

# 迴歸 (Regression)

## 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.線性迴歸簡介
- 2.迴歸分析與繪圖
- 3.複迴歸
- 4.補充篇：R demo
- 5.課程回顧

# 1.線性迴歸簡介

# 迴歸

- 迴歸分析 (Regression Analysis)是以一個或一個以上自變數（預測變項， $X_i$ ），預測一個數值型因變數（被預測變項， $Y$ ）。
- 因變數如果是類別型變數，則稱為邏輯斯迴歸（Logistic Regression）
- 若只有一個自變數稱為簡單迴歸；若使用一組自變數則稱為**多元迴歸**或複迴歸。
- 一般簡單迴歸強調資料具有線性趨勢。
- SPSS的迴歸分析，可獲致很多相關之統計數字。如：相關係數、判定係數、以F檢定判斷因變數與自變數間是否有迴歸關係存在、以t檢定判斷各迴歸係數是否不為0、計算迴歸係數之信賴區間、計算殘差與繪圖。
- 公式推導：[https://github.com/rwepa/DataDemo/blob/master/regression\\_01.pdf](https://github.com/rwepa/DataDemo/blob/master/regression_01.pdf)
- Excel YouTube 示範：[https://youtu.be/i5\\_urp8XzEs](https://youtu.be/i5_urp8XzEs)

# 迴歸模式 (Regression Model)

考慮 X 與 Y 兩個隨機變數，其中的 X 表示「自變數」 independent variables，Y 表示「依變數」 dependent variables， $\varepsilon$  表示誤差項，迴歸方程式表示如下：

$$Y = \alpha + \beta X + \varepsilon$$

上式必須假設以下基本條件：

(1).  $Y_i$  是獨立的常態分佈  $N(\alpha + \beta X_i, \sigma^2), i = 1, 2, \dots, n$

- $\alpha, \beta$  表示迴歸係數 (regression coefficients)
- $H_0: \beta = 0$   
 $H_1: \beta \neq 0$  (研究者目標)

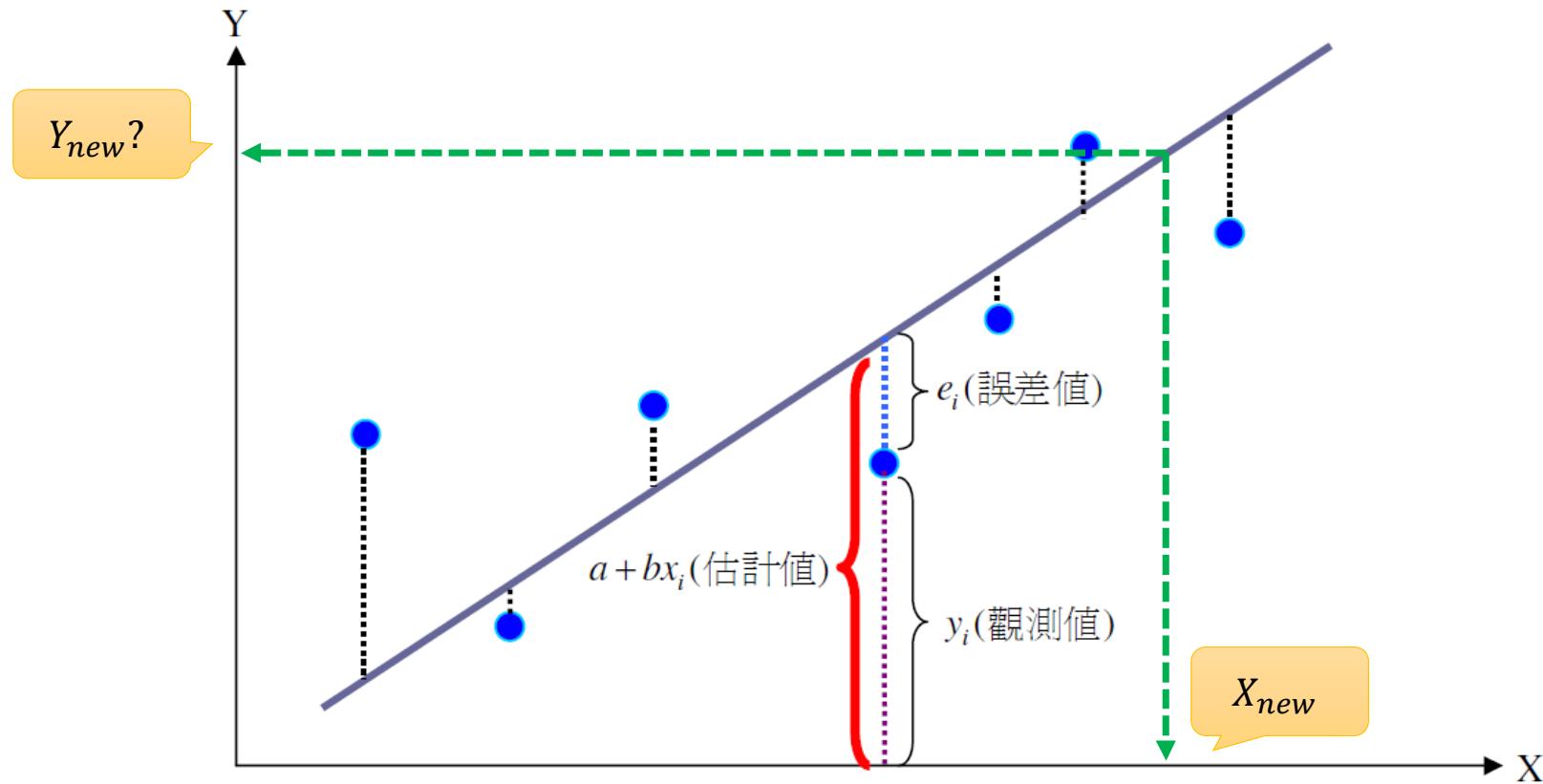
(2).  $\varepsilon_i$  是獨立的常態分佈  $N(0, \sigma^2), i = 1, 2, \dots, n$

即 Y 的估計值為  $\hat{y} = a + bx$ ，其中<sup>^</sup>發音為 hat，而估計值的誤差

$$e_i = \text{觀測值(實際值)} - \text{估計值} = y_i - \hat{y}_i$$

- 迴歸三大假設：
- 1. 常態分配
- 2. 獨立性
- 3. 變異數同質性 ( $\sigma^2$ )

## 迴歸模型



# 最小平方法

重點 2. 最小平方法 Least Squares Method :

考慮  $\text{Min} \left\{ \sum_{i=1}^n e_i^2 \right\} = \text{Min} \left\{ \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right\} = \text{Min} \left\{ \sum_{i=1}^n (y_i - a - b x_i)^2 \right\}$ ，將左式分別對  $a, b$  取微分，並令上式微分等於零，參考以下說明，即可解出  $a, b$

$$\text{令 } w = \sum_{i=1}^n (y_i - a - b x_i)^2$$

解二元一次聯立方程式

$$\frac{dw}{da} = 2 \left( \sum_{i=1}^n (y_i - a - b x_i) \right) \times (-1) = 0$$

$$\frac{dw}{db} = 2 \left( \sum_{i=1}^n (y_i - a - b x_i) \right) \times (-x_i) = 0$$



$$b = \frac{\sum_{i=1}^n x_i y_i - \frac{\left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left( \sum_{i=1}^n x_i \right)^2}{n}}$$

$$a = \bar{y} - b \bar{x}, \quad \text{其中} \quad \bar{y} = \frac{\sum_{i=1}^n y_i}{n}, \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

# 檢定迴歸模型之變異數分析表

- $H_0: \beta_1 = 0$  (此迴歸模型不具解釋能力)
- $H_1: \beta_1 \neq 0$  (此迴歸模型具解釋能力)
- 如果  $f$  值落在拒絕域，即  $\{f > F_\alpha(1, n - 2)\}$ ，即拒絕  $H_0$ ，即此迴歸模型具有解釋能力。

變異來源	平方和	自由度	均方	$f$ 值
迴歸模型	$SSR$	1	$MSR = \frac{SSR}{1}$	$f = \frac{MSR}{MSE}$
隨機誤差	$SSE$	$n - 2$	$MSE = \frac{SSE}{n - 2} = S^2$	
總和	$SST$	$n - 1$		

- $SST = \sum_{i=1}^n (y_i - \bar{y})^2 = SSR + SSE$
- $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = SST - SSE$
- $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 = SST - SSR$

# 平方和計算

- $SST = \sum_{i=1}^n (y_i - \bar{y})^2 = SSR + SSE$
- $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = SST - SSE$
- $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 = SST - SSR$

## 判定係數

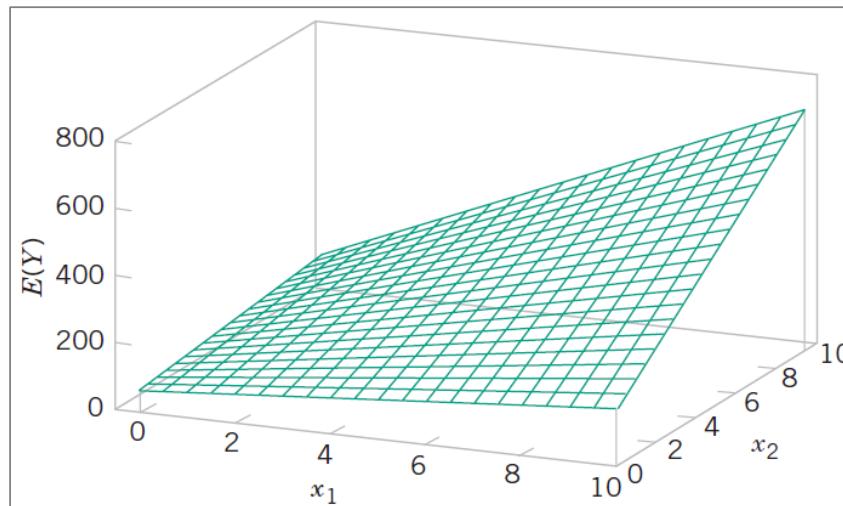
- 考慮  $SSE = 0$  ,  $\frac{SS_R}{SS_T} = 1$  表示總變異  $SST$  完全由迴歸變異解釋，以圖形來表示即資料值剛好可以連成一直線。
- 考慮  $\frac{SS_R}{SS_T}$  接近0 時，總變異值幾乎無法用迴歸模型之變異所解釋，即迴歸模型不具有顯著地解釋能力。
- 判定係數(coefficient of determination) 使用  $R^2$  表示：
- $R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$  ,  $0 \leq R^2 \leq 1$  。
- 當判定係數  $R^2$  越大，則迴歸模型之解釋能力越強， $R^2$  越小，則迴歸模型之解釋能力越弱。

# 常見迴歸模型

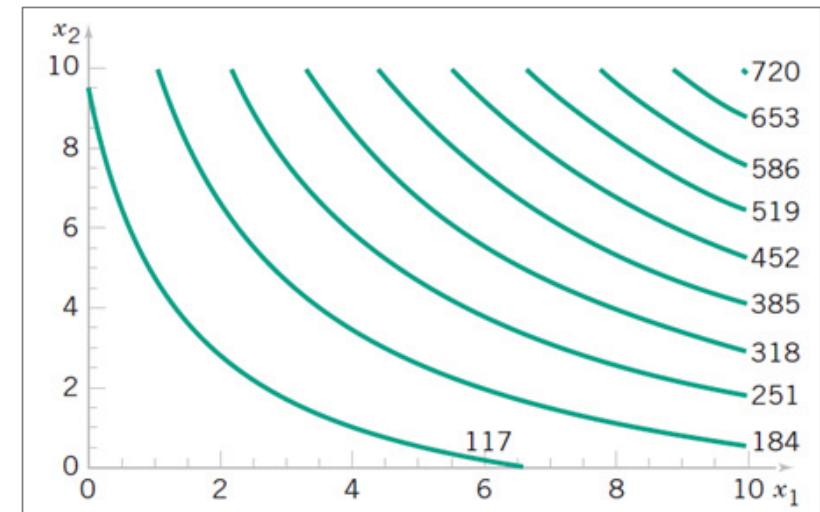
- 簡單迴歸  $y = 10 + 3x$
- 多元迴歸  $y = 10 + 3x_1 + 5x_2$
- 三次多項式模型 (cubic polynomial model)
  - $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$
  - 令  $x_1 = x, x_2 = x^2, x_3 = x^3$ , 則  $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$
- 交互效果 (interaction effect)
  - $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$
  - 令  $x_3 = x_1 x_2$ , 則  $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$
- 二階模型 (second-order model)
  - $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$
  - 同理二階模型亦可轉換為多元迴歸。

## 二階模型範例

$$\bullet Y = 800 + 10x_1 + 7x_2 - 8.5x_1^2 - 5x_2^2 + 4x_1x_2$$



3D regression model plot



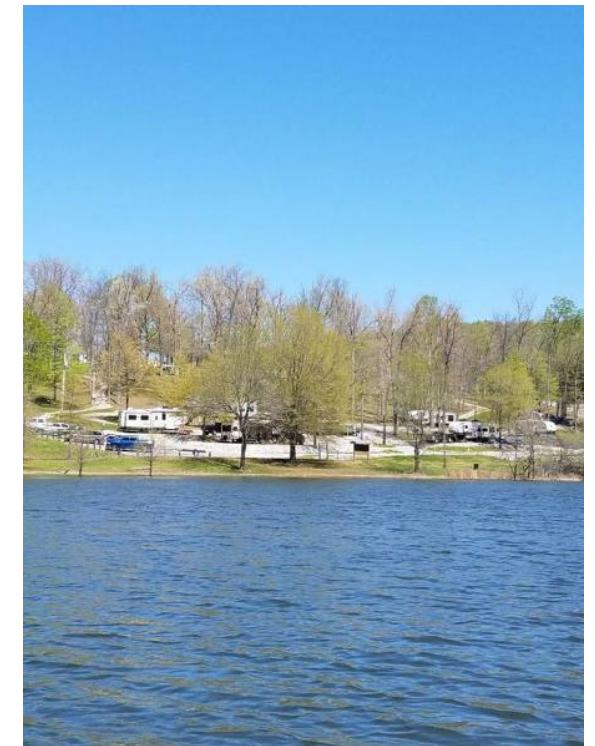
contour plot

參考: <https://industri.fatek.unpatti.ac.id/wp-content/uploads/2019/03/091-Engineering-Statistics-Douglas-C.-Montgomery-George-C.-Runger-Norma-F.-Hubele-Edisi-5-2011.pdf>

## 2.迴歸分析與繪圖

# 年降雨迴歸預測

- 考慮美國肯塔基州-萊星頓市附近Cave Creek :
- 自變數 X : 年降雨 (annual precipitation)
- 因變數 Y : 年逕流 (annual runoff)
- 計算單位皆為英吋 (inches)
- 記錄時間 : 1953~1968,  $n = 16$
- 資料 : Charles T. Haan, Statistical Methods in Hydrology, second edition, 2002 。
- <https://www.amazon.com/Statistical-Methods-Hydrology-Charles-Haan/dp/0813815037>



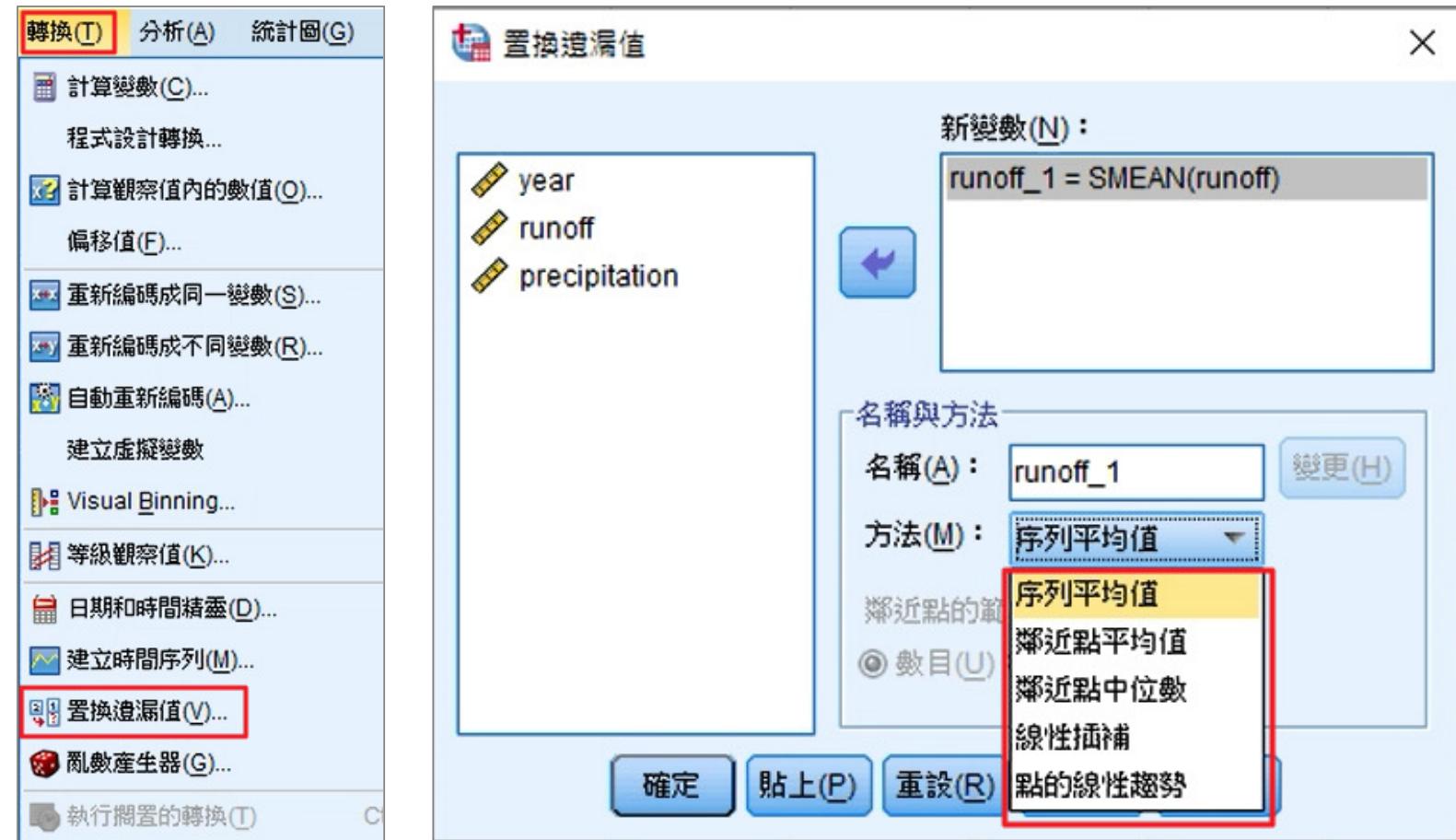
年降雨下載：[https://github.com/rwepa/DataDemo/blob/master/CaveCreek\\_precipitation.csv](https://github.com/rwepa/DataDemo/blob/master/CaveCreek_precipitation.csv)

	A	B	C
1	year	precipitation	runoff
2	1953	42.39	13.26
3	1954	33.48	3.31
4	1955	47.67	5.17
5	1956	50.24	15.5
6	1957	43.28	14.22
7	1958	52.6	21.2
8	1959	31.06	7.7
9	1960	50.02	17.64
10	1961	47.08	22.91
11	1962	47.08	18.89
12	1963	40.89	12.82
13	1964	37.31	11.58
14	1965	37.15	15.17
15	1966	40.38	10.4
16	1967	45.39	18.02
17	1968	41.03	16.25



# 遺漏值的資料預處理

- 轉換 \ 置換遺漏值



# 分析 \ 迴歸 \ 線性



## 線性迴歸



# 線性迴歸-統計資料



# 線性迴歸-選項



如果模型的常數項不顯著，則可以取消此選項。

# 相關

## 迴歸

描述性統計資料

	平均數	標準偏差	N
runoff	14.0025	5.45265	16
precipitation	42.9406	6.16472	16

		相關	
皮爾森 (Pearson) 相關	runoff	1.000	.639
	precipitation	.639	1.000
顯著性 (單尾)	runoff	.	.004
	precipitation	.004	.
N	runoff	16	16
	precipitation	16	16

顯著性 0.004 小於  $\alpha$ ，表示顯著相關，繼續進行後續迴歸分析。

變數已輸入/已移除<sup>a</sup>

模型	變數已輸入	變數已移除	方法
1	precipitation <sup>b</sup>	.	Enter

a. 應變數: runoff

b. 已輸入所有要求的變數。

# 變異數分析

- 判定係數 (R平方) = 0.408。
- 本例判定係數不是很高，可考慮是否有殘差很大之異常樣本？如果有殘差異常大的樣本，可將其排除後再重算新計算並求得更適當之迴歸模型。

模型摘要<sup>b</sup>

模型	R	R 平方	調整後 R 平方	標準偏斜度錯 誤	變更統計資料				
					R 平方變更	F 值變更	df1	df2	顯著性 F 值變 更
1	.639 <sup>a</sup>	.408	.366	4.34200	.408	9.655	1	14	.008

a. 預測值：(常數) · precipitation

b. 應變數: runoff

變異數分析<sup>a</sup>

模型	平方和	df	平均值平方	F	顯著性
1 回歸	182.030	1	182.030	9.655	.008 <sup>b</sup>
殘差	263.941	14	18.853		
總計	445.971	15			

a. 應變數: runoff

b. 預測值：(常數) · precipitation

- 變異數分析用於判斷因變數 (Y) 與自變數 (X) 之間，是否有顯著之迴歸關係。
- 本例之顯著性  $0.008 < \alpha = 0.05$ ，故其結果為棄卻因變數與自變數間無迴歸關係存在之虛無假設  $H_0 : \beta_1 = 0$ 。

# 迴歸係數

係數<sup>a</sup>

模型	非標準化係數		Beta	T	顯著性
	B	標準錯誤			
1 (常數)	-10.263	7.884		-1.302	.214
precipitation	.565	.182	.639	3.107	.008

a. 應變數: runoff

模型 :  $\text{roundoff} = -10.263 + 0.565 \times \text{precipitation}$

殘差統計資料<sup>a</sup>

	最小值	最大值	平均數	標準偏差	N
預測值	7.2890	19.4609	14.0025	3.48358	16
殘差	-11.50499	6.56841	.00000	4.19477	16
標準預測值	-1.927	1.567	.000	1.000	16
標準殘差	-2.650	1.513	.000	.966	16

a. 應變數: runoff

- t檢定，檢定常數項與迴歸係數是否為0。
- 本例常數項之顯著性(0.214)大於 $\alpha=0.05$ ，結果為接受常數項為0之虛無假設  $H_0: \beta_0 = 0$ 。

# 繪圖

---

# 曲線估計

- 分析 \ 迴歸 \ 曲線估計



# 曲線估計視窗



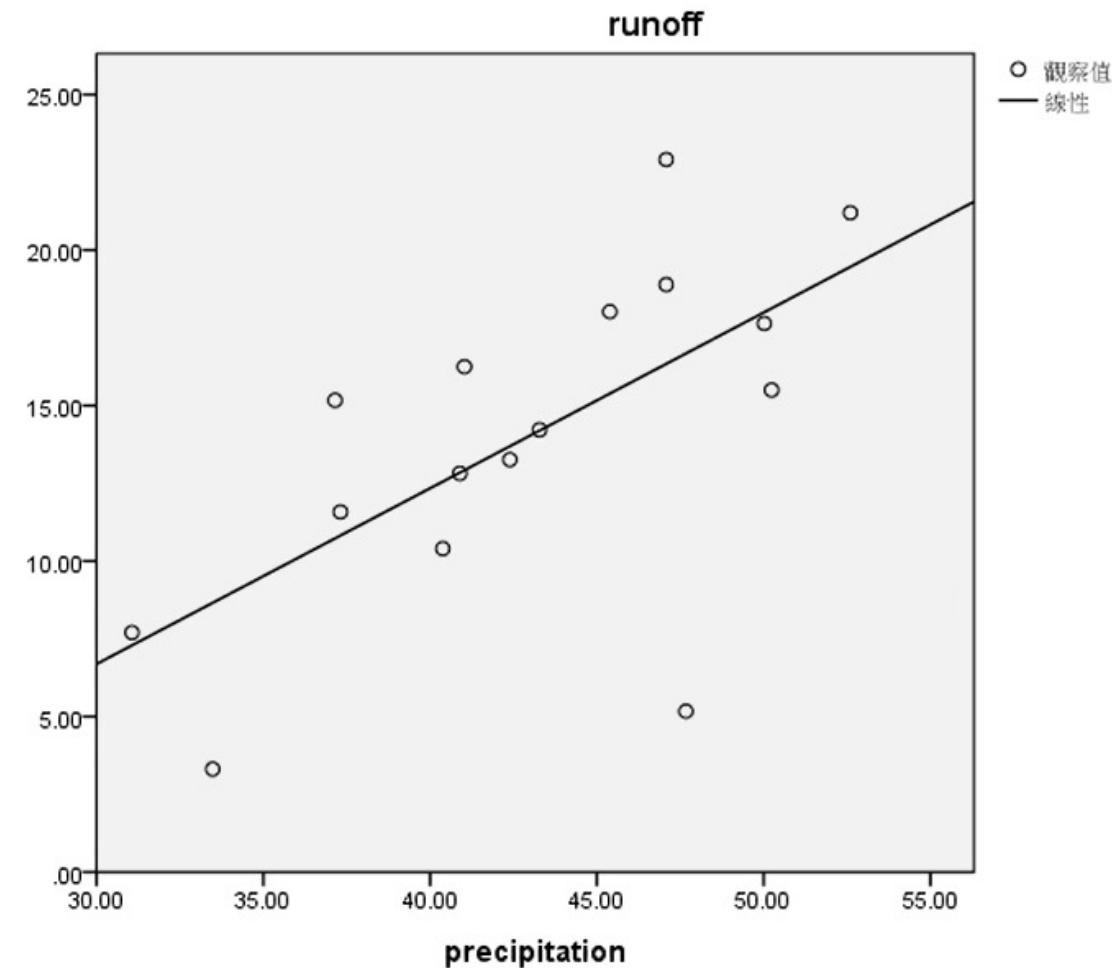
# 自動曲線估計：儲存



新增二個變數

	year	precipitation	runoff	FIT_1	ERR_1
1	1953	42.39	13.26	13.69135	-.43135
2	1954	33.48	3.31	8.65646	-5.34646
3	1955	47.67	5.17	16.67499	-11.50499
4	1956	50.24	15.50	18.12726	-2.62726
5	1957	43.28	14.22	14.19428	.02572
6	1958	52.60	21.20	19.46086	1.73914
7	1959	31.06	7.70	7.28895	.41105
8	1960	50.02	17.64	18.00294	-.36294
9	1961	47.08	22.91	16.34159	6.56841
10	1962	47.08	18.89	16.34159	2.54841
11	1963	40.89	12.82	12.84373	-.02373
12	1964	37.31	11.58	10.82073	.75927
13	1965	37.15	15.17	10.73031	4.43969
14	1966	40.38	10.40	12.55553	-2.15553
15	1967	45.39	18.02	15.38660	2.63340
16	1968	41.03	16.25	12.92284	3.32716

# 迴歸模型繪圖



## 迴歸三大假設

---

- 1.常態分配
- 2.獨立性
- 3.變異數同質性 ( $\sigma^2$ )

# 1. 常態分配 (Shapiro-Wilk常態性檢定)

- $H_0$ : 資料符合常態分配 (研究者希望  $p$  值  $> 0.05$ ，接受  $H_0$ )
- $H_1$ : 資料不符合常態分配
- 分析 \ 描述性統計資料 \ 探索 \ 圖形 \ 常態機率圖附檢定



# Shapiro-Wilk常態性檢定

- p值  $0.916 > 0.05$ ，接受  $H_0$ ，即資料符合常態分配。

常態檢定

	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	統計資料	df	顯著性	統計資料	df	顯著性
runoff	.102	16	.200*	.975	16	.916

\*. 這是 true 顯著的下限。

a. Lilliefors 顯著更正

## 2. 獨立性

- 分析\迴歸\線性\線性迴歸-統計資料\Durbin-Watson 打勾
- 用來檢定觀察值是否獨立（沒有自我相關），Durbin-Watson檢定值分佈在0~4之間，越接近2，觀測值相互獨立的可能性越大。



模型摘要 <sup>b</sup>					
模型	R	R 平方	調整後 R 平方	標準偏斜度錯誤	Durbin-Watson
1	.639 <sup>a</sup>	.408	.366	4.34200	1.156

a. 預測值：(常數), precipitation

b. 應變數: runoff

DW=1.156, 結論是?



### 3. 變異數同質性 ( $\sigma^2$ )

- 轉換 \ 計算變數
- 目標變數：ERR\_1\_平方
- 數丘表示式：ERR\_1\*\*2

year	precipitation	runoff	FIT_1	ERR_1	ERR_1_平方
1953	42.39	13.26	13.69135	-.43135	.19
1954	33.48	3.31	8.65646	-5.34646	28.58
1955	47.67	5.17	16.67499	-11.50499	132.36
1956	50.24	15.50	18.12726	-2.62726	6.90
1957	43.28	14.22	14.19428	.02572	.00
1958	52.60	21.20	19.46086	1.73914	3.02
1959	31.06	7.70	7.28895	.41105	.17
1960	50.02	17.64	18.00294	-.36294	.13
1961	47.08	22.91	16.34159	6.56841	43.14

**計算變數**

目標變數(I)：  
ERR\_1\_平方

= 數值表示式(E)：  
ERR\_1 \*\* 2

類型和標籤(L)...

year  
precipitation  
runoff  
在 CURVEFIT、MO...  
在 CURVEFIT、MO...

# 變異數同質性(續)

變異數分析<sup>a</sup>

模型	平方和	df	平均值平方	F	顯著性
1 迴歸	449.181	1	449.181	.391	.542 <sup>b</sup>
殘差	16075.814	14	1148.272		
總計	16524.995	15			

a. 應變數: ERR\_1\_平方

b. 預測值: (常數), precipitation

- $H_0$ : 變異數具有同質性
- $H_1$ : 變異數沒有同質性
- p值=0.542 >  $\alpha$ ，接受 $H_0$

### 3. 複迴歸

# 複迴歸

- 模型： $Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_k + \varepsilon$
- $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$  (此迴歸模型不具解釋能力)
- $H_1: \beta_1, \beta_2, \dots, \beta_k$ 不全為0 (此迴歸模型具解釋能力)

變異來源	平方和	自由度	均方	$f$ 值
迴歸模型	$SSR$	$k$	$MSR = \frac{SSR}{k}$	$f = \frac{MSR}{MSE}$
隨機誤差	$SSE$	$n - k - 1$	$MSE = \frac{SSE}{n - k - 1}$	
總和	$SST$	$n - 1$		

## marketing.csv



# 相關

描述性統計資料

	平均數	標準偏差	N
sales	16.8488	6.26909	199
youtube	177.0693	102.91219	199
facebook	27.8201	17.80841	199
newspaper	36.5771	26.17073	199

相關

		sales	youtube	facebook	newspaper
皮爾森 (Pearson) 相關	sales	1.000	.782	.582	.231
	youtube	.782	1.000	.062	.061
	facebook	.582	.062	1.000	.352
	newspaper	.231	.061	.352	1.000
顯著性 (單尾)	sales	.	.000	.000	.001
	youtube	.000	.	.193	.196
	facebook	.000	.193	.	.000
	newspaper	.001	.196	.000	.
N	sales	199	199	199	199
	youtube	199	199	199	199
	facebook	199	199	199	199
	newspaper	199	199	199	199

# 迴歸模型-變異數分析

模型摘要									
模型	R	R 平方	調整後 R 平方	標準偏斜度錯 誤	變更統計資料				
					R 平方變更	F 值變更	df1	df2	顯著性 F 值變 更
1	.947 <sup>a</sup>	.898	.896	2.02081	.898	570.190	3	195	.000

a. 預測值：(常數) , newspaper, youtube, facebook

變異數分析 <sup>a</sup>					
模型	平方和	df	平均值平方	F	顯著性
1	迴歸 6985.389	3	2328.463	570.190	.000 <sup>b</sup>
	殘差 796.314	195	4.084		
	總計 7781.703	198			

a. 應變數: sales

b. 預測值：(常數) , newspaper, youtube, facebook

係數 <sup>a</sup>					
模型	非標準化係數		Beta	T	顯著性
	B	標準錯誤			
1	(常數) 3.539	.374		9.460	.000
	youtube .046	.001	.749	32.598	.000
	facebook .189	.009	.537	21.941	.000
	newspaper -.001	.006	-.004	-.147	.883

a. 應變數: sales

## 第2次迴歸模型



先刪除newspaper變數

# 第2次迴歸模型-優化結果

模型摘要

模型	R	R 平方	調整後 R 平方	標準偏斜度錯 誤	變更統計資料				
					R 平方變更	F 值變更	df1	df2	顯著性 F 值變 更
1	.947 <sup>a</sup>	.898	.897	2.01576	.898	859.565	2	196	.000

a. 預測值：(常數), facebook, youtube

變異數分析<sup>a</sup>

模型	平方和	df	平均值平方	F	顯著性
1 回歸	6985.301	2	3492.650	859.565	.000 <sup>b</sup>
殘差	796.402	196	4.063		
總計	7781.703	198			

a. 應變數: sales

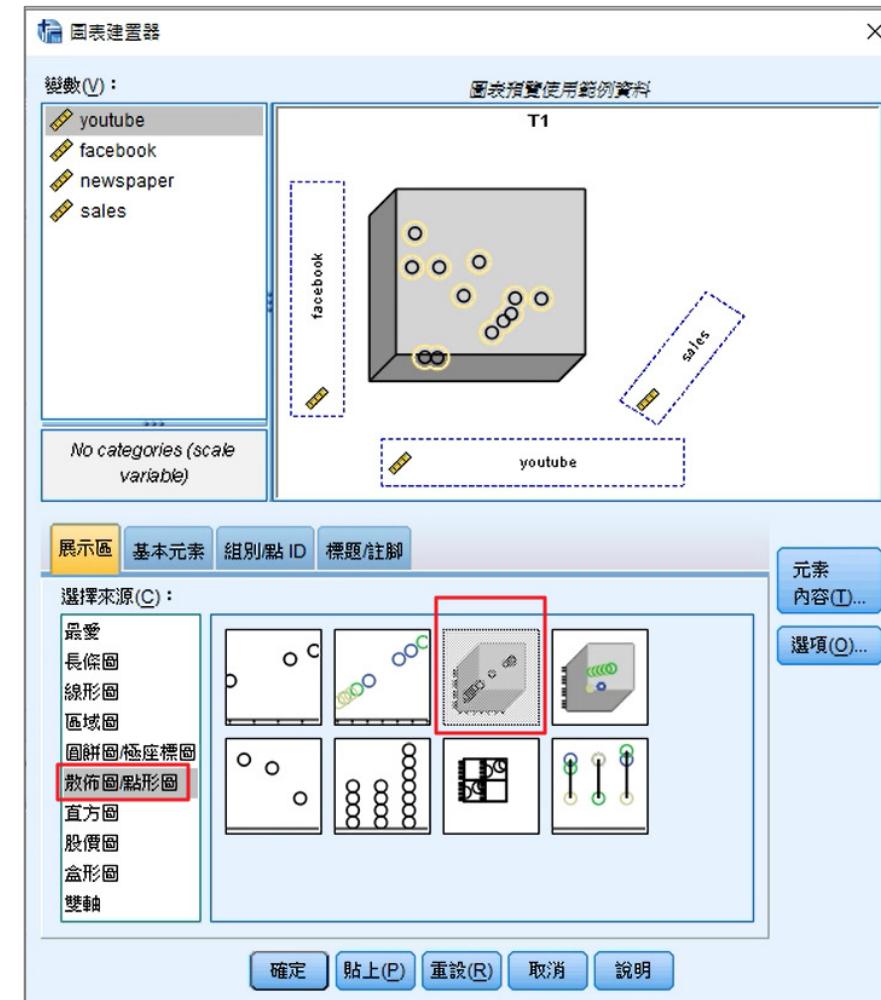
b. 預測值：(常數), facebook, youtube

係數<sup>a</sup>

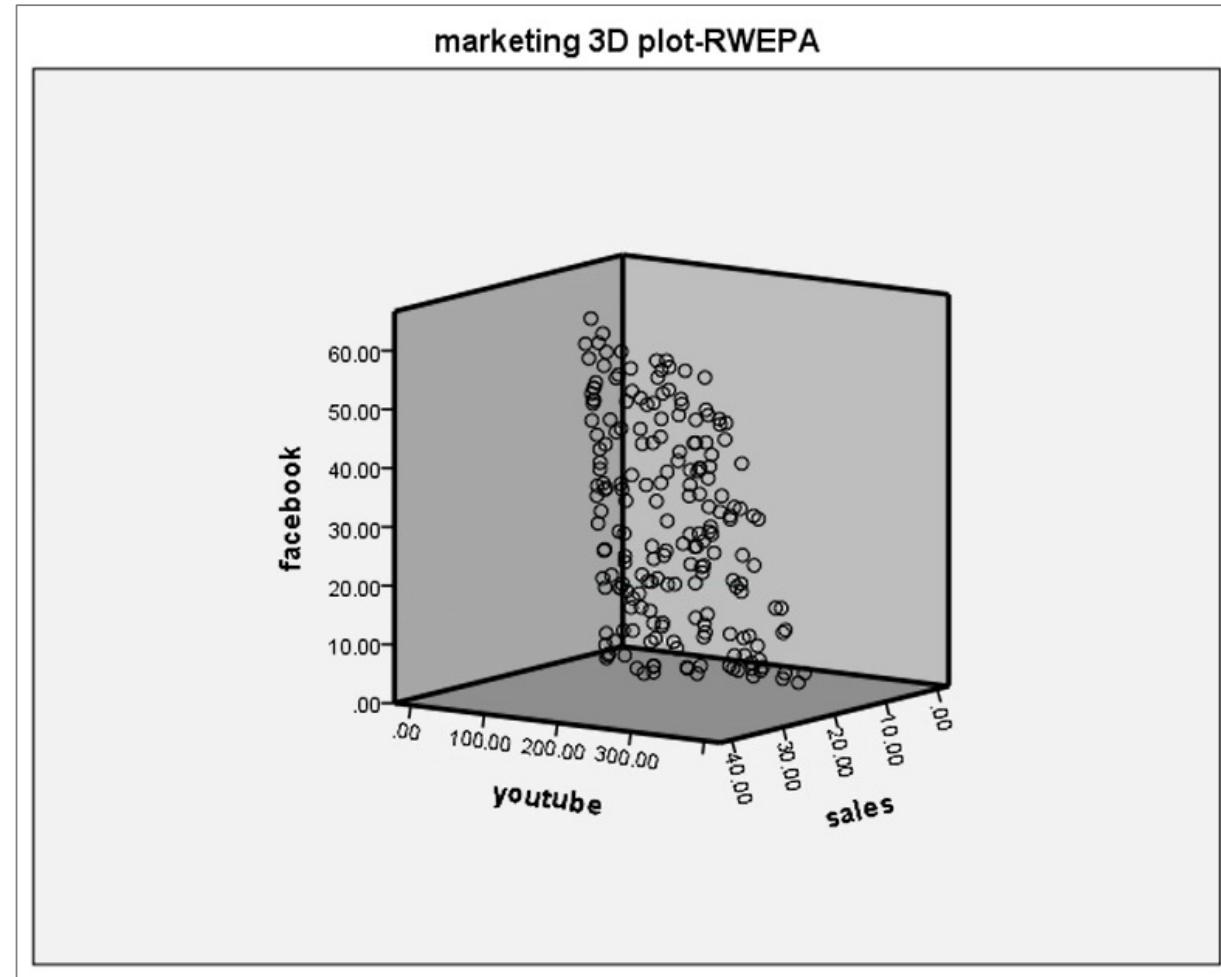
模型	非標準化係數		標準化係數 Beta	T	顯著性
	B	標準錯誤			
1 (常數)	3.521	.353		9.966	.000
youtube	.046	.001	.749	32.703	.000
facebook	.189	.008	.536	23.421	.000

a. 應變數: sales

# 簡易3D散佈圖



# 簡易3D散佈圖-完成圖



## 4.補充篇：R demo

# 獨立性檢定-R

```
> # 模型摘要
> summary(df_lm)

Call:
lm(formula = runoff ~ precipitation, data = df)

Residuals:
    Min      1Q   Median      3Q     Max 
-11.5050 -0.8624  0.2184  2.5697  6.5684 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -10.2625    7.8842  -1.302  0.21404  
precipitation  0.5651    0.1819   3.107  0.00772 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

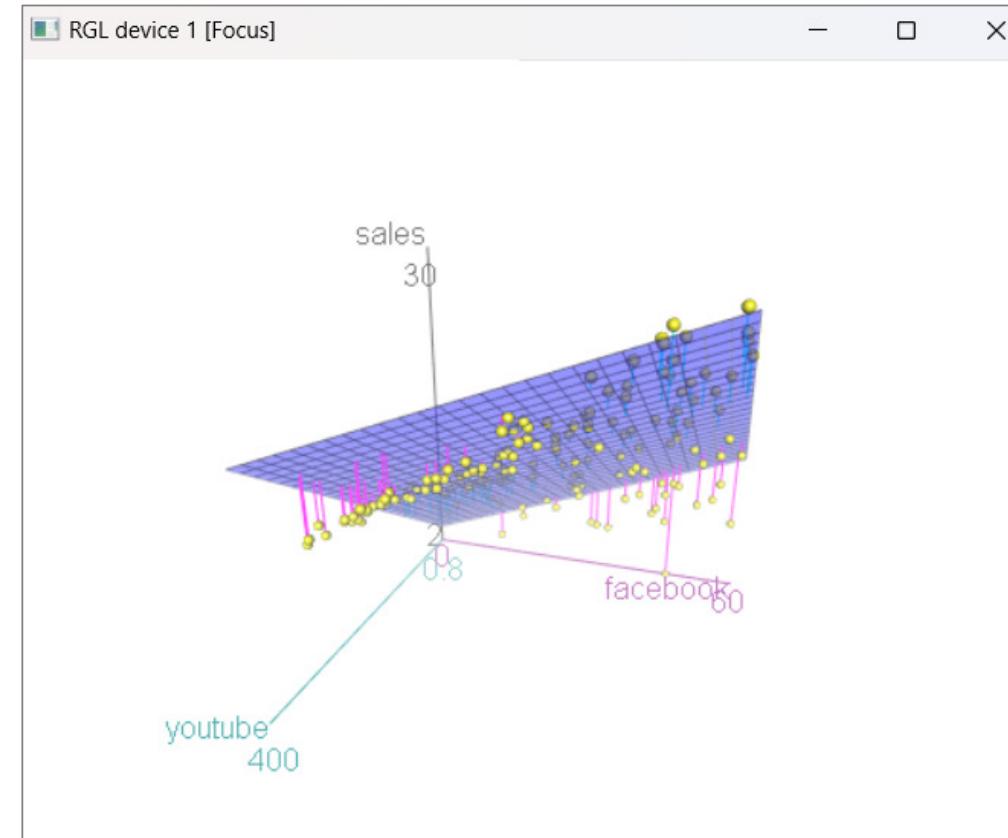
Residual standard error: 4.342 on 14 degrees of freedom
Multiple R-squared:  0.4082,    Adjusted R-squared:  0.3659 
F-statistic: 9.655 on 1 and 14 DF,  p-value: 0.00772

> # 建立 Durbin-Watson 檢定
> set.seed(168)
> durbinWatsonTest(df_lm)
  lag Autocorrelation D-W Statistic p-value
    1          0.4006555     1.156043  0.072
Alternative hypothesis: rho != 0
```

R demo

下載: [https://github.com/rwepa/market\\_survey\\_research/blob/main/regression\\_model.R](https://github.com/rwepa/market_survey_research/blob/main/regression_model.R)

# 迴歸模型-3D繪圖



- 3D自由旋轉
- 放大/縮小

## 5. 課程回顧

1. SPSS敘述統計分析
2. 研究方法與開放資料
3. 平均數估計 (Mean Estimation)
4. 平均數檢定 (Hypothesis Test for Mean)
5. 變異數分析(ANOVA)
6. 相關(Correlation)
7. 迴歸 (Regression)
8. R demo

# 謝謝您的聆聽

## Q & A



李明昌

*alan9956@gmail.com*

**<http://rwepa.blogspot.tw/>**