STATISTICAL METHODS FOR SCIENCE

วิธีเชิงสถิติสำหรับวิทยาศาสตร์

คณาจารย์สาขาวิชาสถิติ
ภาควิชาคณิตศาสตร์ มหาวิทยาลัยบูรพา

BUU BURAPHA UNIVERSITY WISDOM OF THE EAST | FSCI Faculty of Science Science Innovation for Better Sustainable Society

วิธีเชิงสถิติสำหรับวิทยาศาสตร์
STATISTICAL METHODS FOR SCIENCE

คณาจารย์สาขาวิชาสถิติ
ภาควิชาคณิตศาสตร์ มหาวิทยาลัยบูรพา

วิธีเชิงสถิติสำหรับวิทยาศาสตร์

ISBN (e-book): 978-974-384-621-2

จัดทำโดย ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยบูรพา

ราคา 150 บาท

คำนำ

หนังสือเล่มนี้จัดทำขึ้นเพื่อให้นิสิตระดับปริญญาตรี คณะวิทยาศาสตร์ มหาวิทยาลัยบูรพา ใช้ประกอบการเรียนการสอนรายวิชาวิธีเชิงสถิติสำหรับวิทยาศาสตร์ ซึ่งคณะผู้เขียนได้ประมวลเอาความรู้และประสบการณ์ในการสอนวิชาดังกล่าวเป็นเวลาหลายปีมาเรียบเรียงเป็นภาษาที่เข้าใจง่าย โดยเนื้อหาในเล่มมีรายละเอียดครอบคลุมเกี่ยวกับวิธีการวิเคราะห์เชิงสถิติที่สำคัญ โดยแบ่งออกเป็น 3 ส่วน ได้แก่ การวิเคราะห์สหสัมพันธ์และการวิเคราะห์การถดถอย การวิเคราะห์ความแปรปรวน และสถิติศาสตร์ไม่อิงพารามิเตอร์ อีกทั้งยังมีวิธีการวิเคราะห์โดยใช้โปรแกรม R commander ให้ทุกหัวข้อเพื่อให้นิสิตสามารถฝึกการใช้โปรแกรมทางสถิติมาช่วยในการวิเคราะห์ข้อมูลได้อย่างถูกต้อง และเกิดความเข้าใจในบทเรียนอย่างถ่องแท้

ขอขอบคุณทุกท่านที่มีส่วนร่วมในการปรับปรุงแก้ไขหนังสือเล่มนี้ให้สำเร็จด้วยดี ขอบคุณภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยบูรพาที่ให้การสนับสนุน และหวังเป็นอย่างยิ่งว่าหนังสือเล่มนี้จะเป็นประโยชน์กับนิสิต นักศึกษา นักวิจัย และผู้สนใจวิธีเชิงสถิติ และสามารถนำความรู้เกี่ยวกับวิธีการเชิงสถิติไปประยุกต์ใช้ได้อย่างถูกต้องต่อไป

คณาจารย์สาขาวิชาสถิติ

มิถุนายน 2567

สารบัญ

คำนำ III

I การวิเคราะห์การถดถอย 1

1 การวิเคราะห์สหสัมพันธ์และการถดถอยเชิงเส้นอย่างง่าย 2

1.1 การวิเคราะห์สหสัมพันธ์ 2

1.1.1 คุณสมบัติของสัมประสิทธิ์สหสัมพันธ์ 4

1.1.2 การอนุมานเกี่ยวกับสหสัมพันธ์ประชากร 5

1.1.3 การใช้โปรแกรม R commander 8

1.1.4 การวิเคราะห์สหสัมพันธ์ด้วยโปรแกรม R commander 12

1.2 การถดถอยเชิงเส้นอย่างง่าย 17

1.2.1 ลักษณะของข้อมูลที่ใช้ในการวิเคราะห์การถดถอย 17

1.2.2 ตัวแบบการถดถอยเชิงเส้นตรงอย่างง่าย 17

1.2.3 การประมาณค่าสัมประสิทธิ์การถดถอย β_0 และ β_1 18

1.2.4 การประมาณค่า σ^2 23

1.2.5 การทดสอบสมมุติฐาน 24

1.2.6 การวิเคราะห์ความแปรปรวน 26

1.2.7 สัมประสิทธิ์การกำหนด 28

1.2.8 การประมาณค่าเฉลี่ยของ y 29

1.2.9 การวิเคราะห์การถดถอยด้วยโปรแกรม R commander 31

2 การถดถอยเชิงเส้นพหุคูณ 33

2.1 ตัวแบบการถดถอยเชิงเส้นพหุคูณ 33

2.2 การประมาณค่าสัมประสิทธิ์การถดถอย 35
2.3 ผลบวกกำลังสอง 36
2.4 ค่าสถิติ s^2, R^2 และ F 36
2.5 ขั้นตอนในการทดสอบสมมุติฐาน 38
2.6 การวิเคราะห์การถดถอยพหุคูณด้วยโปรแกรม R commander 43
2.6.1 การสร้างไฟล์ข้อมูล 43
2.6.2 การสร้างแผนภาพการกระจาย 43
2.6.3 การวิเคราะห์การถดถอยพหุคูณ 46
2.7 การถดถอยพหุคูณเมื่อตัวแปรอิสระเป็นตัวแปรเชิงคุณภาพ 48
2.7.1 ตัวแปรหุ่น 48
2.7.2 การวิเคราะห์ข้อมูล 50
2.7.3 การวิเคราะห์ด้วยโปรแกรม R commander 55

3 การตรวจสอบข้อสมมุติเบื้องต้น 59
3.1 ข้อสมมุติเบื้องต้น 59
3.2 แผนภาพส่วนเหลือ 59
3.3 การตรวจสอบการแจกแจงปรกติของค่าคลาดเคลื่อน 60
3.3.1 ตรวจสอบด้วยแผนภาพส่วนเหลือ 60
3.3.2 ตรวจสอบด้วยการทดสอบ Anderson-Darling 61
3.4 การตรวจสอบความแปรปรวนคงที่ของค่าคลาดเคลื่อน 64
3.4.1 การตรวจสอบด้วยแผนภาพส่วนเหลือ 64
3.4.2 การแก้ปัญหาการเกิด Heteroscadastic 68
3.5 การตรวจสอบความเป็นอิสระกันของค่าคลาดเคลื่อน 70
3.5.1 ตรวจสอบด้วยแผนภาพ 70
3.5.2 การทดสอบ Durbin-Watson 71
3.6 การตรวจสอบความเป็นอิสระกันของตัวแปรอิสระ 74
3.7 การใช้โปรแกรม R commander เพื่อตรวจสอบข้อสมมุติเบื้องต้น 80
3.7.1 การใช้โปรแกรม R commander ตรวจสอบการแจกแจงปรกติของค่าคลาดเคลื่อน 80
3.7.2 การใช้โปรแกรม R commander ตรวจสอบความเป็นอิสระกันของค่าคลาดเคลื่อน 83
3.7.3 การใช้โปรแกรม R commander ตรวจสอบความเป็นอิสระกันของตัวแปรอิสระ 89

4 การเลือกตัวแบบการถดถอยที่ดีที่สุด 92

4.1 การเลือกแบบไปข้างหน้า . . . 92
4.2 การกำจัดแบบถอยหลัง . . . 93
4.3 การถดถอยทีละขั้น . . . 93
4.4 การวิเคราะห์ด้วยโปรแกรม R commander . . . 98

II การวิเคราะห์ความแปรปรวน 103

5 หลักการพื้นฐานของการวางแผนการทดลอง 104

5.1 คำจำกัดความ . . . 105
5.1.1 วิธีการวางแผนการทดลอง . . . 105
5.1.2 หน่วยทดลอง . . . 105
5.1.3 ทรีตเมนต์ . . . 105
5.1.4 การทำซ้ำ . . . 105
5.1.5 การสุ่ม . . . 106
5.1.6 ความคลาดเคลื่อนจากการทดลอง . . . 107
5.2 แนวทางในการวางแผนการทดลอง . . . 108
5.3 การใช้วิธีการทางสถิติในการทำการทดลอง . . . 109

6 แผนแบบสุ่มสมบูรณ์และการวิเคราะห์ 111

6.1 แผนแบบสุ่มสมบูรณ์ . . . 111
6.2 การวิเคราะห์ข้อมูล . . . 114
6.3 การวิเคราะห์ด้วยโปรแกรม R commander สำหรับ CRD . . . 118
6.4 การเปรียบเทียบค่าเฉลี่ยทรีตเมนต์ . . . 121
6.4.1 การทดสอบความแตกต่างนัยสำคัญน้อยสุด . . . 121
6.4.2 การทดสอบความแตกต่างนัยสำคัญตรงของทูคีย์ . . . 122
6.4.3 การทดสอบพิสัยพหุคูณของดันแคน . . . 122
6.4.4 การเปรียบเทียบพหุคูณด้วยโปรแกรม R commander . . . 127
6.5 การตรวจสอบข้อสมมุติเบื้องต้น . . . 131
6.5.1 การตรวจสอบการแจกแจงปรกติ . . . 131
6.5.2 การตรวจสอบความเท่ากันของความแปรปรวน . . . 131
6.5.3 แนวทางการแก้ไขเมื่อข้อสมมุติไม่เป็นไปตามที่กำหนด . . . 132
6.5.4 การตรวจสอบข้อสมมุติเบื้องต้นด้วยโปรแกรม R commander . . . 135

7 แผนแบบบล็อคสมบูรณ์เชิงสุ่ม **140**

7.1 แผนแบบบล็อคสมบูรณ์เชิงสุ่ม 140

7.2 การสุ่ม 141

7.3 การวิเคราะห์ข้อมูล 143

7.4 การวิเคราะห์ด้วยโปรแกรม R commander 152

8 แผนแบบจัตุรัสละติน **161**

8.1 แผนแบบจัตุรัสละติน 161

8.2 วิธีการสุ่ม 163

8.3 การวิเคราะห์ข้อมูล 165

8.4 การวิเคราะห์ด้วยโปรแกรม R commander 173

9 การทดลองแฟกทอเรียล **179**

9.1 การทดลองแฟกทอเรียล 179

9.2 การทดลองแฟกทอเรียล 2 ปัจจัย ใน CRD 181

9.3 การทดลองแฟกทอเรียลใน RCBD 197

10 การวิเคราะห์ความแปรปรวนร่วม **212**

III สถิติศาสตร์ไม่อิงพารามิเตอร์ **223**

11 บทนำ **224**

12 การทดสอบสมมุติฐานกรณีตัวอย่างหนึ่งกลุ่ม **227**

13 การทดสอบสมมุติฐานกรณีตัวอย่าง 2 กลุ่ม ที่เป็นอิสระกัน **233**

14 การทดสอบสมมุติฐานกรณีตัวอย่าง 2 กลุ่ม ไม่เป็นอิสระกัน **239**

15 การทดสอบสมมุติฐานกรณีตัวอย่าง k กลุ่ม เป็นอิสระกัน **244**

16 การทดสอบสมมุติฐานกรณีตัวอย่าง k กลุ่มที่มีความสัมพันธ์กัน **253**

17 การทดสอบเกี่ยวกับความสัมพันธ์ **262**

บรรณานุกรม 268

ส่วนที่ I

การวิเคราะห์การถดถอย

บทที่ 1

การวิเคราะห์สหสัมพันธ์และการถดถอยเชิงเส้นอย่างง่าย

การวิเคราะห์สหสัมพันธ์และการถดถอย (Correlation and regression analysis) เป็นวิธีเชิงสถิติที่สำคัญมาก โดยมักจะถูกนำมาใช้ในการศึกษาถึงความสัมพันธ์ระหว่างตัวแปรตั้งแต่สองตัวแปรขึ้นไป เช่น ผู้วิจัยต้องการศึกษาความสัมพันธ์ระหว่างปริมาณการใช้ไฟฟ้าและอุณหภูมิในแต่ละวัน หรือนักการศึกษาต้องการศึกษาถึงความสัมพันธ์ระหว่างเกรดเฉลี่ยที่นิสิตได้และคะแนนสอบเข้าของนิสิต ซึ่งในการศึกษาเหล่านี้ผู้ศึกษาต้องบันทึกข้อมูลที่เกี่ยวข้อง ได้แก่ ปริมาณการใช้ไฟฟ้า (y) และอุณหภูมิในแต่ละวัน (x) สำหรับตัวอย่างแรก และสำหรับตัวอย่างที่สอง ผู้ศึกษาต้องบันทึกเกรดเฉลี่ยของนิสิต (y) และคะแนนสอบเข้าของนิสิต (x) ดังนั้นในการศึกษาข้อมูลที่บันทึกจึงมีลักษณะเป็นคู่คือ $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ โดยในหลาย ๆ กรณีเมื่อนำข้อมูลเป็นคู่ที่ได้จากการศึกษามาสร้างแผนภาพการกระจาย (Scatter plot) เพื่อพิจารณาถึงความสัมพันธ์ระหว่างตัวแปรเหล่านี้ มักจะได้ลักษณะของแผนภาพที่จุดข้อมูลในแผนภาพมีการกระจายใกล้เคียงกับเส้นตรง ในกรณีเหล่านี้ผู้ศึกษาสามารถนำข้อมูลที่ได้มาคำนวณหาสมการทางคณิตศาสตร์สำหรับเส้นตรงนี้ และสมการเชิงเส้นตรงที่ได้จะมีประโยชน์มาก เช่น จากตัวอย่างในข้างต้นผู้ศึกษาสามารถนำสมการที่ได้ไปทำนายปริมาณการใช้ไฟฟ้าในอนาคตเมื่อกำหนดอุณหภูมิอากาศ หรือในอีกตัวอย่างหนึ่งผู้ศึกษาสามารถนำสมการที่ได้ไปประมาณเกรดเฉลี่ยของนิสิตเมื่อทราบคะแนนสอบเข้า

ในบทนี้ได้กล่าวถึงการวิเคราะห์สหสัมพันธ์และการถดถอยเชิงเส้นอย่างง่าย (Simple linear correlation and regression analysis) ซึ่งเป็นวิธีเชิงสถิติที่ใช้ในการวิเคาะห์ความสัมพันธ์ของตัวแปรสองตัวแปร เพื่อหาว่าตัวแปรคู่นี้มีความสัมพันธ์กันมากหรือน้อยเพียงใด และทิศทางของความสัมพันธ์เป็นไปในทิศทางใด และยังสามารถสร้างสมการเชิงเส้นที่แสดงความสัมพันธ์ระหว่างตัวแปรทั้งคู่ เพื่อใช้ในการทำนายค่าของตัวแปรที่สนใจเมื่อกำหนดค่าของอีกตัวแปรหนึ่งได้

1.1 การวิเคราะห์สหสัมพันธ์

การศึกษาเกี่ยวกับความสัมพันธ์ระหว่างตัวแปรในทางสถิติมีมานานมากแล้ว โดยเริ่มจากการศึกษาถึงความสัมพันธ์ระหว่างลักษณะทางกายภาพของประชากรมนุษย์ ซึ่งในการศึกษานี้นักสถิติได้สร้างค่าสถิติขึ้นมาค่าหนึ่งเรียกว่า**สัมประสิทธิ์สหสัมพันธ์ (Correlation coefficient)** ซึ่งเป็นค่าสถิติที่ใช้ในการอธิบายว่าลักษณะทาง

กายภาพของมนุษย์มีความสัมพันธ์กันมากหรือน้อยเพียงใด ต่อมาได้มีการนำค่าสัมประสิทธิ์สหสัมพันธ์มาใช้อย่างแพร่หลายในแขนงวิชาต่าง ๆ เพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรที่สนใจ เช่น การศึกษาความสัมพันธ์ระหว่างน้ำหนักและส่วนสูงของผู้ชายในประเทศไทย และความสัมพันธ์ระหว่างคะแนนสอบวิชาคณิตศาสตร์และวิชาสถิติ เป็นต้น

สัมประสิทธิ์สหสัมพันธ์ (Correlation coefficient) คือค่าวัดทางสถิติที่ใช้ในการวัดขนาดของความสัมพันธ์เชิงเส้นระหว่างตัวแปรสองตัวแปร โดยมักใช้สัญลักษณ์คือ r

ให้ $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ เป็นข้อมูล n คู่ ที่ได้จากการศึกษา ดังนั้น

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}}$$

เมื่อ

$$\begin{aligned} S_{xy} &= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\overline{xy} \\ S_{xx} &= \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \\ S_{yy} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \end{aligned}$$

และ

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

โดยหลักการแล้วเราสามารถคำนวณหาค่าสัมประสิทธิ์สหสัมพันธ์สำหรับข้อมูลที่ประกอบด้วยตัวแปรสองตัวแปรชุดใด ๆ ก็ได้ อย่างไรก็ตาม ในหลาย ๆ กรณีข้อมูลเหล่านี้มักเป็นข้อมูลตัวอย่างที่สุ่มมาจากประชากรเป้าหมาย ซึ่งในกรณีนี้เรียกสัมประสิทธิ์สหสัมพันธ์นี้ว่า **สหสัมพันธ์ตัวอย่าง (Sample correlation)** ซึ่งเป็นตัว

ประมาณค่าของสหสัมพันธ์ประชากร (Population correlation, ρ)

1.1.1 คุณสมบัติของสัมประสิทธิ์สหสัมพันธ์

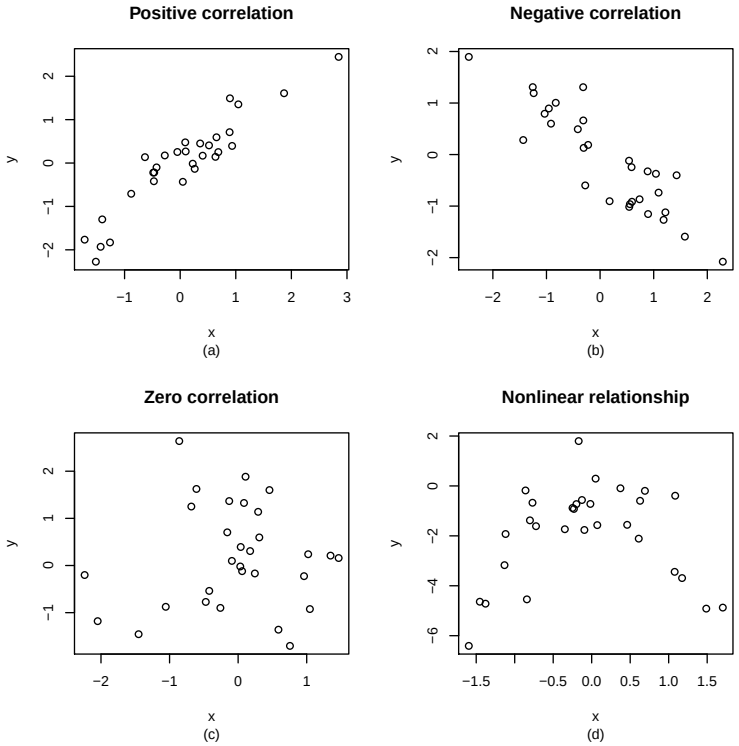
1. สัมประสิทธิ์สหสัมพันธ์เป็นค่าวัดที่ไม่มีหน่วย และเมื่อไม่คำนึงถึงเครื่องหมายค่านี้จะมีค่าระหว่าง 0 และ 1

2. สัมประสิทธิ์สหสัมพันธ์มีค่าเป็นได้ทั้งค่าบวกและค่าลบ โดยเครื่องหมายของสัมประสิทธิ์สหสัมพันธ์จะแสดงถึงทิศทางของความสัมพันธ์ระหว่างตัวแปรทั้งสองตัวแปร หากเครื่องหมายเป็นบวกแสดงว่าตัวแปรทั้งสองมีความสัมพันธ์ในทิศทางเดียวกัน นั่นคือหากตัวแปรหนึ่งมีค่าเพิ่มขึ้นอีกตัวแปรหนึ่งจะมีค่าเพิ่มขึ้นด้วย หากเครื่องหมายเป็นลบแสดงว่าตัวแปรทั้งสองมีความสัมพันธ์ในทิศทางตรงกันข้าม นั่นคือหากตัวแปรหนึ่งมีค่าเพิ่มขึ้นอีกตัวแปรหนึ่งจะมีค่าลดลง

3. ถ้าสัมประสิทธิ์สหสัมพันธ์มีค่าเป็น 0 หรือเข้าใกล้ 0 แสดงว่าตัวแปรทั้งสองไม่มีความสัมพันธ์กัน หรือมีความสัมพันธ์กันน้อย แต่ถ้ามีค่าเท่ากับ ± 1 หรือมีค่าเข้าใกล้ ± 1 แสดงว่าตัวแปรทั้งสองมีความสัมพันธ์กันอย่างสมบูรณ์หรือมีความสัมพันธ์กันมาก

ลักษณะของความสัมพันธ์ระหว่างตัวแปรสองตัวแปรที่กล่าวถึง เป็นลักษณะของความสัมพันธ์เชิงเส้นตรง ดังนั้น เมื่อสัมประสิทธิ์สหสัมพันธ์มีค่าเท่ากับ 0 หรือมีค่าเข้าใกล้ 0 ไม่ได้หมายความว่าตัวแปรสองตัวแปรนั้นไม่มีความสัมพันธ์กัน แต่ความสัมพันธ์นั้นอาจอยู่ในรูปแบบอื่น ๆ ที่ไม่ใช่เส้นตรงก็ได้ (เช่น ความสัมพันธ์เชิงเส้นโค้ง เป็นต้น)

รูปที่ 1.1 แสดงตัวอย่างของสหสัมพันธ์ในรูปแบบต่าง ๆ โดยรูปในแถวบนทางด้านซ้ายเป็นลักษณะของสหสัมพันธ์ในทางบวก (Positive correlation) ส่วนทางด้านขวาเป็นสหสัมพันธ์ในทางลบ (Negative correlation) สำหรับแถวล่างรูปทางด้านซ้ายแสดงลักษณะที่ตัวแปรทั้งสองไม่มีความสัมพันธ์กัน ส่วนรูปทางด้านขวาเป็นลักษณะของสหสัมพันธ์ที่ไม่ใช่เส้นตรง



รูปที่ 1.1: ตัวอย่างของสหสัมพันธ์ในลักษณะต่าง ๆ

1.1.2 การอนุมานเกี่ยวกับสหสัมพันธ์ประชากร

เมื่อค่าสังเกต (x_i, y_i) เป็นข้อมูลตัวอย่างที่สุ่มจากประชากรและเป็นค่าของตัวแปรสุ่ม X และ Y แล้ว สัมประสิทธิ์สหสัมพันธ์ตัวอย่าง r เป็นค่าประมาณของสัมประสิทธิ์สหสัมพันธ์ประชากร ρ

หากตัวแปรสุ่ม X และ Y มีการแจกแจงร่วมที่เป็นการแจกแจงปรกติแล้ว สหสัมพันธ์ตัวอย่าง r สามารถนำมาใช้ในการอ้างอิงเกี่ยวกับสหสัมพันธ์ประชากรได้ ซึ่งการอ้างอิงที่นิยมใช้ในการศึกษาต่าง ๆ คือ การทดสอบสมมุติฐานเกี่ยวกับค่า ρ เพื่อจะได้ทราบว่าตัวแปรสุ่ม X และ Y มีความสัมพันธ์กันในเชิงของประชากรหรือไม่ ขั้นตอนของการทดสอบสมมุติฐานมีดังนี้

การทดสอบสมมุติฐานเกี่ยวกับค่าสัมประสิทธิ์สหสัมพันธ์ประชากร

1. $H_0 : \rho = 0$ (ตัวแปรสองตัวไม่มีความสัมพันธ์กัน) $H_a : \rho \neq 0$ (ตัวแปรสองตัวมีความสัมพันธ์กัน)
2. กำหนดระดับนัยสำคัญ α
3. ค่าวิกฤต: $t_{\alpha/2,n-2}$
4. ค่าสถิติทดสอบ:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

5. ปฏิเสธ H_0 เมื่อ $|T| > t_{\alpha/2,n-2}$ หรือ p-value $< \alpha$

   ยอมรับ H_0 เมื่อ เมื่อ $|T| \leq t_{\alpha/2,n-2}$ หรือ p-value $\geq \alpha$

ตัวอย่างที่ 1.1 จากการศึกษาทางด้านการแพทย์ทราบว่า ในการสูบบุหรี่เมื่อนิโคตินถูกดูดซึมเข้าไปในร่างกาย จะมีการสร้างโคติไนน์ขึ้นในร่างกาย ดังนั้น ปริมาณโคติไนน์ในร่างกายจึงเป็นดัชนีชี้วัดที่ดีว่าคน ๆ หนึ่งมีการสูบบุหรี่มากน้อยเพียงใด ในการศึกษาถึงความสัมพันธ์ระหว่างปริมาณโคติไนน์และจำนวนบุหรี่ที่สูบ ผู้ศึกษาได้สุ่มคนสูบบุหรี่มาจำนวน 12 คน และบันทึกข้อมูลเกี่ยวกับจำนวนบุหรี่เขาสูบต่อวัน, x และปริมาณโคติไนน์ที่ตรวจพบในร่างกาย, y ดังตารางต่อไปนี้

คน	x	y
1	18	283.00
2	4	75.60
3	15	174.00
4	10	209.00
5	1	9.51
6	20	350.00
7	8	105.00
8	7	43.40
9	10	25.10
10	10	150.00
11	20	344.00

แผนภาพการกระจายและผลการวิเคราะห์สหสัมพันธ์ด้วย R commander แสดงดังรูปที่ 1.2 และ 1.3

1. จงหาค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างจำนวนบุหรี่ที่สูบและปริมาณโคติไนน์ในร่างกายของคนกลุ่มนี้

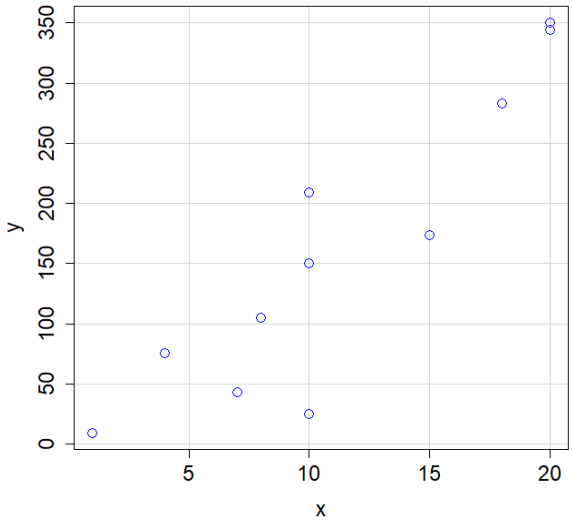
   จากรูปที่ 1.3 ได้ค่าสัมประสิทธิ์สหสัมพันธ์ $r = 0.9059$ หมายความว่า สำหรับข้อมูลชุดนี้ จำนวนบุหรี่ที่สูบและปริมาณโคติไนน์ในร่างกายมีความสัมพันธ์กันขนาด 0.9059 ซึ่งถือว่าสูงมาก และมีทิศทางความสัมพันธ์ในทิศทางบวก นั่นคือเมื่อจำนวนบุหรี่ที่สูบมากขึ้นปริมาณโคติไนน์ในร่างกายก็มากขึ้นด้วย

2. จากข้อมูลที่มีอยู่นี้ จงทดสอบสมมุติฐานปริมาณโคติไนน์ในร่างกายเป็นดัชนีชี้วัดที่ดีของปริมาณการสูบบุหรี่ (ปริมาณโคติไนน์และจำนวนบุหรี่ที่สูบมีความสัมพันธ์กันในเชิงประชากรหรือไม่) ที่ระดับนัยสำคัญ 0.05

   - $H_0 : \rho = 0$ (ปริมาณโคติไนน์และจำนวนบุหรี่ที่สูบไม่มีความสัมพันธ์กัน)

     $H_a : \rho \neq 0$ (ปริมาณโคติไนน์และจำนวนบุหรี่ที่สูบมีความสัมพันธ์กัน)

- $\alpha = 0.05$
- จากรูปที่ 1.3 ได้ค่า p-value = 0.0001 ซึ่งมีค่าน้อยกว่า $\alpha = 0.05$
- ปฏิเสธ H_0 สรุปได้ว่า ที่ระดับนัยสำคัญ 0.05 ปริมาณโคติไนน์และจำนวนบุหรี่ที่สูบมีความสัมพันธ์กัน



รูปที่ 1.2: แผนภาพการกระจายระหว่างจำนวนบุหรี่ที่สูบและปริมาณโคติไนน์ในร่างกาย

```
 Pearson's product-moment correlation

data:  x and y
t = 6.4184, df = 9, p-value = 0.0001226
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.6703438 0.9756120
sample estimates:
cor
0.9059257
```

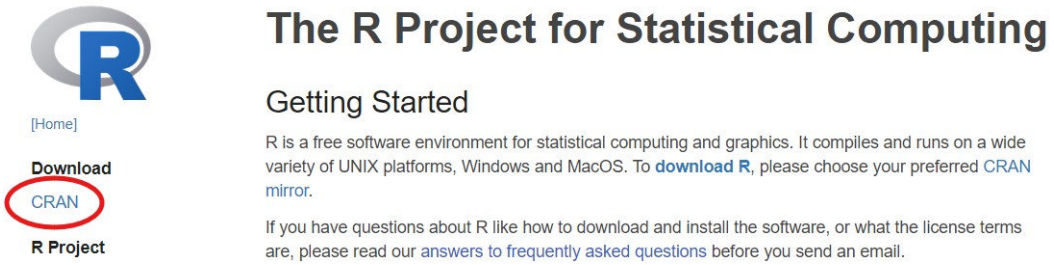
รูปที่ 1.3: ผลการวิเคราะห์สหสัมพันธ์ด้วย R commander ของตัวอย่างที่ 1.1

1.1.3 การใช้โปรแกรม R commander

การติดตั้งโปรแกรม R

โปรแกรม R สามารถดาวน์โหลดได้ฟรีจาก https://www.r-project.org

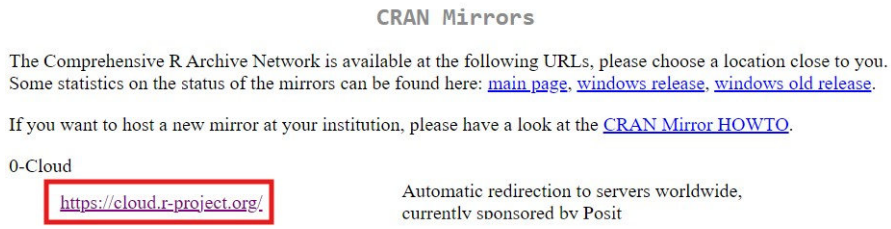
1. เมื่อเข้าในเว็ปไซด์จะได้หน้าต่างดังรูปที่ 1.4



รูปที่ 1.4

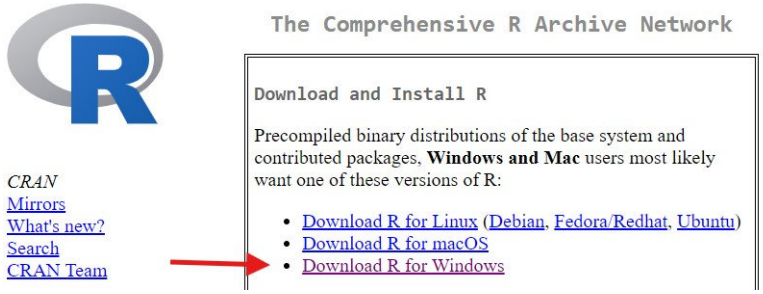
2. เมื่อเข้าไปตามลิงค์ให้เลือกที่ CRAN ในแถบหัวข้อ Download ทางด้านซ้ายมือ

3. จากนั้นเลือก Mirrors ที่ต้องการ ในที่นี้เลือก “0-Cloud” (บนสุด) ตามรูปที่ 1.5

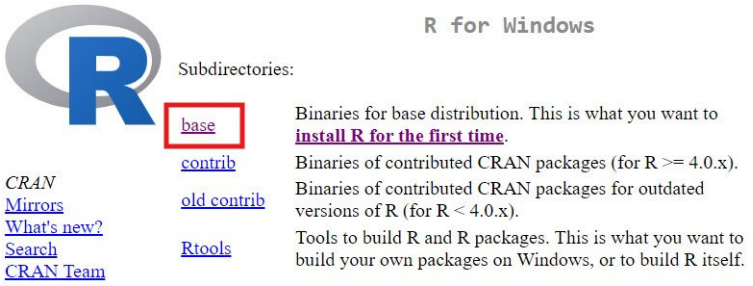


รูปที่ 1.5

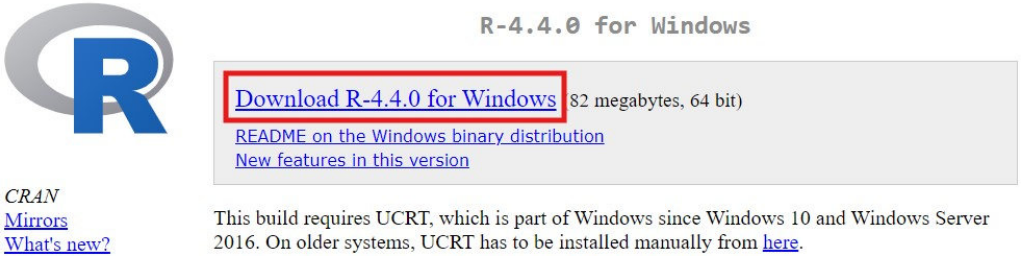
4. เลือก Download R for Windows → base → Download R 4.4.0 for Windows ตามรูปที่ 1.6 - 1.8



รูปที่ 1.6

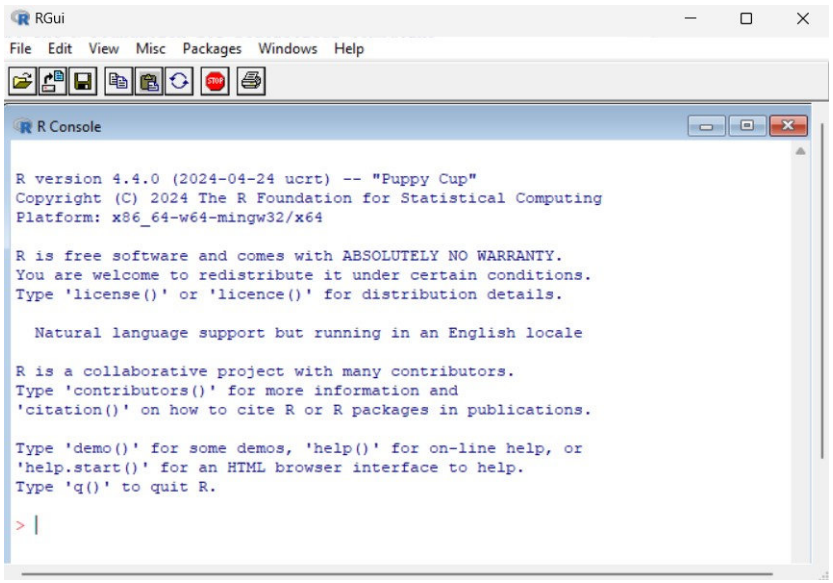


รูปที่ 1.7



รูปที่ 1.8

5. เมื่อติดตั้งเสร็จและเปิดโปรแกรมจะได้หน้าจอดังรูปที่ 1.9

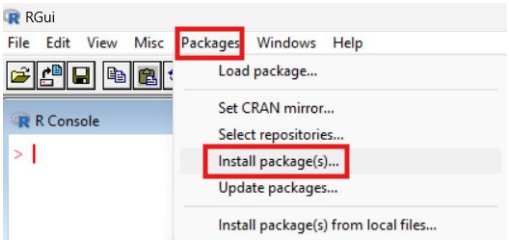


รูปที่ 1.9

การติดตั้ง R commander

1. เปิดโปรแกรม R

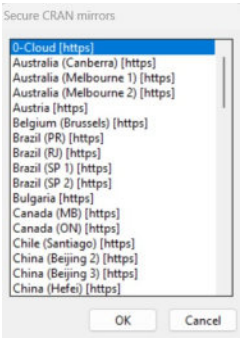
2. เลือกเมนู Packages → Install package(s)... จะได้หน้าจอดังรูปที่ 1.10



รูปที่ 1.10

3. เลือก Mirrors ที่ต้องการ ในที่นี้เลือก “0-Cloud” (บนสุด) ดังรูปที่ 1.11 แล้วคลิก OK

4. เลือกแพคเกจชื่อ “Rcmdr” จากนั้นคลิก OK

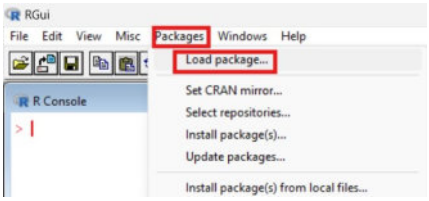


รูปที่ 1.11

การเรียกใช้ R commander

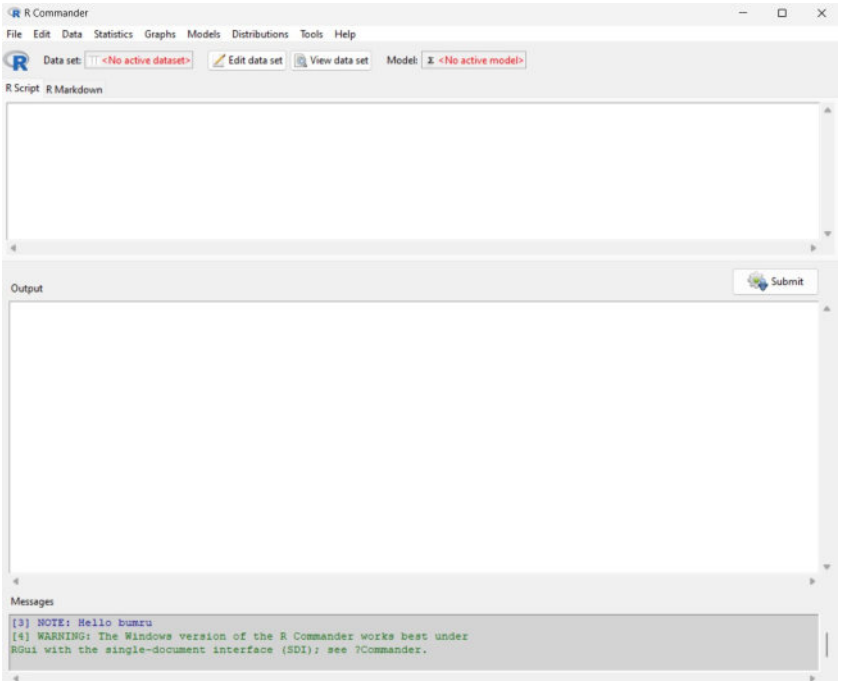
1. เปิดโปรแกรม R

2. เลือกเมนู Packages ⟶ Load package... ดังรูปที่ 1.12



รูปที่ 1.12

3. เลือก Rcmdr จะได้หน้าต่าง R Commander ดังรูปที่ 1.13


R Commander
File Edit Data Statistics Graphs Models Distributions Tools Help
Data set: <No active dataset>
Edit data set
View data set
Model: <No active model>
R Script R Markdown
Output
Submit
Messages
[3] NOTE: Hello bumru
[4] WARNING: The Windows version of the R Commander works best under
RGui with the single-document interface (SDI); see ?Commander.

รูปที่ 1.13

1.1.4 การวิเคราะห์สหสัมพันธ์ด้วยโปรแกรม R commander

การสร้างแผนภาพการกระจาย

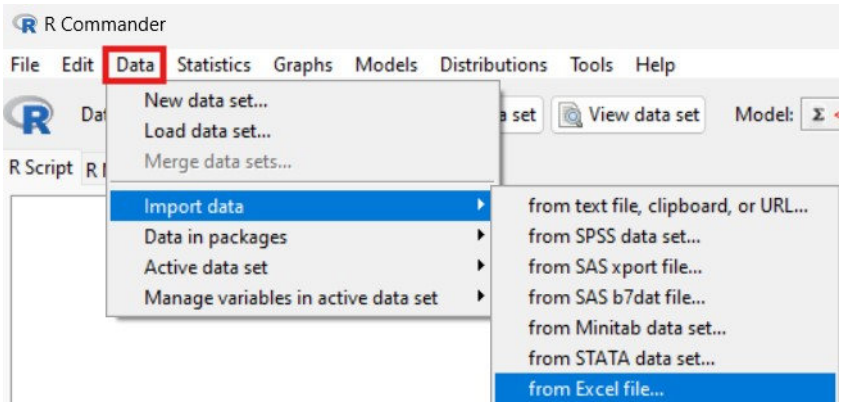
1. สร้างข้อมูลในโปรแกรม MS excel โดยกำหนดตัวแปรดังรูปที่ 1.14 และตั้งชื่อเป็น ”smoke.xlsx”

   - x แทน จำนวนบุหรี่ที่สูบ
   - y แทน ปริมาณโคติไนน์

	A	B
1	**x**	**y**
2	18	283.00
3	4	75.60
4	15	174.00
5	10	209.00
6	1	9.51
7	20	350.00
8	8	105.00
9	7	43.40
10	10	25.10
11	10	150.00
12	20	344.00

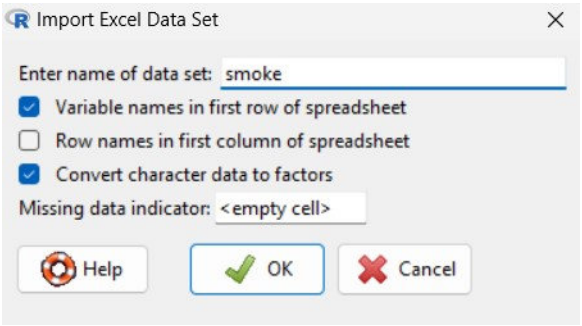
รูปที่ 1.14: การสร้างข้อมูลในโปรแกรม MS excel ของตัวอย่างที่ 1.1

2. นำข้อมูลเข้าโปรแกรม R commander โดยเลือกเมนู Data → Import data → from Excel file... ดังรูปที่ 1.15



รูปที่ 1.15

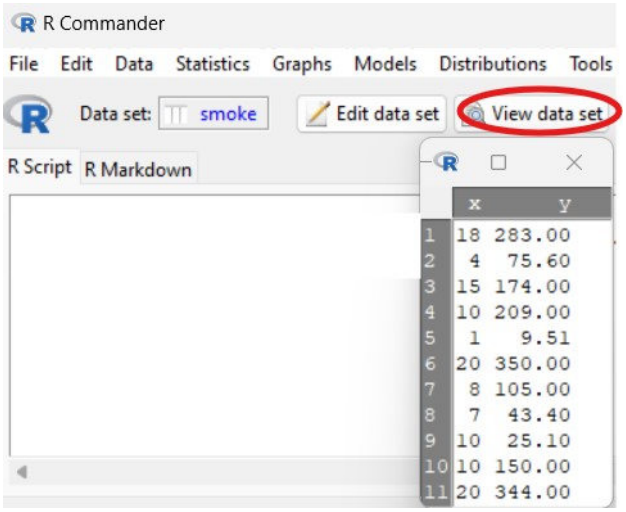
3. จะได้หน้าต่างรูปที่ 1.16 ให้เปลี่ยนชื่อชุดข้อมูลในช่อง Enter name of data set: เป็น ”smoke” แล้วคลิก OK



รูปที่ 1.16

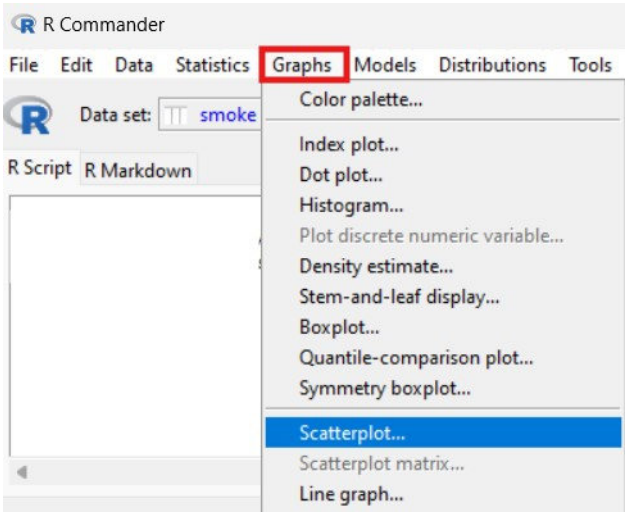
4. เลือกไฟล์ชื่อ ”smoke.xlsx” ที่จัดเก็บไว้ จากนั้นคลิก Open

5. เลือก View data set ดังรูปที่ 1.17 เพื่อตรวจดูข้อมูลที่นำเข้าสำเร็จ



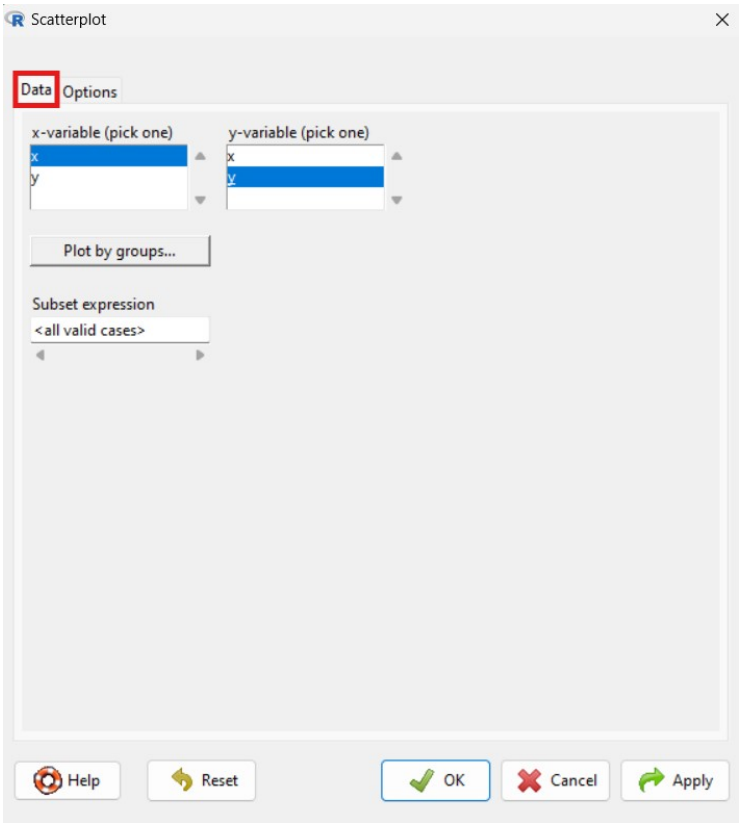
รูปที่ 1.17

6. คลิก Graphs → Scatterplot... ตามรูปที่ 1.18

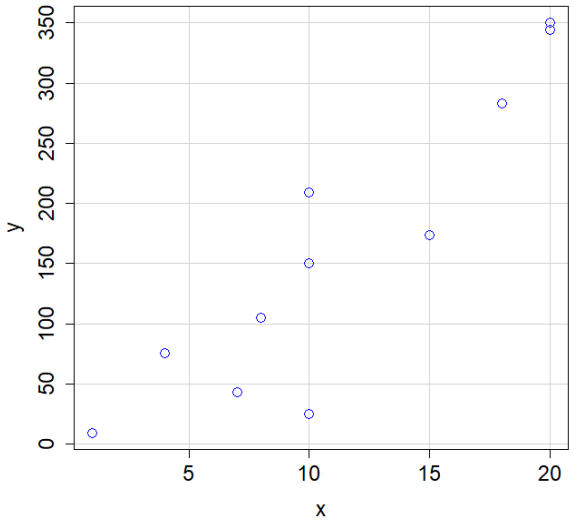


รูปที่ 1.18

7. จะได้หน้าจอดังรูปที่ 1.19 ในส่วนของ Data เลือกตัวแปรตามแกน x และ แกน y ตามรูป และ คลิก OK จะได้แผนภาพดังรูปที่ 1.20



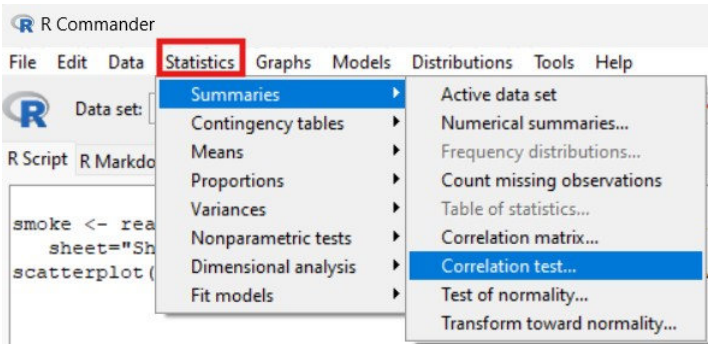
รูปที่ 1.19



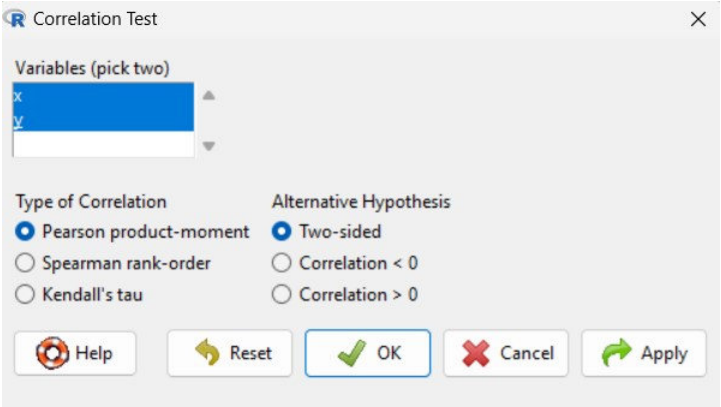
รูปที่ 1.20: แผนภาพการกระจายระหว่างจำนวนบุหรี่ที่สูบและปริมาณโคติไนน์ในร่างกาย

การวิเคราะห์สหสัมพันธ์

1. เลือกเมนู Statistics ⟶ Summaries ⟶ Correlation test... ตามรูปที่ 1.21 จะได้หน้าจอดังรูปที่ 1.22



รูปที่ 1.21



รูปที่ 1.22

2. ในส่วนของ Variables (pick two) เลือกตัวแปรทั้ง x และ y

3. ในส่วนของ Type of Correlation เลือก Pearson product-moment

4. ในส่วนของ Alternative Hypothesis เลือก Two-sided

5. คลิก OK จะได้ผล ดังรูปที่ 1.23

```
Pearson's product-moment correlation

data:  x and y
t = 6.4184, df = 9, p-value = 0.0001226
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.6703438 0.9756120
sample estimates:
cor
0.9059257
```

รูปที่ 1.23: ผลการวิเคราะห์สหสัมพันธ์ด้วย R commander ของตัวอย่างที่ 1.1

1.2 การถดถอยเชิงเส้นอย่างง่าย

การถดถอยเชิงเส้นอย่างง่ายเป็นวิธีเชิงสถิติที่ใช้ศึกษาถึงรูปแบบของความสัมพันธ์ระหว่างตัวแปรสองตัวแปร โดยวิธีนี้จะแยกตัวแปรออกเป็นสองชนิด คือ ตัวแปรตาม (Dependent variable, Response variable) และตัวแปรอิสระ (Independent variable, Explanatory variable, Predictor variable) ซึ่งตัวแปรตามนั้นใช้สัญลักษณ์ y และตัวแปรอิสระใช้สัญลักษณ์ x

ถึงแม้ว่าการวิเคราะห์สหสัมพันธ์และการถดถอยจะเป็นวิธีที่ใช้ในการศึกษาความสัมพันธ์ระหว่างตัวแปรสองตัวแปรทั้งสองวิธี แต่วิธีทั้งสองนี้มีวัตถุประสงค์และกระบวนการที่แตกต่างกัน โดยการถดถอยมีวัตถุประสงค์เพื่อสร้างสมการถดถอย (Regression equation) ซึ่งเป็นฟังก์ชันทางคณิตศาสตร์ที่แสดงความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ และนำสมการถดถอยที่ได้ไปใช้ในการในการอธิบาย ทำนาย หรือควบคุมค่าของตัวแปรตามเมื่อกำหนดค่าของตัวแปรอิสระ

1.2.1 ลักษณะของข้อมูลที่ใช้ในการวิเคราะห์การถดถอย

ข้อมูลที่ใช้ในการวิเคราะห์การถดถอยสามารถเป็นได้ทั้งข้อมูลที่ได้จากการสังเกต (Observation data) หรือข้อมูลที่ได้จากการทดลอง (Experimental data) นอกจากนั้นยังอาจเป็นข้อมูลอนุกรมเวลา (Time series data) หรือข้อมูลที่เก็บรวบรวม ณ ช่วงเวลาใดเวลาหนึ่งที่เรียกว่าข้อมูลภาคตัดขวาง (Cross-sectional data) ก็ได้ การที่จะบอกว่าข้อมูลที่ศึกษาเป็นข้อมูลประเภทใดนั้นสามารถพิจารณาได้จากค่าของตัวแปรอิสระ ถ้าค่าของตัวแปรอิสระเป็นค่าที่ไม่สามารถควบคุมได้แต่สามารถวัดค่าได้โดยไม่มีความคลาดเคลื่อนเกิดขึ้น ข้อมูลแบบนี้เป็นข้อมูลที่ได้จากการสังเกต หากค่าของตัวแปรอิสระเป็นค่าที่ถูกกำหนดไว้ก่อนที่จะมีการสังเกตค่าของตัวแปรตามข้อมูลนี้คือข้อมูลที่ได้จากการทดลอง หากข้อมูลถูกสังเกตตามลำดับเวลาเรียกข้อมูลแบบนี้ว่าข้อมูลอนุกรมเวลา และเมื่อข้อมูลถูกสังเกตมา ณ เวลาหนึ่งเรียกข้อมูลแบบนี้ว่าข้อมูลภาคตัดขวาง

โดยทั่วไปการได้ความรู้จากการวิเคราะห์การถดถอยมากหรือน้อยเพียงใดขึ้นอยู่กับตัวแบบการถดถอยที่เหมาะสม ปริมาณของข้อมูลที่รวบรวมมาได้ และค่าของตัวแปรอิสระที่นำมาใช้ ดังนั้นหากผู้ศึกษาสามารถควบคุมค่าของตัวแปรอิสระได้ก็ควรใช้การวางแผนการทดลอง ในการสังเกตค่าของตัวแปรตาม ซึ่งการใช้การวางแผนการทดลองจะทำให้ได้ความรู้จากการวิเคราะห์การถดถอยเพิ่มขึ้น

1.2.2 ตัวแบบการถดถอยเชิงเส้นตรงอย่างง่าย

ตัวแบบการถดถอยเชิงเส้นตรงอย่างง่ายเป็นตัวแบบการถดถอยที่ง่ายที่สุด ซึ่งมีรูปแบบดังนี้

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

เมื่อ

y_i คือ ค่าสังเกตที่ i ของตัวแปรตาม

x_i คือ ค่าสังเกตที่ i ของตัวแปรอิสระ

β_0 และ β_1 คือ พารามิเตอร์ ที่เรียกว่าสัมประสิทธิ์การถดถอย (Regression coefficients)

ε_i คือ ค่าคลาดเคลื่อน (Error) ของค่าสังเกตที่ i ซึ่งเป็นผลต่างระหว่างค่าที่แท้จริงของตัวแปรตาม y_i และค่าประมาณของ y_i ที่ได้จากสมการถดถอยเมื่อกำหนดค่าของตัวแปรอิสระ x_i ซึ่งเกิดจากอิทธิพลของปัจจัยอื่น ๆ ที่ไม่ใช่อิทธิพลของตัวแปรอิสระที่มีต่อตัวแปรตาม

โดยที่ตัวแปรอิสระ x เป็นตัวแปรที่ทราบค่าหรือมีค่าคงที่ และเป็นค่าที่ควบคุมได้โดยผู้ศึกษา ในขณะที่ y เป็นตัวแปรสุ่มที่ไม่ทราบค่าล่วงหน้า

ข้อสมมุติเบื้องต้น

ในการวิเคราะห์การถดถอยตัวแบบการถดถอยเชิงเส้นจะเป็นตัวแบบที่เหมาะสมเมื่อข้อสมมุติเบื้องต้น (Assumptions) ของค่าคลาดเคลื่อน (ε_i) เหล่านี้เป็นจริง

1. ε_i เป็นตัวแปรสุ่มที่มีการแจกแจงปรกติ ค่าเฉลี่ยเท่ากับศูนย์
2. ε_i มีความแปรปรวนคงที่ (Constant variance หรือ Homoscedasticity) ที่ทุกค่าของตัวแปรอิสระ โดยมีค่าเท่ากับ σ^2
3. ε_i และ ε_j เป็นอิสระกันในทุก ๆ ค่า i และ j เมื่อ $i \neq j$

จากตัวแบบการถดถอยเชิงเส้นอย่างง่ายจะเห็นได้ว่าตัวแปรตาม y และค่าคลาดเคลื่อน ε เป็นตัวแปรสุ่ม ดังนั้นค่าเฉลี่ยของ y สำหรับแต่ละค่าของ x คือ

$$\mu_{y|x} = \beta_0 + \beta_1 x$$

เนื่องจาก ε มีค่าเฉลี่ยเท่ากับ 0 และ $\beta_0 + \beta_1 x$ เป็นค่าคงที่ และเรียกสมการนี้ว่า **สมการถดถอย (Regression equation)** และเรียกแผนภาพของสมการถดถอยนี้ว่า **เส้นถดถอย (Regression line)** โดยเรียก β_0 และ β_1 ว่า**สัมประสิทธิ์การถดถอย (Regression coefficients)** เมื่อ β_0 คือระยะตัดแกน y ของประชากร หรืออีกนัยหนึ่งคือค่าเฉลี่ยของตัวแปรตาม y เมื่อ $x = 0$ และ β_1 คือความชันของเส้นถดถอยของประชากร หรืออีกนัยหนึ่งคือค่าที่เปลี่ยนไปของค่าเฉลี่ยของ y เมื่อ x มีค่าเปลี่ยนไปหนึ่งหน่วย

1.2.3 การประมาณค่าสัมประสิทธิ์การถดถอย β_0 และ β_1

เนื่องจากพารามิเตอร์ β_0 และ β_1 ในตัวแบบการถดถอยเชิงเส้นอย่างง่าย

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n$$