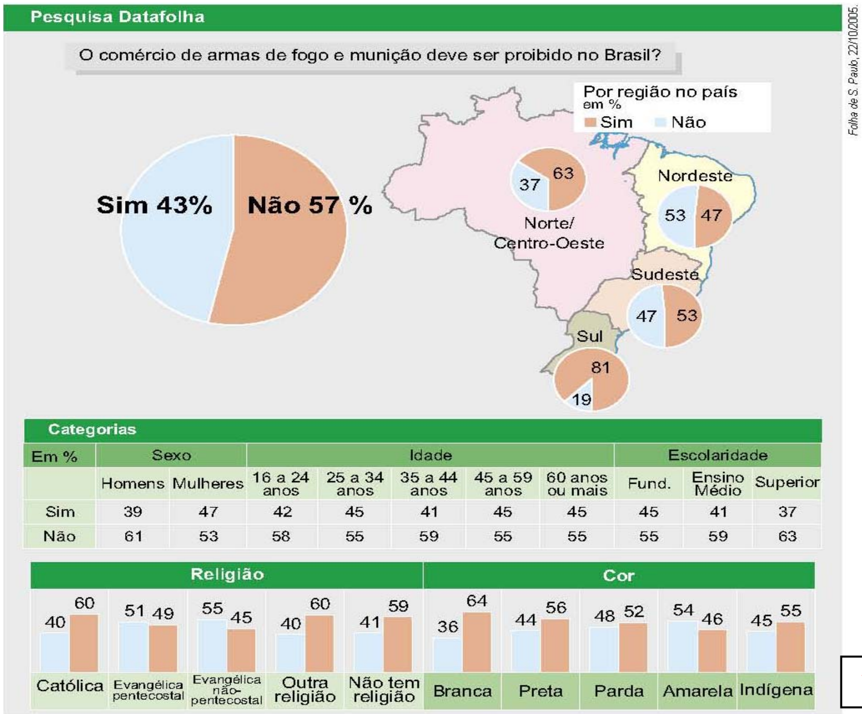
IMES-FAFICA

Tópicos de Matemática Aplicada

Estatística Aplicada no Excel

Ciência da Computação

Bertolo, L.A.

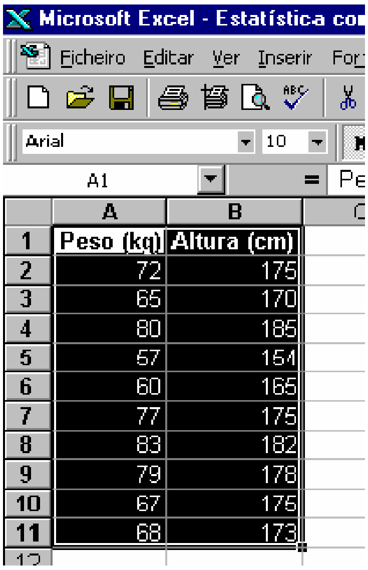


Versão BETA

Outubro 2008

Capítulo 3 – Dados Bivariados

São pares de valores correspondente a um dado indivíduo ou resultado experimental.
Para ilustrar o estudo de dados bivariados, recorreu-se ao exemplo de altura (cm) e peso (kg) de 10 alunos do curso de Ciência da Computação do IMES-FAFICA.



2.1 – Diagrama de Dispersão ou de Espalhamento (*scatter plot*)

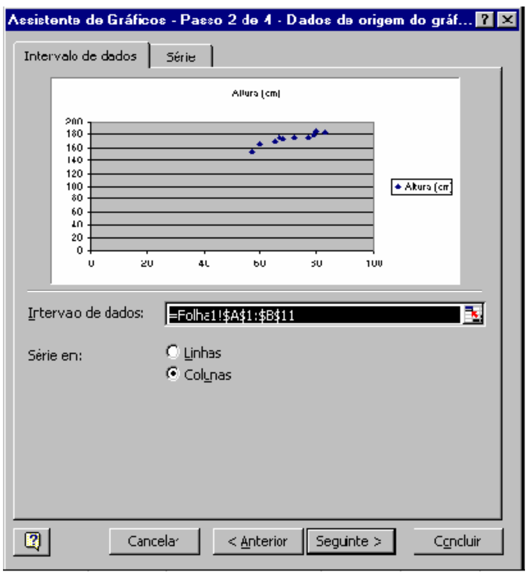
É uma representação gráfica para os dados bivariados, em que cada par de dados (xi, yi) é representado por um ponto de coordenadas (xi, yi), num sistema de eixos cartesianos.

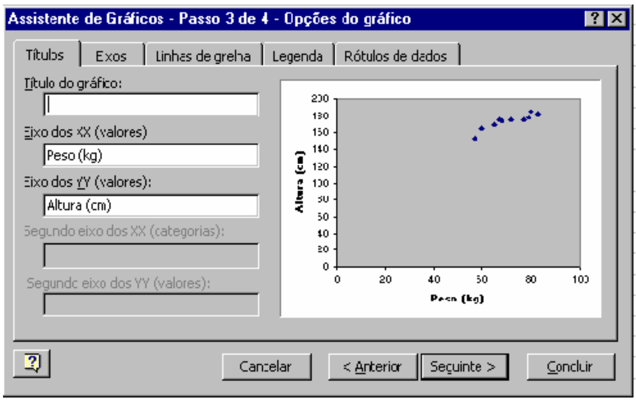
Pode-se obter com facilidade a representação gráfica de dados bivariados, através do **Assistente de Gráficos** [*Chart Wizard*].

Comece por selecionar as células contendo os dados e os respectivos títulos e clique no ícone  da **Barra de ferramentas**.

Na primeira **Caixa de diálogo** selecione a opção **Dispersão (xy)**.

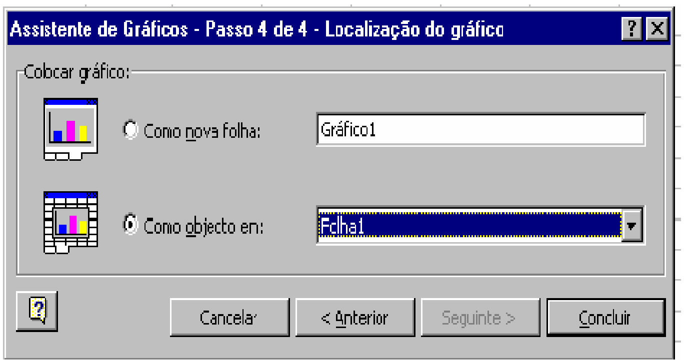
Para continuar a construção do gráfico, e para passar ao **Passo** seguinte, clique no botão **Seguinte >**.





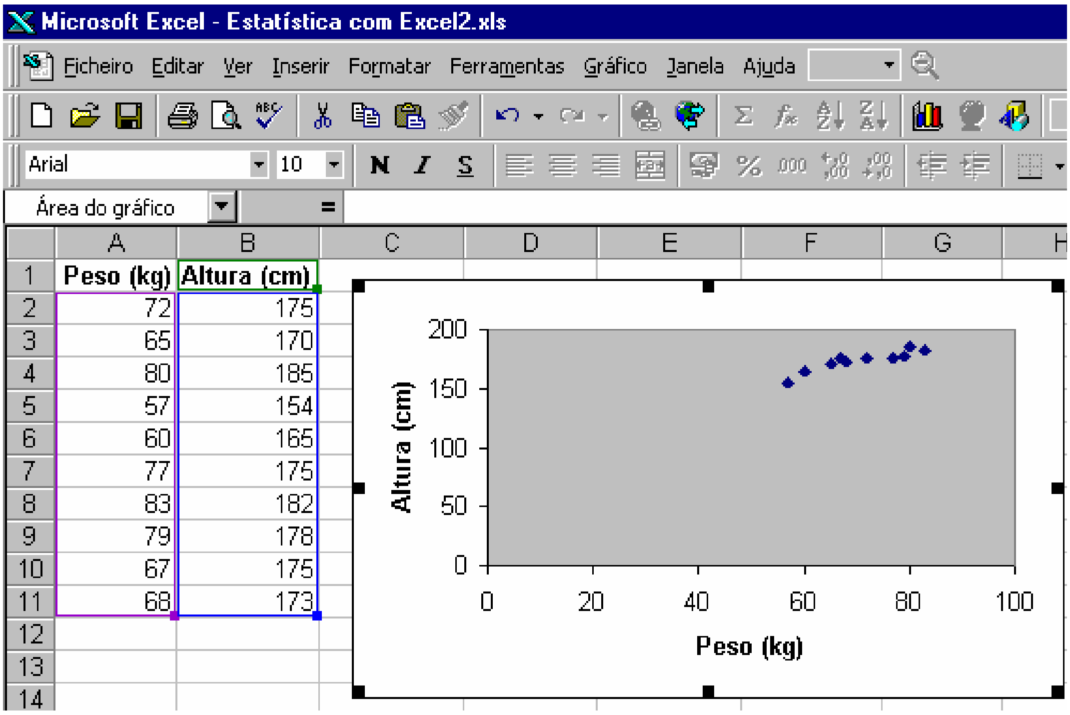
No terceiro passo, a **Caixa de diálogo** apresenta várias opções que permitem formatar o gráfico:
- Em **Títulos** siga o exemplo apresentado.
- Em **Linhas de grade**, desmarque a seleção da opção de grade.
- Em **Legenda**, desmarque a seleção da opção da legenda.

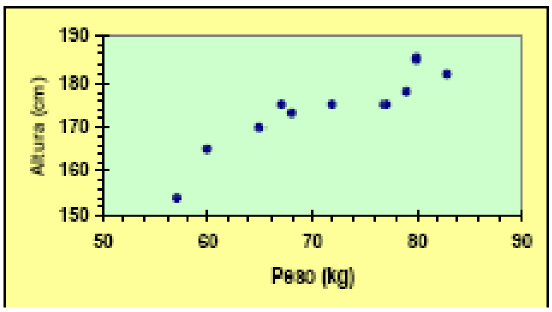
Para continuar a construção do gráfico, e para passar ao **Passo** seguinte, clique no botão **Seguinte >**.



No último passo pode escolher se o gráfico é colocado numa **nova folha de cálculo** ou numa folha já existente.

Clique em **Concluir** e obterá o seguinte resultado:





São múltiplas as opções de formatação para os gráficos de Excel, desde o aspecto geral, aos tipos de letras, à formatação dos eixos, etc. Eis um exemplo do que poderá obter.

2.2 – Covariancia e Correlação

Nós usamos regressão e correlação para descrever a variação em uma ou mais variáveis.

A. A **variação** é a soma dos desvios quadrados de uma variável de sua média.

$$\text{Variação} = \sum_{i=1}^{N} (x - \bar{x})^2$$

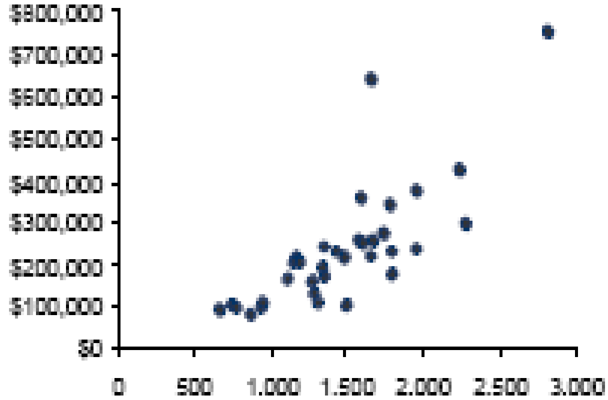
B. A variação é o numerador da **variância** de uma amostra:

$$\text{Variância} = \frac{\sum_{i=1}^{N} (x - \bar{x})^2}{N - 1}$$

C. Ambas, a variação e a variância são **medidas de dispersão** de uma amostra, já estudadas.

Exemplo1: Preços de vendas de casas e pés quadrados

Preços de venda de casas (eixo vertical) v. pés quadrados para uma amostra de 34 casas em Setembro de 2005 em St. Lucie County.



2.2.1 – A Covariância

A **covariância** entre duas variáveis aleatórias é uma medida estatística do grau para o qual as duas variáveis se movem juntas.

A. A covariância captura o quanto uma variável fica diferente da sua média quando a outra variável ficar diferente da sua média.

B. Uma covariância positiva indica que as variáveis tendem a se moverem juntas; uma covariância negativa indica que as variáveis tendem a se moverem em direções opostas.

C. A covariância é calculada como a razão da **co-variação** pelo tamanho da amostra menos um:

$$\text{Covariância} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

onde N é o tamanho da amostra
x_i é a i-ésima observação da variável x,
$\bar{x}$ é a média das observações da variável x,
y_i é a i-ésima observação da variável y, e
$\bar{y}$ é a média das observações da variável y.

D. O valor real da covariância não é significante porque ele não é afetado pela a escala das duas variáveis. Isto é o porquê de se calcular o coeficiente de correlação – para tornar algo interpretável da informação da covariância.

2.2.2 – A função COVAR do Excel

O Excel disponibiliza uma função embutida chamada COVAR que retorna a covariância, a média dos produtos dos desvios de cada par de ponto de dados em dois conjuntos de dados.

A sua sintaxe é:

COVAR(matriz1; matriz2)

2.2.3 – Exemplo 1 – Usando a função COVAR do Excel

Com os dados dos Pesos e Alturas da 10 feras do curso de Ciência da Computação (incluindo o Aderbal, por que não? Ele é uma fera ferida!!!!) encontre a covariância entre as grandezas peso e altura. Para tanto vá à célula C2 e digite =COVAR(A2:A11;B2:B11). O valor encontrado será:

	A	B	C	D	E	F
1	**Peso (kg)**	**Altura (cm)**				
2	72	175	63,44	<--=COVAR(A2:A11;B2:B11)		
3	65	170				
4	80	185				
5	57	154				
6	60	165				
7	77	175				
8	**83**	182				
9	79	178				
10	67	175				
11	68	173				

2.2.4 – Coeficiente de Correlação

O **coeficiente de correlação**, **r**, é uma medida da intensidade da relação entre ou dentre as variáveis.

Cálculo:

$$r = \frac{\text{covariância entre x e y}}{\left(\begin{matrix}\text{Desvio padrão}\\ \text{de x}\end{matrix}\right)\left(\begin{matrix}\text{Desvio padrão}\\ \text{de y}\end{matrix}\right)}$$

Nota: A correlação não implica que um causa o outro. Podemos dizer que duas variáveis X e Y estão correlacionadas, mas não que X causa Y ou que Y causa X, na média – eles simplesmente estão relacionados ou associados um com o outro.

$$r = \frac{\dfrac{\left(\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})\right)}{N-1}}{\sqrt{\dfrac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}\sqrt{\dfrac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}}}$$

2.2.5 – Exemplo 2

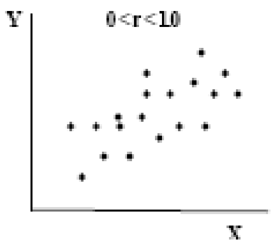
	A	B	C	D	E	F	G	H
1	**Observação**	**x**	**y**	**Desvio de x** $x - x_{Médio}$	**Desvio Quadrado de x** $(x - x_{Médio})^2$	**Desvio de y** $y - y_{Médio}$	**Desvio Quadrado de y** $(y - y_{Médio})^2$	**Produto dos desvios** $(x - x_{Médio})(y - y_{Médio})$
2	1	12	50	-1,50	2,25	8,40	70,56	-12,60
3	2	13	54	-0,50	0,25	12,40	153,76	-6,20
4	3	10	48	-3,50	12,25	6,40	40,96	-22,40
5	4	9	47	-4,50	20,25	5,40	29,16	-24,30
6	5	20	70	6,50	42,25	28,40	806,56	184,60
7	6	7	20	-6,50	42,25	-21,60	466,56	140,40
8	7	4	15	-9,50	90,25	-26,60	707,56	252,70
9	8	22	40	8,50	72,25	-1,60	2,56	-13,60
10	9	15	35	1,50	2,25	-6,60	43,56	-9,90
11	10	23	37	9,50	90,25	-4,60	21,16	-43,70
12	Soma	135	416	0,00	374,50	0,00	2342,40	445,00
13	Cálculos							
14	$x_{Médio}$=	135/10 =	13,5					
15	$y_{Médio}$=	416/10 =	41,6					
16	s^2_x=	374,5/9 =	41,611					
17	s^2_y=	2.342,4/9 =	260,267					
18	r =	$(445/9)/((41,611)^{1/2}(260,267)^{1/2})$ = 49,444/(6,451*16,133) = 0,475						

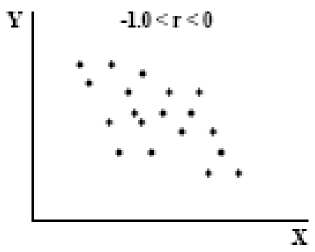
i. O tipo de relação está representada pelo coeficiente de correlação:

r =+1 correlação perfeitamente positiva
+1 >r > 0 relação positiva
r = 0 nenhuma relação
0 > r > –1 relação negativa
r = –1 correlação perfeitamente negativa

ii. Você pode determinar o grau de correlação observando o gráfico de espalhamento.

- Se a relação é para cima existe **correlação positiva**.
- Se a relação é para baixo existe **correlação negativa**.





iii. O coeficiente de correlação está limitado por –1 e +1. Quanto mais próximo o coeficiente estiver de –1 ou +1, mais forte é a correlação.

iv. Com a exceção dos extremos (isto é, r = 1,0 ou r = -1), nós não podemos realmente falar acerca da intensidade de uma relação indicada pelo coeficiente de correlação sem um teste estatístico de significância.

2.2.6 – A função CORREL do Excel

O Excel disponibiliza uma função embutida chamada CORREL que retorna o coeficiente de correlação entre duas variáveis de dois conjuntos de dados.

A sua sintaxe é:

CORREL(matriz1; matriz2)

2.2.7 – Exemplo – Usando a função CORREL do Excel

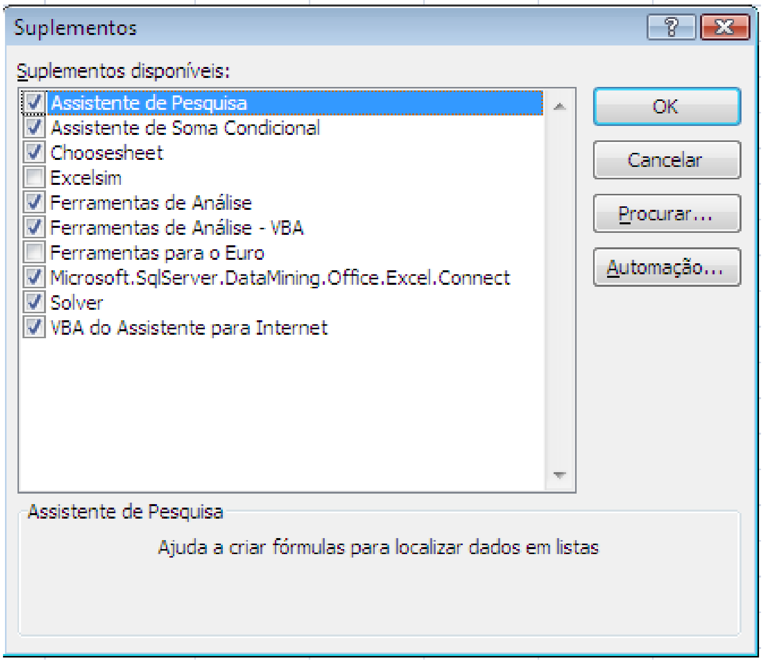
Determina-se o coeficiente de correlação através da função **CORREL** do Excel para as variáveis peso e altura das feras do truco da Computação (com o Aderbal é claro!).

O valor encontrado será:

	A	B	C	D	E	F
1	Peso (kg)	Altura (cm)				
2	72	175				
3	65	170				
4	80	185	0,906819	<--=CORREL(A2:A11;B2:B11)		
5	57	154				
6	60	165				
7	77	175				
8	**83**	182				
9	79	178				
10	67	175				
11	68	173				

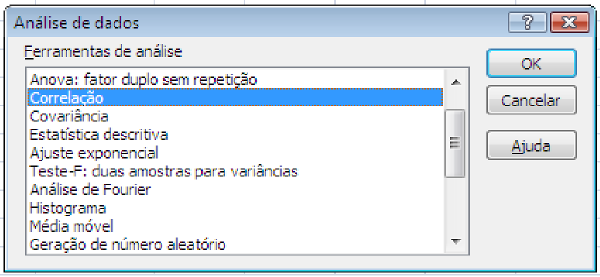
2.2.8 – Exemplo – Usando a ferramenta Análise de dados do Excel

Alternativamente poderíamos usar a ferramenta **Análise de dados.**
Para ativá-la no Office 2007 clique no botão do Office , daí em **Opções do Excel**. Na janela *Opções do Excel*, clique em **Suplementos** e vá até o final desta janela, na caixa de combinação **Gerenciar**, clique no botão **Ir...** para fazer aparecer a caixa *Suplementos*:

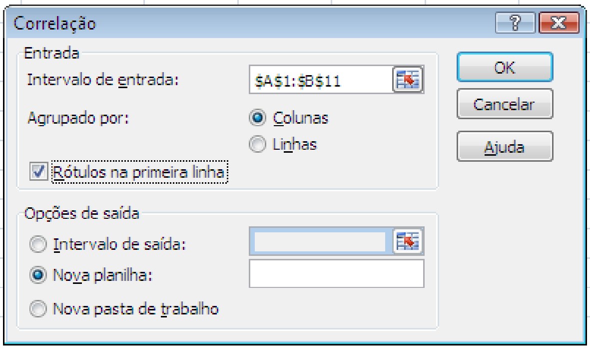


Assinale a caixa de verificação **Ferramentas de Análise**. Faça isto sempre para carregar os suplementos que às vezes podem não estar instalados.

A seguir vá a guia **Dados** e no grupo **Análise**



Clicando o botão OK aparecerá uma nova janela:



Configure nesta janela a Entrada dos dados, o Agrupamento, se deseja ou não os Rótulos na primeira linha e as Opções de saída. Faça tudo como mostra a figura. Depois aperte o botão OK e terás:

	A	B	C	D
1		Peso (kg)	Altura (cm)	
2	Peso (kg)	1		
3	Altura (cm)	0,90681871	1	
4				

2.3 – Regressão Linear Simples

Regressão é a análise da relação entre uma variável e alguma outra variável(s), assumindo uma relação linear. Também referida como **regressão dos mínimos quadrados** e **mínimos quadrados ordinários** (*ordinary least squares - OLS*).

Isto acontece quando a correlação entre as duas variáveis é <u>elevada</u> (quer seja positiva, quer seja negativa), isso significa que se conhecer o valor de uma das variáveis, então é possível ter uma idéia do valor que a outra variável irá tomar. Em linguagem estatística, diz-se que se pode **inferir** o valor de outra variável.

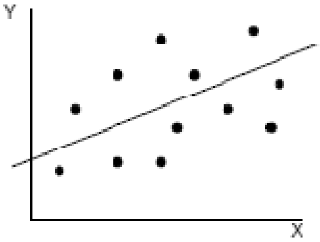
A. O propósito é explicar a variação numa variável (isto é, como uma variável difere do seu valor médio) usando a variação em uma ou outras mais variáveis.

B. Suponha que queremos descrever, explicar, ou predizer porque uma variável difere de sua média. Seja a i-ésima observação desta variável representada como Y_i, e seja n indicando o número de observações.

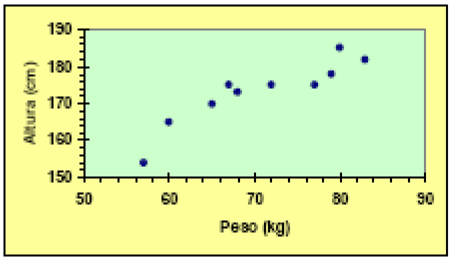
   A variação nos Y_i's (os quais queremos explicar) é:

$$\begin{matrix}\text{Variação}\\ \text{do Y}\end{matrix} = \sum_{i=1}^{N} (y_i - \bar{y})^2 = SS_{Total}$$

C. O princípio dos mínimos quadrados é que a linha de regressão é determinada minimizando a soma dos quadrados das distâncias verticais entre os valores reais de Y e os valores previstos de Y.



Uma linha é um ajuste através dos pontos XY tal que a soma dos resíduos quadráticos (isto é, a soma dos quadrados da distância vertical entre as observações e a linha) seja minimizada.



Voltando ao exemplo das alturas e dos pesos das feras e ao seu diagrama de dispersão, pode-se observar uma associação linear entre o peso e a altura. **Será que é possível prever a altura de um aluno que pese 70 kg?**

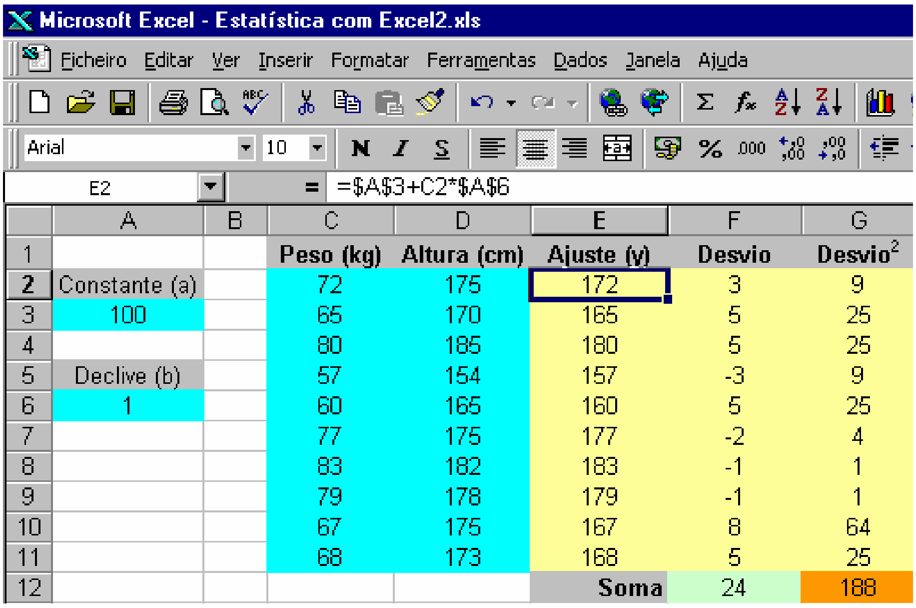
Quando perante uma situação análoga, em que tenhamos um conjunto de dados bivariados (xi, yi), i=1, ..., n, que seguem um padrão linear, poderá ter interesse ajustar uma reta da forma:

$$y = a + bx$$

que dê a informação de como se refletem em y, as mudanças processadas em x.

2.3.1 – O Exemplo 1 – Brincando com os dados

Retomando o exemplo, prepare uma tabela idêntica à que se apresenta. Os valores do Ajuste, do Desvio e do Desvio^2, poderão ser calculados com as seguintes expressões:



	A	B	C	D	E	F	G
1			Peso (kg)	Altura (cm)	Ajuste (y)	Desvio	Desvio^2
2	Constante (a)		72	175	172	3	9
3	100		65	170	165	5	25
4			80	185	180	5	25
5	Declive (b)		57	154	157	-3	9
6	1		60	165	160	5	25
7			77	175	177	-2	4
8			83	182	183	-1	1
9			79	178	179	-1	1
10			67	175	167	8	64
11			68	173	168	5	25
12					Soma	24	188

- Ajuste (y)
1º valor (**célula E2**)

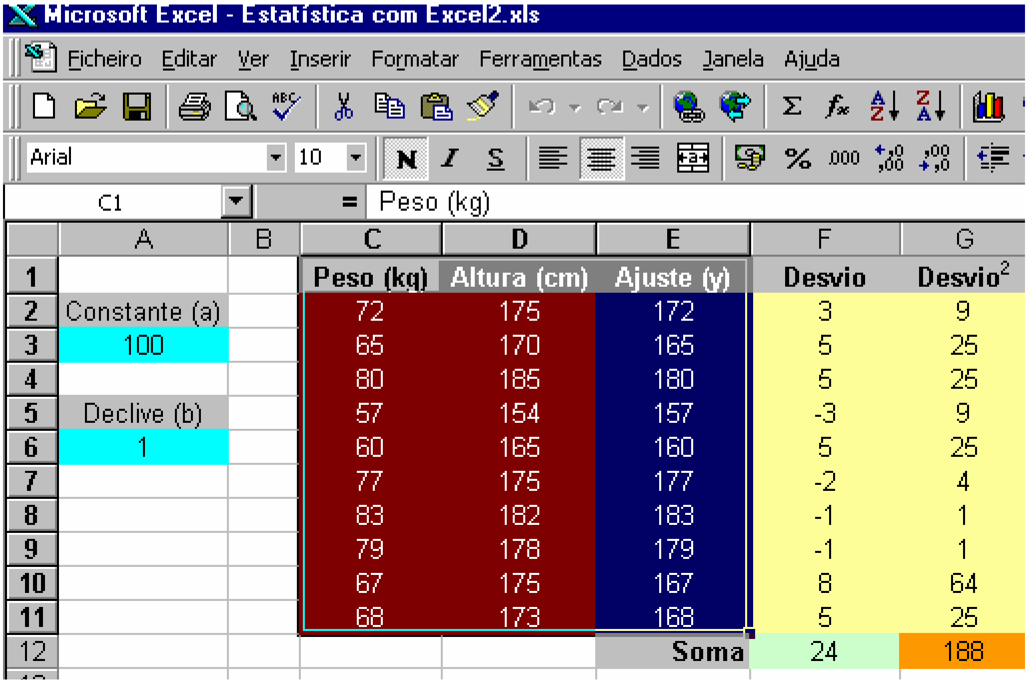
=A3+C2*A6

Copie esta expressão para **as células E3 a E11**.

- Desvio
1º valor (**célula F2**)

=D2-E2

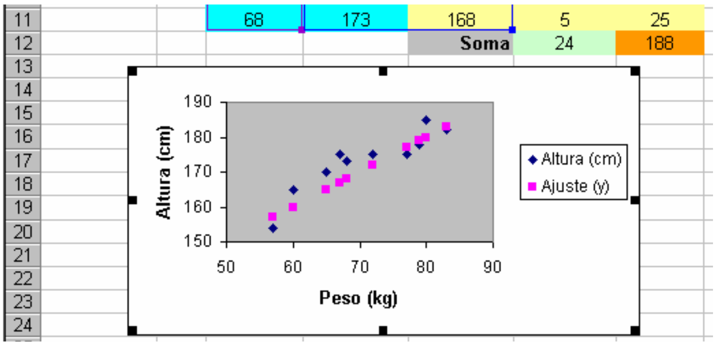
Copie esta expressão para **as células F3 a F11**.
- Desvio^2
1º valor (**célula G2**)
=F2^2
Copie esta expressão para **as células G3 a G11**.



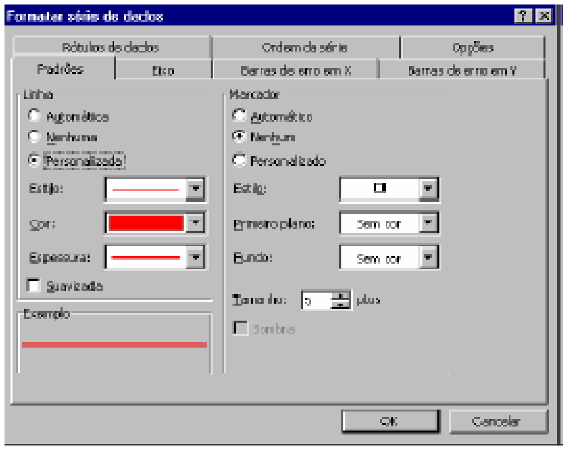
	A	B	C	D	E	F	G
1			Peso (kg)	Altura (cm)	Ajuste (y)	Desvio	Desvio^2
2	Constante (a)		72	175	172	3	9
3	100		65	170	165	5	25
4			80	185	180	5	25
5	Declive (b)		57	154	157	-3	9
6	1		60	165	160	5	25
7			77	175	177	-2	4
8			83	182	183	-1	1
9			79	178	179	-1	1
10			67	175	167	8	64
11			68	173	168	5	25
12					Soma	24	188

Selecione as células das três primeiras colunas contendo os dados e os respectivos títulos e clique no ícone da **Barra de ferramentas**.

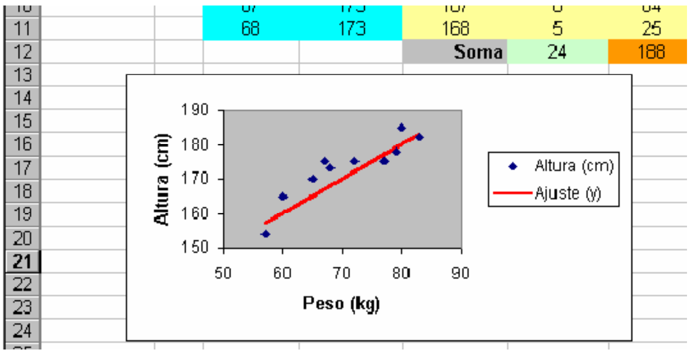
Siga os procedimentos anteriormente descritos e construa um diagrama de dispersão.



Selecione a série de dados correspondente ao "Ajuste (y)" e clique duas vezes, para abrir o menu **Formatar série de dados**.



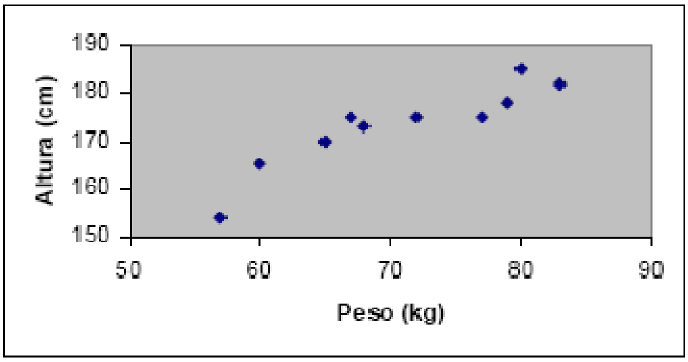
Na opção **Linha**, personalize de acordo com o exemplo.
Na opção **Marcador**, selecione: **Nenhum**



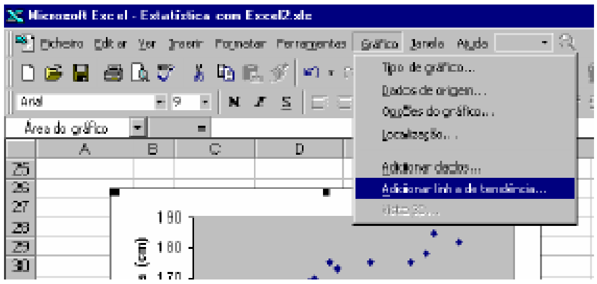
Experimente agora alterar os valores da "Constante(a)" e do "Declive (b)" e observe como se comporta a reta...

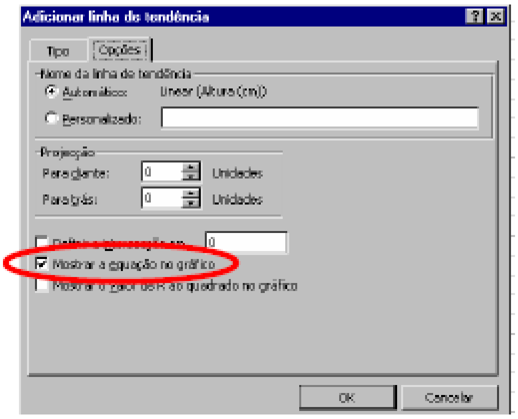
2.3.2 – O Exemplo 1 – Fazendo a sua Regressão Linear

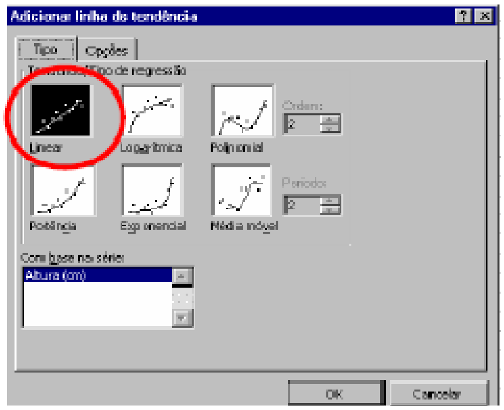
Um dos métodos mais conhecidos de ajustar uma reta a um conjunto de dados é o método dos mínimos quadrados, que consiste em determinar a reta que minimiza a soma dos quadrados dos desvios (ou erros) entre os verdadeiros valores de y e os obtidos a partir da reta que se pretende ajustar.
Construa novamente o diagrama de dispersão.

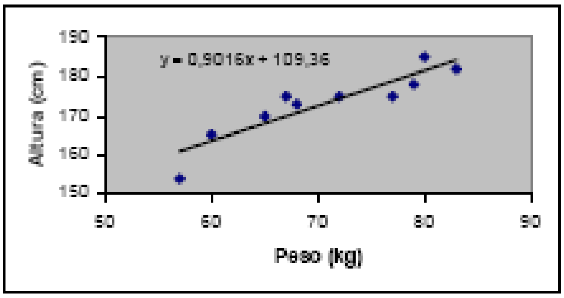


Selecionando o diagrama, clique no menu **Gráfico**, selecione o comando **Adicionar linha de tendência** e siga as opções.









A equação desta reta traduz-se em:

Altura = 109,36 + 0,9016 x Peso

Substituindo na equação o Peso por 70, obtém-se o valor de 172,472, pelo que a **altura esperada para um aluno que pese 70 kg , é de cerca de 172,5 cm.**

2.3.3 – Coeficiente de determinação R^2

O **coeficiente de determinação**, R^2, é a porcentagem da variação da variável dependente (variação dos Y_i's ou a soma dos quadrados total, SST) explicada pela variável independente(s).

A. O coeficiente de determinação é calculado como:

Observação	x	y	^y	y-^y	e^2
1	12	50	39,82	10,18	103,63
2	13	54	41,01	12,99	168,74
3	10	48	37,44	10,56	111,51
4	9	47	36,25	10,75	115,56
5	20	70	49,32	20,68	427,66
6	7	20	33,88	-13,88	192,65
7	4	15	30,31	-15,31	234,40
8	22	40	51,70	-11,70	136,89
9	15	35	43,38	-8,38	70,22
10	23	37	52,89	-15,89	252,49
				0,00	1.813,77

$$R^2 = \frac{\text{Variação explicada}}{\text{Variação total}} = \frac{\text{Variação total} - \text{Variação explicada}}{\text{Variação total}} = \frac{SS_{\text{Total}} - SS_{\text{Residual}}}{SS_{\text{Total}}} = \frac{SS_{\text{Regressão}}}{SS_{\text{Total}}}$$

Voltando ao exemplo 2.2.5 temos:

Observe que: (20-4) + (20-15) + (20 – 24) + (20 – 27) + (20 – 30) = 0

B. Um R^2 de 0,49 indica que as variáveis independentes explicam 49% da variação da variável dependente.

Exemplo 2, continuação

Continuando o exemplo de regressão anterior, podemos calcular o R^2.

x	y	$(y - y_{Médio})^2$	^y	y-^y	$(\hat{y} - y_{Médio})^2$	ε^2
12	50	70,56	39,82	10,18	3,17	103,63
13	54	153,76	41,01	12,99	0,35	168,74
10	48	40,96	37,44	10,56	17,31	111,51
9	47	29,16	36,25	10,75	28,62	115,56
20	70	806,56	49,32	20,68	59,60	427,66
7	20	466,56	33,88	-13,88	59,60	192,65
4	15	707,56	30,31	-15,31	127,46	234,40
22	40	2,56	51,70	-11,70	102,01	136,89
15	35	43,56	43,38	-8,38	3,17	70,22
23	37	21,16	52,89	-15,89	127,46	252,49
	416	2.342,40	416,00	0,00	528,75	1.813,77

R^2 = 528,77/ 2.342,40 = **22,57%** ou

R^2 = 1 – (1.813,63/2.342,40) = 1 – 0,7743 = **22,57%**.

2.4 – Trabalho Final

Parte A –

a. Fazer a mesma coisa da seção 2.2.3 para os dados do exemplo 2
b. Faça mesma coisa da seção 22.7 para os dados do exemplo 2
c. Faça mesma coisa da seção 22.8 para os dados do exemplo 2

Parte B –

Faça a mesma coisa da seção 2.3.2 – Regressão Linear Simples para os dados do exemplo 2, encontrando no final a equação da reta. Resposta $y_i = 25,559 + 1,188\ x_i$

Parte C –

Dada a amostra da planilha abaixo:
Análise de precificação de casas, repita os exercícios 1 e 2 e a secção 2.3.3 (coeficiente de determinação R^2)

	A	B	C	D	E	F	G	H	I	J	K	L
1	Análise de precificação de casas											
2	preços de casas	Pés quadradc	de quartos	de banheiros	vagas na garagem	tem piscina	sobre um lago	um campo de golfe				
3	$274.900	237	3	2	2	1	0	0		1 se sim, 0 se não		
4	$98.000	145	2	2	0	0	0	0				
5	$379.900	282	3	2	2	1	0	0				
6	$575.000	348	4	3	3	1	0	0				
7	$253.990	281	3	2	2	0	0	0				
8	$347.000	288	4	2	2	1	0	0				
9	$529.900	232	4	3,5	2	0	1	0				
10	$226.900	142	3	2	2	0	0	0				
11	$225.000	134	3	2	1	0	0	0				
12	$248.900	111	3	2	2	1	0	0				
13	$789.000	382	4	3	2	1	1	0				
14	$599.000	307	3	3 1/2	3	0	0	0				
15	$499.000	232	4	3	2	1	0	0				
16	$277.977	173	3	2	2	0	0	0				
17	$299.000	164	3	2	2	0	0	0				
18	$329.900	167	3	2	2	0	0	0				
19	$399.999	221	4	2	2	0	0	0				
20	$185.900	154	3	2	2	0	0	0				
21	$294.900	259	4	2	2	0	0	0				
22	$449.900	302	4	3,5	2	1	0	0				
23	$384.990	324	6	4	2	1	0	0				
24	$210.000	126	2	2	2	1	1	0				
25	$75.000	88	2	2	1	0	0	0				
26	$179.000	89	2	2	2	1	0	0				
27	$1.400.000	405	4	4	2	1	1	0				
28	$218.000	144	3	1	2	1	0	0				
29	$176.000	111	2	2	1	0	0	0				
30	$222.000	143	3	2	2	0	0	0				
31	$299.000	238	3	2	2	1	0	0				
32	$429.000	301	4	2	2	0	0	0				
33	$499.000	275	3	2	3	1	1	0				
34	$1.295.000	223	3	2,5	2	1	0	0				
35	$248.900	111	3	2	2	1	0	0				
36	$269.000	195	4	2	2	0	0	0				
37	$347.000	288	4	2	2	1	0	0				
38	$315.000	200	4	3	2	1	0	0				
39	$505.000	355	4	3	2	1	0	0				
40	$525.000	284	4	2	2	0	0	0				
41	$298.900	164	3	2	0	0	0	0				
42	$169.900	173	3	2	0	0	0	0				
43	$159.900	122	3	2	0	0	0	0				
44	$366.000	186	3	2	2	1	0	1				
45	$459.000	259	3	2	2	0	1	0				
46	$389.000	279	4	3	2	1	0	0				
47	$269.000	201	3	2	2	0	0	0				
48	$268.900	151	3	2	2	1	0	0				
49	$798.500	242	4	2	2	1	1	0				
50	$550.000	325	5	3	2	1	0	0				
51	$299.999	168	3	2	2	0	0	0				
52	$200.000	109	3	2	0	0	0	0				
53	$159.000	85	2	1	1	0	0	0				
54	$5.200.000	702	5	6,5	3	1	1	0				
55	$4.300.000	465	4	6,5	3	1	1	0				
56	$4.000.000	367	3	5,5	3	1	1	0				
57	$2.385.000	459	3	3,5	2	1	1	0				
58	$1.650.000	290	3	3	2	1	0	1				
59												
60												

Plan8 Plan1

Pronto