4. REGRESSÃO E CORRELAÇÃO

4.1- DADOS BIVARIADOS

Por vezes os investigadores realizam experiências em que mais do que uma variável é observada.

Por exemplo, um economista pode estar interessado em observar a quantia dispendida por família em artigos de mercearia e também o número de pessoas dessa família. Um agente imobiliário pode observar o preço das casas e a sua área, um médico mede a pressão sistólica e diastólica de um paciente, etc.

Quando duas variáveis são observadas para a mesma unidade experimental o resultado da experiência é uma **variável bivariada**. Como se devem representar estas variáveis? Estas variáveis são importantes quando estudadas separadamente, mas também podemos estar interessados em explorar *a relação entre as duas.* Há representações gráficas que permitem o estudo em conjunto das duas variáveis. Tal como no caso univariado há diferentes representações gráficas para diferentes tipos de variáveis.

4.1.1- GRÁFICOS PARA VARIÁVEIS QUALITATIVAS

Quando pelo menos um das variáveis é *qualitativa*, podemos usar representações em gráficos circulares e diagramas de barras. Por vezes temos uma variável quantitativa e outra qualitativa, medidas em duas populações ou grupos diferentes. Neste caso, podemos representar os dados por diagramas circulares colocados lado a lado ou por gráficos de barras nos quais estas são colocadas lado a lado para as duas populações que podem assim ser comparadas. Uma outra maneira é colocar as barras referentes a cada população em cima uma da outra. Iremos exemplificar estes procedimentos.

EXEMPLOS: 1- *Serão os professores das universidades privadas mais bem pagos do que os das universidades públicas?*

Os dados da tabela seguinte referem-se a uma amostra de 400 professores de universidades Americanas para os quais foram registados a categoria, o tipo de universidade e o salário médio auferido em milhares de dólares.

	Professor Catedrático	**Professor Associado**	**Professor Auxiliar**
Pública	8	7,5	6,8
Privada	8,5	7,8	7

Para representar graficamente estes dados podemos usar digramas de barras colocados lado a lado:

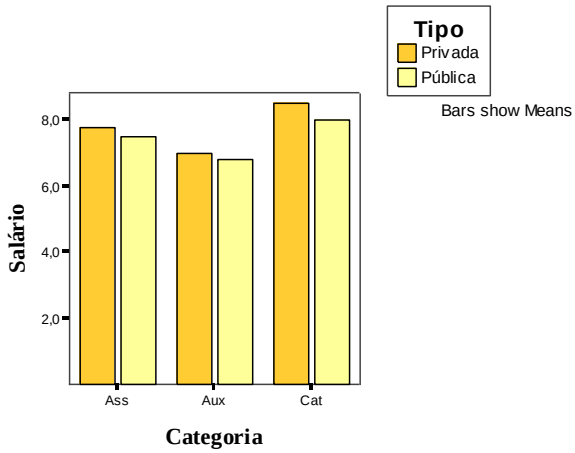

Figura-4.1

2- *Será que as escolas privadas empregam tantos professores qualificados como as públicas?*

Para responder a esta questão registaram-se duas variáveis qualitativas para cada professor: categoria na carreira e tipo de universidade, obtendo-se os seguintes resultados:

	Professor Catedrático	**Professor Associado**	**Professor Auxiliar**	**Total**
Pública	24	57	69	150
Privada	60	78	112	250

Note-se que os valores da tabela não representam os valores de uma variável quantitativa observada para cada professor, mas a *frequência absoluta* ou número de professores que caem em cada categoria. Para comparar estes números entre escolas públicas e privadas, vamos fazer a sua representação em dois diagramas circulares e colocá-los lado a lado.

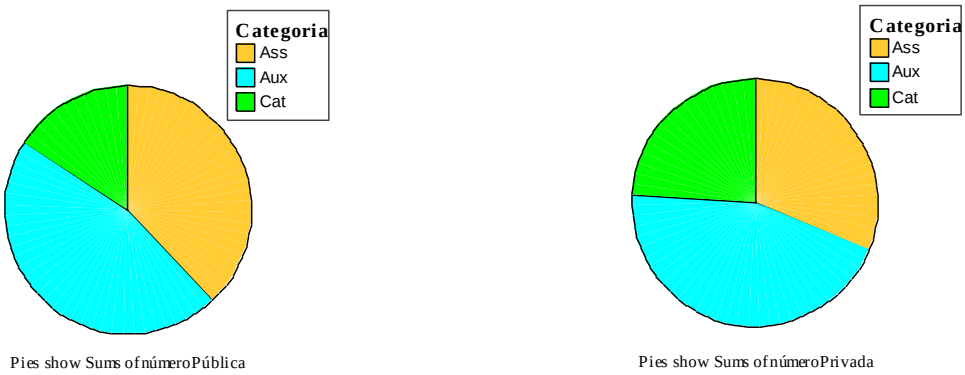

Figura-4.2

Também podemos calcular medidas numéricas para ajudar a comparar a distribuição dos professores nas escolas públicas e privadas.

	Professor Catedrático	**Professor Associado**	**Professor Auxiliar**	**Total**
Pública	$\frac{24}{150} = 0{,}16$	$\frac{57}{150} = 0{,}38$	$\frac{69}{150} = 0{,}46$	1,00
Privada	$\frac{60}{250} = 0{,}24$	$\frac{78}{250} = 0{,}31$	$\frac{112}{250} = 0{,}45$	1,00

Podemos ainda fazer uma representação gráfica em diagrama de barras "empilhadas".

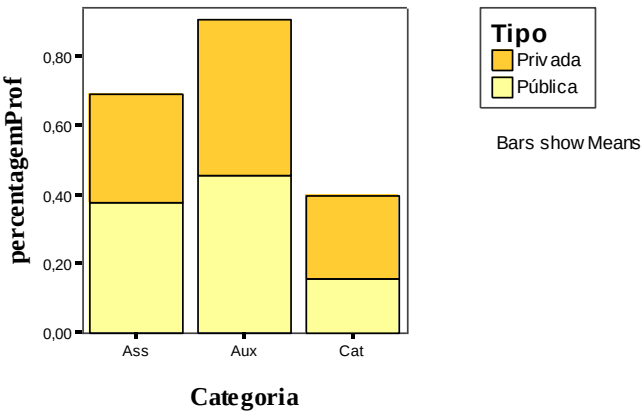

Figura-4.3

Chegamos à conclusão de que as escolas públicas têm menos professores catedráticos e mais professores associados do que as privadas. Não sabemos as razões para estas diferenças. Talvez que as escolas privadas atraiam os professores mais graduados por lhes pagarem melhor ou as escolas públicas não abram lugares para promover os professores associados?

4.1.2- DIAGRAMAS PARA DUAS VARIÁVEIS QUANTITATIVAS

Quando duas variáveis são quantitativas podemos representá-las graficamente, uma no eixo dos *x* e a outra no eixo dos *y* (num sistema de eixos cartesiano). A um gráfico destes chama-se **"diagrama de dispersão"**ou em inglês **"scatterplot".**

Podemos descrever a relação existente entre as variáveis *x* e *y* através do aspecto (padrão) exibido pela nuvem de pontos do gráfico.

Que tipo de padrão se vê? Há alguma tendência ascendente ou descendente que siga um padrão linear nas observações? Não existe qualquer tipo de padrão, mas somente uma distribuição aleatória dos pontos?

Quão acentuado é o padrão? Todas as observações seguem exactamente o mesmo padrão ou a relação visível é fraca?

Existem observações aberrantes? Um *outlier* é uma observação que se afasta das outras. As observações distribuem-se por grupos? Há alguma razão para que isto aconteça?

4.1.3- O COEFICIENTE DE CORRELAÇÃO

Exemplo 1: **níveis de enzima no sangue**

Para efeitos de um estudo médico sobre níveis de concentração de diferentes tipos de enzimas no sangue, recolheram-se amostras de sangue de mulheres com idades compreendidas entre os 40 e os 60 anos de idade. Gostaríamos de saber se há ligações entre os níveis destas enzimas, cuja existência poderia ajudar a identificar reacções biomédicas que poderiam estar a ocorrer com estes doentes.

Consideremos então os valores observados e que se encontram na tabela seguinte:

Tabela-1

Testosterona (A)	SHBG	AND
5.85	3.50	0.92
5.91	3.81	0.88
6.20	3.89	1.16
6.39	3.14	1.22
6.63	3.14	0.88
6.63	3.09	1.10
6.32	2.64	1.13
6.30	3.37	1.03
6.20	3.40	1.13
6.41	3.26	0.83
6.40	2.94	0.69
5.89	3.30	0.74
6.43	3.00	1.36
6.48	3.00	1.06
5.83	3.81	1.16
6.12	3.47	0.79
6.23	3.58	0.69
6.39	3.53	1.16
6.20	3.33	1.13
6.49	3.56	1.16
5.96	3.54	0.96

Cientificamente não há indicação precisa de que a ocorrência de um determinado nível de enzima influencie o nível de outro enzima e o procedimento experimental também nada sugere neste sentido, dado que todas as observações são provenientes de amostras aleatórias. Nestas circunstâncias as variáveis desempenham papéis idênticos. O nosso objectivo é definir uma grandeza que nos permita saber se estas duas variáveis estão ou não relacionadas ou associadas. O termo *correlação* é usualmente utilizado neste contexto. Esta associação pode ocorrer, por exemplo, da seguinte forma: uma variável tende a aumentar quando a outra também aumenta, fenómeno que se denomina *correlação positiva*, ou uma variável aumenta quando a outra diminui, tendo-se então uma *correlação negativa.*

A representação gráfica dos dados é útil para visualizarmos uma possível relação entre as variáveis, e motiva a construção de uma medida numérica da correlação presente nos dados.

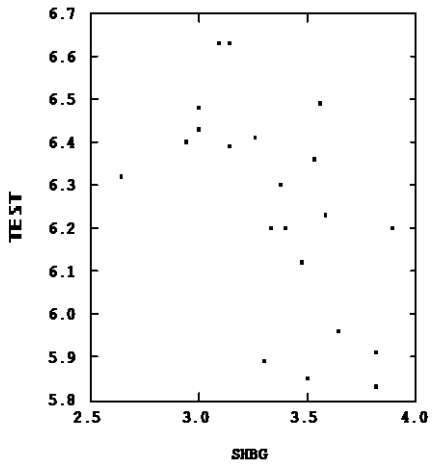

Figura 4.4

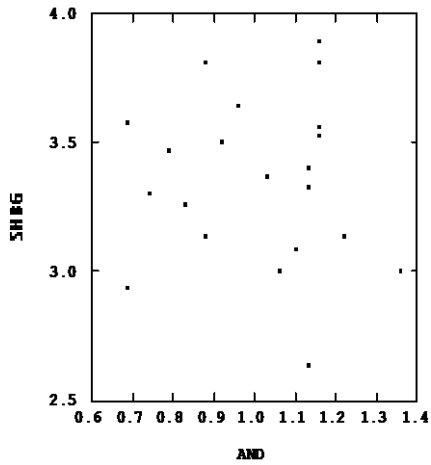

Figura 4.5

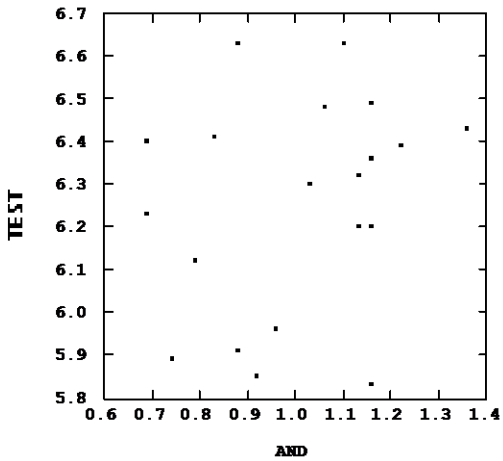

Figura-4.6

Transformemos as variáveis iniciais *x* (SHBG) e *y* (Test) nas novas variáveis definidas por:

$$x' = \frac{x - \bar{x}}{s_x} \qquad y' = \frac{y - \bar{y}}{s_y}$$

Esta transformação remove o efeito de localização e escala de cada variável, assim uma medida de associação baseada nas variáveis x' e y' é independente das unidades de medida de x e y

Vimos que uma correlação positiva entre as variáveis significa que estas tendem a aumentar ou a decrescer simultâneamente e uma correlação negativa significa que elas variam em sentidos opostos. Assim, os pontos do gráfico tendem a apresentar-se no *1º e 3º* quadrantes se houver correlação *positiva* entre as variáveis e no *2º e 4º* quadrantes quando a correlação é *negativa*.

O quadro seguinte indica o sinal de x' e y' em cada caso.

2º Quadrante	**1º Quadrante**
x' < 0	x' > 0
y' > 0	y' > 0
3º Quadrante	**4º Quadrante**
x' < 0	x' > 0
y' < 0	y' < 0

Donde se a correlação é positiva o produto x'y' tenderá a ser positivo, e pelo contrário se a correlação é negativa o produto x'y' tenderá a ser negativo. Se não há associação entre as variáveis este produto tomará valores próximos de zero. A soma dá uma medida da correlação. Valores positivos indicam correlação positiva, valores negativos indicam correlação negativa, e valores perto de zero sugerem ausência de correlação. Costuma designar-se $\frac{1}{n-1}\sum_{i=1}^{n} x_i' y_i'$ por ***r*** e chama-se *coeficiente de correlação empírico*, isto é,

$$r = \frac{\Sigma\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\Sigma\left(x_i - \bar{x}\right)^2 \Sigma\ \left(y_i - \bar{y}\right)^2}} \qquad (1)$$

Pode mostrar-se que ***r*** só toma valores entre -1 e +1. Estes últimos valores só podem ser atingidos por observações que caiam exactamente numa linha recta, com declive positivo e negativo respectivamente.O coeficinte de correlação empírico estima o valor do coeficiente de *correlação linear* da população ρ que mede a relação *linear* existente entre as variáveis X e Y. Para as variáveis do exemplo anterior temos correlações respectivamente iguais a: -0.591, -0.066 e 0.235 como os gráficos sugerem.

Exemplo 2: Recolheram-se amostras de solo do estuário do rio Tejo a 8 profundidades distintas e mediram-se os respectivos graus de humidade (gramas de água/ 100g solo) obtendo-se os seguintes resultados:

Tabela-2

Profundidade (pés)	**Humidade** (gr. água/100g solo)
0	124
5	78
10	54
15	35
20	30
25	21
30	22
35	18

Representando os dados graficamente obtém-se:

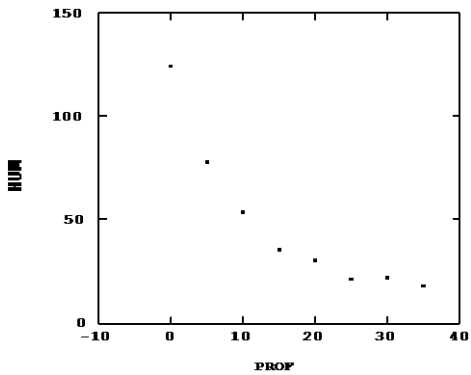

Figura – 4.7

e tem-se ainda o valor da estatística **r** = - 0.891, sugerindo uma relação linear entre as variáveis X e Y, profundidade e humidade respectivamente.

Observações: 1- Uma correlação elevada indica apenas a existência de uma associação estatística e não mais do que isso, isto é, não estabelece uma relação de causa e efeito. Quando se observa uma correlação em valor absoluto perto de 1, convém investigar se a associação entre as variáveis não é espúria.

Em Inglaterra uma publicação anticlerical mostrava claramente que o aumento de crimes nas cidades inglesas tinha crescido com o aumento do número de pastores anglicanos, durante o século XIX. Ainda que os dados fossem correctos tirar tal conclusão é um disparate. Devido à revolução industrial houve um aumento populacional importante que levou muita gente para as cidades. Portanto é razoável considerar que o número de crimes aumentou com a concentração populacional, assim como o número de padres (de médicos, advogados, polícias, etc.)

2- Como já referimos anteriormente a correlação só indica a existência de relação *linear* entre as variáveis X e Y. Por outro lado, $r \approx 0$ não significa mais do que a ausência de

um padrão linear. No exemplo que se segue, **r** = 0 e, no entanto, as variáveis X e Y estão relacionadas pela relação determinística <u>não linear</u> $X^2 + Y^2 = 4$

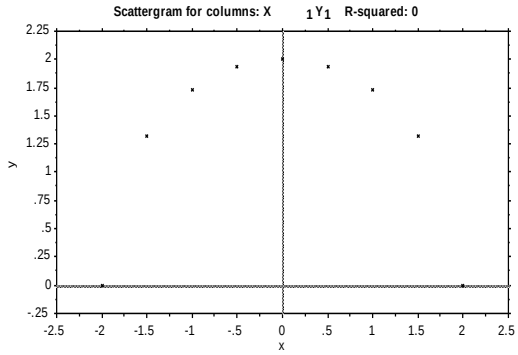

Figura-4.8

4 . 2 REGRESSÃO LINEAR SIMPLES

4.2.1 INTRODUÇÃO

Por vezes as duas variáveis *x* e y estão relacionadas de uma forma particular. A variável *x* explica de alguma forma a variável *y*. Por exemplo o preço de uma casa (y) pode depender da área desta (x), o peso de uma pessoa (y) pode depender da altura (x), no exemplo 2 da secção anterior, a humidade (y) pode depender da profundidade (x), etc. Vejamos mais alguns exemplos que nos ilustram este tipo de relações:

Exemplo 1: Proteína na gravidez

Um grupo de investigadores está interessado em saber se (e no caso afirmativo, de que modo) o nível de uma proteína se altera, nas futuras mães, ao longo da gravidez. Seleccionou-se para o estudo 19 mulheres, todas em estado diferente de gravidez (gestação), e mediu-se o nível de proteína em cada uma delas, tendo-se obtido os seguintes resultados:

Tabela-1

nível de proteína (mg ml^{-1}), **y**	**Gestação** (semanas), **x**	**nível de proteína** ($mgml^{-1}$), **y**	**Gestação** (semanas), **x**
0.38	11	0.65	27
0.58	12	0.74	28
0.51	13	0.83	29
0.38	15	0.99	30
0.58	17	0.84	31
0.67	18	1.04	33
0.84	19	0.92	34
0.56	21	1.18	35
0.78	22	0.92	36
0.86	25		

O objectivo desta experiência é averiguar como é que uma variável (nível de proteína) é afectada por uma outra variável (gestação).

Exemplo 2: Apanha automática de uvas

As vinhas estão geralmente dispostas de uma maneira muito regular, com longas filas de videiras dispostas paralelamente e separadas por um estreito arruamento. Isto permite que máquinas automáticas passem pelos arruamentos para a apanha da uva. A apanha é feita por um braço rotativo. De modo a estudar a eficiência da máquina, registou-se o nº de cachos não retirados, fazendo variar a velocidade de rotação do braço, enquanto a máquina viajava através do arruamento a uma velocidade constante. O resultado daexperiência encontra-se na

Tabela -2

prop. de cachos não apanhados **y**	**velocidade do motor**(r.p.m.), **x**
0.100	3.16
0.067	3.16
0.168	3.16
0.132	3.16
0.051	3.66
0.093	3.66
0.027	3.66
0.025	3.66
0.034	4.16
0.026	4.16
0.016	4.16
0.008	4.16
0.009	4.66
0.014	4.66
0.002	4.66
0.003	4.66

O objectivo é averiguar como é que a velocidade do motor afecta a proporção de cachos não apanhados, para poder decidir, por exemplo, qual a velocidade adequada.

Exemplo 3: O uso de radiocarbono na atribuição de datas

Pode-se estimar a idade de materiais orgânicos através da medição de um elemento radioactivo (o radiocarbono). Contudo, verificou-se através da amostragem de madeiras de idades conhecidas, que a idade por radiocarbono não é equivalente à idade verdadeira e portanto é necessário fazer-se um ajustamento. A tabela 3 dá a idade de radiocarbono de amostras de sub-fósseis de carvalhos juntamente com a informação da idade *relativa* verdadeira obtida através da informação dada pelos anéis das árvores.

Tabela-3

Idade por radiocarbono, (anos antes 1950) **y**	**Idade por anéis da árvore**, (anos numa escala flutuante) **x**
3604	0
3731	60
3714	120
3792	180
3856	240
3878	300
3883	360
4007	420
4017	480
4107	540
4125	600
4133	660
4179	720
4203	780
4304	820
4390	900
4456	960
4541	1120

O nosso objectivo é averiguar como podemos converter a data por radiocarbono de modo a encontrarmos a data verdadeira. Por exemplo, se obtivermos uma data por radiocarbono de 4300, qual é a data verdadeira?

Exemplo 4: Capacidade física de estudantes

Mediu-se a distância atingida no salto por cada um de 11 estudantes de educação física. Os resultados encontram-se na tabela 4 juntamente com medições da altura, peso do corpo, e gordura

Tabela-4

Altura (cm) x_1	**Gordura** (kg), x_2	**Peso** (kg) x_3	**Dist.salto** (cm), **y**
173.1	12.9	46.7	187.5
182.5	17.9	51.3	182.5
166.7	13.8	48.0	214.0
167.7	19.0	48.0	147.0
165.2	15.5	44.1	167.0
166.0	11.6	42.4	157.5
148.9	9.4	33.3	170.0
181.4	14.3	53.7	198.5
164.3	20.7	46.2	145.0
172.0	17.1	48.7	166.5
160.9	16.1	48.4	189.0

O objectivo do estudo é saber se (e em caso afirmativo como) é que a distância do salto é afectada pelo peso, gordura e altura do estudante.

Resumindo temos:

Experiência	**Resultado**	**Condição**
experiência clínica	nível de proteína	tempo de gestação
experiência agrícola	proporção de cachos não apanhados	velocidade do motor
experiência histórica	data por radiocarbono	data por anéis
experiência desportiva	distância de salto	altura, peso dos estudantes

Recapitulando estes exemplos podemos verificar que há algo de comum entre eles. Com efeito, em todos pretendemos averiguar como é que o resultado de uma experiência é afectado pelas condições sob as quais a experiência é efectuada. No 1º exemplo, queremos saber como é que o tempo de gestação afecta o nível de uma proteína nas futuras mães. No 2º exemplo, o conhecimento de como é que a velocidade do motor afecta a proporção de cachos não apanhados, pode permitir a selecção da velocidade adequada. No exemplo do radiocarbono, a compreensão da relação existente entre a idade por radiocarbono e a idade verdadeira (medida por um outro processo) permite-nos usar aquele método em situações futuras para datar novos elementos. Por fim a existência de uma possível relação entre a distância do salto e outras características dos estudantes, pode permitir ao professor uma selecção mais adequada de desportistas.

Continuando esta análise podemos avançar mais formalmente e dizer que temos em questão, essencialmente, dois tipos de variáveis consoante o papel que desempenham na experiência. Uma 'variável resposta' (nível da proteína no exemplo 1, proporção de cachos não apanhados no exemplo 2, idade por radiocarbono no exemplo 3 e distância de salto no exemplo 4) e uma (ou mais) 'variáveis explicativas' (tempo de gestação no exemplo 1, velocidade do motor no exemplo 2, idade por anéis no exemplo 3 e no exemplo 4 altura e dois tipos de peso). O objectivo é a descrição de um tipo de relação particular entre estes dois tipos de variáveis. [Reparemos nos gráficos que se seguem e pensemos um pouco se conseguimos descobrir alguma relação especial].

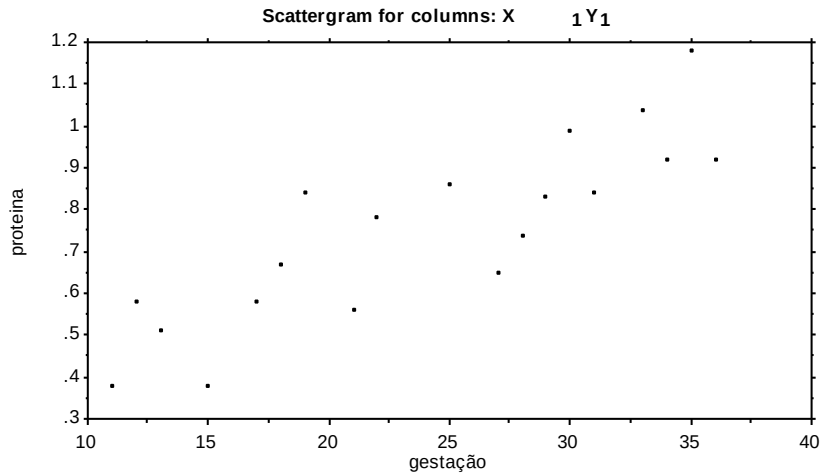

Figura 4.9 Nuvem de pontos para os dados de proteína

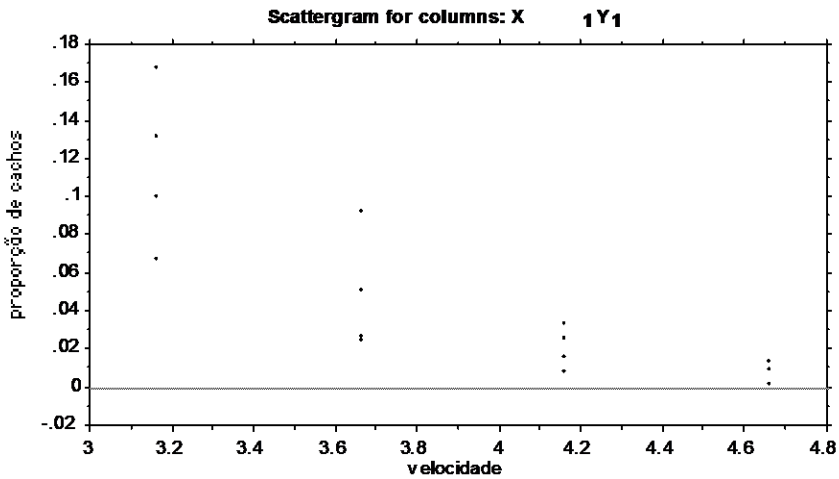

Figura 4.10 Nuvem de pontos para os dados da apanha da uva

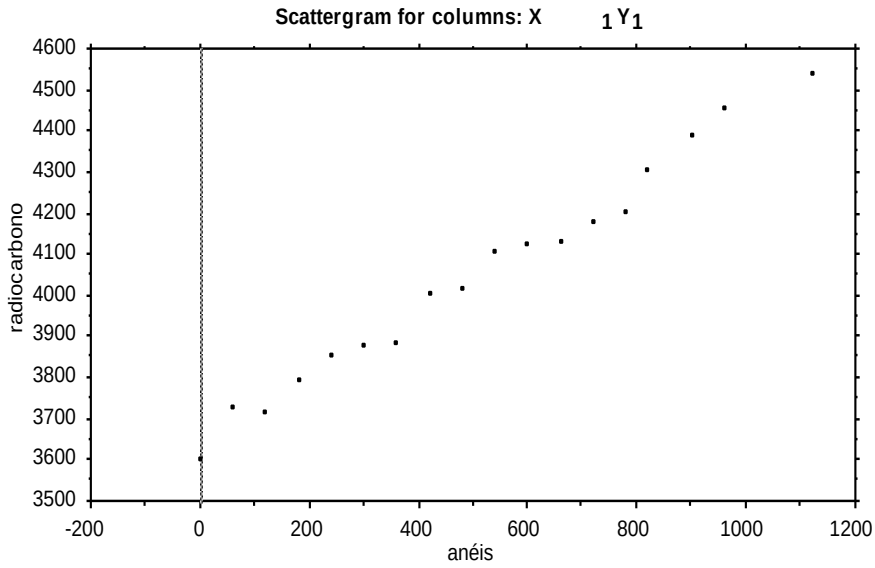

Figura 4.11 Nuvem de pontos para os dados da atribuição de idade por radiocarbono

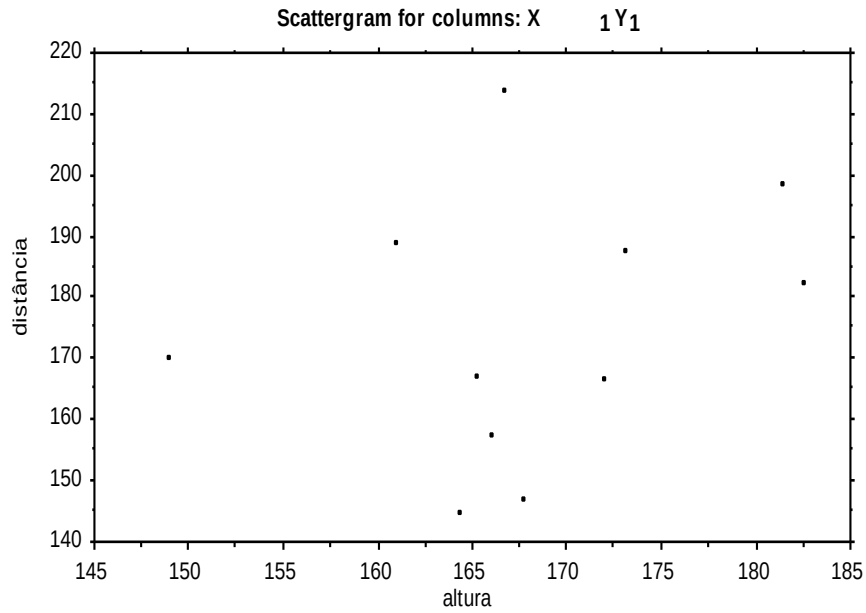

Figura 4.12 Nuvem de pontos para os dados de desporto relacionando distância e altura

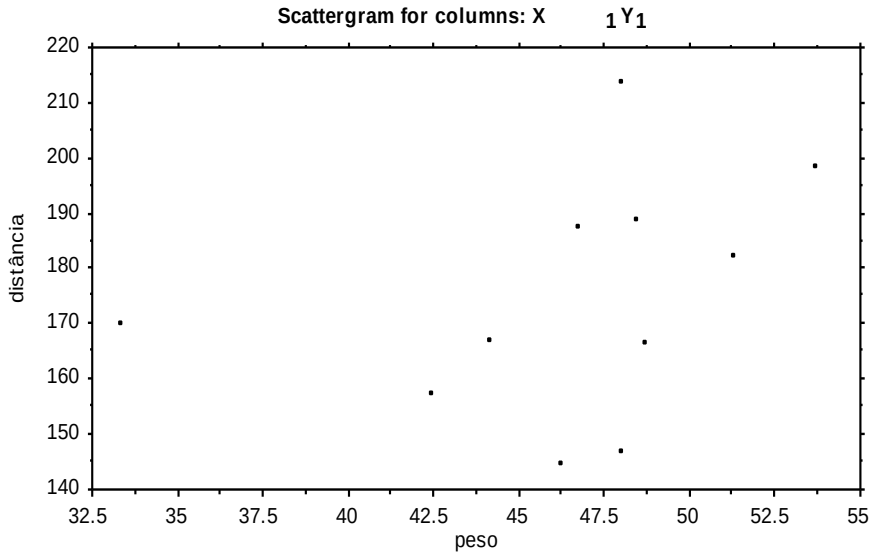

Figura 4.13 Nuvem de pontos para os dados de desporto relacionando peso e distância

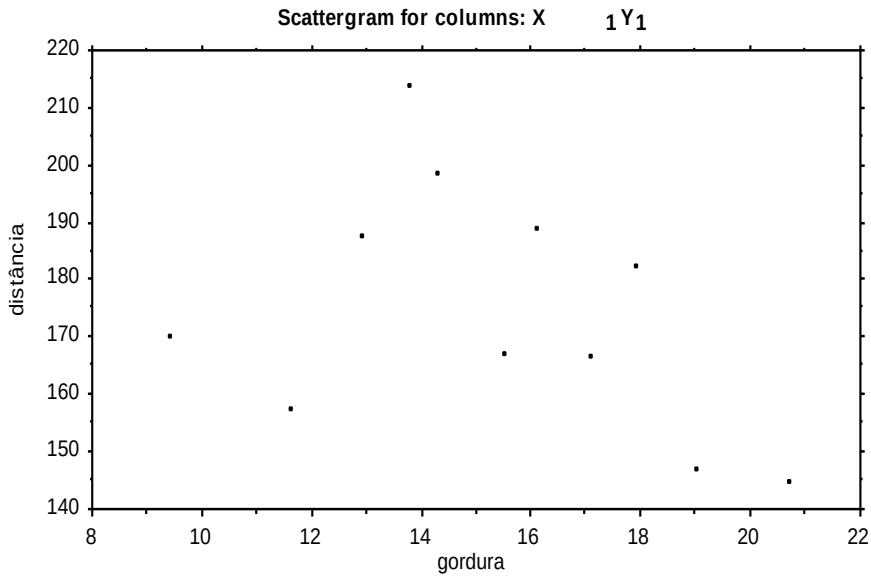

Figura 4.14 Nuvem de pontos para os dados de desporto relacionando gordura e distância

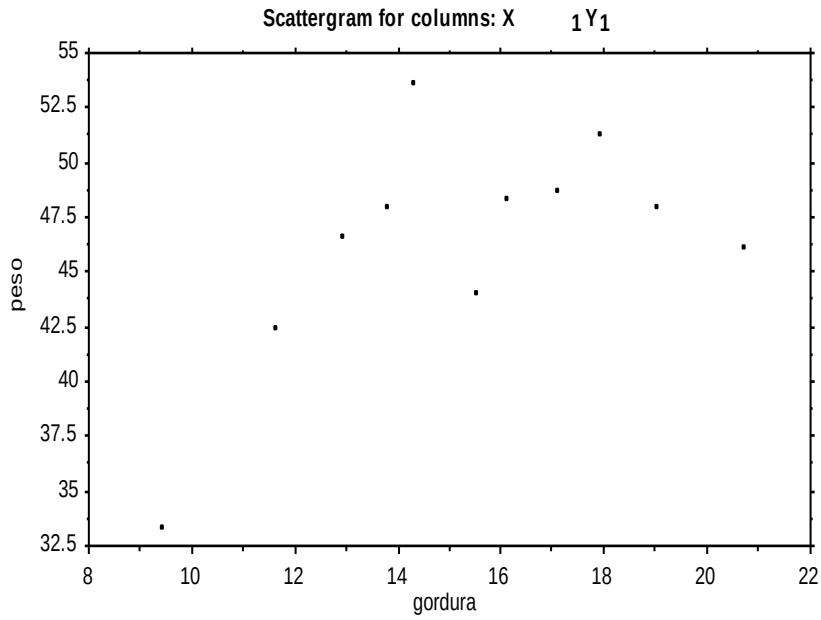

Figura 4.15 Nuvem de pontos para os dados de desporto relacionando gordura e peso

Dá-se o nome de REGRESSÃO à técnica estatística que serve para explorar a relação entre uma variável 'resposta' e uma ou mais variáveis 'explicativas'. Um *modelo* é uma descrição de um tipo de relação particular entre diferentes *variáveis.*

Um exemplo bem conhecido de um *modelo*, é aquele que descreve a relação entre a distância *s* percorrida por uma partícula e o tempo *t* que leva a percorrer, nomeadamente $s = \alpha + \beta t$, em que α é a posição inicial da particula no instante $t = 0$ e β é a velocidade média. Se α e β forem desconhecidos, basta observar *s* para dois valores distintos de *t* e resolver as equações resultantes para obter α e β. Se por qualquer razão a distância não puder ser medida exactamente, havendo um erro de medição (*e*) de natureza aleatória, então o que observamos é uma quantidade *y* (e não *s*) que podemos no entanto admitir ser tal que $y = s + e$. A relação entre *y* e *t* não é então exacta, mas apenas aproximada. Sendo agora α e β desconhecidos não podemos obter estes valores observando apenas dois valores de *t* e respectivos *y*, pois não há uma *relação funcional* exacta entre *y* e *t*, mas apenas uma *relação funcional com erro de medição* (desconhecido). Observando no entanto vários valores de *y* para diferentes valores de *t*, métodos estatísticos permitem-nos obter valores aproximados (estimativas) para os verdadeiros valores de α e β.

As situações que nos interessam são exactamente deste tipo. Os modelos que nós vamos estudar não pretendem pois descrever a realidade exactamente, mas apenas aproximadamente. O objectivo é procurar, para cada situação, os modelos mais simples que melhor descrevem a realidade. Damos o nome de **modelos de regressão** a modelos estocásticos (por oposição a determinísticos) que exprimem relações entre uma variável resposta e uma ou mais variáveis explicativas. Esta relação pode ser linear ou não linear. O modelo de regressão é *simples* se houver apenas uma variável explicativa e é *múltiplo* se houver mais do que uma variável explicativa. Nós vamos aqui iniciar apenas o estudo do modelo de regressão linear simples.

4.2.2 O MODELO DE REGRESSÃO LINEAR SIMPLES

Suponhamos que temos dados da forma (y_i, x_i), i = 1,..., n, e que queremos explorar a relação entre a variável explicativa x e a variável resposta y . Um modelo de regressão linear simples pode ser escrito na forma:

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad (1)$$

onde ε_i representa o erro associado à i-ésima observação. Admite-se que os erros têm uma média 0 e uma variância constante desconhecida.

Várias questões se podem pôr:

1º Como obter os valores de α e β (parâmetros desconhecidos)?

2º Como se pode decidir se o modelo descreve bem a realidade?

3º Como obter outro modelo que a descreva melhor?

4º Como utilizar o modelo para responder a questões sobre o problema em causa?

Um primeiro passo, informal mas extremamente útil, para tentar descobrir a relação existente entre duas variáveis é fazer uma representação gráfica. Consideremos então o exemplo 1. Façamos um gráfico em que indicamos em ordenadas os valores da variável resposta (nível da proteína) e em abcissas o valor da variável explicativa (tempo de gestação). Podemos começar por observar que quando o tempo de gestação aumenta, também aumenta o nível da proteína.

Esta relação não está, no entanto, muito bem determinada. Há grande quantidade de ruído (erro), ou variabilidade nas medições. Adaptar o modelo (1) a estes dados

significa que admitimos que o nível de proteína exacto, digamos y* é tal que $y^* = \alpha + \beta x$. No entanto nós observamos y e não y*, mas admitimos que $y = y^* + \varepsilon$. Se conseguirmos determinar valores adequados para α e β, então podemos deduzir qual a relação linear que exprime y* o nível da proteína em função do tempo de gestação. Adaptemos então várias rectas ($\tilde{y} = \tilde{\alpha} + \tilde{\beta} x$) a estes dados e para cada recta adaptada calculemos o valor da expressão:

$$SS = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 \qquad (2)$$

Analisemos os gráficos que se seguem:

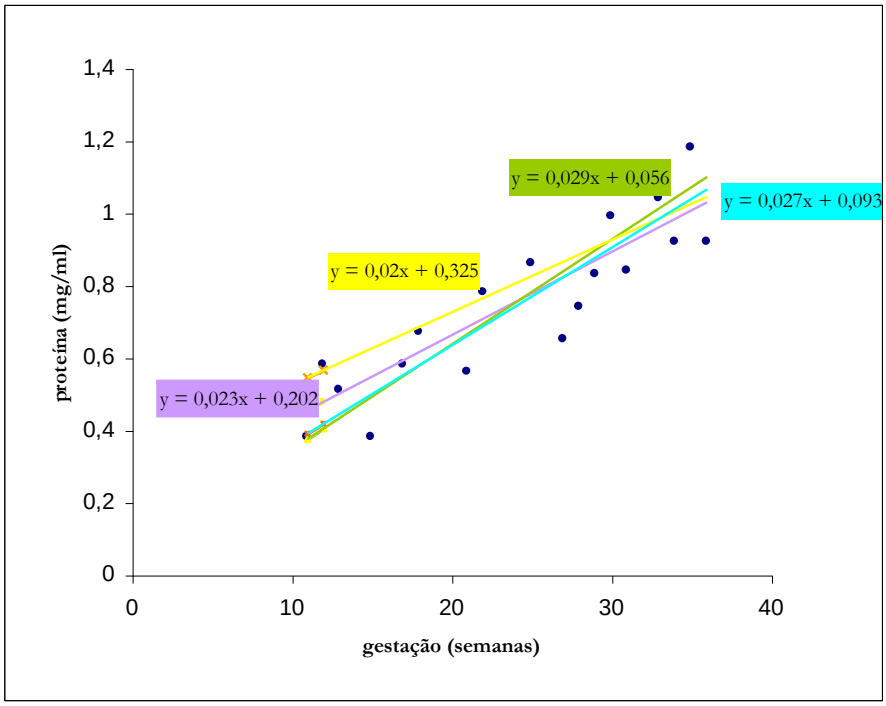

Fig. 4.16 Várias rectas adaptadas aos dados da proteína

$$\tilde{y}^{(1)} = 0.202 + 0.023x \qquad SS^{(1)} = 0.225$$

$$\tilde{y}^{(2)} = 0.056 + 0.029x \qquad SS^{(2)} = 0.27$$

$$\tilde{y}^{(3)} = 0.325 + 0.020x \qquad SS^{(3)} = 0.284$$

$$\tilde{y}^{(4)} = 0.093 + 0.027x \qquad SS^{(4)} = 0.251$$

Observamos que de todas as rectas calculadas a "melhor" parece ser aquela para a qual SS tem menor valor. Com efeito ao calcular SS estamos a calcular a soma dos quadrados dos valores estimados dos erros ε_i, isto é, a soma dos quadrados dos resíduos $e_i = y_i - \tilde{y}_i$, e assim é pois natural escolher a que tem menor SS. Sob esta perspectiva, a recta óptima será a que tiver menor valor de SS entre todas as rectas possíveis. A este método de obter a recta óptima chama-se método dos mínimos quadrados.

4.2.3 MÉTODO DE MÍNIMOS QUADRADOS

Seja

$$SS(\alpha,\beta)=\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2 \qquad (1)$$

a soma dos quadrados dos resíduos que vamos minimizar como função de α e β. Temos um problema de estimação pontual e o método que vamos utilizar é, como sugerido, o *Método de Mínimos Quadrados.*

Definição 1: Sejam $(x_i, y_i), i = 1,\ldots,n$, n pares de observações satisfazendo a condição:

I) para cada x_i, valor de uma variável <u>não</u> aleatória, as v.a.'s y_i são iguais a $y_i = \alpha + \beta x_i + \varepsilon_i$, onde ε_i são v.a.'s com $E(\varepsilon_i)=0, Var(\varepsilon_i)=\sigma^2$. Então, aos valores de α e β que minimizam a soma de quadrados (1)

$$SS(\alpha,\beta)=\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2$$

chamam-se estimadores de mínimos quadrados de α e β, e a este método de estimação chama-se *Método de Mínimos Quadrados.*

Para obter estes estimadores vamos derivar a soma de quadrados (1) em ordem a cada um dos parâmetros, obtendo as seguintes equações

$$\begin{cases} \dfrac{\partial SS(\alpha,\beta)}{\partial\alpha}=\displaystyle\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)=0 \\ \dfrac{\partial SS(\alpha,\beta)}{\partial\beta}=\displaystyle\sum_{i=1}^{n}x_i(y_i-\alpha-\beta x_i)=0 \end{cases} \qquad (2)$$

a que se chamam *Equações Normais.* E resolvendo o sistema de equações anteriores em ordem a α e β obtém-se

$$\begin{cases} \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \\ \hat{\beta} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{cases} \tag{3}$$

Efectuando alguns cálculos, obtemos ainda uma expressão simplificada para $\hat{\beta}$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4}$$

A recta dos mínimos quadrados é então:

$$\hat{y} = \hat{\alpha} + \hat{\beta}x \tag{5}$$

Às diferenças entre os valores observados e os valores adaptados, $e_i = y_i - \hat{y}_i,\ i = 1,\ldots,n$, dá-se o nome de resíduos e à quantidade $SS(\hat{\alpha},\hat{\beta}) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i{}^2$, soma dos quadrados dos resíduos e costuma designar-se mais vulgarmente por SS_e.

Aplicando este método aos exemplos propostos:

Exemplo 1:

$$\bar{x} = 24.0,\ \bar{y} = 0.75,\ \hat{\alpha} = 0.2018,\ \hat{\beta} = 0.02284,\ \text{então}$$
$$\hat{y} = 0.2018 + 0.02284x \tag{6}$$

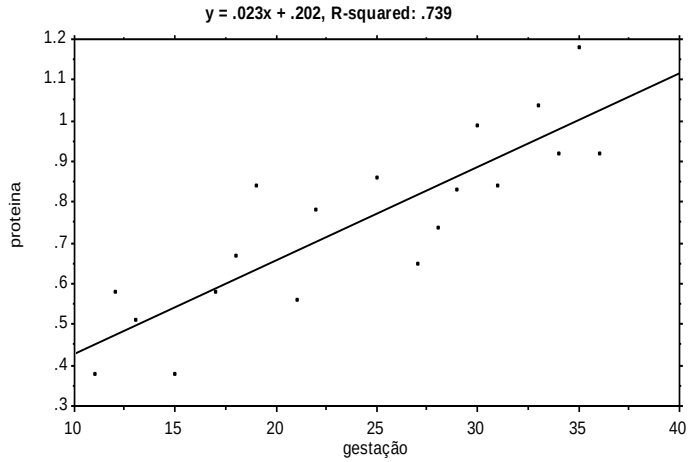

Fig. 4.17 Recta dos minimos quadrados para os dados de gestação.

Se quisermos saber, por exemplo, qual o valor esperado do nível da proteína em uma mulher com 24 semanas de gestação, basta substituir o valor de x por 24 em (6) e obtemos 0.75. Podemos perguntar: Qual a confiança que temos nesse valor? Métodos estatísticos adequados permitem-nos responder a essa e outras questões relevantes relativamente ao modelo. Neste momento a única coisa que podemos fazer é obter a "melhor" recta. Uma análise apropriada dos resíduos também nos permite averiguar da validade da hipótese da linearidade.

Como já foi dito, os resíduos são as diferenças entre os valores observados da variável resposta e os correspondentes valores sobre a recta de regressão. São, estimativas dos erros ε_i associados a cada observação.

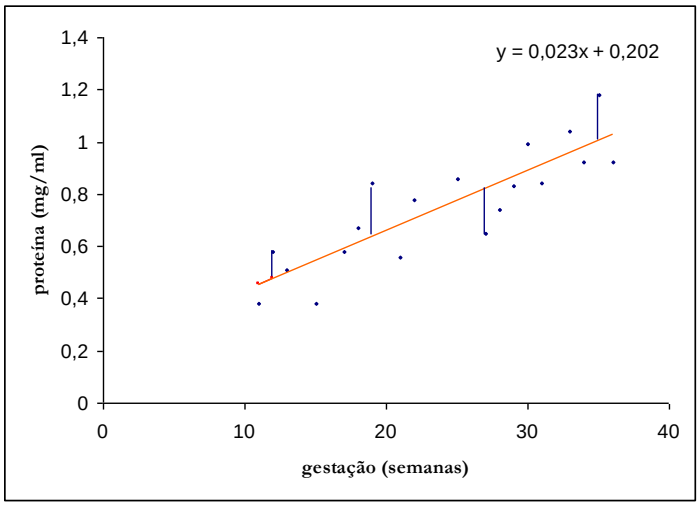

Fig 4.18. Recta dos mínimos quadrados e resíduos para os dados de gestação

Os segmentos de rectas verticais que ligam cada ponto à recta de regressão adaptada representam os resíduos (alguns). Pela observação do gráfico o que se pode concluir?

Exemplo 2

$\bar{x} = 3.91, \bar{y} = 0.048, \hat{\alpha} = 0.328, \hat{\beta} = -0.071,$ e a recta de regressão é

$\hat{y} = 0.328 - 0.071x$ (7)

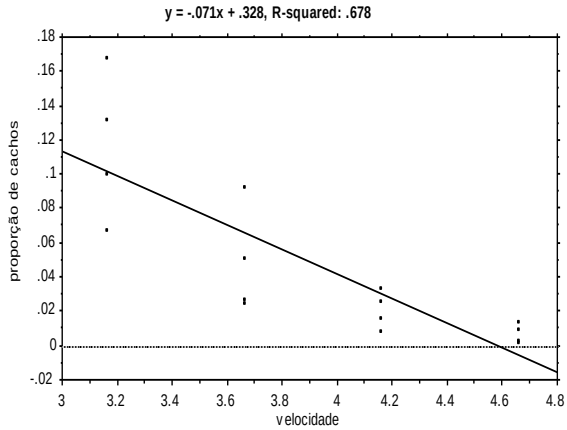

Fig 4.19. Recta dos mínimos quadrados para o exemplo da apanha da uva

Note-se agora que a recta já não parece adaptar-se tão bem. Poderemos pôr dúvidas inclusivamente sobre a linearidade da relação entre as variáveis em questão. Aliás a utilização desta recta ia-nos sugerir para uma velocidade de 4.6 uma proporção negativa de cachos não apanhados! Ora isto é manifestamente impossível. Consideremos a seguinte transformação da proporção: $h(y) = \ln \frac{y}{1-y}$ façamos gora o estudo considerando como variável resposta z = h(y) e variável explicativa a velocidade.

Tabela -5

dados transformados da apanha de uvas

ln(y/(1-y))	velocidade
-2.197	3.16
-2.634	3.16
-1.6	3.16
-1.883	3.16
-2.924	3.66
-2.278	3.66
-3.585	3.66
-3.664	3.66
-3.347	4.16
-3.623	4.16
-4.119	4.16
-4.82	4.16
-4.701	4.66
-4.225	4.66
-6.213	4.66
-5.806	4.66

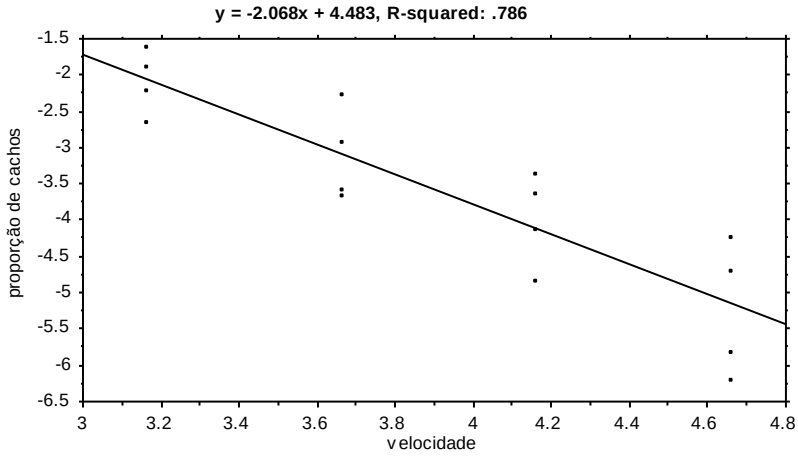

Fig 4.20. Recta de mínimos quadrados para os dados transformados da apanha de uvas

A recta $\hat{z} = \hat{\alpha}_z + \hat{\beta}_z x$=4.483-2.068x, adapta-se agora bastante bem. Note-se que

$$\hat{z} = h(\hat{y}) = \ln \frac{\hat{y}}{1-\hat{y}} \Leftrightarrow \frac{\hat{y}}{1-\hat{y}} = e^{\hat{z}} \Leftrightarrow \hat{y} = \frac{e^{\hat{z}}}{1+e^{\hat{z}}}$$

finalmente a relação entre x e y é da forma

$$\hat{y} = \frac{e^{4.483-2.068x}}{1+e^{4.483-2.068x}} \qquad (8)$$

Da relação (8) infere-se para uma velocidade de 4.66 uma proporção de cachos não apanhados igual a 0.0057.

Exemplo 3:

$$\bar{x} = 514.44,\ \bar{y} = 4051.11,\ \hat{\alpha} = 3636,\ \hat{\beta} = 0.808,\ \text{e}\quad \hat{y} = 3636 + 0.808x \qquad (9)$$

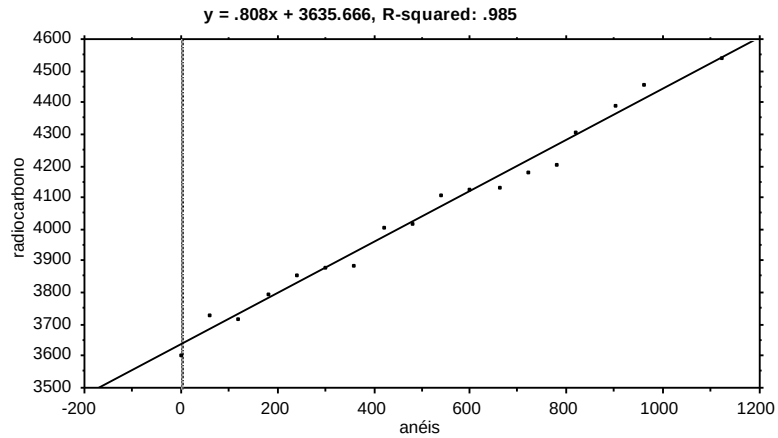

Fig 4.21. Recta dos mínimos quadrados relativa aos dados de atribuição de idade por radiocarbono.

Vemos como se adapta tão bem uma recta. Lembremo-nos que o objectivo aqui era o de encontrar a relação entre a atribuição de idade por radiocarbono e idade real, para poder, com base na idade obtida pelo método de radiocarbono, inferir a idade real. Este é um problema de "regressão inversa" ou "calibração". Suponhamos então que observávamos uma data por radiocarbono de 4300 num objecto de interesse. Usando a relação obtida obteríamos uma idade real, na escala flutuante, de 822.

O exemplo 4 difere dos apresentados até agora pois temos mais do que uma variável explicativa. O método adequado para tratar de problemas desta natureza é utilizar a técnica de regressão múltipla. Embora tenhamos apresentado os gráficos 4.12, 4.13, 4.14 isto não significa que seja adequada uma análise separada da relação de y com cada uma das variáveis explicativas. Apenas como exercício didático podemos fazer essa análise, mas não com o propósito de tirar quaisquer conclusões sobre as possíveis relações existentes.

Uma vez apresentado este método tem interesse indagar sobre a qualidade dos estimadores que obtivemos. Juntemos às hipóteses feitas sobre o modelo linear sintetizadas na condição da definição do método de MQ, ainda outra

II) As v.a.'s ε_i são <u>não</u> correlacionadas duas a duas, isto é, $Cov\left(\varepsilon_i, \varepsilon_j\right) = 0, i \neq j, i, j = 1,..., n$.

As boas propriedades destes estimadores são enunciadas no seguinte resultado que apresentamos sem demonstração.

Teorema de Gauss-Markov: Consideremos o modelo linear definido pelas condições I) e II). Então, os estimadores de mínimos quadrados de α e β dados pelas equações (3) são *lineares centrados de variância mínima* (BLUE- best linear unbiased estimator).

O método de mínimos quadrados não dá um estimador do parâmetro σ^2 mas um estimador deste parâmetro baseado nos estimadores de MQ de α e β é

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}\, x_i\right)^2 = \frac{SSe}{n-2} = MSe \qquad (10)$$

Observações: 1. Voltemos aos dados do exemplo 2 da secção 4.1.3 (Tabela-2), e consideremos o modelo de regressão linear *simples* que designaremos por modelo 1

$y = \alpha + \beta x + \varepsilon$ *modelo 1*

Adaptemos a estes dados a recta de mínimos quadrados:

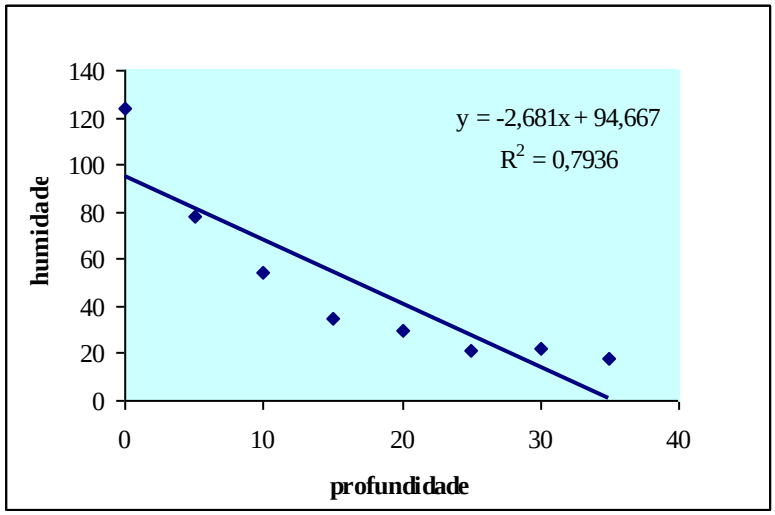

Figura-4.22

Note-se que a figura 4.7 já sugeria uma certa curvatura na relação entre X e Y, o que é mais patente depois da adaptação da recta de mínimos quadrados. A figura seguinte mostra que os resíduos são predominantemente positivos para valores pequenos de X, negativos para valores intermédios de X e de novo positivos para valores grandes de X.

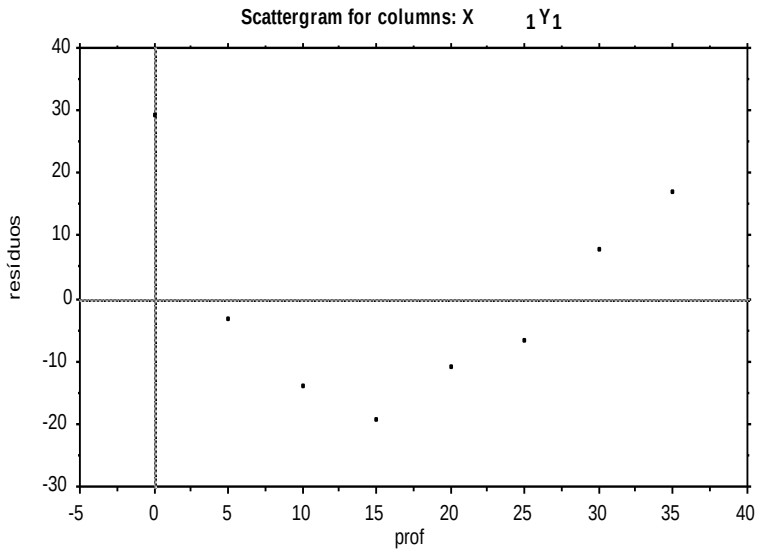

Figura-4.23- Gráfico dos resíduos do *modelo 1*

A relação linear existente entre estas duas variáveis deve ser de tipo mais geral do que a regressão linear simples estudada até aqui.

Tal como as figuras anteriores sugerem tentemos adaptar a estes dados uma curva do 2º grau, isto é, consideremos o modelo com as variáveis explicativas X e X^2 da forma:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon \qquad \textit{modelo 2}$$

tendo-se ainda a variável resposta Y como função *linear* nos parâmetros β_i i = 0, 1, 2. Este é um exemplo de modelo de regressão linear mais geral (regressão polinomial).

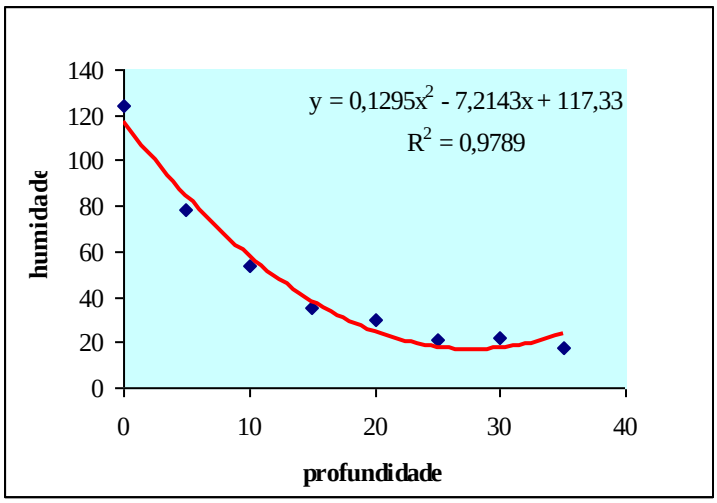

Figura-4.24

A representação gráfica dos resíduos versus a variável profundidade mostra que se distribuem agora aleatóriamente em torno do ponto zero e numa banda horizontal.

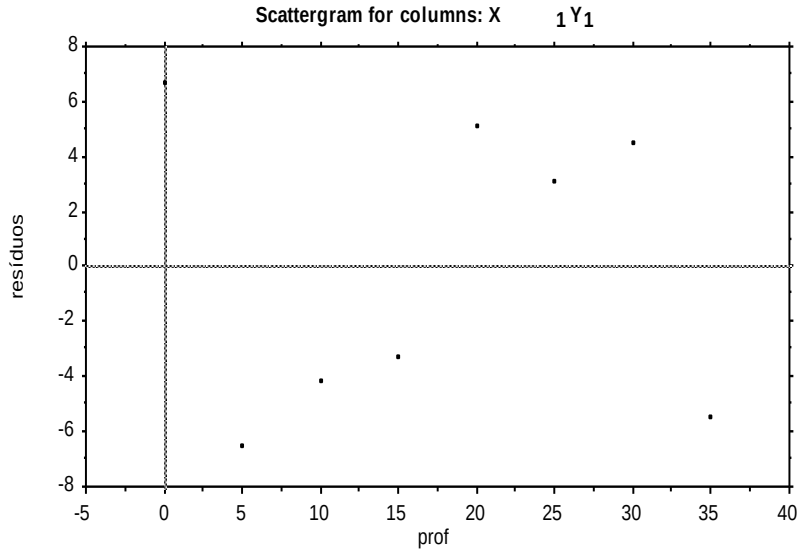

Figura-4.25- Gráfico dos resíduos do *modelo 2*

O modelo linear que estudámos é um caso particular do modelo linear mais geral

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} + \varepsilon_i \qquad (11)$$

Onde $\beta_0, \beta_1, ..., \beta_p$, $p \geq 1$ são parâmetros desconhecidos e

ε_i v.a.'s com $E(\varepsilon_i) = 0$ e $Var(\varepsilon_i) = \sigma^2$. No caso de p> 1 temos mais do que uma variável explicativa, e ao modelo (5) chama-se um modelo de ***regressão múltipla.***

A denominação de modelo *linear* deve-se ao facto da parte determinística do modelo ser uma função linear nos parâmetros $\beta_0, \beta_1, ..., \beta_p$, $p \geq 1$. O modelo

$y = \alpha + \beta e^x + \varepsilon$, $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$ que relaciona a variável explicativa x com a v.a. (resposta) y, não é linear como função da variável explicativa, mas é linear nos parâmetros, logo é um modelo linear. As variáveis explicativas podem ser potências de uma variável, $y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_p x^p + \varepsilon$, a este modelo de regressão múltipla chama-se *regressão polinomial.*

2. Em muitas situações reais a componente determinística do modelo não é linear. Vejamos por exemplo:

i) Certas populações de animais e plantas tendem a crescer exponencialmente. Se Y representa a dimensão da população no instante t, podemos utilizar o modelo

$$E(Y_t) = \alpha_0 e^{\alpha_1 t} \qquad (12)$$

embora esta expressão não seja linear nos parâmetros podemos linearizá-la. Aplicando logaritmos a ambos os membros da igualdade obtemos o modelo

$$\ln Y_t = \ln \alpha_0 + \alpha_1 t + \varepsilon \qquad (13)$$

cuja parte determinística já é linear nos parâmetros que agora se podem estimar pelo método dos MQ $(\alpha = \ln \alpha_0 \; e \; \beta = \alpha_1)$.

ii) Outro modelo que ocorre nas ciências biológicas é aquele que relaciona o peso (ou volume) de um organismo com alguma medida linear, como o comprimento (ou peso). Se *P* é o peso e *c* o comprimento, o modelo

$$E(P) = \alpha_0 c^{\alpha_1} \qquad (14)$$

é muitas vezes utilizado (equação alométrica). Se quisermos relacionar o peso de organismos seleccionados aleatoriamente para comprimentos fixos observados, podemos aplicar logaritmos às observações e obter o modelo

$$\ln P = \ln \alpha_0 + \alpha_1 \ln c + \varepsilon \qquad (15)$$

que é do tipo $\ln P = \alpha + \beta x + \varepsilon$ com $\alpha = \ln \alpha_0$, $\beta = \alpha_1$ e $x = \ln c$.

4.2.4 TESTES DE HIPÓTESES

Outro problema com interesse é o dos testes de hipóteses. Consideremos a seguinte situação:

Exemplo 1: Para estudar o efeito da temperatura (x) na velocidade (y) de certa reacção química, foram efectuadas 8 experiências laboratoriais, que conduziram à seguinte relação linear:

$$\hat{y} = -2.14 + 0.79x \quad \text{com} \quad \hat{\sigma}(\hat{\alpha}) = 2.91 \ \text{ e } \ \hat{\sigma}(\hat{\beta}) = 0.296$$

Há evidência suficiente nos dados de que o aumento de temperatura faça com que a reacção estudada se processe mais rapidamente? Justifique. Tome $\alpha = 0.05$.

O que pretendemos saber com esta pergunta pode ser respondido através de um teste das hipóteses: $H_0 : \beta = 0$ versus $H_1 : \beta > 0$.

Exemplo 2: Efectuou-se um estudo em 9 países africanos em vias de desenvolvimento, para averiguar da possível relação entre o número de habitantes por médico e a esperança de vida (em anos), tendo-se obtido os seguintes resultados:

Tabela-6

Nº hab./médico	**E.média vida**(anos)
1 907	63.00
26 447	48.30
815	52.70
6 411	53.50
10 136	49.05
7 306	38.30
22 291	50.00
18 657	47.35
7 378	52.50

Há evidência suficiente nos dados que mostre que o número de habitantes/médico está linearmente relacionado com a esperança média de vida? Como responder agora a esta questão? O modelo que supostamente se adapta aos dados é:

Esp. Média Vida=α+β (nº hab./médico)+ε (1)

Será que a variável nº hab./médico (x), tem uma contribuição significativa na explicação da variável resposta Esp. Média de Vida (y)? Com base na amostra observada vamos construir um teste para a hipótese $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$.

Exemplos como estes ilustram bem as situações em que pode ter interesse construir um teste para uma hipótese sobre o parâmetro β.

Noutros casos o parâmetro de interesse pode ser a ordenada na origem α. Retomemos então o modelo linear fazendo agora uma hipótese suplementar sobre a distribuição das v.a.'s ε_i, isto é,

$$y_i = \alpha + \beta x_i + \varepsilon_i \text{ com } \varepsilon_i \cap Gau(0,\sigma), \text{i} = 1,\ldots,\text{n} \text{ independentes} \quad (2)$$

Estimando os parâmetros α e β obtêm-se os valores

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i = \bar{y} + \hat{\beta}\left(x_i - \bar{x}\right) \quad i = 1,\ldots,n \quad (3)$$

Desta relação conclui-se que

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right) = \sum_{i=1}^{n} \left(y_i - \bar{y}\right) - \hat{\beta} \sum_{i=1}^{n} \left(x_i - \bar{x}\right) = 0 \quad (4)$$

Por outro lado pode provar-se que:

$$\hat{\beta} \cap Gau\left(\beta, \sigma \Big/ \sqrt{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}\right) \quad (5)$$

E que o estimador de σ^2 tem distribuição qui-quadrado,

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \cap \chi^2_{n-2} \quad (6)$$

Além disso as variáveis (5) e (6) são independentes. Logo

$$\frac{\hat{\beta}-\beta}{\hat{\sigma}}\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\ \cap t_{(n-2)} \tag{7}$$

Então o Intervalo de Confiança de nível (1-α) para o parâmetro β é o seguinte:

$$\left(\hat{\beta}-t_{1-\frac{\alpha}{2};n-2}\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}},\hat{\beta}+t_{1-\frac{\alpha}{2};n-2}\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}}\right) \tag{8}$$

No caso do exemplo anterior obtém-se as seguintes estimativas dos parâmetros e os respectivos intervalos de confiança.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	**95% Confidence Interval for B**	
	B	Std. Error	Beta	Lower Bound	Upper Bound
(Constant)	53,505	3,569		45,066	61,944
habitantesmedico	-,0003	,0003	-,369	-,001	,0003

Dependent Variable: esperança de vida

Conclusão: O I.C. de 95% para esta amostra é (-0,001, 0,0003), contém o zero levando à NÃO rejeição da hipótese nula ao nível de 5%.

NOTA:

Pode mostrar-se facilmente que o estimador do parâmetro β está relacionado com **r** da seguinte forma

$$r=\hat{\beta}\sqrt{\frac{\Sigma\left(x_i-\bar{x}\right)^2}{\Sigma\left(y_i-\bar{y}\right)^2}} \tag{9}$$

Assim, $\hat{\beta}=0$ implica **r**=0 e vice-versa e consequentemente a hipótese nula $H_0:\beta=0$ é equivalente a $H_0:\rho=0$. No entanto, o declive dá-nos informação adicional na quantidade de aumento (decréscimo) em y por cada unidade de aumento em x.

4.2.5 Análise de resíduos e Observações influentes

Depois de adaptarmos a retcta de MQ e antes de fazermos testes nos parâmetros da regressão devemos fazer uma representação gráfica dos resíduos para ver se alguma das hipóteses do modelo linear foi seriamente violada.

Audiências de Programas de Televisão

O sucesso de um programa de uma certa televisão comercial é em parte determinado por um sistema de classificação que indica a capacidade do programa atrair e manter os espectadores atentos. O director de programas está preocupado com a audiência dos noticiários e pretende encontrar os factores que a influenciam. Além das variáveis (factores) óbvias tais como o formato, efeitos especiais, apresentador/a, foi sugerido que poderia existir um efeito de "arrastamento" do programa exibido imediatamente antes das notícias. A classificação do noticiário dependia em parte da classificação do programa anterior, isto é, do programa" indutor". Para quantificar este efeito, foi observada uma amostra aleatória das classificações precedentes para várias regiões e em vários períodos de tempo ao longo dos 2 últimos anos. Os dados consistem de observações na variável *y,* classificação do noticiário, e na variável *x*, que representa as classificações do programa "indutor".

Tabela-1

x	*y*	*x*	*y*
2,50	3,80	5,50	4,35
2,70	4,10	5,70	4,15
2,90	5,80	5,90	4,85
3,10	4,80	6,10	6,20
3,30	5,70	6,30	3,80
3,50	4,40	6,50	7,00
3,70	4,80	6,70	5,40
3,90	3,60	6,90	6,10
4,10	5,50	7,10	6,50
4,30	4,15	7,30	6,10
4,50	5,80	7,50	4,75
4,70	3,80	2,50	1,00
4,90	4,75	2,70	1,20
5,10	3,90	7,30	9,50
5,30	6,20	7,50	9,00

Ajustando aos dados um modelo linear obtém-se:

$\hat{\alpha} = \mathbf{1.707}$ e $\hat{\beta} = \mathbf{0.665}$; $\hat{\sigma} = \sqrt{MSE} = \mathbf{1.402}$ e coeficiente de determinação R^2=***0.396***

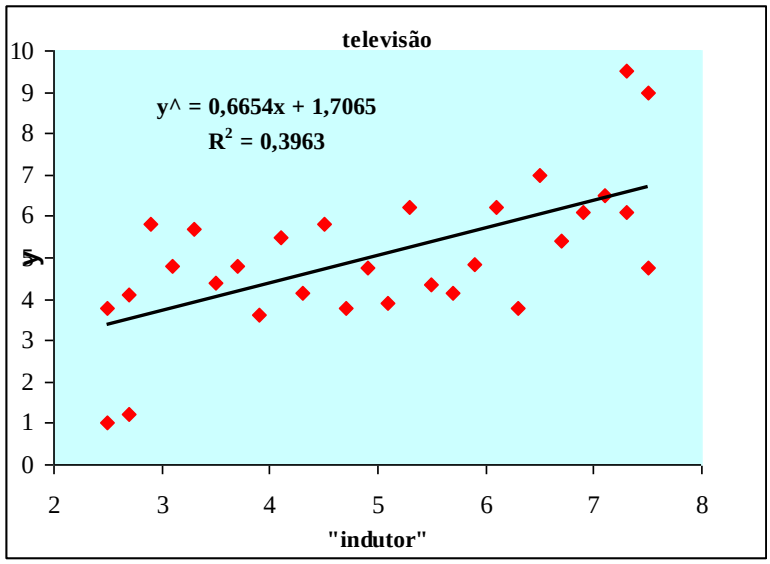

Figura- 1

Ao observar o diagrama de dispersão verificamos que existem 4 observações bastante afastadas das restantes. Representemos agora graficamente os resíduos:

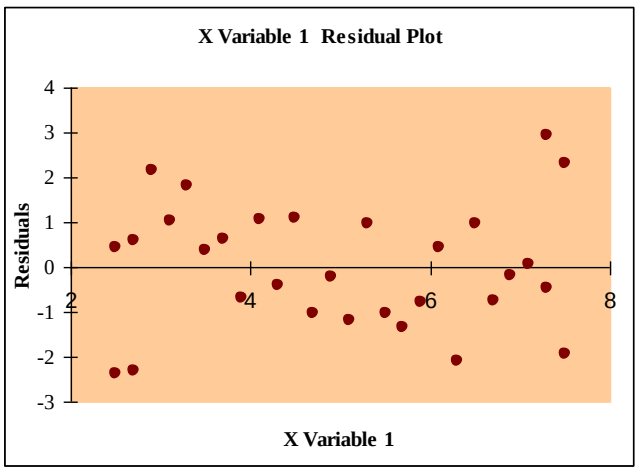

Figura- 2

A Figura-2 mostra que para valores intermédios da variável x ("indutor") os resíduos parecem distribuir-se aleatoriamente em torno da recta *e=0*; no entanto, para valores pequenos de x a maior parte dos resíduos são positivos indicando que o modelo subestima as respostas, mas há dois grandes resíduos negativos sugerindo sobrestimação pelo modelo, a situação inverte-se para valores grandes de x.

Olhando para a Figura-1 verificamos que estes 4 pontos inflacionaram o declive da recta de regressão. Parece que se retirássemos estas 4 observações a recta de MQ deveria ter um declive perto de zero, indicando que provavelmente a variável x não afecta a resposta. Estes pontos vão ser encarados como outliers e devem ser investigados.

Será que estas observações influenciam o modelo?

Analisemos os dados sem estas observações (27, 28, 29 e 30) e vejamos se as estimativas dos coeficientes da recta de MQ vêm muito alteradas.

Sem estas 4 observações obtém-se:

$\hat{\alpha} = \mathbf{3.713}$ $\hat{\beta} = \mathbf{0.260}$ e $R^2 = \mathbf{0,161}$

Para melhor podermos comparar os resultados vamos escrever a tabela:

Quadro Resumo

	Amostra completa	*Amostra reduzida*
$\hat{\beta}$	**0,665**	**,260**
$\hat{\alpha}$	**1,707**	**3,713**
R^2	**,396**	**,161**
s	**1,402**	**,925**
n	**30**	**26**

Tal como esperávamos o declive da recta diminuiu consideravelmente, houve uma redução de cerca de 61%.

O que nos leva a concluir que as observações com resíduos grandes devem ser sujeitas a investigação, pois podem indicar erros de digitalização ou também evidenciar a existência de algum comportamento dos dados que pode passar despercebido numa primeira análise do problema em estudo.

4.3* PREDIÇÃO

Temos ainda a considerar o problema da predição. Assim, suponhamos que se obtiveram os seguintes pares de observações $(x_1, y_1),...,(x_n, y_n)$ com base nas quais desejamos predizer uma observação futura Y_0 para um determinado valor da variável controlada x_0. Note-se que Y_0 é uma v.a. e não um parâmetro, não se trata pois de um problema de inferência sobre parâmetros de uma distribuição como até aqui temos feito. No modelo linear que estamos a estudar assumimos que $Y_0 = \alpha + \beta x_0 + \varepsilon, \quad \varepsilon \cap Gau(0, \sigma)$, i.e., a distribuição de Y_0 está centrada no valor

médio desta, $E(Y_0)=\alpha+\beta x_0$, e assim é natural usar $\hat{Y}_0=\hat{\alpha}+\hat{\beta}x_0$ como predictor de Y_0 e que é simultaneamente o estimador de $E(Y_0)$.

Voltemos aos exemplos apresentados no início do capítulo e suponhamos, por exemplo, que pretendíamos predizer o nível de proteína Y_0 que uma futura mãe deve ter ao fim de x_0=24 semanas de gestação, o que nós pretendemos é predizer um valor particular da v.a. Y_0. Então usando (6) da secção 4.2.2, temos que $\hat{Y}_0=0.2018+0.0228\times 24=0.75$, ao tomar este valor como predictor de Y_0 estamos a cometer um erro de predição que é dado pela diferença $\text{Erro}_P=Y_0-\hat{Y}_0$. Este erro é uma v.a. cuja distribuição é ainda gaussiana de parâmetros

$$E(Y_0-\hat{Y}_0)=E(Y_0)-E(\hat{Y}_0)=0 \quad e$$

$$\begin{aligned}
Var(Y_0-\hat{Y}_0) &= E(\hat{Y}_0-Y_0)^2=E[\hat{Y}_0-\mu(x_0)+\mu(x_0)-Y_0]^2 \\
&= E[\hat{Y}_0-\mu(x_0)]^2+E[Y_0-\mu(x_0)]^2-2Cov(Y_0,\hat{Y}_0)= \\
&= Var(\hat{Y}_0)+Var(Y_0)-0 \\
&= Var(\hat{\alpha}+\hat{\beta}x_0)+\sigma^2= \\
&\qquad\qquad (1) \\
&\sigma^2\left\{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}+1\right\}
\end{aligned}$$

Note-se ainda que as v.a.'s $\hat{Y}_0$ e Y_0 são independentes, por isso $Cov(Y_0,\hat{Y}_0)=0$ uma vez que o predictor $\hat{Y}_0$ só depende das observações $Y_1,...,Y_n$ *independentes de* Y_0, através de $\hat{\alpha}$ e $\hat{\beta}$. Além disso, poderíamos provar que a v.a. $Y_0-\hat{Y}_0$ tem distribuição gaussiana por ser uma combinação linear de gaussianas independentes, de valor médio zero e variância dada por (1) o que nos leva a concluir que a v.a.

$$T(Y_0,\hat{Y}_0)=\frac{\hat{Y}_0-Y_0}{\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}}}\sqrt{\frac{n-2}{n}}\cap t_{n-2} \; com \;\; s_{xx}=\sum(x_i-\bar{x})^2 \qquad (2)$$

Esta v.a. pode ser utilizada na construção de um intervalo de predição para Y_0 (Y_0 é uma v.a. e não um parâmetro, mas o princípio para a construção deste intervalo de predição é o mesmo do utilizado nos I.C. para um parâmetro) de nível (1-α). Vamos optar pelo intervalo de amplitude média mínima o que corresponde a considerar

$$P\left(-t_{1-\alpha/2;n-2} < T < t_{1-\alpha/2;n-2}\right) = 1-\alpha \qquad (3)$$

O intervalo de predição será então,

$$\left\{\hat{Y}_0 - t_{1-\alpha/2}\,\hat{\sigma}\sqrt{\frac{n}{n-2}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right)}, \hat{Y}_0 + t_{1-\alpha/2}\,\hat{\sigma}\sqrt{\frac{n}{n-2}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right)}\right\} \quad (4)$$

Nota: A variância do erro de predição $Y_0 - \hat{Y}_0$ é tanto menor quanto mais próximo de $\bar{x}$ estiver o valor x_0 não observado da variável explicativa, para o qual queremos fazer a predição. Logo, a precisão do predictor aumentará com a proximidade de x_0 a $\bar{x}$, e o mesmo acontece com a precisão do intervalo de predição (em termos de amplitude) como seria de esperar. É portanto arriscado fazer previsões para um futuro longínquo ou relativamente a um passado remoto, para o qual o modelo até pode não ser o "correcto".

NOTA: As secções que têm asterisco não foram dadas este ano.