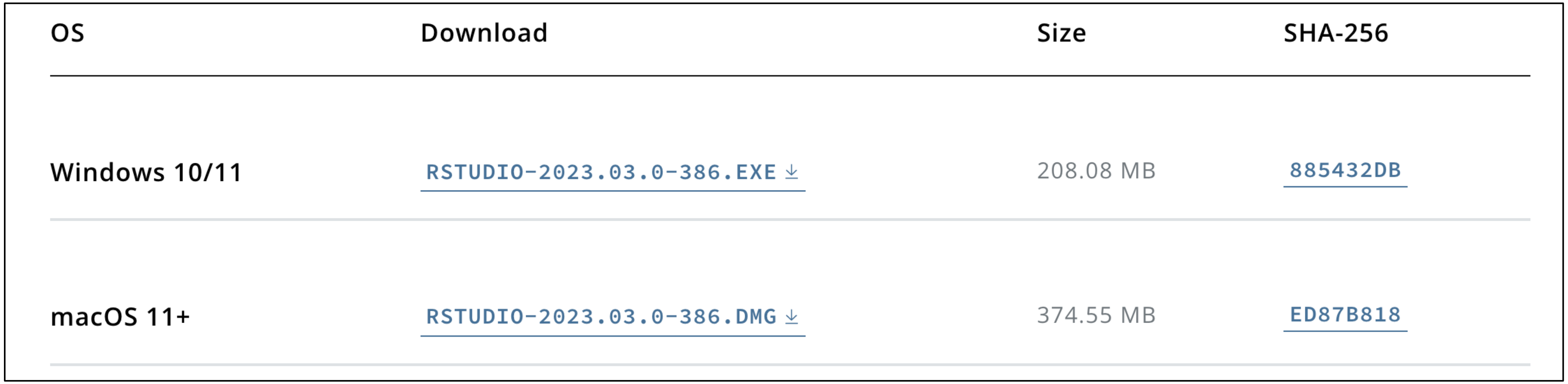
R 語言設計概念與農業應用

台大農藝系 劉力瑜

lyliu@ntu.edu.tw

安裝 R 與 Rstudio

- R 下載:
  - Windows: https://cran.r-project.org/bin/windows/base/
  - Mac: https://cran.r-project.org/bin/macosx/
  - 下載後雙擊滑鼠左鍵下載檔案進入安裝畫面
- Rstudio 下載: https://posit.co/download/rstudio-desktop/

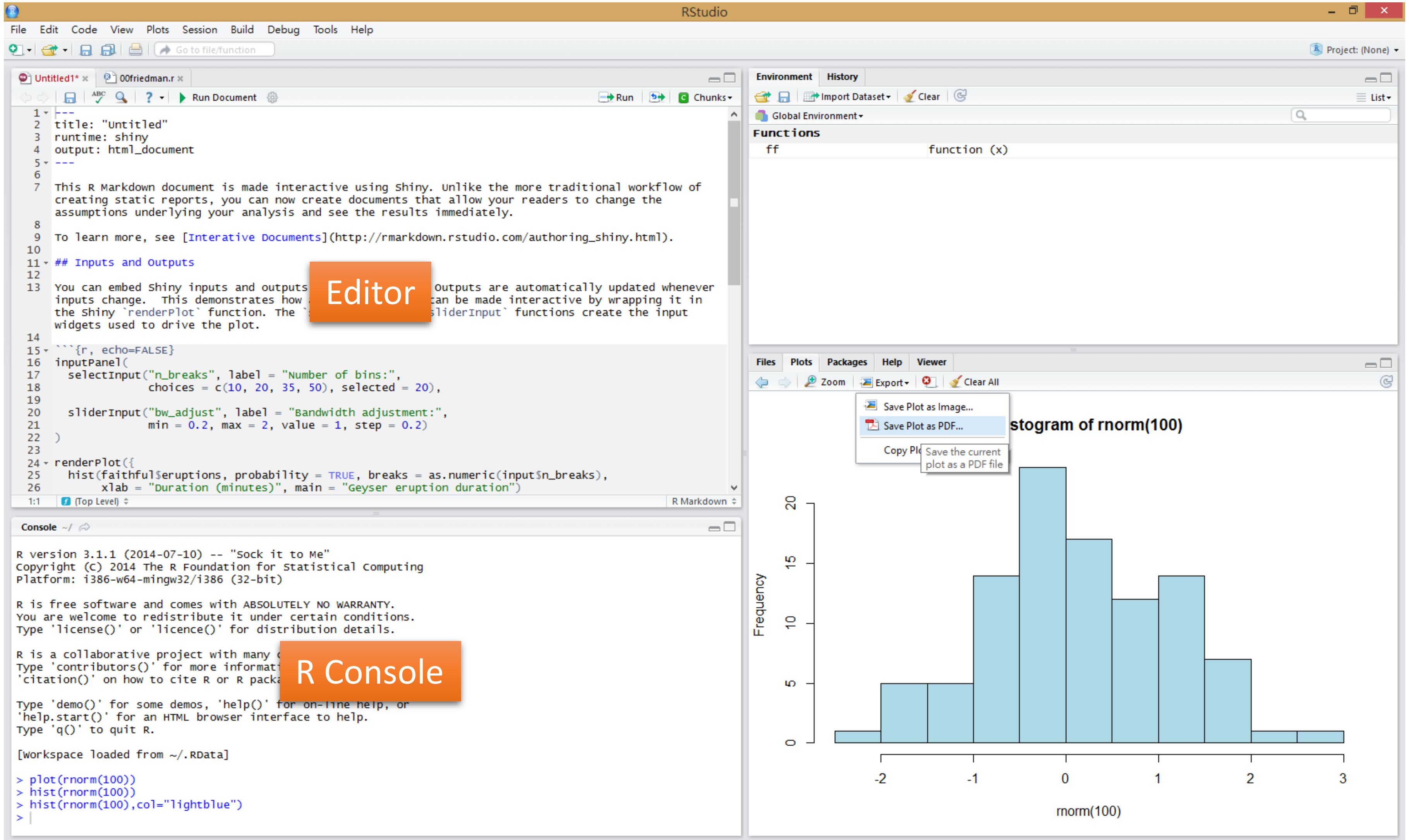


OS	Download	Size	SHA-256
Windows 10/11	RSTUDIO-2023.03.0-386.EXE ⤓	208.08 MB	885432DB
macOS 11+	RSTUDIO-2023.03.0-386.DMG ⤓	374.55 MB	ED87B818

What is R?

- R 並非專用統計軟體, 而是可用來執行分析的環境:
  - 匯入適當的 package (套件)
  - 應用套件內提供之 function (函式)
- Packages 由許多熱心人士編寫並免費提供學術使用。

RStudio 方便使用的 R 介面



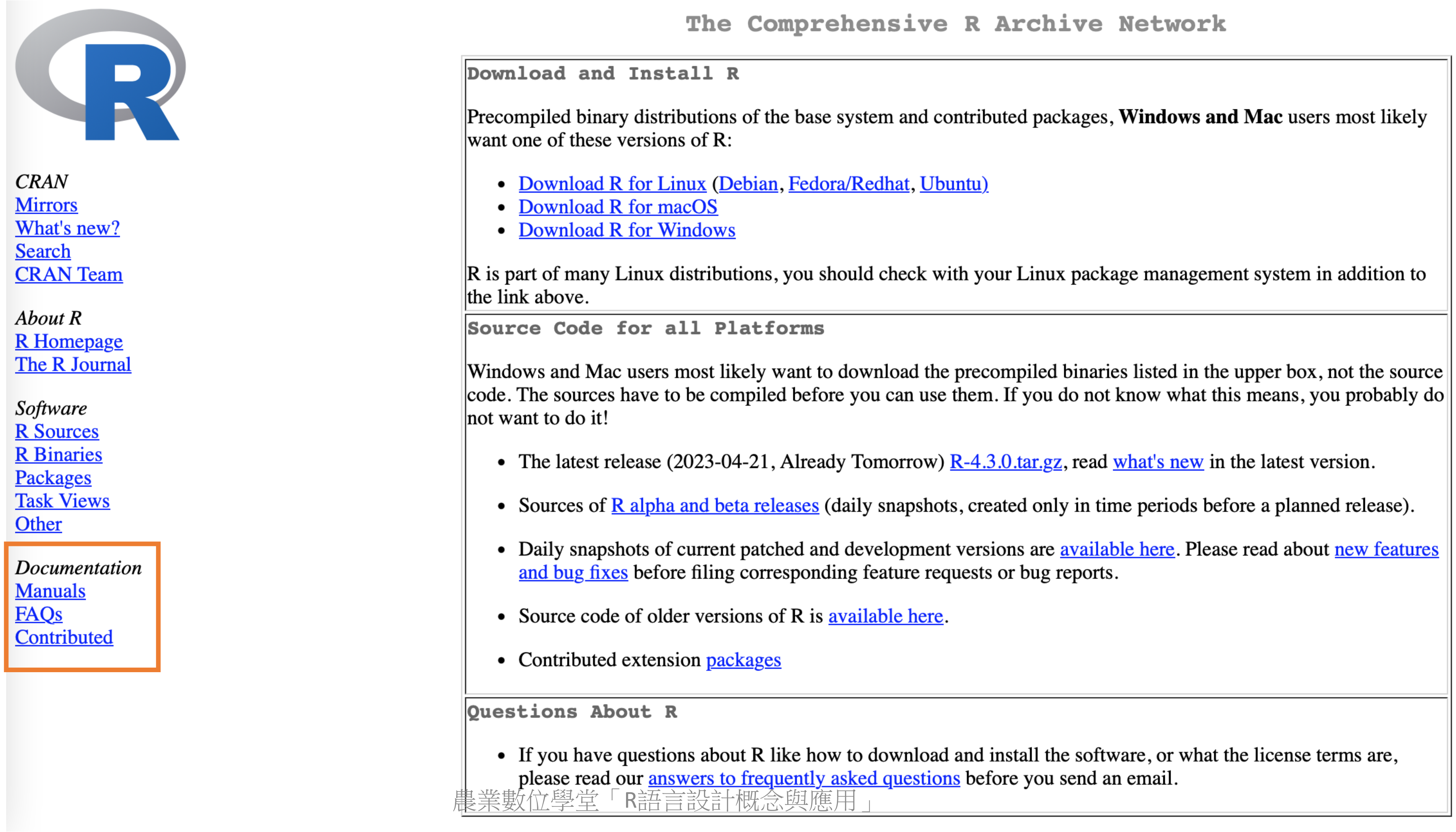
~~~
---
title: "Untitled"
runtime: shiny
output: html_document
---

This R Markdown document is made interactive using Shiny. Unlike the more traditional workflow of
creating static reports, you can now create documents that allow your readers to change the
assumptions underlying your analysis and see the results immediately.

To learn more, see [Interative Documents](http://rmarkdown.rstudio.com/authoring_shiny.html).

## Inputs and Outputs

You can embed Shiny inputs and outputs                 outputs are automatically updated whenever
inputs change.  This demonstrates how                 an be made interactive by wrapping it in
the Shiny `renderPlot` function. The `                 liderInput` functions create the input
widgets used to drive the plot.

```{r, echo=FALSE}
inputPanel(
  selectInput("n_breaks", label = "Number of bins:",
              choices = c(10, 20, 35, 50), selected = 20),

  sliderInput("bw_adjust", label = "Bandwidth adjustment:",
              min = 0.2, max = 2, value = 1, step = 0.2)
)

renderPlot({
  hist(faithful$eruptions, probability = TRUE, breaks = as.numeric(input$n_breaks),
       xlab = "Duration (minutes)", main = "Geyser eruption duration")
~~~

```
R version 3.1.1 (2014-07-10) -- "Sock it to Me"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many c
Type 'contributors()' for more informati
'citation()' on how to cite R or R packa

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> plot(rnorm(100))
> hist(rnorm(100))
> hist(rnorm(100),col="lightblue")
> 
```

R的優缺點

- 優點:
  - 免費軟體
  - 完善的說明文件與討論區
  - 漂亮的圖型介面
  - 程式容易根據使用者需求做修改
- 缺點:
  - 並無 user friendly 之使用者介面
  - 需詳知函式名稱與程式編寫邏輯
  - 說明文件與討論區使用英文

http://cran.csie.ntu.edu.tw/
or https://cran.r-project.org



CRAN
Mirrors
What's new?
Search
CRAN Team

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Task Views
Other

Documentation
Manuals
FAQs
Contributed

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux (Debian, Fedora/Redhat, Ubuntu)
- Download R for macOS
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

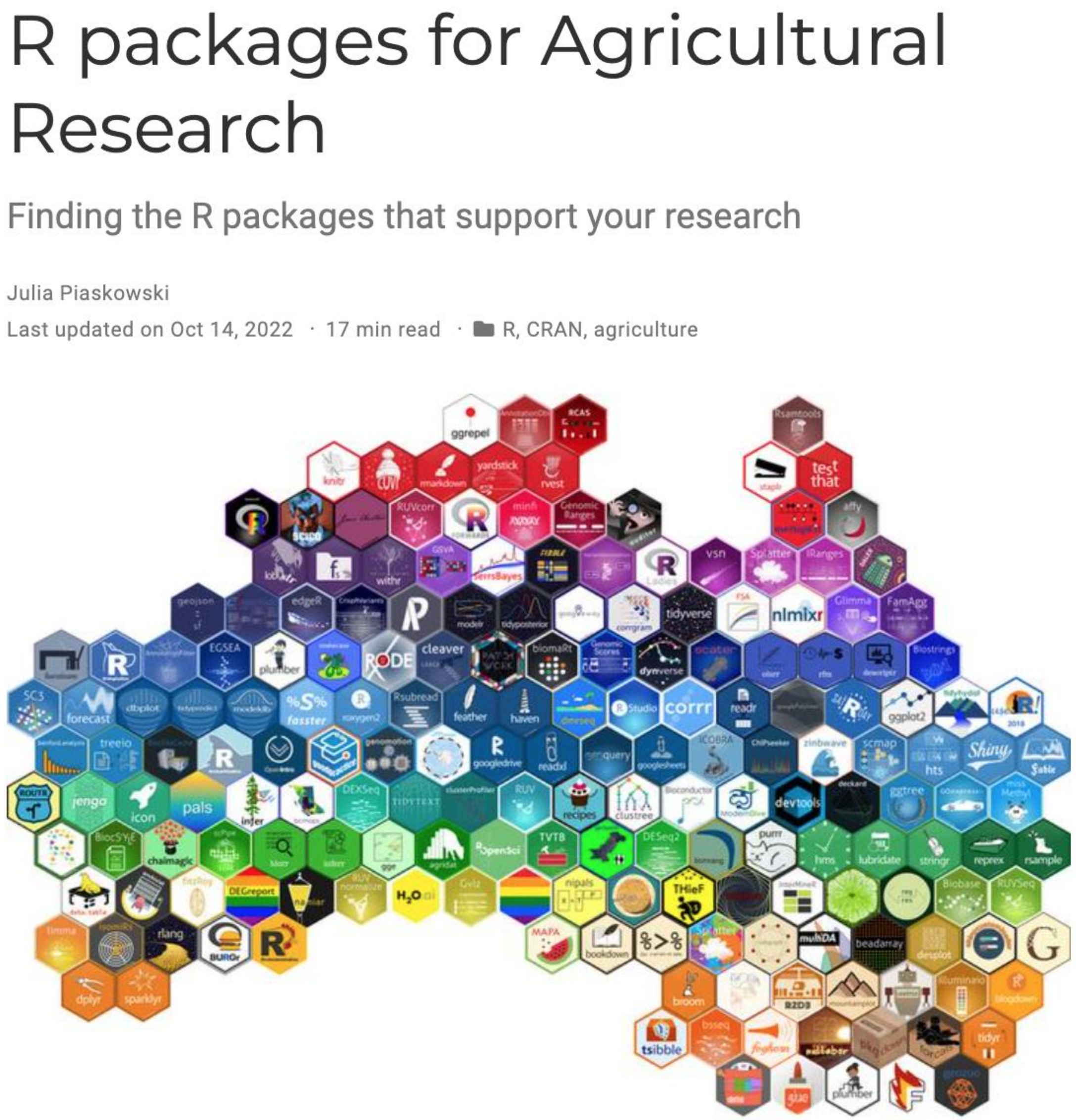
Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2023-04-21, Already Tomorrow) R-4.3.0.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

https://agstats.io/post/keeping-up-with-r/



R packages for Agricultural Research

Finding the R packages that support your research

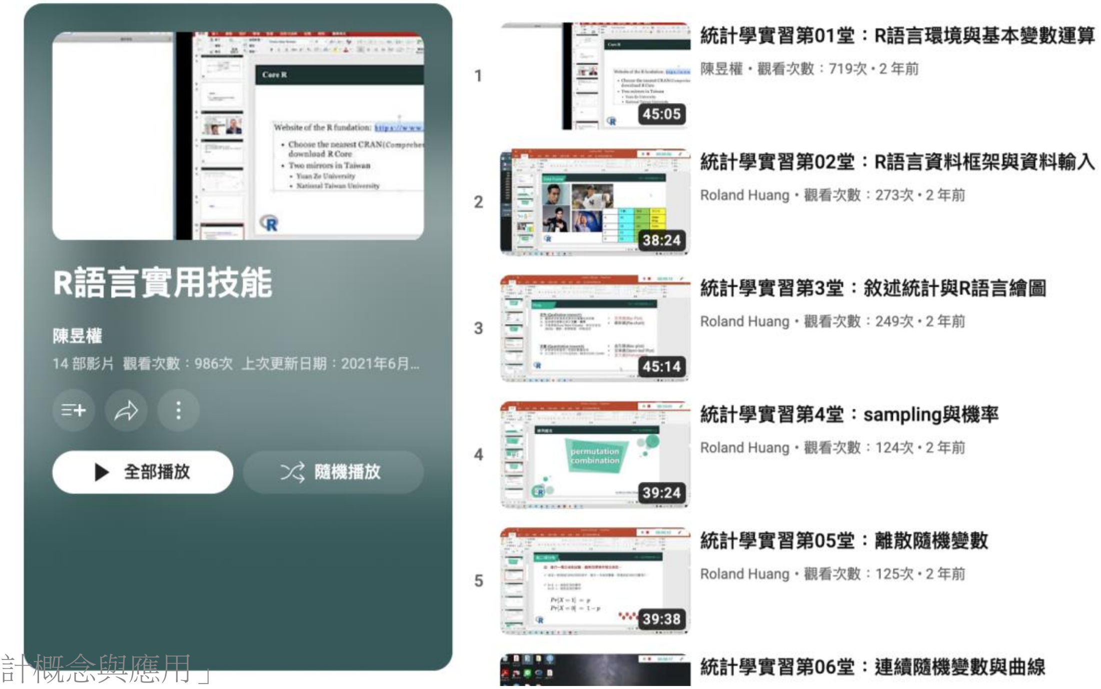
Julia Piaskowski

Last updated on Oct 14, 2022 · 17 min read · R, CRAN, agriculture

http://www.hmwu.idv.tw/index.php/r-software

R統計軟體教學/R語言程式設計/學習講義 (R Software Learning Materials)

https://youtube.com/playlist?list=PL3tdy4h3sD63W3T3JGhBusRcT-uxTM_2P



R語言實用技能

陳昱權

14 部影片 觀看次數：986次 上次更新日期：2021年6月...

1 統計學實習第01堂：R語言環境與基本變數運算
陳昱權・觀看次數：719次・2 年前

2 統計學實習第02堂：R語言資料框架與資料輸入
Roland Huang・觀看次數：273次・2 年前

3 統計學實習第3堂：敘述統計與R語言繪圖
Roland Huang・觀看次數：249次・2 年前

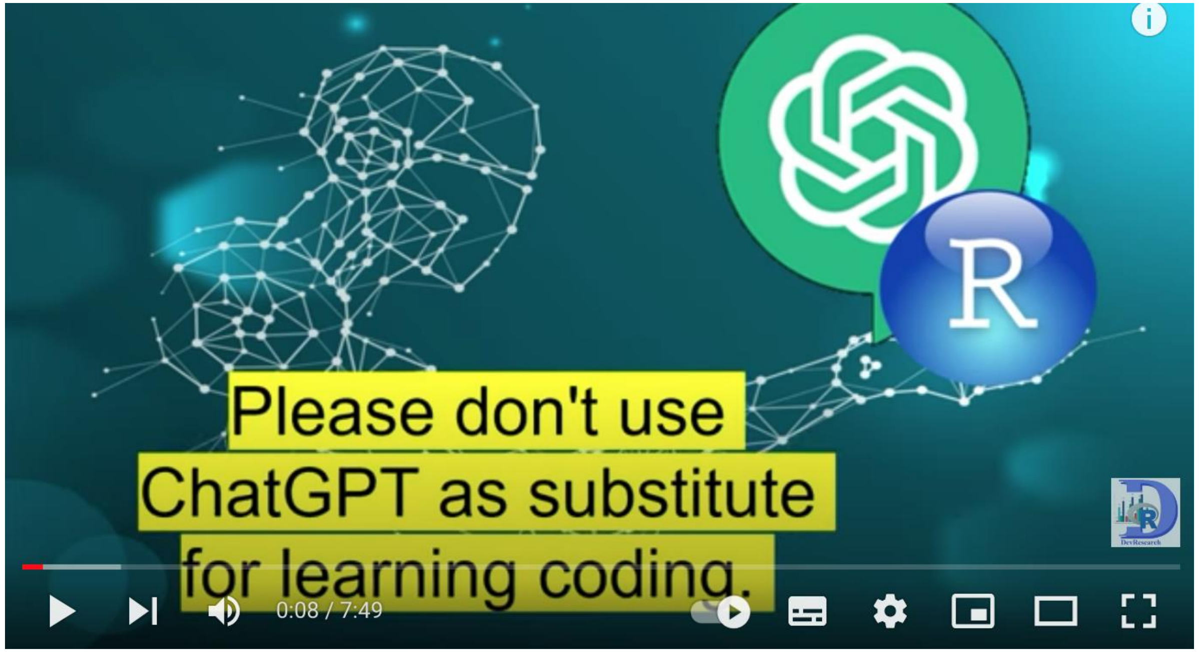
4 統計學實習第4堂：sampling與機率
Roland Huang・觀看次數：124次・2 年前

5 統計學實習第05堂：離散隨機變數
Roland Huang・觀看次數：125次・2 年前

統計學實習第06堂：連續隨機變數與曲線

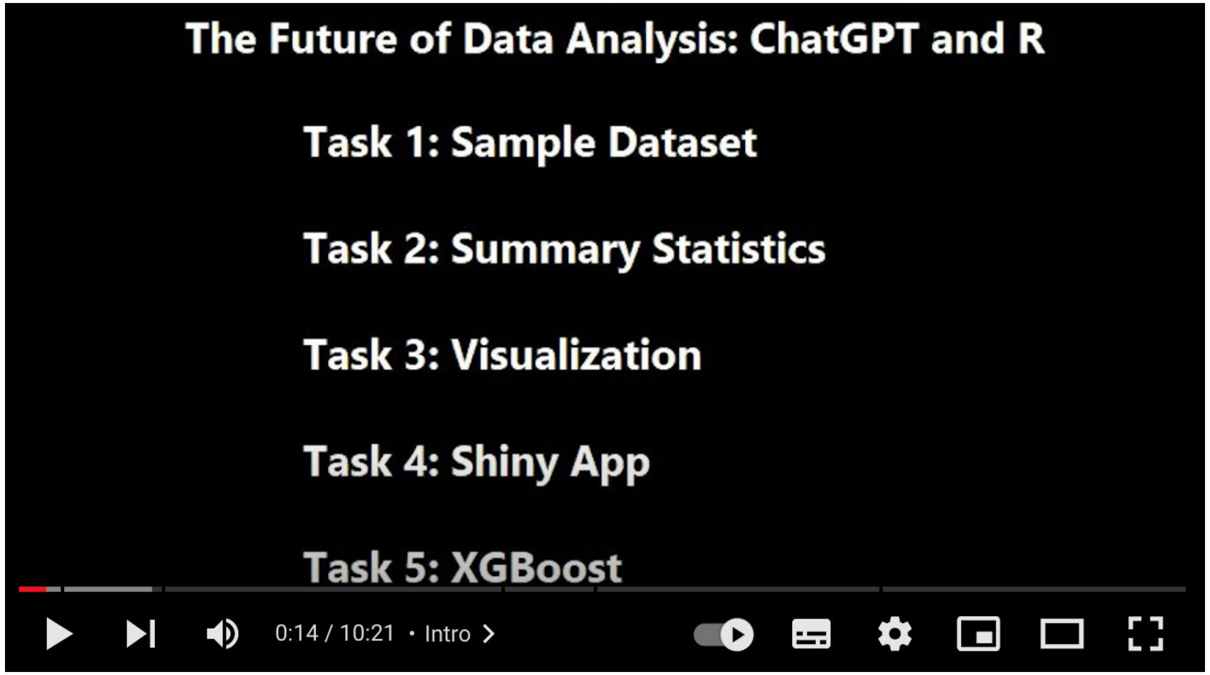
ChatGPT for coding: “程式語言翻譯機”

https://www.youtube.com/watch?v=9FQuy2Q5-1M



Please don't use ChatGPT as substitute for learning coding.

https://www.youtube.com/watch?v=Y3H1yTUPV0Y



The Future of Data Analysis: ChatGPT and R

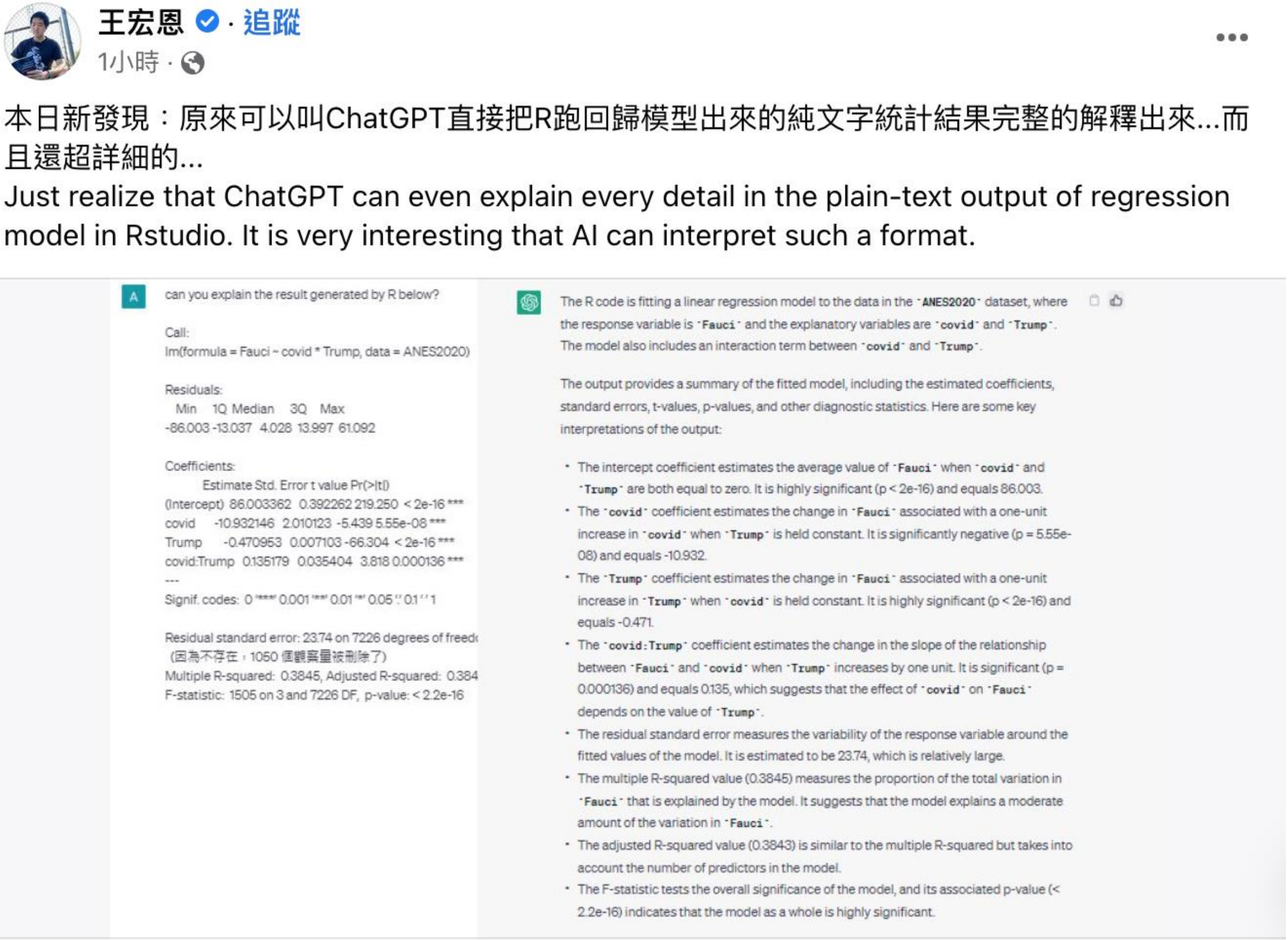
Task 1: Sample Dataset

Task 2: Summary Statistics

Task 3: Visualization

Task 4: Shiny App

Task 5: XGBoost



王宏恩 · 追蹤
1小時 ·

本日新發現：原來可以叫ChatGPT直接把R跑回歸模型出來的純文字統計結果完整的解釋出來...而且還超詳細的...
Just realize that ChatGPT can even explain every detail in the plain-text output of regression model in Rstudio. It is very interesting that AI can interpret such a format.

can you explain the result generated by R below?

```
Call:
lm(formula = Fauci ~ covid * Trump, data = ANES2020)

Residuals:
   Min     1Q Median     3Q    Max
-86.003 -13.037  4.028 13.997 61.092

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 86.003362  0.392262 219.250  < 2e-16 ***
covid      -10.932146  2.010123  -5.439 5.55e-08 ***
Trump       -0.470953  0.007103 -66.304  < 2e-16 ***
covid:Trump  0.135179  0.035404   3.818 0.000136 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.74 on 7226 degrees of freed
  (因為不存在，1050 個觀察量被刪除了)
Multiple R-squared:  0.3845, Adjusted R-squared:  0.384
F-statistic:  1505 on 3 and 7226 DF,  p-value: < 2.2e-16
```

The R code is fitting a linear regression model to the data in the `ANES2020` dataset, where the response variable is `Fauci` and the explanatory variables are `covid` and `Trump`. The model also includes an interaction term between `covid` and `Trump`.

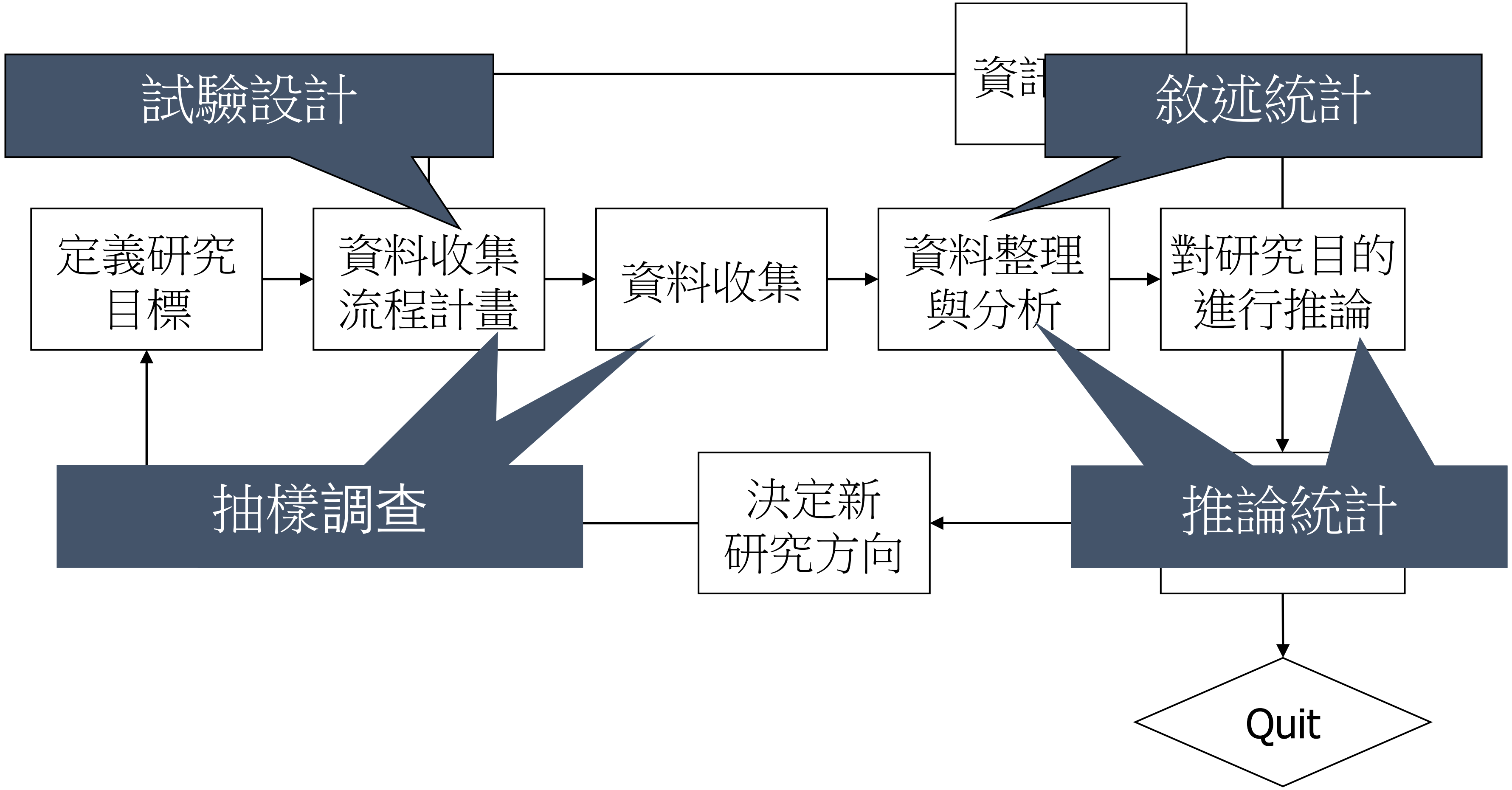
The output provides a summary of the fitted model, including the estimated coefficients, standard errors, t-values, p-values, and other diagnostic statistics. Here are some key interpretations of the output:

- The intercept coefficient estimates the average value of `Fauci` when `covid` and `Trump` are both equal to zero. It is highly significant (p < 2e-16) and equals 86.003.
- The `covid` coefficient estimates the change in `Fauci` associated with a one-unit increase in `covid` when `Trump` is held constant. It is significantly negative (p = 5.55e-08) and equals -10.932.
- The `Trump` coefficient estimates the change in `Fauci` associated with a one-unit increase in `Trump` when `covid` is held constant. It is highly significant (p < 2e-16) and equals -0.471.
- The `covid:Trump` coefficient estimates the change in the slope of the relationship between `Fauci` and `covid` when `Trump` increases by one unit. It is significant (p = 0.000136) and equals 0.135, which suggests that the effect of `covid` on `Fauci` depends on the value of `Trump`.
- The residual standard error measures the variability of the response variable around the fitted values of the model. It is estimated to be 23.74, which is relatively large.
- The multiple R-squared value (0.3845) measures the proportion of the total variation in `Fauci` that is explained by the model. It suggests that the model explains a moderate amount of the variation in `Fauci`.
- The adjusted R-squared value (0.3843) is similar to the multiple R-squared but takes into account the number of predictors in the model.
- The F-statistic tests the overall significance of the model, and its associated p-value (< 2.2e-16) indicates that the model as a whole is highly significant.

R 語言應用

- 資料收集 - 開放資料、試驗設計
- 了解資料 - 圖表與摘要統計值
- 建立模型 - 傳統統計方法、機器學習的分類與預測
- 進階應用 - 影像辨識、生物資訊

研究進行流程



試驗設計

資

敘述統計

定義研究目標

資料收集流程計畫

資料收集

資料整理與分析

對研究目的進行推論

抽樣調查

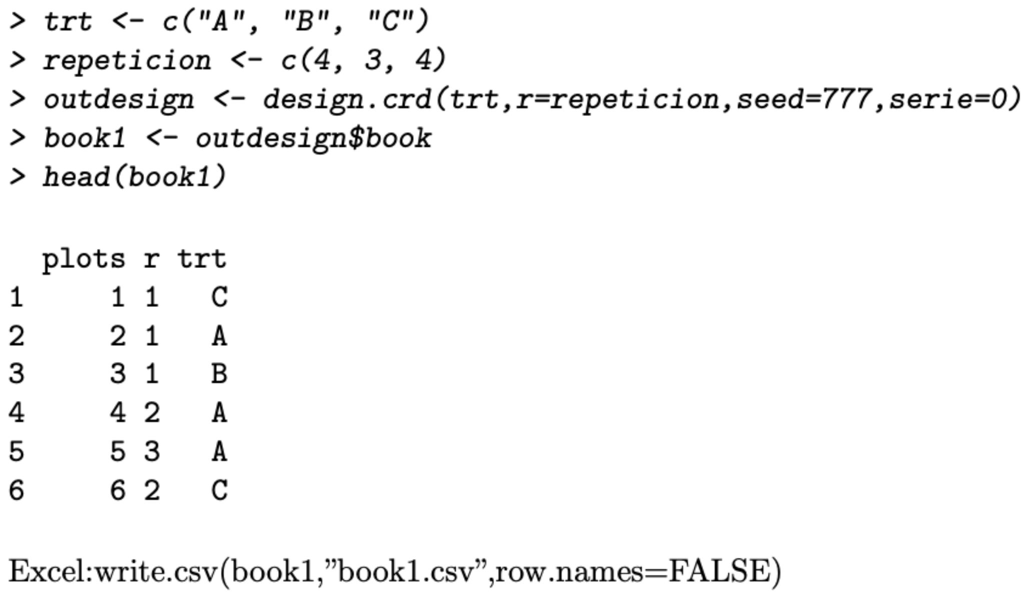
決定新研究方向

推論統計

Quit

Agricolae for Experimental Design

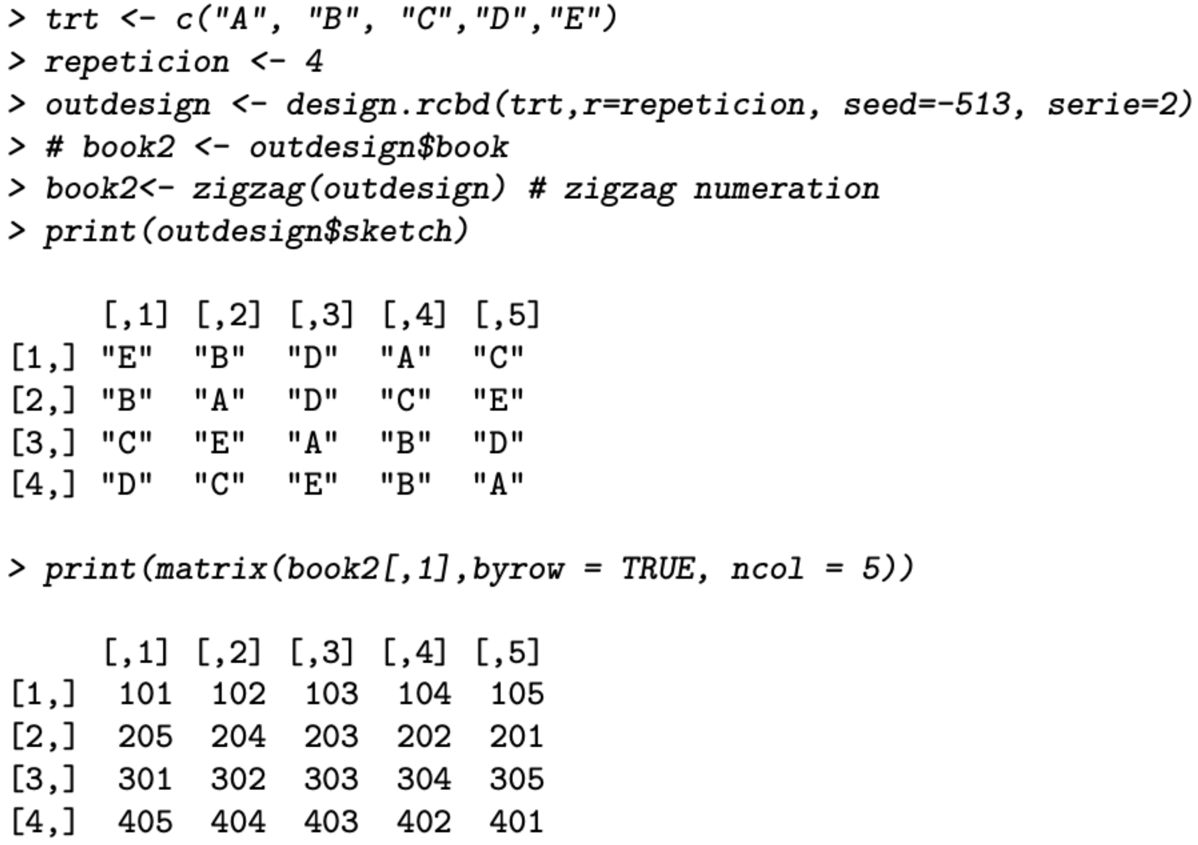
Completely randomized design (CRD)



```
> trt <- c("A", "B", "C")
> repeticion <- c(4, 3, 4)
> outdesign <- design.crd(trt,r=repeticion,seed=777,serie=0)
> book1 <- outdesign$book
> head(book1)

  plots r trt
1     1 1   C
2     2 1   A
3     3 1   B
4     4 2   A
5     5 3   A
6     6 2   C
```

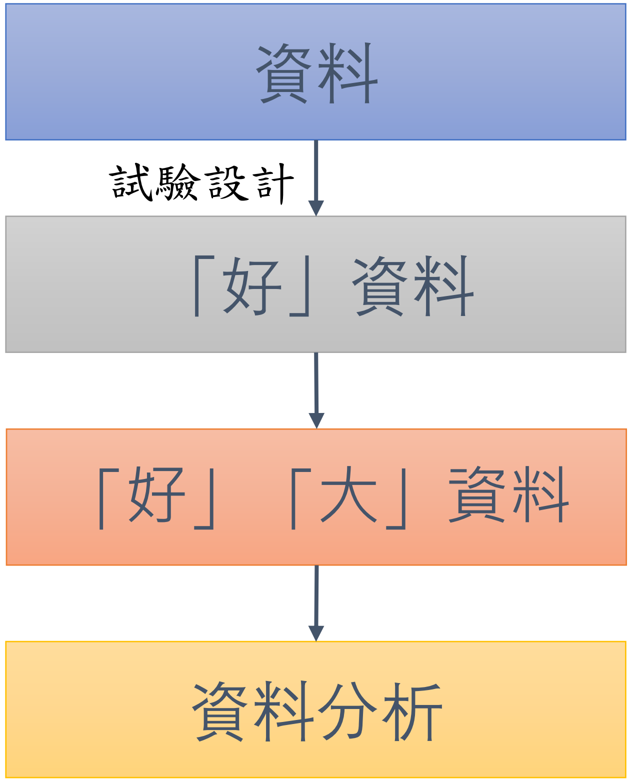
Excel:write.csv(book1,”book1.csv”,row.names=FALSE)

Randomized complete block design (RCBD)



```
> trt <- c("A", "B", "C","D","E")
> repeticion <- 4
> outdesign <- design.rcbd(trt,r=repeticion, seed=-513, serie=2)
> # book2 <- outdesign$book
> book2<- zigzag(outdesign) # zigzag numeration
> print(outdesign$sketch)

     [,1] [,2] [,3] [,4] [,5]
[1,] "E"  "B"  "D"  "A"  "C"
[2,] "B"  "A"  "D"  "C"  "E"
[3,] "C"  "E"  "A"  "B"  "D"
[4,] "D"  "C"  "E"  "B"  "A"

> print(matrix(book2[,1],byrow = TRUE, ncol = 5))

     [,1] [,2] [,3] [,4] [,5]
[1,]  101  102  103  104  105
[2,]  205  204  203  202  201
[3,]  301  302  303  304  305
[4,]  405  404  403  402  401
```

從資料收集出發


資料

試驗設計

「好」資料

「好」「大」資料

資料分析

氮肥用量與產量：

氮肥 (Kg) (X)	0	0.5	1.0	1.5	2.0	2.5
產量 (Kg) (Y)	10	18	32	48	55	62

✓

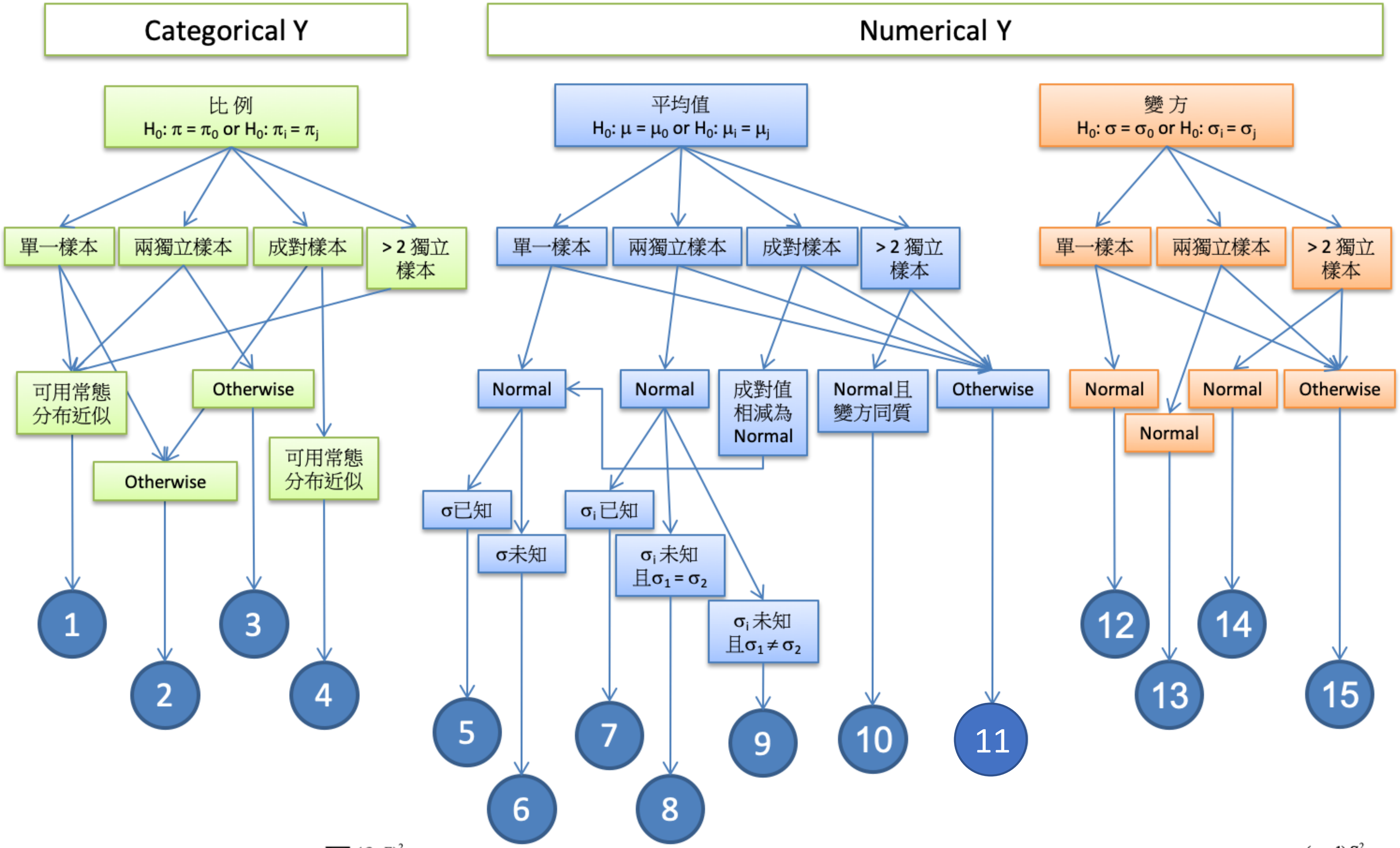
氮肥用量 (X)	低	中	高
產量 (Kg) (Y)	10	18	32

✗

Y (response)	X (predictor)	
	類別變數	連續變數
類別變數	列聯表 / 比率資料分析	廣義線性模式 (logistic regression, etc.)
連續變數	平均數檢定	迴歸與相關

列聯表 / 比率資料分析

平均數檢定



Categorical Y

比 例

$H_0: \pi = \pi_0$ or $H_0: \pi_i = \pi_j$

單一樣本

兩獨立樣本

成對樣本

> 2 獨立樣本

可用常態分布近似

Otherwise

Otherwise

可用常態分布近似

1

2

3

4

Numerical Y

平均值

$H_0: \mu = \mu_0$ or $H_0: \mu_i = \mu_j$

單一樣本

兩獨立樣本

成對樣本

> 2 獨立樣本

Normal

Normal

成對值相減為Normal

Normal且變方同質

Otherwise

σ已知

σ未知

σ_i 已知

σ_i 未知且 $\sigma_1 = \sigma_2$

σ_i 未知且 $\sigma_1 \neq \sigma_2$

5

6

7

8

9

10

11

變 方

$H_0: \sigma = \sigma_0$ or $H_0: \sigma_i = \sigma_j$

單一樣本

兩獨立樣本

> 2 獨立樣本

Normal

Normal

Normal

Otherwise

12

13

14

15

1: z test or chi-squared test $\sum \frac{(O-E)^2}{E}$
2: Binomial test
3: Fisher's exact test
4: McNemar test

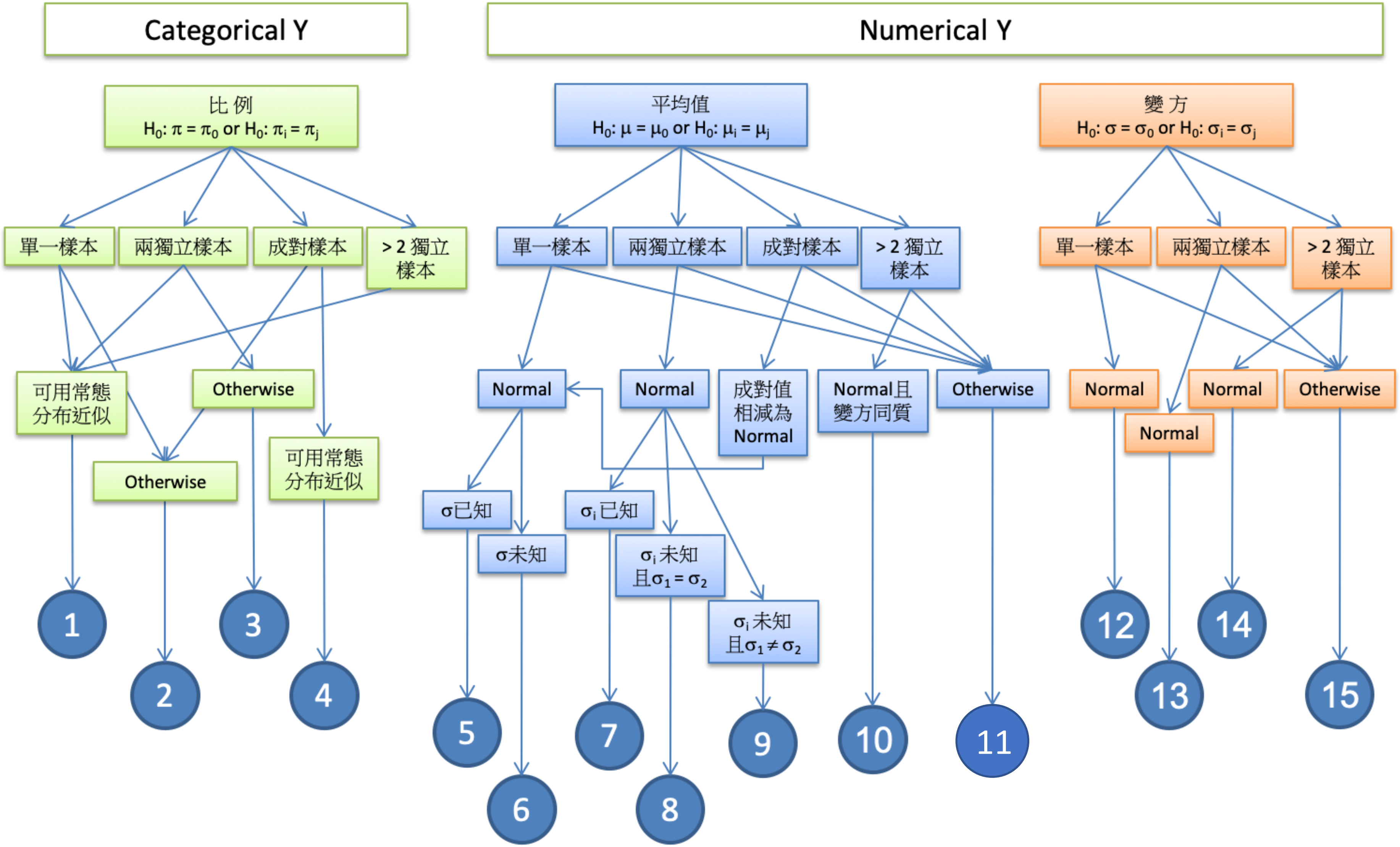
5: one-sample z test
6: one-sample t test
7: two-sample z test
8: two-sample t-test
9: Welch two-sample t-test

10: 變方分析 (ANOVA)
11: 變數轉換或無母數法

12: chi-squared test $\frac{(n-1)S^2}{\sigma^2}$
13: F-test
14: Bartlett's test
15: Levene's test

列聯表 / 比率資料分析

平均數檢定



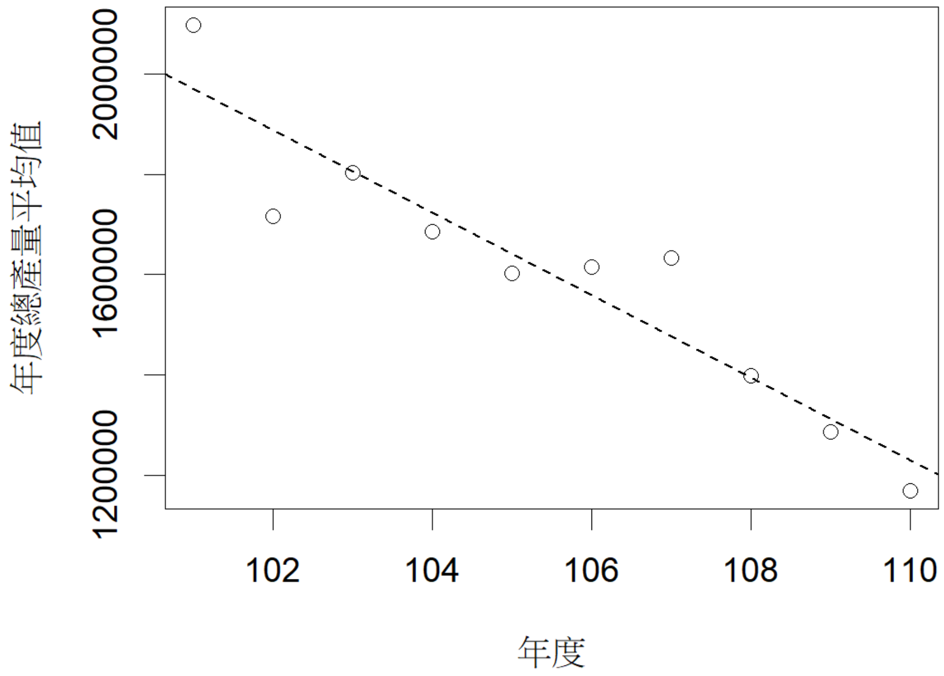
1: prop.test or chisq.test
2: binom.test
3: fisher.test
4: mcnemar.test

5: N/A [qnorm; pnorm]
6: t.test
7: t.test
8: t.test
9: t.test

10: aov or lm
11: boxcox (MASS);
kruskal.test

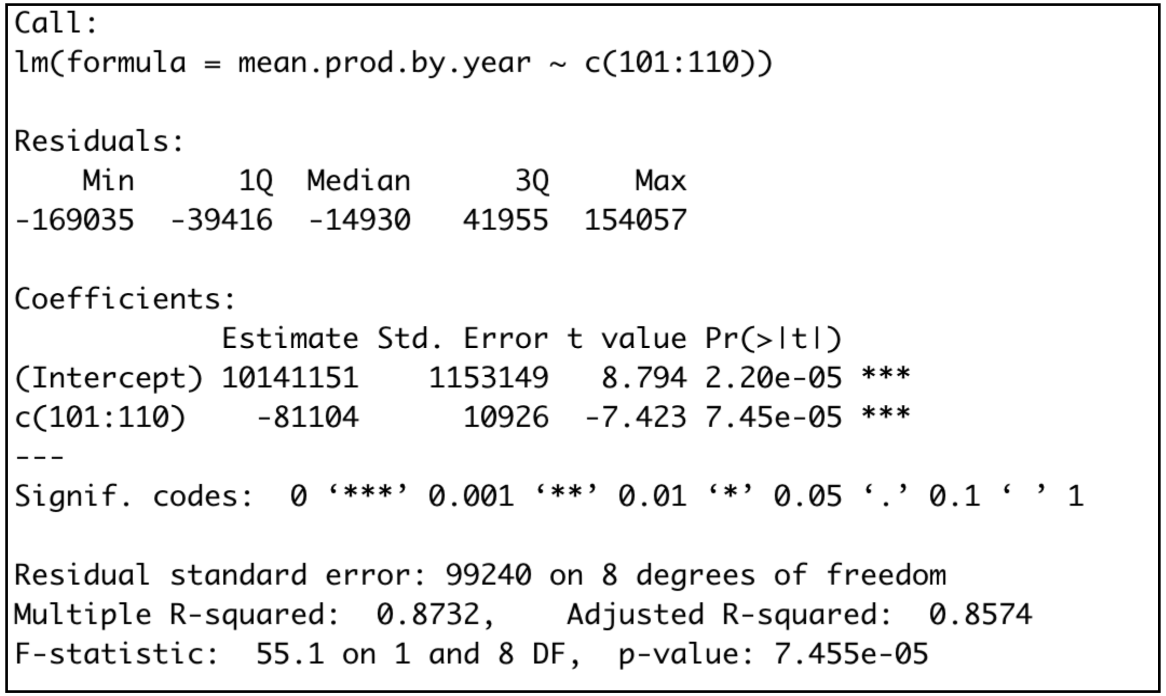
12: N/A [qchisq]
13: var.test
14: bartlett.test
15: leveneTest (car)

迴歸與相關

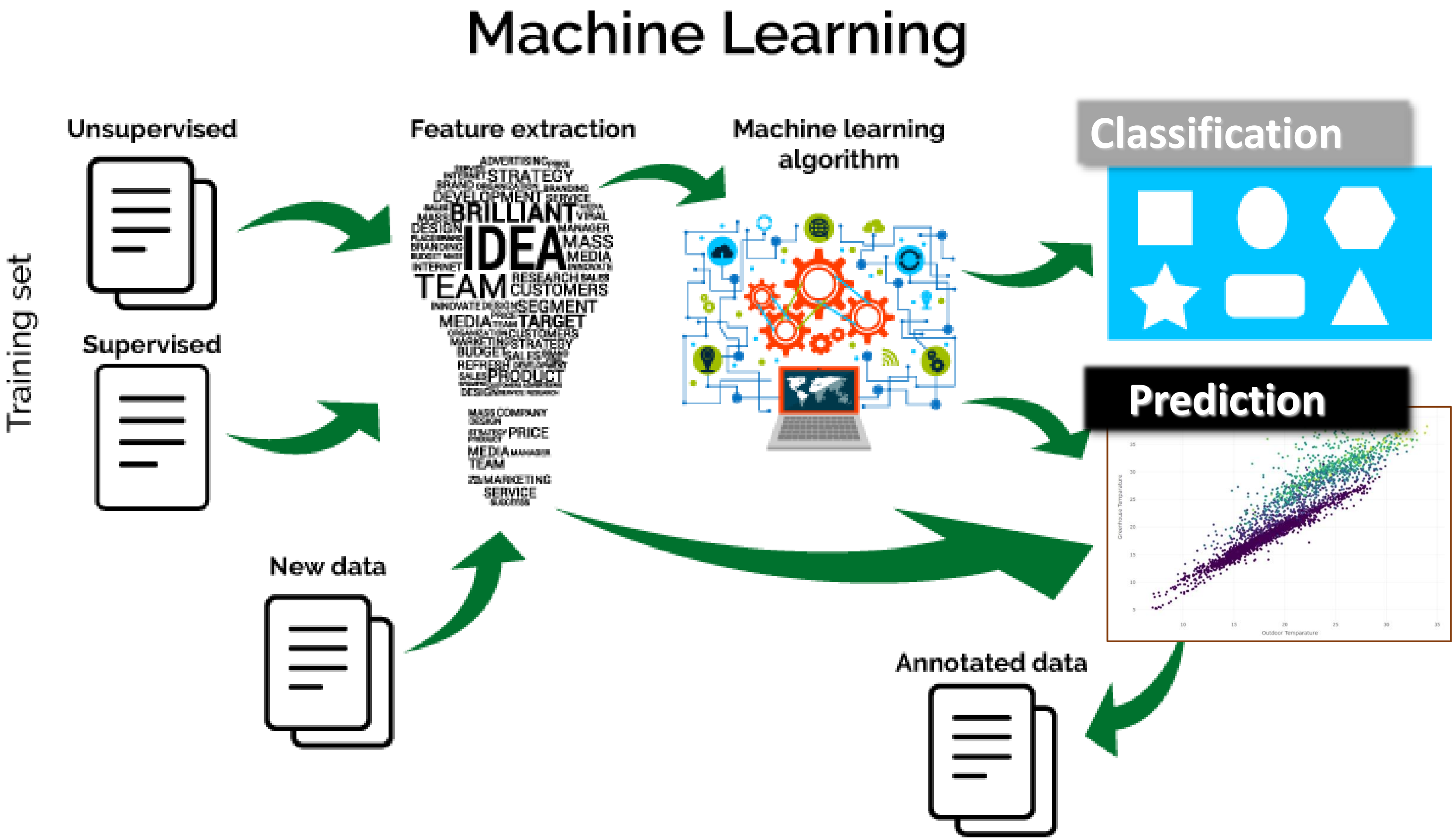


圖、西瓜總收量年度平均值散佈圖。虛線為最小平方法最佳配適直線：年度平均總收量估計值 = 10246280 – 82011 × 年度，相關係數 r = -0.94。

```
cor(mean.yield.by.year, c(101:110))
Out = lm(mean.yield.by.year ~ c(101:110))
summary(Out)
plot(101:110,mean.yield.by.year,
   xlab="年度",ylab="年度單位面積產量平均值")
abline(lm(mean.yield.by.year~c(101:110)),lty=2,lwd=1.5)
```



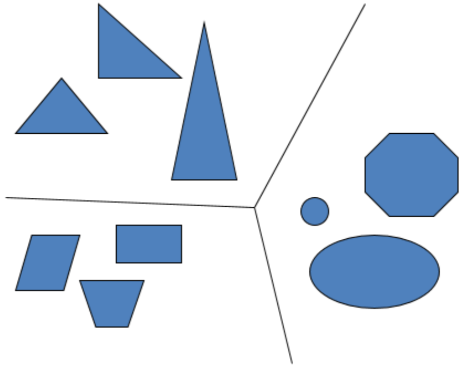
```
Call:
lm(formula = mean.prod.by.year ~ c(101:110))

Residuals:
    Min      1Q  Median      3Q     Max 
-169035  -39416  -14930   41955  154057 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 10141151    1153149   8.794 2.20e-05 ***
c(101:110)    -81104      10926  -7.423 7.45e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 99240 on 8 degrees of freedom
Multiple R-squared:  0.8732,    Adjusted R-squared:  0.8574 
F-statistic:  55.1 on 1 and 8 DF,  p-value: 7.455e-05
```



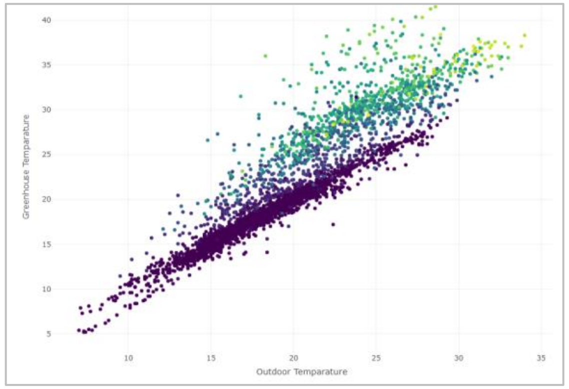
https://www.pantechsolutions.net/blog/machine-learning-projects-and-ideas/

Objectives of Machine Learning

- **Classification:**
  - Discover clusters of samples having similar patterns in features

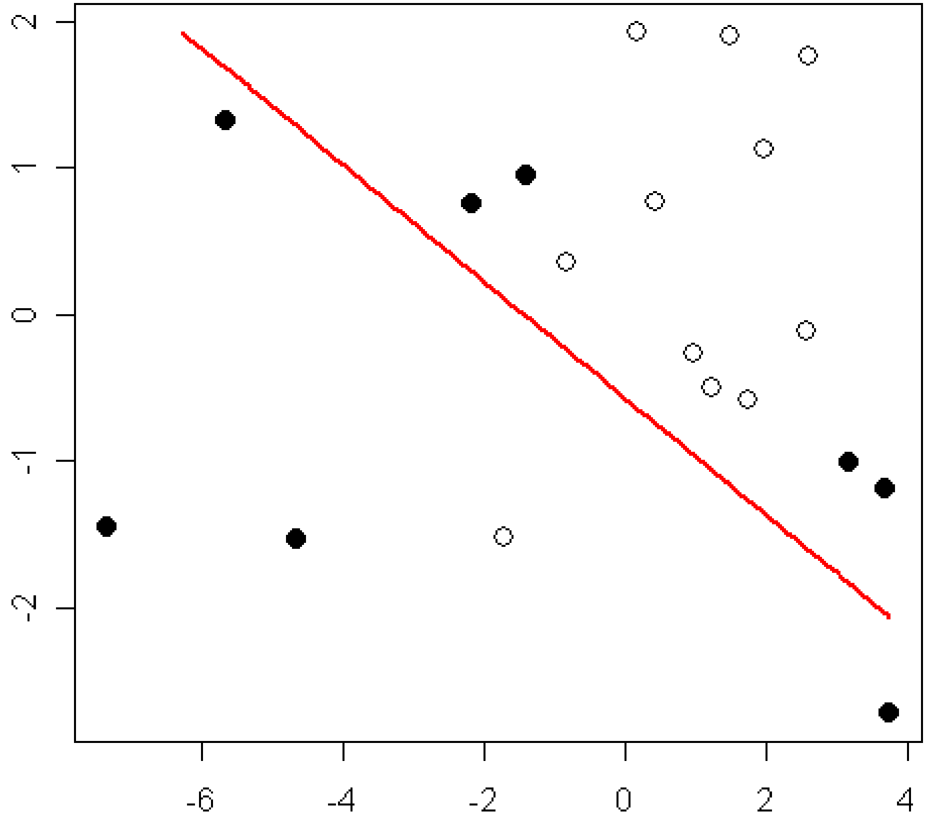


- **Prediction:**
  - Build up a predictive model to represent a continuous response of the target variable.

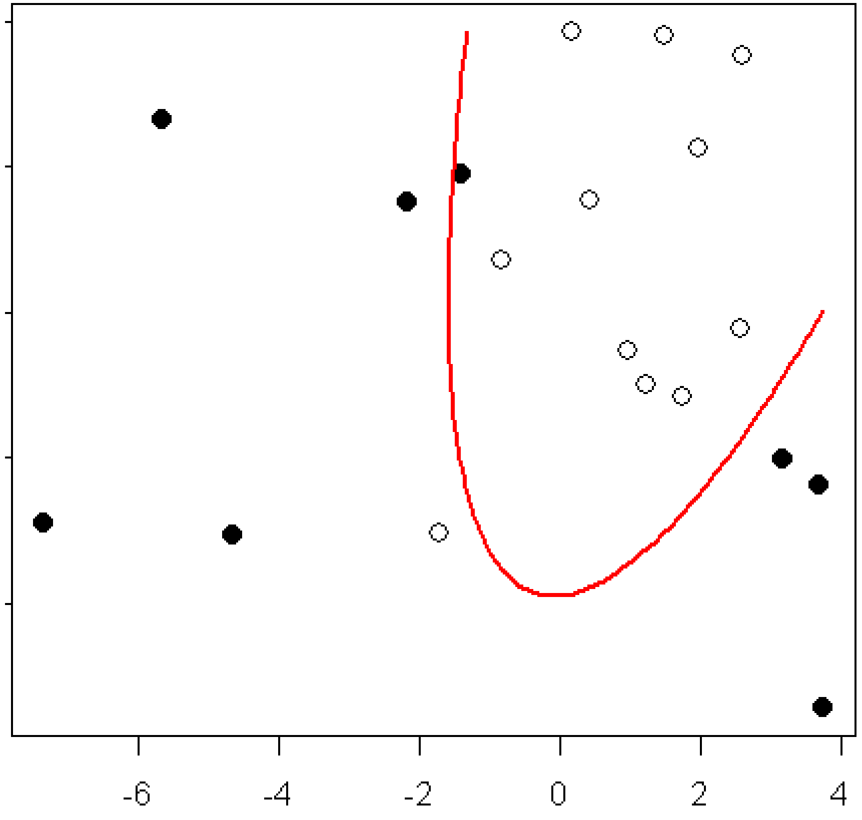


Linear vs Quadratic Discriminant Analysis for Classification

LDA MASS::lda



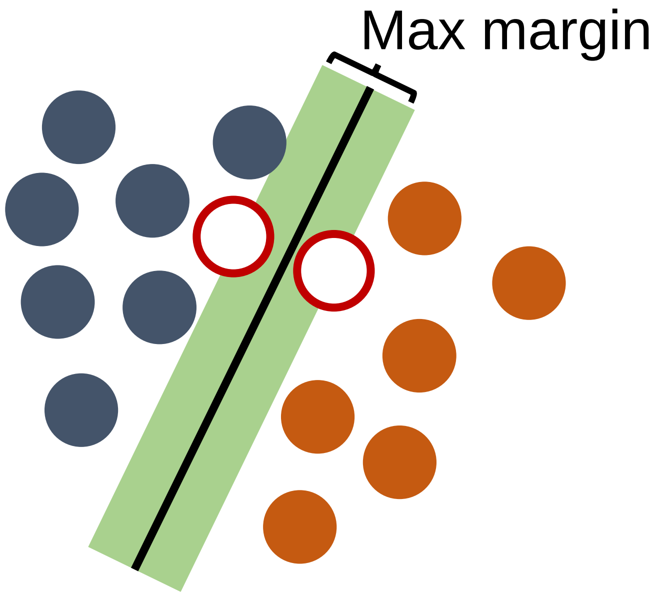
QDA MASS::qda



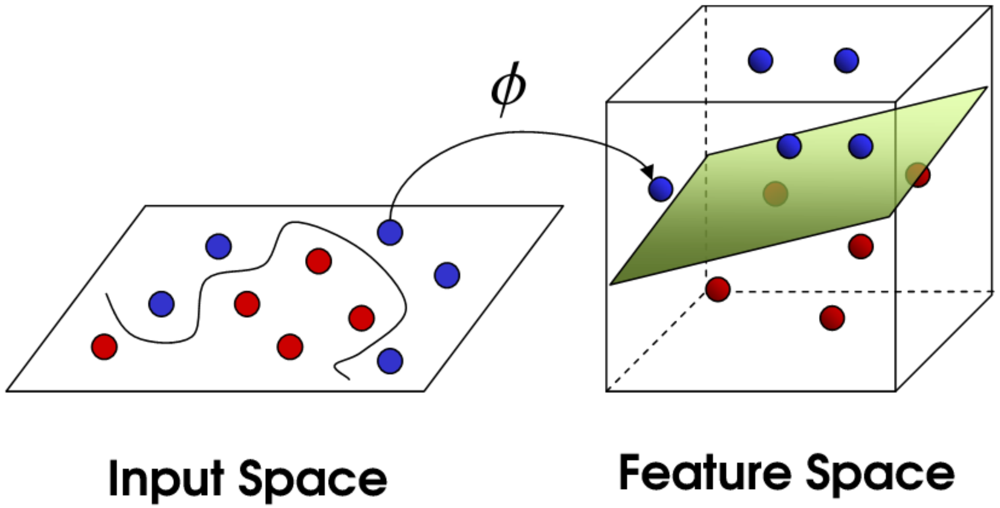
Support Vector Machine (SVM)

e1071::svm

Linear classification:



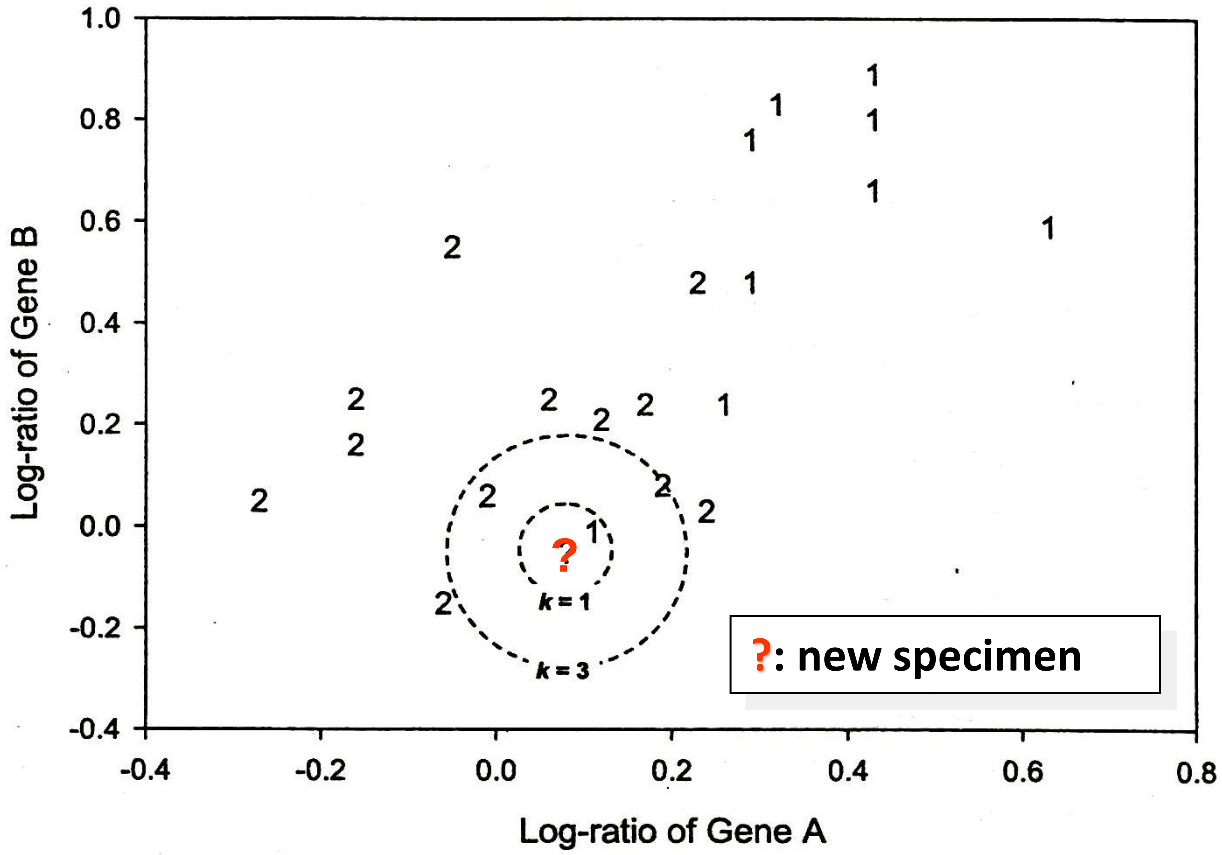
Non-linear classification:



Multiple Kernel learning [Alpaydin, Chapter 13.8]

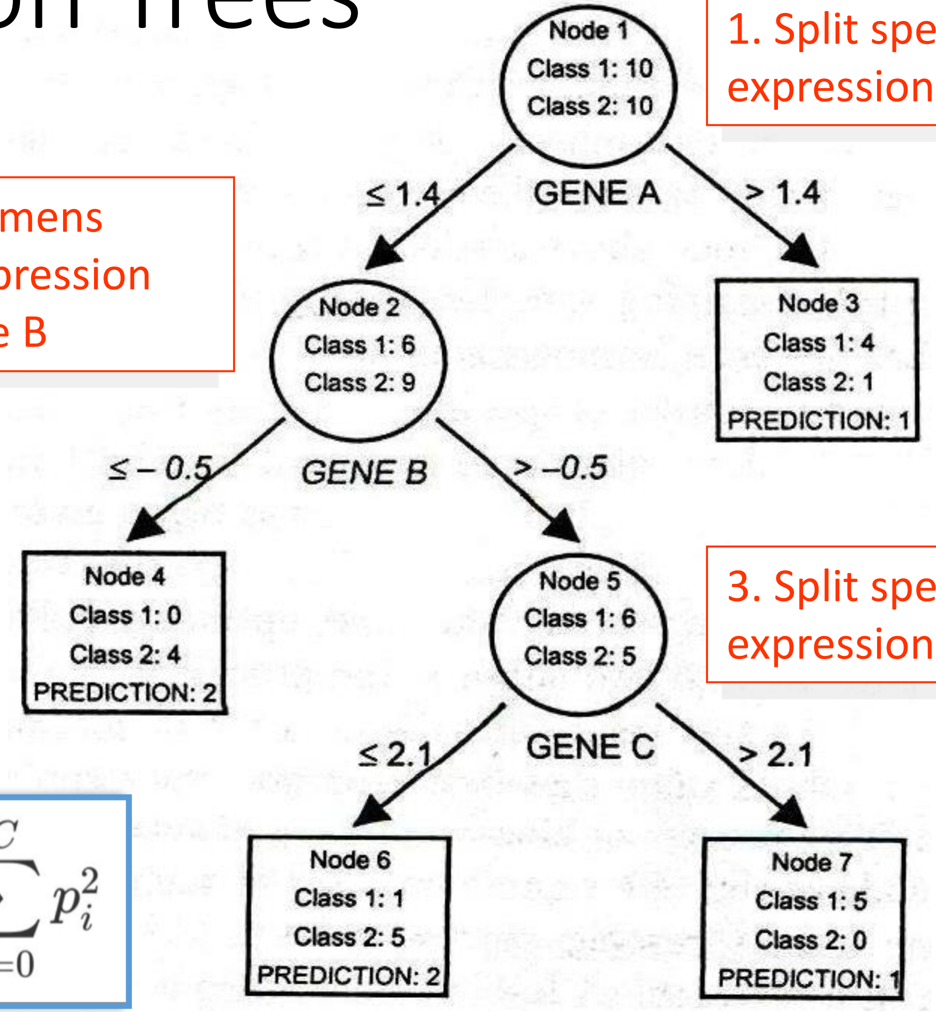
Nearest neighbor classification (kNN)

class::knn



Classification Trees

rpart::rpart



1. Split specimens based on expression level of Gene A

2. Split specimens based on expression level of Gene B

3. Split specimens based on expression level of Gene C

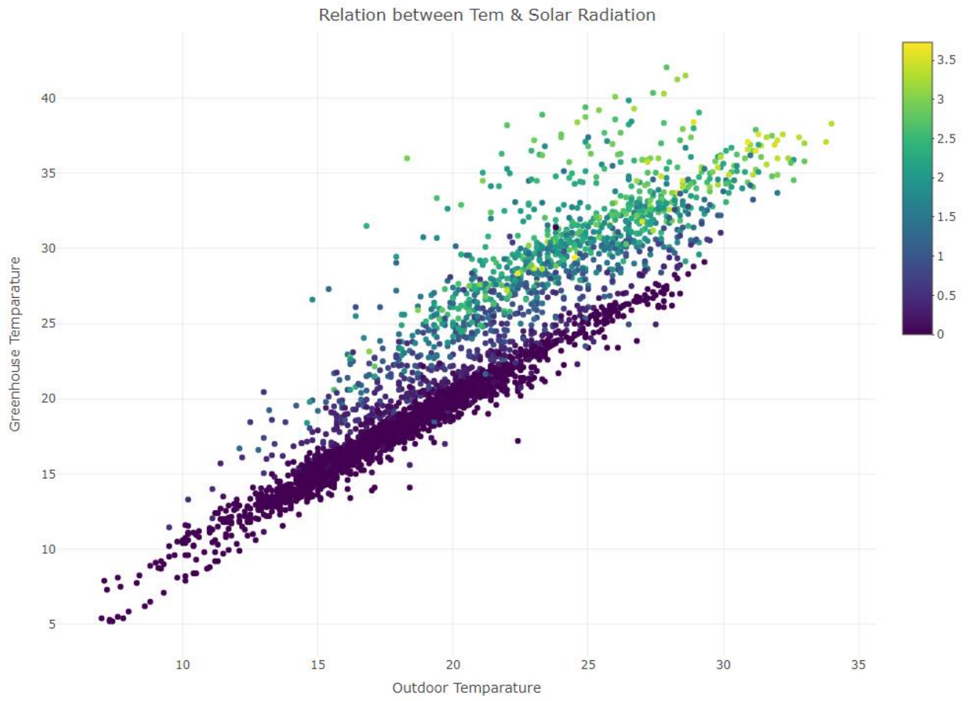
Gini Index

$$G = 1 - \sum_{i=0}^{C} p_i^2$$

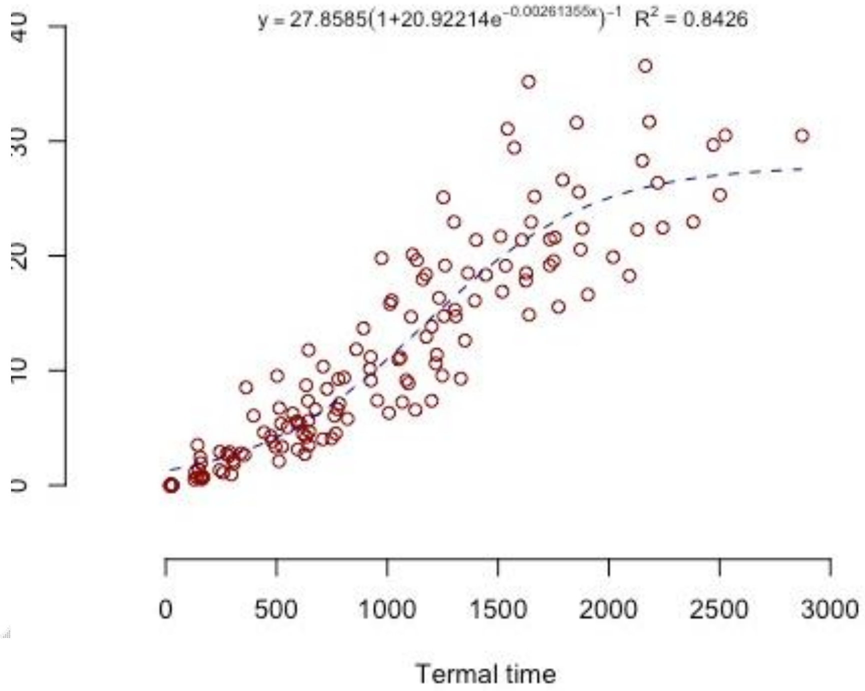
Extensions:
Random Forest
XGBoost

Regression Model (Prediction)

stats::lm



Linear Multiple Regression Model

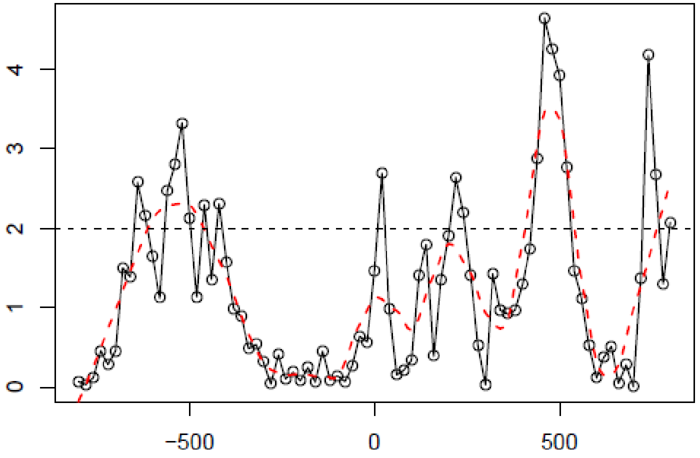


Nonlinear Gompetz Model

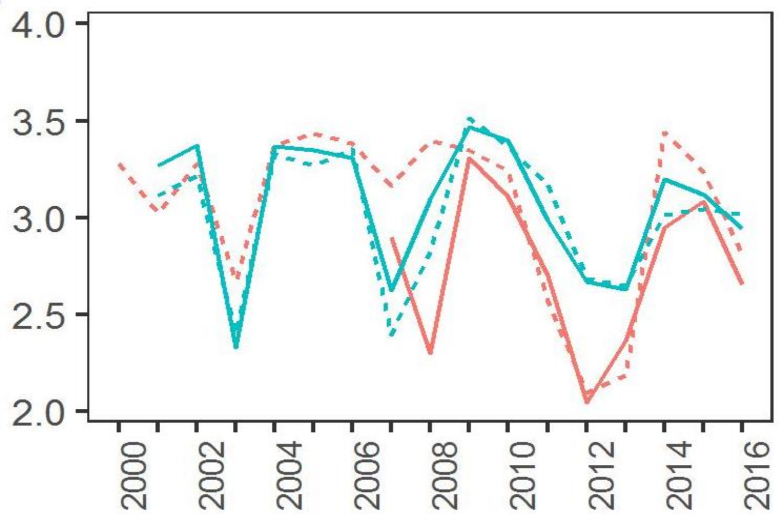
Other Models for Prediction

stats::loess

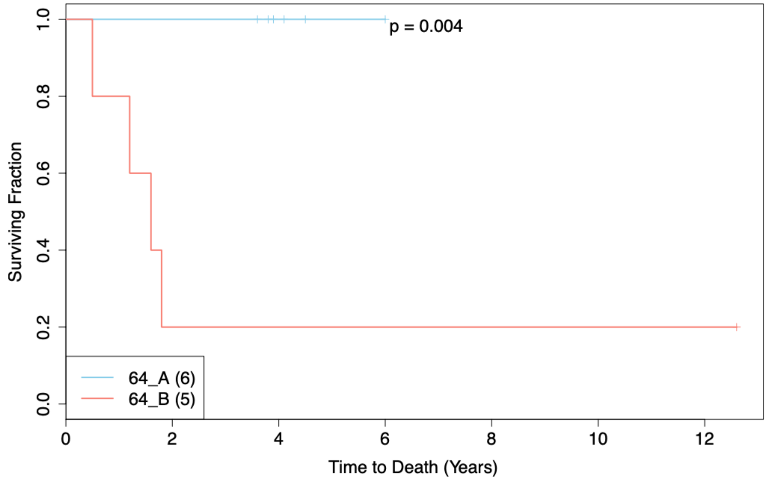
LOESS



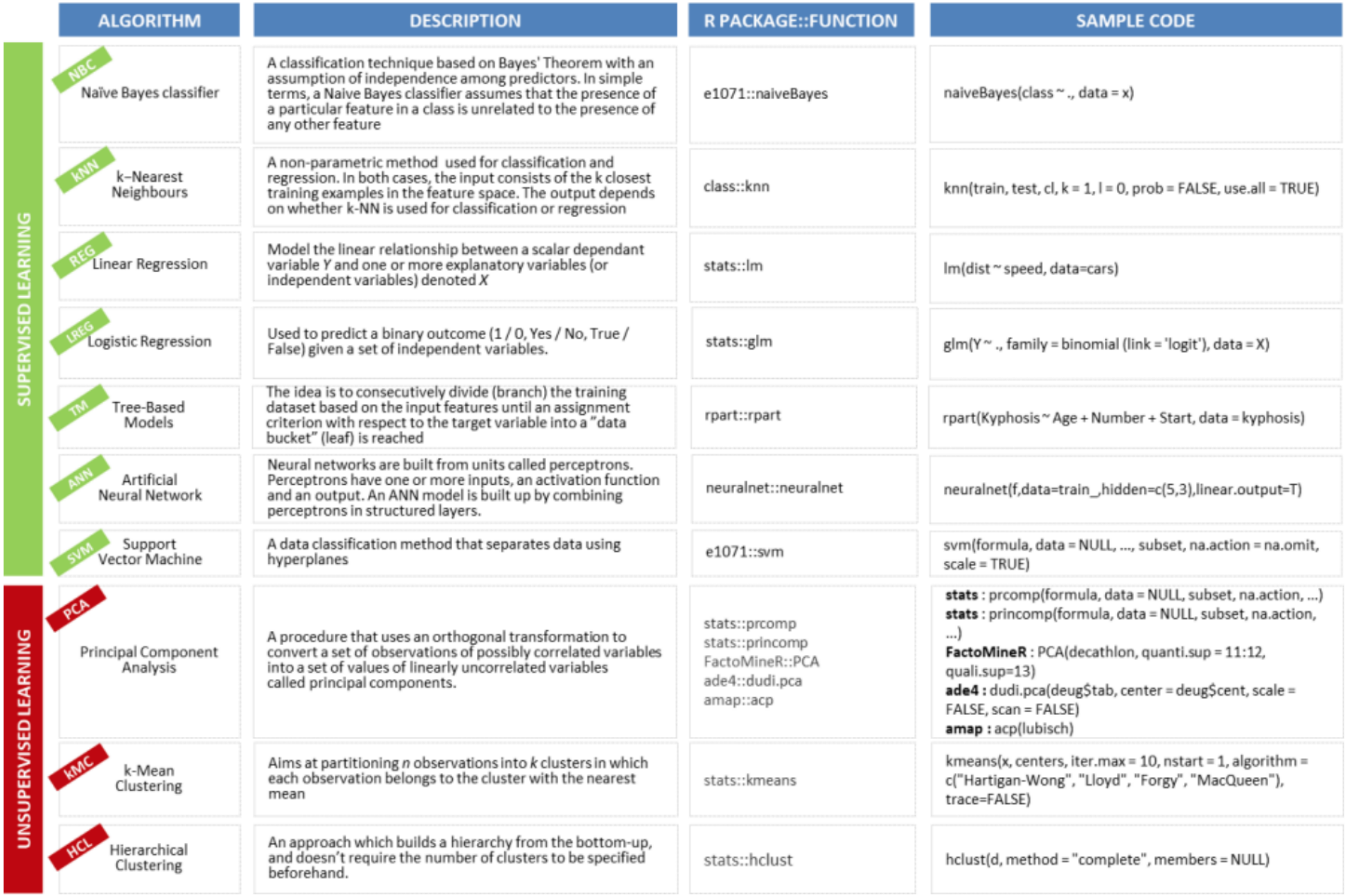
Time-series



Survival

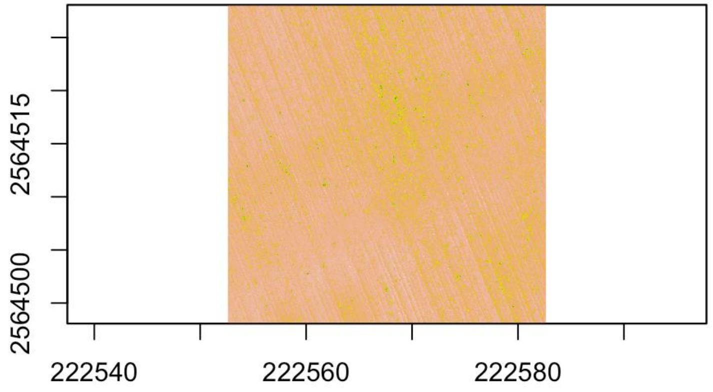


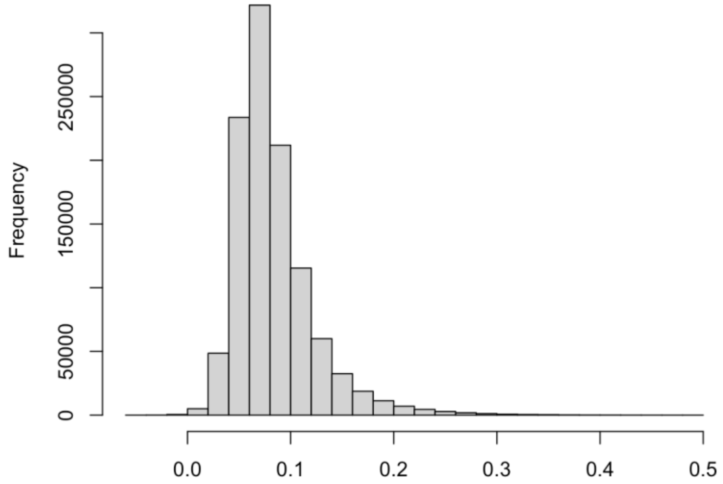
https://raw.githubusercontent.com/rstudio/cheatsheets/main/Machine%20Learning%20Modelling%20in%20R.pdf

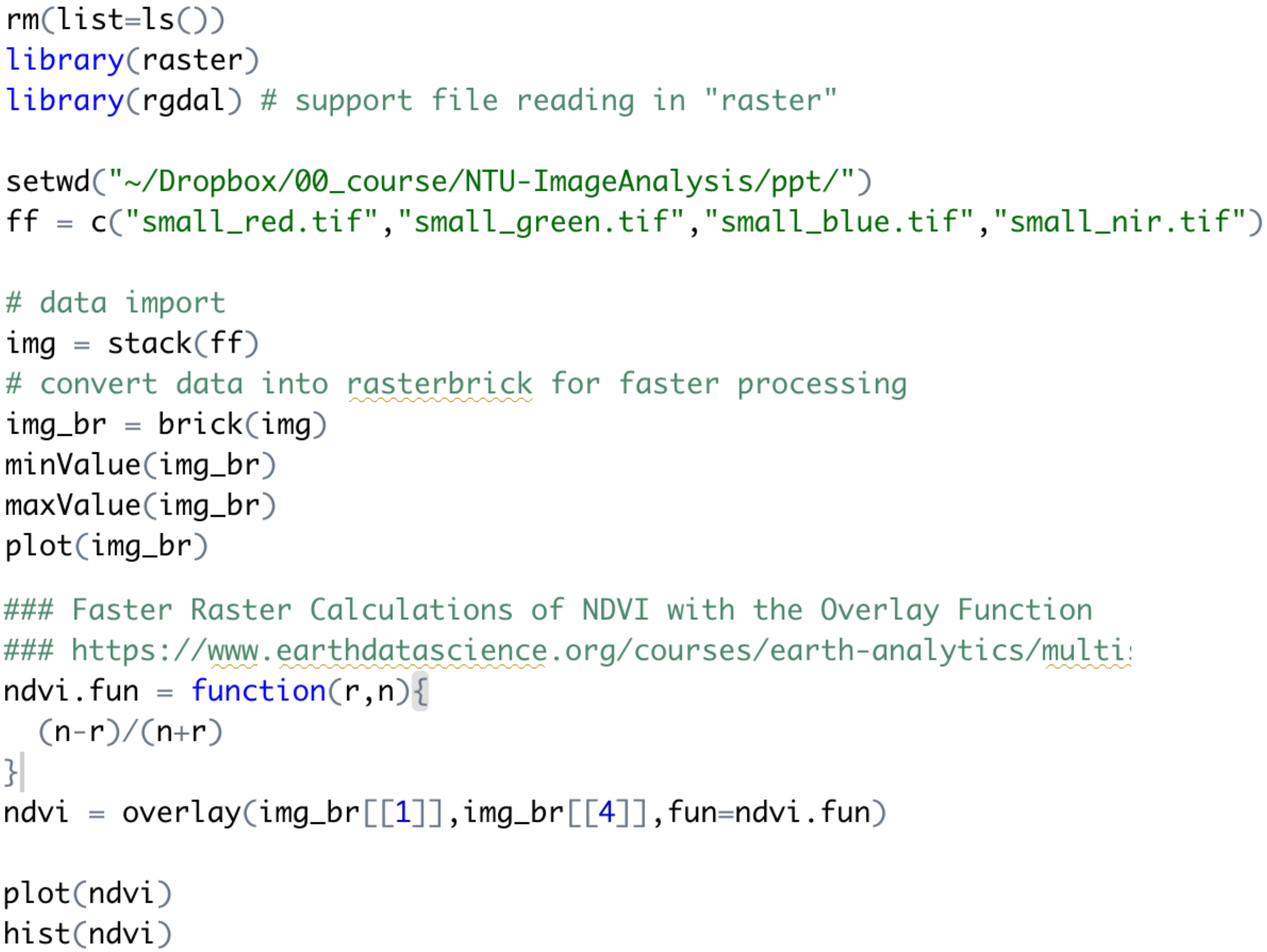


	ALGORITHM	DESCRIPTION	R PACKAGE::FUNCTION	SAMPLE CODE
SUPERVISED LEARNING	NBC Naïve Bayes classifier	A classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature	e1071::naiveBayes	naiveBayes(class ~ ., data = x)
	kNN k–Nearest Neighbours	A non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression	class::knn	knn(train, test, cl, k = 1, l = 0, prob = FALSE, use.all = TRUE)
	REG Linear Regression	Model the linear relationship between a scalar dependant variable *Y* and one or more explanatory variables (or independent variables) denoted *X*	stats::lm	lm(dist ~ speed, data=cars)
	LREG Logistic Regression	Used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables.	stats::glm	glm(Y ~ ., family = binomial (link = 'logit'), data = X)
	TM Tree-Based Models	The idea is to consecutively divide (branch) the training dataset based on the input features until an assignment criterion with respect to the target variable into a "data bucket" (leaf) is reached	rpart::rpart	rpart(Kyphosis ~ Age + Number + Start, data = kyphosis)
	ANN Artificial Neural Network	Neural networks are built from units called perceptrons. Perceptrons have one or more inputs, an activation function and an output. An ANN model is built up by combining perceptrons in structured layers.	neuralnet::neuralnet	neuralnet(f,data=train_,hidden=c(5,3),linear.output=T)
	SVM Support Vector Machine	A data classification method that separates data using hyperplanes	e1071::svm	svm(formula, data = NULL, ..., subset, na.action = na.omit, scale = TRUE)
UNSUPERVISED LEARNING	PCA Principal Component Analysis	A procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.	stats::prcomp stats::princomp FactoMineR::PCA ade4::dudi.pca amap::acp	**stats** : prcomp(formula, data = NULL, subset, na.action, ...) **stats** : princomp(formula, data = NULL, subset, na.action, ...) **FactoMineR** : PCA(decathlon, quanti.sup = 11:12, quali.sup=13) **ade4 :** dudi.pca(deug$tab, center = deug$cent, scale = FALSE, scan = FALSE) **amap :** acp(lubisch)
	kMC k-Mean Clustering	Aims at partitioning *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean	stats::kmeans	kmeans(x, centers, iter.max = 10, nstart = 1, algorithm = c("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen"), trace=FALSE)
	HCL Hierarchical Clustering	An approach which builds a hierarchy from the bottom-up, and doesn't require the number of clusters to be specified beforehand.	stats::hclust	hclust(d, method = "complete", members = NULL)

影像辨識







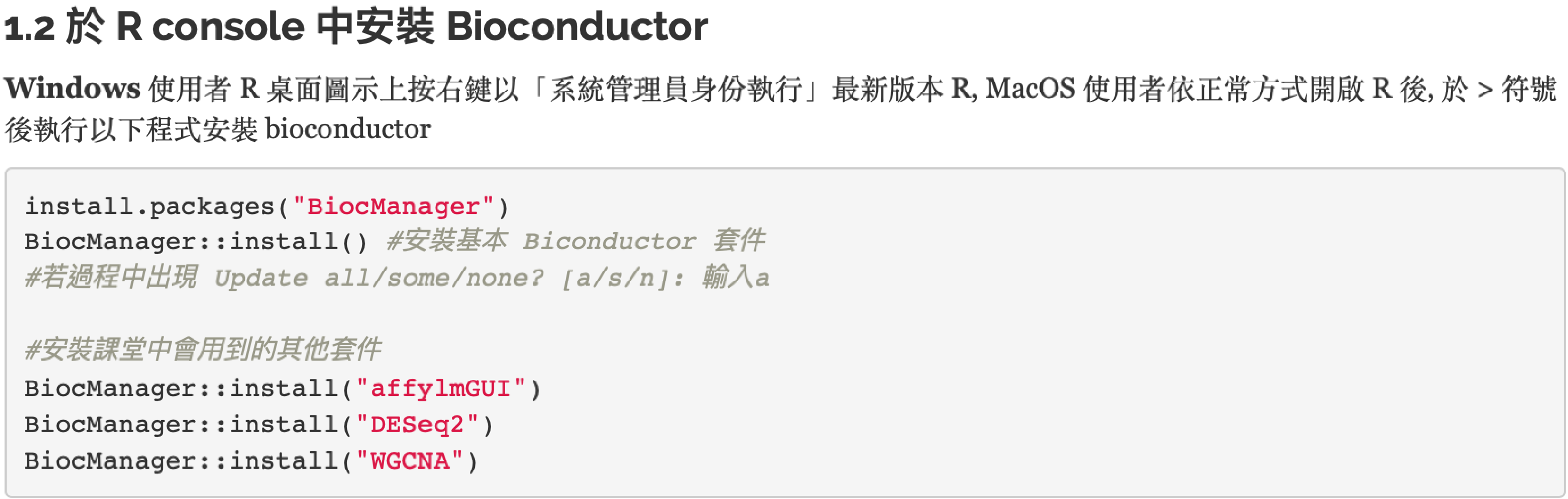
```r
rm(list=ls())
library(raster)
library(rgdal) # support file reading in "raster"

setwd("~/Dropbox/00_course/NTU-ImageAnalysis/ppt/")
ff = c("small_red.tif","small_green.tif","small_blue.tif","small_nir.tif")

# data import
img = stack(ff)
# convert data into rasterbrick for faster processing
img_br = brick(img)
minValue(img_br)
maxValue(img_br)
plot(img_br)

### Faster Raster Calculations of NDVI with the Overlay Function
### https://www.earthdatascience.org/courses/earth-analytics/multi
ndvi.fun = function(r,n){
  (n-r)/(n+r)
}
ndvi = overlay(img_br[[1]],img_br[[4]],fun=ndvi.fun)

plot(ndvi)
hist(ndvi)
```

BioConductor

- BioConductor is the R software project for the analysis of biomedical and genomic data
  - Microarrays
  - Genome sequence data
  - Pathway graphs

https://www.bioconductor.org/

BioConductor



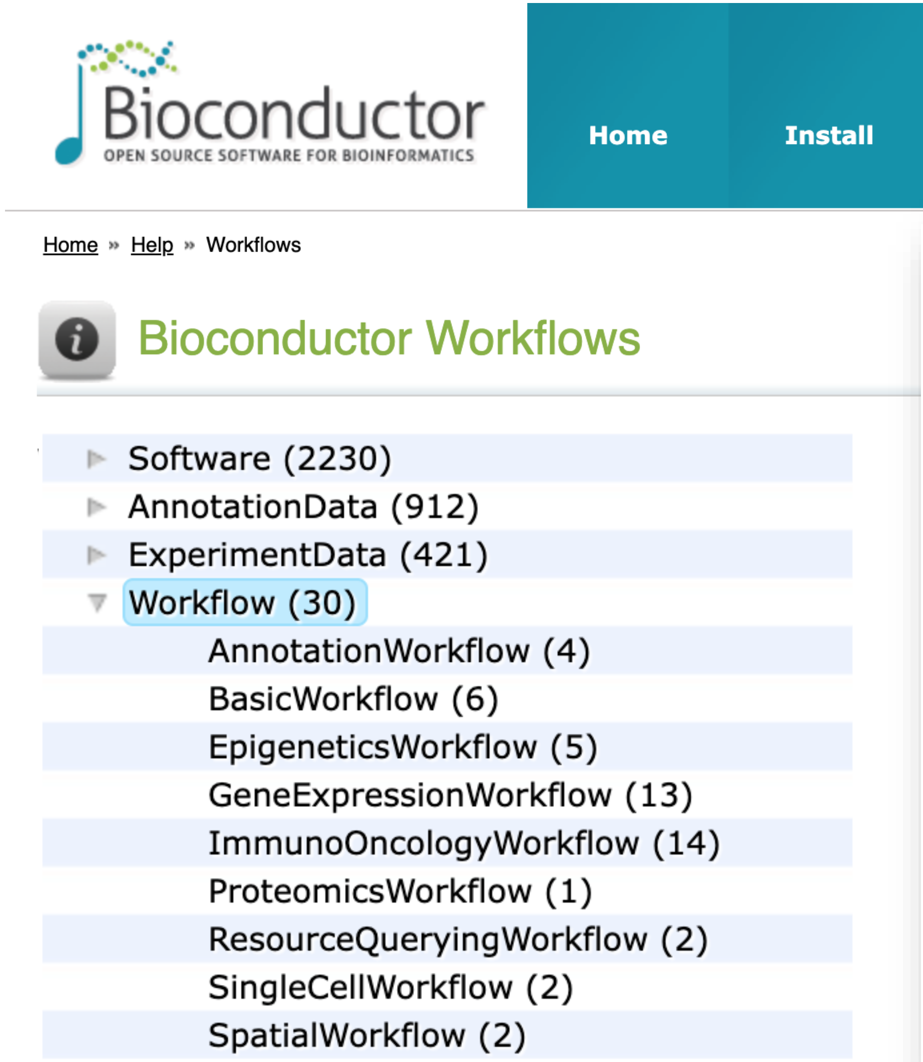
1.2 於 R console 中安裝 Bioconductor

Windows 使用者 R 桌面圖示上按右鍵以「系統管理員身份執行」最新版本 R, MacOS 使用者依正常方式開啟 R 後, 於 > 符號後執行以下程式安裝 bioconductor

```
install.packages("BiocManager")
BiocManager::install() #安裝基本 Biconductor 套件
#若過程中出現 Update all/some/none? [a/s/n]: 輸入a

#安裝課堂中會用到的其他套件
BiocManager::install("affylmGUI")
BiocManager::install("DESeq2")
BiocManager::install("WGCNA")
```

Common Workflows in BioC

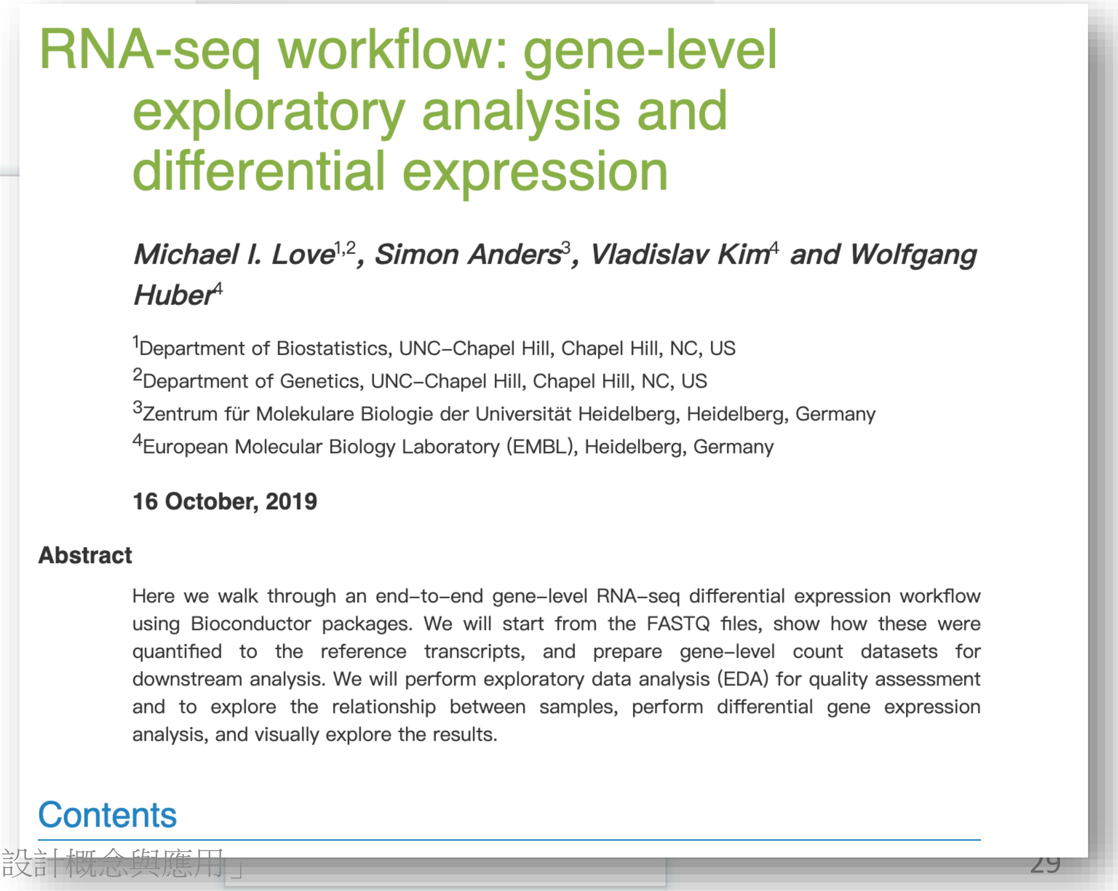
https://www.bioconductor.org/packages/release/BiocViews.html#___Workflow



Home » Help » Workflows

Bioconductor Workflows

- Software (2230)
- AnnotationData (912)
- ExperimentData (421)
- Workflow (30)
  - AnnotationWorkflow (4)
  - BasicWorkflow (6)
  - EpigeneticsWorkflow (5)
  - GeneExpressionWorkflow (13)
  - ImmunoOncologyWorkflow (14)
  - ProteomicsWorkflow (1)
  - ResourceQueryingWorkflow (2)
  - SingleCellWorkflow (2)
  - SpatialWorkflow (2)



RNA-seq workflow: gene-level exploratory analysis and differential expression

Michael I. Love[1,2], *Simon Anders*[3], *Vladislav Kim*[4] *and Wolfgang Huber*[4]

[1]Department of Biostatistics, UNC–Chapel Hill, Chapel Hill, NC, US
[2]Department of Genetics, UNC–Chapel Hill, Chapel Hill, NC, US
[3]Zentrum für Molekulare Biologie der Universität Heidelberg, Heidelberg, Germany
[4]European Molecular Biology Laboratory (EMBL), Heidelberg, Germany

16 October, 2019

Abstract

Here we walk through an end–to–end gene–level RNA–seq differential expression workflow using Bioconductor packages. We will start from the FASTQ files, show how these were quantified to the reference transcripts, and prepare gene–level count datasets for downstream analysis. We will perform exploratory data analysis (EDA) for quality assessment and to explore the relationship between samples, perform differential gene expression analysis, and visually explore the results.

Contents

Summary

- R 語言是具有彈性的工具，現在開始學習、永遠不嫌晚
- 師傅引進門、修行看個人
- 善用網路與AI工具，建立自我學習的能力
- 透過社群共學也是一個好途徑~