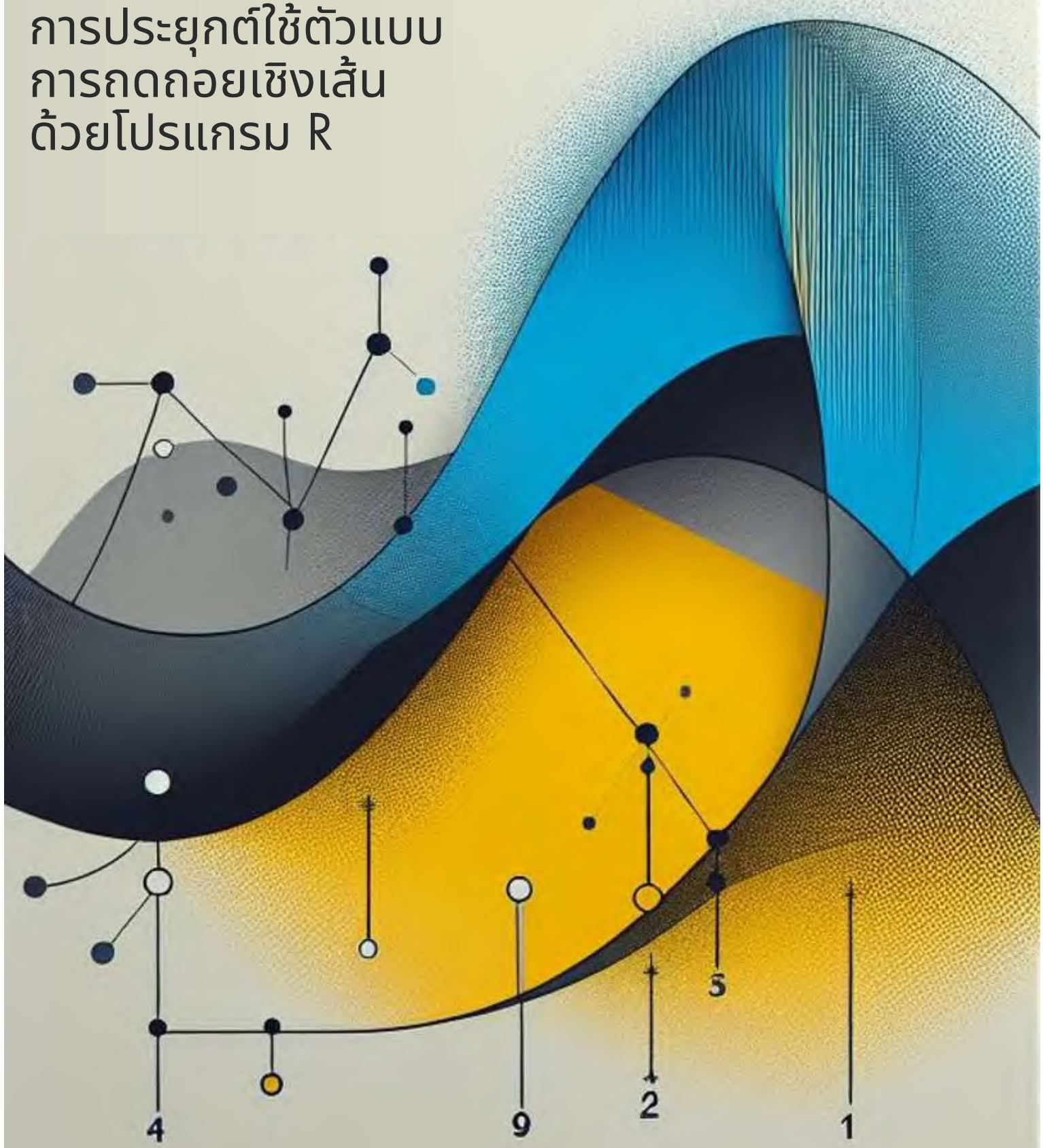


Applied Linear Regression Models with R

การประยุกต์ใช้ตัวแบบ
การถดถอยเชิงเส้น
ด้วยโปรแกรม R



รองศาสตราจารย์ ดร. วิฐรา พึ่งพาพงศ์

ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี
จุฬาลงกรณ์มหาวิทยาลัย

**การประยุกต์ใช้ตัวแบบการถดถอยเชิงเส้น
ด้วยโปรแกรม R**

**Applied Linear Regression Models
with R**

วิฐรา พึ่งพาพงศ์

คำนำ

การประยุกต์ใช้ตัวแบบการถดถอยเชิงเส้นไปใช้นั้น ผู้ใช้จะต้องมีความเข้าใจหลักการวิเคราะห์การถดถอยเชิงเส้นทั้งในเชิงทฤษฎีและปฏิบัติ หนังสือเล่มนี้รวบรวมหลักการวิเคราะห์การถดถอยเชิงเส้นและเน้นการประยุกต์ใช้งานโดยเน้นการประยุกต์ใช้งานกับข้อมูลจริง โดยทั่วไปแล้วข้อมูลจริงนั้นมักแตกต่างจากข้อสมมติในทางทฤษฎีที่ได้กล่าวไว้ ดังนั้นผู้วิเคราะห์ข้อมูลจะต้องเผชิญกับข้อมูลจริงที่มีลักษณะต่าง ๆ หลากหลายรูปแบบ หนังสือเล่มนี้เหมาะสำหรับผู้ที่สนใจประยุกต์การวิเคราะห์การถดถอยเชิงเส้นไปใช้งาน ถึงแม้จะไม่มีความรู้พื้นฐานเกี่ยวกับการวิเคราะห์การถดถอยเชิงเส้นและการเขียนโปรแกรมภาษา R ผู้อ่านสามารถเข้าใจหลักการการวิเคราะห์การถดถอยเชิงเส้นและเรียนรู้จากตัวอย่าง ข้อมูลจริงที่ใช้ในหนังสือเล่มนี้ทั้งหมดถูกเก็บไว้ในแพ็คเกจสำหรับโปรแกรมภาษา R และแสดงคำสั่ง R ทั้งหมด ผู้อ่านสามารถฝึกปฏิบัติตามได้โดยง่าย

หนังสือเล่มนี้จึงกล่าวถึงหลักการวิเคราะห์การถดถอยเชิงเส้นซึ่งรวมถึงทฤษฎีเฉพาะในส่วนที่จำเป็นและเป็นประโยชน์โดยตรงต่อการประยุกต์ใช้งานจริง และเนื่องจากในปัจจุบันข้อมูลมีความซับซ้อนมาก ดังนั้นนอกเหนือจากหัวข้อพื้นฐานในการวิเคราะห์การถดถอยเชิงเส้นที่ปรากฏในหนังสือการวิเคราะห์การถดถอยทั่วไปแล้ว หนังสือเล่มนี้ยังได้ครอบคลุมเนื้อหาการวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูงโดยใช้ตัวแบบการถดถอยที่ปรับด้วยฟังก์ชันการลงโทษ (Penalty Function) ซึ่งเป็นวิธีที่ได้รับความนิยมและมีการประยุกต์ใช้กันอย่างแพร่หลายในปัจจุบันนี้

ผู้เขียนมีประสบการณ์ในการสอนในรายวิชาการประยุกต์ใช้ตัวแบบเชิงเส้นทางสถิติในระดับบัณฑิตศึกษา และมีผลงานตีพิมพ์หลายชิ้นที่เกี่ยวข้องกับการวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูง ทั้งนี้ผู้เขียนหวังเป็นอย่างยิ่งว่าหนังสือเล่มนี้จะเป็นประโยชน์แก่นิสิตนักศึกษา นักวิจัย ตลอดจนผู้สนใจทั่วไปที่ต้องการประยุกต์ตัวแบบการถดถอยเชิงเส้นกับข้อมูลต่าง ๆ ผู้เขียนขอขอบพระคุณผู้ทรงคุณวุฒิ 2 ท่าน ได้แก่ รองศาสตราจารย์ ดร. วินัย โพธิ์สุวรรณ และ ผู้ช่วยศาสตราจารย์ ดร. ชนาพันธุ์ ชนาเนตร ที่ได้สละเวลาในการอ่านและให้คำแนะนำที่เป็นประโยชน์อย่างมากในการปรับปรุงหนังสือเล่มนี้ หากมีความผิดพลาดประการใดที่ปรากฏในหนังสือเล่มนี้ ผู้เขียนยินดีน้อมรับคำแนะนำเพื่อการปรับปรุงในโอกาสต่อไป

ผู้ช่วยศาสตราจารย์ ดร. วิรุรา พึ่งพาพงศ์

8 ธันวาคม 2564

การประยุกต์ใช้ตัวแบบการถดถอยเชิงเส้นด้วยโปรแกรม R

ผู้แต่ง

วิฐรา พึ่งพาพงศ์

พิมพ์ครั้งที่ 1

ธันวาคม 2564

พิมพ์ครั้งที่ 2

ธันวาคม 2566

ISBN e-book

978-616-588-367-2

ราคา

299 บาท

สงวนลิขสิทธิ์ตามพระราชบัญญัติลิขสิทธิ์ (ฉบับเพิ่มเติม) พ.ศ.2558 ห้ามลอกเลียนแบบหรือคัดลอกส่วนใดส่วนหนึ่งของหนังสือเล่มนี้ นอกจากจะได้รับอนุญาตเป็นลายลักษณ์อักษรจากผู้เขียนเท่านั้น

จัดทำโดย

วิฐรา พึ่งพาพงศ์

สั่งซื้อได้ที่

ศูนย์หนังสือจุฬาลงกรณ์มหาวิทยาลัย

ถนนพญาไท เขตปทุมวัน กรุงเทพฯ 10330

โทร.0-2218-9872 โทรสาร 0-2254-9495

Call Center โทร.0-2255-4433

<http://www.chulabook.com>

สารบัญ

บทที่ 1 บทนำ	1
1.1 วัตถุประสงค์ของหนังสือเล่มนี้.....	1
1.2 ประวัติความเป็นมาของการวิเคราะห์การถดถอยเชิงเส้น	1
1.3 การใช้งานการวิเคราะห์การถดถอยเชิงเส้นในปัจจุบัน	2
1.4 รูปแบบความสัมพันธ์จากการวิเคราะห์การถดถอย	3
1.5 บทสรุป.....	4
บทที่ 2 การเขียนโปรแกรม R เบื้องต้น	5
2.1 การติดตั้งโปรแกรม R และ RStudio.....	5
2.1.1 การติดตั้งโปรแกรม R	5
2.1.2 การติดตั้งโปรแกรม RStudio	8
2.2 คำสั่งเบื้องต้นใน R	11
2.3 ประเภทข้อมูลใน R.....	12
2.4 การเรียกใช้ฟังก์ชันใน R	22
2.5 การติดตั้งแพ็คเกจใน R	22
2.5.1 การติดตั้งแพ็คเกจจาก CRAN	22
2.5.2 การติดตั้งแพ็คเกจจาก GitHub	24
2.6 การเปิดและวิธีการอ่านไฟล์ช่วยเหลือ	25
2.7 การอ่านและบันทึกไฟล์.....	28
2.7.1 การอ่านและบันทึกแฟ้มข้อความ	28
2.7.2 การบันทึกวัตถุใน R.....	30
2.7.3 การบันทึกคำสั่งใน R.....	31
2.8 การเขียนคำสั่งโครงสร้างควบคุมใน R.....	31
2.8.1 คำสั่ง If.....	31
2.8.2 คำสั่ง For	34
2.8.3 คำสั่ง While.....	36
2.9 การสร้างฟังก์ชันใน R.....	37
2.10 บทสรุป	38
บทที่ 3 การอนุมานเชิงสถิติเบื้องต้น.....	39
3.1 ประชากรและตัวอย่าง.....	39
3.2 การแจกแจงค่าตัวอย่าง.....	40
3.3 การประมาณค่า.....	42
3.3.1 การประมาณค่าแบบจุด	42

3.3.2 การประมาณค่าแบบช่วง	43
3.4 การทดสอบสมมุติฐาน	45
3.4.1 การตั้งสมมุติฐาน	46
3.4.2 ความผิดพลาดในการทดสอบสมมุติฐาน	46
3.4.3 ขั้นตอนการทดสอบสมมุติฐานด้วยวิธีการหาค่าพี	47
3.5 บทสรุป	52
บทที่ 4 การวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย	53
4.1 ความสัมพันธ์ระหว่างสองตัวแปร	53
4.2 ตัวแบบการถดถอยเชิงเส้นอย่างง่าย	57
4.3 วิธีกำลังสองน้อยสุด	58
4.3.1 การประมาณค่าสัมประสิทธิ์การถดถอย	58
4.3.2 การประมาณค่าแปรปรวน	61
4.4 การอนุมานสำหรับสัมประสิทธิ์การถดถอย	62
4.4.1 การแจกแจงความน่าจะเป็นของสัมประสิทธิ์การถดถอย	62
4.4.2 การทดสอบสมมุติฐานของสัมประสิทธิ์การถดถอย	64
4.4.3 การประมาณค่าแบบช่วงของสัมประสิทธิ์การถดถอย	68
4.5 การวิเคราะห์ความแปรปรวนสำหรับตัวแบบการถดถอยเชิงเส้น	68
4.6 สัมประสิทธิ์การกำหนด	74
4.7 การใช้ตัวแบบการถดถอยในการพยากรณ์	76
4.7.1 การประมาณค่าเฉลี่ยของตัวแปรตาม สำหรับค่าของตัวแปรอิสระที่กำหนด	76
4.7.2 การพยากรณ์ตัวแปรตามสำหรับค่าของตัวแปรอิสระที่กำหนด	77
4.7.3 แถบความเชื่อมั่น	79
4.8 บทสรุป	81
บทที่ 5 การวิเคราะห์การถดถอยเชิงเส้นพหุคูณ	82
5.1 ตัวแบบการถดถอยเชิงเส้นพหุคูณ	82
5.2 วิธีกำลังสองน้อยสุด	83
5.2.1 การประมาณค่าสัมประสิทธิ์การถดถอย	83
5.2.2 การประมาณค่าแปรปรวน	86
5.3 รูปแบบของการวิเคราะห์การถดถอยเชิงเส้นพหุคูณแบบอื่น ๆ	87
5.3.1 ตัวแบบการถดถอยพหุนาม	87
5.3.2 ตัวแบบการถดถอยที่มีตัวแปรหุ่น	89
5.3.3 ตัวแบบการถดถอยที่มีตัวแปรปฏิสัมพันธ์	90
5.3.4 สูตรการเขียนอาร์กิวเมนต์ Formula ในคำสั่ง lm	93
5.4 การอนุมานสำหรับสัมประสิทธิ์การถดถอยหนึ่งตัว	94

4.4.1 การทดสอบสมมุติฐานของสัมประสิทธิ์การถดถอย.....	94
5.4.2 การประมาณค่าแบบช่วงของสัมประสิทธิ์การถดถอย	95
5.5 การวิเคราะห์ความแปรปรวน.....	96
5.5.1 ตาราง ANOVA และตัวสถิติทดสอบเอฟ.....	96
5.5.2 การทดสอบเอฟจากผลบวกกำลังสองที่เพิ่มขึ้น.....	98
5.6 สัมประสิทธิ์การกำหนดและสัมประสิทธิ์การกำหนดแบบปรับ	99
5.7 การใช้ตัวแบบการถดถอยในการพยากรณ์	101
5.7.1 การประมาณค่าเฉลี่ยของตัวแปรตามเมื่อกำหนดค่าตัวแปรอิสระ.....	101
5.7.2 การพยากรณ์ตัวแปรตามเมื่อกำหนดค่าตัวแปรอิสระ	102
5.7.3 แถบความเชื่อมั่น	103
5.8 บทสรุป.....	104
บทที่ 6 การตรวจสอบข้อสมมติและการแก้ไข	105
6.1 ข้อสมมติในการวิเคราะห์การถดถอย.....	107
6.2 การตรวจสอบข้อสมมติในการวิเคราะห์การถดถอย	108
6.2.1 การตรวจข้อสมมติความเป็นเส้นตรง	108
6.2.2 การตรวจข้อสมมติความปกติ	110
6.2.3 การตรวจข้อสมมติความแปรปรวนคงที่	112
6.2.4 การตรวจสอบข้อสมมติความเป็นอิสระ	113
6.3 ความแรงของตัวประมาณจากวิธีกำลังสองน้อยสุด.....	115
6.3.1 ผลกระทบเมื่อข้อสมมติความเป็นเส้นตรงถูกละเมิด	115
6.3.2 ผลกระทบเมื่อข้อสมมติความปกติถูกละเมิด.....	116
6.3.3 ผลกระทบเมื่อข้อสมมติความแปรปรวนคงที่ถูกละเมิด	116
6.3.4 ผลกระทบเมื่อข้อสมมติความเป็นอิสระถูกละเมิด.....	116
6.4 วิธีการแก้ไขกรณีที่ข้อสมมติความเป็นเส้นตรงถูกละเมิด	116
6.4.1 การใช้ตัวแบบพหุนาม	117
6.4.2 การแปลงตัวแปร.....	118
6.5 วิธีการแก้ไขกรณีที่ข้อสมมติความปกติถูกละเมิด	122
6.5.1 การแปลงตัวแปร.....	122
6.6 วิธีการแก้ไขกรณีที่ข้อสมมติความแปรปรวนคงที่ถูกละเมิด	123
6.6.1 การแก้ไขความแปรปรวนที่ไม่คงที่ด้วยการแปลงตัวแปรตาม	124
6.6.2 การแก้ไขความแปรปรวนที่ไม่คงที่ด้วยการวิเคราะห์การถดถอยแบบถ่วงน้ำหนัก.....	126
6.7 วิธีการแก้ไขกรณีที่ข้อสมมติความเป็นอิสระถูกละเมิด	129
6.7.1 การแก้ไขความไม่เป็นอิสระเมื่อมีแนวโน้มเกิดขึ้นตามเวลา	129
6.7.2 การแก้ไขความไม่เป็นอิสระเมื่อความคลาดเคลื่อนมีสหสัมพันธ์ในตัว.....	132

6.8 ปัญหาตัวแปรอิสระมีความสัมพันธ์เชิงเส้นพหุ	135
6.8.1 การตรวจสอบปัญหาตัวแปรอิสระมีความสัมพันธ์เชิงเส้นพหุ	136
6.8.2 การแก้ไขปัญหาตัวแปรอิสระมีความสัมพันธ์เชิงเส้นพหุด้วยตัวแบบการถดถอย แบบบริดจ์	140
6.9 บทสรุป	144
บทที่ 7 การตรวจสอบข้อมูลผิดปกติและข้อมูลที่มีอิทธิพล	145
7.1 ข้อมูลผิดปกติและข้อมูลที่มีอิทธิพล	145
7.2 การตรวจหาความผิดปกติของตัวแปรอิสระ	147
7.2.1 เลฟเวอเรจ	147
7.3 การตรวจหาความผิดปกติของตัวแปรตาม	149
7.3.1 ส่วนเหลือสทิวเดนต์	150
7.3.2 ส่วนเหลือตัดค่าสังเกต	151
7.3.3 ส่วนเหลือสทิวเดนต์ตัดค่าสังเกต	152
7.4 การตรวจหาข้อมูลที่มีอิทธิพล	153
7.4.1 ดีเอฟฟิตส์	153
7.4.2 ระยะทางของคุก	155
7.4.3 ดีเอฟปีตาส์	157
7.4.4 อัตราส่วนความแปรปรวนร่วม	160
7.5 การตรวจหาข้อมูลผิดปกติและข้อมูลที่มีอิทธิพลด้วย olsrr	161
7.6 การแก้ไขปัญหาเมื่อตรวจพบข้อมูลที่มีอิทธิพล	167
7.6.1 วิธีคำนวณฐานของส่วนเหลือนำกำลังสองน้อยสุด	168
7.6.2 วิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนักชนิดทำซ้ำสำหรับการวิเคราะห์การถดถอยที่มี ความแกร่ง	168
7.7 บทสรุป	172
บทที่ 8 กลยุทธ์การเลือกตัวแบบ	173
8.1 กลยุทธ์ในการเลือกตัวแบบ	174
8.1.1 กลยุทธ์ในการเลือกตัวแบบที่สามารถอธิบายตัวแปรตามจากตัวแปรอิสระที่มี จำนวนน้อยที่สุด	174
8.1.2 กลยุทธ์ในการเลือกตัวแบบในการพยากรณ์ข้อมูลในอนาคต	175
8.2 การเลือกเซตย่อยของตัวแปรอิสระที่ดีที่สุด	175
8.2.1 สมประสิทธิ์การกำหนดแบบปรับ	176
8.2.2 มาลโลว์ซีพี	177
8.2.3 เกณฑ์สารสนเทศของอะกะอิเกะ	178
8.2.4 เกณฑ์สารสนเทศของเบส์	179

8.3 วิธีการเลือกตัวแปรเป็นลำดับ	182
8.3.1 การเลือกแบบไปข้างหน้า	182
8.3.2 การกำจัดแบบถอยหลัง	185
8.3.3 การถดถอยทีละขั้น.....	187
8.4 ข้อควรระวังในการเลือกตัวแปรเป็นลำดับด้วยวิธีการอัตโนมัติ	189
8.5 วิธีการเลือกตัวแปรสำหรับการพยากรณ์ในอนาคต	190
8.5.1 วิธีการแบ่งชุดข้อมูล	190
8.5.2 เกณฑ์ในการวัดประสิทธิภาพในการพยากรณ์ของตัวแบบ	193
8.5.3 กรณีศึกษาการเลือกตัวแปรสำหรับการพยากรณ์.....	195
8.6 บทสรุป.....	213
บทที่ 9 การวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูง.....	214
9.1 ปัญหาของการวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูง	214
9.2 การวิเคราะห์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษ	215
9.2.1 การถดถอยแบบบริดจ์	215
9.2.2 การถดถอยลาสโซ่.....	216
9.2.3 การถดถอยลาสโซ่แบบปรับปรุง.....	217
9.2.4 การถดถอยอิลาสติกเน็ต.....	218
9.3 การวิเคราะห์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษด้วย glmnet.....	218
9.3.1 การถดถอยแบบบริดจ์	221
9.3.2 การถดถอยลาสโซ่.....	222
9.3.3 การถดถอยลาสโซ่ปรับปรุง.....	223
9.3.4 การถดถอยอิลาสติกเน็ต.....	224
9.4 การตรวจสอบข้อสมมติเบื้องต้นของตัวแบบการถดถอย	226
9.5 การทดสอบสมมติฐานเกี่ยวกับสัมประสิทธิ์การถดถอย.....	227
9.5.1 วิธีบูตสเตรปในการอนุมานเกี่ยวกับสัมประสิทธิ์การถดถอย	227
9.5.2 วิธีบูตสเตรปแบบปรับสำหรับการวิเคราะห์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษ.....	231
9.5.3 วิธีการบูตสเตรปและการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยแบบอื่น ๆ.....	235
9.6 บทสรุป.....	237
บทที่ 10 บทประยุกต์การใช้ตัวแบบการถดถอยเชิงเส้น.....	238
10.1 กรณีศึกษาน้ำหนักสมองของสัตว์เลี้ยงลูกด้วยนม.....	238
10.2 กรณีศึกษาข้อมูลรถยนต์พลังงานไฟฟ้า.....	257
10.3 บทสรุป.....	277

บรรณานุกรม	278
ดัชนี.....	282

บทที่ 1 บทนำ

1.1 วัตถุประสงค์ของหนังสือเล่มนี้

การวิเคราะห์การถดถอยเชิงเส้นมีประโยชน์ในการวิเคราะห์ข้อมูลดังเห็นได้จากการประยุกต์ใช้งานที่แพร่หลายในปัจจุบัน ตัวแบบที่ได้จากการวิเคราะห์การถดถอยสามารถนำไปศึกษาถึงความสัมพันธ์ของปัจจัยต่าง ๆ ต่อตัวแปรที่ต้องการศึกษา รวมทั้งยังสามารถนำตัวแบบไปใช้ในการพยากรณ์ได้ การจะประสบความสำเร็จในการนำตัวแบบการวิเคราะห์การถดถอยเชิงเส้นไปใช้ในการวิเคราะห์ข้อมูลจริง ผู้ใช้จะต้องมีความเข้าใจหลักการวิเคราะห์การถดถอยเชิงเส้นทั้งในเชิงทฤษฎีและปฏิบัติ หนังสือเล่มนี้เน้นการประยุกต์ใช้งานการวิเคราะห์การถดถอยเชิงเส้น โดยได้ผสมผสานหลักการวิเคราะห์การถดถอยเชิงเส้นและการประยุกต์ใช้งานกับข้อมูลจริง ทั้งนี้ หนังสือเล่มนี้หลีกเลี่ยงกล่าวถึงทฤษฎีเชิงลึกของตัวประมาณการถดถอยเชิงเส้น โดยจะกล่าวถึงทฤษฎีเฉพาะในส่วนที่จำเป็นและมีความสัมพันธ์โดยตรงกับการนำไปประยุกต์ใช้งานต่อ

เนื่องจากในปัจจุบัน มีซอฟต์แวร์หรือโปรแกรมทางสถิติจำนวนมากในการวิเคราะห์ข้อมูลขั้นสูง ซึ่งรวมถึงเทคนิคการวิเคราะห์การถดถอยเชิงเส้น โปรแกรม R เป็นโปรแกรมทางสถิติที่ใช้กันอย่างแพร่หลายทั้งในวงการวิชาการและในภาคธุรกิจ โปรแกรม R มีข้อดีหลายอย่าง คือ การติดตั้งโปรแกรมที่สามารถทำได้โดยง่ายและรองรับระบบปฏิบัติการต่าง ๆ ไม่ว่าจะเป็น Windows MacOS และ UNIX/Linux นอกจากนี้ โปรแกรม R เป็นโปรแกรมฟรีแวร์ (Freeware) ซึ่งผู้ใช้ไม่ต้องเสียค่าใช้จ่ายใด ๆ ในการติดตั้งหรือใช้งานโปรแกรม R

หนังสือเล่มนี้มีวัตถุประสงค์เพื่อให้ผู้อ่านได้เข้าใจถึงหลักการที่สำคัญในการวิเคราะห์การถดถอยเชิงเส้น และนำไปประยุกต์ใช้งานกับข้อมูลจริงได้โดยใช้โปรแกรม R สำหรับข้อมูลที่นำมาใช้เป็นตัวอย่างในการวิเคราะห์ทั้งหมดในหนังสือเล่มนี้เป็นข้อมูลจริงทั้งหมด ซึ่งการใช้ข้อมูลจริงมาแสดงในการวิเคราะห์ข้อมูลจะทำให้ผู้อ่านได้เห็นถึงลักษณะต่าง ๆ รวมถึงความซับซ้อนของข้อมูลที่เกิดขึ้นจริง นอกจากนี้ หนังสือเล่มนี้ยังได้กล่าวถึงหัวข้อที่นอกเหนือจากหัวข้อพื้นฐานในการวิเคราะห์การถดถอยซึ่งสามารถพบเจอได้ในหนังสือทั่วไปแต่มีความสำคัญในการวิเคราะห์ข้อมูลจริง เช่น การวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูง การวิเคราะห์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษ และ วิธีบูตสเตรปในการอนุมานเกี่ยวกับสัมประสิทธิ์การถดถอย

1.2 ประวัติความเป็นมาของการวิเคราะห์การถดถอยเชิงเส้น

ประวัติความเป็นมาของการวิเคราะห์การถดถอยเชิงเส้นเกิดขึ้นในศตวรรษที่ 19 ณ ประเทศอังกฤษ โดยนักวิทยาศาสตร์และนักสถิติที่ชื่อ ฟรานซิส กัลตัน (Francis Galton) ซึ่งในขณะนั้นกำลังศึกษาถึงการถ่ายทอดทางพันธุกรรมลักษณะของมนุษย์ที่สำคัญอย่างหนึ่งนั่นก็คือความสูงของมนุษย์นั่นเอง ในงานวิจัยของกัลตัน เขาได้เก็บข้อมูลความสูงของพ่อแม่ ความสูงของลูก และเพศของลูก ในกรณีของแม่และบุตรสาว ความสูงจะถูกปรับโดยนำไปคูณกับค่า 1.08 (Galton, 1886) กัลตันได้ศึกษาความสัมพันธ์ความสูงเฉลี่ยระหว่างพ่อ

และแม่ และความสูงของลูก และพบว่า ลูกที่สูงมักจะเกิดจากพ่อแม่ที่มีค่าเฉลี่ยความสูงที่มาก และลูกที่เตี้ยมักจะเกิดจากพ่อแม่ที่มีค่าเฉลี่ยความสูงที่ต่ำ นอกจากนี้ กัลตันยังได้พบกับปรากฏการณ์ที่น่าสนใจเป็นอย่างยิ่งมาก กล่าวคือ ในบรรดาพ่อแม่ที่สูงหรือมีความสูงเฉลี่ยเกินกว่าค่าเฉลี่ยของข้อมูล กลับพบว่าลูกจะเตี้ยกว่าความสูงเฉลี่ยพ่อแม่ ในทางตรงกันข้าม สำหรับบรรดาพ่อแม่ที่เตี้ยหรือมีความสูงเฉลี่ยต่ำกว่าค่าเฉลี่ยของข้อมูล กลับพบว่าลูกของพ่อแม่เหล่านั้นจะสูงกว่าความสูงเฉลี่ยพ่อแม่ จากปรากฏการณ์นี้เอง จึงทำให้กัลตันเรียกรูปแบบการถ่ายทอดทางพันธุกรรมรูปแบบนี้ว่า “การถดถอย (Regression)” ซึ่งอธิบายถึงปรากฏการณ์การถ่ายทอดทางพันธุกรรมของความสูงของมนุษย์มีลักษณะถดถอยเข้าสู่ค่าเฉลี่ย หรือ “Regression towards the mean” แม้ว่าในขณะนั้น ผลงานของกัลตันจะแสดงให้เห็นถึงหลักการที่สำคัญทางพันธุศาสตร์ อย่างไรก็ตาม สิ่งที่ยกัลตันค้นพบนำมาสู่การพัฒนาแนวคิดที่สำคัญอย่างมากทางสถิติต่อมา นั่นก็คือ การสร้างสมการคำนวณทางคณิตศาสตร์เพื่ออธิบายความสัมพันธ์ระหว่างตัวแปรเชิงปริมาณ 2 ตัวแปร และนำไปใช้ในการพยากรณ์ค่าของตัวแปรเชิงปริมาณตัวหนึ่ง เมื่อกำหนดค่าของตัวแปรเชิงปริมาณอีกตัวหนึ่ง โดยในอีกไม่กี่ปีต่อมาหลังจากการตีพิมพ์ผลงานวิจัยของกัลตัน การพัฒนาวิธีการวิเคราะห์การถดถอยเชิงเส้นก็เสร็จโดยสมบูรณ์โดยนักสถิติชาวอังกฤษ 3 ท่าน คือ ฟรานซิส เอดจ์เวิร์ธ (Francis Edgeworth) คาร์ล เพียร์สัน (Karl Pearson) และ จอร์จ ยูล (Goerge Yule)

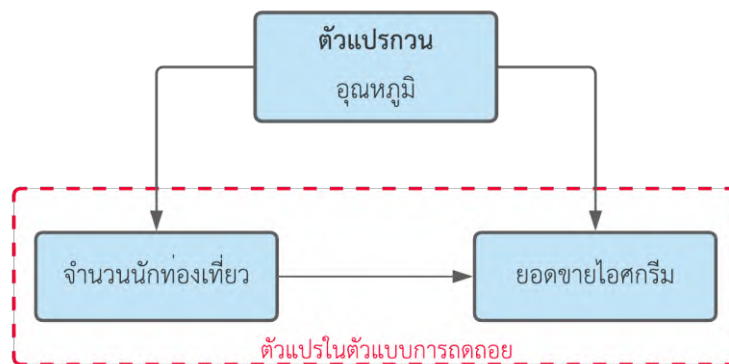
1.3 การใช้งานการวิเคราะห์การถดถอยเชิงเส้นในปัจจุบัน

แม้ว่างานต้นแบบของการวิเคราะห์การถดถอยของกัลตันจะศึกษาความสัมพันธ์ระหว่างตัวแปรเชิงปริมาณ 2 ตัวแปร ต่อมาได้มีขยายขอบเขตการทำงานของงานการวิเคราะห์การถดถอยให้สามารถใช้ได้กับการศึกษาความสัมพันธ์ระหว่างตัวแปรมากกว่า 2 ตัว รวมถึงการใช้งานสำหรับกรณีที่มีตัวแปรเป็นตัวแปรเชิงคุณภาพอีกด้วย การวิเคราะห์การถดถอยเชิงเส้นจึงมีความยืดหยุ่นในการใช้งานเป็นอย่างมากและสามารถนำไปประยุกต์ใช้กับข้อมูลที่หลากหลายได้ ตัวอย่างการใช้งานการวิเคราะห์การถดถอยเชิงเส้นในด้านต่าง ๆ อาทิเช่น ทางด้านธุรกิจ องค์กรสามารถนำตัวแบบการวิเคราะห์การถดถอยเพื่อศึกษาความสัมพันธ์ระหว่างรายได้จากค่าใช้จ่ายที่องค์กรใช้ในการโฆษณา หรือในด้านการแพทย์ ผู้วิจัยสามารถใช้การวิเคราะห์การถดถอยเพื่อศึกษาความสัมพันธ์ระหว่างปริมาณวัคซีนที่ให้กับคนไข้และปริมาณภูมิคุ้มกันที่เพิ่มขึ้นมา เป็นต้น

แม้ว่าการวิเคราะห์การถดถอยเชิงเส้นเป็นหนึ่งในเครื่องมือที่สำคัญและเป็นประโยชน์ในการวิเคราะห์ข้อมูล รวมถึงสามารถนำไปประยุกต์ใช้งานได้หลากหลาย อย่างไรก็ตาม การวิเคราะห์การถดถอยเชิงเส้นยังมีข้อจำกัดอยู่ ทั้งนี้ เนื่องจาก ตัวแบบการถดถอยถูกพัฒนาขึ้นภายใต้ข้อสมมติ (Assumption) ทางสถิติที่กำหนดไว้ ดังนั้น ตามหลักสถิติแล้ว การจะนำตัวแบบการถดถอยไปใช้งานให้มีความถูกต้องและน่าเชื่อถือนั้นจึงจำเป็นต้องตรวจสอบข้อสมมติดังกล่าวก่อนนำตัวแบบไปใช้งาน นอกจากนี้ การวิเคราะห์การถดถอยเชิงเส้นแบบดั้งเดิมนั้น จะประมาณตัวแบบการถดถอยได้ก็ต่อเมื่อขนาดตัวอย่างหรือจำนวนของข้อมูลมากกว่าจำนวนตัวแปรในตัวแบบ แต่ในปัจจุบัน ข้อมูลมีความสลับซับซ้อนมากขึ้น และเป็นไปได้ที่มีจำนวนตัวแปรในตัวแบบจะมีจำนวนมากกว่าขนาดตัวอย่าง การจะวิเคราะห์ข้อมูลลักษณะนี้ จึงต้องใช้เทคนิคการวิเคราะห์การถดถอยเชิงเส้นขั้นสูงอย่างการวิเคราะห์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษ ซึ่งปัจจุบัน การวิจัยที่เกี่ยวข้องกับการวิเคราะห์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษยังเป็นที่น่าสนใจอยู่ ทั้งนี้ เพื่อที่จะพัฒนาตัวแบบการถดถอยเชิงเส้นในรูปแบบใหม่ ๆ ซึ่งรองรับข้อมูลขนาดใหญ่ที่มีความสลับซับซ้อนมากขึ้นในปัจจุบัน

1.4 รูปแบบความสัมพันธ์จากการวิเคราะห์การถดถอย

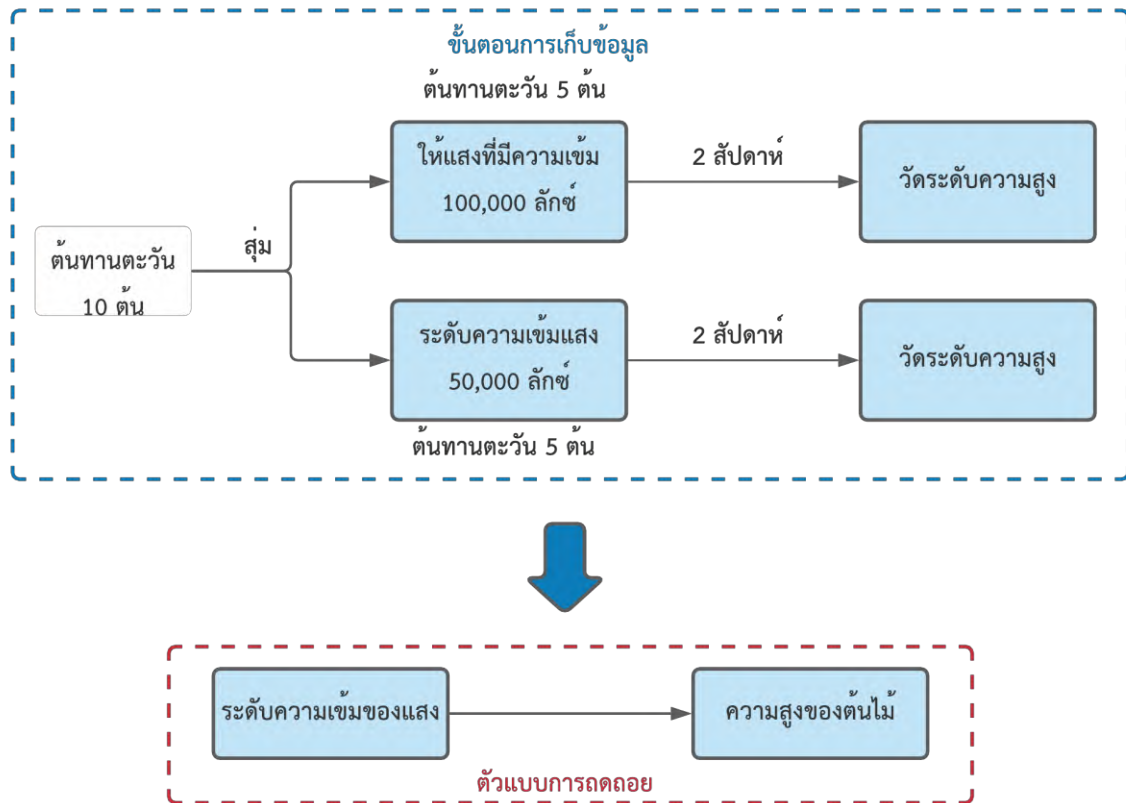
สำหรับข้อควรระวังในการสรุปผลที่ได้จากการใช้การวิเคราะห์การถดถอยเชิงเส้นในการศึกษาความสัมพันธ์ระหว่างตัวแปรโดยทั่วไป คือ การสรุปในรูปแบบความเกี่ยวพัน (Association) ซึ่งมีใช้การบ่งบอกถึงความสัมพันธ์เชิงสาเหตุ (Causal Relationship) ยกตัวอย่าง เช่น หากตัวแบบการถดถอยพบว่าจำนวนนักท่องเที่ยวมีความสัมพันธ์ไปในทิศทางเดียวกันแบบมีนัยสำคัญทางสถิติกับยอดขายสินค้า นั้นไม่จำเป็นว่าการที่จำนวนนักท่องเที่ยวเพิ่มขึ้นเป็นเหตุผลที่ทำให้ยอดขายสูงขึ้นไปด้วย ยกตัวอย่างเช่น ร้านค้าขายไอศกรีมซึ่งยอดขายที่เพิ่มขึ้นมีความสัมพันธ์กับอากาศที่ร้อนหรืออุณหภูมิที่สูงขึ้น แต่ในช่วงฤดูร้อน จำนวนนักท่องเที่ยวก็มักมีจำนวนมากขึ้นด้วย จากตัวอย่างนี้ เห็นได้ชัดว่าการที่ยอดขายไอศกรีมเพิ่มขึ้นไม่ได้เกิดจากสาเหตุที่จำนวนนักท่องเที่ยวเพิ่มขึ้น แต่ยังสรุปความได้ว่ายอดขายไอศกรีมเพิ่มขึ้นมีความสัมพันธ์กับจำนวนนักท่องเที่ยวที่เพิ่มขึ้น ดังนั้น โดยทั่วไป การสรุปผลในรูปแบบความสัมพันธ์ที่ได้จากการวิเคราะห์การถดถอยจึงอยู่ในรูปแบบความเกี่ยวพัน ไม่ใช่รูปแบบความสัมพันธ์เชิงสาเหตุ โดยเฉพาะการใช้ตัวแบบการถดถอยในการวิเคราะห์ข้อมูลที่ได้จากการสังเกต (Observational Data) ที่ไม่ได้พิจารณาปัจจัยอื่น ๆ ที่อาจเกี่ยวข้องกับยอดขายที่เพิ่มขึ้นเข้ามาในตัวแบบ เช่น อุณหภูมิที่สูงขึ้น เป็นต้น ซึ่งปัจจัยอื่น ๆ เหล่านี้ เรียกว่า ตัวแปรกวน (Confounding Variable) รูปที่ 1-1 แสดงความสัมพันธ์ระหว่างตัวแปรในตัวแบบและตัวแปรกวน



รูปที่ 1-1 ความสัมพันธ์ระหว่างตัวแปรในตัวแบบและตัวแปรกวน

อย่างไรก็ตาม หากการเก็บข้อมูลนั้นมาจากการทดลอง (Experimental Data) ซึ่งอาจมีการกำหนดตัวแปรควบคุมที่ชัดเจน และในการเก็บข้อมูลนั้นมีการสุ่ม (Randomization) หรือคือ การกำหนดค่าบางอย่างให้กับหน่วยทดลอง ในกรณีนี้ ข้อสรุปจากตัวแบบการถดถอยจะสามารถสรุปความสัมพันธ์ของตัวแปรได้ในรูปแบบความสัมพันธ์เชิงสาเหตุ ยกตัวอย่างเช่น ในการศึกษาการให้ความเข้มของแสงที่ระดับต่างกันจะมีความสัมพันธ์กับความสูงของต้นทานตะวันหรือไม่ โดยเมื่อเริ่มต้นการศึกษามีการนำต้นกล้าของต้นทานตะวันที่มีความสูงระดับเท่า ๆ กันมา 10 ต้น จากนั้นจึงทำการสุ่มต้นทานตะวันมา 5 ต้นได้รับความเข้มของแสงที่ระดับ 100,000 ลักซ์ และต้นทานตะวันอีก 5 ต้นจะได้รับความเข้มของแสงที่ระดับ 50,000 ลักซ์ โดยตัวแปรอื่น ๆ เช่น ปริมาณน้ำที่ให้ หรือจำนวนชั่วโมงที่มีการให้แสงต่อวัน จะมีการควบคุมให้เหมือนกันสำหรับต้นทานตะวันทั้ง 10 ต้น จากนั้นจึงวัดความสูงของต้นทานตะวันเมื่อผ่านไป 2 สัปดาห์ ซึ่งการศึกษานี้สามารถนำมาเขียนเป็นแผนภาพได้ดังแสดงในรูปที่ 1-2 หากในการศึกษานี้ มีการใช้การวิเคราะห์การถดถอยและพบว่ามีความสัมพันธ์

ระหว่างความเข้มของแสงและความสูงของต้นทานตะวันแบบมีนัยสำคัญทางสถิติ ข้อสรุปที่ได้จะอยู่ในรูปแบบความสัมพันธ์เชิงสาเหตุนั่นเอง



รูปที่ 1-2 การศึกษาความสัมพันธ์ระหว่างระดับความเข้มของแสงและความสูงของต้นทานตะวัน

1.5 บทสรุป

ในบทนี้ได้กล่าวถึงวัตถุประสงค์ของหนังสือเล่มนี้ ซึ่งเน้นหลักการที่สำคัญในการวิเคราะห์การถดถอยเชิงเส้น และแสดงการนำไปประยุกต์ใช้งานกับข้อมูลจริงโดยใช้โปรแกรม R นอกจากนี้ ในบทนี้ยังกล่าวถึงประวัติความเป็นมาของการวิเคราะห์การถดถอยเชิงเส้นซึ่งมีการใช้งานมาตั้งแต่ศตวรรษที่ 19 จนถึงปัจจุบัน ด้วยข้อมูลในยุคสมัยใหม่มีความซับซ้อนมากขึ้น รูปแบบและของเขตการทำงานของ การวิเคราะห์การถดถอยเชิงเส้นในปัจจุบันมีขยายขอบเขตและการเปลี่ยนแปลงไปเพื่อรองรับการวิเคราะห์ข้อมูลที่มีความซับซ้อนมากขึ้น และแม้ว่าเครื่องมือการวิเคราะห์การถดถอยจะเป็นเครื่องมือที่มีประโยชน์และมีการนำไปใช้งานเป็นวงกว้าง การสรุปผลรูปแบบความสัมพันธ์ที่ได้จากตัวแบบการถดถอยเป็นสิ่งที่ผู้ใช้ต้องพึงระวัง ซึ่งโดยทั่วไปรูปแบบความสัมพันธ์ที่ได้จากการวิเคราะห์การถดถอยโดยใช้ข้อมูลที่มาจากการสังเกตจะอยู่ในรูปแบบความเกี่ยวพัน ไม่ใช่รูปแบบความสัมพันธ์เชิงสาเหตุ เนื่องจากไม่สามารถจัดอิทธิพลที่มาจากตัวแปรกวนได้ แต่หากการเก็บข้อมูลนั้นมาจากการทดลองที่มีการสุ่มในการกำหนดค่าบางอย่างให้กับหน่วยทดลอง กรณีนี้ข้อสรุปที่ได้จากตัวแบบการถดถอยจะสามารถสรุปความสัมพันธ์ในรูปแบบความสัมพันธ์เชิงสาเหตุได้

บทที่ 2

การเขียนโปรแกรม R เบื้องต้น

ภาษา R เป็นภาษาที่รู้จักกันดีและใช้กันอย่างแพร่หลายในการวิเคราะห์ข้อมูล ซึ่งมีเครื่องมือวิเคราะห์ข้อมูลทางสถิติให้เลือกใช้อย่างครบถ้วน โปรแกรม R สามารถติดตั้งได้โดยง่ายและรองรับในระบบปฏิบัติการต่าง ๆ ได้แก่ Windows MacOS และ UNIX/Linux สำหรับจุดเด่นของโปรแกรม R คือ การเป็น Open-sourced Software หรือก็คือผู้ใช้สามารถเรียกใช้ฟังก์ชันต่าง ๆ ตลอดจนเรียกดูคำสั่งในฟังก์ชันต่าง ๆ รวมถึงหากผู้ใช้สร้างฟังก์ชันขึ้นใหม่เองก็สามารถนำฟังก์ชันเหล่านั้นให้ผู้อื่นเรียกใช้ได้โดยไม่มีค่าใช้จ่ายใด ๆ นอกจากนี้ ข้อดีของโปรแกรม R อีกอย่างหนึ่ง คือ การมีชุมชนของผู้ใช้โปรแกรม R ที่เข้มแข็งทั่วโลก ซึ่งพร้อมที่จะคอยช่วยเหลือและสนับสนุนหากท่านมีปัญหาในการเขียนโปรแกรม R

สำหรับประวัติความเป็นมาของภาษา R นั้น เริ่มต้นจากในปี ค.ศ. 1991 ศาสตราจารย์ รอส อีฮากา (Ross Ihaka) และ ศาสตราจารย์ โรเบิร์ต เจนเทิลแมน (Robert Gentleman) อาจารย์จากภาควิชาสถิติ มหาวิทยาลัยแห่งโอ๊คแลนด์ ประเทศนิวซีแลนด์ ได้พัฒนาภาษา R ขึ้น โดยมีต้นแบบจากภาษา S¹ ต่อมาในปี ค.ศ. 1995 ศาสตราจารย์ มาร์ติน มัชเลอร์ (Martin Mächler) มีส่วนร่วมให้ ศาสตราจารย์ รอส อีฮากา และ ศาสตราจารย์ โรเบิร์ต เจนเทิลแมนเผยแพร่โปรแกรม R ภายใต้ GNU General Public License² ซึ่งทำให้ผู้ใช้โปรแกรม R สามารถใช้งานได้ฟรีและนี่ก็ถือเป็นก้าวสำคัญของโปรแกรม R ที่ทำให้ผู้ใช้ทั่วโลกเข้าถึงได้ (Peng, 2012) สำหรับรายละเอียดการออกแบบและพัฒนาโปรแกรม R สามารถอ่านได้เพิ่มเติมที่ Ihara and Gentleman (1996)

ในบทนี้ จะได้กล่าวถึงการติดตั้งโปรแกรม R การเขียนโปรแกรมภาษา R ในการจัดการข้อมูลและวิเคราะห์ข้อมูลเบื้องต้น

2.1 การติดตั้งโปรแกรม R และ RStudio

2.1.1 การติดตั้งโปรแกรม R

การติดตั้งโปรแกรม R สามารถทำได้ดังขั้นตอนต่อไปนี้

1. เปิด Web Browser และเข้าไปที่หน้าเว็บไซต์ <https://www.r-project.org/>
2. คลิกที่ download R ดังรูปที่ 2-1

¹ ภาษา S ถูกพัฒนาโดยนักวิจัยใน Bell Lab ซึ่งเป็นหน่วยงานในบริษัท AT&T ประเทศสหรัฐอเมริกา ในปี ค.ศ. 1976 สำหรับใช้ในการวิเคราะห์ข้อมูลภายในองค์กร ซึ่งต่อมาภายหลังมีบริษัทเอกชนซื้อลิขสิทธิ์ นำมาต่อยอดและจัดจำหน่ายเชิงพาณิชย์ภายใต้ชื่อ S-PLUS

² GNU General Public License คือใบอนุญาตการใช้งานโปรแกรม (Software License) ใต้ฟรี นอกจากนี้ผู้ใช้สามารถปรับเปลี่ยนและแบ่งปันโปรแกรมให้ผู้อื่นใช้งานได้ (ที่มา: <https://www.gnu.org>)



[Home]

Download

CRAN

R Project

About R

Logo

Contributors

What's New?

Reporting Bugs

Conferences

Search

Get Involved: Mailing Lists

Developer Pages

R Blog

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).

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.

News

- [R version 4.1.0 \(Camp Pontanezen\)](#) has been released on 2021-05-18.
- [R version 4.0.5 \(Shake and Throw\)](#) was released on 2021-03-31.
- Thanks to the organisers of useR! 2020 for a successful online conference. Recorded tutorials and talks from the conference are available on the [R Consortium YouTube channel](#).
- You can support the R Foundation with a renewable subscription as a [supporting member](#)

รูปที่ 2-1 หน้าแรกของเว็บไซต์ <https://www.r-project.org/>

3. เลือก Comprehensive R Archive Network (CRAN) ซึ่งเป็นที่ตั้งของ R Mirror Server ที่กระจายอยู่ทั่วโลก
4. คลิกที่ลิงค์เพื่อดาวน์โหลดไฟล์ติดตั้งโปรแกรม R ตามระบบปฏิบัติการที่ต้องการ ดังรูปที่ 2-2

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.

รูปที่ 2-2 โปรแกรมติดตั้ง R ในระบบปฏิบัติการต่าง ๆ

ในกรณีที่ต้องการลงโปรแกรม R สำหรับระบบปฏิบัติการ Windows หลังจากคลิกลิงค์ “Download R for Windows” ให้คลิกที่ “Install R for the first time” ดังรูปที่ 2-3

R for Windows

Subdirectories:

[base](#)

[contrib](#)

[old contrib](#)

[Rtools](#)

Binaries for base distribution. This is what you want to [install R for the first time](#).

Binaries of contributed CRAN packages (for R >= 2.13.x; managed by Uwe Ligges). There is also information on [third party software](#) available for CRAN Windows services and corresponding environment and make variables.

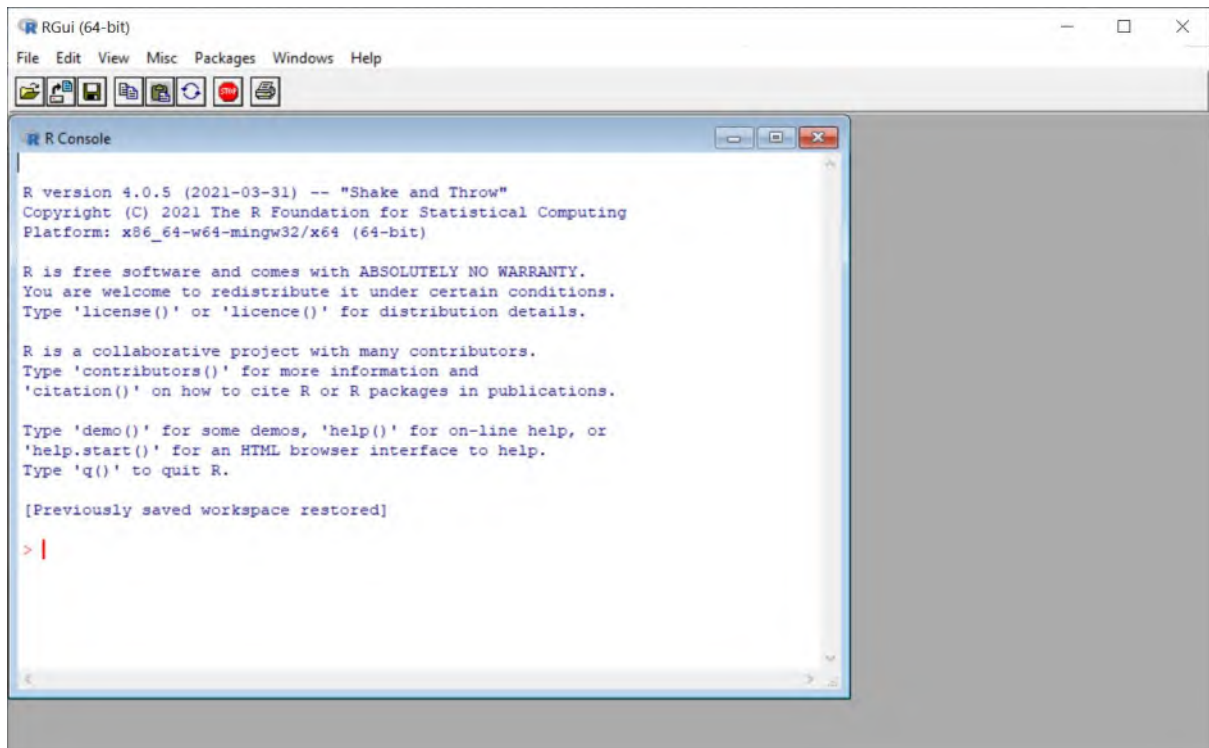
Binaries of contributed CRAN packages for outdated versions of R (for R < 2.13.x; managed by Uwe Ligges).

Tools to build R and R packages. This is what you want to build your own packages on Windows, or to build R itself.

รูปที่ 2-3 การลงโปรแกรม R สำหรับระบบปฏิบัติการ Windows

5. เปิดไฟล์ที่ดาวน์โหลดและทำตามขั้นตอนที่ไฟล์ติดตั้งแนะนำ

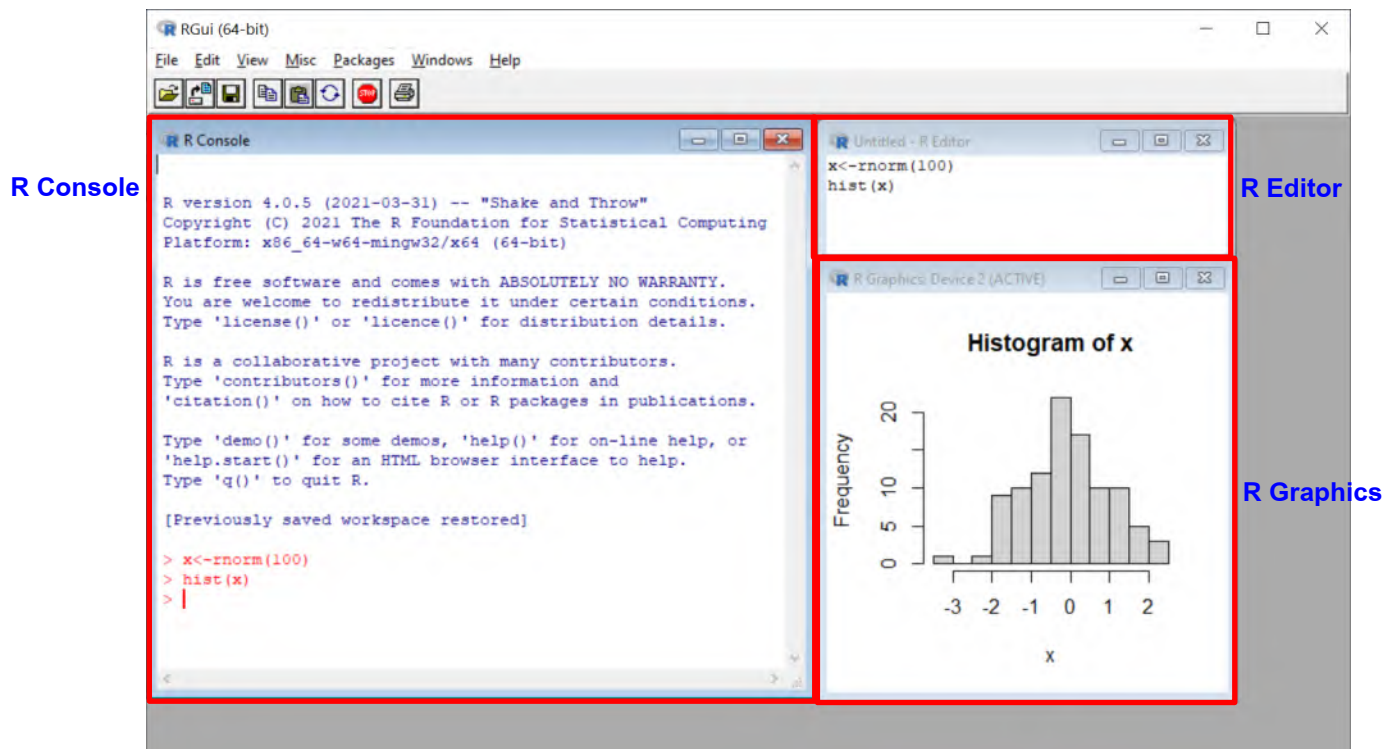
เมื่อเสร็จขั้นตอนที่ 5 โปรแกรม R ได้ถูกลงที่เครื่องคอมพิวเตอร์เรียบร้อยแล้ว หากเปิดโปรแกรม R จะได้หน้าจอแสดงดังรูปที่ 2-4



รูปที่ 2-4 ตัวอย่างหน้าจอโปรแกรม R

รูปที่ 2-4 แสดง R Console ซึ่งผู้ใช้สามารถพิมพ์คำสั่งภาษา R ใน R Console ได้โดยตรง โดยเมื่อพิมพ์คำสั่งหนึ่งเสร็จ ให้กดปุ่ม Enter เพื่อให้โปรแกรม R ประมวลผลและส่งผลลัพธ์ออกมา

การใช้ Text Editor ใน R จะช่วยให้การเขียนโปรแกรมทำได้สะดวกขึ้น การเปิด Text Editor ใน R สามารถทำได้โดยคลิกที่ "File -> New script" ซึ่งจะทำให้การเปิด R Editor ขึ้นมา ซึ่งผู้ใช้สามารถพิมพ์คำสั่งต่างๆ ได้ใน R Editor และเมื่อต้องการประมวลผลคำสั่งให้ Highlight ชุดคำสั่งที่ต้องการ (อาจจะประกอบด้วยคำสั่งหลายคำสั่งหรือหลายบรรทัด) และคลิกที่ปุ่ม  Run line or selection ที่อยู่แถบด้านบน เพื่อส่งคำสั่งจาก R Editor ไปยัง R Console เพื่อประมวลผล หากมีการส่งคำสั่งการวาดกราฟต่าง ๆ หน้าต่าง R Graphics ก็จะปรากฏภาพตามคำสั่งขึ้นมา ดังรูปที่ 2-5



รูปที่ 2-5 ส่วนต่าง ๆ ในโปรแกรม R

2.1.2 การติดตั้งโปรแกรม RStudio

แม้ว่าโปรแกรม R จะมีหน้าต่าง R Editor เพื่อช่วยอำนวยความสะดวกในการเขียนโปรแกรม อย่างไรก็ตาม เครื่องมือในการอำนวยความสะดวกใน R Editor ยังมีข้อจำกัดอยู่ เช่น การเขียนคำสั่งใน R Editor ผู้ใช้งานจำเป็นต้องจำคำสั่งในการเรียกฟังก์ชันใน R ต่าง ๆ ด้วยตนเอง เป็นต้น โปรแกรม RStudio เป็นโปรแกรมที่ได้รับความนิยมเป็นอย่างมาก เนื่องจากมีเครื่องมือในการอำนวยความสะดวกในการเขียนโปรแกรมภาษา R ที่ครบถ้วน ปัจจุบัน RStudio มีหลากหลายเวอร์ชันและรองรับได้ทั้งภาษา R และ Python ข้อดีของ RStudio อีกอย่างคือ การที่ RStudio มีเวอร์ชัน Open Source Edition ซึ่งผู้ใช้งานสามารถใช้งานได้โดยไม่เสียค่าใช้จ่ายใด ๆ ปัจจุบัน RStudio เป็นผลิตภัณฑ์ของบริษัท Posit และสามารถใช้งานผ่านการติดตั้งโปรแกรมบนเครื่องคอมพิวเตอร์ (RStudio Desktop) หรือบน Server (RStudio Server) รวมไปถึงการใช้งานผ่าน RStudio บนเว็บเบราว์เซอร์ (Web Browser) ผ่านเทคโนโลยีคลาวด์ได้ (Posit Cloud)

สำหรับการติดตั้งโปรแกรม RStudio Desktop เมื่อผู้ใช้ทำการติดตั้งโปรแกรม R เป็นที่เรียบร้อยแล้ว สามารถทำตามขั้นตอนด้านล่างเพื่อติดตั้งโปรแกรม RStudio (เวอร์ชัน Open Source Edition) บนคอมพิวเตอร์ที่ใช้งานได้

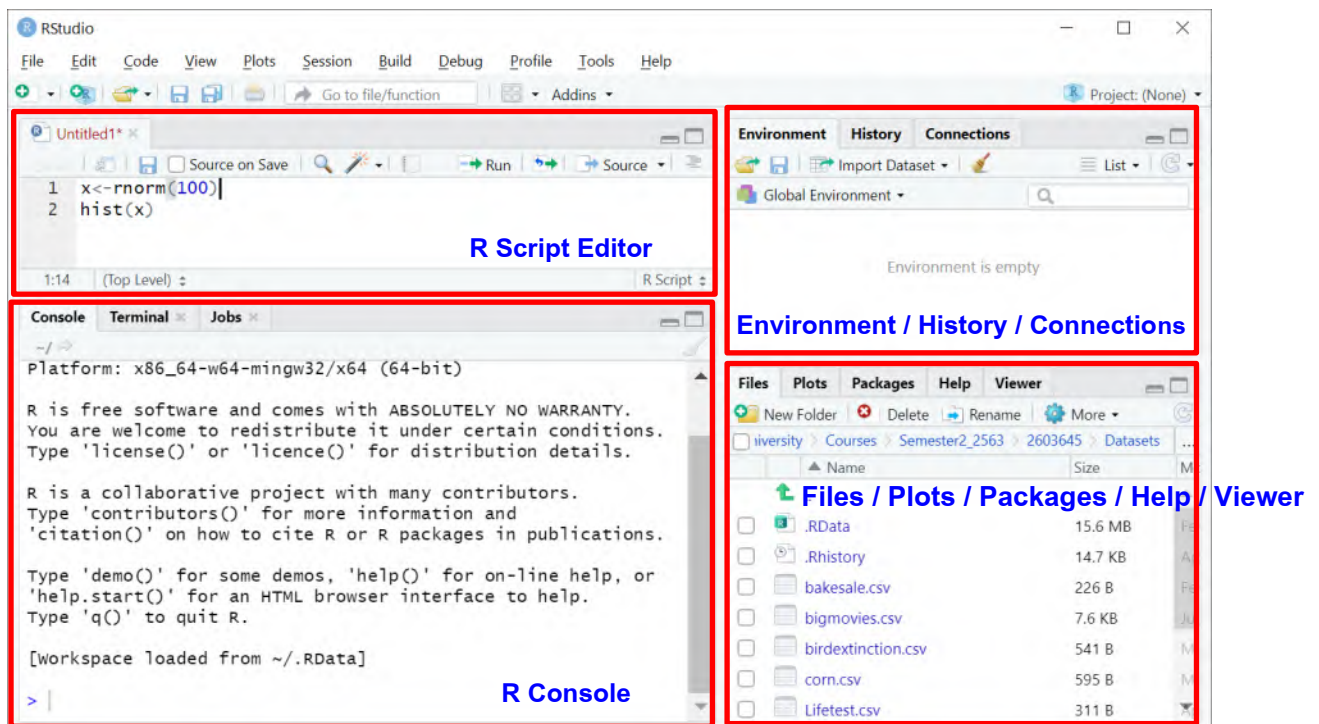
1. เปิดเว็บเบราว์เซอร์และเข้าไปที่หน้าเว็บไซต์ <https://posit.co/download/rstudio-desktop/>
2. เลื่อนมาด้านล่างและคลิกที่ลิงค์เพื่อดาวน์โหลดไฟล์ติดตั้งโปรแกรม RStudio ตามระบบปฏิบัติการที่ต้องการ ดังรูปที่ 2-6

OS	Download	Size	SHA-256
Windows 10/11	RSTUDIO-2023.09.1-494.EXE	214.34 MB	FE62B784
macOS 12+	RSTUDIO-2023.09.1-494.DMG	385.88 MB	25A2CC51
Ubuntu 20/Debian 11	RSTUDIO-2023.09.1-494-AMD64.DEB	147.81 MB	89054E48
Ubuntu 22/Debian 12	RSTUDIO-2023.09.1-494-AMD64.DEB	148.61 MB	B04D8132
Fedora 19/Red Hat 7	RSTUDIO-2023.09.1-494-X86_64.RPM	164.41 MB	87CD0A0A
OpenSUSE 15	RSTUDIO-2023.09.1-494-X86_64.RPM	148.46 MB	1E9AEA34
Fedora 34/Red Hat 8	RSTUDIO-2023.09.1-494-X86_64.RPM	165.63 MB	DE43E4A5
Fedora 36/Red Hat 9	RSTUDIO-2023.09.1-494-X86_64.RPM	158.17 MB	2CAD4D99

รูปที่ 2-6 โปรแกรมติดตั้ง RStudio Desktop ในระบบปฏิบัติการต่าง ๆ

3. เปิดไฟล์ที่ดาวน์โหลดและทำตามขั้นตอนที่ไฟล์ติดตั้งแนะนำ

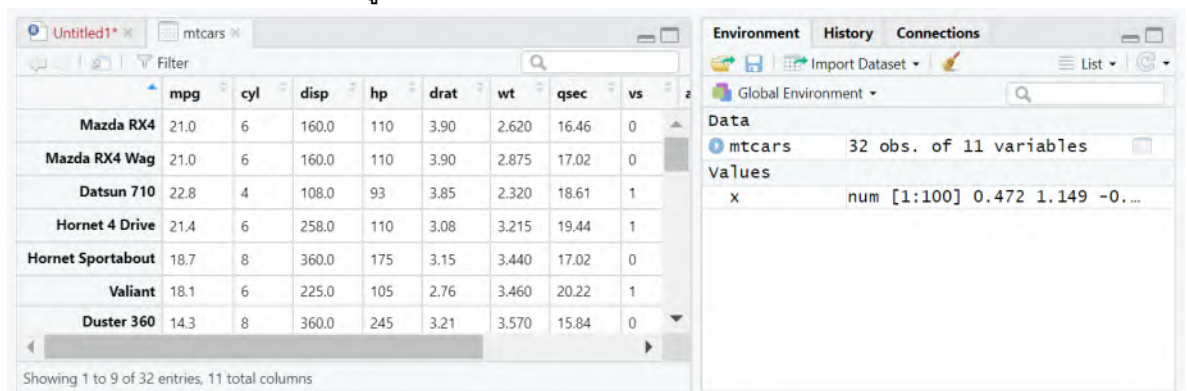
เมื่อเสร็จขั้นตอนที่ 5 โปรแกรม RStudio จะถูกติดตั้งที่เครื่องคอมพิวเตอร์เป็นที่เรียบร้อยแล้ว หากเปิดโปรแกรม RStudio ขึ้นมา ให้ผู้ใช้เลือกที่ “File -> New File -> R Script” เพื่อเปิดหน้าต่าง R Editor ขึ้นมา ซึ่งโดยทั่วไปหน้าจอของ RStudio จะประกอบไปด้วย 4 ส่วน ดังรูปที่ 2-7



รูปที่ 2-7 ส่วนประกอบต่างๆ ในโปรแกรม RStudio

ในการส่งผ่านชุดคำสั่งใน R Script Editor มาประมวลผลใน R Console สามารถทำได้โดย Highlight ชุดคำสั่งใน R Script Editor และคลิกที่ปุ่ม  เพื่อส่งคำสั่งไปประมวลผล สำหรับแท็บ (Tab) ต่าง ๆ แต่ละแท็บทางด้านขวามือ มีรายละเอียดต่าง ๆ ดังนี้

แท็บ Environment แสดงวัตถุ (Object) ต่าง ๆ ที่อยู่ในพื้นที่ทำงาน (Workspace) ปัจจุบัน กรณีที่วัตถุเป็นตาราง สามารถคลิกที่ชื่อวัตถุเพื่อแสดงข้อมูลที่เก็บในตารางดังกล่าวได้ ซึ่งจะแสดงทางด้านซ้ายมือ ดังรูปที่ 2-8



รูปที่ 2-8 Environment Tab ใน RStudio

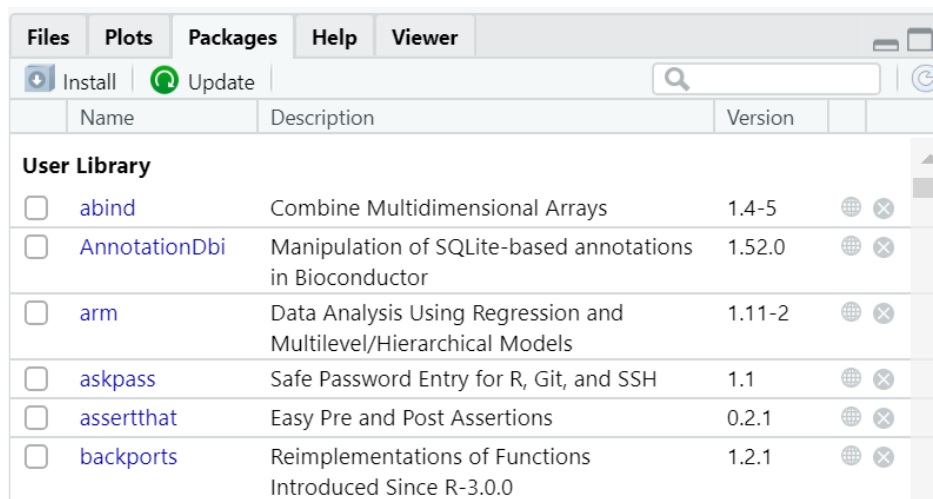
แท็บ History แสดงประวัติของคำสั่งที่ถูกประมวลผล

แท็บ Connections ใช้สำหรับเชื่อมต่อไปยังแหล่งข้อมูลต่าง ๆ เช่น การเชื่อมต่อฐานข้อมูลผ่าน Open Database Connectivity (ODBC) เป็นต้น

แท็บ **Files** แสดงไฟล์ต่าง ๆ ที่อยู่ในโฟลเดอร์ที่ใช้งานปัจจุบัน (Working Directory)

แท็บ **Plots** แสดงกราฟต่าง ๆ ที่สร้างขึ้น

แท็บ **Packages** แสดงแพ็คเกจที่ติดตั้งอยู่ในเครื่องเรียบร้อยแล้ว เพื่อให้ผู้ใช้เรียกใช้ใน
พื้นที่ทำงานปัจจุบัน ตลอดจนการติดตั้งและอัปเดตแพ็คเกจให้ทันสมัย



	Name	Description	Version		
User Library					
☐	abind	Combine Multidimensional Arrays	1.4-5	⌐	✕
☐	AnnotationDbi	Manipulation of SQLite-based annotations in Bioconductor	1.52.0	⌐	✕
☐	arm	Data Analysis Using Regression and Multilevel/Hierarchical Models	1.11-2	⌐	✕
☐	askpass	Safe Password Entry for R, Git, and SSH	1.1	⌐	✕
☐	assertthat	Easy Pre and Post Assertions	0.2.1	⌐	✕
☐	backports	Reimplementations of Functions Introduced Since R-3.0.0	1.2.1	⌐	✕

รูปที่ 2-9 Packages Tab ใน RStudio

แท็บ **Help** แสดงไฟล์ช่วยเหลือ

แท็บ **Viewer** ใช้สำหรับแสดงหน้าเว็บ HTML ที่ถูกจัดเก็บในเครื่องคอมพิวเตอร์ โดยใช้คำสั่ง “viewer”

2.2 คำสั่งเบื้องต้นใน R

คำสั่งเบื้องต้นใน R จะขอก้าวถึงการคำนวณทางคณิตศาสตร์ การสร้างวัตถุ และการใส่คำอธิบาย (Comment) ใน R

การคำนวณทางคณิตศาสตร์

ในภาษา R สามารถใช้เครื่องหมายทางคณิตศาสตร์ทั่วไปในการคำนวณได้ ซึ่งได้แก่ +(บวก) -(ลบ) *(คูณ) และ /(หาร) นอกจากนี้ ยังสามารถใช้วงเล็บในคำสั่งการคำนวณได้ ตัวอย่างในการคำนวณทางคณิตศาสตร์ เช่น

```
> 4+5/2
[1] 6.5
> (4+5)/2
[1] 4.5
```

การสร้างวัตถุ (Object)

การสร้างวัตถุใน R สามารถทำได้โดยใช้สัญลักษณ์ `=` หรือ `<-` หรือ `->` หากผู้ใช้ใช้สัญลักษณ์ `<-` ในการสร้างวัตถุ ชื่อวัตถุที่สร้างจะต้องอยู่ทางซ้าย แต่หากผู้ใช้ใช้สัญลักษณ์ `->` ชื่อวัตถุที่สร้างจะอยู่ทางขวายกตัวอย่างเช่น

```
> x=5
> x
[1] 5
> y<-10
> y
[1] 10
> x+y->z
> z
[1] 15
```

เนื่องจากภาษา R มีลักษณะไวต่ออักษรพิมพ์ใหญ่/เล็ก (Case Sensitive) ซึ่งหมายความว่า การใช้ตัวอักษรพิมพ์ใหญ่ (Capital Letter) หรือเล็ก (Lower Case Letter) มีความแตกต่างกัน เช่น

```
> y<-2
> Y<-3
> y+Y
[1] 5
```

นอกจากการสร้างวัตถุแล้ว หากผู้ใช้ต้องการทราบว่าในพื้นที่ทำงาน (Workspace) ปัจจุบันมีวัตถุใดบ้างสามารถใช้คำสั่ง `ls` ในการแสดงวัตถุทั้งหมด และใช้คำสั่ง `rm` ในการลบวัตถุ ตัวอย่างเช่น

```
> ls()
[1] "x" "y" "Y" "z"
> rm(z)
> ls()
[1] "x" "y" "Y"
```

ในกรณีที่ใช้ RStudio สามารถแสดงวัตถุในพื้นที่ทำงานในแท็บ Environment ได้

การใส่คำอธิบาย (Comment)

ผู้ใช้สามารถใส่คำอธิบายได้หลังสัญลักษณ์ `#` โดยข้อความดังกล่าวจะไม่ถูกประมวลผล ตัวอย่างเช่น

```
> 10^5 # ^ = power/exponent
[1] 1e+05
> 11 %% 2 # %% = Modulo
[1] 1
```

2.3 ประเภทข้อมูลใน R

R จัดเก็บข้อมูลได้หลายหลายประเภท ในที่นี้จะขอกล่าวถึงประเภทข้อมูลเบื้องต้นใน R ดังนี้

ตัวเลข (Numeric)

ข้อมูลประเภทตัวเลขใน R ผู้ใช้ไม่จำเป็นต้องระบุหรือประกาศเพิ่มเติมว่าตัวเลขนั้นเป็นตัวเลขจำนวนเต็ม หรือตัวเลขจำนวนจริง ซึ่งข้อมูลประเภทตัวเลขสามารถนำไปคำนวณทางคณิตศาสตร์ได้

```
> a<-5.2
> a
[1] 5.2
> b<-5
> b
[1] 5
```

ตัวอักษร (Characters)

ตัวอักษรใน R ผู้ใช้จะต้องใส่ข้อมูลตัวอักษรภายในเครื่องหมายัญประกาศ ' (Single Quote) หรือ " (Double Quotation Quote) ตัวอย่างเช่น

```
> text1<-"Hello world!"
> text1
[1] "Hello world!"
> text2<-'123'
> text2
[1] "123"
```

จากตัวอย่างข้างต้น จะสังเกตได้ว่าการแสดง `text2` จะแสดงข้อมูลอยู่ในเครื่องหมายัญประกาศ ซึ่งมีความแตกต่างจากข้อมูลประเภทตัวเลขที่กล่าวถึงก่อนหน้านี้ หากต้องการเปลี่ยนข้อมูลใน `text2` ให้เป็นตัวเลขสามารถทำได้โดยใช้คำสั่ง `as.numeric` เช่น

```
> as.numeric(text2)
[1] 123
```

ในทางกลับกัน หากต้องการเปลี่ยนตัวเลขให้เป็นอักษร สามารถทำได้โดยใช้คำสั่ง `as.character` เช่น

```
> as.character(100)
[1] "100"
```

ตรรกะ (Logical)

ข้อมูลประเภทตรรกะ จะเก็บข้อมูลอยู่สองอย่างคือ TRUE หรือ FALSE ซึ่งฟังก์ชันหลายอย่างจะส่งผลลัพธ์ออกมาเป็นตรรกะ ยกตัวอย่างเช่น

```
> is.numeric(123)
[1] TRUE
> is.character(123)
[1] FALSE
> is.na(123)
[1] FALSE
> is.na(NA)
[1] TRUE
```

จากตัวอย่างด้านบนจะเห็นว่าคำสั่ง `is.numeric(123)` ใช้ในการตรวจสอบว่าข้อมูลที่อยู่ภายในวงเล็บเป็นตัวเลขหรือไม่ สำหรับคำสั่ง `is.character(123)` ใช้ในการตรวจสอบว่าข้อมูลเป็นประเภทตัวอักษรหรือไม่ นอกจากนั้น ในตัวอย่างนี้ จะเห็นว่าสามารถใช้ `NA` แทนข้อมูลที่ขาดหายไป (Missing Data) และสามารถใช้ฟังก์ชัน `is.na` ในการตรวจสอบว่าข้อมูลเป็นข้อมูลที่ขาดหายไปหรือไม่

เวกเตอร์ (Vector)

เวกเตอร์ใน R เป็นการเก็บข้อมูลประเภทเดียวกันทั้งหมดออกเป็นแถว โดยอาจจะเป็นเวกเตอร์ของตัวเลข เวกเตอร์ของตัวหนังสือ หรือเวกเตอร์ของตรรกะ ตัวอย่างการจัดเก็บข้อมูลเวกเตอร์ใน R สามารถแสดงได้ดังรูปที่ 2-10



รูปที่ 2-10 การจัดเก็บเวกเตอร์ขนาด k

ในการสร้างเวกเตอร์สามารถทำได้หลายรูปแบบเช่นใช้ฟังก์ชัน `c` ในการเชื่อมต่อข้อมูล เช่น

```
> x<-c(1,2,3)
> x
[1] 1 2 3
> y<-c("A", "B", "C", "D", "E")
> y
[1] "A" "B" "C" "D" "E"
> z<-c(TRUE, FALSE)
> z
[1] TRUE FALSE
```

นอกจากนี้ ในการสร้างเวกเตอร์ของตัวเลข ยังสามารถใช้คำสั่ง `seq` ในการสร้างเวกเตอร์ตามลำดับ (Sequence) เช่นการสร้างเวกเตอร์ของตัวเลขโดยเริ่มต้นจาก 1 ถึง 10 โดยบวกเพิ่มทีละ 2 สามารถทำได้โดย

```
> x<-seq(from=1, to=10, by=2)
> x
[1] 1 3 5 7 9
```

กรณีที่ต้องการใช้คำสั่ง `seq` โดยที่ `by=1` สามารถทำได้โดยง่าย ดังนี้

```
> 1:10
[1] 1 2 3 4 5 6 7 8 9 10
```

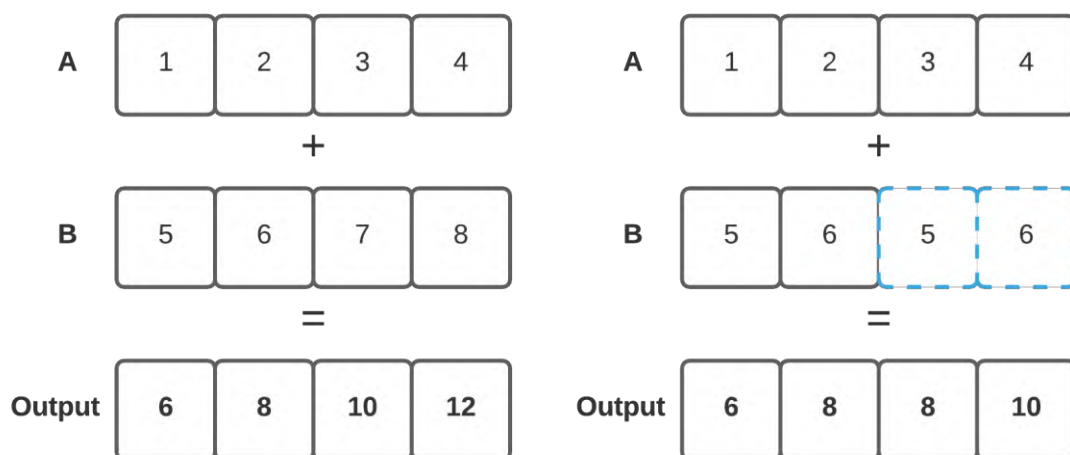
สำหรับข้อมูลที่ถูกจัดเก็บอยู่ในรูปเวกเตอร์ ผู้ใช้สามารถตรวจสอบขนาดของเวกเตอร์ได้โดยใช้ฟังก์ชัน **length** ดังนี้

```
> length(1:10)
[1] 10
```

สำหรับข้อมูลที่ถูกจัดเก็บอยู่ในรูปเวกเตอร์ของตัวเลข สามารถคำนวณเชิงคณิตศาสตร์ได้ เช่น

```
> x<-1:5
> x+3
[1] 4 5 6 7 8
```

นอกจากนี้ ยังสามารถคำนวณเชิงคณิตศาสตร์ระหว่างเวกเตอร์กับเวกเตอร์ได้ โดยลักษณะการทำงานในโปรแกรม R จะดำเนินการทางคณิตศาสตร์ของข้อมูลที่อยู่ในตำแหน่งเดียวกันของ 2 เวกเตอร์ ยกตัวอย่าง ดังรูปที่ 2-11 ทางซ้ายมือ จะเห็นได้ว่าผลลัพธ์ที่ได้จากการบวก 2 เวกเตอร์ A และ B จะได้เวกเตอร์ขนาด 4 โดยผลลัพธ์ที่ได้แต่ละช่องจะเกิดจากผลรวมของ 2 ตัวเลข ที่อยู่ตำแหน่งเดียวกันของเวกเตอร์ A และ B ในกรณีที่เวกเตอร์ A และ B มีขนาดไม่เท่ากัน ผลลัพธ์ที่ได้จะเป็นเวกเตอร์ที่มีขนาดเท่ากับขนาดที่มากที่สุดระหว่างขนาดของเวกเตอร์ A และเวกเตอร์ B โดย R จะทำการต่อเวกเตอร์ที่สั้นกว่าโดยซ้ำชุดข้อมูลในเวกเตอร์ดังกล่าว ตัวอย่างเช่น ในรูปที่ 2-11 ทางด้านขวามือ จะพบว่าเวกเตอร์ B มีขนาด 2 ซึ่งมีขนาดน้อยกว่าเวกเตอร์ B ซึ่งมีขนาด 4 ดังนั้น ในการดำเนินการทางคณิตศาสตร์ระหว่าง 2 เวกเตอร์ดังกล่าว โปรแกรม R จะทำการต่อเวกเตอร์ B โดยซ้ำชุดข้อมูลใน B ตั้งแต่ตำแหน่งแรก ดังที่เห็นในช่องรอยประสีฟ้า แล้วจึงทำการคำนวณค่าจาก 2 เวกเตอร์ในตำแหน่งเดียวกัน



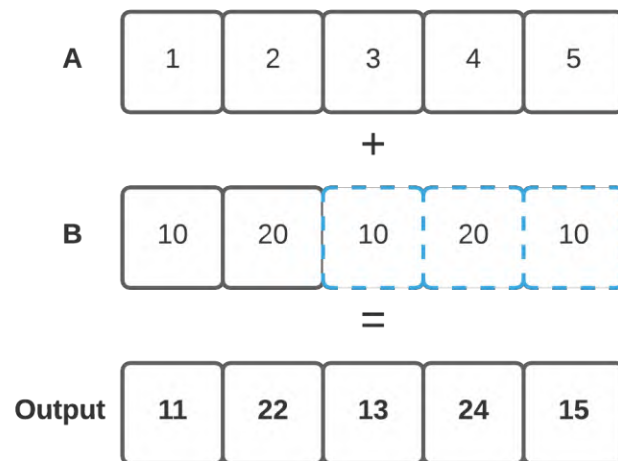
รูปที่ 2-11 การดำเนินการทางคณิตศาสตร์ระหว่างสองเวกเตอร์

สำหรับคำสั่งในการดำเนินการระหว่างสองเวกเตอร์ ดังตัวอย่างที่แสดงในรูปที่ 2-11 สามารถเขียนได้ดังนี้

```
> # Left panel
> A<-1:4
> A
[1] 1 2 3 4
> B<-5:8
> A
[1] 1 2 3 4
> A+B
[1] 6 8 10 12
>
> # Right panel
> A<-1:4
> A
[1] 1 2 3 4
> B<-5:6
> B
[1] 5 6
> A+B
[1] 6 8 8 10
```

ตัวอย่างถัดมา แสดงให้เห็นว่าในกรณีที่ดำเนินการระหว่างสองเวกเตอร์ที่มีขนาดแตกต่างกัน โดยที่ขนาดของเวกเตอร์ที่สั้นกว่าเมื่อนำมาซ้ำชุดข้อมูลเพื่อจะดำเนินการคำนวณทางคณิตศาสตร์แล้ว ไม่สามารถซ้ำชุดข้อมูลได้ครบตามขนาดของเวกเตอร์ที่สั้นกว่า กรณีนี้จะเกิดข้อความเตือน (Warning Message) ขึ้นมาอย่างไรก็ตาม การดำเนินการระหว่างสองเวกเตอร์ที่มีขนาดแตกต่างกันยังสามารถทำได้ โดยหลักการทำงานสองเวกเตอร์ดังกล่าวจะแสดงในรูปที่ 2-12 ในกรณีนี้จะเห็นได้ว่า จะต้องทำการซ้ำเวกเตอร์ A ไปสองรอบ แต่ในรอบที่สอง การซ้ำชุดข้อมูลจะเป็นการซ้ำที่ไม่ครบตามขนาดของเวกเตอร์ B ดังแสดงในช่องรอยประสีฟ้า

```
> A<-1:5
> A
[1] 1 2 3 4 5
> B<-c(10,20)
> B
[1] 10 20
> A+B
[1] 11 22 13 24 15
warning message:
In A + B : longer object length is not a multiple of shorter
object length
```



รูปที่ 2-12 การดำเนินการทางคณิตศาสตร์ระหว่างสองเวกเตอร์ที่มีขนาดต่างกัน

สำหรับการเข้าถึงข้อมูลบางส่วนของเวกเตอร์ ผู้ใช้สามารถระบุตำแหน่งของเวกเตอร์ได้ในเครื่องหมาย `[]` หรือหากใส่ตัวเลขที่ติดลบในเครื่องหมาย `[]` จะหมายความว่า จะไม่รวมข้อมูลในตำแหน่งดังกล่าว ยกตัวอย่างเช่น

```
> A<-1:5
> A
[1] 1 2 3 4 5
> A[1]
[1] 1
> A[2:4]
[1] 2 3 4
> A[-c(1,3)]
[1] 2 4 5
```

นอกจากนี้ ผู้ใช้งานยังสามารถใส่ข้อมูลประเภทตรรกะภายในเครื่องหมาย `[]` โดยจะแสดงข้อมูลในกรณีที่ข้อมูลตรรกะนั้นเป็น TRUE เท่านั้น เช่น

```
> A<-1:5
> A
[1] 1 2 3 4 5
> A[c(TRUE, TRUE, FALSE, TRUE, FALSE)]
[1] 1 2 4
```

เมทริกซ์ (Matrix)

เมทริกซ์ คือ ตารางสี่เหลี่ยมที่บรรจุตัวเลขในแต่ละช่อง รูปที่ 2-13 แสดงภาพเมทริกซ์ขนาด $r \times c$ ซึ่งมี r แถว (Row) และ c หลัก (Column) เนื่องจากเมทริกซ์เป็นตารางสองมิติของแถวกับหลัก ดังนั้นเวลาเข้าถึงข้อมูลบางส่วนในเมทริกซ์สามารถเข้าถึงได้โดยการระบุตำแหน่งของแถวและหลักในเครื่องหมาย `[]` โดยมีเครื่องหมายจุลภาค (Comma) คั่นระหว่างเลขแถวและหลัก ดังแสดงในรูปที่ 2-13

	หลักที่ 1	2	...	c
แถวที่ 1	[1, 1]	[1, 2]	...	[1, c]
2	[2, 1]			
...				
r	[r, 1]			[r, c]

รูปที่ 2-13 เมทริกซ์

สำหรับการสร้างเมทริกซ์เบื้องต้นใน R สามารถทำได้โดยใช้ฟังก์ชัน `matrix` โดยสามารถระบุข้อมูลเข้าเพิ่มเติมหรือที่เรียกว่าอาร์กิวเมนต์ (Argument) ได้ คือ เวกเตอร์ของข้อมูล (`data`) จำนวนแถว (`nrow`) จำนวนหลัก (`ncol`) และตรรกะว่าจะใส่ข้อมูลตามแถวในเมทริกซ์หรือไม่ (`byrow`)

```
> A<-matrix(data=c(1,2,3,4,5,6), ncol=3, nrow=2, byrow=TRUE)
> A
  [,1] [,2] [,3]
[1,]  1   2   3
[2,]  4   5   6
> B<-matrix(data=c(1,2,3,4,5,6), ncol=3, nrow=2, byrow=FALSE)
> B
  [,1] [,2] [,3]
[1,]  1   3   5
[2,]  2   4   6
```

สำหรับตัวอย่างในการเข้าถึงข้อมูลในเมทริกซ์สามารถทำได้ดังตัวอย่างด้านล่าง

```
> A<-matrix(c(1,2,3,4,5,6), ncol=3, nrow=2, byrow=TRUE)
> A
  [,1] [,2] [,3]
[1,]  1   2   3
[2,]  4   5   6
> A[1,2] # show element in the 1st row 2nd column
[1] 2
```

```
> A[1,] # Show elements in the 1st row
[1] 1 2 3
> A[,2] # Show element in the 2nd column
[1] 2 5
```

ในการดำเนินการทางคณิตศาสตร์ระหว่างเมทริกซ์ สามารถใช้เครื่องหมายทางคณิตศาสตร์ ได้แก่ บวก (+) ลบ (-) และคูณ (%) นอกจากนี้ ยังมีฟังก์ชันที่สำคัญที่เกี่ยวข้องกับเมทริกซ์ เช่น ฟังก์ชัน `dim` ใช้ในการแสดงขนาดของเมทริกซ์ ฟังก์ชัน `t` ใช้ในสลับเปลี่ยนเมทริกซ์ (Matrix Transpose) และฟังก์ชัน `solve` ใช้ในการหาเมทริกซ์ผกผัน (Inverse Matrix) เป็นต้น

แฟคเตอร์ (Factor)

แฟคเตอร์ใช้สำหรับเก็บข้อมูลเชิงคุณภาพ (Qualitative Data) โดย R สามารถเก็บข้อมูลได้ทั้งข้อมูลลักษณะนาม (Nominal Data) และข้อมูลเชิงอันดับ (Ordinal Data) วิธีการสร้างแฟคเตอร์สามารถทำได้ตัวอย่างดังนี้

```
> # Example of nominal data
> gender<-factor(x=c("Male", "Male", "Female"),
+ levels=c("Male", "Female"))
> gender
[1] Male Male Female
Levels: Male Female
>
> # Example of ordinal data
> rating<-factor(x=c("Fair", "Good", "Good"),
+ levels=c("worst", "Bad", "Fair", "Good", "Excellent"),
+ ordered=TRUE)
> rating
[1] Fair Good Good
Levels: worst < Bad < Fair < Good < Excellent
```

จากตัวอย่างคำสั่งด้านบน จะสังเกตได้ว่าข้อมูลประเภทแฟคเตอร์เวลาแสดงผลจะแสดงข้อมูลออกมาเป็นตัวหนังสือแต่จะไม่อยู่ในเครื่องหมาย “ ” (Double Quote) นอกจากนี้ ยังแสดงในส่วนของระดับ (Level) ซึ่งบ่งบอกถึงชื่อกลุ่ม (Category ทั้งหมดที่เป็นไปได้ กรณีผู้ใช้ใช้คำสั่ง `ordered=TRUE` จะเป็นการเก็บข้อมูลเชิงอันดับ และในส่วนของระดับ (Level) ก็จะมีการระบุการเรียงอันดับที่ชัดเจน

เดต้าเฟรม (Data Frame)

เดต้าเฟรมในโปรแกรม R เปรียบเสมือนตารางสองมิติเช่นเดียวกับเมทริกซ์ แต่เดต้าเฟรมสามารถเก็บข้อมูลได้หลากหลายประเภท โดยที่ข้อมูลประเภทเดียวกันจะอยู่ในหลักเดียวกัน ยกตัวอย่างดังรูปที่ 2.14 จะเห็นได้ว่าหลักที่ 1 เป็นข้อมูลตัวหนังสือ หลักที่ 2 เป็นข้อมูลแฟคเตอร์ หลักที่ 3 เป็นข้อมูลตัวเลข เป็นต้น

	Name	Gender	Age
แถวที่ 1	John	Male	26
2	Ken	Male	32
3	Mary	Female	30
4	Fred	Male	25
5	Jane	Female	24

รูปที่ 2-14 ตัวอย่างเดต้าเฟรม

สำหรับคำสั่ง R ในการสร้างเดต้าเฟรมตามรูปที่ 2-14 สามารถทำได้ดังนี้

```
> n<-c("John", "Ken", "Mary", "Fred", "Jane")
> g<-factor(x=c("Male", "Male", "Female", "Male", "Female"),
+ levels=c("Male", "Female"))
> a<-c(25, 32, 30, 25, 24)
> customer<-data.frame(Name=n, Gender=g, Age=a)
> customer
  Name Gender Age
1 John   Male  25
2 Ken    Male  32
3 Mary Female  30
4 Fred   Male  25
5 Jane  Female  24
```

เนื่องจากเดต้าเฟรมเป็นตารางสองมิติเช่นเดียวกับเมทริกซ์ ดังนั้น ผู้ใช้สามารถเข้าถึงข้อมูลในเดต้าเฟรมได้เช่นเดียวกับเมทริกซ์ นอกจากนี้ หากผู้ใช้ต้องการเข้าถึงข้อมูลในบางหลักสามารถใช้คำสั่ง ชื่อเดต้าเฟรม\$ชื่อหลัก ดังตัวอย่างด้านล่าง

```
> customer[3:4, ]
  Name Gender Age
3 Mary Female  30
4 Fred   Male  25
> customer$Age
[1] 25 32 30 25 24
> customer[customer$Gender=="Female", ]
  Name Gender Age
3 Mary Female  30
5 Jane Female  24
```

เดต้าเฟรมถือเป็นประเภทข้อมูลที่พบบ่อยที่สุดในการวิเคราะห์ข้อมูล นอกจาก R จะมีฟังก์ชันในการวิเคราะห์ข้อมูลต่าง ๆ แล้ว R ยังมีข้อมูล ซึ่งผู้ใช้สามารถเรียกขึ้นมาได้ด้วยโดยใช้คำสั่ง `data` ในการโหลดข้อมูลเข้ามาในพื้นที่ทำงานปัจจุบัน หรือจะเรียกชื่อชุดข้อมูลได้เช่นกัน ดังตัวอย่างด้านล่าง

```
> data(trees) # load data frame 'trees' into the working space
```