O coeficiente de correlação

A correlação mede a direção e a força da relação linear entre duas variáveis quantitativas, sendo usualmente escrita como r.

Suponha que dispomos de dados sobre as variáveis x e y de n indivíduos. Para o primeiro indivíduo, os valores são x1 e y1; para o segundo, x2 e y2, e assim por diante. As médias e os desvios padrões dessas duas variáveis são μ_x e σ_x para os valores x e μ_y e σ_y para os valores y. A correlação r entre x e y é:

$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left[\left(\frac{x_i - \mu_x}{\sigma_x} \right) * \left(\frac{y_i - \mu_y}{\sigma_y} \right) \right]$$

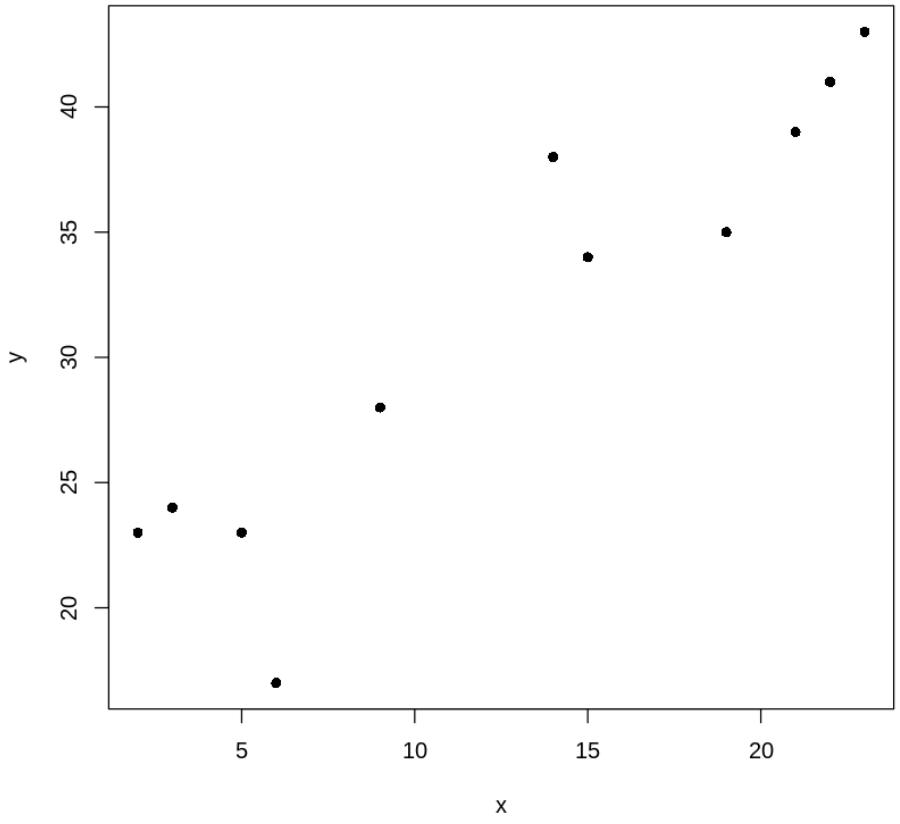
O valor de r varia de -1 a +1, onde os extremos indicam correlação perfeita e 0 significa nenhuma correlação.

O sinal é negativo quando valores grandes de uma variável estão associados a valores pequenos da outra e positivo se ambas as variáveis tendem a ser grandes ou pequenas simultaneamente.

- $|r| < 0{,}3$: Nenhuma correlação ou muito fraca
- $0{,}3 \leq |r| < 0{,}5$: Correlação fraca
- $0{,}5 \leq |r| < 0{,}7$: Correlação moderada
- $0{,}7 \leq |r|$: Correlação forte

As funções `cor` e `cor.test` podem ser usadas para calcular a correlação entre dois ou mais vetores.

```
> x <- c(2, 3, 3, 5, 6, 9, 14, 15, 19, 21, 22, 23)
> y <- c(23, 24, 24, 23, 17, 28, 38, 34, 35, 39, 41, 43)
> plot(x, y, pch=16)
```



Vamos realizar um teste de correlação entre os dois vetores

```
> cor.test(x, y)

        Pearson's product-moment correlation

data:  x and y
t = 7.8756, df = 10, p-value = 1.35e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7575203 0.9799783
sample estimates:
      cor
0.9279869
```

Para mudar o nível de confiança do teste

```
> cor.test(x, y, conf.level = 0.8)

        Pearson's product-moment correlation

data:  x and y
t = 7.8756, df = 10, p-value = 1.35e-05
alternative hypothesis: true correlation is not equal to 0
80 percent confidence interval:
 0.8386211 0.9687074
sample estimates:
      cor
0.9279869
```

Múltiplo R, R^2 e R^2 ajustado

- **R múltiplo:** O coeficiente de correlação múltipla entre três ou mais variáveis.
- **R-Quadrado:** É calculado como (Múltiplo R)^2 e representa a proporção da variância na variável resposta de um modelo de regressão que pode ser explicada pelas variáveis preditoras. Este valor varia de 0 a 1.

Generalizando, podemos calcular o R^2 de duas maneiras:

$$r^2 = \frac{variação\ explicada}{variação\ total} = \frac{\sum(y_{est} - \mu_y)^2}{\sum(y - \mu_y)^2}$$

$$r = \frac{\sum(x - \mu x)(y - \mu y)}{\sqrt{\sum(x - \mu x)^2}\sqrt{\sum(y - \mu y)^2}}$$

Na prática, estamos frequentemente interessados no valor de R ao quadrado porque nos diz quão úteis são as variáveis preditoras na previsão do valor da variável de resposta.

No entanto, cada vez que adicionamos uma nova variável preditora ao modelo, é garantido que o R-quadrado aumentará, mesmo que a variável preditora não seja útil.

O R-quadrado ajustado é uma versão modificada do R-quadrado que se ajusta ao número de preditores em um modelo de regressão. É calculado como:

$$r^2 ajustado = 1 - \frac{(1 - r^2) * (n - 1)}{n - k - 1}$$

O R-quadrado ajustado compara o poder explicativo dos modelos de regressão que contêm diferentes números de preditores (k).

Suponha que você compare um modelo de cinco preditores que tem um R-quadrado mais alto a um modelo com um preditor. O modelo de cinco preditores têm um R-quadrado mais alto por que é melhor? Ou o R-quadrado é mais alto porque tem mais preditores? Basta comparar os valores do R-quadrado ajustados para descobrir!

O R-quadrado ajustado é uma versão modificada do R-quadrado que foi ajustada para o número de preditores no modelo. O R-quadrado ajustado aumenta somente se o novo termo melhorar o modelo mais do que seria esperado pelo acaso. Ele diminui quando um preditor melhora o modelo menos do que o esperado por acaso. O R-quadrado ajustado pode ser negativo, mas geralmente não é. É sempre menor que o R-quadrado.

Case: O desempenho dos fundos mútuos

Diversos fundos mútuos comparam seu próprio desempenho com o de uma referência (benchmark), um índice dos retornos de todos os papeis que esse tipo de fundo compra. Por exemplo, o Vanguard International Growth Fund estabelece como uma referência, o índice Morgan Stanley para a Europa, a Austrália e o Extremo Oriente (EAFE-Europe, Australia and Far East), que mede o desempenho das ações fora dos Estados Unidos. Apresentam-se a seguir os retornos percentuais do fundo em comparação com os do índice EAFE para o período entre 1982 até 2000.

Ano	Fundo	EAFE	Ano	Fundo	EAFE
1982	5,27	-0,86	1992	-5,79	-11,85
1983	43,08	24,61	1993	44,74	32,94
1984	-1,02	7,86	1994	0,76	8,06
1985	56,94	56,72	1995	14,89	11,55
1986	56,71	69,94	1996	14,65	6,36
1987	12,48	24,93	1997	4,12	2,06
1988	11,61	28,59	1998	16,93	20,33
1989	27,76	10,80	1999	26,34	27,30
1990	-12,05	-23,20	2000	-8,60	-13,96
1991	4,74	12,50			

Faça um diagrama de dispersão adequado para fazer previsões dos rendimentos do fundo baseando-se nos rendimentos do EAFE. Nesse diagrama, existe um padrão bem definido de linha reta? Quão forte é esse padrão (forneça uma medida numérica)? Há nele outliers em relação ao padrão linear?

Referências

MOORE et al., A prática da estatística empresarial como usar dados para tomar decisões. Tradução Luís Antônio Fajardo, Rio de Janeiro, LTC. 2006

https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor.test

http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r

https://statsandr.com/blog/correlation-coefficient-and-correlation-test-in-r/