

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.複迴歸
- 3.Rcmdr demo
- 4.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
- Excel YouTube 示範 : https://youtu.be/i5_urp8XzEs

迴歸模式 (Regression Model)

考慮 X 與 Y 二個隨機變數，其中的 X 表示「自變數」 independent variables， Y 表示「依變數」 dependent variables， ε 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 $N(\alpha + \beta X_i, \sigma^2), i = 1, 2, \dots, n$

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

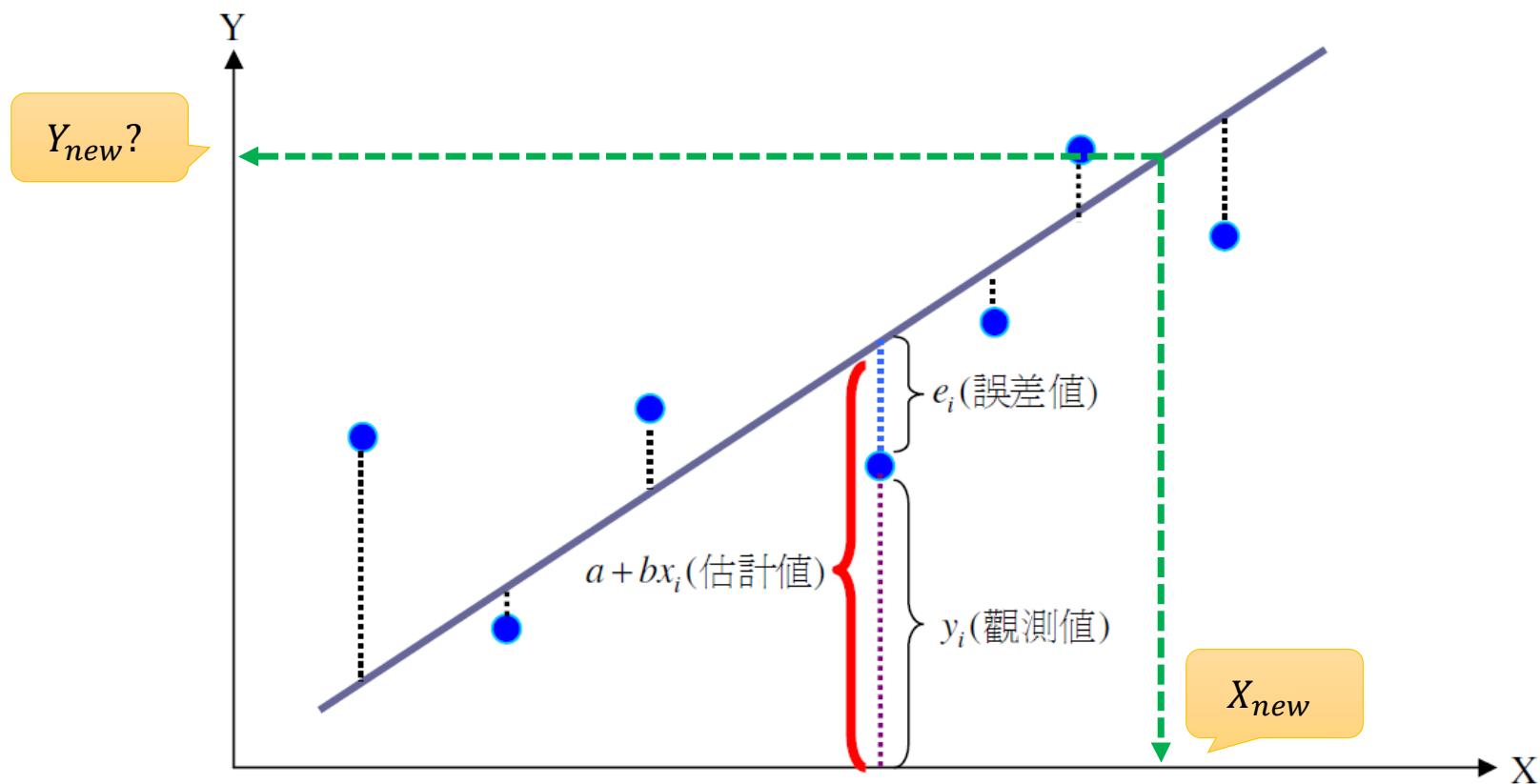
(2). ε_i 是獨立的常態分佈 $N(0, \sigma^2), i = 1, 2, \dots, n$

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

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

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

迴歸模型



最小平方法

重點 2. 最小平方法 Least Squares Method :

考慮 $Min \left\{ \sum_{i=1}^n e_i^2 \right\} = Min \left\{ \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right\} = Min \left\{ \sum_{i=1}^n (y_i - a - b x_i)^2 \right\}$ ，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

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

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

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

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

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

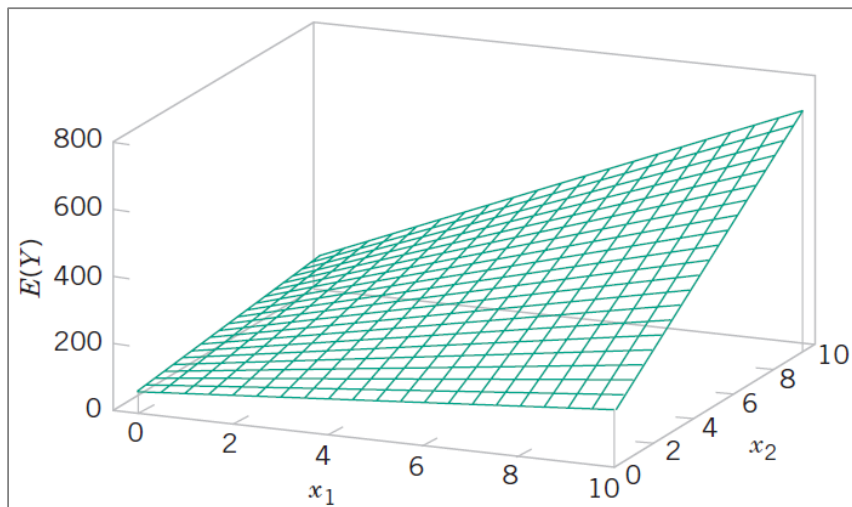
- 考慮 $SSE = 0$, $\frac{SS_R}{SS_T} = 1$ 表示總變異 SST 完全由迴歸變異解釋，以圖形來表示即資料值剛好可以連成一直線。
- 考慮 $\frac{SS_R}{SS_T}$ 接近 0 時，總變異值幾乎無法用迴歸模型之變異所解釋，即迴歸模型不具有顯著地解釋能力。
- 判定係數(coefficient of determination) 使用 R^2 表示：
- $R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$,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

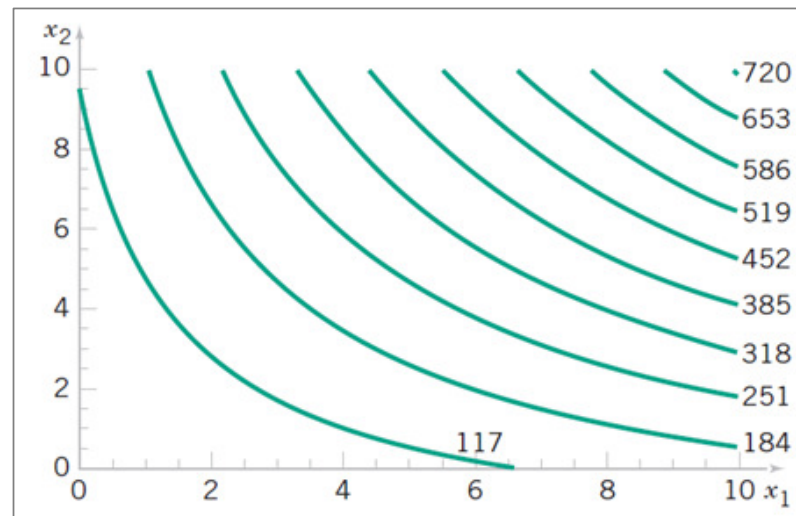
- 簡單迴歸 $y = 10 + 3x$
- 多元迴歸 $y = 10 + 3x_1 + 5x_2$
- 三次多項式模型 (cubic polynomial model)
 - $Y = \beta_0 + \beta_1x + \beta_2x^2 + \beta_3x^3 + \varepsilon$
 - 令 $x_1 = x, x_2 = x^2, x_3 = x^3$, 則 $Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \varepsilon$
- 交互效果 (interaction effect)
 - $Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_{12}x_1x_2 + \varepsilon$
 - 令 $x_3 = x_1x_2$, 則 $Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \varepsilon$
- 二階模型 (second-order model)
 - $Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_{11}x_1^2 + \beta_{22}x_2^2 + \beta_{12}x_1x_2 + \varepsilon$
 - 同理二階模型亦可轉換為多元迴歸。

二階模型範例

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



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. 複迴歸

複迴歸

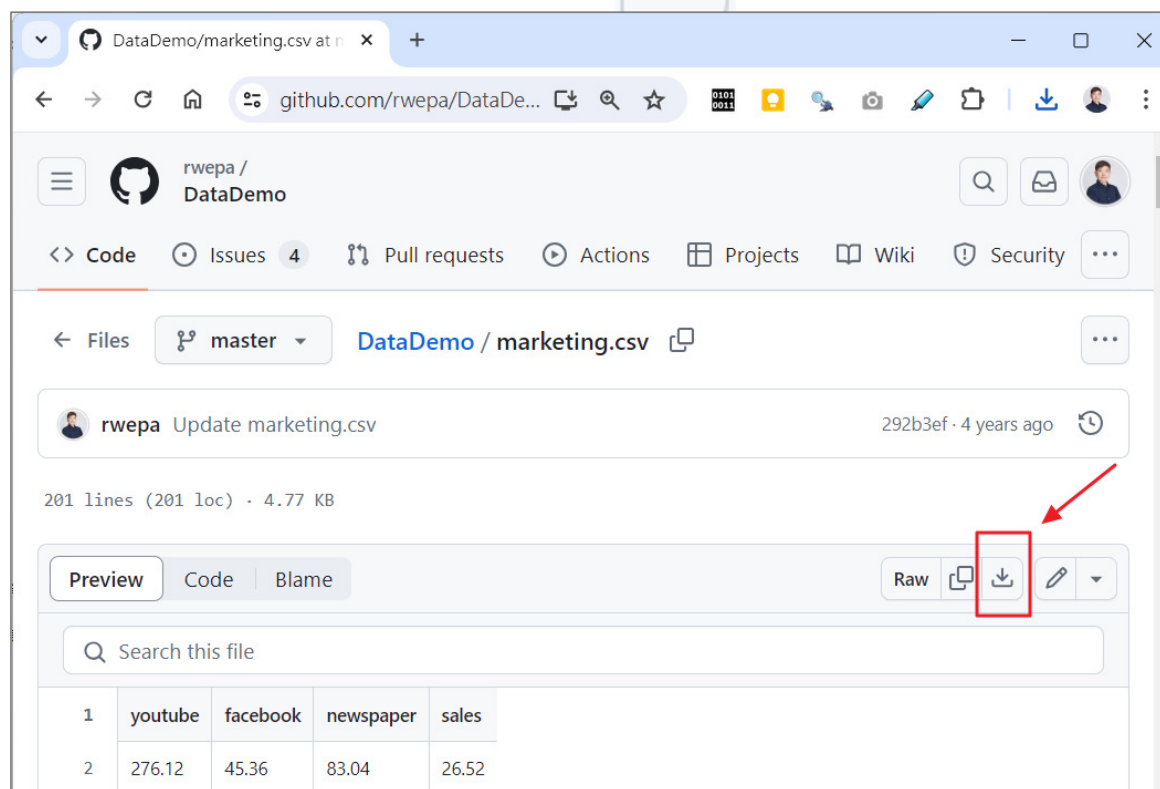
- 模型： $Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_k + \varepsilon$
- $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ （此迴歸模型不具解釋能力）
 $H_1: \beta_1, \beta_2, \dots, \beta_k$ 不全為0 （此迴歸模型具解釋能力）

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	$n - 1$		

3.Rcmdr demo

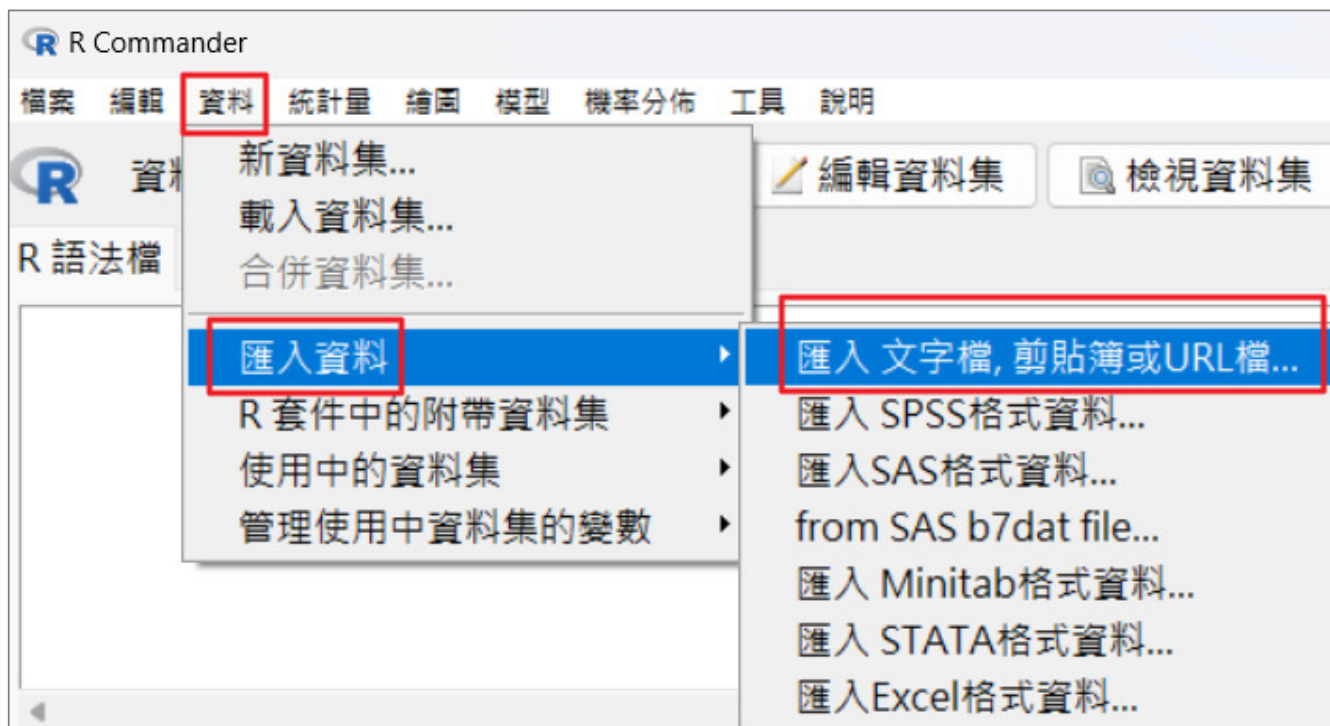
下載 marketing.cav

- <https://github.com/rwepa/DataDemo/blob/master/marketing.csv>
- 按 下載 按鈕 [Download raw file] ，預設儲存在 下載 資料夾。



Rcmdr -

- 啟動Rcmdr→library(Rcmdr)
- 資料 \ 匯入資料 \ 匯入文字檔, 剪貼簿或URL檔...



讀取文字檔

讀取 文字檔, 剪貼簿或URL檔

請輸入資料集名稱：

檔案中的變數名稱：☒

將字串變數轉成因子：☒

遺漏資料標示符號：

資料檔位置

☒ 本機檔案

☐ 剪貼簿

☐ 網際網路位址 URL

欄位分隔字元

☒ 空白鍵

☒ Commas [,]

☐ Semicolons [;]


☐ Tab 鍵


☐ 其他 指定符號：


小數點符號

☒ 點 [.]

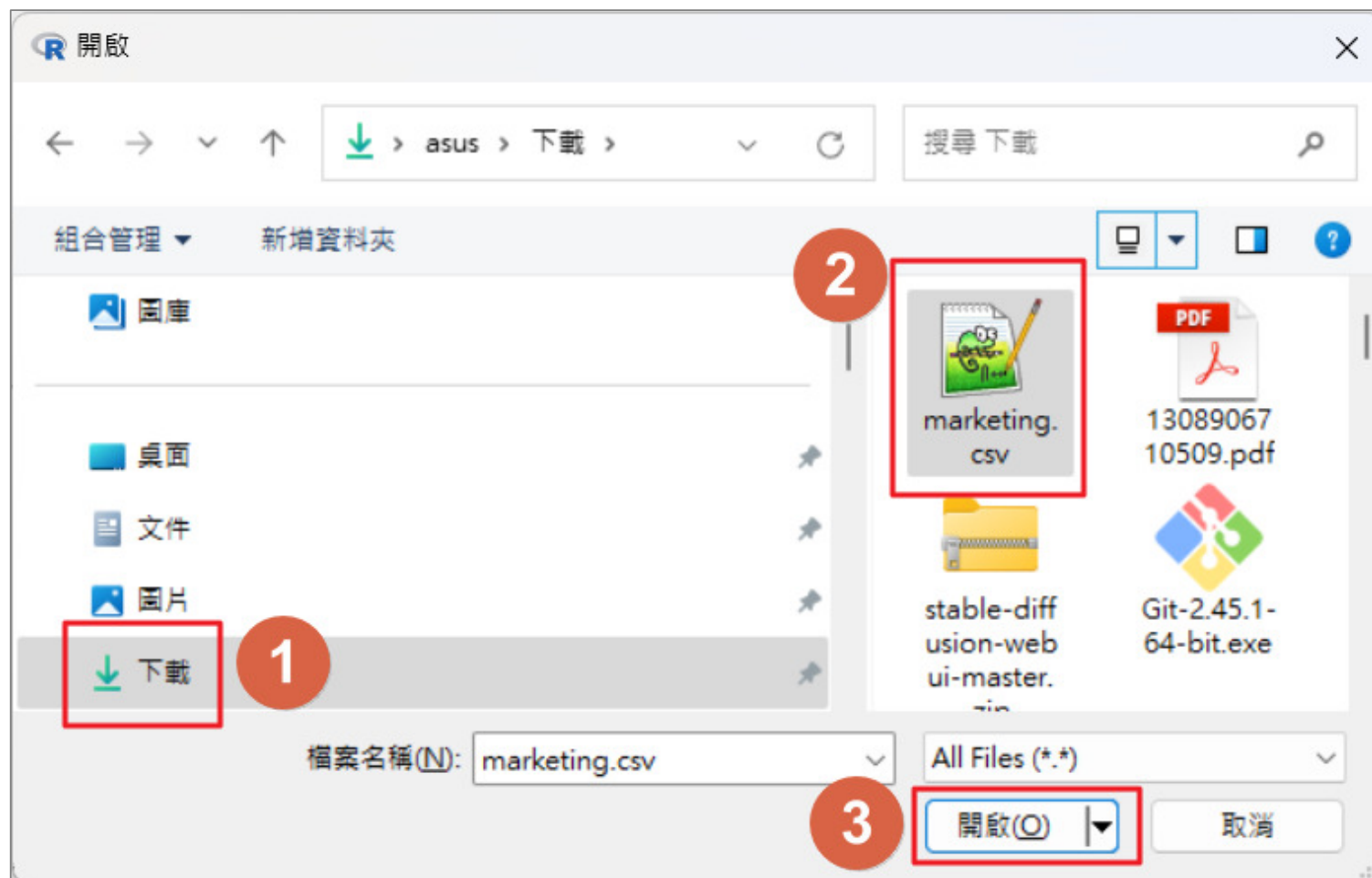
☐ 逗號 [,]

 說明

 OK

 取消

開啟視窗



資料結構 str

R 語法檔 R的Markdown文件


```
df_168 <- read.table("C:/Users/asus/Downloads/marketing.csv",  
  header=TRUE, stringsAsFactors=TRUE, sep=";", na.strings="NA",  
  dec=".", strip.white=TRUE)
```

str(df_168)

1

2

Output

 執行語法

```
> df_168 <- read.table("C:/Users/asus/Downloads/marketing.csv",  
+   header=TRUE, stringsAsFactors=TRUE, sep=";", na.strings="NA",  
+   dec=".", strip.white=TRUE)
```

```
> str(df_168)  
'data.frame': 200 obs. of  4 variables:  
 $ youtube  : num  276.1 53.4 20.6 181.8 217 ...  
 $ facebook : num  45.4 NA 55.1 49.6 13 ...  
 $ newspaper: num  83 54.1 83.2 70.2 70.1 ...  
 $ sales    : num  26.5 12.5 11.2 22.2 15.5 ...
```

3

- 資料結構 str
- 資料物件 data.frame
- 列數: 200
- 變數: 4個

摘要 summary

- 統計量 \ 摘要 \ 使用中的資料集

```
> summary(df_168)
```

youtube	facebook	newspaper	sales
Min. : 0.84	Min. : 0.00	Min. : 0.36	Min. : 1.92
1st Qu.: 89.25	1st Qu.: 11.94	1st Qu.: 15.30	1st Qu.: 12.45
Median : 179.70	Median : 27.00	Median : 30.90	Median : 15.48
Mean : 176.45	Mean : 27.82	Mean : 36.66	Mean : 16.83
3rd Qu.: 262.59	3rd Qu.: 43.68	3rd Qu.: 54.12	3rd Qu.: 20.88
Max. : 355.68	Max. : 59.52	Max. : 136.80	Max. : 32.40
	NA's : 1		

- facebook 變數有NA

填補NA

- `df_168$facebook[is.na(df_168$facebook)] <- median(df_168$facebook, na.rm = TRUE)`

R 語法檔 R的Markdown文件

```
df_168$facebook[is.na(df_168$facebook)] <- median(df_168$facebook, na.rm = TRUE)
summary(df_168)
```

Output

```
> df_168$facebook[is.na(df_168$facebook)] <- median(df_168$facebook, na.rm = TRUE)
> summary(df_168)
```

youtube	facebook	newspaper	sales
Min. : 0.84	Min. : 0.00	Min. : 0.36	Min. : 1.92
1st Qu.: 89.25	1st Qu.: 11.97	1st Qu.: 15.30	1st Qu.: 12.45
Median : 179.70	Median : 27.00	Median : 30.90	Median : 15.48
Mean : 176.45	Mean : 27.82	Mean : 26.66	Mean : 16.82
3rd Qu.: 262.59	3rd Qu.: 43.62		
Max. : 355.68	Max. : 59.52		

- facebook 變數沒有NA

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：

$$\bullet \text{ 母體相關係數} = \frac{\text{共變異數(Covariance)}}{\text{標準差}_X \times \text{標準差}_Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

$$\bullet \text{ 樣本相關係數 } \gamma = \frac{S_{xy}}{S_x S_y} = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

相關係數特性

- 相關係數值介於-1到+1之間， $-1 \leq \gamma \leq 1$
- 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。

相關矩陣

- 統計量 \ 摘要 \ 相關矩陣



相關係數 cor

```
> cor(df_168[,c("facebook", "newspaper", "sales", "youtube")], use="complete")
```

	facebook	newspaper	sales	youtube
facebook	1.0000000	0.35134059	0.5818842	0.06178210
newspaper	0.3513406	1.00000000	0.2282990	0.05664787
sales	0.5818842	0.22829903	1.00000000	0.78222442
youtube	0.0617821	0.05664787	0.7822244	1.00000000

- sales, newspaper 相關係數較小

線性迴歸

- 統計量 \ 模型配適 \ 線性迴歸

R 線性迴歸

輸入模型名稱: RegModel.1

依變數〈反應變數〉〈選取1個〉

facebook
newspaper
sales
youtube

自變數〈解釋變數〉〈選取1個或多個〉

facebook
newspaper
sales
youtube

Indices or names of row(s) to remove
<use all valid cases>

子樣本選取之條件
<所有有效觀察值>

說明 重新選擇 OK 取消 採用

線性迴歸 (Linear Model, lm) – 完成圖

```
R 語法檔 R的Markdown文件
RegModel.1 <- lm(sales~facebook+newspaper+youtube, data=df_168)
summary(RegModel.1)
```

Output

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	3.5561270	0.3728689	9.537	<2e-16 ***	
facebook	0.1891022	0.0086111	21.960	<2e-16 ***	
newspaper	-0.0006339	0.0058512	-0.108	0.914	
youtube	0.0455313	0.0013923	32.702	<2e-16 ***	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.018 on 196 degrees of freedom
Multiple R-squared: 0.8977, Adjusted R-squared: 0.8961
F-statistic: 573 on 3 and 196 DF, p-value: < 2.2e-16

線性模型

- 統計量 \ 模型配適 \ 線性模型

線性模型

輸入模型名稱: LinearModel.2

變數 (雙擊滑鼠左鍵加入公式)

facebook
newspaper
sales
youtube

Model Formula

運算子 (雙擊滑鼠左鍵加入公式) + * : / %in% - ^ ()

Splines/Polynomials:
(選擇變數並點擊)

B-spline natural spline 正交多項式 原形式多項式

splines 之自由度: 5
多項式之階次: 2

sales ~ facebook + newspaper + youtube

模型公式協助

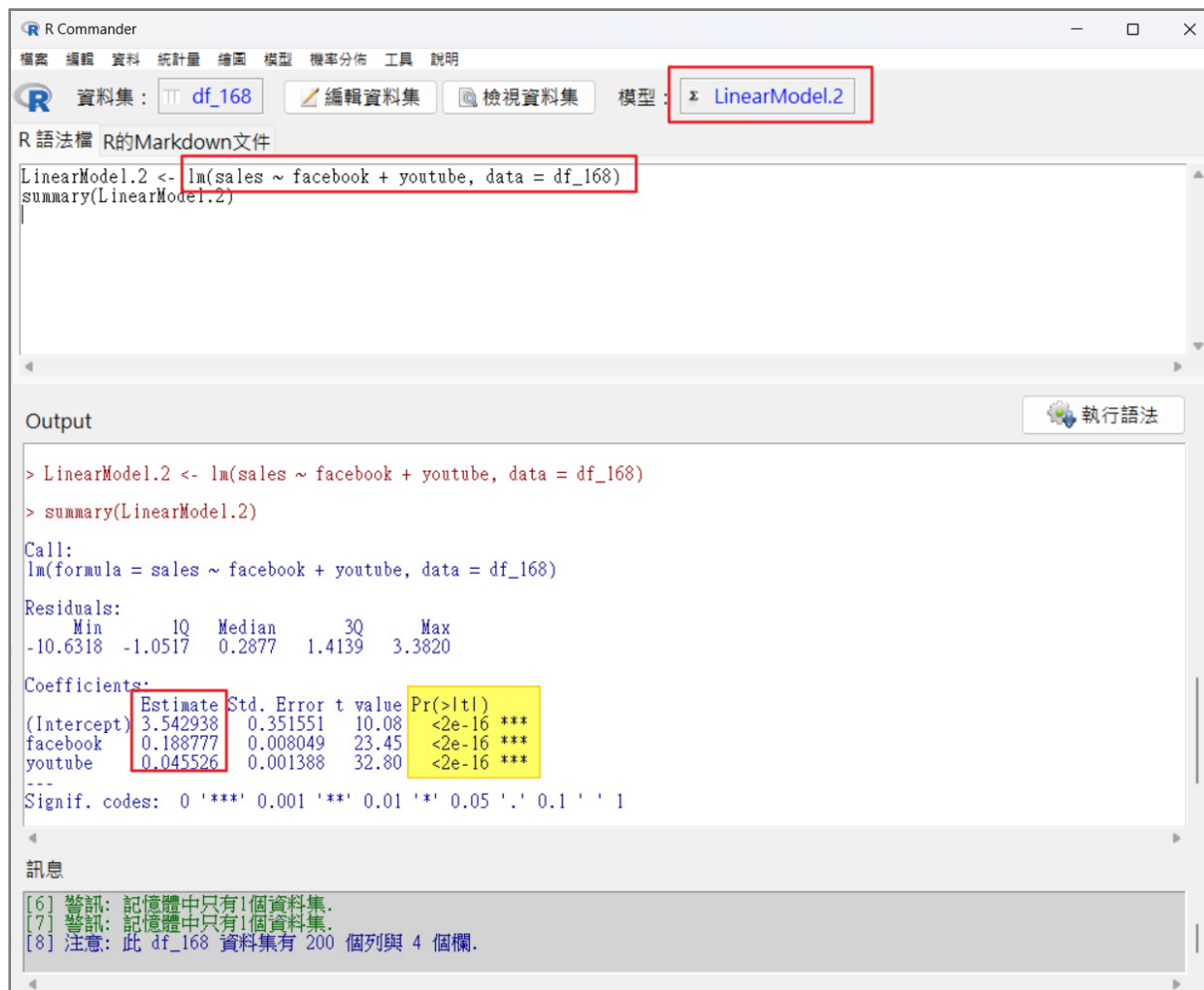
Indices or names of row(s) to remove
<use all valid cases>

子樣本選取之條件
<所有有效觀察值>

Weights
沒有選擇任何變數

說明 重新選擇 OK 取消 採用

線性模型 – 完成圖



The screenshot shows the R Commander interface with the following elements:

- Menu Bar:** 檔案, 編輯, 資料, 統計量, 繪圖, 模型, 機率分佈, 工具, 說明
- Buttons:** 資料集: df_168, 編輯資料集, 檢視資料集, 模型: LinearModel.2
- R 語法欄 (R's Markdown File):**

```
LinearModel.2 <- lm(sales ~ facebook + youtube, data = df_168)
summary(LinearModel.2)
```
- Output:**

```
> LinearModel.2 <- lm(sales ~ facebook + youtube, data = df_168)
> summary(LinearModel.2)

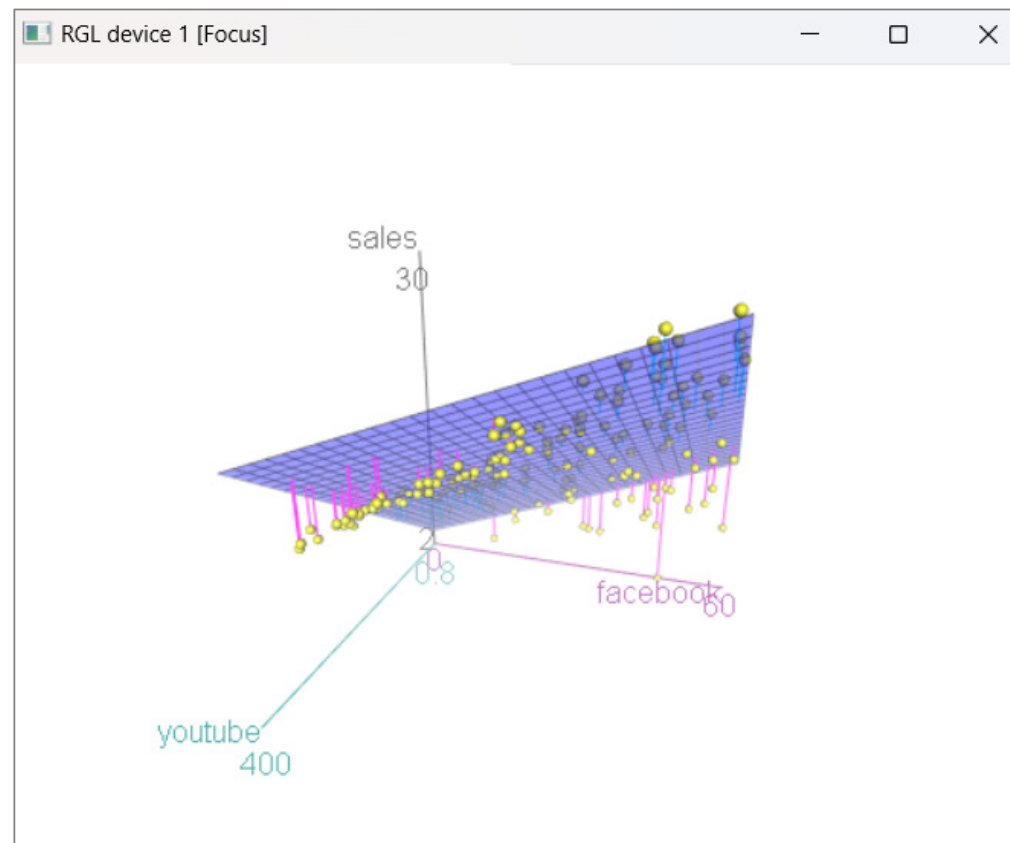
Call:
lm(formula = sales ~ facebook + youtube, data = df_168)

Residuals:
    Min       1Q   Median       3Q      Max
-10.6318  -1.0517   0.2877   1.4139   3.3820

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.542938   0.351551  10.08   <2e-16 ***
facebook     0.188777   0.008049  23.45   <2e-16 ***
youtube      0.045526   0.001388  32.80   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
- 訊息 (Messages):**

```
[6] 警訊: 記憶體中只有1個資料集.
[7] 警訊: 記憶體中只有1個資料集.
[8] 注意: 此 df_168 資料集有 200 個列與 4 個欄.
```

迴歸模型-3D繪圖



- 3D自由旋轉
- 放大/縮小

4.課程回顧

Reviews

- R, RStudio
- Rcmdr
- Import Tab File 匯入Tab檔案
- 宗教社會服務概況資料分析
- 研究方法與開放資料
- ggplot2 簡介
- 套件 package
- 變異數分析(ANOVA)
- 迴歸 (Regression)

謝謝您的聆聽

Q & A



李明昌

alan9956@gmail.com

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