



บทที่ 5

การวิเคราะห์สหสัมพันธ์และการถดถอยเชิงเส้น



เนื้อหาบทที่ 5

- สัมประสิทธิ์สหสัมพันธ์
- สัมประสิทธิ์ของการตัดสินใจ
- การวิเคราะห์ถดถอย

การหาความสัมพันธ์ระหว่างข้อมูลเชิงปริมาณ 2 ชุด หรือมากกว่านั้นถ้าใช้เทคนิคการทดสอบไคสแควร์จะต้องแบ่งข้อมูลเชิงปริมาณนั้นออกเป็นกลุ่ม ๆ เสียก่อนแล้วจึงนับความถี่ ซึ่ง การทำเช่นนี้จะสูญเสียรายละเอียดของข้อมูลไป เทคนิคการวิเคราะห์สหสัมพันธ์เป็นเทคนิคการหาความสัมพันธ์ระหว่างข้อมูลเชิงปริมาณ ตั้งแต่ 2 ชุดขึ้นไป และสามารถบอกได้ว่าข้อมูลที่มี นั้นมีความสัมพันธ์กัน และถ้าสัมพันธ์มีความสัมพันธ์กันในทิศทางใด

การวิเคราะห์สหสัมพันธ์

ค่าสัมประสิทธิ์สหสัมพันธ์ เป็นตัวเลขตัวเดียวที่บอกให้ทราบว่า 2 สิ่งมีความสัมพันธ์กันหรือไม่ และมีความสัมพันธ์กันมากน้อยเพียงใด ซึ่งเป็นรากฐานของการทำนาย เช่น ถ้าเราอยากทราบว่า คนที่ได้คะแนนคนที่ได้วิชาสถิติสูง จะทำงานได้บัญชีได้ดี เราก็จะสามารถใช้คะแนนวิชาสถิติทำนายการปฏิบัติงานด้านการบัญชีในภายหลังได้ ถ้าการทำนายเป็นไปอย่างถูกต้องมาก ก็กล่าวได้ว่ามี สหสัมพันธ์ระหว่างคะแนนวิชาสถิติและความสำเร็จในการทำงานบัญชี เป็นบวก ซึ่งหาได้โดยการคำนวณหาค่าสัมประสิทธิ์ของคะแนนของคนหลายๆคน คะแนนของคนเดียว จะคำนวณหาค่าสัมประสิทธิ์สหสัมพันธ์ไม่ได้ และการที่เรารู้ค่าสหสัมพันธ์และใช้ทำนายได้ เป็นประโยชน์มากทั้งในด้านการคัดเลือกหรือการจัดคนให้เหมาะสมกับงานโดยใช้การทดสอบ ในเรื่องอื่นๆก็เช่นเดียวกัน

การวิเคราะห์สหสัมพันธ์

ถ้าให้ X เป็นตัวแปรแทน ความสูง (cm)

Y เป็นตัวแปรแทนน้ำหนัก (kg)

ลักษณะของข้อมูลที่เก็บรวบรวมมาจะต้องเป็นข้อมูลที่มาจากหน่วยตัวอย่างเดียวกัน เช่นคนเดียวกัน ดังนี้

คนที่	น้ำหนัก (Y)	ความสูง (X)
1	y_1	x_1
2	y_2	x_2
$\vdots$		
n	y_n	x_n

จากตารางจะเห็นว่าค่าของตัวแปร Y และ X ได้มาจากคน ๆ เดียวกัน ซึ่งอาจจะมีหน่วยเดียวกันหรือหน่วยต่างกันก็ได้

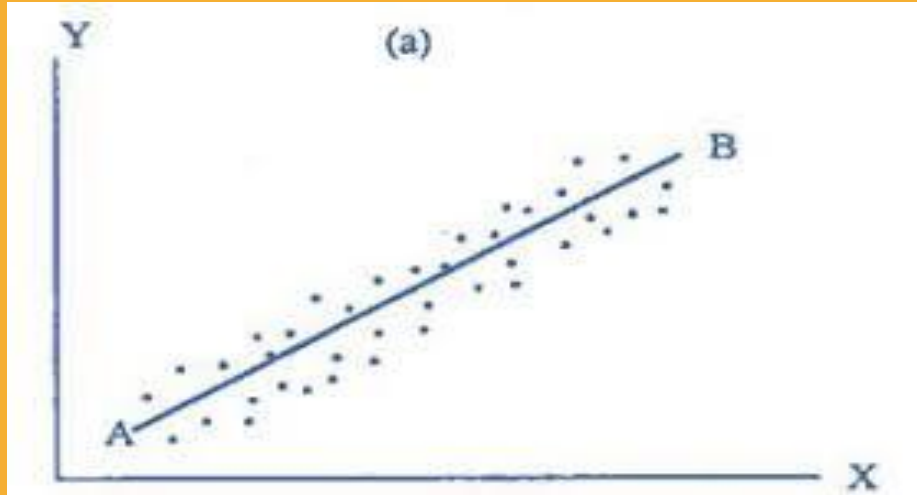
การวิเคราะห์สหสัมพันธ์

วิธีการดูลักษณะความสัมพันธ์

การดูลักษณะความสัมพันธ์ของข้อมูล หรือตัวแปรสามารถทำได้หลายวิธี

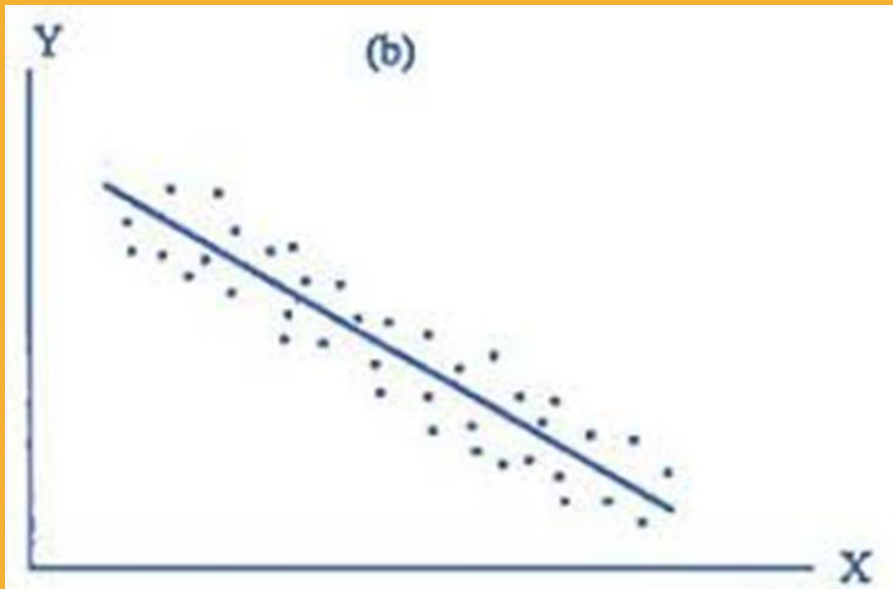
1. แผนภาพการกระจาย (scatter diagram)
2. ค่าสัมประสิทธิ์สหสัมพันธ์ (correlation coefficient)

แผนภาพการกระจาย



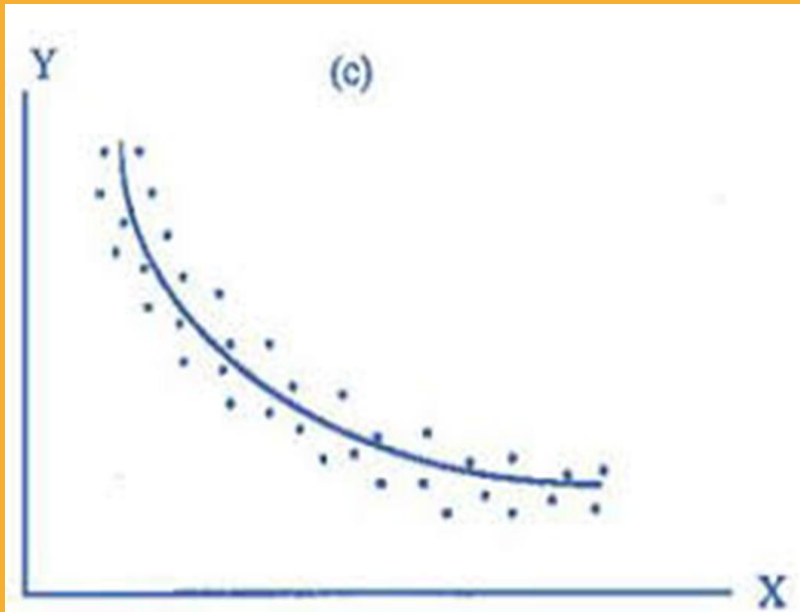
ในรูป (a) แนวโน้มของจุดชี้ขึ้นด้านขวาตามแนวเส้นตรง เมื่อ X มีค่ามาก y มีค่ามาก เมื่อ X มีค่าน้อย y มีค่าน้อย เรียกว่ามีความสัมพันธ์เชิงเส้นเชิงบวก (positive and linear correlation) หรือความสัมพันธ์แบบแปรตามกัน

แผนภาพการกระจาย



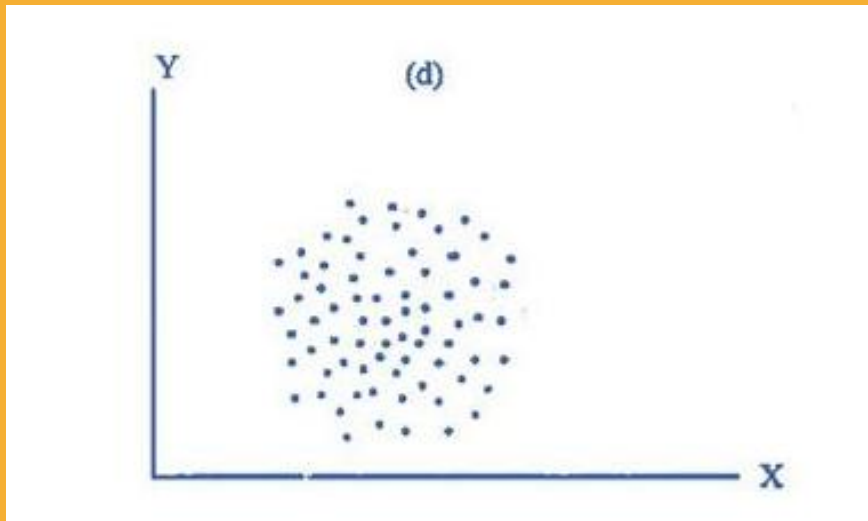
ในรูป (b) แนวโน้มของจุดซึ่งด้านขวาตามแนวเส้นตรงเมื่อ X มีค่ามาก y มีค่าน้อย เมื่อ X มีค่าน้อย y มีค่ามาก เรียกว่ามีความสัมพันธ์เชิงเส้นเชิงลบ (negative and linear correlation) หรือความสัมพันธ์แบบแปรผกผัน

แผนภาพการกระจาย



ในรูป (C) แนวโน้มของจุดชี้ลงด้านขวา
ตามแนวเส้นโค้ง เรียกว่ามีความสัมพันธ์ ไม่
เชิงเส้นเชิงลบ (negative and nonlinear
correlation)

แผนภาพการกระจาย



ในรูป (d) แนวโน้มของจุดกระจายออกไม่มีแนว
เส้นตรง เรียกว่าไม่มี ความสัมพันธ์
(no correlation)

ค่าสัมประสิทธิ์สหสัมพันธ์

สำหรับตัวสถิติที่ใช้วัดค่าสหสัมพันธ์อย่างง่าย ว่ามีความสัมพันธ์มากหรือน้อยเพียงใด คือ สัมประสิทธิ์สหสัมพันธ์ (Correlation coefficient) ซึ่งในกรณีของสหสัมพันธ์อย่างง่าย ตัวสถิตินี้เรียกว่า สัมประสิทธิ์สหสัมพันธ์อย่างง่าย (Simple Correlation Coefficient) เขียนแทนด้วย สัญลักษณ์ ρ หรือ ρ_{xy} ในกรณีที่เป็นค่าพารามิเตอร์ และ r หรือ r_{xy} ในกรณีที่เป็นค่าสถิติโดยที่ ρ และ r จะไม่มีหน่วยและมีค่าตั้งแต่ -1 ถึง 1

ค่าสัมประสิทธิ์สหสัมพันธ์

ความหมายของค่าสัมประสิทธิ์สหสัมพันธ์

1. ถ้า ρ และ r มีค่าเป็นบวก แสดงว่าตัวแปร X และ Y มีความสัมพันธ์เชิงบวก หมายความว่าถ้าตัวแปร X มีค่าเพิ่มขึ้นตัวแปร Y จะมีค่าเพิ่มขึ้นหรือถ้าตัวแปร X มีค่าลดลง ตัวแปร Y จะมีค่าลดลง
2. ถ้า ρ และ r มีค่าเป็นลบแสดงว่าตัวแปร X และ Y มีความสัมพันธ์เชิงลบ หมายความว่า ถ้าตัวแปร X มีค่าเพิ่มขึ้นตัวแปร Y จะมีค่าลดลง หรือถ้าตัวแปร X มีค่าลดลง ตัวแปร Y จะมีค่าเพิ่มขึ้น
3. ถ้า ρ และ r มีค่าเท่ากับ 1 แสดงว่าตัวแปร X และ Y มีความสัมพันธ์เชิงบวก อย่างสมบูรณ์
4. ถ้า ρ และ r มีค่าเท่ากับ -1 แสดงว่าตัวแปร X และ Y มีความสัมพันธ์เชิงลบ อย่างสมบูรณ์

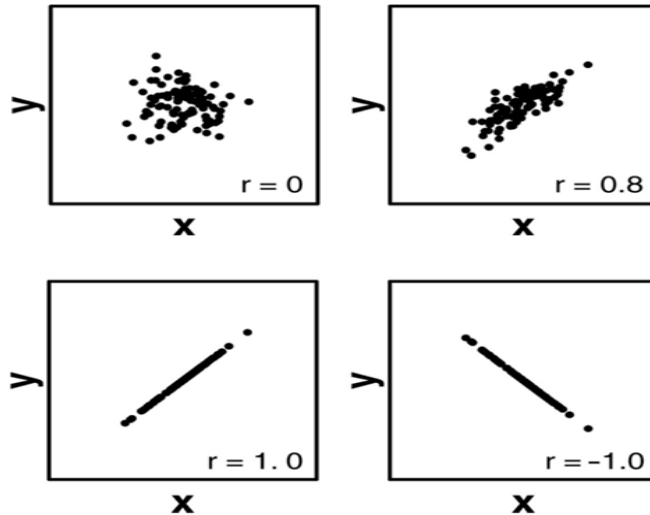
ค่าสัมประสิทธิ์สหสัมพันธ์

5. ถ้า ρ และ r มีค่าเข้าใกล้ 1 แสดงว่าตัวแปร X และ Y มีความสัมพันธ์เชิงบวก และมีความสัมพันธ์มาก
6. ถ้า ρ และ r มีค่าเข้าใกล้ -1 แสดงว่าตัวแปร X และ Y มีความสัมพันธ์เชิงลบ และมีความสัมพันธ์มาก
7. ถ้า ρ และ r มีค่าเข้าใกล้ 0 แสดงว่าตัวแปร X และ Y มีความสัมพันธ์กันน้อย
8. ถ้า ρ และ r มีค่าเท่ากับ 0 แสดงว่าตัวแปร X และ Y ไม่มีความสัมพันธ์เชิงเส้น

ค่าสัมประสิทธิ์สหสัมพันธ์

ตัวอย่าง ถ้าค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างความสูงกับน้ำหนักเท่ากับ 0.85 หมายความว่าความสูงมีความสัมพันธ์เชิงบวกอย่างมากกับน้ำหนัก ถ้าความสูงมากน้ำหนักจะมากด้วย และถ้าความสูงน้อยน้ำหนักจะน้อยด้วย เป็นต้น

ความหมายของ ρ และ r สามารถแสดงด้วยแผนภาพการกระจาย ดังนี้



การคำนวณค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน

วิธีนี้ใช้ในการคำนวณค่าเป็นการหาสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรสองตัวที่ใช้กันมากและรู้จักกันดีที่สุด ทั้งยังเป็นรากฐานของสหสัมพันธ์แบบอื่นๆ บางแบบเขียนแทนสัญลักษณ์ r หรือ r_{xy}

การคำนวณค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน

สูตรที่ 1 คำนวณโดยใช้คะแนนมาตรฐาน

$$r_{xy} = \frac{\sum Z_x Z_y}{n}$$

เมื่อ	r_{xy}	แทน	ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปร X และตัวแปร Y
	Z_x	แทน	คะแนนมาตรฐานแต่ละค่าของตัวแปร X
	Z_y	แทน	คะแนนมาตรฐานแต่ละค่าของตัวแปร Y
	n	แทน	จำนวนคู่ของข้อมูล

สูตรที่ 2 คำนวณโดยใช้คะแนนที่เป็นค่าเบี่ยงเบนจากค่าเฉลี่ย

$$r_{xy} = \frac{n \sum XY - \sum X \sum Y}{\sqrt{(n \sum X^2 - (\sum X)^2)(n \sum Y^2 - (\sum Y)^2)}}$$

เมื่อ r_{xy} แทน ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปร X กับตัวแปร Y

n

แทน จำนวนคู่ของข้อมูล

$\sum XY$

แทน ผลรวมทั้งหมดของผลคูณของตัวแปร X กับตัวแปร Y

$\sum X$

แทน ผลรวมของตัวแปร X

$\sum Y$

แทน ผลรวมของตัวแปร Y

$\sum X^2$

แทน ผลรวมของตัวแปร X แต่ละค่ายกกำลังสอง

$\sum Y^2$

แทน ผลรวมของตัวแปร Y แต่ละค่ายกกำลังสอง

สูตรที่ 2 เป็นสูตรที่นิยมใช้มากที่สุด เพราะข้อมูลที่เก็บรวบรวมมาได้ส่วนใหญ่อยู่ในรูปคะแนนดิบ

ตัวอย่าง จงหาความสัมพันธ์จากคะแนนของนักเรียน 10 คนที่กำหนดให้ต่อไปนี้

คนที่	X	Y	XY	x^2	y^2
1	14	19	266	196	361
2	11	11	121	121	121
3	6	5	30	36	25
4	17	14	238	289	196
5	8	10	80	64	100
6	13	16	208	169	256
7	6	14	84	36	196
8	9	10	90	81	100
9	16	15	240	256	225
10	12	14	168	144	196
	112	128	1,525	1,392	1,776

แทนค่าสูตร

$$r_{xy} = \frac{n \sum XY - \sum X \sum Y}{\sqrt{(n \sum X^2 - (\sum X)^2)(n \sum Y^2 - (\sum Y)^2)}}$$

$$= \frac{10(1525) - (112)(128)}{\sqrt{10(1392) - (112)^2} \sqrt{10(1776) - (128)^2}}$$

$$= \frac{15,250 - 14,336}{\sqrt{(13,390 - 12,544)(17,760 - 16,384)}}$$

$$= \frac{914}{\sqrt{(1,376)(1,376)}}$$

$$= \frac{914}{1376}$$

$$= 0.6642 \#$$

การแปลผล

1. ความสัมพันธ์ที่ได้เป็นการแสดงแต่เพียงว่าเกิดของสองอย่างที่มีสัมพันธ์กันเป็นไปตามกันหรือกลับกัน มิได้หมายความว่าอะไรเป็นเหตุเป็นผลซึ่งกันและกัน
2. ถ้าจะแปลผลในแง่ของปริมาณ ให้หาค่า r^2 ซึ่งเป็นสัดส่วนของความแปรปรวนรวมกันที่ใช้ในการพยากรณ์ได้ว่า เมื่อทราบค่าของตัวแปรหนึ่งการทำนายหรือพยากรณ์ตัวแปรที่สองจะถูกต้องเท่ากับ $100r^2\%$

สัมประสิทธิ์สหสัมพันธ์แบบอันดับของสเปียร์แมน หรือเรียกอีกอย่างหนึ่งว่า rank correlation เป็นวิธีการทางสถิติในการศึกษาความสัมพันธ์ของ 2 ตัวแปร จากกลุ่มตัวอย่างเดียวกัน โดยข้อมูลที่ได้จากตัวแปรทั้งสองอยู่ในมาตราเรียงอันดับ(ordinal scale) ซึ่งเขียนแทนสัญลักษณ์ ρ และคำนวณหาได้จากสูตร

$$\rho = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

เมื่อ

N = จำนวนคู่ของข้อมูล

D = ผลต่างของอันดับของข้อมูลแต่ละคู่

- การหาความสัมพันธ์จากอันดับที่ของข้อมูลจะมีประโยชน์มาก ในกรณีที่ข้อมูลที่นำมาหาค่าสหสัมพันธ์นั้นไม่อาจวัดค่าออกมาเป็นตัวเลขหรือไม่สามารถวัดเชิงปริมาณได้แต่เราสามารถจัดอันดับได้เช่น ความชอบหรือความนิยม ในการใช้สินค้าต่างๆ

ค่าสัมประสิทธิ์สหสัมพันธ์

ข้อตกลงเบื้องต้น

1. ข้อมูลหรือตัวแปรทั้งสองมีการวัดในระดับมาตราเรียงอันดับ
2. ในกรณีที่ข้อมูลมีจำนวนน้อย(น้อยกว่า30) เราอาจจะเปลี่ยนข้อมูลดิบให้เป็นแบบอันดับที่แล้วหาค่าสหสัมพันธ์โยใช้อันดับที่

ตัวอย่าง ผลการสอบวิชาคณิตศาสตร์และวิชาวิทยาศาสตร์ของนักศึกษา 10 คนปรากฏดังนี้
จงหาค่าสหพันธ์ระหว่างอันดับที่ของผลสอบวิชาทั้งสอง

คนที่	1	2	3	4	5	6	7	8	9	10
คะแนนคณิต	10	17	12	19	11	14	15	16	13	18
คะแนนวิทย์	11	15	13	19	12	16	14	18	10	17

วิธีทำ ก่อนอื่นต้องจัดอันดับคะแนนทั้งสองวิชาดังตารางต่อไปนี้

คนที่	คณิต R_x	วิทย์ R_y	$D=R_x - R_y$	D^2
1	10	9	1	1
2	3	5	-2	4
3	8	7	1	1
4	1	1	0	0
5	9	8	1	1
6	6	4	2	4
7	5	6	-1	1
8	4	2	2	4
9	7	10	-3	9
10	2	3	-1	1
			$\sum D = 0$	$\sum D^2 = 26$

ค่าสัมประสิทธิ์สหสัมพันธ์

จากตารางจะได้ $N = 10$

$$\sum D^2 = 26$$

สูตร

$$\begin{aligned}\rho &= 1 - \frac{6 \sum D^2}{N(N^2 - 1)} \\ &= 1 - \frac{6(26)}{10(100 - 1)} \\ &= 1 - 0.1576 \\ &= 0.8424\end{aligned}$$

หมายเหตุ ในกรณีที่อันดับของข้อมูลซ้ำกันให้ใช้อันดับเฉลี่ย เช่นคนที่ 3,4,5 ได้คะแนน เท่ากันก็ให้ทั้ง 3 คนมีอันดับที่

เท่ากับอันดับเฉลี่ยคือ $\frac{3+4+5}{3} = 4$ เป็นต้น

ค่าสัมประสิทธิ์สหสัมพันธ์

สัมประสิทธิ์การตัดสินใจ

ค่าสัมประสิทธิ์การกำหนด ซึ่งมีสัญลักษณ์คือ r^2 เป็นสถิติที่วัดความพอดีของแบบจำลองการถดถอย

ค่าสัมประสิทธิ์การกำหนดแสดงให้เห็นว่าแบบจำลองการถดถอยเหมาะสมกับชุดข้อมูลได้ดีเพียงใด กล่าวคือ เป็นตัวระบุเปอร์เซ็นต์ที่อธิบายโดยแบบจำลองการถดถอย

ดังนั้น ยิ่งค่าสัมประสิทธิ์การตัดสินใจสูง โมเดลการถดถอยก็จะยิ่งดีขึ้นเท่านั้น แม้ว่าจะไม่เป็นไปตามเงื่อนไขนี้เสมอไป แต่โดยหลักการแล้วเราสนใจที่จะกำหนดค่าสัมประสิทธิ์การตัดสินใจให้มากที่สุดเท่าที่จะเป็นไปได้

$$r^2 = \frac{(\sum xy)^2}{(\sum x^2)(\sum y^2)}$$

ดังนั้น ค่าของ r^2 จะมีค่า 0 ถึง 1 และค่า r^2 จะเป็นค่าที่บอกค่าสมการของเส้นถดถอยที่วิเคราะห์มาได้นั้นสามารถอธิบายความแปรปรวนที่เกิดขึ้นใน y ได้มากน้อยเพียงใด

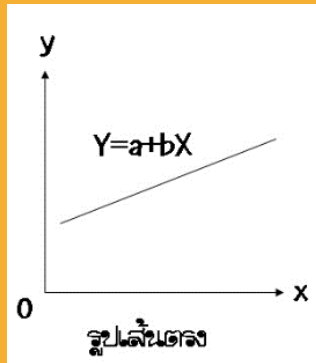
การวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย

การวิเคราะห์การถดถอย เป็นวิธีการหาสมการที่สามารถใช้ในการพยากรณ์ หรือคาดคะเนเกี่ยวกับตัวแปรที่ต้องการศึกษา โดยอาศัยความรู้เกี่ยวกับค่าของตัวแปรที่เกี่ยวข้อง นั่นคือ จากค่าของตัวแปรที่ต้องการศึกษาซึ่งจะเรียกว่า ตัวแปรตาม กับค่าของตัวแปรที่เกี่ยวข้อง ซึ่งเราจะเรียกว่า ตัวแปรอิสระ เราจะหาสมการซึ่งเพียงพอในการกำหนดความสัมพันธ์ระหว่างตัวแปรทั้งสอง สมการนี้ เรียกว่า **สมการเส้นถดถอย**

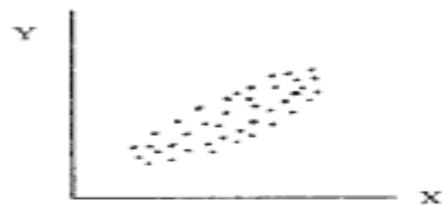
การวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย

การวิเคราะห์ข้อมูลถดถอยเชิงเส้นแบบง่ายคืออะไร

การวิเคราะห์ข้อมูลถดถอยเชิงเส้นแบบง่ายคือ การวิเคราะห์ความสัมพันธ์เชิงเส้นระหว่างตัวแปรต้น (แทนด้วย x) และตัวแปรตาม (แทนด้วย y) โดยมีตัวแปรต้นเพียง 1 ตัว ซึ่งมีความสัมพันธ์เชิงเส้นต่อกัน โดยจุดมุ่งหมายของการวิเคราะห์ข้อมูลคือการสร้างสมการเส้นตรงแบบถดถอยเชิงเส้นที่สามารถอธิบายความสัมพันธ์ของทั้งสองตัวแปรได้ดีที่สุด



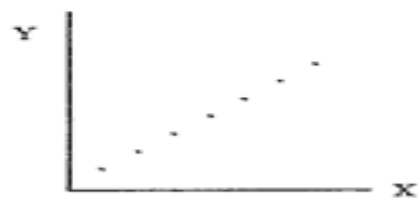
เมื่อ a และ b เป็นค่าคงที่ สมการถดถอยของกลุ่มตัวอย่าง



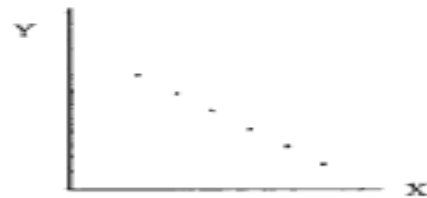
มีความสัมพันธ์ทางบวก
(r เป็นบวก)



มีความสัมพันธ์ทางลบ
(r เป็นลบ)



มีความสัมพันธ์ทางบวกอย่างสมบูรณ์
($r = +1.00$)



มีความสัมพันธ์ทางลบอย่างสมบูรณ์
($r = -1.00$)

ภาพของการกระจาย

สมการถดถอยเชิงเส้นอย่างง่าย

สูตรสมการถดถอยเชิงเส้นอย่างง่าย

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} \longrightarrow 1$$

$$a = \bar{y} - b\bar{x} \longrightarrow 2$$

สมการถดถอยเชิงเส้นอย่างง่าย

$$\hat{y} = a + bx_t \longrightarrow 3$$

เมื่อ $\hat{y}$ คือค่าแนวโน้มที่แปรตามตัวแปรเวลาที่เราต้องการหา

a คือระยะทางจากจุดกำเนิดถึงจุดตัดบนแกน y

b คือความชันของสมการเส้นตรง

x_t คือตัวแปรเวลา ณ เวลาที่ t

ตัวอย่าง จากข้อมูลในตารางเป็นคะแนนสอบกลางภาควิชาสถิติและวิชาอังกฤษของนักศึกษาแผนกวิชาคอมพิวเตอร์วิทยาลัย
แห่งหนึ่งจำนวน 10 คน

คะแนนสอบ สถิติ	75	55	71	72	80	95	95	98	65	82
คะแนนสอบ อังกฤษ	80	65	78	35	48	85	98	98	66	83

1. จงสร้างสมการการถดถอยที่แสดงความสัมพันธ์ระหว่างคะแนนสอบวิชาสถิติกับวิชาภาษาอังกฤษ
2. ถ้านักศึกษาคนหนึ่งสอบวิชาสถิติได้ 70 คะแนน จะสอบวิชาภาษาอังกฤษได้ประมาณกี่คะแนน

คะแนนสอบวิชาสถิติ(X)	คะแนนสอบ วิชาภาษาอังกฤษ(Y)	X ²	Y ²	XY
75	80	5,625	6,400	6,000
55	65	3,025	4,225	3,575
71	78	5,041	6,084	5,538
72	35	5,184	1,225	2,520
80	48	6,400	2,304	3,840
95	85	9,025	7,225	8,075
95	98	9,025	9,604	9,310
98	98	9,604	9,604	9,604
65	66	4,225	4,356	4,290
82	83	6,724	6,889	6,806
$\Sigma=788$	$\Sigma=736$	$\Sigma=63,878$	$\Sigma=5,7916$	$\Sigma=59,558$

วิธีทำ

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$



1

$$= \frac{10 (59,558) - 788 (736)}{10 (63,878) - (788)^2}$$

$$= \frac{15,612}{17,836}$$

$$= 0.8753$$

จากนั้นหา

$$\bar{x} = \frac{\sum x}{N}$$

และ

$$\bar{y} = \frac{\sum y}{N}$$

$$= \frac{788}{10}$$

$$= 78.8$$

$$= \frac{736}{10}$$

$$= 73.6$$

วิธีทำ

$$a = \bar{y}_t - b\bar{x}_t \longrightarrow 2$$

$$= 73.6 - 0.8753(78.8)$$

$$= 73.6 - 68.97364$$

$$= 4.6264$$

1. สมการการถดถอยที่แสดงความสัมพันธ์ระหว่างคะแนนสอบวิชาสถิติกับวิชาภาษาอังกฤษ คือ

$$\hat{Y}_t = a + bx_t \longrightarrow 3$$

$$= 4.6264 + 0.87x$$

2. นักศึกษาคนหนึ่งสอบวิชาสถิติได้ 70 คะแนน $X = 70$

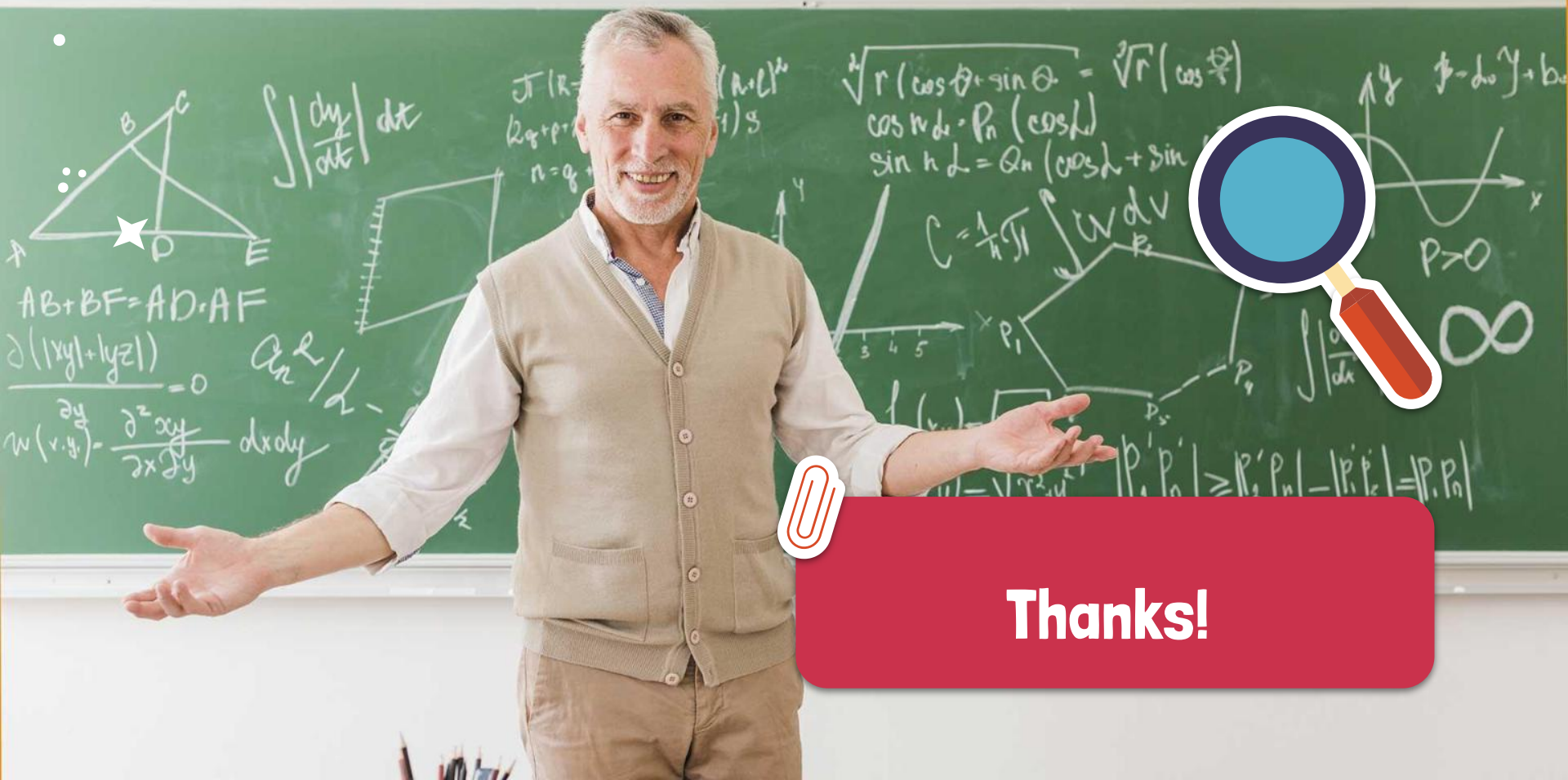
$$\hat{y} = 4.6264 + 0.8753x$$

$$= 4.6264 + 0.8753(70)$$

$$= 4.6264 + 61.261$$

$$= 65.8974$$

ดังนั้น นักศึกษาคนหนึ่ง que สอบวิชาสถิติได้ 70 คะแนน คาดว่าจะสอบวิชาภาษาอังกฤษได้ประมาณ 66 คะแนน



Thanks!