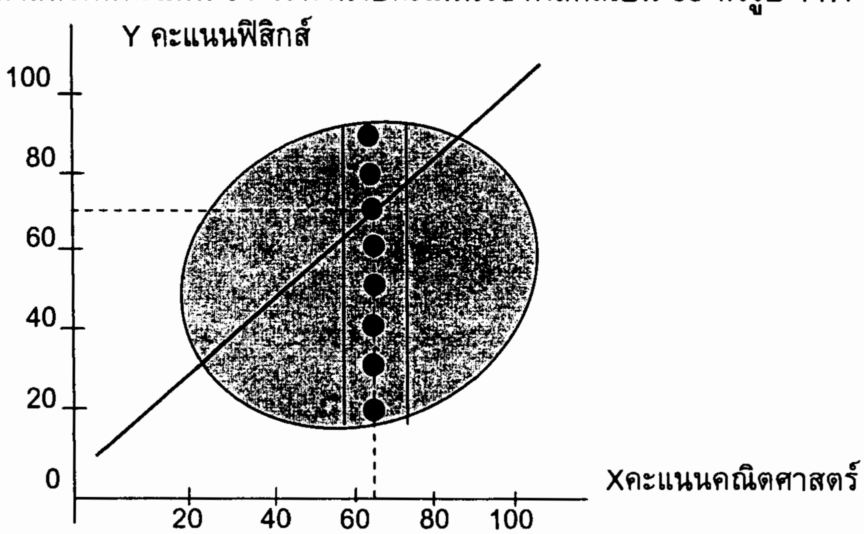
บทที่ 11

การทำนาย

ถ้าเราหาความสัมพันธ์ระหว่างคะแนนเช่นคะแนนวิชาคณิตศาสตร์ และคะแนนวิชาฟิสิกส์ อาจพบว่าคะแนนวิชาคณิตศาสตร์ สามารถทำนายผลของวิชาฟิสิกส์ได้ ซึ่งมีสองวิธีที่จะใช้ทำนายโดยการพิจารณาแผนภาพการกระจาย (scatter diagram) ซึ่งเราสามารถเห็นความเกี่ยวข้องทางบวกระหว่างคะแนน คณิตศาสตร์ และคะแนนวิชาฟิสิกส์สำหรับกลุ่มตัวอย่างของนักเรียน

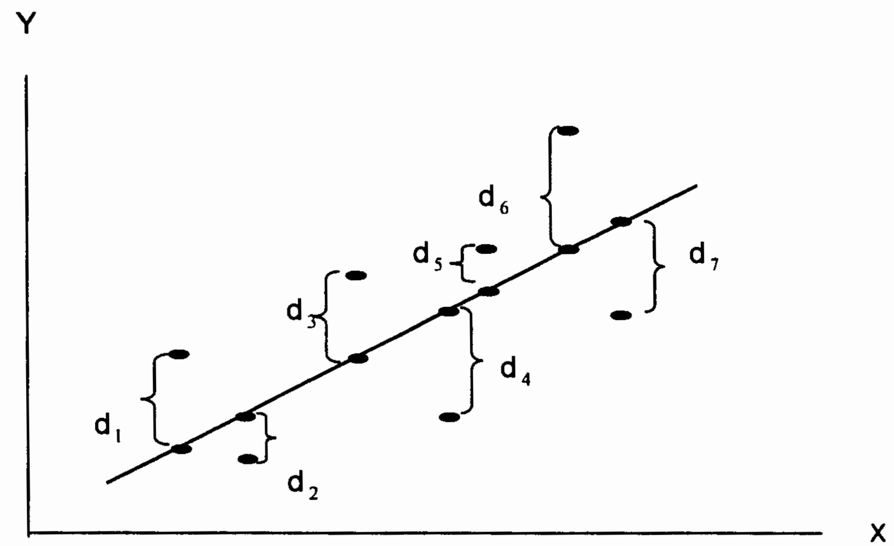
ถ้าสมเหตุสมผลที่จะทำนายว่า X และ Y มีความสัมพันธ์เชิงส้น เราสามารถปรับปรุงการทำนายของ Y จาก X โดยหาเส้นตรงที่ดีที่สุดที่เหมาะสมกับข้อมูล Y ซึ่งเส้นตรงนี้จะใช้ตรวจสอบทุกคะแนนในกลุ่มตัวอย่าง นักคณิตศาสตร์เรียกเส้นตรงที่ดีที่สุดที่เหมาะสมว่าเส้นถดถอย (regression line) สมการที่สอดคล้องกับเส้นตรงดังกล่าวเรียกว่าสมการถดถอย (regression equation) เส้นถดถอยดังแสดงในรูป 11.1 สามารถใช้ทำนายในกรณีใหม่ โดยเริ่มจากคะแนนคณิตศาสตร์บนแกนนอน (abscissa) ไปบนเส้นและไปยังแกนตั้ง (ordinate) โดยวิธีการนี้ การทำนายที่ดีที่สุดสำหรับนักเรียนด้วยคะแนนคณิตศาสตร์ที่มีคะแนน 64 จะทำนายคะแนนวิชาฟิสิกส์เป็น 68 ดังรูป 11.1



รูป 11.1

1. เกณฑ์ของความเหมาะสมที่สุด

เป็นเรื่องดีที่จะพูดถึงการหาเส้นตรงที่เหมาะสมที่สุด แต่เราจะรู้ได้อย่างไรว่าเราใช้วิธีที่ดีที่สุด วิธีที่ดีที่สุดสามารถนิยามได้มากกว่าหนึ่งวิธี ผลเฉลยของ Karl Pearson กับปัญหานี้คือประยุกต์เกณฑ์กำลังสองน้อยที่สุด (the least-square criterion) ดังรูป 11.2 แสดงการทำนาย Y จาก X



รูป 11.2

ภาพประกอบแสดงการแจกแจงของสองตัวแปร (bivariate distribution) และเส้นถดถอยที่เป็นไปได้ การทำนายจากเส้นนี้ดีอย่างไร สำหรับเจ็ดกรณีแสดงในแผนภาพการกระจาย (scatter diagram) ความคลาดเคลื่อนของการทำนายปรากฏที่เส้นในแนวตั้ง (vertical line)

ให้ d_Y เป็นผลต่างระหว่างค่าจริงของ Y และค่าทำนาย $d_Y = Y - Y'$ เกณฑ์กำลังสองที่น้อยที่สุดสำหรับเส้นถดถอยของเพียร์สันคือ: เส้นที่เหมาะสมที่สุดคือทำให้ผลบวกของกำลังสองผลความแตกต่างมีค่าน้อยที่สุด ดังนั้น $\sum d_Y^2$ เป็นค่าที่น้อยที่สุดเท่าที่จะเป็นไปได้

อาจมีข้อสงสัยว่าทำไมจึงไม่หาค่าต่ำสุดผลบวกของขนาดค่าสัมบูรณ์ของผลต่างแทนที่จะหาจากผลบวกของกำลังสอง คำตอบมีสองส่วน (1) เป็นเรื่องยากที่จะจัดการเชิง

คณิตศาสตร์กับค่าสัมบูรณ์ของความแตกต่างของค่า Y ในขณะที่กำลังสองของความแตกต่างของค่า Y ทำให้เกิดพัฒนาการทางคณิตศาสตร์ของค่าทางปฏิบัติ (2) ตำแหน่งของเส้นถดถอยและค่าของสัมประสิทธิ์สหสัมพันธ์จะเปลี่ยนแปลงภายใต้อิทธิพลของการสุ่มตัวอย่างน้อยกว่าที่เกิดขึ้นกับการใช้เกณฑ์อื่นๆ

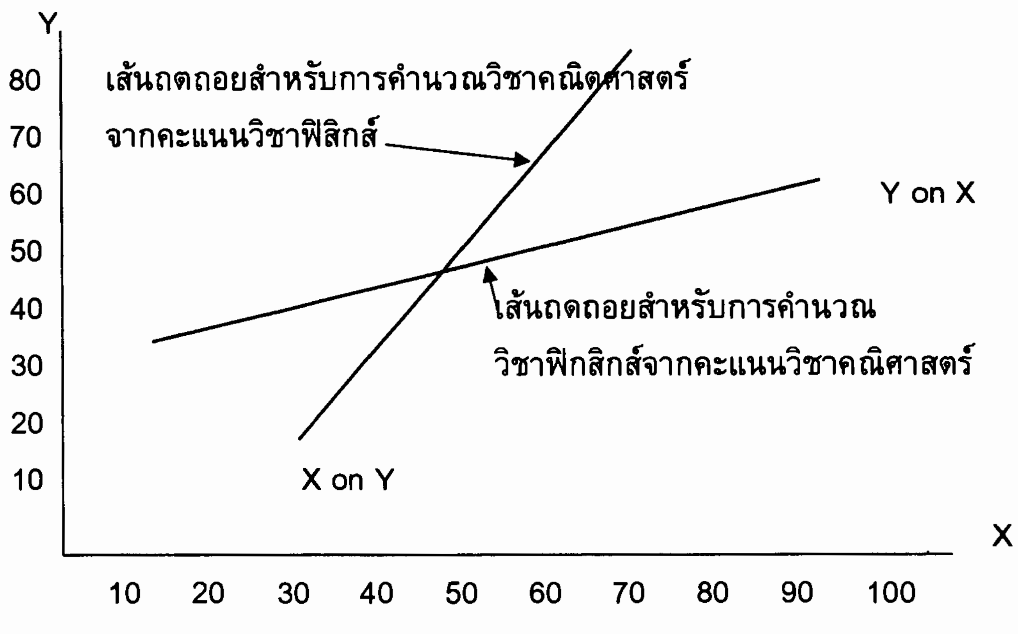
เราควรทราบว่าผลบวกของส่วนเบี่ยงเบนกำลังสองจากค่าเฉลี่ย $\sum(X-\overline{X})^2$ เป็นค่าต่ำสุด มีความสัมพันธ์ระหว่างเส้นถดถอยของเพียร์สันและค่าเฉลี่ย

ขั้นแรกค่าเฉลี่ยเป็นผลเฉลยของกำลังสองที่น้อยที่สุดของการวัดแนวโน้มเข้าสู่ส่วนกลาง ดังนั้นเส้นถดถอยเป็นผลเฉลยที่ยกกำลังสองที่น้อยที่สุดกับปัญหาของการหาเส้นตรงที่เหมาะสมที่สุด

ข้อที่สอง เส้นถดถอยแท้ที่จริงเป็นชนิดหนึ่งของค่าเฉลี่ย สำหรับแต่ละค่าของ X เส้นถดถอยบอกเราเกี่ยวกับค่าเฉลี่ย หรือค่าคาดหมายของ Y ในขณะที่ $\overline{Y}$ เป็นค่าเฉลี่ยของทุกค่า Y และ Y' เป็นค่าประมาณของค่าเฉลี่ยของ Y กำหนดเงื่อนไขที่ X มีค่าเฉพาะ

ดังนั้น เราจะคิดเฉพาะการทำนาย Y จาก X แต่ถ้าเราจะทำนาย X จาก Y สมการที่ได้จะเป็นคนละสมการกัน เนื่องจากเราต้องหากำลังสองที่น้อยที่สุด ใน X, $\sum(X-X')^2$ ถ้า r ไม่เท่ากับ ± 1.00 สองเส้นนั้นจะแตกต่างกัน รูป 11.3

วิชาฟิสิกส์



รูป 11.3 สองเส้นถดถอยสำหรับวิชาคณิตศาสตร์และฟิสิกส์

2. สมการถดถอย : รูปคะแนนมาตรฐาน (The Regression Equation: Standard Form)

ในการปฏิบัติ เราไม่จำเป็นต้องทดลองกับเส้นต่าง ๆบนแผนภาพการกระจาย เพื่อหาเส้นที่เหมาะสมที่สุดโดยเกณฑ์ของเพียร์สัน และเราไม่จำเป็นต้องลากเส้นกราฟ เพื่อใช้ในการทำนาย แต่เราจะหาเส้นตรงตามหลักกำลังสองที่น้อยที่สุด

การถดถอยของ Y บน X :

สูตรคะแนนมาตรฐาน

$$z'_Y = r z_X \qquad \text{.......... (11.1)}$$

เมื่อ z'_Y คือคะแนนมาตรฐานที่เป็นค่าทำนายของ Y

r คือสัมประสิทธิ์สหสัมพันธ์ระหว่าง X และ Y

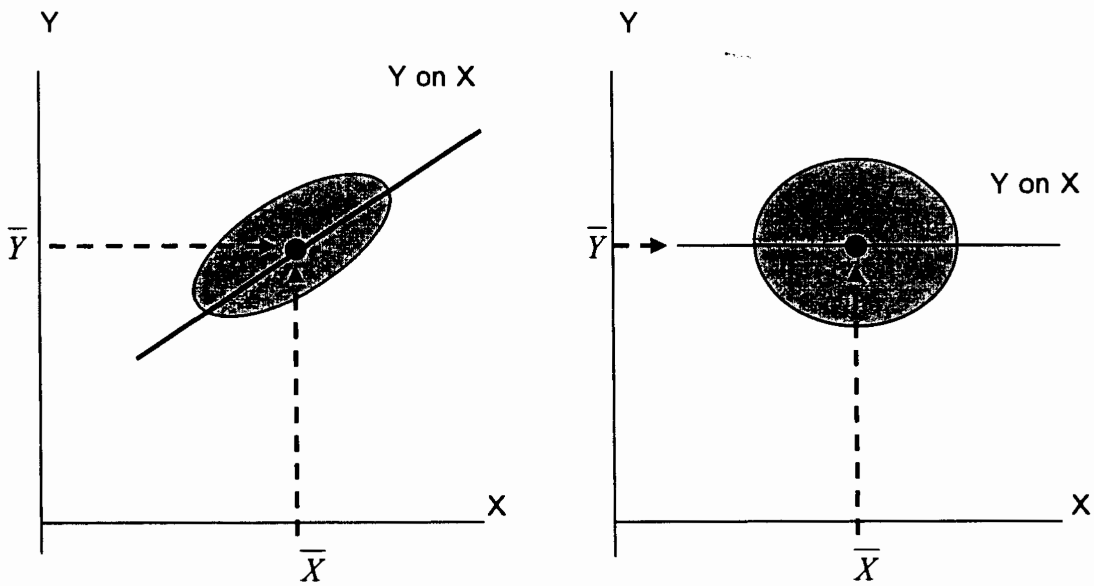
z_X คือคะแนนมาตรฐานของ X ซึ่งมี z'_Y เป็นค่าจากการทำนาย

ในกรณีที่เราทำนาย Y จาก X (predict Y from X) เราอาจเรียกว่าการทำนายของ Y บน X (regression of Y on X)

สมการถดถอยคะแนนมาตรฐาน ไม่ค่อยใช้สำหรับการทำนายที่แท้จริง เพราะว่าคะแนนจากการทำนายอยู่ในรูปของคะแนนดิบ

ขั้นที่หนึ่ง สมมุติว่าเราต้องการทำนาย Y จากค่าเฉลี่ยของ X เพราะว่าค่าเฉลี่ยของชุดของคะแนนแสดงในรูปคะแนนมาตรฐานปกติจะเป็นศูนย์ ค่ามาตรฐานของตัวทำนายบน Y จะเป็นศูนย์ด้วย นั่นคือ $z'_Y = r\ (0) = 0$ (ค่าเฉลี่ยของ Y) ซึ่งไม่เกี่ยวกับค่าของ r ดังนั้น (สำหรับทุกค่าของ r) สมการถดถอยทำนายว่ากรณีที่ค่าเฉลี่ยของ X จะเป็นคะแนนที่ค่าเฉลี่ยของ Y บนกราฟ เส้นถดถอยจะผ่านจุดที่แสดงค่าเฉลี่ยของ X และค่าเฉลี่ยของ Y เสมอ ดังแสดงในรูป 11.4 (ก)

ขั้นที่สอง ถ้า r = 0 ค่ามาตรฐานที่เป็นตัวทำนายของ Y จะเป็นศูนย์เสมอ : $z'_Y = 0(z_X) = 0$ ในพจน์ของคะแนนดิบ ถ้าสหสัมพันธ์เป็นศูนย์ แล้วค่าทำนายของY เป็นค่าเฉลี่ยของ Y ไม่ว่าค่าอะไรของ X จะใช้ทำนาย Y การประยุกต์เป็นสิ่งน่าสนใจมีเหตุผลเชิงสหัชญาณของการทำนาย แต่มันยังสอดคล้องกับเกณฑ์กำลังสองที่น้อยที่สุด ผลบวกของกำลังสองของความคลาดเคลื่อนของการทำนายเป็นค่าที่น้อยที่สุด รูป11.4 (ข) แสดงเส้นถดถอยของ Y บน X เมื่อ r = 0



รูป 11.4 (ก) รูป11.4 (ข)

รูป 11.4 แผนภาพการกระจายแสดงเส้นถดถอยของ Y บน Xเมื่อ (1) $r > 0$ และ (2) $r = 0$ ทั้งสองเส้นผ่านจุด $X = \overline{X}$, $Y = \overline{Y}$

3. สมการถดถอย : รูปคะแนนดิบ

แม้ว่าสูตรคะแนนมาตรฐานช่วยให้เราเข้าใจเส้นตรงที่เหมาะสมที่สุด ส่วนใหญ่ของปัญหาในการทำนายที่แสดงในพจน์ของคะแนนดิบ การแปลงคะแนนมาตรฐานของสมการถดถอยสู่รูปคะแนนดิบที่มีประโยชน์ในการปฏิบัติงาน ขั้นที่หนึ่ง แทนคะแนน z ด้วยคะแนนดิบที่สมมูลกัน นั่นคือ

$$\frac{Y' - \bar{Y}}{s_Y} = r\frac{X - \bar{X}}{s_x}$$

ดังนั้น แก้สมการหา Y' จะได้ในรูปคะแนนดิบของสมการถดถอย

สมการถดถอยของ Y บน X

สูตรคะแนนดิบ

$$Y' = r\frac{s_Y}{s_X}(X - \bar{X}) + \bar{Y} \qquad \text{..........(11.2)}$$

เมื่อ Y' คือคะแนนดิบที่เป็นค่าทำนายใน Y

s_X และ s_Y คือส่วนเบี่ยงเบนมาตรฐานของ X และ Y ตามลำดับ

$\bar{X}$ และ $\bar{Y}$ คือค่าเฉลี่ยของ X และ Y ตามลำดับ

r คือสัมประสิทธิ์สหสัมพันธ์ระหว่าง X และ Y

ในการทำนาย Y' เราจำเป็นต้องรู้ค่าของ X ที่จำเพาะเจาะจง ค่าเฉลี่ยสองค่า ส่วนเบี่ยงเบนมาตรฐานสองค่า และค่าสัมประสิทธิ์สหสัมพันธ์ ยกตัวอย่าง เมื่อ X เป็นคะแนนวิชาคณิตศาสตร์ Y เป็นคะแนนวิชาฟิสิกส์ ของผลสอบปลายภาคของนักเรียนชั้นมัธยมศึกษาปีที่ 5 ซึ่งได้ข้อมูลดังนี้

กำหนด	X คือคะแนนวิชาคณิตศาสตร์	Y คือคะแนนวิชาฟิสิกส์
	$\bar{X}$ = 64	$\bar{Y}$ = 68
	s_X = 14	s_Y = 16
	r = .38	

ตัวอย่าง ถ้าหากนักเรียนอยู่ในชั้นนี้และได้คะแนนวิชาคณิตศาสตร์ 67 จงทำนายคะแนนวิชาฟิสิกส์ของนักเรียนคนนี้

วิธีทำ ขั้นที่ 1 แทนค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน และสัมประสิทธิ์สหสัมพันธ์ในสูตรคะแนนดิบและเขียนในรูปอย่างง่ายจะได้

$$Y' = r\frac{s_Y}{s_X}(X - \bar{X}) + \bar{Y}$$

$$= 0.38\frac{16}{14}(X - 64) + 68$$

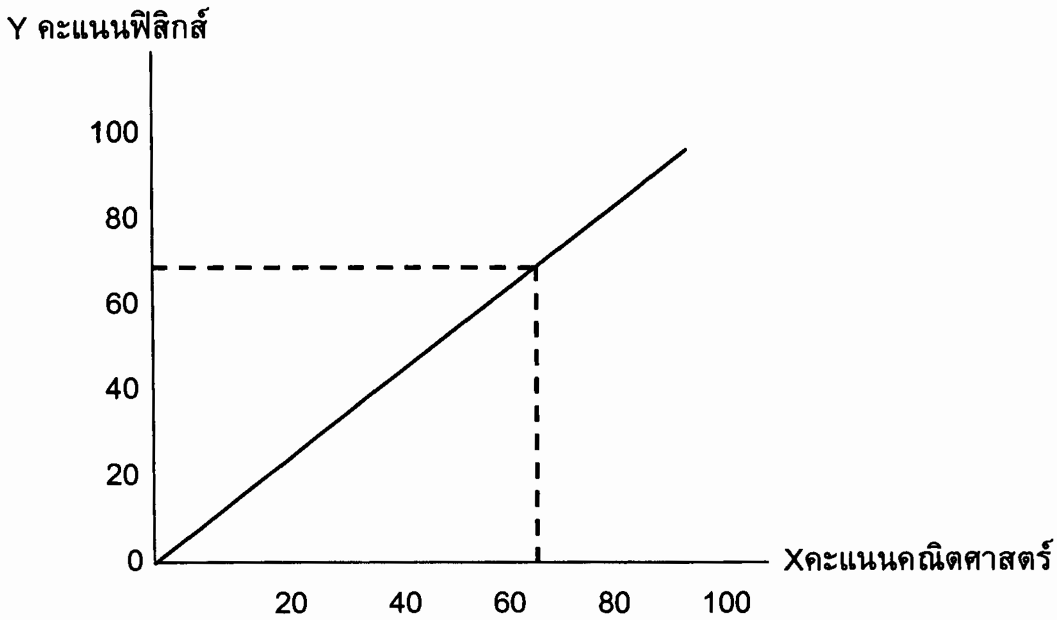
= 0.43429X - 27.79429 + 68

= 0.43429X + 40.20572

ขั้นที่ 2 แทนค่า X = 67 เพื่อหาค่าทำนาย Y'

Y' = 0.43429(67) + 40.20572

= 69.30



รูป 11.5 กราฟของสมการถดถอยสำหรับทำนายผลการสอบวิชาฟิสิกส์จากคะแนนวิชาคณิตศาสตร์

4. ความคลาดเคลื่อนจากการทำนาย: ความคลาดเคลื่อนมาตรฐานของการทำนาย

สมการถดถอยกำหนดอะไรเป็นค่าของ Yที่ถูกคาดหวัง (Y') เมื่อX เป็นค่าเฉพาะแน่ละ Y' ไม่เหมือนกับเป็นค่าที่แท้จริงของ Y ที่สมนัยกับค่าเฉพาะของ X เช่นถ้าผู้ชายสูง 6 ฟุต สมการถดถอยที่เหมาะสมอาจทำนายน้ำหนักของเขาเป็น 78 กิโลกรัม แต่เราไม่ได้คาดหมายให้คนที่สูง 6 ฟุตที่กำหนดให้มีน้ำหนัก 78 กิโลกรัม อย่างแน่นอน จำไว้ว่า ค่า

ทำนายเป็นเพียงค่าประมาณของค่าเฉลี่ยของ Y สำหรับกรณีของคะแนน X ที่กำหนดให้ ดังนั้น 78 กิโลกรัมเป็นเพียงค่าประมาณที่ดีที่สุดของน้ำหนักของคนที่สูง 6 ฟุต ถ้าสหสัมพันธ์ต่ำ จะมีการเบี่ยงเบนของค่าที่แท้จริงจากค่าทำนายมาก ถ้าค่าสหสัมพันธ์สูงค่าที่แท้จริงจะมีค่าใกล้เคียงกับค่าทำนายมาก เมื่อสหสัมพันธ์เป็นหนึ่ง(unity) ค่าที่แท้จริงจะเท่ากับค่าทำนาย

เราจะใช้อะไรวัดความคลาดเคลื่อนของการทำนาย ความเบี่ยงเบนของค่า Y ที่แท้จริงจากค่าทำนาย (Y') เช่นการวัดจะมีสมบัติตามต้องการถ้ามันถูกกำหนดในรูปของส่วนเบี่ยงเบนมาตรฐาน ความคลาดเคลื่อนมาตรฐานของการประมาณ s_{XY} เป็นชนิดหนึ่งของการวัด เปรียบเทียบสูตรของมันกับสูตรของส่วนเบี่ยงเบนมาตรฐาน

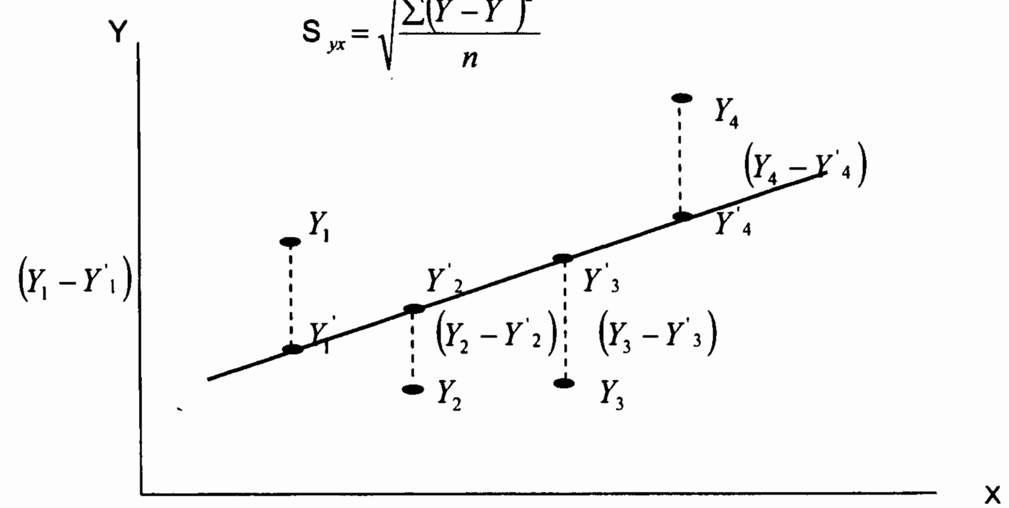
ส่วนเบี่ยงเบนมาตรฐานของ Y

$$s_Y = \sqrt{\frac{\sum (Y-\bar{Y})^2}{n}}$$

ความคลาดเคลื่อนมาตรฐานของ Y บน X

$$s_{YX} = \sqrt{\frac{\sum (Y-Y')^2}{n}} \qquad \text{..........(11.3)}$$

รูป 11.6 แสดงความแตกต่าง *(Y - Y')* ซึ่งสูตร 11.3 เป็นฐาน สิ่งนี้เป็นค่าเดียวกับที่เราเรียก d_Y ในหัวข้อ 11.2 และแสดงในภาพประกอบ 11.2



รูป 11.6

เมื่อสหสัมพันธ์เป็นความบริบูรณ์(perfect) ทุกค่าของ *(Y - Y')* เป็นศูนย์ และเพราะฉะนั้น s_{YX} เป็นศูนย์ ซึ่งสรุปว่าไม่มีข้อคลาดเคลื่อนของการทำนาย เมื่อสหสัมพันธ์เป็นศูนย์ $Y' = \overline{Y}$ สำหรับทุกค่าของ X ดังข้อสังเกตในหัวข้อ11.3 เราอาจแทน Y' ด้วย $\overline{Y}$ ในสูตร 11.3 จะได้ $\sqrt{\frac{\sum(Y-\overline{Y})^2}{n}}$ ซึ่งคือ s_Y ดังนั้นเราสามารถเห็นว่าค่า s_{YX} พิสัยจากศูนย์เมื่อสหสัมพันธ์เป็นบริบูรณ์กับ s_Y เมื่อไม่มีสหสัมพันธ์ทั้งหมด

เราจะแสดงบางลักษณะเหล่านี้โดยใช้สิบคู่ของคะแนนจากตาราง 11.2 จากตาราง 11.2 มีค่า X และ Y ดังตาราง 11.1 และมีค่าต่างที่จะนำมาคำนวณ Y' ดังนี้

$$\overline{X} = \frac{146}{10} = 14.6$$

$$\overline{Y} = \frac{114}{10} = 11.4$$

$$s_x = \sqrt{\sum(X-\overline{X})^2/n} = \sqrt{490.40/10} = 7.017$$

$$s_Y = \sqrt{\sum(Y-\overline{Y})^2/n} = \sqrt{662.40/10} = 8.193$$

$$r = \sum(X-\overline{X})(Y-\overline{Y})/ns_X s_Y = 385.60/\ (10)\ (7.017)\ (8.139) = 0.675$$

ดังนั้น $$Y' = r\frac{s_Y}{s_X}(X-\overline{X}) + \overline{Y}$$

$$= 0.675\frac{8.193}{7.017}(X-14.6)+11.4$$

$$= 0.7881X - 11.5066 + 11.4$$

$$= 0.7881X - 0.1066$$

ตาราง 11.1 คำนวณ s_Y และ s_{YX} (ข้อมูลจากตาราง 10.2)

X	Y	Y - $\bar{Y}$	$(Y-\bar{Y})^2$	Y'	Y - Y'	$(Y-Y')^2$
15	7	-4.4	19.36	11.71	-4.72	22.2784
6	11	-0.4	0.16	4.62	6.38	40.7044
6	5	-6.4	40.96	4.62	0.38	0.1444
18	16	4.6	21.16	14.08	1.92	3.6864
12	9	-2.4	5.76	9.35	-0.35	0.1229
5	0	-11.4	129.96	3.83	-3.83	14.6988
19	3	-8.4	70.56	14.86	-11.86	140.66
20	13	1.6	2.56	15.66	-2.67	7.1289
17	24	12.6	158.76	13.29	10.71	114.70
28	26	14.6	213.16	21.96	4.04	16.3216
$\sum$	114	0.0	662.4	113.98	0.00	360.5391
$\sum/n$	11.4		66.24	11.49		36.0539

คำนวณ s_Y

$$s_Y = \sqrt{\sum(Y-\bar{Y})^2/n} = \sqrt{66.24} = 8.139$$

คำนวณ s_{YX}

$$s_{YX} = \sqrt{\sum(Y-Y')^2/n} = \sqrt{36.0539} = 6.00$$

สำหรับสูตรของ s_{YX} มองดูคล้ายกับสูตรสำหรับหาส่วนเบี่ยงเบนมาตรฐาน ซึ่ง s_{YX} เป็นชนิดหนึ่งของส่วนเบี่ยงเบนมาตรฐาน โดยที่ s_{YX} เป็นส่วนเบี่ยงเบนมาตรฐานของคะแนนที่ได้รับ (obtained) Y รอบค่าทำนายคะแนน Y

เพราะว่า s_{YX} แท้จริงแล้วคือส่วนเบี่ยงเบนมาตรฐาน ซึ่งควรมีสมบัติเช่น ผลบวกของกำลังสองของความเบี่ยงเบนของแต่ละคะแนนจากค่าเฉลี่ย $\sum(Y-\bar{Y})^2$ มีค่าน้อยที่สุด ซึ่งเป็นความคลาดเคลื่อนมาตรฐานของการประมาณด้วย เส้นถดถอยจะต้องก่อให้เกิดค่าน้อยที่สุดของ $\sum(Y-Y')^2$ ซึ่งเป็นผลบวกของกำลังสองของผลต่างระหว่างแต่ละค่าของ Y และค่าที่สมนัยของ Y'

สมบัติข้อที่สองของส่วนเบี่ยงเบนมาตรฐานที่ s_{YX} มีด้วยเช่นกันคือผลบวกของความเบี่ยงเบนของคะแนนที่ไม่ได้ยกกำลังสองจากค่าเฉลี่ยของคะแนนต้องเป็นศูนย์) เมื่อเราดูในสดมภ์ที่สามจะเห็นว่า $\sum(Y-\bar{Y})$ เป็นศูนย์ที่บรรทัดล่างของสดมภ์ ดังนั้น $\sum(Y-Y')$ เป็นศูนย์ในบรรทัดล่างของสดมภ์ที่หก ซึ่งในตอนท้ายบทนี้จะได้แสดงสมบัตินี้ในรูปทั่วไปอีกครั้ง

ความเกี่ยวข้องของ s_{YX} และ s_Y

ความคลาดเคลื่อนมาตรฐานของการประมาณ Y บน X

$$s_{YX} = s_Y\sqrt{1-r^2} \qquad \text{.......... (9.4)}$$

สูตร 11.4 เป็นความสมนัยทางพีชคณิตของสูตร 11.3 สังเกตว่า ถ้า $r = 0$ แล้ว $s_{YX} = s_Y$ และ ถ้า $r = 1$ แล้ว $s_{YX} = 0$ เราสามารถใช้สูตร 11.4 คำนวณ s_{YX} สำหรับข้อมูลของตาราง 11.1

$$\begin{aligned} s_{YX} &= 8.139\sqrt{1-(.675)^2} \\ &= 8.139 \times 0.7378 \\ &= 6.005 \end{aligned}$$

ซึ่งต่างจากคำตอบที่ได้รับจากสูตร 11.3 (ดูตาราง11.1) เล็กน้อยเพราะว่าเกิดจากความคลาดเคลื่อนในการปัดเศษ

ความคลาดเคลื่อนในการประมาณ Y จาก X

ถ้าเรามีเหตุผลพอที่จะตั้งสมมุติฐานว่าคะแนนที่แท้จริงบน Y เป็นการแจกแจงปกติจากคะแนนทำนาย เราสามารถใช้ความรู้ของโค้งปกติตอบคำถามเกี่ยวกับความคลาดเคลื่อนของการทำนาย ยกตัวอย่าง การใช้หลักในการอธิบายในบทที่ผ่านมาและข้อสมมุติฐานของโค้งปกติ เราสามารถตรวจสอบว่า

68% ของคะแนน Y ตกระหว่างช่วง $\bar{Y} \pm 1.00 s_Y$

95% ของคะแนน Y ตกระหว่างช่วง $\bar{Y} \pm 1.96 s_Y$

99% ของคะแนน Y ตกระหว่างช่วง $\bar{Y} \pm 2.58 s_Y$

เพราะว่าคะแนนทำนาย Y' เป็นชนิดของค่าเฉลี่ยและความคลาดเคลื่อนมาตรฐานของการประมาณ s_{YX} ซึ่งเป็นชนิดหนึ่งของส่วนเบี่ยงเบนมาตรฐาน เป็นจริงด้วยในการแจกแจงสองตัวแปรปกติ (normal bivariate distribution) ที่

68% ของค่า Y ที่แท้จริงตกระหว่างช่วง $\bar{Y} \pm 1.00 s_{YX}$

95% ของค่า Y ที่แท้จริงตกระหว่างช่วง $\bar{Y} \pm 1.96 s_{YX}$

99% ของค่า Y ที่แท้จริงตกระหว่างช่วง $\bar{Y} \pm 2.58 s_{YX}$

จากที่กำหนดในตอนต้น

X คือคะแนนวิชาคณิตศาสตร์ Y คือคะแนนวิชาฟิสิกส์

$\bar{X} = 64$ $\bar{Y} = 68$

$s_X = 14$ $s_Y = 16$

$r = .38$

หา s_{YX} จะได้

$$s_{YX} = s_Y\sqrt{1 - r^2}$$

$$= 16\sqrt{1 - (.38)^2}$$

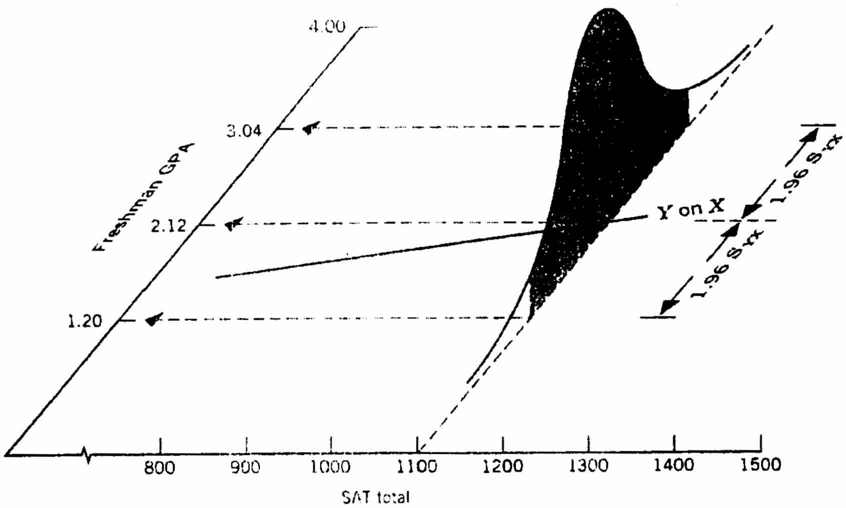
$$= 16(.9250)$$

$$= 14.8$$

ดังนั้น 95% ของคะแนนวิชาฟิสิกส์ที่แท้จริงอยู่ในช่วง

ขอบล่าง (lower limit) : $Y' - 1.96 s_{YX} = 68 - (1.96)(14.8) = 68 - 29.00 = 39$

ขอบบน (upper limit) : $Y' + 1.96 s_{YX} = 68 + (1.96)(14.8) = 68 + 29.00 = 97$



รูปที่ 11.7

หลักที่กล่าวมาแล้วสามารถใช้ตอบคำถาม เช่นสัดส่วนเท่าใดของนักเรียนที่ได้คะแนนคณิตศาสตร์ 67 คะแนนอาจคาดหวังว่าจะได้คะแนนฟิสิกส์ 65 หรือต่ำกว่า ในการแก้ปัญหานี้ ขั้นที่หนึ่งหาค่าทำนายของ Y เช่นที่เราทำในหัวข้อ 4 เพื่อหา Y' จะได้ Y' = 69.30

เราแทนค่าเฉลี่ย (ของนักเรียนที่ได้คะแนนคณิตศาสตร์เป็น 67) เป็น 69.30 และส่วนเบี่ยงเบนมาตรฐาน s_{YX} เป็น 14.8

จาก $z = \dfrac{Y - Y'}{s_{YX}}$

$= \dfrac{65 - 69.3}{14.8}$

$= -0.2905$

เปิดตารางจะได้พื้นที่ที่อยู่ต่ำกว่า z = -.2905 คือ 0.3857 นั่นคือผู้ที่ได้คะแนนคณิตศาสตร์ 67 คะแนน มีโอกาสที่จะได้คะแนนฟิสิกส์ 65 คะแนนหรือต่ำกว่าอยู่ 38.57%

ข้อควรระวังที่เกี่ยวข้องกับการประมาณของความคลาดเคลื่อนจากการทำนาย

เงื่อนไขต่าง ๆเป็นสิ่งที่พบเห็นมาแล้วในบทก่อนหน้านี้

1. ความสัมพันธ์ระหว่างสองตัวแปรต้องเป็นเชิงเส้น

2. ความเบี่ยงเบนของ Y ที่แท้จริงรอบๆ Y' ต้องเหมือนกันสำหรับทุกค่าของ X นั่นคือ ตั้งสมมุติฐานว่า การเบี่ยงเบนใน Y เหมือนกันจากสดมภ์สู่สดมภ์ ในการนิยามสูตรสำหรับความคลาดเคลื่อนมาตรฐานของการประมาณ $s_{YX} = \sqrt{(Y-Y')^2/n}$ เราสามารถเห็นว่า s_{YX} เป็นฟังก์ชันของขนาดเฉลี่ยของระยะทางกำลังสองของความแตกต่างระหว่างแต่ละค่า Y และ Y' ที่สมนัย s_{YX} จะประมาณส่วนเบี่ยงเบนมาตรฐานของค่า Y เพียงค่ากลางของ X มันจะเป็นความคลาดเคลื่อนจากการประมาณมากเกินไป ใน Y สำหรับค่าต่ำของ X และประมาณน้อยเกินไปสำหรับค่าสูง

3. คะแนน Y ที่แท้จริงต้องเป็นการแจกแจงปกติสำหรับทุกค่าของ X โปรดจำไว้ว่าสำหรับปัญหาคล้ายกับหัวข้อที่กล่าวมาแล้ว เราใช้ตารางภายใต้โค้งปกติเพื่อหาสมบัติของกรณีคาดหมายตกเหนือหรือต่ำกว่าจุดที่กำหนดให้

ข้อเตือนใจอีกข้อหนึ่งคือ เนื่องจากในการทำนายสถานการณ์ในชีวิตจริงเราจะเกี่ยวข้องกับคะแนนของกลุ่มตัวอย่าง จากประชากร ไม่ใช่คะแนนของประชากร(เพราะถ้ารู้คะแนนของประชากรเราจะต้องทำนายทำไม) ดังนั้นค่าสหสัมพันธ์ที่แท้จริง เส้นถดถอยและความคลาดเคลื่อนมาตรฐานของการประมาณจึงอาจแตกต่างไปจากกลุ่มตัวอย่าง และขึ้นอยู่กับความเบี่ยงเบนของการชักกลุ่มตัวอย่างโดยสุ่ม ถ้าขนาดของกลุ่มตัวอย่างเป็น 100 และเราใช้กระบวนการที่ถูกต้อง ช่วงคาดหมายที่จะบรรจุ 95% ของค่า Y ที่แท้จริงจะกว้างกว่าสิ่งที่เราต้องการประมาณ 4% ถ้าในกรณีที่กลุ่มตัวอย่างน้อยกว่า ความคลาดเคลื่อนจะมากขึ้น ในรูปทั่วไปการทำนายและการประมาณความคลาดเคลื่อนจะทำได้ดีที่สุดเมื่อขนาดของกลุ่มตัวอย่างใหญ่พอที่จะลดขนาดของความคลาดเคลื่อน

นอกจากนั้นเราควรทราบว่า : มีสมการเส้นถดถอยที่แท้จริงสองสมการ สมการหนึ่งสำหรับทำนาย Y จาก X และอีกสมการสำหรับทำนาย X จาก Y

ข้อสังเกตทางคณิตศาสตร์

แทนคะแนนดิบที่สมมูลกับ Y' ใน $\sum(Y-Y')$

$$\sum(Y-Y') = \sum\left[Y-\left(r\frac{s_Y}{s_X}\right)X+\left(r\frac{s_Y}{s_X}\right)\overline{X}-\overline{Y}\right]$$

$$= \sum\left[(Y-\overline{Y})-\left(r\frac{s_Y}{s_X}\right)(X-\overline{X})\right]$$

$$= \sum(Y-\overline{Y})-r\frac{s_Y}{s_X}\sum(X-\overline{X})$$

เพราะว่า $\sum(Y-\overline{Y}) = 0$ และ $\sum(X-\overline{X}) = 0$

แบบฝึกหัด 11

1. ถ้า r = 1.00 อะไรคือค่าของ $\sum d_Y^2$ ทำไม
2. ถ้า r = -1.00 อะไรคือค่าของ $\sum d_Y^2$ ทำไม
3. สมมุติค่าสหสัมพันธ์ระหว่าง วิชาคณิตศาสตร์และวิทยาศาสตร์ระดับมหาวิทยาลัยปีหนึ่งเป็น r = +.50 จงหาคะแนน z บนวิชาคณิตศาสตร์ที่คุณทำนายสำหรับบางคน เมื่อคะแนน z บนวิชาฟิสิกส์เป็น

   3.1 +3.00　　3.2 -2.00　　3.3 0.00

   3.4 -1.00　　3.5 +1.00　　3.6 0.5
4. สมมุติ X คือคะแนนคณิตศาสตร์ ซึ่ง $\overline{X}$ = 64 และ S_X = 8 ในขณะที่ Y คือคะแนนฟิสิกส์ ซึ่ง $\overline{Y}$ = 58 และ S_Y = 11 จงใช้สูตรคะแนนมาตรฐานสำหรับเส้นถดถอยของ Y บน X เพื่อทำนาย Y' (คะแนนดิบ) สำหรับค่าต่อไปนี้ของ X สมมุติ r = .50 (ในการเขียนคำตอบให้แสดงทั้งคะแนน z และคะแนนดิบ)

   4.1 58　　4.2 69　　4.3 74　　4.4 52
5. สมมุติ $\overline{X}$ = 50 และ S_X = 10 $\overline{Y}$ = 200 และ S_Y = 30 จงหาค่าของr สำหรับแต่ละข้อต่อไปนี้

   5.1 สำหรับ X = 40 , สมการถดถอยการทำนาย Y' = 170

   5.2 สำหรับ X = 40 , สมการถดถอยการทำนาย Y' = 230

   5.3 สำหรับ X = 60 , Y' = 215
6. สมมุติ $\overline{X}$ = 55 และ s_X = 12 $\overline{Y}$ = 118 และ s_Y = 27 สำหรับกรณีที่ X เป็น 55 สมการทำนายของ Y' = 118 นักศึกษาจะบอกค่าของ r ได้หรือไม่ จงอธิบาย
7. จากสมการเส้นถดถอย เมื่อ X = 81, Y' = 317 เมื่อ X = 92, Y' = 317

   7.1 จงหาค่า r

   7.2 จงหา $\overline{Y}$