

資料倉儲導論-第3章 套件,函數,資料物件

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

3.1 R使用環境

3.2 套件 package

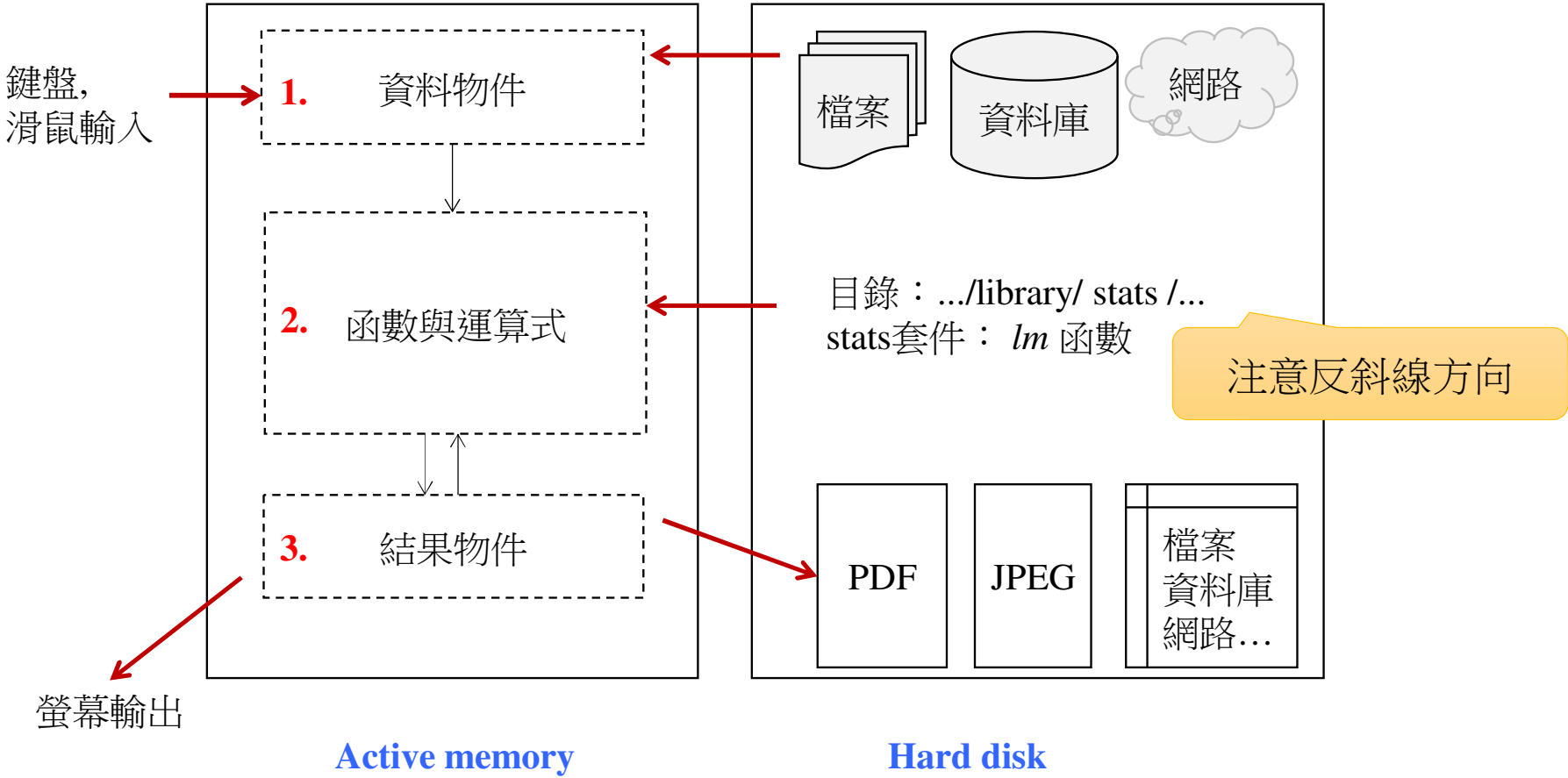
3.3 輔助說明 help

3.4 數學運算

3.5 資料物件

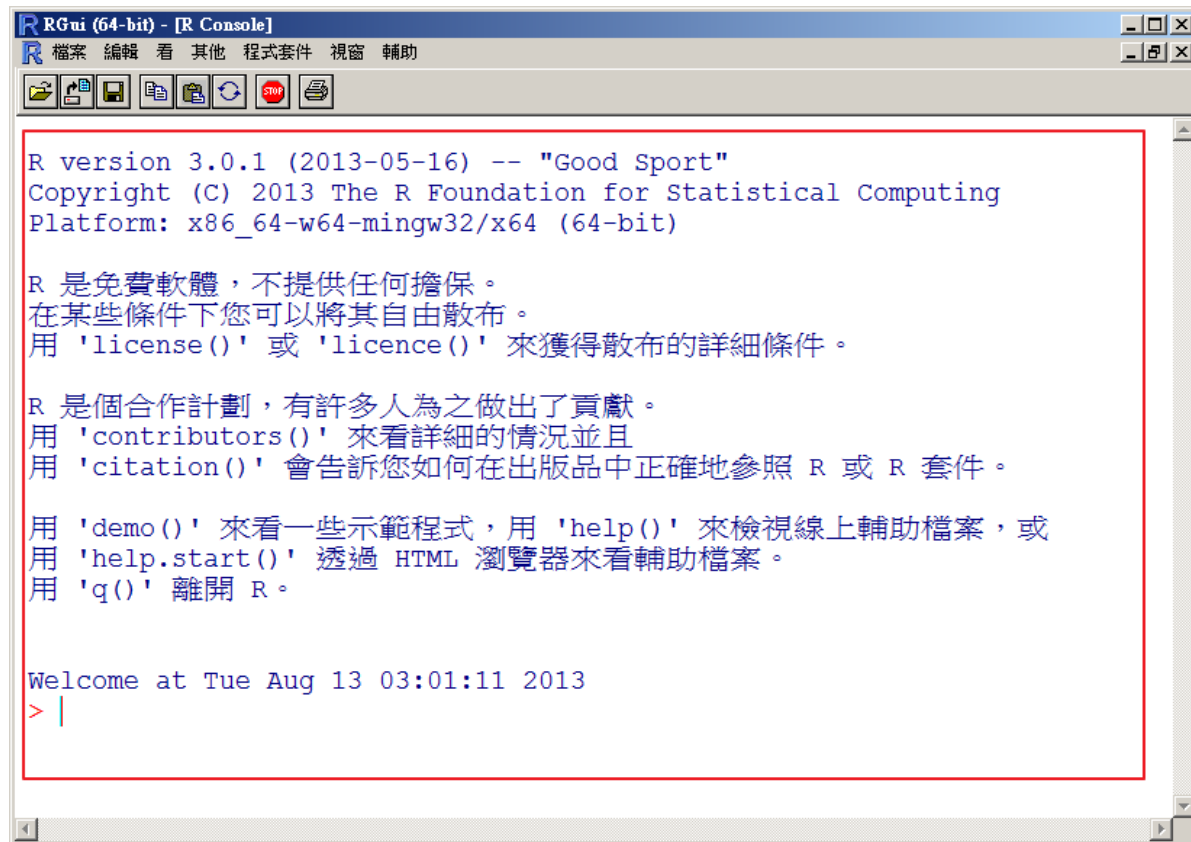
3.1 R使用環境

R運作方式



基本觀念

- 控制台(console)
- 歷程
 - xxx.Rhistory
- 套件(package)
- 工作空間(workspace)
 - xxx.RData
- 物件(object)



```
R version 3.0.1 (2013-05-16) -- "Good Sport"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R 是免費軟體，不提供任何擔保。
在某些條件下您可以將其自由散布。
用 'license()' 或 'licence()' 來獲得散布的詳細條件。

R 是個合作計劃，有許多人為之做出了貢獻。
用 'contributors()' 來看詳細的情況並且
用 'citation()' 會告訴您如何在出版品中正確地參照 R 或 R 套件。

用 'demo()' 來看一些示範程式，用 'help()' 來檢視線上輔助檔案，或
用 'help.start()' 透過 HTML 瀏覽器來看輔助檔案。
用 'q()' 離開 R。

Welcome at Tue Aug 13 03:01:11 2013
> |
```

控制台的特定符號

- 命令提示字元(大於) > (等待使用者輸入資料)
- 指令未完提示字元(加號) + (表示尚未輸入完成)
- 註解提示字元(井字號) # (不會編譯註解)
- 結果行列顯示編號

```
> iris$sepal.Length
```

```

[1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9 5.4 4.8 4.8 4.3 5.8 5.7
[17] 5.4 5.1 5.7 5.1 5.4 5.1 4.6 5.1 4.8 5.0 5.0 5.2 5.2 4.7 4.8 5.4
[33] 5.2 5.5 4.9 5.0 5.5 4.9 4.4 5.1 5.0 4.5 4.4 5.0 5.1 4.8 5.1 4.6
[49] 5.3 5.0 7.0 6.4 6.9 5.5 6.5 5.7 6.3 4.9 6.6 5.2 5.0 5.9 6.0 6.1
[65] 5.6 6.7 5.6 5.8 6.2 5.6 5.9 6.1 6.3 6.1 6.4 6.6 6.8 6.7 6.0 5.7
[81] 5.5 5.5 5.8 6.0 5.4 6.0 6.7 6.3 5.6 5.5 5.5 6.1 5.8 5.0 5.6 5.7
[97] 5.7 6.2 5.1 5.7 6.3 5.8 7.1 6.3 6.5 7.6 4.9 7.3 6.7 7.2 6.5 6.4
[113] 6.8 5.7 5.8 6.4 6.5 7.7 7.7 6.0 6.9 5.6 7.7 6.3 6.7 7.2 6.2 6.1
[129] 6.4 7.2 7.4 7.9 6.4 6.3 6.1 7.7 6.3 6.4 6.0 6.9 6.7 6.9 5.8 6.8
[145] 6.7 6.7 6.3 6.5 6.2 5.9
```


物件命名原則

- R的大小寫有差異: *a* 與 *A* 是不同的物件
- R 也保留一些物件與指令名稱, 如 *c*, *C*, *T*, *F* 等為保留字 (“reserved words”), 命名時避免重覆, 以免引起人類困擾.
- 物件名稱起始位置須以文字或 “.” (句點)
- 如果物件名稱以 “.” (句點) 為起始, 名稱第二個位置需為文字, 物件名稱其餘位置, 以文字 (A-Z 或 a-z), 數字 (0-9), */ . -*, 皆可.
- 中間不可有空格
- 建議名稱不要使用中文

Google's R Style Guide

- <https://google.github.io/styleguide/Rguide.html>
- 函數使用 BigCamelCase

```
# Good
DoNothing <- function() {
  return(invisible(NULL))
}
```

- 不要使用 attach 函數
- 使用 `x <- 1`, 不要使用 `x = 1`
- = 用於函數之參數設定 `plot(..., type = "b")`
- 不要使用句點 `Customer.Sales`  改為 `CustomerSales`

3.2 套件 package

套件

- 使用套件兩部曲 - 先安裝, 再載入套件
 - **install.packages**(“套件名稱”) # 安裝套件(一生一次)
 - **library**(套件名稱) # 載入套件(每次使用)
- 範例: 新增與載入 e1071 套件(machine learning)

```
> install.packages("e1071")
trying URL 'http://cran.cs.pu.edu.tw/bin/windows/contrib/3.0/e1071_1.6-1.zip'
Content type 'application/zip' length 514468 bytes (502 Kb)
opened URL
downloaded 502 Kb

package 'e1071' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
C:\Users\Administrator\AppData\Local\Temp\RtmpoHS0Ak\downloaded_packages
> library(e1071)
Loading required package: class
>
```


example(svm, package="e1071")

已載入的套件 search()

```
> # 已載入套件
```

```
> search()
```

```
[1] ".GlobalEnv" "package:e1071"  
[3] "tools:rstudio" "package:stats"  
[5] "package:graphics" "package:grDevices"  
[7] "package:utils" "package:datasets"  
[9] "package:methods" "AutoLoads"  
[11] "package:base"  
>
```



R套件 - 40類別

(2022.9.22)

- <https://cran.csie.ntu.edu.tw/web/packages/index.html>

Contributed Packages

Available Packages

Currently, the CRAN package repository features 18728 available packages.

[Table of available packages, sorted by date of publication](#)

[Table of available packages, sorted by name](#)

[CRAN Task Views](#) aim to provide some guidance which packages on CRAN are relevant for tasks related to a certain topic. They provide tools to automatically install all packages from each view. Currently, 40 views are available.

40類別 - 中文說明

Task Views - R套件

RWEPA → task

更新日期: 2022.9.29 - 40個套件類別

CRAN Task View:

<https://cran.csie.ntu.edu.tw/web/views/>

CRAN (Taiwan):

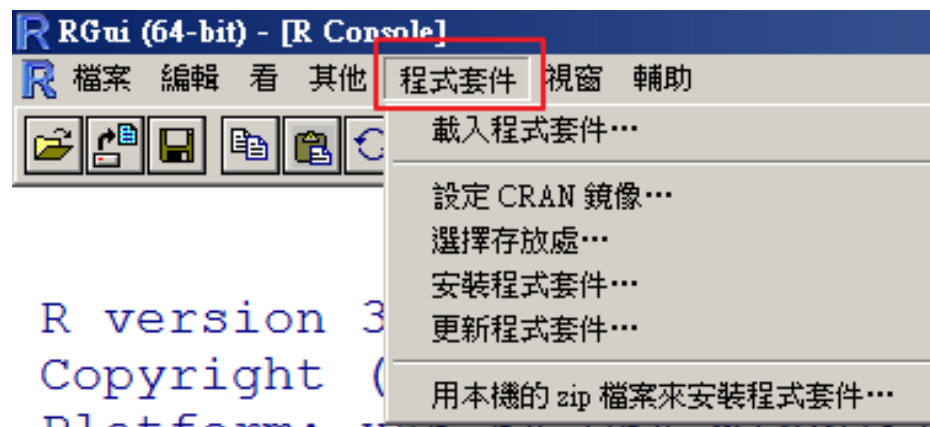
<https://cran.csie.ntu.edu.tw/>

選取 CRAN 網站左側 [Packages] , 套件區分成以下類別, 中文說明如下:

-----	-----	-----	-----
編號	主題	英文說明	中文說明
-----	-----	-----	-----

- 01. **Agriculture**, Agricultural Science, 農業科學
- 02. **Bayesian**, Bayesian Inference, 貝氏統計
- 03. **CausalInference**, Causal Inference, 因果推論
- 04. **ChemPhys**, Chemometrics and Computational Physics, 計量化學, 計算物理
- 05. **ClinicalTrials**, Clinical Trial Design, Monitoring, and Analysis, 臨床試驗設計, 監測和分析

R 套件選單



- `update.packages("xxx")` # 更新套件
- `detach("package : xxx")` # 卸離套件
- `remove.packages("xxx")` # 移除已安裝套件
- 上述指令大部份可在 R / RStudio 執行

R對話資訊

- `sessionInfo()` → 理解R安裝訊息

```
> sessionInfo()
R version 4.2.0 (2022-04-22 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: windows 10 x64 (build 19044)

Matrix products: default

locale:
[1] LC_COLLATE=Chinese (Traditional)_Taiwan.utf8
[2] LC_CTYPE=Chinese (Traditional)_Taiwan.utf8
[3] LC_MONETARY=Chinese (Traditional)_Taiwan.utf8
[4] LC_NUMERIC=C
[5] LC_TIME=Chinese (Traditional)_Taiwan.utf8

attached base packages:
[1] stats      graphics  grDevices  utils      datasets
[6] methods    base
```

套件安裝目錄

- .Library

```
> # 預設套件安裝目錄  
> .Library  
[1] "C:/PROGRA~1/R/R-42~1.0/library"
```

- .libPaths()

```
> # 套件安裝目錄  
> .libPaths()
```

```
[1] "C:/Users/asus/AppData/Local/R/win-library/4.2"  
[2] "C:/Program Files/R/R-4.2.0/library"
```

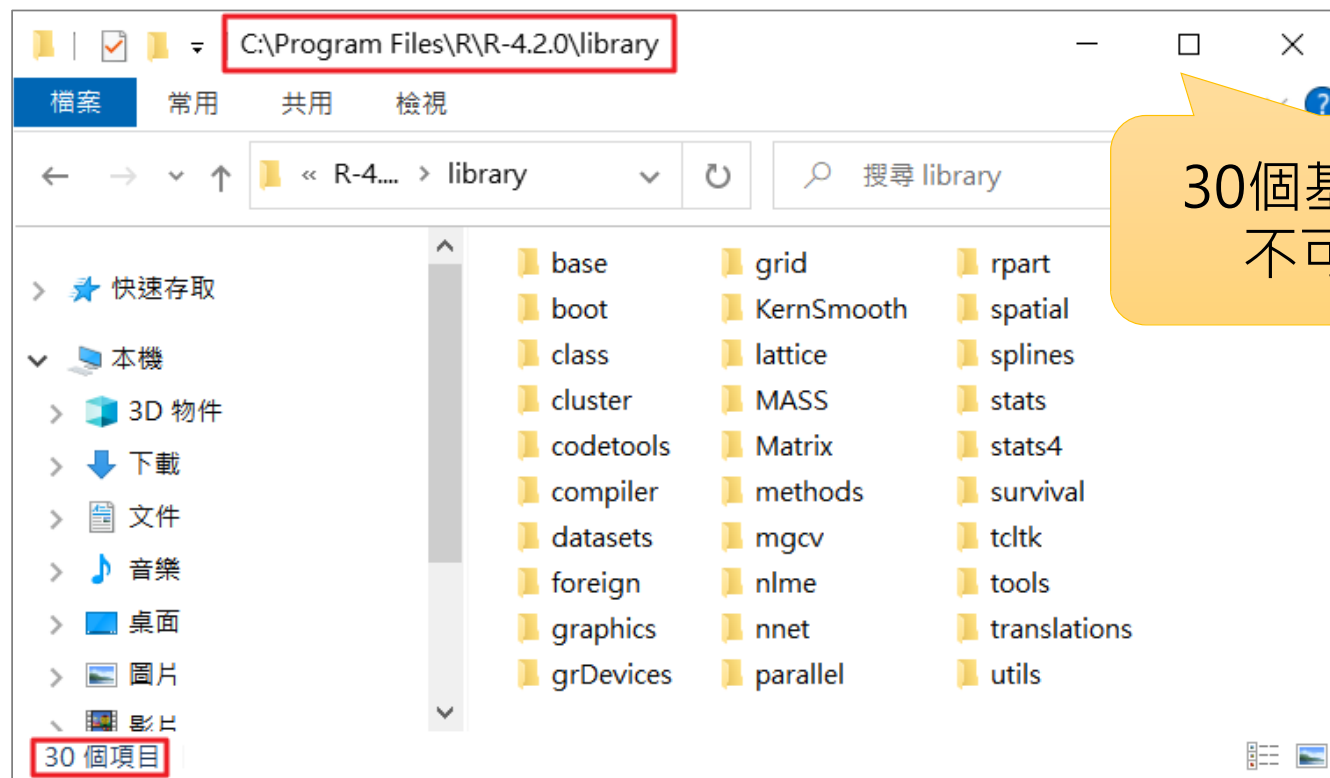


已安裝套件

```
> # 已安裝套件
> x <- installed.packages()
> class(x) # "matrix" "array"
[1] "matrix" "array"
> dim(x)
[1] 1616 16
> mypackage = x[, 1] # matrix[列, 行]
> mypackage[1:5]
      abind      actuar      ada
"abind"  "actuar"  "ada"
      adabag  additivityTests
"adabag"  "additivityTests"
> library() # same as installed.packages()
```

套件安裝目錄1

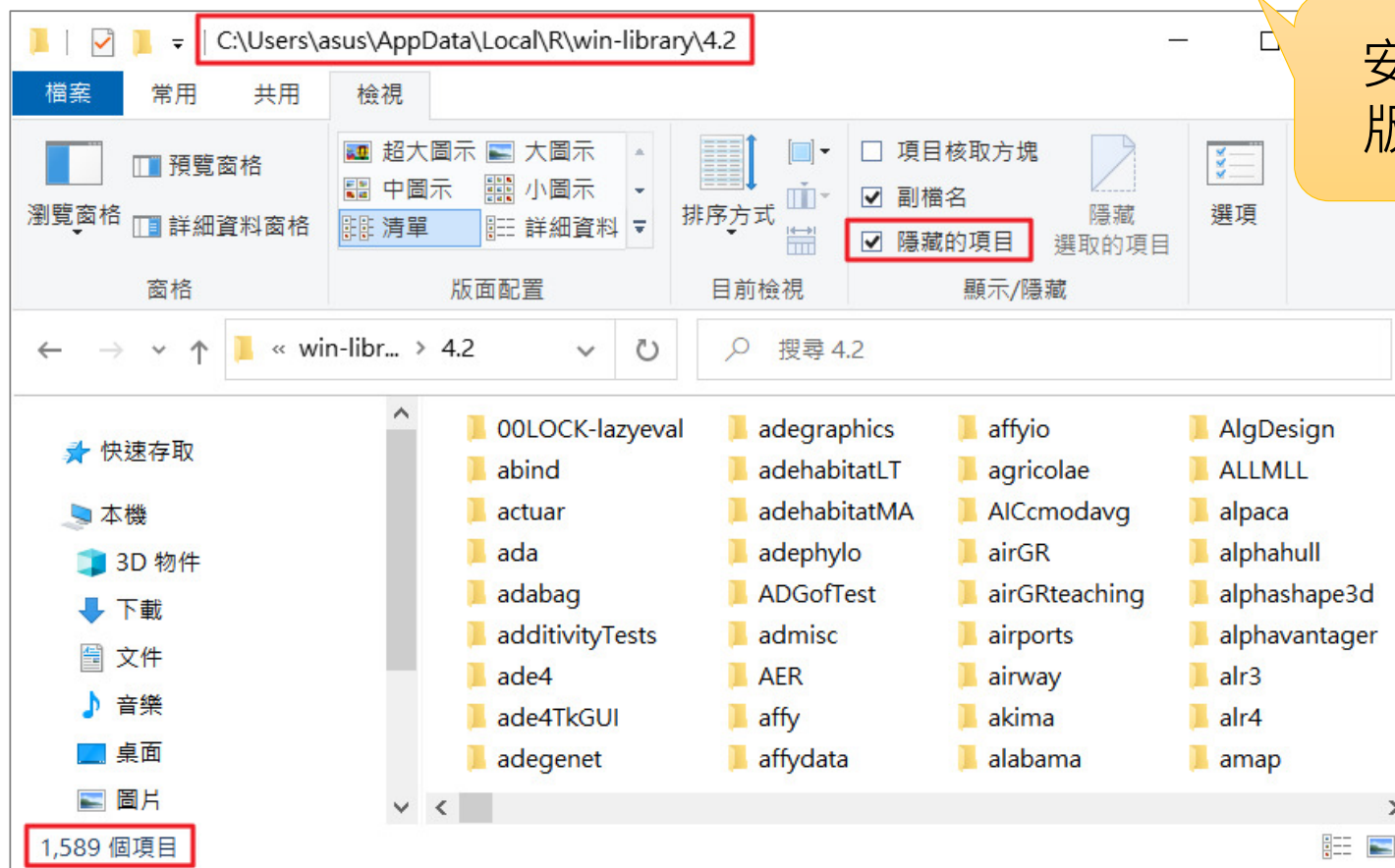
- C:\Program Files\R\R-4.2.0\library



30個基礎套件,
不可刪除.

套件安裝目錄2

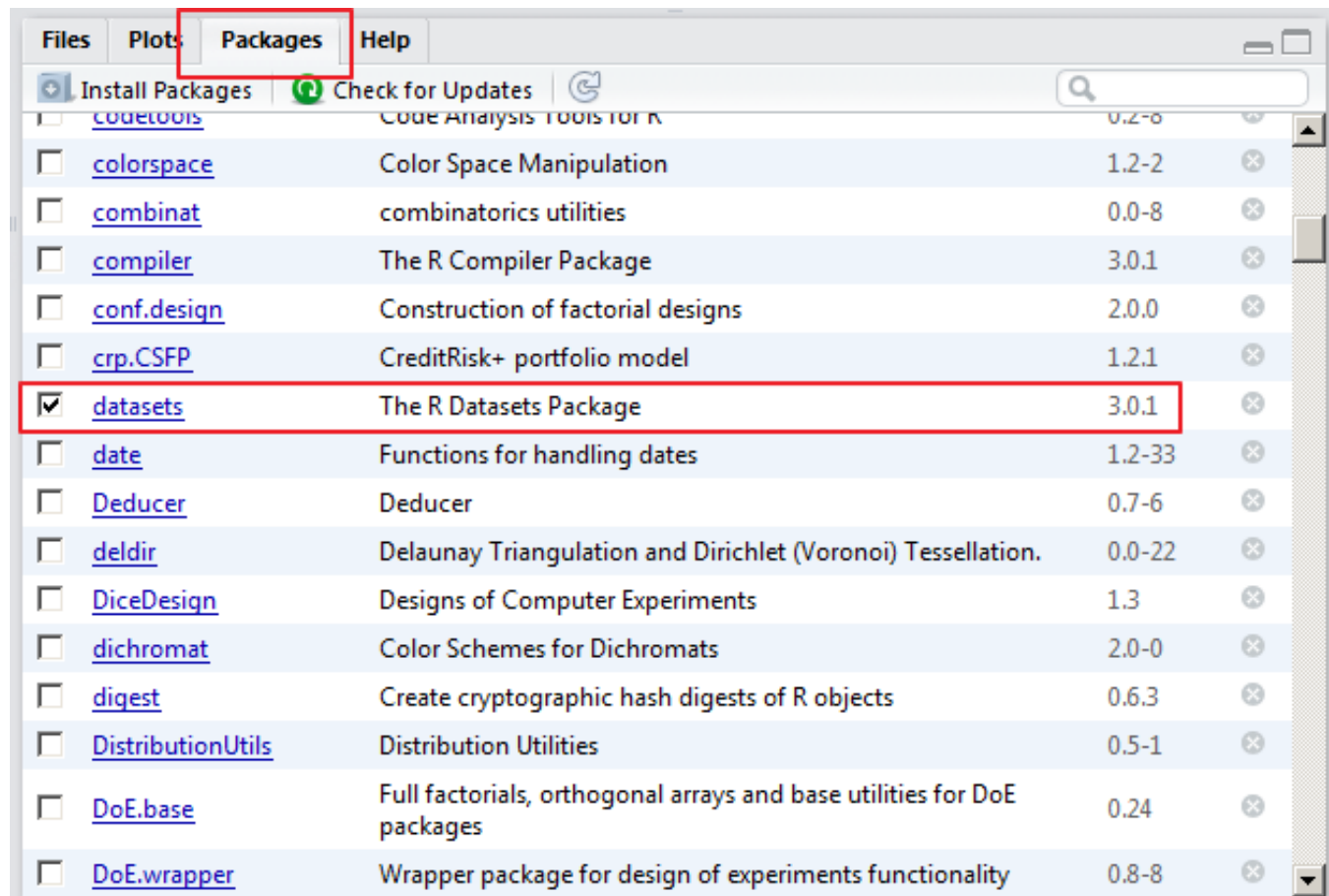
• C:\Users\asus\AppData\Local\R\win-library\4.2



安裝目錄會隨著
版本不同而改變

RStudio 套件管理

打勾表示已經載入



3.3 輔助說明 help

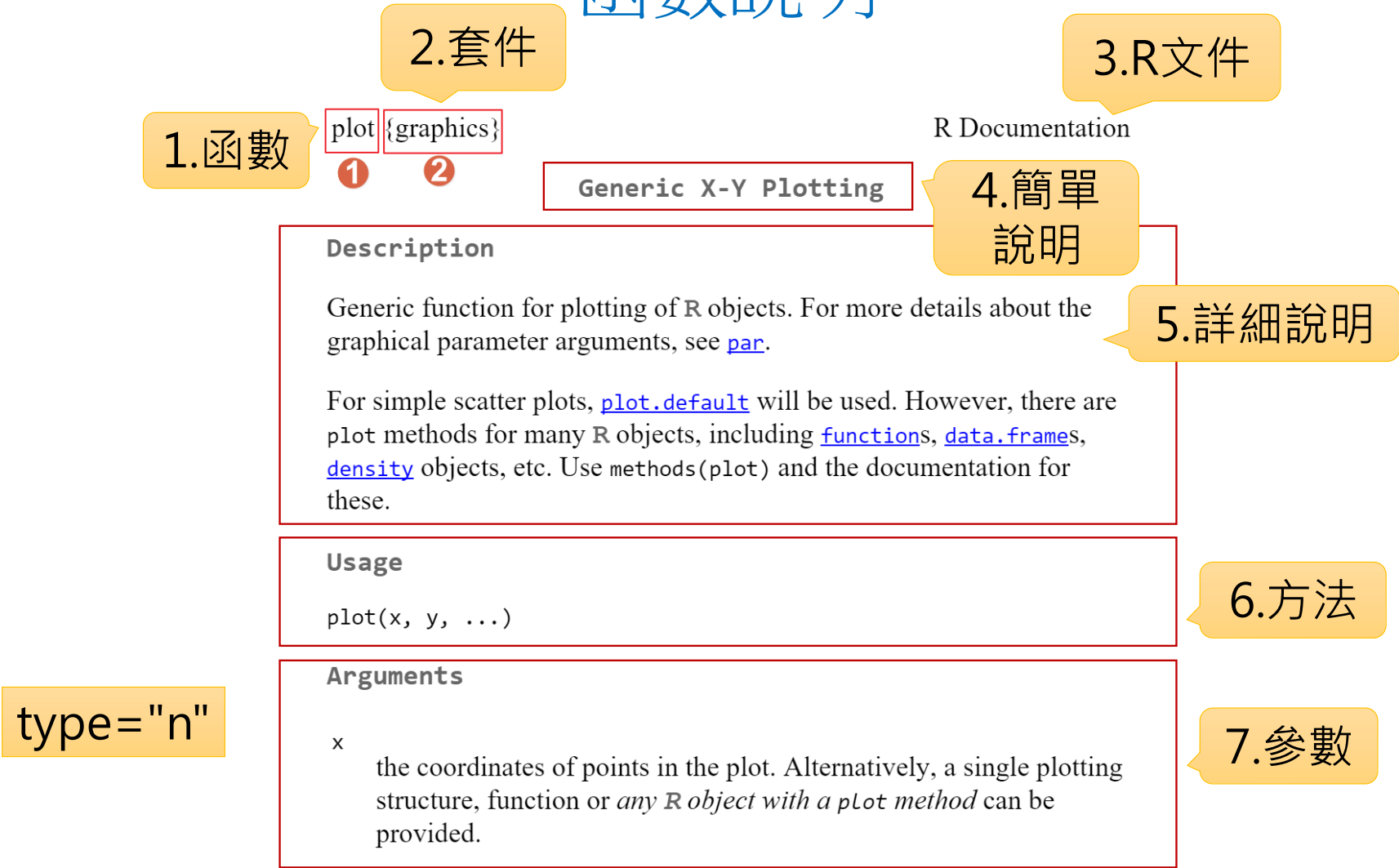
輔助說明

- 常用輔助說明方法
 - `help.start()`
 - `?plot`
 - `help(plot)`
 - 選取 `plot` 按 F1
 - `help.search()`



```
> help.search("regression")  
>  
> ??regression
```

函數說明



3.4 數學運算

數學運算



- R 即是計算機
 - `log`, `exp`
- 算術操作 (arithmetic operator)
 - `+`, `-`, `*`, `/`, `^`, `%%`, `%/%`, `%*%`
- 關係比較操作 (relation/comparison operator)
 - `==`, `!=`, `<`, `<=`, `>`, `>=`
- 邏輯操作 (logical operator)
 - `!`, `&`, `|`

- `x == "台北市"`
- `x == '台北市'`
- `y == 3.14`

特殊數值

- R 可以正確表示無窮大數值:
 - $+\infty$ (正無窮大): **Inf**
 - $-\infty$ (負無窮大): **-Inf**
- **NaN**: 不是一個數值(數學上無定義,例:0/0)
- **NA**: 表示遺漏值(missing values)或(Not Available)
- **is.finite(x)** 判定是否為有限的
- **is.infinite(x)** 判定是否為無窮大
- **is.nan(x)** 判定是否為NaN
- **pi, letters, LETTERS, month.abb, month.name**

R進行微分,積分

英文月份

3.5 資料物件

資料型別

- 整數
- 數值
- 字串: 須使用 `台北市` 或 "台北市" 符號
- 邏輯值: 包括 TRUE, FALSE

資料物件 (5個)

向量 vector

北部	中部	南部
----	----	----

矩陣matrix

1	3	5
2	4	6

陣列array

1.1	4.4	7.7
2.2	5.5	8.8
3.3	6.6	9.9

資料框data.frame

1	男	62
2	女	50
3	女	54
4	男	72

資料框：以串列方式儲存，但其長度相同。

串列list

北部	中部	南部
1	3	5
2	4	6
1	男	62
2	女	50
3	女	54
4	男	72

向量

矩陣

資料框

串列：
每一個元素其資料型別與長度可以不相同。

矩陣預設採用
直行填入資料

Excel 工作表

資料物件重要觀念

- 向量是最基本的物件
 - 數值向量、字元向量
 - 所有資料的資料型別須相同
 - 因子是一種較特別的向量, 儲存類別型變數
- 矩陣與陣列
 - 矩陣是二維陣列, 陣列允許大於或等於2個維度
 - 所有資料的資料型別須相同
- 資料框、時間序列、串列均可同時存入數字與字串, 但資料框與時間序列內的向量長度都相等
- 串列可包含陣列與資料框(二維, 可多種資料型別), 函數回傳值以串列(list)物件為主

資料物件名稱中, 英文皆可, 建議英文, 不可用數字開頭.

建立向量函數 c (concatenate)

- 將一群數字、字串、邏輯值結合成向量
- 將多個向量結合成向量
- 向量物件具有屬性長度 **length** 與型式 **mode**
- 向量會將所有元素將強制(**coercion**)轉換成單一相同型態
- 因子 (**factor**) 是一種特別的向量，用於將資料依離散型變數做成分群 (**group**)

因子 factor - levels, labels

```
> # 因子 factor
> f1 <- factor(1:3)
> f2 <- factor(1:3, levels=1:5)
> f1
[1] 1 2 3
Levels: 1 2 3
> f2
[1] 1 2 3
Levels: 1 2 3 4 5
> f2[4] <- 5
> f2[5] <- 10
Warning message:
In `[<-.factor`(`*tmp*`, 5, value = 10) :
  invalid factor level, NA generated
> f2
[1] 1     2     3     5    <NA>
Levels: 1 2 3 4 5
>
```


factor 範例1

```
> eye.colors <- factor(c("brown", "blue", "blue", "green", "brown", "brown", "brown"))
>
> eye.colors
[1] brown blue  blue  green brown brown brown
Levels: blue brown green
>
> levels(eye.colors)
[1] "blue" "brown" "green"
>
> labels(eye.colors)
[1] "1" "2" "3" "4" "5" "6" "7"
>
```

factor 範例2

```
> gender <- factor(c("男", "女", "男", "男", "女"))
>
> gender
[1] 男 女 男 男 女
Levels: 女 男
>
> levels(gender)
[1] "女" "男"
>
> str(gender)
Factor w/ 2 levels "女","男": 2 1 2 2 1
>
```

有序因子 (ordered factor)

- 有序因子表示有大小順序, 例: {大, 中, 小}

```
> ClothSize <- ordered(c("L", "H", "L", "M", "H"),
+                       levels = c("L", "M", "H"))
>
> ClothSize
[1] L H L M H
Levels: L < M < H
>
> levels(ClothSize)
[1] "L" "M" "H"
>
> str(ClothSize)
Ord.factor w/ 3 levels "L"<"M"<"H": 1 3 1 2 3
>
```

因子轉換

- `as.factor` 轉換為因子
- `as.numeric()` 轉換為數值
- `as.character()` 轉換為字串
- 因子內部儲存為 整數 $\{1, 2, 3, \dots\}$, 整數表示顏色或使用 **`colors()`**

- 使用時機-例: 縣市別, 性別
- 使用 **`cut`** 函數: 數值 → 類別

向量 vector

- 類似 Excel 的一行或是一列
- 整數, 實數, 字元, 數值+字元?

```
> # 向量 vector -----  
> # 整數  
> v0 <- c(1:10)  
> v0  
[1] 1 2 3 4 5 6 7 8 9 10  
> class(v0)  
[1] "integer"  
> typeof(v0)  
[1] "integer"  
>
```

R demo

矩陣 matrix

- 矩陣是將向量擴充至二個(或保持一個)維度。
- 建立矩陣
 - `matrix(data = NA, nrow = 1, ncol = 1, byrow = FALSE, dimnames = NULL)`
- 轉換為矩陣
 - `as.matrix(x)`
- 判斷是否為矩陣
 - `is.matrix(x)`

預設採用直行
填入資料

矩陣的運算

- 判斷是否為矩陣 `is.matrix()`
- 轉換為矩陣 `as.matrix()`
- 矩陣運算 `+`, `-`, `%*%`
- 矩陣轉置 `t()`
- 取出對角線值 `diag()`

矩陣相乘

$a * X = b$
?solve

陣列 array

- 陣列是將向量擴充至二個(或以上)維度。
- 陣列表示多重維度且為相同資料型態。
- 產生陣列
 - `array(data = NA, dim = length(data), dimnames = NULL)`
- 轉換為陣列
 - `as.array(x)`
- 判斷是否為陣列
 - `is.array(x)`

資料框 data.frame

- 資料框是一種重要的資料物件型態。
- 一般R模型計算以資料框作為資料輸入。
- 資料框是二維資料物件，每一橫列表示一個觀測值，每一直行表示一個變數，變數資料型態可能不相同，但個數相同。

```
cars[2]           # data.frame  
cars["dist"]      # data.frame  
cars[,2]          # vector
```

認識 iris, 150*5

1

2

3

4

5

```
> iris
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa

資料結構 str

- head(iris) # 前6筆
- tail(iris) # 後6筆
- str(iris) # 資料結構

```
> str(iris)
'data.frame': 150 obs. of 5 variables:
 $ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species : Factor w/ 3 levels "setosa","versicolor",...: 1 1 1 1 1 1 1 1 1 1 ...
>
```

資料摘要 summary

```
> summary(iris)
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
Min. :4.300	Min. :2.000	Min. :1.000	Min. :0.100	setosa :50
1st Qu.:5.100	1st Qu.:2.800	1st Qu.:1.600	1st Qu.:0.300	versicolor:50
Median :5.800	Median :3.000	Median :4.350	Median :1.300	virginica :50
Mean :5.843	Mean :3.057	Mean :3.758	Mean :1.199	
3rd Qu.:6.400	3rd Qu.:3.300	3rd Qu.:5.100	3rd Qu.:1.800	
Max. :7.900	Max. :4.400	Max. :6.900	Max. :2.500	

```
>
```

Sepal.Length	欄位名稱
Min. :4.300	最小值 Minimum
1st Qu.:5.100	25百分位數 First Quantile 第一四分位數 Q1
Median :5.800	50百分位數 Second Quantile 第二四分位數 Q2
Mean :5.843	平均值
3rd Qu.:6.400	75百分位數 Third Quantile 第三四分位數 Q3
Max. :7.900	最大值 Maximum

謝謝您的聆聽

Q & A

李明昌

EMAIL: alan9956@gmail.com

WEB: <http://rwepa.blogspot.com/>

