
14		25	480		45	555,29	555,29
15							

Figura 16.2 • Resolução dos exemplos usando as funções estatísticas

Uma outra forma de estimar valores de y para diversos valores de x é usando a função estatística TENDÊNCIA.

TENDÊNCIA(*série_y; série_x; x's; constante*)
A função estatística TENDÊNCIA[8]dá como resultado o valor projetado $\hat{y}$ da reta de regressão linear simples, para um único ou um grupo de valores de x, denominado como $x's$, quando são conhecidas as duas séries de dados ou observações,

[7] Em inglês, a função estatística PREVISÃO é *FORECAST*.
[8] Em inglês, a função...

série_y e *série_x*. Ao usar a função TENDÊNCIA 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*. Quando o argumento *constante* for:

- VERDADEIRO (ou omitido) a função fornecerá um único ou um grupo de valores da reta de regressão $\hat{y} = a + bx$.
- FALSO, a função TENDÊNCIA fornece os resultados da reta de regressão $\hat{y} = bx$, isto é, com a=0.

A função TENDÊNCIA tem mais aplicações das apresentadas nesta parte, sugerimos que o leitor veja a *Ajuda on-line* do Excel para conhecer todas as aplicações possíveis da função TENDÊNCIA. A função PREVISÃO é equivalente à função TENDÊNCIA apenas para uma projeção por vez. Uma das vantagens da função TENDÊNCIA é poder formar *matrizes* de resultados conforme registrado na planilha **Funções** a partir da célula I2.

MEDINDO A UTILIDADE DA RETA DE REGRESSÃO

A reta obtida pelo método dos quadrados mínimos não descreve os dados perfeitamente, a reta é um resumo útil da tendência. Quão útil é essa reta de regressão? A resposta está baseada em duas medições importantes: *o erro padrão da estimativa* e no *coeficiente de determinação-r^2*.

ERRO PADRÃO DA ESTIMATIVA

O *erro padrão da estimativa* S_e informa de forma aproximada quão grande são os erros de estimativa, denominados como *resíduos*, do grupo de dados, medido na própria unidade de *y*. O objetivo é conseguir um valor de S_e tão pequeno quanto possível, podendo-se interpretar o valor de S_e como um desvio padrão de forma que se os resíduos tem distribuição normal, podemos esperar que 68% dos pontos se encontrem dentro de um desvio padrão, isto é, no intervalo $\pm 1 \times S_e$. O valor de S_e é obtido da definição de variância da amostra S_e^2 com (n-2) graus de liberdade:

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

Ao ajustar uma reta, se espera que ela *explique* o grupo de dados. Se os dados estiverem contidos numa reta, se obterá uma reta coincidente com os pontos observados e, dessa maneira, a soma dos quadrados dos desvios será igual a zero e a *explicação* da reta ajustada é completa. Portanto, devemos entender que o valor de *SSE*[9] é a parte *não explicada* pela regressão.

A partir de S_e^2 obtemos o valor do desvio padrão S_e conhecido como *erro padrão da estimativa*, que mede a dispersão dos desvios ao redor da reta de regressão:

$$S_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}}$$

Sendo cumpridas as premissas da regressão linear, se espera que aproximadamente 95% dos dados observados *y* se encontrem dentro do intervalo $\pm 2 \times S_e$ de seus respectivos valores projetados pela reta de regressão $\hat{y}$.

COEFICIENTE DE DETERMINAÇÃO - r^2

Sabemos que o coeficiente de correlação *r* mede o grau de associação linear de duas variáveis de um mesmo experimento. Como já foi visto, a covariância está relacionada também com o coeficiente *b* do modelo de regressão. Agora, vamos supor que escolhemos como modelo de regressão a reta de regressão horizontal $\hat{y} = \overline{y}$, isto é, a equação que representa a média de *y*. Neste caso, o coeficiente *b* da reta de regressão tem valor zero, concluindo que o valor da covariância é igual a zero e, conseqüentemente, o coeficiente de correlação também é nulo. Embora a reta da média não explique nada, é um ponto interessante de partida.

Analisando a reta de regressão com os coeficientes *a* e *b*, pode-se ver que a maioria dos dados estão distribuídos ao redor da reta. Definindo como, Figura 16.3:

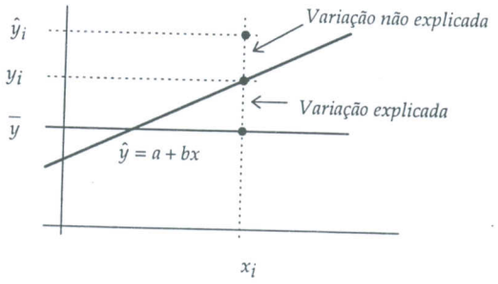


Figura 16.3 • Coeficiente de determinação

[9] Em inglês, SSE é *Sum of Square Errors*.