Por: Áurea Sousa
Professora Auxiliar do Departamento de Matemática e Estatística
da Faculdade de Ciências e Tecnologia da Universidade dos Açores
aurea.st.sousa@uac.pt

Diagrama de dispersão, correlação e regressão linear

A representação gráfica no âmbito da análise estatística de dados é sobejamente conhecida e tem vindo a ser sublinhada ao longo do tempo. Além da deteção de padrões nos dados, a representação gráfica permite observar a eventual existência de valores muito grandes ou muito pequenos (outliers) quando comparados com os restantes valores da coleção de dados.

A representação dos pares de valores correspondentes a duas variáveis, X e Y, num sistema cartesiano dá origem a um gráfico, conhecido como diagrama de dispersão, o qual mostra o comportamento conjunto dessas variáveis, permitindo observar, por exemplo, se a relação existente entre estas é ou não linear (isto é, se os pontos projetados se tendem a aproximar ou não de uma reta imaginária).

Se quando os valores de uma variável aumentam os da outra variável também tendem a aumentar, dizemos que há uma correlação positiva entre as duas variáveis. Em contrapartida, se quando os valores de uma variável aumentam os da outra variável tendem a diminuir, dizemos que há uma correlação negativa (nesse caso, o sinal do coeficiente de correlação é negativo) entre as duas variáveis. Por outro lado, se os pontos estiverem demasiado dispersos conclui-se que não há correlação linear entre as variáveis ou que esta é muito baixa. Assim, a partir da observação de um diagrama de dispersão, podemos verificar, por exemplo, se a correlação linear entre as duas variáveis é mais ou menos forte, tendo em atenção a proximidade dos pontos projetados em relação a uma reta imaginária.

O coeficiente de correlação de Pearson, denotado por ρ na população e por r na amostra, varia entre -1 e 1 e permite quantificar a intensidade da associação linear entre duas variáveis quantitativas, informando-nos ainda sobre a direção da relação existente entre essas variáveis. Segundo Franzblau (1958), os valores deste coeficiente podem ser interpretados de acordo com o apresentado na tabela seguinte, onde || significa módulo ou valor absoluto do valor referente ao coeficiente de correlação, sendo de ressaltar que quanto mais próximo de 1 for o módulo do coeficiente de correlação, maior é a intensidade da associação linear entre as duas variáveis.

O modelo de regressão linear tem maior alcance do que aquele que parece ter à primeira vista, porque é possível linearizar diversos modelos não lineares, utilizando transformações matemáticas apropriadas a cada caso.

A Análise de regressão tem diversas aplicações em diversas áreas, tais como as Ciências da Saúde, as Ciências Sociais e Humanas, a Economia e a Gestão. A Econometria, conciliando a Matemática, a Estatística e a Economia, ocupa-se dos problemas de medida e quantificação das relações económicas, sendo de salientar que o modelo clássico de regressão linear constitui o alicerce principal da Econometria. Atualmente a análise estatística de regressão constitui um domínio que se tem tornado cada vez mais relevante para a formação de economistas e gestores e de outros profissionais.

Os gráficos aqui apresentados representam três situações que merecem uma especial referência: o caso em que há uma correlação perfeita positiva entre as duas variáveis (r=1); o caso em que há uma correlação perfeita negativa entre as duas variáveis (r= -1) e o caso em que há uma correlação nula entre as duas variáveis (isto é, o caso em que não há correlação linear).

Neste contexto, nunca é demais recordar que a inexistência de uma relação linear entre duas variáveis não significa que não se verifique outro tipo de relação entre estas (e.g., exponencial) e que a existência de correlação não significa causalidade.

No estudo de um determinado fenómeno, o comportamento de uma variável Y em função de outra variável X pode ser descrito através de um modelo matemático (linear, quadrático, cúbico, exponencial, logarítmico, entre outros). A escolha do modelo apropriado deve ser feita com base na observação do tipo de curva e na correspondente escolha da equação de um modelo matemático que mais se aproxime dos pontos exibidos no correspondente diagrama de dispersão. Porém, convém ter em mente que os pontos projetados se encontram geralmente um pouco distantes dos que se situam sobre a curva referente ao modelo ajustado.

Afirma-se que existe uma relação linear entre as variáveis se os dados se aproximarem de uma linha reta. Quando isso acontece, tem muitas vezes interesse a realização de uma análise de regressão linear simples (ajustamento de uma reta aos dados), a qual permite verificar a existência de uma relação funcional entre uma variável dependente (variável a explicar/ variável resposta, geralmente designada por Y) e uma variável independente ou explicativa (X), ambas quantitativas. Neste contexto, os resíduos são dados pela diferença (desvios) entre os valores observados da variável dependente e os correspondentes valores estimados (valores situados sobre a reta de regressão), e têm um papel fundamental na verificação da adequação do modelo de regressão considerado.

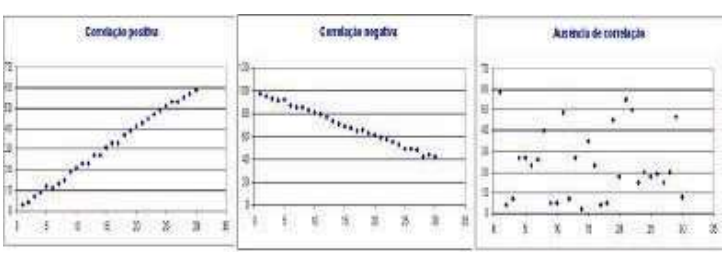
Supondo que existe efetivamente uma relação linear entre duas variáveis quantitativas, X e Y, Karl Gauss, entre 1777 e 1855, propôs a estimação dos parâmetros da reta de regressão, mediante a minimização da soma dos quadrados dos desvios, processo conhecido como método dos mínimos quadrados. Assim, o leitor deverá notar que ambas as técnicas (correlação e regressão linear simples), apesar de interligadas, diferem, uma vez que no caso da correlação as variáveis são aleatórias e desempenham um papel idêntico, podendo não haver nenhuma dependência entre estas, ao contrário do que acontece no caso da regressão. Quanto mais próximo de 1 for o quadrado do coeficiente de correlação de Pearson (coeficiente de determinação), maior é a percentagem da variação de Y que é explicada pela reta de regressão estimada, e consequentemente, maior é a qualidade do ajustamento.

Valores de r	Interpretação
r=0	Não existe relação linear
|r| < 0,20	Correlação negligenciável
0,20 < |r| < 0,40	Correlação é fraca
0,40 < |r| < 0,60	Correlação moderada
0,60 < |r| < 0,80	Correlação é forte
|r| > 0,80	Correlação é muito forte

Um modelo geral de regressão linear visa estudar o relacionamento entre uma variável dependente, Y, e uma ou várias variáveis independentes, X, (X1, X2, …, XP). No caso particular em que há apenas uma variável independente a regressão é designada por regressão linear simples. Em contrapartida a existência de duas ou mais variáveis independentes remete-nos para um modelo de regressão linear múltipla. De um modo geral, a análise de regressão permite verificar a existência de uma relação funcional entre uma variável dependente (variável a explicar/ variável resposta) e uma ou mais variáveis independentes ou explicativas, sendo de salientar que a equação de regressão visa explicar a variação da variável dependente com base na variação da(s) variável(eis) independentes.

Os resíduos devem ser normalmente distribuídos, terem variância constante (homocedasticidade) e serem independentes; também não devem existir outliers influentes. No caso da regressão linear múltipla, para além da avaliação destes pressupostos, é preciso ainda verificar se existe colinearidade (correlação elevada entre duas variáveis independentes) ou multicolinearidade (mais do que duas variáveis independentes fortemente correlacionadas) entre as variáveis independentes.

No estudo da regressão linear (simples ou múltipla), tem especial interesse a indicação do modelo teórico, a verificação dos seus pressupostos, a estimação dos parâmetros do modelo e a avaliação da qualidade do ajustamento. Por vezes são utilizadas variáveis binárias ou variáveis dummy (variáveis que tomam apenas dois valores (0 e 1)), com o intuito de quantificar efeitos de ordem qualitativa sobre a variável dependente.



Polícia Marítima trabalha apenas com metade dos efectivos que são necessários

Grupo Parlamentar do PSD reuniu com a Associação Sócio Profissional da Polícia Marítima

O deputado do PSD/Açores na Assembleia da República, António Ventura, questionou o Governo de António Costa sobre "a falta de agentes da Polícia Marítima na Região, que actualmente tem um efectivo que é apenas metade do necessário", avançou.

O social democrata falava após uma audiência da Associação Sócio Profissional da Polícia Marítima (ASPPM) com o grupo parlamentar do PSD, tendo explicado que pretende conhecer "o ponto da situação da Polícia Marítima na Região, assim como quantos agentes estão previstos serem colocados nas nossas ilhas", explicou.

"Queremos igualmente informações sobre os investimentos previstos, em termos de equipamentos da Polícia Marítima nos Açores", acrescentou o deputado.

Segundo aquela associação, o efectivo regional da Polícia Marítima "encontra-se reduzido a cerca de metade do previsto, podendo estar em risco determinadas funções de cumprimento da lei no arquipélago", reforçou António Ventura.

A Polícia Marítima, como força policial na dependência da Autoridade Marítima, "é essencial para garantir o cumprimento das leis e dos regulamentos nos espaços públicos integrantes do espaço marítimo, áreas portuárias ou zonas balneares", refere o deputado do PSD/Açores.

O social democrata adianta que "a insularidade e a geografia dos Açores requerem uma ação e uma atenção acrescidas e permanentes, por parte da Polícia Marítima", uma vez que os Açores "possuem um conjunto significativo de reservas naturais e recursos piscatórios, que devem ser preservados", disse.

António Ventura conclui, alertando para o facto de "o crescente aumento do turismo em todas as ilhas justificar uma cada vez maior atenção por parte das autoridades policiais, daí as questões colocadas ao Governo, que consideramos de fulcral importância".