5.2. Regressão linear

Até agora, vimos técnicas estatísticas (gráficos e medidas) que nos permitem estudar uma variável de cada vez, isoladamente. Contudo, muitas vezes é mais interessante estudar as relações que existem entre diversas variáveis. Por exemplo, é óbvio que o peso de uma pessoa deve estar relacionado com sua altura; quanto mais alta a pessoa, mais pesada ela deve ser. Veremos agora algumas técnicas para medir quão forte é a relação entre duas variáveis (*coeficiente de correlação*), e para criar modelos que descrevam esta relação (*modelos de regressão*)

5.2.1. Tipos de relações entre variáveis

Se analisamos dados sobre duas ou mais variáveis, há quatro perguntas que podemos fazer:

1. Existe alguma relação entre estas variáveis?
2. Que tipo de relação é esta?
3. Quão 'forte' é esta relação?
4. Que modelo pode ser usado para descrever esta relação?

O primeiro passo para responder a estas perguntas é fazer o *diagrama de dispersão* dos dados das duas variáveis – o gráfico que mostra os valores observados de uma variável no eixo horizontal, e os da outra no eixo vertical, e representa cada par de observações por um ponto no plano cartesiano. A forma da *nuvem* formada por estes pontos indicará se há ou não relação estatística entre as variáveis, e também o tipo de relação.

As Figs. **1** a **4** mostram alguns tipos possíveis de relação. A Fig. **1** mostra duas variáveis que têm uma relação *linear*, que pode ser representada graficamente por uma linha reta (no exemplo, corrente × tensão num circuito elétrico); a Fig. **2** mostra variáveis que estão relacionadas de forma não-linear (no exemplo, a lei de Boyle, que relaciona o volume ocupado por um gás e a pressão a que ele está submetido). As relações mostradas nestes gráficos são exatas, e é possível escrever modelos *determinísticos* para descrevê-las; dizemos que um modelo é "determinístico" quando a partir de uma das variáveis é possível determinar o valor da outra. Estes modelos são muito comuns na Física clássica, tanto lineares (como $V=RI$, $F=ma$, $E=mc^2$), quanto não-lineares (como $e=e_o+v_ot+at^2/2$, $P=V^2/R$).

Este tipo de modelo, porém, tem pouca aplicação em áreas como a Biologia ou a Medicina, porque nos organismos vivos a *variação* entre os valores observados é sempre muito grande, e é difícil chegar a estas relações simples; em geral, é preciso construir modelos *probabilísticos*, que não determinam o valor que será assumido pela variável, mas indicam qual será o valor médio que ela irá assumir, depois de muitas repetições, e também um intervalo que contém os valores mais prováveis da variável.

As Figs. **3** e **4** mostram pares de variáveis que têm relações probabilísticas. A Fig. **3** mostra duas variáveis com relação *probabilística linear* (comprimento × peso de 500 recém-nascidos num hospital de São Paulo). Pode ser visto que os pontos da nuvem tendem a se agrupar em torno de uma linha reta; veremos a seguir o modelo usado para descrever este tipo de relação. A Fig. **4** mostra duas variáveis com relação *probabilística não-linear* (temperatura × consumo de energia numa cidade). Os pontos parecem agora se agrupar em

torno de uma curva, próxima de uma parábola; os modelos para descrever este tipo de relação são bem mais complicados, e não serão vistos neste texto.

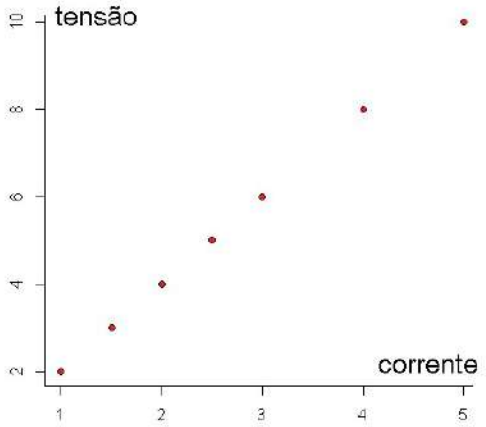

Fig. 1

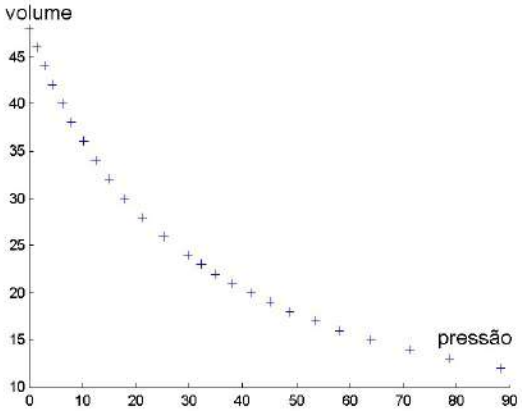

Fig. 2

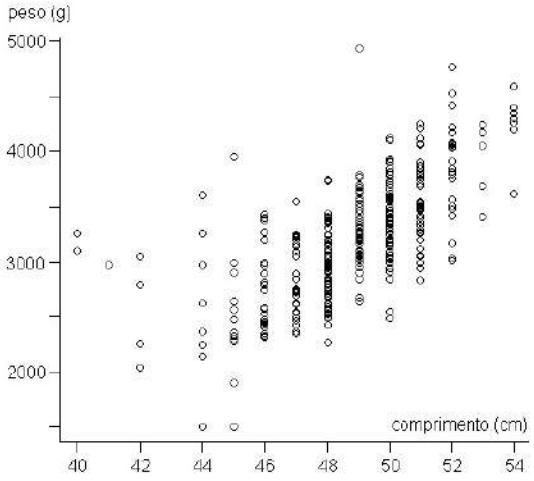

Fig. 3

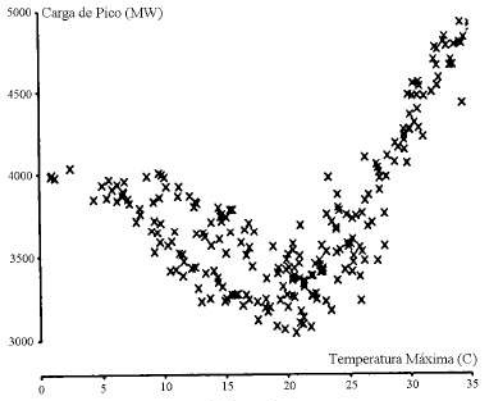

Fig. 4

Se existe relação probabilística linear entre duas variáveis, dizemos que elas estão *correlacionadas*, e a intensidade desta relação pode ser medida pelo *coeficiente de correlação linear de Pearson* (ver seção **2.2.4**). A relação pode ser descrita por um modelo estatístico, o *modelo de regressão linear*, que relaciona as duas variáveis. Este modelo é *probabilístico*; não determina o valor exato que será assumido pela variável Y, dado o valor de X, mas indica qual será o valor médio de Y, dado X; além disso, permite determinarmos intervalos em torna desta média onde há uma probabilidade conhecida de encontrarmos o valor de Y, dado X.

5.2.2. Modelo de regressão linear

5.2.2.1. Introdução

Toda ciência procura investigar as relações entre diversas variáveis e criar *modelos matemáticos* que descrevam estas relações. Um "modelo" matemático é uma equação, um gráfico, uma tabela, ou qualquer outro método que mostre como uma variável se compor-

ta, ou como duas ou mais variáveis se relacionam entre si. A Física, por exemplo, usa modelos como $V=RI$ (eletricidade), $F=ma$ (mecânica).

De onde vêm estes modelos? Alguns podem ter ser criados para descrever resultados observados experimentalmente. Por exemplo, Galileu Galilei (1564-1642), depois de fazer vários experimentos sobre corpos em queda livre, descobriu que os espaços percorridos por estes corpos eram proporcionais ao quadrado do tempo de queda, o que pode ser representado pelo modelo $h = kt^2$, onde h é o espaço percorrido, t é o tempo de queda, e k é uma constante a ser determinada. Outro exemplo é o de Johannes Kepler (1571-1630), que analisou os dados obtidos pelo astrônomo Tycho Brahe, e concluiu que os planetas giravam em torno do sol descrevendo uma elipse (e não um círculo, como se acreditava na época).

Modelos também podem vir a ser *deduzidos* matematicamente, como conseqüência de uma teoria mais ampla. Por exemplo, partindo da teoria da gravitação de Newton, foi possível demonstrar que o modelo geométrico para a órbita dos planetas deveria de fato ser uma elipse (confirmando o que tinha sido proposto empiricamente por Kepler), e também que o espaço percorrido por um corpo em queda livre é proporcional ao quadrado do tempo de queda (como tinha sido observado por Galileu).

Nas ciências, há várias vantagens em usar modelos matemáticos. A primeira delas é que o modelo torna explícita a relação entre as variáveis, e deixa claro quanto uma variável influencia (ou não) uma outra. Na época de Galileu, por exemplo, acreditava-se que a velocidade de queda de um corpo era proporcional ao *peso* deste corpo, e portanto um corpo mais pesado deveria cair mais rápido do que um corpo mais leve; o modelo de Galileu porém não inclui o peso como variável, e deixa bem claro que o peso do corpo não tem nenhuma influência no tempo de queda. Além disso, o modelo nos permite fazer previsões – usando o modelo de elipse, Kepler pode prever a posição futura de planetas, depois de algumas observações de posições no passado.

Estes dois modelos (de Kepler e de Galileu) são *determinísticos* (conhecido o valor de X, o valor de Y pode ser *determinado* com exatidão), e são obviamente simplificados, pois desprezam outros fatores; por exemplo, na queda livre, despreza-se a resistência do ar. Por usar estes modelos determinísticos, a Física costumava ser chamada de "ciência exata" (uma denominação que hoje está um tanto fora de moda). Quem já teve alguma experiência com aulas num laboratório de Física, porém, já teve ter percebido que, na prática, as coisas nunca acontecem exatamente como os modelos prevêem; os planetas, por exemplo, nunca são encontrados exatamente onde se esperava; há sempre um erro, uma diferença entre o que foi observado e o que tinha sido previsto (foi para justificar estas discrepâncias, aliás, Gauss criou a *Teoria dos Erros* e a curva normal, que se tornaram bases da Estatística). O que a Física clássica fazia era desprezar estes erros; podia fazer isto porque estes erros eram em geral muito pequenos, e não importavam muito no resultado final.

Outras ciências, porém, não podem ignorar os erros, tais como a Biologia, a Medicina, e qualquer outra que estude organismos vivos. Por exemplo, embora exista uma relação entre *peso* e *altura* de homens adultos, não é possível criar uma fórmula que nos permita prever o peso de um homem, dada a sua altura. Se tomarmos uma amostra de homens, medirmos suas alturas e pesos, e representarmos estes pares de valores num gráfico (onde cada ponto representa os dados de um homem), veremos que os pontos não estarão dispostos ao longo de uma linha reta ou curva; ao invés disto, se espalharão numa área do gráfico, formando uma *nuvem* de pontos. Não é possível fazer um modelo *determinístico* para a relação entre peso e altura, mas é possível fazermos um modelo *probabilístico*, que não irá *determinar* o valor do peso para uma dada altura, mas fará uma estimativa do valor *médio*

deste peso. Por exemplo, homens de 180 cm de altura têm pesos variados; assim também os homens de 150 cm. Contudo, os homens de 180 cm provavelmente terão *em média* pesos maiores que os homens de 150 cm; estas médias podem ser calculadas pelo modelo. Além disto, o modelo permite também estimar um *intervalo de previsão*, onde haja uma probabilidade estipulada (geralmente de 0,95) de encontrarmos o valor do peso de um homem, dada sua altura. Por serem baseados em probabilidades, estes modelos são chamados de *probabilísticos*.

5.2.2.2. Modelo de regressão linear

Se a relação entre duas variáveis é *probabilística linear* (como nas Figs. **3** e **5**), o modelo que pode representá-la é o chamado de *modelo de regressão linear*. Graficamente, ele consiste em uma linha reta, à qual é adicionada uma parcela aleatória, da forma

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad (1)$$

Esta reta é chamada de *reta de regressão de Y em X*. O que caracteriza um modelo probabilístico e o torna diferente de um modelo determinístico é a presença da parcela de erro aleatório e_i. Será possível estimar aproximadamente o valor de y_i quando conhecemos x_i, mas não prever este valor com exatidão, pois uma parcela e_i de erro sempre estará presente; este erro será representado por uma variável aleatória, e será portanto imprevisível.

Não *existe* uma terminologia padrão para designar as variáveis nos modelos de regressão. A variável *Y* pode ser chamada de variável *dependente* (porque depende da variável *X*), variável *resposta*, ou *desfecho*; a variável *X* pode ser chamada de variável *independente*, variável *explicativa*, ou *covariável*.

Existem técnicas matemáticas que permitem encontrar os valores dos parâmetros β de modo que o erro e_i seja o menor possível (ou, equivalentemente, que a reta de regressão esteja o mais "próxima" possível do centro do conjunto de dados). Por exemplo, a Fig. **5** mostra o diagrama de dispersão que relaciona o *peso* e o *comprimento* de 20 cães de uma amostra. A reta central é a *reta de regressão*, determinada por uma técnica chamada de *método dos mínimos quadrados ordinários (MQO)*.

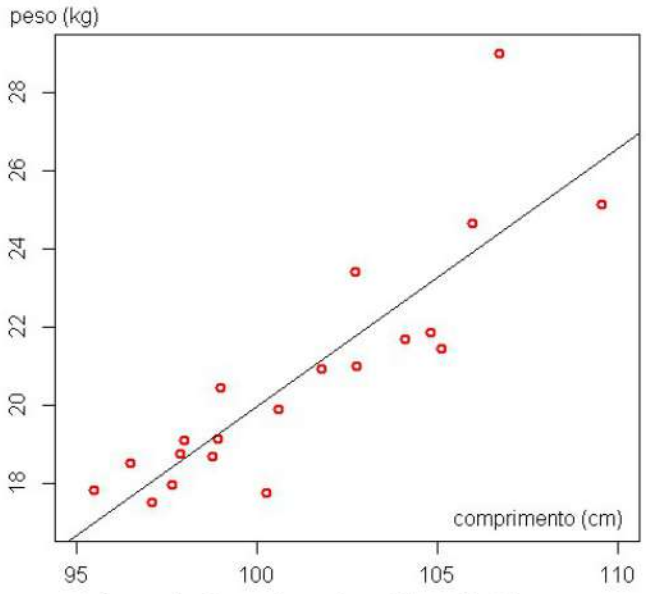

Figura 5. Comprimento x altura de 20 cães

A equação desta reta é:

$$peso_i = \beta_0 + \beta_1 \times comprimento_i + e_i$$

A reta mostra o valor *médio* que o peso assume para cada valor dado do comprimento; i.e., o peso Y que se esperaria, em média, que um cão tivesse, dado o seu comprimento X. Os valores estimados para os coeficientes, a partir da amostra (este exemplo será desenvolvido a seguir, na seção **5.2.4**) foram:

$$\hat{\beta}_0 = -46{,}1$$

$$\hat{\beta}_1 = 0{,}67$$

Podemos portanto prever que os animais de 100 cm de comprimento devem ter em média um peso de:

$$\text{peso médio} = -46{,}19 + 0{,}67 \times 100 = 20{,}8 \text{ kg}$$

É claro, porém, que nem todos os animais de 100 cm pesarão exatamente 20,8 kg; seus pesos irão variar acima e a abaixo deste valor médio. A diferença entre os pesos realmente encontrados e o peso médio previsto é representada no modelo pelo erro aleatório e_i. Não podemos prever o valor deste erro, mas se usarmos para ele um modelo de *variável aleatória contínua*, poderemos prever faixas onde é mais provável que o valor se encontre. Geralmente pressupõe-se que o erro seja uma VAC *gaussiana* (esta pressuposição terá que ser confirmada depois, por meios de testes). A partir daí, podemos determinar um *intervalo de predição*, isto é, um intervalo onde há uma probabilidade conhecida (geralmente 0,95) de se encontrarem os valores do peso.

É importante, contudo, lembrar que este modelo não pretende simplesmente descrever o que acontece *nesta amostra*, e sim, descrever o que acontece *na população* que esta amostra representa; não nos interessa afirmar que 95% dos cães desta amostra têm pesos nesta faixa, mas sim afirmar que 95% de todos os cães desta raça têm pesos nesta faixa.