

# 目錄

Practical Data  
Processing and Analysis  
Using R

## 本書結構

CHAPTER 1	搭建R程式設計環境	1
1.1	為什麼是R	2
1.2	安裝R	3
1.2.1	在Windows作業系統中安裝R	4
1.2.2	在Linux系統下安裝R	6
1.2.3	在Mac OS X中安裝R	13
1.3	啟動R	17
1.4	輔助示例	18
1.5	整合開發環境觀察	22
1.6	批次執行	24
1.7	使用套裝軟體	25
	參考文獻	27
CHAPTER 2	資料類型	29
2.1	變數	30
2.1.1	變數命名規則	30
2.1.2	變數值分配	30
2.2	呼叫函數時指定參數	31
2.3	純量	33
2.3.1	數值	33
2.3.2	NA	33
2.3.3	NULL	34
2.3.4	字串	36

2.3.5	布林值	36
2.3.6	因數	37
2.4	向量	41
2.4.1	向量產生	41
2.4.2	存取向量中的資料	43
2.4.3	向量運算	46
2.4.4	連續數字組成的向量	49
2.4.5	保存重複值的向量	50
2.5	串列	51
2.5.1	串列產生	52
2.5.2	存取串列中的資料	53
2.6	矩陣	54
2.6.1	矩陣產生	54
2.6.2	存取矩陣中的資料	57
2.6.3	矩陣運算	59
2.7	陣列	63
2.7.1	陣列產生	63
2.7.2	存取陣列資料	64
2.8	資料框架	65
2.8.1	資料框架產生	66
2.8.2	存取資料框架	69
2.8.3	實用工具函數	71
2.9	類型判別	74
2.10	類型轉換	76
	參考文獻	78

## CHAPTER 3 R語言程式設計 79

3.1	R的特徵	80
3.2	流程控制（條件敘述與迴圈敘述）	80
3.2.1	if 語句	81
3.2.2	迴圈敘述	82

3.3	運算	84
3.3.1	數值運算	85
3.3.2	向量運算	85
3.3.3	NA處理	87
3.4	定義函數	90
3.4.1	基本定義	90
3.4.2	可變長參數	91
3.4.3	巢套函數	92
3.5	作用域	93
3.6	按值傳遞	97
3.7	物件的不變性	98
3.8	模組樣式	101
3.8.1	佇列	101
3.8.2	編寫佇列模組	102
	參考文獻	104

## CHAPTER 4 資料操作 I：以向量為基礎的處理與外部資料處理 107

4.1	鳶尾花資料集	108
4.2	讀 / 寫檔案	110
4.2.1	讀 / 寫CSV檔案	111
4.2.2	讀 / 寫物件檔案	115
4.3	合併資料框架的行與列	116
4.4	apply系列函數	118
4.4.1	apply()	118
4.4.2	lapply()函數	121
4.4.3	sapply()	125
4.4.4	tapply()	127
4.4.5	mapply()	130
4.5	資料分組並呼叫函數	132
4.5.1	summaryBy()	133
4.5.2	orderBy()	136

4.5.3	<code>sampleBy()</code>	138
4.6	資料拆分與合併	141
4.6.1	<code>split()</code>	141
4.6.2	<code>subset()</code>	143
4.6.3	資料合併	145
4.7	資料排序	147
4.7.1	<code>sort()</code>	147
4.7.2	<code>order()</code>	148
4.8	存取資料框架中的列	149
4.8.1	<code>with()</code>	150
4.8.2	<code>within()</code>	151
4.8.3	<code>attach()</code> 與 <code>detach()</code>	153
4.9	查找符合條件的資料索引	155
4.10	分組運算	157
4.11	更易處理的資料表現形式	158
4.12	與MySQL聯動	161
4.12.1	安裝MySQL及RMySQL	161
4.12.2	使用RMySQL存取MySQL資料庫	170
	參考文獻	172

## CHAPTER 5 資料操作 II：資料處理及加工 173

5.1	資料處理及加工相關套裝軟體	174
5.2	使用SQL處理資料	174
5.3	資料分析：拆分、應用、合併	177
5.3.1	<code>adply()</code> 函數	178
5.3.2	<code>ddply()</code> 函數	180
5.3.3	輕鬆進行每組運算	182
5.3.4	<code>mdply()</code>	186
5.4	資料結構變形與彙總	187
5.4.1	<code>melt()</code>	189
5.4.2	<code>cast()</code>	191

5.4.3	資料彙總	192
5.5	資料表：更快、更方便的資料框架	194
5.5.1	資料表產生	194
5.5.2	資料存取與分組運算	196
5.5.3	使用key快速存取資料	199
5.5.4	使用key合併資料表	201
5.5.5	利用引用修改資料	203
5.5.6	將串列轉換為資料框架	204
5.6	更好的迴圈敘述	207
5.7	並行處理	209
5.7.1	設置進程數	209
5.7.2	plyr並行化	211
5.7.3	foreach並行化	213
5.8	單元測試與除錯	214
5.8.1	testthat	215
5.8.2	使用test_that()進行測試分組	217
5.8.3	測試檔案的結構	218
5.8.4	除錯 (debugging)	220
5.9	測定程式碼執行時間	227
5.9.1	測定命令語句執行時間	227
5.9.2	程式碼效能測試	229
	參考文獻	231

## CHAPTER 6 繪圖 233

6.1	散布圖	234
6.2	圖形選項	236
6.2.1	坐標軸名稱 (xlab, ylab)	236
6.2.2	圖形標題 (main)	237
6.2.3	點的類型 (pch)	238
6.2.4	點的大小 (cex)	239
6.2.5	顏色 (col)	239

0.2.6	坐標軸的取值範圍 (xlim, ylim)	240
0.2.7	圖形類型 (type)	242
0.2.8	線型 (lty)	245
0.2.9	圖形排列	246
0.2.10	抖動	247
0.3	基本圖形	249
0.3.1	點	249
0.3.2	折線	251
0.3.3	直線	253
0.3.4	曲線	255
0.3.5	多邊形	256
0.4	字串	260
0.5	識別圖形中的資料 (identify)	262
0.6	圖例	263
0.7	繪製矩陣中的資料 (matplot、matlines、matpoints)	264
0.8	應用圖形	266
0.8.1	盒形圖	266
0.8.2	直方圖	270
0.8.3	密度圖	273
0.8.4	長條圖	275
0.8.5	圓餅圖	276
0.8.6	馬賽克圖	278
0.8.7	散布圖矩陣	281
0.8.8	透視圖、等高線圖	283
	參考文獻	287

## CHAPTER 7 統計分析 289

- 7.1 亂數產生與分布函數 290
- 7.2 基本統計量 293
  - 7.2.1 樣本平均數、樣本變異數、樣本標準差 294
  - 7.2.2 五數彙總 295
  - 7.2.3 眾數 297
- 7.3 抽樣 298
  - 7.3.1 簡單隨機抽樣 298
  - 7.3.2 考量權數的抽樣 299
  - 7.3.3 分層隨機抽樣 300
  - 7.3.4 系統抽樣 304
- 7.4 列聯表 305
  - 7.4.1 列聯表產生 306
  - 7.4.2 求和與百分比 307
  - 7.4.3 獨立性檢定 309
  - 7.4.4 費雪精準檢定 316
  - 7.4.5 McNemar檢定 317
- 7.5 適合度檢定 321
  - 7.5.1 卡方檢定 321
  - 7.5.2 Shapiro Wilk檢定 322
  - 7.5.3 科摩哥洛夫—史密諾夫檢定 322
  - 7.5.4 分位數—分位數點圖 325
- 7.6 相關分析 329
  - 7.6.1 皮爾森相關係數 329
  - 7.6.2 Spearman秩相關係數 334
  - 7.6.3 肯德爾級相關係數 336
  - 7.6.4 相關係數檢定 336
- 7.7 推論與檢定 338
  - 7.7.1 單樣本平均數 339
  - 7.7.2 兩獨立樣本平均數 342
  - 7.7.3 兩配對樣本平均數 347
  - 7.7.4 兩樣本變異數 349

7.7.5	單樣本比例	351
7.7.6	兩樣本比例	353
	參考文獻	354

## CHAPTER 8 線性迴歸 357

8.1	線性迴歸的基本假設	358
8.2	簡單線性迴歸	359
8.2.1	模型產生	359
8.2.2	提取線性迴歸結果	360
8.2.3	預測與信賴區間	362
8.2.4	模型評估	364
8.2.5	變異數分析及模型間比較	368
8.2.6	模型診斷圖形	370
8.2.7	迴歸直線的視覺化	372
8.3	多元迴歸	374
8.3.1	模型產生及評估	374
8.3.2	類別變數	375
8.3.3	多元迴歸模型的視覺化	378
8.3.4	使用函數I()	379
8.3.5	變數的變換	382
8.3.6	交互作用	382
8.4	離群值	389
8.5	變數選擇	391
8.5.1	選擇變數的方法	391
8.5.2	比較所有情形	396
	參考文獻	399

## CHAPTER 9 分類演算法 I：資料探索、預處理、模型評估方法 401

9.1	資料探索	402
9.1.1	技術統計	402

9.1.2	資料視覺化	408
9.2	預處理	412
9.2.1	資料變換	412
9.2.2	遺漏值 (missing values) 處理	419
9.2.3	變數選擇	423
9.3	模型評估方法	436
9.3.1	評估指標	436
9.3.2	ROC曲線	440
9.3.3	交叉確認	446
	參考文獻	459

## CHAPTER 10 分類演算法 II：機器學習演算法 461

10.1	邏輯迴歸模型	462
10.2	多項邏輯迴歸分析	466
10.3	決策樹	470
10.3.1	決策樹模型	470
10.3.2	分類與迴歸樹	472
10.3.3	條件推斷決策樹	475
10.3.4	隨機森林	478
10.4	類神經網路	485
10.4.1	類神經網路模型	485
10.4.2	類神經網路模型學習	487
10.5	支持向量機	492
10.5.1	支持向量機模型	492
10.5.2	支持向量機學習	494
10.6	類別不平衡	500
10.6.1	放大取樣、縮小取樣	500
10.6.2	SMOTE	503
10.7	文件分類	505
10.7.1	語料庫與文件	505
10.7.2	文件變換	507

10.7.3	文件的矩陣表示	508
10.7.4	高頻詞	513
10.7.5	詞語之間的相關關係	514
10.7.6	文件分類	514
10.7.7	從檔案產生語料庫	517
10.7.8	詮釋資料	519
10.8	caret套裝軟體	523
	參考文獻	528

## CHAPTER 11 利用鐵達尼號資料練習機器學習 531

11.1	鐵達尼號資料格式	532
11.2	讀入資料	532
11.2.1	轉換資料類型	533
11.2.2	分離測試資料	535
11.2.3	準備交叉確認	536
11.3	資料探索	539
11.4	評估指標	544
11.5	決策樹模型	544
11.5.1	rpart的交叉確認	545
11.5.2	精確度分級 (accuracy rating)	547
11.5.3	條件推斷決策樹	547
11.6	發現其他特徵	549
11.6.1	使用ticket識別家庭	549
11.6.2	預測生還機率	551
11.6.3	增加家庭ID	551
11.6.4	合併家庭成員的生還機率	554
11.6.5	使用家庭資訊建立ctree()模型	556
11.6.6	效能評估	558

11.7	交叉確認並行化	560
11.7.1	反復執行三次10層交叉確認	560
11.7.2	使用foreach()與%dopar%進行並行化	562
11.8	開發更好的演算法	563
	參考文獻	564

中文索引 565

英文索引 569