Métodos Quantitativos (e Qualitativos)

6. Correlações e regressão linear

Salvatore Cosentino
D.Mat. U.Minho

12 jan 2021

Modelização

Uma lei física é uma relação entre um certo número de observáveis.

Um exemplo simples e ideal é

$$\boxed{y = f\,(x, a)}$$

onde y, x, a são certos observáveis, e f é uma função.

Os objectivos das experiências, em que observamos valores (x_k, y_k), podem ser:

decidir se y depende mesmo de x.

conjeturar a lei, ou seja, a forma da função f,

estimar os valores dos parâmetros livres $a = (a_1, a_2, \dots)$ que mais concordam com as observações,

fazer previsões sobre valores de y em correspondência de valores de x ainda não testados.

Associação

Nas C.S., um objetivo típico é decidir se há alguma forma de "associação", ou seja, dependência, entre duas ou mais variáveis

(por exemplo, se a taxa de criminalidade cresce com o incremento da pobreza, . . .)

Naturalmente, uma associação entre as variáveis x e y não implica necessariamente uma relação de causa-efeito, podendo as duas ser dependentes de uma terceira variável z . . .

Experiências

Uma experiência típica para testar uma lei

$$y = f(x, a)$$

consiste em observar os valores

$$y_1 \quad y_2 \quad y_3 \quad \dots \quad y_n$$

da variável y em correspondência de um certo número de valores

$$x_1 \quad x_2 \quad x_3 \quad \dots \quad x_n$$

da variável x, considerada como variável independente.

Numa experiência ideal temos um bom controlo, e possivelmente nenhum erro significativo, do observável x.

Em correspondência de cada valor x_k temos muitas observações de y, e portanto uma estimação da média $\overline{y_k}$ e do desvio padrão S_{y_k}.

Em geral, cada x_k é observado muitas vezes e estimado com a sua média $\overline{x_k}$ e o seu desvio padrão S_{x_k}

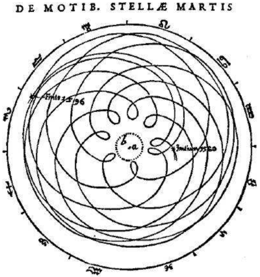
Leis simples

A lei pode ser uma previsão de uma teoria física que queremos testar, ou simplesmente uma conjectura sugerida pelos resultados das experiências.

Uma função f suficientemente irregular e um número grande de parâmetros livres permite ajustar com ótima precisão qualquer dado experimental!

(basta, por exemplo, que f seja um polinómio de grau superior ou igual ao número das observações)

Famoso é o caso dos epicíclos que os gregos usavam para modelar as órbitas dos planetas.



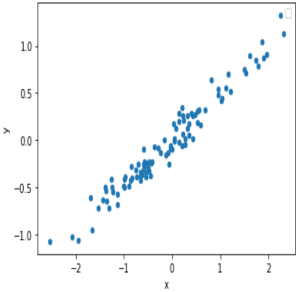
É boa ideia experimentar leis simples, possivelmente com poucos parâmetros livres.

Diagramas de dispersão

Um diagrama de dispersão (em inglês, scatter plot) dos

$$\boxed{x_k \qquad \text{versus} \qquad y_k}$$

pode sugerir, ou não, uma correlação entre x e y, e possivelmente a forma da lei.



Mais honesto é um diagrama de dispersão que tenha em consideração os erros, obtido ao fazer mais observações para cada k, ou considerando as sensibilidades dos instrumentos utilizados para medir os x_k e os y_k, logo do género

$$\overline{x_k} \pm S_{x_k} \qquad \text{versus} \qquad \overline{y_k} \pm S_{y_k}$$

Tabelas de contingências

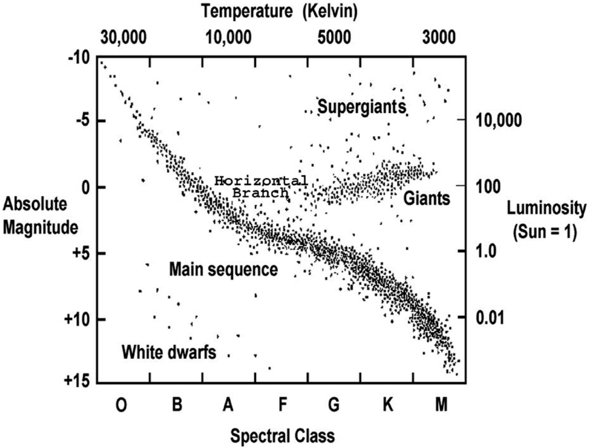
Quando as duas variáveis são classificadas num número pequeno de classes, por exemplo são variáveis dicotómicas, os diagramas de dispersão são substituídos por tabelas de contingências (em inglês, cross-tab(ulation)),

matrizes com as frequências observadas

	direito	canhoto
macho	2	26
fêmea	5	32

Exemplo: o diagrama de Hertzsprung e Russell

O diagrama de Hertzsprung[1] e Russell[2] relaciona a magnitude absoluta (ou seja, a luminosidade) com a temperatura das estrelas (ou seja, a classe espetral)



[1] E. Hertzsprung, On the Use of Photographic Effective Wavelengths for the Determination of Color Equivalents *Publications of the Astrophysical Observatory in Potsdam* **22** (1911)

[2] H.N. Russell, Relations Between the Spectra and Other Characteristics of the Stars, *Popular Astronomy* **22** (1914), 275-294.

Exemplo: a lei de Hubble

A lei linear mais famosa da história da física é tal vez a lei de Hubble [3]

$$v = H \cdot d$$

que mostra a proporcionalidade entre a velocidade v de afastamento das galáxias e as distâncias d entre as galáxias e a nossa Via Láctea, evidência da expansão do universo, efeito do provável big bang.

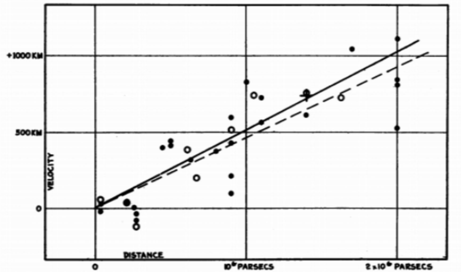


FIGURE 1
Velocity-Distance Relation among Extra-Galactic Nebulae.

[3] E. Hubble, A relation between distance and radial velocity among extra-galactic nebulae, *Proceedings of the National Academy of Sciences* **15** (1929) 168-173.

Exemplo: a "mouse to elephant curve"

Uma lei não linear em biologia é a lei de Kleiber [4]

$$r \sim m^{3/4}$$

que diz que a taxa metabólica r de um mamífero é proporcional a $\frac{3}{4}$-ésima potência da sua massa m.

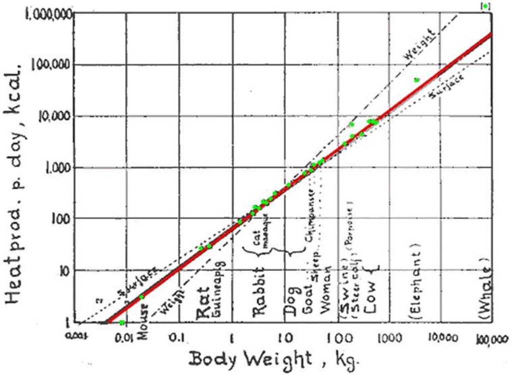


Fig. 1. Log. metabol. rate/log body weight

[4] M. Kleiber, Body size and metabolic rate, *Physiological Reviews* **27** (1947) 511-541.

Mínimos quadrados

O método dos mínimos quadrados (em inglês, least-square fitting) é uma receita que consiste em escolher os estimadores α para os parâmetros livres a de maneira tal que a soma dos erros quadráticos

$$Q_a^2 = \sum_{k=1}^{n} \frac{\left(\overline{y_k} - f\left(x_k, a\right)\right)^2}{S_{y_k}^2}$$

seja a menor possível.

Observe que cada erro quadrático

$$\varepsilon_k^2 = \left(\overline{y_k} - f\left(x_k, a\right)^2\right)$$

é pesado com um fator inversamente proporcional à incerteza S_{y_k},

Em particular, se n é grande, um dado incerto, por exemplo com $S_{y_{13}}$ muito maior que os outros S_{y_k}, não influencia significativamente a estimação.

Máxima verosimilhança

Uma hipótese de trabalho razoável é a hipótese gaussiana: cada y_k tem lei normal com esperança $f(x_k, a)$ e variância $S_{y_k}^2$.

Neste caso, a densidade de probabilidade de obter o resultado $\overline{y_k}$ é

$$p(y_k) = \frac{1}{S_{y_k}\sqrt{2\pi}}\, e^{-\frac{1}{2}\frac{(\overline{y_k} - f(x_k, a))^2}{S_{y_k}^2}}$$

Na hipótese de que as diferentes observações são independentes, a densidade de probabilidade de obter os resultados $\overline{y_1}, \overline{y_2}, \ldots, \overline{y_n}$ é proporcional a

$$\exp\left(-\frac{1}{2}\sum_{k=1}^{n}\frac{(\overline{y_k} - f(x_k, a))^2}{S_{y_k}^2}\right)$$

O método dos mínimos quadrados portanto maximiza a densidade de probabilidade, e por esta razão é também chamado princípio da máxima verosimilhança.

Na prática

Em teoria, desde que a função f seja diferenciável, os valores de

$$a = (a_1, a_2, \ldots, a_m)$$

são obtidos calculando as derivadas parciais $\partial Q_a^2 / \partial a_j$ e resolvendo o sistema de m equações

$$\frac{\partial Q_a^2}{\partial a_j} = 0$$

com $j = 1, 2, ..., m$.

Na prática, se a forma de f não é simples, este é um problema difícil.

O melhor é procurar soluções aproximadas, por exemplo utilizando técnicas de análise numérica.

A propagação dos erros permite também estimar as incertezas nos parâmetros, na forma

$$a = \alpha \pm S_\alpha$$

Qualidade do ajuste

O método dos mínimos quadrados estima os parâmetros livres a e portanto produz a conjectura $y = f(x, \alpha)$, que os estatísticos chamam curva de regressão.

O problema é que o método dos mínimos quadrados funciona sempre, independentemente da forma de f e dos valores das observações!

A posteriori, convém avaliar a qualidade do ajuste, com base no bom senso e na honestidade do cientista.

O ajuste pode ser considerado bom se os valores de $f(x_k, \alpha)$ pertencem aos intervalos

$$\overline{y_k} \pm S_{y_k}$$

ou se pelo menos não se afastam dos $\overline{y_k}$ por mais de que multiplos pequenos de S_{y_k}.

Também, é boa norma verificar que a sequência dos sinais dos erros $\overline{y_k} - f(x_k, \alpha)$ não mostra um padrão suspeito.

Teste qui-quadrado

O valor de

$$Q_\alpha^2 = \sum_{k=1}^{n} \frac{\left(\overline{y_k} - f\left(x_k, \alpha\right)\right)^2}{S_{y_k}^2} = \min_a \sum_{k=1}^{n} \frac{\left(\overline{y_k} - f\left(x_k, a\right)\right)^2}{S_{y_k}^2}$$

é uma medida da qualidade do ajuste.

De fato, se α são os valores verdadeiros dos a, então a hipótese gaussiana implica que Q_α^2 tem lei qui-quadrado $\chi^2(n-m)$ com $n-m$ graus de liberdade.

Uma tabela/software fornece então a probabilidade

$$q = \mathrm{Prob}\left(Q^2 > Q_\alpha^2\right)$$

onde Q^2 é uma variável com lei $\chi^2(n-m)$.

A interpretação é: q é a probabilidade de observar um qui-quadrado maior do que foi observado na hipótese "$y = f\left(x, \alpha\right)$".

Decisão

Se a lei conjeturada é a hipótese conservadora, então os cientistas consideram aceitáveis valores

$$q \geq 0.1 \qquad \text{ou até} \qquad \geq 0.01$$

(esta regra é equivalente a aceitar a hipótese nula "a lei $y = f(x, \alpha)$ é verdadeira" com nível de significância da ordem de 5% ou 1%, valores típico de um teste sobre uma hipótese conservadora).

Se as variâncias $S^2_{y_k}$ foram subestimadas, o se os dados não são gaussianos, pode até acontecer que bons modelos levem a valores

$$q \simeq 0.001$$

Por outro lado, valores grandes de Q^2_α, tais que

$$\boxed{q \ll 0.001}$$

são fortes indícios de que a conjectura $y = f(x, \alpha)$ não é uma lei que descreve bem os dados observados.

Se . . .

Se só temos uma observação de y_k para cada valor x_k, não temos uma estimação credível das variâncias $S^2_{y_k}$.

O que os físicos fazem nesse caso é pôr as variâncias iguais a 1 nas fórmulas acima, portanto minimizar

$$\sum_{k=1}^{n} \left(y_k - f\left(x_k, \alpha\right)\right)^2 = \min_{a} \sum_{k=1}^{n} \left(y_k - f\left(x_k, a\right)\right)^2$$

e depois estimar

$$S^2_{y_k} \simeq \frac{1}{n-m} \sum_{k=1}^{n} \left(y_k - f\left(x_k, \alpha\right)\right)^2$$

(o que significa fazer a hipótese de que as $S^2_{y_k}$ são todas iguais).

A partir destas variâncias é possível, usando a fórmula da propagação dos erros, estimar os erros nos parâmetros α.

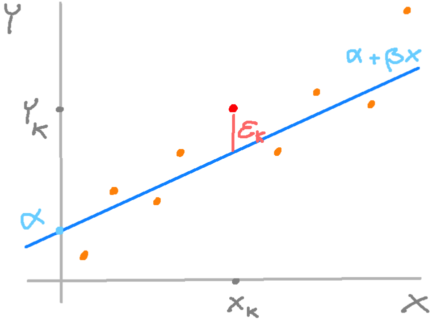
Regressão linear

Modelos que são tratáveis analiticamente são os modelos lineares, tais que $f(x, a)$ depende linearmente dos parâmetros a, porque minimizar os desvios quadráticos é equivalente a resolver um sistema de equações lineares.

Um exemplo simples é uma lei linear (tecnicamente, "afim")

$$\boxed{y = a + bx}$$

entre os observáveis x e y (mas também é possível considerar mais variáveis independentes $x', x'', \ldots$).



Mínimos quadrados

De acordo com a receita dos mínimos quadrados, os estimadores α e β são os valores dos parâmetros a e b que minimizam a soma dos erros quadráticos

$$Q^2_{a,b} = \sum_{k=1}^{n} (a + bx_k - y_k)^2$$

Resolvendo o sistema de equações

$$0 = \frac{\partial Q^2_{a,b}}{\partial a} \qquad \Rightarrow \qquad 0 = n\overline{y} - n\alpha - \beta\overline{x}$$

$$0 = \frac{\partial Q^2_{a,b}}{\partial b} \qquad \Rightarrow \qquad 0 = \sum_{k=1}^{n} y_k x_k - n\alpha\overline{x} - \beta \sum_{k=1}^{n} x_k x_k$$

obtemos a resposta

$$\beta = \frac{S^2_{xy}}{S^2_{xx}} \qquad \text{e} \qquad \alpha = \overline{y} - \beta\,\overline{x}$$

onde

$$S^2_{xy} = \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y}) \quad \text{e} \quad S^2_{xx} = \sum_{k=1}^{n} (x_k - \overline{x})(x_k - \overline{x})$$

Reta de regressão

A reta estimada

$$\boxed{y = \alpha + \beta x}$$

é chamada reta de regressão.

Os seus parâmetros, o declive (e inglês, slope) e a ordenada na origem (em inglês, intercept) são

$$\boxed{\beta = \frac{S_{xy}^2}{S_{xx}^2}} \qquad \text{e} \qquad \boxed{\alpha = \overline{y} - \beta\,\overline{x}}$$

respetivamente,

Também utilizado é o coeficiente β estandardizado

$$\beta^* = \frac{S_{xx}}{S_{yy}}\,\beta$$

que é adimensional.

Média dos quadrados dos resíduos

Na hipótese gaussiana, α e β são bons estimadores de a e b respectivamente, porque α tem lei normal $N\left(a, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}^2}\right)\right)$ e β tem lei normal $N\left(b, \sigma^2/S_{xx}^2\right)$.

Naturalmente não sabemos o valor de σ^2,

mas um seu estimador é a média dos quadrados dos resíduos

$$\boxed{S^2 = \frac{1}{n-2}\sum_{k=1}^{n}(\alpha + \beta x_k - y_k)^2}$$

Sempre na hipótese gaussiana, a variável

$$\frac{S^2}{\sigma^2}$$

tem lei qui-quadrado $\chi^2(n-2)$.

Estimação dos parâmetros

Se definimos

$$S_\alpha = S\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}^2}} \qquad \text{e} \qquad S_\beta = \frac{S}{S_{xx}}$$

o modelo diz que

$$\frac{\alpha - a}{S_\alpha} \qquad \text{e} \qquad \frac{\beta - b}{S_\beta}$$

têm lei de Student $T(n-2)$.

Intervalos de confiança de nível $1-\varepsilon$ pelos parâmetros da lei linear são portanto

$$\boxed{a = \alpha \pm t_{1-\varepsilon/2} \cdot S_\alpha} \qquad \text{e} \qquad \boxed{b = \beta \pm t_{1-\varepsilon/2} \cdot S_\beta}$$

onde $t_{1-\varepsilon/2}$ é o quantil da lei de Student $T(n-2)$.

Teste sobre a independência linear

Como decidir que $y = a + bx$ é mesmo uma lei ?

Uma primeira ideia é testar a hipótese nula

$$b = 0$$

ou seja a hipótese conservadora de que não há evidência experimental de dependência linear entre x e y.

Fixado um nível de significância ε, a região crítica do teste é

$$\left|\beta/S_\beta\right| > t_{1-\varepsilon/2}$$

onde $t_{1-\varepsilon/2}$ é o quantil da lei de Student $T(n-2)$.

Portanto, admitimos que a variável y depende (e linearmente) de x se encontramos um valor

$$|\beta| > t_{1-\varepsilon/2} \cdot S_\beta$$

Para valores típicos do nível de significância, 5% ou 1%, este limite é da ordem de duas ou três vezes $S_\beta = S_{\text{res}}/S_{xx}$, a razão entre as incertezas nas variáveis $(y_k - \alpha + \beta x_k)$ e x_k, o que é muito razoável.

Simetria

A falta de simetria das fórmulas acima reflecte o fato de considerar x como variável independente da lei $y = a + bx$.

A regressão tipicamente é utilizada quando temos um bom controlo do observável x, e por isto podemos pensar que os erros na sua determinação são desprezáveis.

Caso contrário, ao escrever a lei na forma $x = a' + b'y$, o argumento acima produz a reta de regressão

$$x = \alpha' + \beta' y$$

onde agora os estimadores de b' e a' são

$$\boxed{\beta' = \frac{S^2_{xy}}{S^2_{yy}}} \qquad \text{e} \qquad \boxed{\alpha' = \overline{x} - \beta' \overline{y}}$$

onde

$$S^2_{yy} = \sum_{k=1}^{n} (y_k - \overline{y})(y_k - \overline{y})$$

Coeficiente de determinação

A relação teórica entre os declives b e b' é

$$bb' = 1$$

O produto dos seus estimadores

$$\boxed{R^2 = \beta\beta' = \frac{S_{xy}^4}{S_{xx}^2\, S_{yy}^2}}$$

é dito coeficiente de determinação, e assume valores no intervalo

$$0 \leq R^2 \leq 1$$

A qualidade do ajuste pode ser considerada boa se $R^2 \simeq 1$.

Por outro lado, é razoável suspeitar que observar $R^2 \simeq 0$ é indício de que a lei linear não descreve bem os dados das experiências.

Variabilidade explicada

Os estatísticos também dizem que R^2 é "a proporção de variabilidade de y explicada pela regressão", pois é a razão

$$\boxed{R^2 = \frac{S^2_{\text{reg}}}{S^2_{\text{tot}}}}$$

entre a variabilidade explicada pela regressão

$$\begin{aligned} S^2_{\text{reg}} := \sum_{k=1}^{n} (\alpha + \beta x_k - \overline{y})^2 \quad &= \quad \sum_{k=1}^{n} (\overline{y} - \beta\overline{x} + \beta x_k - \overline{y})^2 \\ &= \quad \beta^2 \sum_{k=1}^{n} (x_k - \overline{x})^2 = \frac{S^4_{xy}}{S^4_{xx}} S^2_{xx} = \frac{S^4_{xy}}{S^2_{xx}} \end{aligned}$$

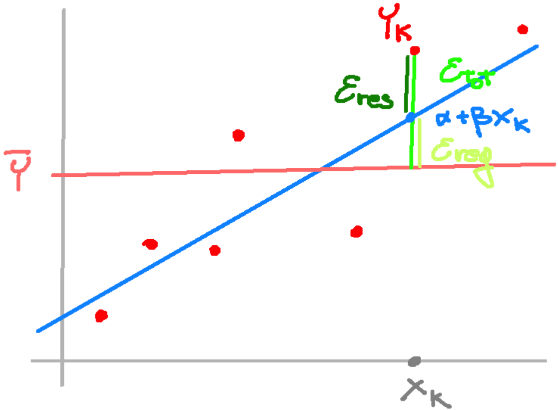
e a variabilidade total da variável dependente

$$S^2_{\text{tot}} := S^2_{yy} = \sum_{k=1}^{n} (y_k - \overline{y})^2$$

Variabilidade residual

Fica "por explicar" a variabilidade residual

$$S^2_{\text{res}} = S^2_{\text{tot}} - S^2_{\text{reg}} = \sum_{k=1}^{n} (\alpha + \beta x_k - y_k)^2$$



Análise da variância

É claro que um modelo linear $y = \beta x + \alpha$, com 2 parâmetros livres, deve poder ajustar melhor os dados de que um modelo $y = \alpha$, sem dependência entre as variáveis, que tem apenas 1 parâmetro livre.

É natural, no entanto, testar a hipótese nula de que "o modelo linear não é significativamente melhor".

É razoável rejeitar a hipótese nula, logo aceitar o modelo linear, apenas se a variabilidade explicada pela regressão

$$S^2_{\text{reg}} := \sum_{k=1}^{n} (\alpha + \beta x_k - \overline{y})^2$$

é grande, ou seja, significativa, quando comparada com a variabilidade residual

$$S^2_{\text{res}} = \sum_{k=1}^{n} (\alpha + \beta x_k - y_k)^2$$

Para fazer esta comparação é necessário estimar os valores esperados destas duas variabilidades na hipótese nula ...

... e este é um caso particular de análise da variância (o acrónimo inglês é ANOVA), desenvolvida por Fisher.

Teste F

O quociente

$$\boxed{F = \frac{S^2_{\text{reg}}}{\frac{1}{n-2}\, S^2_{\text{res}}}}$$

é chamado estatística F.

Na hipótese nula, F tem uma distribuição conhecida (um quociente entre duas variáveis qui-quadrado normalizadas) chamada distribuição de Fisher-Snedecor $F(1, n-2)$ com $2-1$ e $n-2$ graus de liberdade.

Os software de estatística calculam diretamente o p-value, a probabilidade

$$p = \text{Prob}\,(f > F)$$

de uma variável de Fisher-Snedecor ser superior ao valor observado F.

Assim, rejeitamos a hipótese nula, logo consideramos razoável uma lei linear, se o p-value for significativamente inferior ao nosso nível de significância preferido (tipicamente 5% ou 1%).

Coeficiente de correlação (linear)

Uma medida adimensional da "correlação linear" entre x e y é o coeficiente de correlação (empírico), ou coeficiente de correlação de Pearson,

$$\boxed{R = \frac{S_{xy}^2}{S_{xx}\, S_{yy}}}$$

ou seja,

$$R = \frac{\sum_k (x_k - \overline{x})(y_k - \overline{y})}{\sqrt{\sum_k (x_k - \overline{x})^2}\sqrt{\sum_k (y_k - \overline{y})^2}}$$

que assume valores no intervalo

$$-1 \leq R \leq 1$$

e não depende das médias e das variâncias dos observáveis (logo da origem a da escada usada para as medir).

O seu quadrado é o coeficiente de determinação,

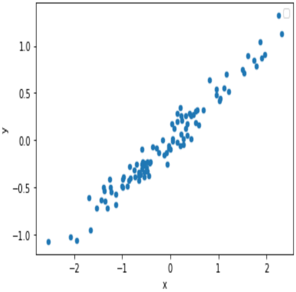
e o seu sinal é o sinal do declive β.

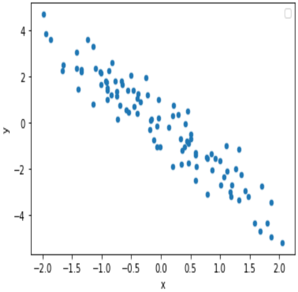
R próximo de mais ou menos um

Um valor de

$$\boxed{R \simeq \pm 1}$$

é indício de correlação linear efectiva entre as variáveis.





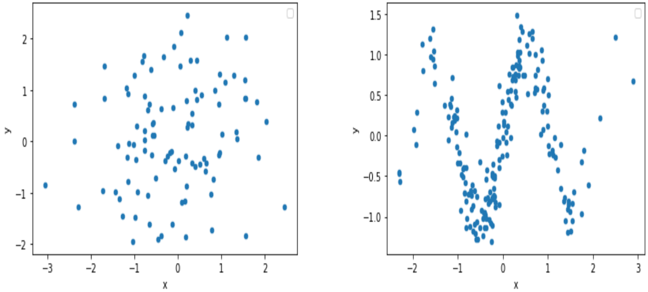
R próximo de zero

Um valor de

$$\boxed{R \simeq 0}$$

é indício de que as variáveis podem ser independentes

ou, pelo menos, não relacionadas por uma lei linear.



p-value

Um cientista honesto testa a hipótese nula de que as variáveis x e y são independentes.

Os livros/software de estatística permitem calcular o p-value

$$\boxed{p = \mathrm{Prob}(|\rho| \geq |R| \,|\, x \text{ e } y \text{ são independentes})}$$

a probabilidade de observar os n dados com coeficiente de correlação ρ superior ao observado R se a hipótese nula for verdadeira.

Um valor de p suficientemente pequeno, por exemplo

$$p \leq 0.05 \qquad \text{ou} \qquad p \leq 0.01$$

é considerado evidência de que a hipótese conservativa pode ser rejeitada (num teste com nível de significância 5% ou 1%), assim que podemos aceitar a lei linear.

Valores aceitáveis de R

A seguinte tabela mostra o limite inferior da região crítica $|R| > r$ deste teste para níveis de significância 5% e 1% em função do números de observações $n = 10, 20, 30, 40, 60, 80, 100$

	10	20	30	40	60	80	100
5%	0.63	0.44	0.36	0.31	0.26	0.22	0.20
1%	0.76	0.56	0.46	0.40	0.34	0.29	0.26

Por exemplo, se $n = 10$, a correlação linear é considerada efetiva, com nível de significância 5%, se é observado um coeficiente de correlação

$$|R| \geq 0.63$$

Se $n = 100$, a correlação linear é considerada efetiva, com nível de significância 5%, a partir de

$$|R| \geq 0.20$$

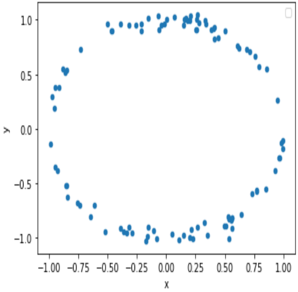
Atenção !

É importante lembrar que este teste apenas avalia a correlação linear entre as variáveis!

Por exemplo, se os observáveis x e y verificam a identidade

$$x^2 + y^2 = \text{constante}$$

e portanto os resultados das experiências estão distribuídos ao longo de uma circunferência, então o coeficiente de correlação esperado é $R = 0$ (por razões de simetria), embora as variáveis não sejam independentes ...



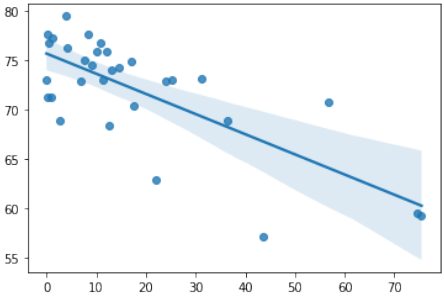
Previsões

A regressão linear estima os valores mais prováveis de a e b, e portanto a lei na forma da reta de regressão

$$y = \alpha + \beta x$$

Pode ser utilizada para fazer uma previsão do valor de y em correspondência de um certo valor x da variável independente,

desde que o valor x não se afaste muito do intervalo $[x_{\text{min}}, x_{\text{max}}]$ onde fizemos as experiências.



Previsão com intervalos

A hipótese gaussiana implica que a a variância de $y - \alpha - \beta x$ é igual a

$$\sigma^2 \cdot \left(1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}^2}\right)$$

e portanto que a variável

$$\frac{y - \alpha - \beta x}{S\sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}^2}}}$$

tem lei de Student T_{n-2}.

Um intervalo de confiança de nível $1 - \varepsilon$ para o valor $y = a + bx$ é portanto

$$\boxed{y = \alpha + \beta x \pm t_{1-\varepsilon/2} \cdot S\sqrt{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}^2}}}$$

onde $t_{1-\varepsilon/2}$ é o quantil da lei de Student $T(n-2)$.

Observe que, como esperado, o intervalo cresce quando x se afasta da média $\overline{x}$ dos valores utilizados na regressão.

Coeficiente de correlação de Spearman

Outra medida da correlação entre x e y é obtida ao substituir, na definição de Pearson, os valores das observações pelas respetivas ordens,

ou seja, se x'_k é a ordem de x_k (1 se é o menor dos x_i's, 2 se é o segundo, ...) e se y'_k é a ordem de y_k,

então o coeficiente de correlação de Spearman é

$$\boxed{\rho = \frac{S^2_{x'y'}}{S_{x'x'}\, S_{y'y'}}}$$

que também pode ser calculado pela fórmula

$$\rho = 1 - \frac{6}{n^3 - n} \sum_{k=1}^{n} d_k^2$$

onde

$$d_k = y'_k - x'_k$$

Correlação de ordem

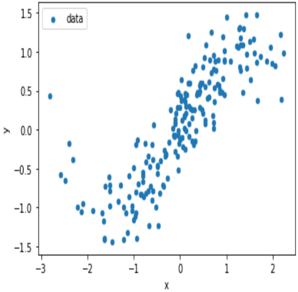
Um valor do coeficiente de correlação de Spearman próximo de

$$\rho \pm 1$$

é sinal de correlação monótona (crescente ou decrescente), e não necessariamente linear!, entre as duas variáveis.

Não é sensível aos outliers.

É mais pesado computacionalmente, pois ordenar é uma tarefa que demora.



Por exemplo, estes dados têm um $\rho \simeq 0.887$ com um p-value $p \simeq 2.1 \times 10^{-68}$.

R.A.

Diagramas de dispersão.

Mínimos quadrados.

Regressão linear.

Reta de regressão.

Coeficiente de determinação, variabildade explicada.

Coeficiente de correlação de Pearson.

Teste F de Fisher-Snedecor.

Coeficiente de correlação de Spearman.