相關 (Correlation)

大數據分析

- R/Python/Julia/SQL程式設計與應用 (R/Python/Julia/SQL Programming and Application)
- 資料視覺化 (Data Visualization)
- 機器學習 (Machine Learning)
- 統計品管 (Statistical Quality Control)
- 最佳化 (Optimization)

李明昌博士
alan9956@gmail.com
http://rwepa.blogspot.com/

大綱

- 1.相關簡介
- 2.雙變數的相關係數
- 3.繪製散佈圖
- 4.偏相關

1.相關簡介

相關

- 相關 (Correlation) 表示變數間相互發生之關聯，通常以線性相關為主。
- 分析兩組資料間之相關，稱之為簡單相關；若是分析多組資料間之相關，則稱之為複相關 (Multiple Correlation)。
- 簡單相關有二種方式：1. 繪製資料散佈圖 2.計算簡單相關係數（包括相關程度大小及正負之數值）。
- 簡單相關係數之計算公式為：
  - 母體相關係數=$\frac{\text{共變異數(Covariance)}}{\text{標準差}_X \times \text{標準差}_X} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$
  - 樣本相關係數 $\gamma = \frac{S_{xy}}{S_x S_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$

相關係數特性

- 相關係數值介於-1到+1之間，$-1 \le \gamma \le 1$
- 相關係數值其情況可有下列三種：
  1. $\gamma = 0$ 無線性相關，可能有非線性關係
  2. $\gamma > 0$ 正相關
  3. $\gamma < 0$ 負相關
- 當相關係數之絕對值小於0.3 時，為低度相關。
- 絕對值介於 0.3~0.7時，為中度相關。
- 達到 0.7~0.8時，為高度相關。
- 若達到 0.8以上時，即為非常高度相關。

2.雙變數的相關係數

雙變數的相關係數

- 考慮 marketing.csv 銷售資料集
- https://github.com/rwepa/DataDemo/blob/master/marketing.csv
- 計算相關係數並進行檢定，其虛無假設與對立假設為：
  - H_0: $\rho=0$（無關）
  - H_1: $\rho\neq0$（相關）

相關係數-SPSS

- 分析 \ 相關 \ 雙變數

相關係數-SPSS (續)

- 兩個星號（**）表示於α=0.01之顯著水準下兩者顯著相關，其顯著性為0.000。
- 一個星號（*）而已，表示於α=0.05之顯著水準下兩者顯著相關；若無星號則表示兩者無顯著相關。

相關

		youtube	facebook	newspaper	sales
youtube	皮爾森 (Pearson) 相關	1	.062	.057	.782**
	顯著性（雙尾）		.386	.426	.000
	N	200	199	200	200
facebook	皮爾森 (Pearson) 相關	.062	1	.352**	.582**
	顯著性（雙尾）	.386		.000	.000
	N	199	199	199	199
newspaper	皮爾森 (Pearson) 相關	.057	.352**	1	.228**
	顯著性（雙尾）	.426	.000		.001
	N	200	199	200	200
sales	皮爾森 (Pearson) 相關	.782**	.582**	.228**	1
	顯著性（雙尾）	.000	.000	.001	
	N	200	199	200	200

**. 相關性在 0.01 層上顯著（雙尾）。

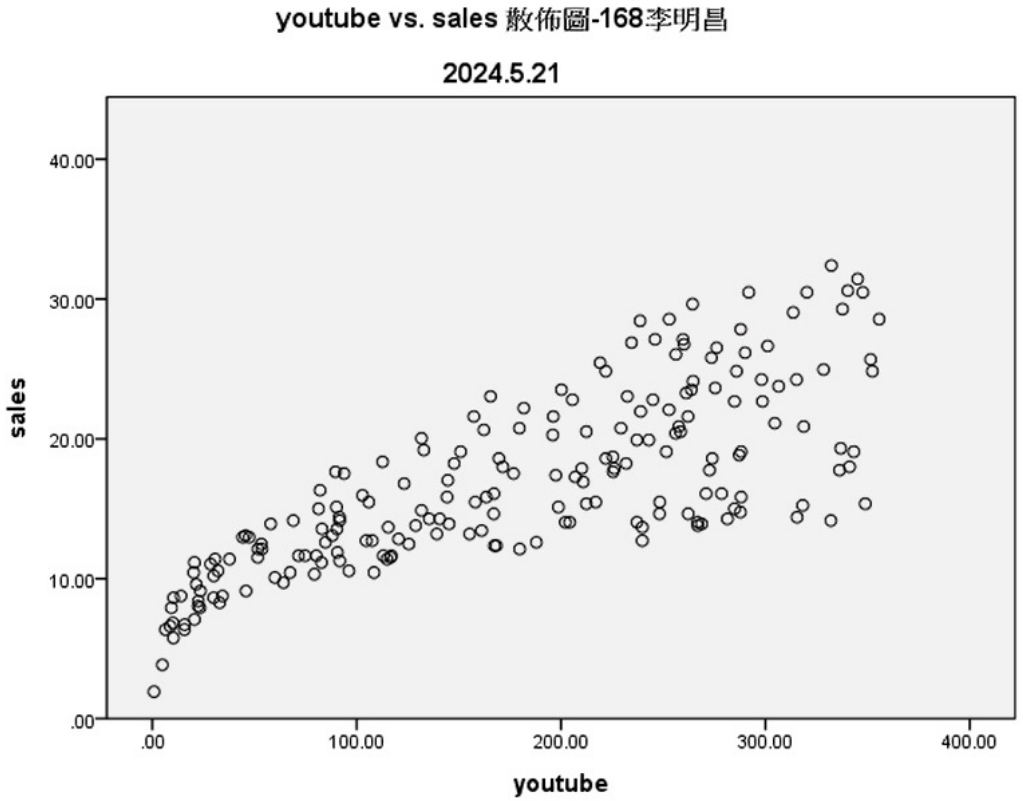
3.繪製散佈圖

繪製資料散佈圖

- 散佈圖通常用以探討兩數值資料之相關情況。例：
  - 廣告費(X)與銷售量(Y)之關係
  - 年齡與所得之關係
  - 所得與購買能力之關係
  - 每月所得與信用分數之關係
- 在X軸之資料稱為自變數；Y軸之資料稱為因變數（依變 數）；利用散佈圖即可判讀出：當X軸資料變動後，對Y 軸資料之影響程度。例：隨廣告費逐漸遞增，銷售量將如何變化？

散佈圖-SPSS

- 統計圖 \ 圖表建置器 \ 散佈圖 / 點形圖 \ 簡易散佈圖



4.偏相關

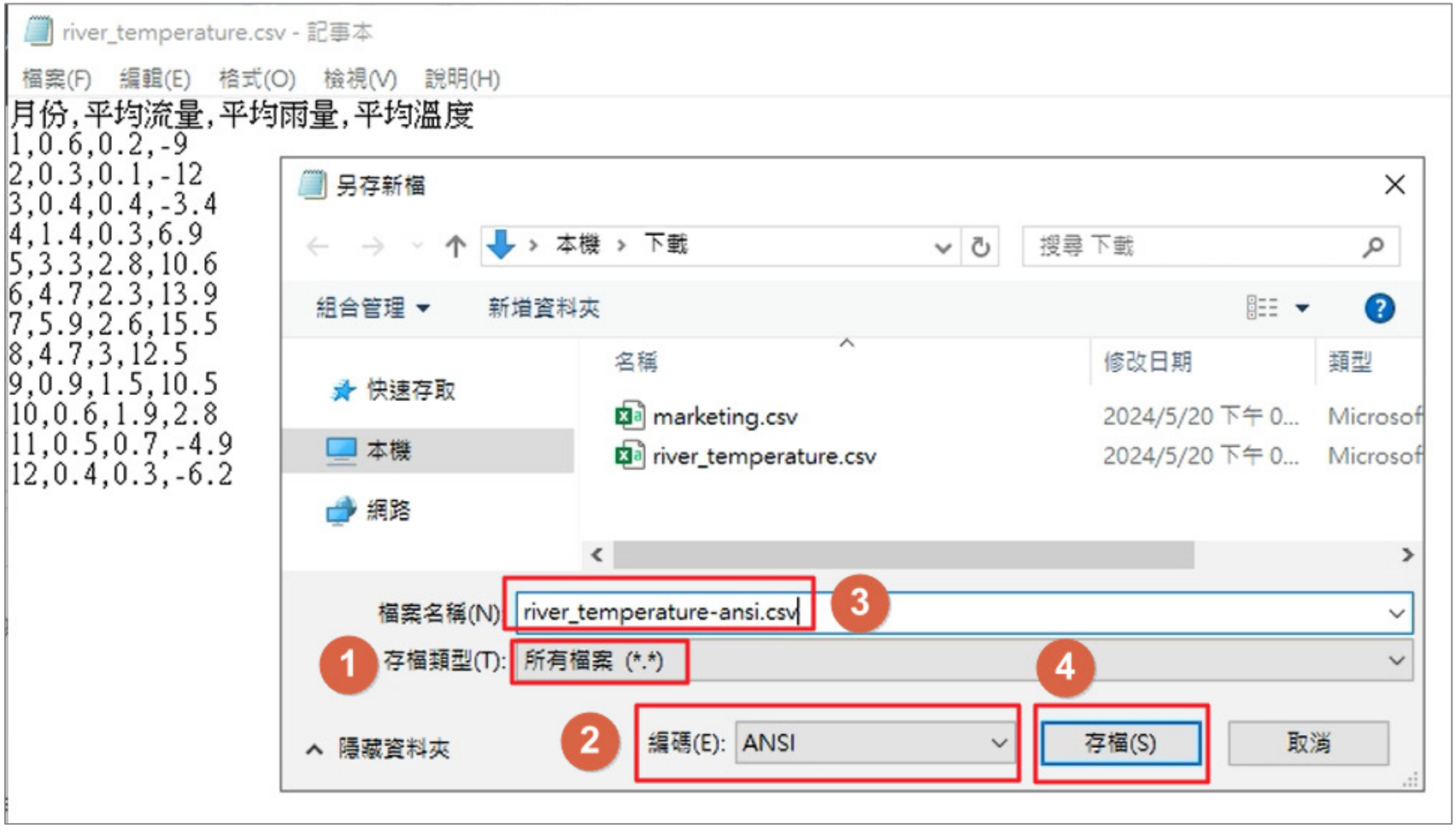
偏相關 (Partial Correlation)

- 真實世界的很多情況，不是簡單的兩個變數就能解釋清楚。且其間各變數相互牽扯，彼此間夾雜很多相互影響力。結果使得我們無法看清某兩個變數間的真正關係。
- 偏相關就是在其他變數固定的條件下（排除第三變數影響），而去檢定兩組變數間是否有關係。由於排除了其他變數之影響，故又稱為「淨相關」。
- 虛無假說(Null hypothesis)：H_0: $\gamma = 0$ 兩變數之間無淨相關
  對立假說(Alternative hypothesis)： H_0: $\gamma \neq 0$ 兩變數之間有淨相關

偏相關

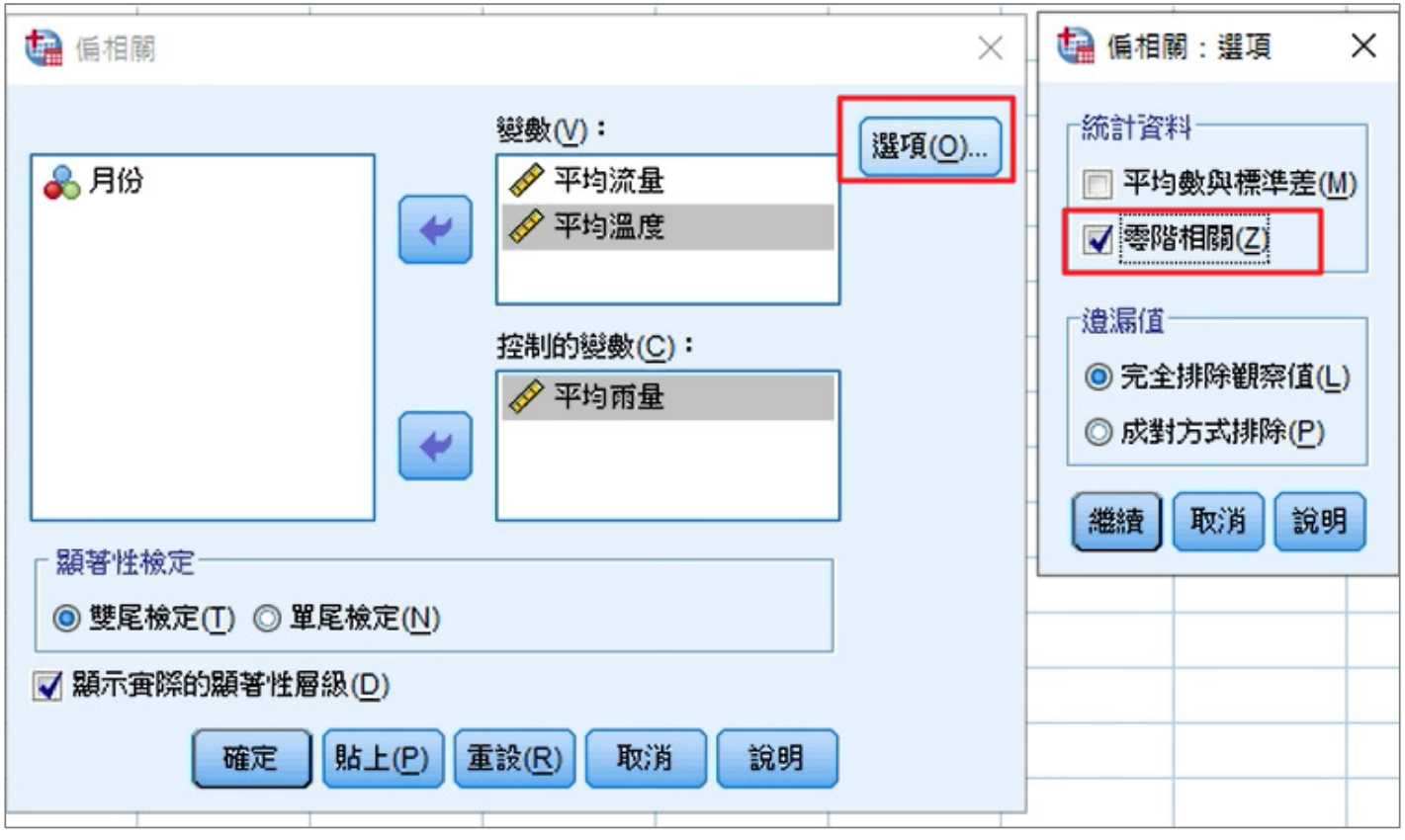
- 下載：https://github.com/rwepa/DataDemo/blob/master/river_temperature.csv
- 分析河水溫度與河水流量之間的相關關係，排除雨量變數。
- 在沒有控制任何變項，看單純兩個變項之間的關係時，稱為零階相關 (zero-order correlations)。
- 當控制第三個變項後，再看兩個變項之間的關係，稱為一階淨相關 (first-order correlation)，即上述偏相關。

river_temperature.csv



偏相關-SPSS

- 分析 \ 相關 \ 偏相關



偏相關-SPSS (續)

相關

控制變數			平均流量	平均溫度	平均雨量
-無-[a]	平均流量	相關	1.000	.819	.831
		顯著性（雙尾）	.	.001	.001
		df	0	10	10
	平均溫度	相關	.819	1.000	.848
		顯著性（雙尾）	.001	.	.000
		df	10	0	10
	平均雨量	相關	.831	.848	1.000
		顯著性（雙尾）	.001	.000	.
		df	10	10	0
平均雨量	平均流量	相關	1.000	.388	
		顯著性（雙尾）	.	.238	
		df	0	9	
	平均溫度	相關	.388	1.000	
		顯著性（雙尾）	.238	.	
		df	9	0	

a. 儲存格包含零階皮爾森 (Pearson) 相關。

謝謝您的聆聽

Q & A

李明昌

alan9956@gmail.com

http://rwepa.blogspot.tw/