บทที่ 14

สหสัมพันธ์และการถดถอย

(Correlation and Regression)

14.1 **สหสัมพันธ์** (Correlation) เป็นการหาความสอดคล้องหรือความสัมพันธ์ของตัวแปร 2 ชุด โดยดูว่าตัวแปรทั้งสองนั้นมีความสัมพันธ์กันในลักษณะใด

ถ้าตัวแปรสุ่ม X และ Y มีความสัมพันธ์กัน เราสามารถหาค่าความสัมพันธ์นั้นได้ ค่าที่แสดงความสัมพันธ์นี้เรียกว่าสัมประสิทธิ์สหสัมพันธ์ (Correlation coefficient) เขียนสัญลักษณ์แทนสัมประสิทธิ์สหสัมพันธ์ด้วย r_{xy}

โดยทั่วไปจะนิยามสัมประสิทธิ์สหสัมพันธ์ว่าเป็นค่าเฉลี่ยของผลคูณของคะแนนมาตรฐานในตัวแปร 2 ชุด ซึ่งอาจเขียนได้ดังนี้

$$r_{xy} = \frac{\sum Z_x Z_y}{N} \qquad \text{......................(14–1)}$$

เมื่อ r_{xy} แทน ค่าสัมประสิทธิ์สหสัมพันธ์

Z_x แทน คะแนนมาตรฐานในตัวแปร X

Z_y แทน คะแนนมาตรฐานในตัวแปร Y

N แทน จำนวนคู่ของความสัมพันธ์

เนื่องจาก $Z_x = \frac{X-\bar{X}}{S_x}$ และ $Z_y = \frac{Y-\bar{Y}}{S_y}$

ดังนั้นสมการ (14–1) สามารถเขียนใหม่ได้ดังนี้

$$r_{xy} = \frac{\sum (X-\bar{X})\ (Y\text{-}k)}{NS_xS_y} \qquad \text{......................(14–2)}$$

ถ้าให้ x แทน $X-\bar{X}$

y แทน Y - Y

$$\text{จะได้ว่า } r_{xy} = \frac{\sum xy}{NS_xS_y} \qquad \text{......................(14–3)}$$

r_{xy} จะมีค่าอยู่ระหว่าง -1.00 ถึง + 1.00

ถ้า r เป็นบวก หมายความว่าตัวแปรทั้งสองนั้น มีความสัมพันธ์กันในทางตรง คือถ้า X มีค่าน้อย Y จะมีค่าน้อย และถ้า X มีค่ามาก Y จะมีค่ามาก

ถ้า r เป็นลบ หมายความว่าตัวแปรทั้งสองนั้น มีความสัมพันธ์ในทางกลับกัน คือถ้า X มีค่าน้อย Y จะมีค่ามาก และถ้า X มีค่ามาก Y จะมีค่าน้อย

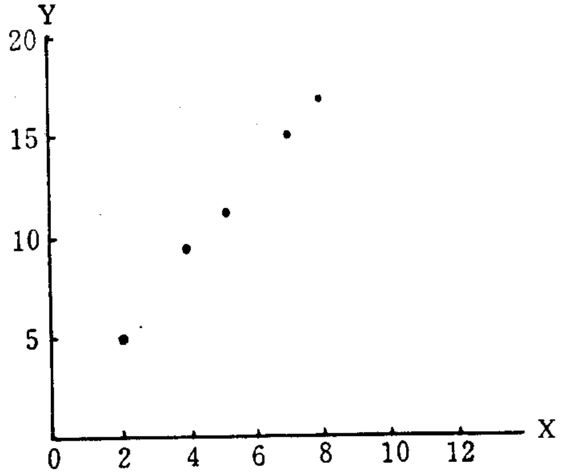
ถ้า r มีค่าอยู่ใกล้ศูนย์ (0) หมายความว่าตัวแปรทั้งสองนั้นมีความสัมพันธ์กันน้อยมาก หรือเกือบไม่มีเลย

14.2 กราฟแสดงความสัมพันธ์ของ X กับ Y

ถ้าตัวแปร X และ Y มีความสัมพันธ์กัน เราสามารถเขียน Scatter diagram แสดงความสัมพันธ์ของตัวแปรทั้งสองได้ ดังตัวอย่าง

1. ถ้า r_{xy} มีค่าเท่ากับ + 1.00 แสดงว่าตัวแปร X และ Y มีความสัมพันธ์กันทางบวกอย่างสมบูรณ์ กล่าวคือ หากตัวแปร X มีค่ามาก ตัวแปร Y ก็จะมีค่ามากด้วยและในทางตรงกันข้ามหากตัวแปร X มีค่าน้อย ตัวแปร Y ก็จะมีค่าน้อยด้วย ซึ่งสามารถเขียนกราฟแสดงความสัมพันธ์ระหว่างตัวแปร X กับ Y เมื่อ $r_{xy} = +1.00$ ได้ดังนี้

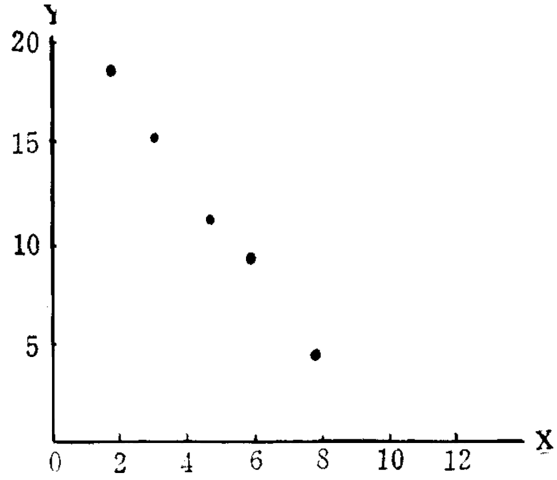
X	Y
2	5
4	9
4	9
5	11
7	15
8	17



r = 1.00

2. ถ้า r_{xy} มีค่าเท่ากับ–1.00 แสดงว่าตัวแปร X และ Y มีความสัมพันธ์กลับกัน กล่าวคือถ้าตัวแปร X มีค่ามาก ตัวแปร Y จะมีค่าน้อย แต่ถ้าตัวแปร X มีค่าน้อยตัวแปร Y จะมีค่ามาก ซึ่งสามารถเขียนกราฟแสดงความสัมพันธ์ระหว่างตัวแปร X กับ Y เมื่อ $r_{xy} = -1.00$ ได้ดังนี้

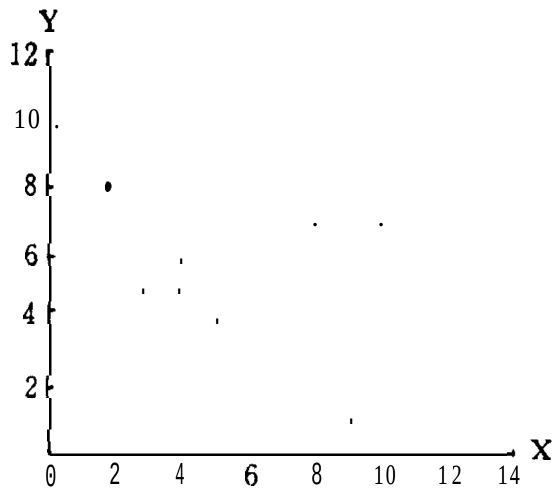
X	Y
2	17
3	15
5	11
6	9
6	9
8	5



r = -1.00

3. ถ้า r_{xy} มีค่าเท่ากับ 0 แสดงว่าตัวแปร X และ Y ไม่มีความสัมพันธ์กันเลย กล่าวคือ ถ้าตัวแปร X มีค่ามากตัวแปร Y อาจมีค่ามากหรือน้อยก็ได้เอาแน่นอนอะไรไม่ได้ ทำนองเดียวกัน ถ้าตัวแปร X มีค่าน้อย ตัวแปร Y อาจมีค่ามากหรือน้อยก็ได้ เอาแน่นอนไม่ได้เช่นกัน สามารถเขียนกราฟแสดงว่าตัวแปรทั้งสองไม่สัมพันธ์กัน (r_{xy} = 0) ได้ดังนี้

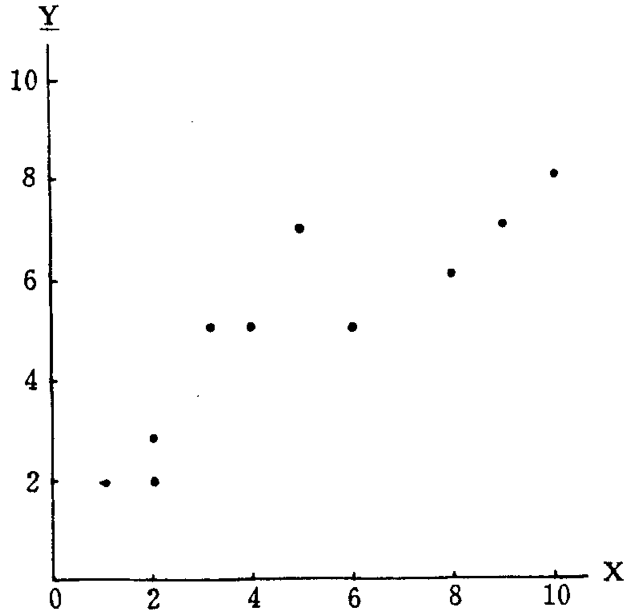
X	Y
0	3
2	8
3	5
4	6
4	5
5	4
5	4
8	7
9	1
10	7



r = 0

4. ถ้า r_{xy} มีค่าน้อยกว่า + 1.00 และมีค่าสูงมากในทางบวก อาจเขียนกราฟแสดงความสัมพันธ์ระหว่างตัวแปรทั้งสองได้ดังนี้

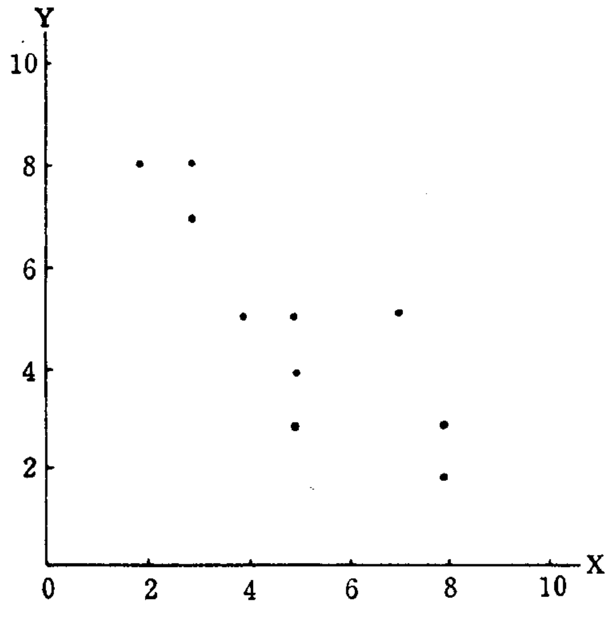
X	Y
1	2
2	3
2	2
3	5
4	5
5	7
6	5
8	6
9	7
10	8



r = 0.88

5. ถ้า r_{xy} มีค่าเข้าใกล้ − 1.00 อาจเขียนกราฟแสดงความสัมพันธ์ระหว่างตัวแปรทั้งสองได้ดังนี้

X	Y
2	8
3	7
3	8
4	5
5	4
5	5
5	3
7	5
8	3
8	2



r = − 0.85

14.3 วิธีหาค่าสัมประสิทธิ์สหสัมพันธ์

วิธีการในการคำนวณหาค่าสัมประสิทธิ์สหสัมพันธ์มีหลายวิธี แต่ในที่นี้จะขอกล่าวเฉพาะบางวิธีเท่านั้น โดยจะกล่าวถึงวิธีที่เป็นที่นิยมใช้กันทั่ว ๆ ไป หรือเป็นวิธีที่เกี่ยวข้องกับเรื่องการวัดผลการศึกษา ซึ่งนักศึกษาจะต้องนำไปใช้บ่อย ๆ ดังนี้

1. Pearson Product Moment Correlation

ในกรณีที่ตัวแปร X และ Y เป็นผลที่ได้จากการวัดในมาตราอันตรภาค (Interval scale) ขึ้นไป สามารถหาค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปร X และ Y โดยวิธีของ Karl Pearson ซึ่งได้มีการแปลง (derived) สูตร (14—2) ให้อยู่ในรูปที่ง่ายต่อการคำนวณ ดังนี้

$$r_{xy} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}} \qquad \text{..................(14-4)}$$

ตัวอย่าง 1 ในการสอบวิชาคณิตศาสตร์ (X) และวิทยาศาสตร์ (Y) กับนักเรียนจำนวน 10 คน ปรากฏคะแนนดังตารางข้างล่าง จงคำนวณหาค่าสัมประสิทธิ์สหสัมพันธ์ ของคะแนนวิชาคณิตศาสตร์และวิทยาศาสตร์

คนที่	X	Y
1	1	2
2	2	3
3	2	2
4	3	5
5	4	5
6	5	7
7	6	5
8	8	6
9	9	7
10	10	8

วิธีทำ

คนที่	X	Y	X^2	Y^2	XY
1	1	2	1	4	2
2	2	3	4	9	6
3	2	2	4	4	4
4	3	5	9	25	15
5	4	5	16	25	20
6	5	7	25	49	35
7	6	5	36	25	30
8	8	6	64	36	48
9	9	7	81	49	63
10	10	8	100	64	80
Σ	50	50	340	290	303

แทนค่าจากสูตร

$$r = \frac{10\ (303) - (50)\ (50)}{\sqrt{[10\ (340) - (50)^2]\ [10\ (290) - (50)^2]}}$$

$$= \frac{530}{\sqrt{(900)\ (400)}}$$

$$= \frac{530}{600}$$

$$= 0.883$$

2. Spearman Rank Correlation Coefficient

ในกรณีที่คะแนนของตัวแปร X และ Y อยู่ในมาตราอันดับ (Ordinal Scale) เราหาค่าสหสัมพันธ์โดยใช้ Spearman Rank Correlation Coefficient ดังสูตร

$$r_{xy} = 1 - \frac{6 \sum d_i^2}{N^3 - N} \qquad \text{.....................(14–5)}$$

เมื่อ d แทน ความแตกต่างของอันดับ

N แทน จำนวนคนในกลุ่มตัวอย่าง

3. Point Biserial Correlation Coefficient

เป็นการหาค่าสัมประสิทธิ์สหสัมพันธ์ที่พัฒนามาจาก Pearson Product Moment Correlation โดยมีหลักว่าตัวแปรหนึ่งจะต้องเป็นตัวแปรต่อเนื่อง (Continuous) (เป็นผลการวัดในระดับอันตรภาคขึ้นไป) ส่วนอีกตัวแปรหนึ่งเป็นผลการวัดในระดับนามบัญญัติ (Nominal Scale) ที่เป็น Dichotomus คือมีเพียง 2 กลุ่ม (Categories) โดยให้ Category หนึ่งเป็น 1 อีก Category หนึ่งเป็น 0 ตัวอย่างตัวแปร Dichotomus ได้แก่ตัวแปรเกี่ยวกับเพศ (ชาย–หญิง) ตัวแปรเกี่ยวกับผลการสอบ (ได้–ตก) ตัวแปรเกี่ยวกับอาชีพ (ชาวนา–ไม่ใช่ชาวนา) ถ้าตัวแปร X และ Y มีลักษณะดังกล่าวข้างต้น เราสามารถคำนวณหาค่าสัมประสิทธิ์สหสัมพันธ์โดยใช้ Point Biserial Correlation Coefficient ซึ่งมีสูตรดังนี้

$$r_{pbi} = \frac{M_p - M_q}{\sigma_x} . \sqrt{pq} \qquad \text{....................................(14–6)}$$

เมื่อ σ_x แทน ความเบี่ยงเบนมาตรฐานในตัวแปรต่อเนื่อง

M_p แทน คะแนนเฉลี่ยของตัวแปร Dichotomus ที่ได้คะแนน 1

M_q แทน คะแนนเฉลี่ยของตัวแปร Dichotomus ที่ได้คะแนน 0

p แทน สัดส่วนระหว่างจำนวนคนที่ได้คะแนน 1 ในตัวแปร Dichotomus กับจำนวนคนทั้งหมด

q แทน สัดส่วนระหว่างจำนวนคนที่ได้คะแนน 0 ในตัวแปร Dichotomus กับจำนวนคนทั้งหมด

ตัวอย่าง 2 ในการทดสอบนักเรียนจำนวน 8 คน ได้คะแนนรวม (X) และคะแนนในข้อที่ 5 (Y) ดังแสดงในตาราง จงคำนวณหาค่าสหสัมพันธ์ของคะแนนในข้อที่ 5 กับคะแนนรวมของข้อสอบฉบับนี้

คนที่	X	Y
1	1	1
2	1	1
3	2	0
4	6	1
5	6	1
6	7	0
7	8	0
8	9	0

แทนค่าจากสูตร

$$r_{pbi} = \frac{3.5-6.5}{3.0} \sqrt{\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)}$$

$$= 0.50$$

ค่า r ในที่นี้ก็คืออำนาจจำแนกของข้อสอบข้อที่ 5 นั่นเอง

สูตร (14–6) อาจแปลงสูตรให้อยู่ในอีกรูปหนึ่ง ซึ่งให้ค่าในการคำนวณเท่ากัน ได้ดังนี้

$$r_{pbi} = \frac{M_p - M_t}{\sigma_x} \sqrt{\frac{p}{q}} \qquad \text{....................(14–7)}$$

เมื่อ M_t แทนคะแนนเฉลี่ยของคะแนนรวม

จากตัวอย่าง จะได้ว่าคะแนนเฉลี่ยของคะแนนรวม $(M_t) = 5$

$$r_{pbi} = \frac{3.5-5.0}{3.0} \sqrt{\frac{.50}{.50}}$$

$$= \frac{-1.5}{3.0}$$

$$= -0.50$$

4. Biserial Correlation Coefficient

ค่าสัมประสิทธิ์สหสัมพันธ์แบบนี้เกิดจากผลการวัดในตัวแปร X และ Y ซึ่งเดิมอยู่ในระดับอันตรภาคขึ้นไป แต่ทว่าตัวแปรหนึ่งได้ถูกลดลงให้เหลือเป็นเพียงตัวเลขในมาตรานามบัญญัติ โดยลดลงเหลือเพียง 2 กลุ่ม เช่น ตัวแปรเกี่ยวกับผลสัมฤทธิ์ทางการเรียน ซึ่งแต่เดิมจะมีลักษณะเป็นตัวแปรต่อเนื่อง แต่หากถูกนำมาลดลงทำให้เหลือเพียง 2 กลุ่มคือผ่านกับไม่ผ่าน เราสามารถหาค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรทั้งสอง โดยใช้ biserial correlation ซึ่งมีสูตรดังนี้

$$r_{bis} = \frac{M_p - M_q}{\sigma_x} \cdot \frac{pq}{y} \qquad \text{..............(14—8)}$$

หรือ

$$r_{bis} = \frac{M_p - M_t}{\sigma_x} \cdot \frac{p}{y} \qquad \text{..............(14—9)}$$

เมื่อ X แทน ตัวแปรต่อเนื่อง

σ_x แทน ความเบี่ยงเบนมาตรฐานของตัวแปร X

M_p แทน ค่าเฉลี่ยของตัวแปร Dichotomus ที่เป็น 1

M_q แทน ค่าเฉลี่ยของตัวแปร Dichotomus ที่เป็น 0

M_t แทน คะแนนเฉลี่ยของคะแนนรวม

p แทน สัดส่วนระหว่างจำนวนคนในกลุ่มที่ ได้คะแนน 1 กับจำนวนคนทั้งหมด .

q แทน สัดส่วนระหว่างจำนวนคนในกลุ่มที่ ได้คะแนน 0 กับจำนวนคนทั้งหมด

y แทน Ordinate ของโค้งปกติ ซึ่งแบ่งโค้งออกเป็น 2 ส่วนคือ p กับ q

ตัวอย่าง 3 ในสถาบันแห่งหนึ่งมีคนจบการศึกษา 60 % และไม่จบการศึกษา 40% ผู้ที่จบการศึกษามีค่าเฉลี่ย I.Q. เท่ากับ 120 และผู้ที่ไม่จบการศึกษามีค่าเฉลี่ยของ I.Q. เท่ากับ 110 ความเบี่ยงเบนมาตรฐานเท่ากับ 15 จงหาค่าสหสัมพันธ์ระหว่าง I.Q. กับผลการศึกษา (จบ—ไม่จบ)

แทนค่าจากสูตร

$$r_{bi} = \frac{120 - 110}{15} \cdot (0.621)$$

$$= 0.41$$

ค่า $\frac{pq}{y}$ หาได้จากตาราง Ordinate of the Normal Curve ในภาคผนวกของหนังสือเล่มนี้

ค่า r_{pbi} จะต่ำกว่าค่าของ r_{bis} เสมอ ซึ่งสามารถเขียนความสัมพันธ์ ระหว่าง r_{pbi} กับ r_{bis} ได้ดังนี้

$$\frac{r_{bis}}{r_{pbi}} = \sqrt{\frac{pq}{y}} \qquad \text{...............(14–10)}$$

หรือ $$r_{pbi} = r_{bis} \cdot \frac{y}{\sqrt{pq}} \qquad \text{...... ....(14–11)}$$

5. Phi Coefficient

ในกรณีที่คะแนนที่ได้จากตัวแปร X และ Y เป็น Dichotomus ทั้งคู่ คือมีค่าเป็น 0 กับ 1 จะสามารถหาค่าสัมประสิทธิ์สหสัมพันธ์โดยใช้ Phi Coefficient (Ø)

ตาราง 14–1 แสดงคะแนนของนักเรียน 20 คน ในข้อสอบข้อที่ 5 และข้อที่ 6

ข้อที่ \ นักเรียน	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	p	q
5	0	0	0	1	0	0	1	0	1	1	1	1	1	1	0	1	0	1	0	1	0.55	0.45
6	0	0	0	1	0	0	1	0	1	1	1	0	1	0	0	0	0	1	1	0	0.40	0.60

จากตาราง 14–1 จะเห็นได้ว่ามีนักเรียน 4 คน ที่ตอบข้อ 5 ถูก แต่ตอบข้อ 6 ผิด มีนักเรียน 7 คน ที่ตอบถูกทั้งสองข้อ มีนักเรียน 8 คน ที่ตอบผิดทั้งสองข้อ และมีนักเรียนเพียง 1 คน เท่านั้นที่ตอบข้อ 5 ผิด แต่ตอบข้อ 6 ถูก จากข้อมูลนี้สามารถสรุปเป็นตารางได้ดังนี้

ตาราง 14–2 สรุปจำนวนคนตอบถูกและผิดในแต่ละข้อ

ข้อ 5 \ ข้อ 6	ผิด	ถูก	รวม
ถูก	4	7	11
ผิด	8	1	9
รวม	12	8	20

สามารถสรุปเป็นตารางทั่วไปได้ดังนี้

ตาราง 14–3

ข้อ i \ ข้อ k	ผิด	ถูก	รวม
ถูก	A (a)	B (b)	$a + b = p_i$
ผิด	c (c)	D (d)	$c + d = q_i$
รวม	$a + c = q_k$	$h + d = p_k$	1.00

จากตาราง 14–3 ให้ A, B, C, D แทนความถี่ของคะแนนดิบในแต่ละช่องและ a, b, c, d แทนสัดส่วนในแต่ละช่อง ซึ่ง $a+b+c+d = 1.00$

จากสมการ (14–3) จะได้ว่าค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างข้อ i และ k เป็น

$$r_{ik} = \frac{\Sigma (X_i - M_i)(X_k - M_k)}{N S_i S_k} \qquad \text{..................(14–12)}$$

พิจารณาเทอมขวามือโดยไม่พิจารณาค่าความเบี่ยงเบนมาตรฐาน

$$\frac{\sum X_i X_k}{N} - \frac{M_k \sum X_i}{N} - \frac{M_i \sum X_k}{N} + \frac{\sum M_i M_k}{N}$$

แต่ $\sum X_i X_{k/N} = p_{ik}$

(p_{ik} คือสัดส่วนของจำนวนคนที่ตอบถูกทั้งสองข้อ ซึ่งก็คือ ๖ ในตาราง)

และ $\frac{\sum X_i}{N} = M_i$, $\frac{\sum X_k}{N} = M_k$

ส่วนเทอมสุดท้าย $\frac{\sum M_i M_k}{N} = \frac{N M_i M_k}{N} = M_i M_k$

นั่นคือถ้าจะเขียนให้อยู่ในรูปง่าย ๆ จะได้เป็น

$$p_{ik} - M_i M_k - M_i M_k + M_i M_k = p_{ik} - M_i M_k$$

เนื่องจาก $M_i = p_i$ และ $M_k = p_k$

ดังนั้น $r_{ik} = \frac{p_{ik} - p_i p_k}{S_i S_k}$ ....(14–13)

ค่าความเบี่ยงเบนมาตรฐานของข้อ i คือ $\sqrt{p_i q_i}$

$$r_{phi} = \frac{p_{ik} - p_i p_k}{\sqrt{p_i q_i p_k q_k}} \qquad \text{..........................(14–14)}$$

จากตาราง 14–2 จะได้ว่า

$$r_{phi} = \frac{\frac{7}{20} - \left(\frac{11}{20}\right)\left(\frac{8}{20}\right)}{\sqrt{\left(\frac{11}{20}\right)\left(\frac{9}{20}\right)\left(\frac{8}{20}\right)\left(\frac{12}{20}\right)}}$$

$$= \frac{0.35 - 0.22}{\sqrt{0.0594}}$$

$$= 0.53$$

หรืออาจคำนวณค่า r_{phi} ได้โดยตรงจากความถี่ ดังสูตร

$$r_{phi} = \frac{BC - AD}{\sqrt{(A+B)(C+D)(A+C)(B+D)}} \qquad \text{.................(14–15)}$$

ดังนั้นค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างข้อ 5 และข้อ 6

$$r_{phi} = \frac{7 \times 8 - 4 \times 1}{\sqrt{11 \times 9 \times 12 \times 8}}$$

$$= \frac{56 - 4}{\sqrt{9504}}$$

$$= 0.53$$

14.4 การทดสอบนัยสำคัญของ r_{xy}

การคำนวณหาค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation coefficient) นั้น จำเป็นต้องมีการทดสอบให้แน่ใจก่อนว่า ค่า r_{xy} นั้นมีนัยสำคัญทางสถิติ (significance) จริงคือตัวแปรทั้งสองนั้นมีความสัมพันธ์กันจริงอย่างเชื่อมั่นได้ เราสามารถทดสอบนัยสำคัญของ r_{xy} ได้ 2 วิธีคือ

1. นำค่า r_{xy} ที่คำนวณได้ ไปเทียบกับค่า r ที่ได้จากการเปิดตารางค่านัยสำคัญของสัมประสิทธิ์สหสัมพันธ์ ที่ df = n−2 ในกรณีที่กลุ่มตัวอย่างมีไม่เกิน 100 คน

2. ทดสอบนัยสำคัญของ r_{xy} โดยใช้ t ดังสูตร

$$t = r_{xy}\sqrt{\frac{N-2}{1-r^2}} \qquad \text{.....................(14–16)}$$

$$df = n-2$$

ค่า t ที่ได้จากการคำนวณนี้ ถ้ามีค่ามากกว่าค่า t ในตาราง แสดงว่า r_{xy} นั้นมีนัยสำคัญทางสถิติ (significance) กล่าวคือตัวแปรทั้งสองนั้นมีความสัมพันธ์กันอย่างเชื่อมั่นได้

ถ้าค่า t ที่ได้จากการคำนวณมีค่าน้อยกว่าหรือเท่ากับค่า t ในตารางแสดงว่า r_{xy} นั้นไม่มีนัยสำคัญทางสถิติ กล่าวคือเราไม่พบว่าตัวแปรทั้งสองนั้นมีความสัมพันธ์กันอย่างแท้จริง

ตัวอย่าง 4 ในการตรวจวิชาเรียงความนักเรียนจำนวน 15 คน ค่าสัมประสิทธิ์สหสัมพันธ์ (r_{xy}) ของครู 2 คนเท่ากับ 0.75 จงทดสอบนัยสำคัญของค่าสหสัมพันธ์

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

$\alpha = .05$

แทนค่าจากสูตร

$$t = .75\sqrt{\frac{15-2}{1-.75^2}}$$

$$= \mathbf{14.74}$$

จากตารางที่ $\alpha = .05$, df $= 13$, $t = 2.16$

$\therefore$ เราปฏิเสธสมมติฐานเป็นกลาง (**reject** H_0)

นั่นคือ ครูทั้งสองคนตรวจข้อสอบวิชาเรียงความของนักเรียนได้สอดคล้องกัน อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

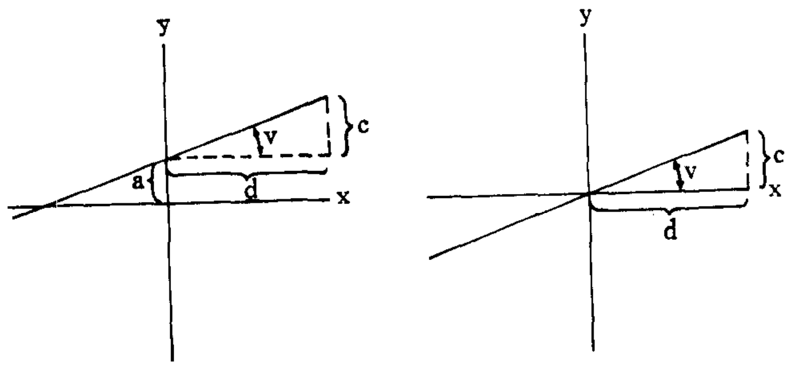
14.5 เส้นถดถอย (Regression line)

ถ้า X และ Y เป็นตัวแปรสุ่มที่มีความสัมพันธ์กัน ความสัมพันธ์นั้นอาจแสดงด้วยสมการเส้นตรง เช่น $Y = 5X + 4$ หรือหากเขียนให้อยู่ในรูปของสมการทั่วไปของเส้นตรง จะได้ว่า

$$Y = a + bX \qquad \text{..............(14-17)}$$

โดยที่ a แทนความยาวตั้งแต่จุด (0, 0) ถึงเส้นที่ตัดกับแกน Y

b แทนความลาดของเส้นตรง



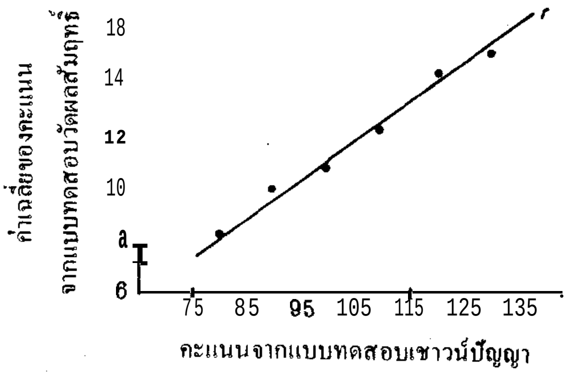
ภาพ แสดงสมการของเส้นตรง

ในทางการศึกษามักจะไม่พบตัวแปรทั้งสองที่มีความสัมพันธ์กันอย่างสมบูรณ์ ($r = 1.00$) เหมือนในตัวอย่างข้างต้น แต่อย่างไรก็ตาม สามารถที่จะหาเส้นตรงที่เหมาะสมมาแทนความสัมพันธ์นั้น ๆ ได้ เช่น ความสัมพันธ์ระหว่างเชาวน์ปัญญา และผลสัมฤทธิ์ทางการเรียน ดังตาราง

ตาราง 14—4 แสดงค่าเฉลี่ยของคะแนนผลสัมฤทธิ์ทางการเรียนของนักเรียน ซึ่งแบ่งตามช่วงระดับ I.Q.

I.Q.	75—84	85—94	95—104	105—114	115—124	125—134
M	8.33	10.00	10.80	12.20	14.25	15.00

ความสัมพันธ์นี้ สามารถนำมาเขียนกราฟได้ดังภาพ



ภาพแสดงความสัมพันธ์ระหว่างเชาวน์ปัญญากับผลสัมฤทธิ์ทางการเรียน

เส้นตรงที่เหมาะสมที่สุดที่จะอธิบายความสัมพันธ์ของตัวแปรทั้งสองนั้น เรียกว่า เส้นถดถอย (Regression line) ซึ่งอยู่ในรูปสมการ

$$Y' = a + b X \quad \text{.................. ....}(14\text{—}18)$$

โดยที่ Y' เป็นค่า Y ที่ได้จากการทำนาย

a, b เป็นค่าคงที่

การทำนายโดยอาศัยสมการเส้นตรงที่เหมาะสม นิยมใช้หลักการของวิธีกำลังน้อยที่สุด (The method of least squares) ซึ่งทำให้ผลบวกทั้งหมดยกกำลังสอง (Sum of square) ของระยะทางระหว่างจุด (x, y) กับเส้นตรงเส้นหนึ่งมีค่าน้อยที่สุด หรืออาจจะสรุปหลักของวิธีกำลังสองน้อยที่สุด ได้ดังนี้

1. ผลรวมของส่วนเบี่ยงเบนจากเส้นตรงนี้ยกกำลังสอง $(\Sigma(Y-Y'))^2$ จะมีค่าน้อยกว่าส่วนเบี่ยงเบนที่เกิดจากเส้นตรงอื่นใด

2. โดยเฉลี่ยแล้วส่วนเบี่ยงเบนที่อยู่เหนือเส้นจะเท่ากับส่วนเบี่ยงเบนที่อยู่ใต้เส้น สรุปแล้วจะมีเส้นตรงเพียงเส้นเดียวที่เป็นไปตามเงื่อนไขดังกล่าว

ค่า a และ b ในสมการของเส้นถดถอย คำนวณได้จากสูตรดังนี้

$$b = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} \qquad \text{..........} (14-19)$$

และ $a = \bar{Y} - b\bar{X}$ ....(14–20)

ตาราง 14—5 แสดงคะแนนที่ได้จากการสอบคัดเลือก (X) และเปอร์เซ็นต์ที่ได้ (Y) ของนักเรียนจำนวน 12 คน

ที่	X	Y	X^2	XY
1	63	87	3969	5481
2	50	74	2500	3700
3	55	76	3025	4180
4	65	90	4225	5850
5	55	85	3025	4675
6	70	87	4900	6090
7	64	92	4096	5888
8	70	98	4900	6860
9	58	82	3364	4756
10	68	91	4624	6188
11	52	77	2704	4011
12	60	78	3600	4680
$N = 12$	$\Sigma X = 730$	$\Sigma Y = 1017$	$\Sigma X^2 = 4493;$	$\Sigma XY = 62352$

$$\bar{X} = 60.83$$
$$\bar{Y} = 84.75$$
$$N = 12$$

แทนค่าจากสูตร

$$b = \frac{12\ (62352) - (730)\ (1017)}{12\ (44932) - (730)^2}$$
$$= 0.925$$
$$a = 84.75 - (0.925)\ (60.83)$$
$$= 28.482$$

∴ สมการถดถอย คือ

$$Y' = 0.925\ X + 28.482$$

สมการนี้ทำนายเปอร์เซ็นต์ได้ ถ้าทราบคะแนนสอบคัดเลือก

เช่น นักเรียนคนหนึ่งได้คะแนนสอบคัดเลือก (X) เท่ากับ 40 คะแนน อยาก ทราบว่านักเรียนคนนี้จะได้เปอร์เซ็นต์ (Y) เท่าไร

$$Y' = 0.925\ (40) + 28.482$$
$$= 37 + 28.482$$
$$= 65.482$$

ในการทำนายแต่ละครั้ง ย่อมเกิดความคลาดเคลื่อนได้ ทั้งนี้ เพราะตัวแปรทั้งสอง มักไม่ค่อยมีความสัมพันธ์กันอย่างสมบูรณ์ ความคลาดเคลื่อนในการทำนาย เรียกว่า Standard error of estimate เขียนแทนด้วย $S.E._{est.}$

ถ้า $S.E._{est}$ มีค่ามาก แสดงว่าการทำนายครั้งนี้มีความคลาดเคลื่อนเกิดขึ้นมาก แต่ถ้า $S.E._{est.}$ มีค่าน้อย แสดงว่าการทำนายครั้งนี้มีความคลาดเคลื่อนเกิดขึ้นน้อย

ถ้าค่า Y ที่ได้จากการทำนาย (Y') มีค่าต่างไปจากค่า Y_i ซึ่งได้จากการวัดจริง ๆ มาก ค่า $S.E._{est.}$ ก็จะมากด้วย

ถ้าค่า Y ที่ได้จากการทำนาย (Y') มีค่าต่างไปจากค่า Y_i ซึ่งได้จากการวัดจริง ๆ น้อย ค่า $S.E._{est}$ ก็จะน้อยด้วย

และถ้าค่า Y ที่ได้จากการทำนาย (Y') ไม่ต่างจากค่า Y_i' ซึ่งได้จากการวัดจริง ๆ เลย $(Y_i'-Y')^2 = 0$ ค่า $S.E._{est.}$ ก็จะเป็น 0 นั่นคือไม่มีความคลาดเคลื่อนในการทำนายเลย

ค่าความคลาดเคลื่อนในการทำนายคำนวณได้จากสูตร

$$S.E._{est.} = \sqrt{\frac{\Sigma(Y_i - Y')^2}{N-2}} \qquad \text{.........................(14–21)}$$

ถ้าเส้นตรงตัดกับแกน x และ y ที่จุดกำเนิด (0,0) ค่าของ a จะเป็น 0 จะได้สมการเป็น

$$Y = bX \qquad \text{.........................(14–22)}$$

ในรูปคะแนนดิบอาจมีปัญหาเรื่องการแจกแจงของคะแนนทั้ง 2 ชุดนั้น ฉะนั้นเพื่อตัดปัญหาเรื่องนี้ออกไปก็สามารถทำได้โดยแปลงคะแนนดิบให้เป็นคะแนนมาตรฐาน ทั้งนี้เพื่อให้ค่าเฉลี่ยของการแจกแจงทั้งสองเป็นศูนย์ นั่นคือถ้าเส้นตรงผ่านจุดกำเนิด และค่า X อยู่ในรูปของคะแนนมาตรฐาน จะได้

$$Z_y' = bZ_x \qquad \text{.........................(14–23)}$$

เนื่องจากคะแนนทุก ๆ ตัวในที่นี้ได้แปลงเป็นคะแนนมาตรฐาน ดังนั้นค่า b (ค่า tangent ของมุม v) จะเท่ากับค่าสหสัมพันธ์ระหว่างตัวแปรทั้งสอง (r_{xy})

$$Z_y' = r_{xy}Z_x \qquad \text{.........................(14–24)}$$

ในทำนองเดียวกันก็สามารถทำนายค่า x จากค่า Y ได้เช่นกัน

$$Z_x' = r_{xy}Z_y \qquad \text{.........................(14–25)}$$

แบบฝึกหัด 14

Y	82	66	78	34	47	85	99	99	68

ก. จงเขียน scatter diagram

ข. จาก scatter diagram จงเดาดูว่าค่า r จะเข้าใกล้ 1, —1 หรือ 0

ค. คำนวณค่า r

3. จงคำนวณหาค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างน้ำหนัก (X) กับความสูง (Y) ซึ่งมีข้อมูลดังตาราง

X	Y	X	Y	X	Y
10.0	18.7	12.5	20.8	14.5	19.6
10.5	21.5	13.0	21.6	15.0	23.8
11.0	18.5	13.5	22.4	15.5	21.7
11.5	19.6	14.0	23.3	16.0	23.2
12.0	18.2				