附件一

輔仁大學 107 年高教深耕計畫
【程式設計融入課程補助計畫】授課成效報告

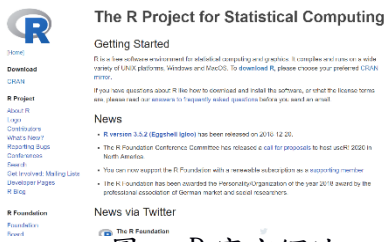
基本資料

開課學院	管理學院	開課系所	統計資訊學系
學年度/學期	107 學年度 / 第 1 學期	學制別	大學 ☑日間部 □進修部
課程名稱	迴歸分析	上課時間	星期_一_，09：10~12：00
開課單位	統資三乙	修課人數	81
授課教師	黃孝雲	聯絡電話	(手機)0963-239-576 (研究室分機)3940
電郵信箱	058029@mail.fju.edu.tw		

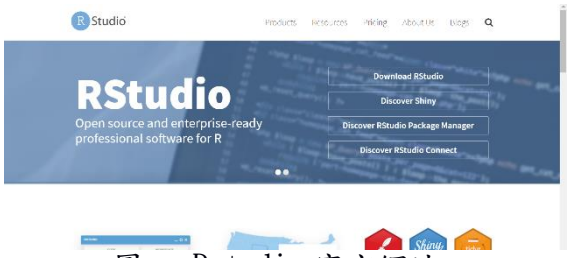
整體教學設計

跨域特色

大數據時代的來臨，資料分析成為最熱門的話題之一，迴歸分析為非常經典的資料分析的必修課程，與其他資料分析相同，若沒有程式設計做輔助，實際上純粹只是個概念與理論的課程，讓學生產生實際運用的能力則需搭配軟體，本計畫採用時下大數據十分新興的統計軟體 R(如圖一所示)，將 R 語言融入課程中，並以 Rstudio 作為我們的編輯器(如圖二所示)。



圖一 R 官方網站



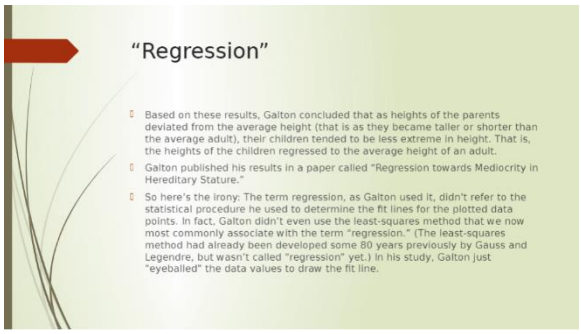
圖二 Rstudio 官方網站

程式語言

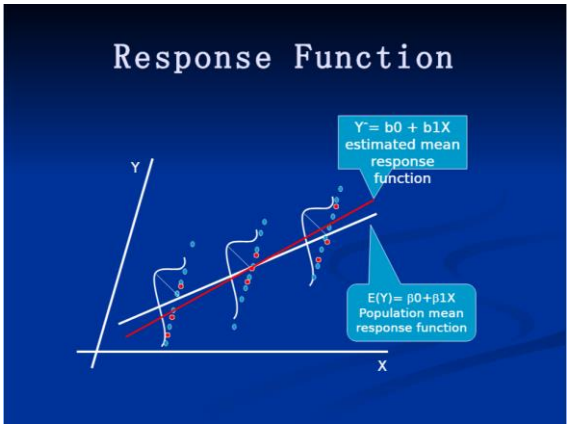
□ Python □ APP Inventor 2 ■ R □ Javascript □其他

教學目標

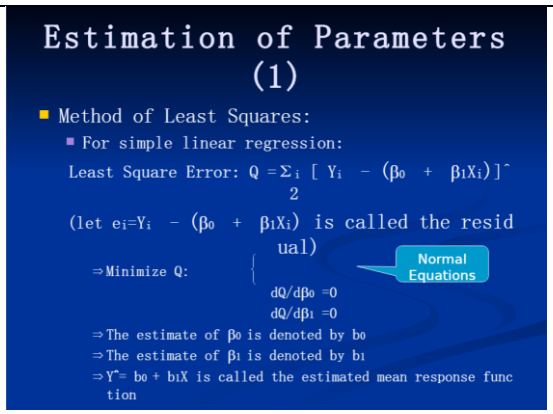
- **知識面目標** (期望學習者透過課程能習得哪些知識)：
  介紹迴歸模型的概念(如圖三至圖七所示)與推導(如圖八所示)，實際操作 R 語言建立迴歸模型。



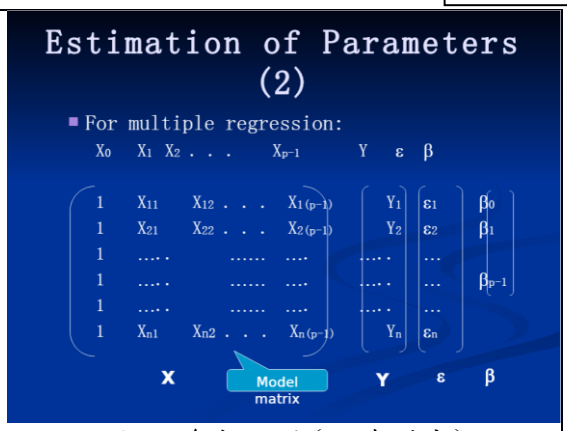
圖三 迴歸介紹



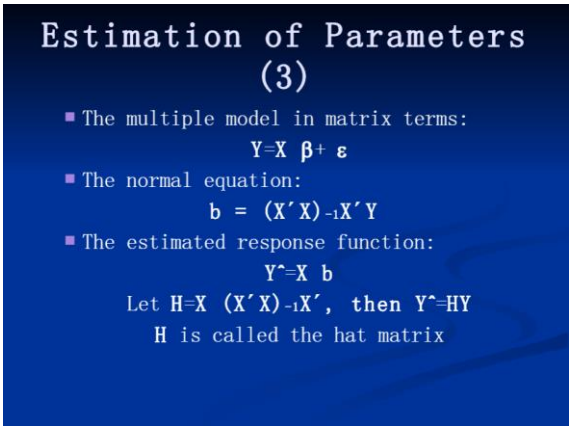
圖四 迴歸方程式



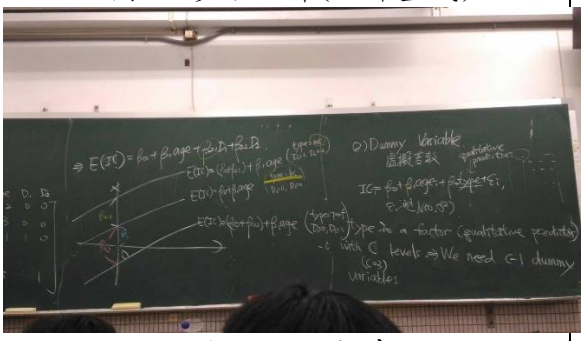
圖五 參數估計



圖六 參數估計(矩陣型式)

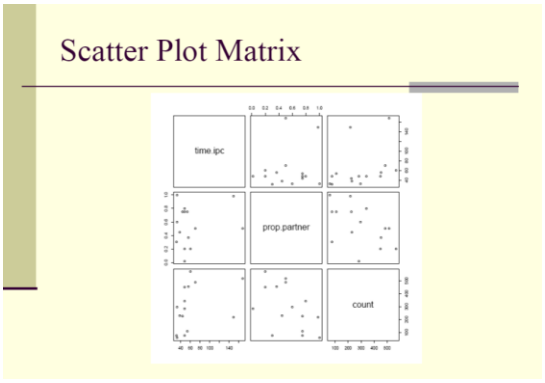


圖七 迴歸方程式(矩陣型式)

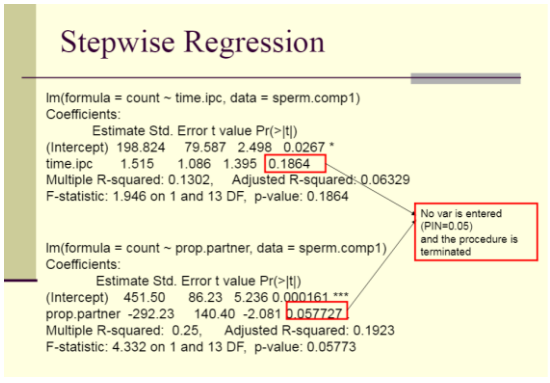


圖八 上課板書

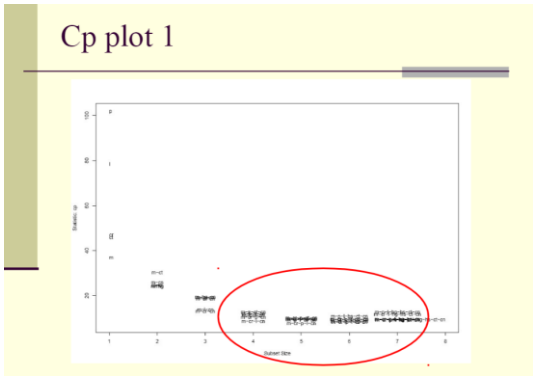
- **學科專業技能目標** (期望學習者透過課程能展現哪些學科專業技能)：
  學習如何建立迴歸模型，並將結果視覺化，合理的解釋迴歸模型及預測的結果(如圖九至圖十三所示)。



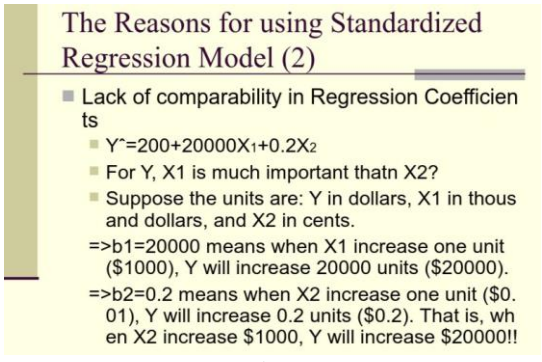
圖九 散佈圖矩陣



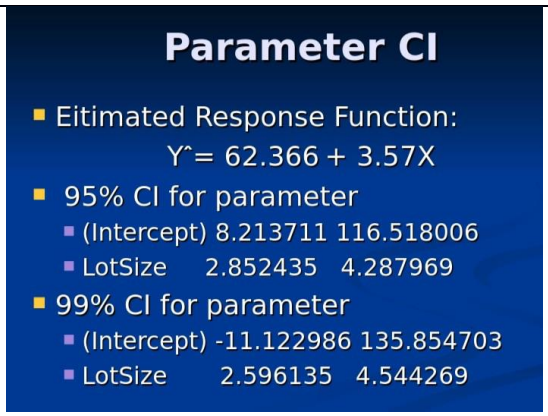
圖十 逐步回歸



圖十一 最佳子集法

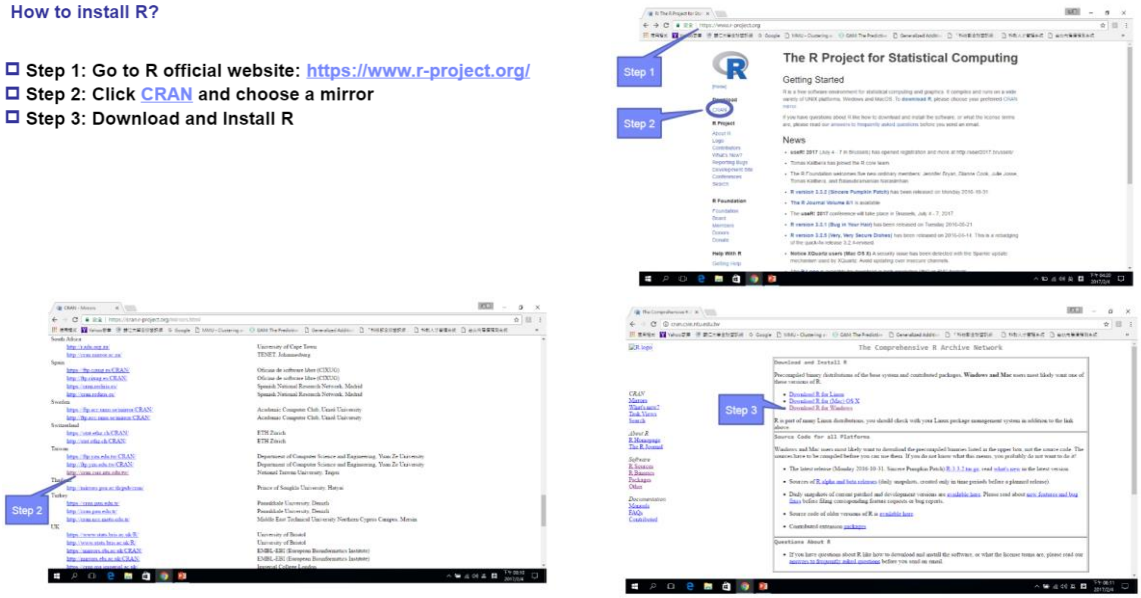


圖十二 標準化迴歸模型

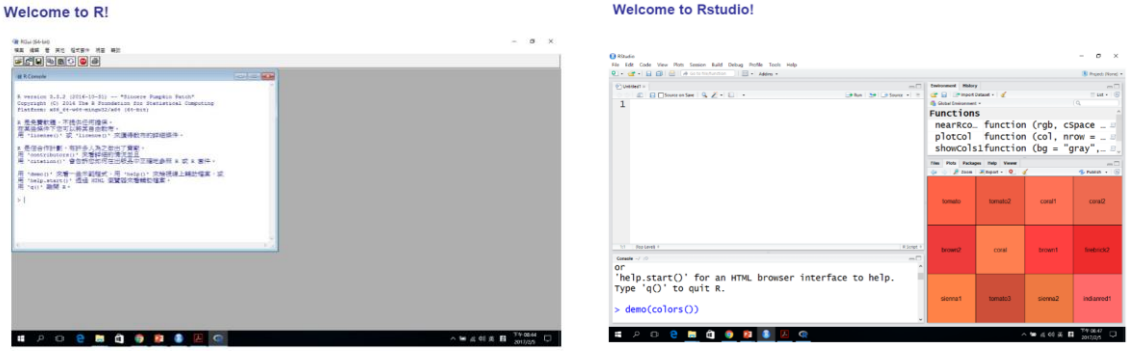


圖十三 參數估計信賴區間

- **程式設計技能目標** (期望學習者透過課程能展現那些程式設計技能)：
學習 R 語言的安裝(如圖十四所示)、操作介面(如圖十五所示)及基本語法(如圖十六至圖十八所示)，藉由 R 語言對真實資料進行資料預處理(如圖十九所示)，並使用 R 語言詳細討論建立迴歸模型的流程(如圖二十所示)。



圖十四 R 安裝步驟



圖十五 R 及 Rstudio 使用介面

Categorical data: Frequencies

- table() #Count factor variable frequency; K way table
- addmargins() # Adding row/col margins

- prop.table(,1) # Row proportions
- round(prop.table(,1), 2) # Round row prop to 2 digits
- round(100*prop.table(,1), 2) # Round row prop to 2 digits (percents)
- addmargins(round(prop.table(,1), 2),2) # Round row prop to 2 digits

- prop.table(,2) # Column proportions
- round(prop.table(,2), 2) # Round column prop to 2 digits
- round(100*prop.table(,2), 2) # Round column prop to 2 digits (percents)
- addmargins(round(prop.table(,2), 2),1) # Round col prop to 2 digits

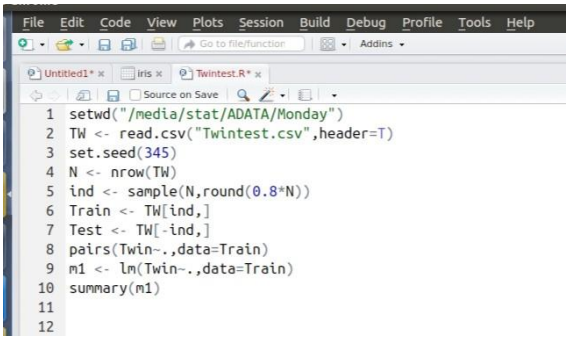
- prop.table() # Tototal proportions
- round(prop.table(),2) # Tototal proportions rounded
- round(100*prop.table(),2) # Tototal proportions rounded

圖十六 常用語法-次數分配

Descriptive Statistics For Continuous Variable

- mean() # Mean of all numeric variables
- median() #Median
- var() # Variance
- sd() # Standard deviation
- max() # Max value
- min() # Min value
- range() # Range
- quantile() # Quantiles 25%
- quantile(mydata$SAT, c(.3,.6,.9)) # Customized quantiles
- length(mydata$SAT) # Num of observations when a variable is specify
- length(mydata) # Number of variables when a dataset is specify
- which.max(mydata$SAT) # Determines the location, i.e., index of the (first) minimum or maximum of a numeric vector"
- which.min(mydata$SAT) # From help: "Determines the location, i.e., index of the (first) minimum or maximum of a numeric vector"

圖十八常用語法-敘述統計量



圖二十 切割資料集

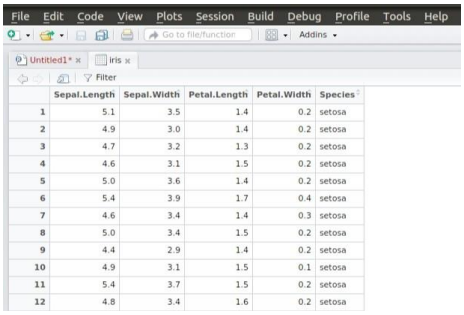
Categorical data: Frequencies

- chisq.test() # Do chisq test Ho: no relathionship
- fisher.test() # Do fisher'exact test Ho: no relationship
- assocstats() #association measures and degree of association.

- X2(chi - square) tests for relationships between nominal variables. The null hypothesis (Ho) is that there is no relationship.
- Cramer's V is a measure of association between two nominal variables. It goes from 0 to 1 where 1 indicates strong
- Fisher's exact test is used when there are very few cases in the cells (usually less than 5). It tests the relationship between two nominal variables.
- The null is that variables are independent.

- Source: http://dss.princeton.edu/training/StataTutorial.pdf

圖十七常用語法-檢定



圖十九 資料集檢視

- **態度面目標** (期望學習者修習完課程後能有哪些態度轉變)：
  在這大數據的時代，資訊更新非常迅速，資料分析並不是只能用 R，Python(如圖二十一所示)、SQL(圖二十二所示)等都是分析常用的工具，期望同學能與時俱進。

圖二十一 Python

圖二十二 SQL

作業設計	個人報告：□書面 □簡報 ___ 次 小組報告：□書面 □簡報 ___ 次 程式設計(個人)：3 次 程式設計(小組)：___ 次 □其他 _____ ___ 次
評量設計	• **形成性評量之規劃** (隨堂練習或小考等)： 小考 1、小考 2

附件一

(第一部分為測試學生的學科基本知識，第二部分則是測試學生是否會用 R 跑回歸模型。)

PART I

1. 何謂迴歸模型之模型矩陣(model matrix) (亦稱為設計矩陣(design matrix));請詳細說明其重要性.(5%)
2. 請以矩陣方程式的方式寫下最小平方法的迴歸參數估計式; 並清楚寫下矩陣方程式中各符號之定義(10%)
3. 請以一般方程式的形式寫下複迴歸(multiple linear regression model) 也就是一般線性迴歸模型(general linear regression model)之模型(要寫完整,包括分配等)(5%)
4. 請以矩陣方程式的方式寫下複迴歸模型,並清楚寫下矩陣方程式中各符號之定義(要寫完整,包括分配等) (10%)

PART II: 請以 Rstudio 回答以下問題,將所有你所使用的語法寫在答案卷上,並將最終語法檔上傳至 ICAN.

請利用 ICAN 上的資料 "HOSPITAL.csv"回答以下問題.

1. 請將資料讀入 Rstudio 並命名為 HOSTIPAL; (5%)
2. 請將 MSA 及 Region 改為 factor 資料型態並將 Region 的參考類別改成'3' (S). (5%)
3. 請以 Lstay 為第一個變數,Age,Infect,Region,ADC,Nnurse 為其他變數繪製散佈圖矩陣;並解釋此矩陣之意義. (5%)
4. 請以 Lstay 為應變數,Age,Infect,Region,ADC,Nnurse 為自變數在不經任何轉換下建立一迴歸模型;請列出你估計出的迴歸模型. (5%)
5. 請解釋 Age 的迴歸係數之 95%信賴區間及其迴歸上的解釋意義. (10%)
6. 請解釋與 Region 有關的迴歸係數之 95%信賴區間及其迴歸上的解釋意義. (10%)
7. 在顯著水準為 0.05 下，以一般線性檢定(General linear test),檢定 Region 是否應存在於模型中;請詳列你的整個檢定的 R 語法. (10%)
8. 請診斷模型的函數型式假設是否合理? (5%)
9. 請診斷模型的變異數一致性假設是否合理? (5%)
10. 請診斷模型的獨立性假設是否合理? (5%)
11. 請診斷模型的常態性假設是否合理? (5%)

圖二十三 小考 1

PART I

1. 請寫下簡單線性迴歸模型(simple linear regression model)之模型(5%)
2. 令 $C_i = \frac{x_i - \bar{x}}{S_{xx}}$，請證明(20%)
   a. $\sum_{i=1}^{n} C_i = 0$
   b. $\sum_{i=1}^{n} C_i x_i = 1$
   c. $\sum_{i=1}^{n} C_i^2 = 1/S_{xx}$
   d. $E(\hat{\beta}_1) = \beta_1$
3. 請詳細寫下簡單迴歸模型之四大基本假設 (20%)
4. 請寫下簡單線性迴歸模型中,經由最大概似法所得到的迴歸係數估計式.(5%)
5. 請寫下殘差(residual)的定義(5%)

PART II: 請以 Rstudio 回答以下問題,將所有你所使用的語法寫在答案卷上,並將最終語法檔上傳至 ICAN.

請利用 ICAN 上的資料 "gasconsumption.txt"回答以下問題.

1. 請將資料請入 Rstudio 並命名為 GasC.(5%)
2. 請問 GasC 屬於 R 的何種資料型態? (5%)
3. 其中變數 ET 表該車的引擎型態(0 表垂直,1 表水平),應為因子變數;但資料讀入時資料型態為整數(int). 請 GasC 中的 ET 變數改為因子資料型態. (10%)
4. 在眾變數中;MPG 表該車的油耗表現,HP 表該車之馬力,請用 R 算出這兩個變數的平均數及標準差. (10%)
5. 接 4., 請將這兩個變數以 MPG 為 Y 軸,HP 為 X 軸,繪出散佈圖(scatter plot) (寫出語法即可,不用將圖用手繪出).. (10%)
6. 請以 MPG 為依變數,HP 為自變數,建立一簡單迴歸模型,請將你估計出的樣本迴歸方程式寫出. (10%)

圖二十四 小考 2

- **總結性評量之規劃** (期中考、期末考或專題成果等)：

期中考、期末考

請以 Rstudio 回答以下問題,將所有你所使用的語法寫在答案卷上,並將最終語法檔上傳至 ICAN.

請利用 ICAN 上的資料 "pga2004test.csv"回答以下問題.

所有變數皆為連續變數,AveWin 為我們有興趣的應變數.

1. 請將資料讀入 Rstudio 並命名為 PGA(5%)
2. 請使用 set.seed(234)並隨機抽取 PGA 之 80%資料做為建模資料,取出並命名為 Train,同時將剩下之資料做為測試資料並命名為 Test. (10%)
3. 以 Train 為資料, AveWin 為應變數,其他變數為自變數,繪製一散佈圖矩陣,並解釋此圖的訊息. (10%)
4. 以 Train 為資料, AveWin 為應變數,其他變數為自變數, 在不經任何轉換下建立一迴歸模型,命名為 m1;請列出你估計出的迴歸模型並解釋其 coefficient of determination 之意義. (10%)
5. 請針對 m1 進行四大假設之診斷,請詳列你的診斷結論及理由! (16%)
6. 請使用 boxcox 轉換法,針對 m1 做適當的矯正,並將矯正後之模型命名為 m2. (9%)
7. 請利用最佳子集法及 adjusted R square (Ra^2)指標挑選出最佳變數,請列出你所選出的變數及原因. (10%)
8. 請利用向後逐步迴歸(backward regression)法,依 AIC 準則進行變數選擇,請列出你所選出的變數. (10%)
9. 請診斷 m2 的共線性問題是否嚴重,詳列你的理由(10%)
10. 請利用測試資料 Test 評估 m1 及 m2 的預測能力,寫下你所使用的評估標準及結果. (10%)

圖二十五 期中考

PART I 請將你的的答案寫在答案卷上

1. 在已知出β_1之最大概似估計式為$\hat{\beta}_1$且簡單迴歸模型之對數概似函數(log-likelihood function)為

$$\ln L(\beta_0, \beta_1, \sigma^2) = \frac{-n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1)^2$$

   的情況下,
   a. 請推導出β_0之最大概似估計式. (10%)
   b. 並請依此概似函數說明β_0及β_1之最大概似估計式及最小平方估計式之間的關係(10%)
2. 請寫下"coefficient of Determination"之定義,所有定義中的符號都要有明確說明其定義(統計量算式). (20%)
3. 請寫下"fitted value"之定義.(5%)
4. 請寫下"residual"之定義.(5%)

PART II: 請以 Rstudio 回答以下問題,將所有你所使用的語法寫在答案卷上,並將最終語法檔上傳至 ICAN.

請利用 ICAN 上的資料 "grapeJuice.csv"回答以下問題. 此資料包含兩變數: price (果汁單價,單位為元); sales (果汁銷售量,單位為鐘)

1. 請將資料請入 Rstudio 並命名為 grapeJuice.csv .請以 sales 為依變數, price 為自變數,建立一不經任何轉換之簡單迴歸模型,請將你估計出的樣本迴歸方程式寫出. (10%)
2. 根據模型,請將以下 ANOVA 表中的 A,B,…I 之值寫在答案卷上(18%)

	Sum of squares	d.f.	Mean (MSE,MSR)	F	P-value
Regression	A	D	F	H	I
Residual	B	E	G		
Total	C				

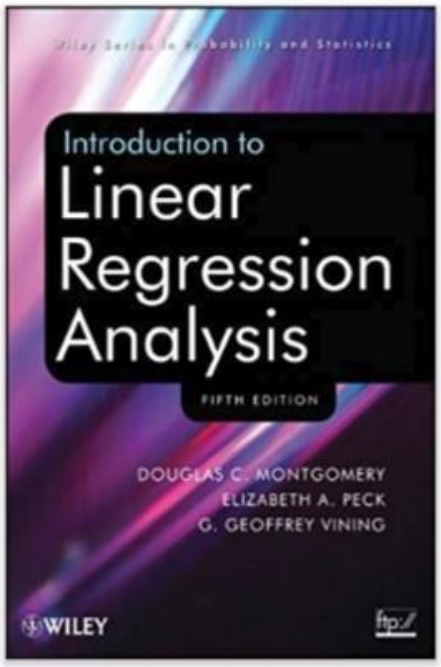
3. 續上題,請寫下 ANOVA 表中 P-value 相應的虛無假設,並在顯著水準為 0.05 下,寫下你的檢定結論。(12%)
4. 請詳述模型之迴歸係數的意義(要用變數的單位來說明). (5%)
5. 請求出模型之迴歸係數的 95%信賴區間;並詳述$\hat{\beta}_1$相應之信賴區間之意義. (5%)

圖二十六 期末考

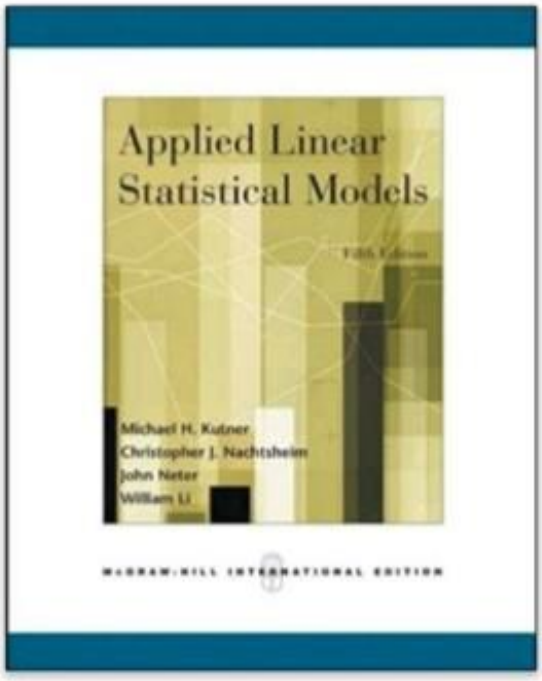
學習輔助資源	線上資源：□Codecademy □Coursera □Code school ■其他 Tronclass 實體資源：□專題演講 ■其他 電腦教室(ES301)
參考與延伸學	Title: Introduction to Linear Regression Analysis (Fifth Edition)(如圖二十七所示) Author: Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining Title: Applied Linear Statistical Models (Fifth Edition) (如圖二十八所示)

附件一

習資料	Author: Kutner, Nachtsheim, Neter, and Li



圖二十七 參考資料 1

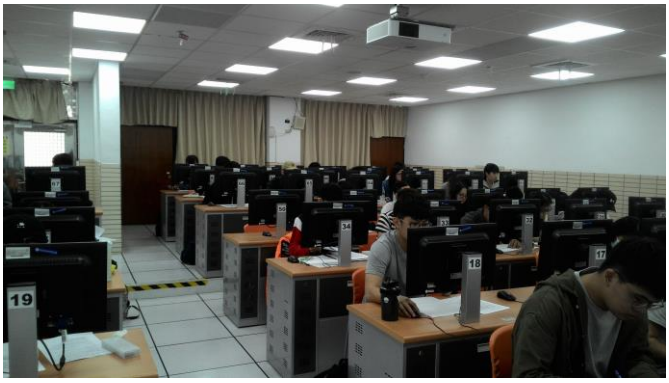


圖二十八參考資料 2

教學設計

週別	課程單元名稱	學習目標	教學設計重點
1	R 語言基本介紹	熟悉 R 語言的安裝與介面	R 語言的推廣
2	R 語言基本語法	熟悉 R 語言的常用語法	建立 R 語言的基礎
3-5	簡單迴歸	以 R 語言建立簡單迴歸模型	學習建立模型
6-8	複迴歸	以 R 語言建立複迴歸模型	學習建立模型
10-14	模型矯正	以 R 語言對模型做矯正	學習修正模型
15-16	模型選擇	以 R 語言選出最終模型	以最終模型做預測

課堂活動剪影 (至少 2 張)



授課心得感想

本迴歸分析課程強調理論與資訊並重，利用板書將迴歸分析的基礎概念讓學生了解，並融入業界常用來做資料分析的 R 語言來輔助迴歸模型的建立，R 語言是一套免費的統計軟體，並且擁有相當多的資料集(如圖二十九所示)，學生即使不在學校也能在任何一台電腦下載使用，並使用不同的資料集自己做練習。

What is R

- R is a statistical-computing environments of the S language that was developed at AT&T Bell Laboratories by Rick Becker, John Chamberr and Allan Wilks. (S-PLUS is a commercial system based on the S language.)
- In Aug. 1993, Ross Ihaka and Robert Gentleman announce their first R source codes in the s-news mailing list.
  - At that time, Ross and Robert are colleagues of University of Auckland, Auckland, New Zealand.
- The R source codes available by ftp under the terms of the Free Software Foundation's GUN general license in June of 1995.
- Till Feb/2/2017, the newest R version 3.3.2 is released
- The official R website: https://www.r-project.org/

An Example

- Perform following code in Rstudio:
  - data(mtcars)
  - View(mtcars)
  - mpg_ave <- mean(mtcars$mpg)
  - mpg_sd <- sd(mtcars$mpg)
  - hist(mtcars$mpg)
  - boxplot(mtcars$disp,col="red")
  - pairs(mpg~disp+hp+wt
  - ,panel = panel.smooth
  - ,data=mtcars,col="blue")
  - colors()
  - demo(colors())

Why use R?

- R is powerful: R is the leading tool for statistics, data analysis, and machine learning. There are over 2,000 cutting-edge, user-contributed packages available on CRAN. To get an idea of what packages are out there, just take a look at these Task Views.
- R is portable: You can easily use it anywhere. It's platform-independent, so you can use it on any operating system.
- R is open-source: That means anyone can examine the source code to see exactly what it's doing. This also means that you, or anyone, can fix bugs and/or add features, rather than waiting for the vendor to find/fix the bug and/or add the feature–at their discretion–in a future release.
- R has a large, active, and growing community of users. You can get help easily.
- R is free: You can use it at any employer without having to persuade your boss to purchase a license.
- Real-time updated, sate-of-art graphic, …

How to get help?

- help() or ?: help on function
- Help.search() or ??: Searches the help system for instances of the string
- Google!

圖二十九 R 的優勢與內建資料集

在程式設計的第一步，我們必須對資料集做資料整理，這也是資料分析最重要的一個步驟，利用這個階段，帶領同學熟悉 R 語言的界面以及一些資料分析常用的指令，並利用內建資料集讓學生實際操作，訓練學生邏輯思考，最後利用 R 語言強大的繪圖功能將其結果視覺化，

以顯現出資料整理前後的差異。

當資料整理完成後才是建立迴歸模型，因 R 語言有內建的指令可直接呼叫，故在建模方面相對容易，建模後須檢驗模型是否符合迴歸分析的基本假設以及如何校正模型，以此與理論相呼應，最後再利用測試資料評估模型的預測能力，除此之外，模型建立完成後，針對該模型及資料集要如何去說故事說服他人，才是這門課真正的價值。

在課堂練習與測驗中，可以觀察到大部分的同學已能熟悉 R 語言的操作並利用 R 語言建立適當的迴歸模型，以視覺化的方式讓對於迴歸分析不是那麼熟悉的人不需要懂複雜的方程式或數字也能瞭其大意。

在這門迴歸分析，融入大量的 R 語言，希望學生學會這個工具，在往後能善用 R 語言的輔助學習更多統計方法並應用在實務上，將統計與資訊做結合。