บทที่ 12
การถดถอยเชิงเส้นและสหสัมพันธ์
(Linear Regression and Correlation)

12.1 การถดถอย

การถดถอยเป็นวิธีการอันหนึ่งที่จะใช้เป็นเครื่องมือตรวจหาลักษณะธรรมชาติของความสัมพันธ์ระหว่างตัวแปร 2 ตัวแปร ขึ้นไป ทั้งนี้เพื่อประโยชน์ในการประมาณค่าที่เราไม่ทราบโดยอาศัยค่าสังเกตหรือข้อมูลจากปรากฏการณ์ที่ผ่านมา อย่างเช่น การศึกษาอิทธิพลของปุ๋ยที่ใช้ใน 1 ไร่ ของข้าวสาลี (X) ต่อผลผลิตที่ได้ใน 1 ไร่ (Y) สิ่งที่เราต้องการทราบคืออิทธิพลของปุ๋ย (X) ที่มีต่อผลผลิต (Y) ในที่นี้ X เป็นตัวแปรอิสระ Y เป็นตัวแปรตามสังเกตว่าตามความเป็นจริงแล้วผลผลิตข้าวสาลี (Y) ไม่ได้ขึ้นอยู่กับจำนวนปุ๋ย (X_1) เพียงอย่างเดียว ยังขึ้นอยู่กับตัวแปรอิสระอื่น ๆ เช่น ดินฟ้าอากาศ (X_2) พันธุ์ (X_3) น้ำ (X_4) การดูแลรักษา (X_5) ฯลฯ แต่ถ้าเราพิจารณาอิทธิพลเพียงปัจจัยเดียวคือตัวแปรอิสระเพียงตัวเดียว เรียกการถดถอยนั้นว่าการถดถอยอย่างง่าย เมื่อใดที่เราพิจารณาอิทธิพลของปัจจัยมากกว่าหนึ่งตัว การถดถอยนั้นเรียกว่าการถดถอยเชิงซ้อน

ความสัมพันธ์ของ X กับ Y อาจเป็นรูปเส้นตรงหรือเส้นโค้งฯ แต่ในที่นี้เราจะอธิบายเฉพาะความสัมพันธ์ในรูปของเส้นตรงที่เรียกว่า "การถดถอยอย่างง่ายที่เป็นเส้นตรง" การศึกษาเรื่องการถดถอยนั้น เราจะต้องอาศัยความรู้ในอดีตหรือรายละเอียดเกี่ยวกับข้อมูลที่จะนำมาวิเคราะห์เพื่อที่จะพิจารณาให้เห็นว่า ข้อมูลชุดใดเป็นตัวแปรอิสระ (X) หรือเป็นสาเหตุ ข้อมูลชุดใดเป็นตัวแปรตาม (Y) หรือเป็นผล

12.2 การถดถอยอย่างง่ายที่เป็นเส้นตรง

การถดถอยอย่างง่ายเป็นเรื่องของการวิเคราะห์ความสัมพันธ์ของตัวแปร 2 ชุด โดยที่ตัวแปรชุดหนึ่งเป็นตัวแปรอิสระหรือสาเหตุเราใช้อักษร X_i ซึ่งเป็นตัวที่มีอิทธิพลต่อตัวแปรอีกชุดหนึ่ง เป็นตัวแปรตามหรือผลใช้อักษร Y_i ค่าของ Y_i จะมากหรือน้อยหรือเปลี่ยนแปลงไปในรูปใดต้องขึ้นอยู่กับค่าของ X_i โดยเราจะศึกษาว่า ถ้า X_i เปลี่ยนไป 1 หน่วย จะทำให้ค่าของ Y_i เปลี่ยนไปกี่หน่วย อย่างเช่น จำนวนเงินรายได้กับจำนวนเงินออมทรัพย์ ผู้มีรายได้มากย่อมมีโอกาสออมทรัพย์ได้

มาก ผู้มีรายได้น้อยย่อมมีโอกาสออมทรัพย์ได้น้อย จำนวนเงินรายได้เป็น X_i จำนวนออมทรัพย์เป็น Y_i หรือเรามักจะได้ยินอยู่เสมอว่าผู้มีรายได้มาก X_i ย่อมมีรายจ่ายมาก Y_i เป็นเงาตามตัว

การวิเคราะห์การถดถอยและสหสัมพันธ์นี้ถ้าเราวิเคราะห์โดยใช้ข้อมูล 2 ชุด ชุดหนึ่งเป็นสาเหตุและอีกชุดหนึ่งเป็นผล เราถือว่าเป็นการวิเคราะห์แบบการถดถอยอย่างง่าย และสหสัมพันธ์อย่างง่าย เช่น ปุ๋ยเป็นสาเหตุหรือตัวแปรอิสระกับผลิตผลเป็นผลหรือตัวแปรตาม ความตั้งใจเรียนเป็นสาเหตุหรือตัวแปรอิสระกับผลสอบเป็นผลหรือตัวแปรตาม แต่ถ้าเราวิเคราะห์การถดถอยหรือสหสัมพันธ์ โดยใช้ข้อมูลมากกว่า 2 ชุด โดยที่ให้ชุดหนึ่งเป็นผลและชุดอื่น ๆ เป็นสาเหตุที่มีมากกว่า[1] สาเหตุเราถือว่าเป็นการวิเคราะห์การถดถอยเชิงซ้อนหรือสหสัมพันธ์เชิงซ้อนดังตัวอย่าง เราต้องการวิเคราะห์ความสัมพันธ์ระหว่างผลิตผลของดอกกล้วยไม้เพื่อดูว่า จะมีปริมาณมากน้อยเท่าใด ปริมาณเพิ่มขึ้นหรือลดลงเท่าใดขึ้นอยู่กับสาเหตุใดบ้าง ซึ่งสาเหตุดังกล่าวอาจจะเป็น ปุ๋ย พันธุ์ น้ำ การรักษาดูแล ดินฟ้าอากาศ ฯลฯ

การวิเคราะห์การถดถอยและสหสัมพันธ์จะต้องคำนึงถึงรูปลักษณะของเส้นที่แสดงความสัมพันธ์ที่คาดว่า ความสัมพันธ์จะเป็นไปในรูป ซึ่งอาจจะเป็นเส้นตรงหรือเส้นโค้ง เพื่อที่จะเลือกแบบของสมการที่จะนำมาเป็นหลักในการวิเคราะห์ได้ถูกต้อง จึงพอสรุปได้ว่า ในเนื้อเรื่องของคำบรรยายเกี่ยวกับการวิเคราะห์การถดถอยและสหสัมพันธ์นี้อาจจะแยกหัวข้อได้ดังนี้

การวิเคราะห์การถดถอยแบ่งออกเป็น

(1) การวิเคราะห์การถดถอยอย่างง่าย
  (ก) การถดถอยอย่างง่ายที่เป็นเส้นตรง
  (ข) การถดถอยอย่างง่ายที่ไม่เป็นเส้นตรง
(2) การวิเคราะห์การถดถอยเชิงซ้อน
  (ก) การถดถอยเชิงเส้นที่มีหลาย ๆ ตัวแปรอิสระ
  (ข) การถดถอยไม่เชิงเส้นที่มีหลาย ๆ ตัวแปรอิสระ

การวิเคราะห์สหสัมพันธ์แบ่งออกเป็น

(1) การวิเคราะห์สหสัมพันธ์อย่างง่ายที่เป็นเส้นตรง
  (ก) สหสัมพันธ์อย่างง่ายที่เป็นเส้นตรง
  (ข) สหสัมพันธ์อย่างง่ายที่ไม่เป็นเส้นตรง
(2) การวิเคราะห์สหสัมพันธ์เชิงซ้อน

(ก) สหสัมพันธ์เชิงเส้นที่มีหลาย ๆ ตัวแปรอิสระ

(ข) สหสัมพันธ์ไม่เชิงเส้นที่มีหลาย ๆ ตัวแปรอิสระ

แต่สำหรับในบทนี้จะอธิบายแต่วิธีวิเคราะห์การถดถอยอย่างง่ายที่เป็นเส้นตรงและวิธีวิธีวิเคราะห์สหสัมพันธ์อย่างง่ายที่เป็นเส้นตรงเท่านั้น

การวิเคราะห์การถดถอยอย่างง่ายที่เป็นเส้นตรง เรามีตัวแบบทางคณิตศาสตร์หรือสมการถดถอยของประชากร (Population Regression) ซึ่งเป็นสมการที่แสดงถึงความสัมพันธ์ที่แท้จริงของตัวแปร 2 ตัว สมการที่แสดงถึงความสัมพันธ์ที่แท้จริงของการถดถอยเชิงเส้นของประชากรเขียนได้เป็น

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

ในที่นี้ ค่าของ Y แต่ละค่าจะประกอบด้วยส่วนใหญ่ ๆ 2 ส่วน คือ $\beta_0 + \beta_1 X$ กับ ε

$\beta_0 + \beta_1 X$ หมายถึงค่าที่คาดหวังว่า Y ควรจะเป็นโดยผลของ X โดยที่

β_0 เป็นจุดตัดบนแกน Y เมื่อ X มีค่าเป็นศูนย์

β_1 เป็นความชันของเส้นถดถอย

ε เป็นค่าคาดหวังที่ทำให้ Y คลาดเคลื่อนไปจากสาเหตุอื่นนอกเหนือไปจาก X

แต่ในทางปฏิบัติ เราไม่สามารถสังเกตค่าของประชากรได้หมด นั่นคือเราไม่สามารถสังเกตความสัมพันธ์จริงนี้ได้ สิ่งที่เราจะทำได้ก็เพียงหาค่าประมาณเพื่อที่จะเป็นตัวแทนของประชากร ตัวประมาณค่าที่ดีนั้นจะต้องมีคุณสมบัติคือไม่มีความเอนเอียงมีความแปรปรวนน้อยที่สุด วิธีการกำลังสองน้อยที่สุดเป็นวิธีการที่เข้ากับคุณสมบัติข้างต้นในการกำหนดเส้นตรงกับข้อมูล สมการของการถดถอยเชิงเส้นเขียนได้ดังนี้

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + e$$

ค่าของ Y แต่ละตัวจะประกอบด้วย $\hat{\beta}_0 + \hat{\beta}_1 X$ กับ e ซึ่งมีความหมายเช่นเดียวกับสมการถดถอยของประชากร คือ $\hat{\beta}_0 + \hat{\beta}_1 X$ แสดงถึงส่วนหนึ่งของค่า Y ที่คาดหวังจากผลของ X

$\hat{\beta}_0$ เป็นจุดตัดบนแกน Y ซึ่งเป็นตัวค่าประมาณของ β_0

$\hat{\beta}_1$ เป็นความชันของเส้นถดถอยซึ่งเป็นตัวค่าประมาณของ β_1

และ e เป็นค่าความแตกต่างหรือค่าคาดเคลื่อนระหว่างค่า Y กับเส้นถดถอยซึ่งเกิดขึ้นเนื่องมาจากสาเหตุอื่นที่นอกเหนือไปจาก X และเป็นตัวค่าประมาณของ ε

จาก $Y = \hat{\beta}_0 + \hat{\beta}_1 X + e$

$Y - e = \hat{\beta}_0 + \hat{\beta}_1 X$

ให้ $Y - e$ เป็น $\hat{Y}$ ซึ่งหมายถึงตัวค่าประมาณของ

$\therefore \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

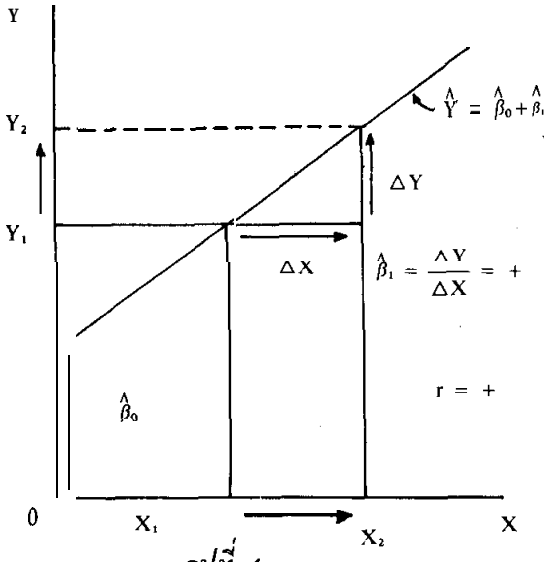
เพราะฉะนั้น สมการถดถอยอย่างง่ายที่เป็นเส้นตรงของตัวอย่างที่เราต้องการวิเคราะห์คือสมการข้างต้น

สมการ $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ อาจเรียกว่าสมการเส้นถดถอย Y on X, $\hat{Y}$ อาจเขียน Y_c, Y' ได้เช่นเดียวกัน

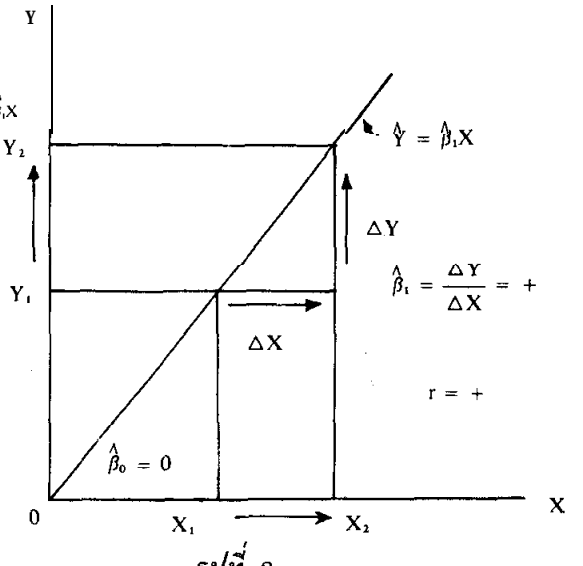
$\hat{\beta}_0$ เป็นจุดตัดบนแกน Y มีค่าอาจเป็น +, − หรือ 0

$\hat{\beta}_1$ เป็นตัววัดค่าความชันของเส้นถดถอย เรียกว่าสัมประสิทธิ์ การถดถอยค่า $\hat{\beta}_1$ นี้อาจจะเป็น +, − หรือ 0

ถ้า $\hat{\beta}_1$ มีค่าเป็นบวกแสดงว่าค่าของ X มีผลที่ทำให้ Y มีค่าเปลี่ยนแปลงไปในทางเดียวกัน คือ X มีค่าเพิ่ม Y จะมีค่าเพิ่มตาม และถ้า X มีค่าลดลง Y ก็จะมีค่าลดตาม ถ้าเขียนกราฟเส้นถดถอย จะได้ดังภาพ

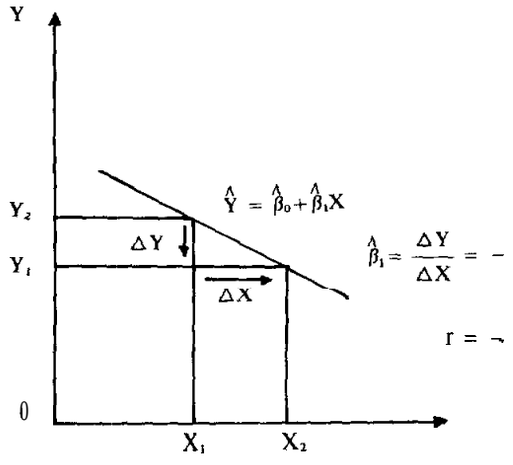


รูปที่ 1

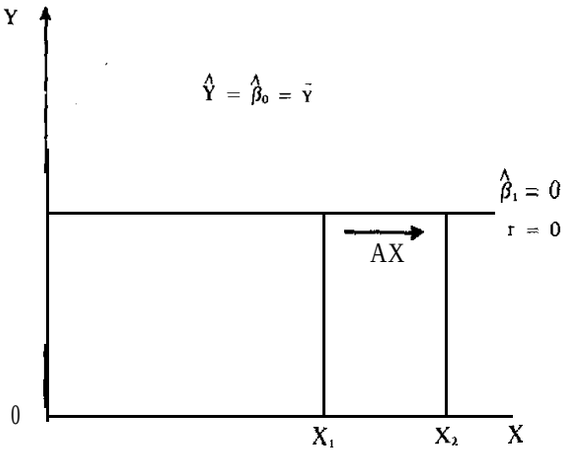


รูปที่ 2

ถ้า $\hat{\beta}_1$ มีค่าเป็นลบแสดงว่าค่าของ X มีผลที่ทำให้ Y มีค่าเปลี่ยนแปลงไปในทางตรงข้ามกันคือ X มีค่าเพิ่มขึ้น Y จะมีค่าลดลงหรือ X มีค่าลดลง Y จะมีค่าเพิ่มขึ้น ถ้าเขียนกราฟเส้นถดถอยจะได้ดังภาพ



รูปที่ 3



รูปที่ 4

ถ้า $\hat{\beta}_1 = 0$ แสดงว่า X ไม่มีผลต่อ Y และ กล่าวคือ X จะเปลี่ยนอย่างไรก็ไม่ทำให้ Y มีการเปลี่ยนแปลงเลย ถ้าเราเขียนกราฟเส้นถดถอยจะได้ดังรูปที่ 4

สมการเส้นถดถอยมีสองชนิดด้วยกัน หนึ่ง สำหรับทำนาย Y จาก X สมการเขียนได้เป็น $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ สองสำหรับทำนาย X จาก Y สมการเขียนได้เป็น $\hat{X} = \hat{\beta}_0' + \hat{\beta}_1' Y$ โดยทั่ว ๆ ไป เรามักจะพบสมการที่หนึ่ง ซึ่งใช้กันมากดังรายละเอียดจะกล่าวต่อไป

12.3 คุณสมบัติของการถดถอยเชิงเส้นที่คำนวณได้โดยวิธีการกำลังสองน้อยที่สุด

(1) $\bar{X}$, $\bar{Y}$ จะอยู่บนเส้นถดถอยหรือเส้นถดถอยจะผ่านจุด $(\bar{X}, \bar{Y})$

(2) ผลบวกของส่วนเบี่ยงเบนจากเส้นถดถอยจะมีค่าเป็นศูนย์ $\Sigma (Y - \hat{Y}) = 0$

(3) ผลบวกของส่วนเบี่ยงเบนจากเส้นถดถอยกำลังสอง $\Sigma(Y - \hat{Y})^2$ มีค่าน้อยที่สุด จากคุณสมบัติข้อที่ 3 ถ้าเรา differentiate เทียบกับ $\hat{\beta}_0$ และ $\hat{\beta}_1$ แล้วทำให้เท่ากับศูนย์ จะได้สมการปกติสำหรับหาค่า $\hat{\beta}_0$, $\hat{\beta}_1$ ที่จะทำให้ $\Sigma(Y - \hat{Y})^2$ มีค่าน้อยที่สุด สมการปกตินั้นเขียนได้เป็น

$$n\hat{\beta}_0 + \hat{\beta}_1 \Sigma X = \Sigma Y \quad \text{.....(1)}$$

$$\hat{\beta}_0 \Sigma X + \hat{\beta}_1 \Sigma X^2 = \Sigma XY \quad \text{.....(2)}$$

ΣX, ΣX^2, ΣXY และ ΣY คำนวณหาได้จากข้อมูลที่เราได้มาแล้ว นำไปแทนค่าลงในสมการ (1) และ (2) จากนั้นก็ทำการแก้คำไขหาค่า $\hat{\beta}_0$ และ $\hat{\beta}_1$ ได้โดยวิธีการทางพีชคณิตได้

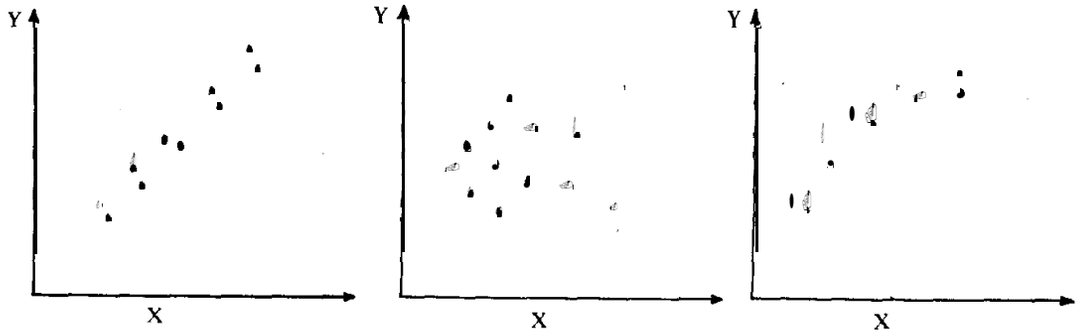
$$\hat{\beta}_1 = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2} = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{\Sigma (X - \bar{X})^2}$$

และ

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

เมื่อคำนวณหาค่า $\hat{\beta}_0$, $\hat{\beta}_1$ ได้แล้วก็นำไปแทนค่าในสมการเส้นถดถอย $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ เมื่อต้องการประมาณค่า Y ก็หาค่า $\hat{Y}$ โดยแทนค่า X ลงในสมการนี้ ค่า $\hat{\beta}_0$ เป็นจุดตัดแกน Y เมื่อ X มีค่าเป็นศูนย์ ส่วนค่า $\hat{\beta}_1$ เป็นความชันหรืออัตราการเปลี่ยนแปลงระหว่าง X กับ Y คือ เมื่อ X เปลี่ยนแปลงไป 1 หน่วย Y จะเปลี่ยนแปลงไป 1 หน่วย ดูจาก รูปที่ 1 $\hat{\beta}_1 = \frac{\Delta Y}{\Delta X}$ ถ้า $\Delta X = 1$ ค่า $\hat{\beta}_1$ จะเท่ากับ ΔY นั่นคือการเปลี่ยนแปลงของ Y จะมีค่าเท่ากับ $\hat{\beta}_1$ หน่วยในทางเพิ่มขึ้น เมื่อการเปลี่ยนแปลงของ $X = 1$ หน่วย แต่ถ้ามาดูรูปที่ 3 การเปลี่ยนแปลงของ Y จะมีค่าเท่ากับ $\hat{\beta}_1$ หน่วยในทางลดลง เมื่อการเปลี่ยนแปลงของ $X = 1$ หน่วย

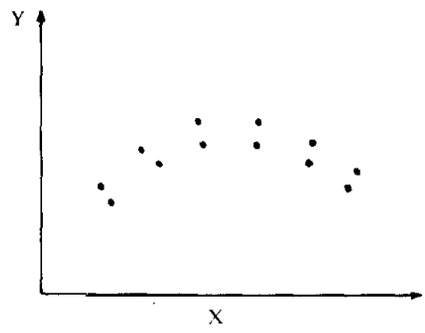
12.4 ลักษณะตัวแบบที่ใช้กำหนดกับข้อมูล



กรณีที่ 1 ใช้ตัวแบบ $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

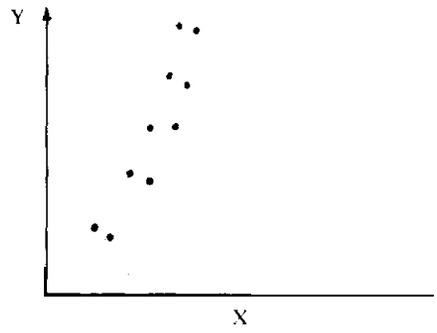
กรณีที่ 2 ใช้ตัวแบบ $\hat{Y} = \bar{Y}$

กรณีที่ 3 พยายามใช้ตัวแบบ $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_{11} X^2$



กรณีที 4

พยายามใช้ตัวแบบ $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_{11} X^2$



กรณีที 5

พยายามใช้ตัวแบบ $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_{11} X^2$

ตัวอย่างการวิเคราะห์การถดถอยอย่างง่ายที่เป็นเส้นตรง

ในการสำรวจนักศึกษา 12 คน เกี่ยวกับความสัมพันธ์ว่าคะแนนทดสอบเชาวน์ปัญญามีอิทธิพลต่อเกรดเฉลี่ยมากน้อยแค่ไหน ดังตาราง

เกรดเฉลี่ย	2.1	2.2	3.1	2.3	3.4	2.9	2.9	2.7	2.1	1.7	3.3	3.5
เชาวน์ปัญญา	116	129	123	121	131	134	126	122	114	118	132	129

วิธีทำ สร้างตารางเพื่อคำนวณหา ΣX, ΣY, ΣX^2, ΣXY แทนลงในสูตรเพื่อหาค่า $\hat{\beta}_0$, $\hat{\beta}_1$

จากสูตร

$$\hat{\beta}_1 = \frac{n\Sigma XY \quad (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

ตัวแบบ

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

เกรดเฉลี่ย (Y)	เชาวน์ปัญญา (X)	X^2	Y^2	XY	$\hat{Y}$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
2.1	116	13456	4.41	243.6	2.1192	−0.0192	.0004
2.2	129	16641	4.84	283.8	2.9733	−0.7733	.5929
3.1	123	15129	9.61	381.3	2.5791	0.5209	.2704
2.3	121	14641	5.29	278.3	2.4477	−0.1477	.0225
3.4	131	17161	11.56	445.4	3.1047	0.2953	.0900
2.9	134	17956	8.41	388.6	3.3018	−0.4018	.1600
2.9	126	15876	8.41	365.4	2.7762	0.1238	.0144
2.7	122	14884	7.29	329.4	2.5134	0.1866	.0361
2.1	114	12996	4.41	239.4	1.9878	0.1122	.0121
1.7	118	13924	2.89	200.6	2.2506	−0.5506	.3025
3.3	132	17424	10.89	435.6	3.1704	0.1296	.0169
3.5	129	16641	12.25	451.5	2.9733	0.5267	.2769
32.2	1495	186729	90.26	4042.9			2.0651

ในที่นี้เราได้ $\Sigma X = 1495$, $\Sigma Y = 32.2$. $\Sigma XY = 4042.9$ แทนลงในสูตร

$$\hat{\beta}_1 = \frac{12(4042.9) - (1495)(32.2)}{12(186729) - (1495)^2}$$

$$= \frac{48514.8 - 48139}{2240748 - 2235025} = \frac{375.8}{5723}$$

$$= 0.0657$$

$$\hat{\beta}_0 = \frac{32.2}{12} - (0.0657)\left(\frac{1495}{12}\right)$$

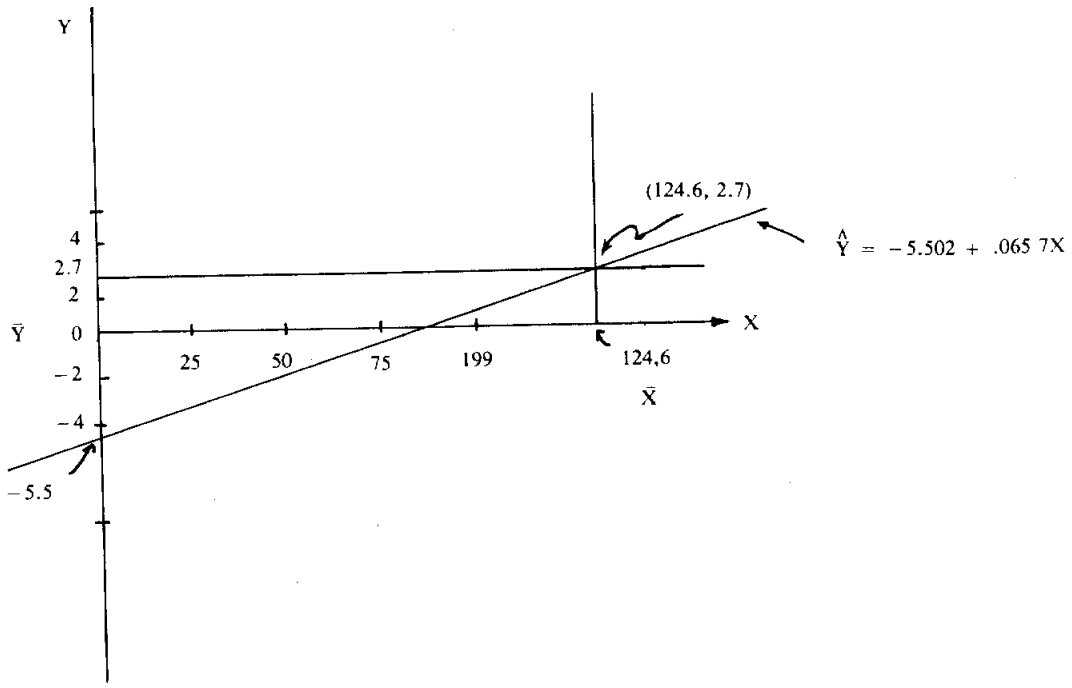
$$= 2.683 - 8.185$$

$$= -5.502$$

แทนค่า $\hat{\beta}_0$ และ $\hat{\beta}_1$ ลงในตัวแบบ

$\therefore$ สมการถดถอย $\hat{Y} = -5.502 + 0.0657X$

จะเห็นว่าค่า $\hat{\beta}_1$ แสดงถึงความสัมพันธ์ของ X กับ Y $\hat{\beta}_1$ มีค่าเป็นบวก แสดงว่าความสัมพันธ์ของ X กับ Y ไปทางเดียวกัน $\hat{\beta}_1 = 0.0657$ หมายความว่าเกรดเฉลี่ย (Y) จะเพิ่มขึ้นโดยเฉลี่ย 0.0657 หน่วยต่อเชาวน์ปัญญา (X) ที่เพิ่มขึ้น 1 หน่วย ถ้าหากว่านักศึกษามีเชาวน์ปัญญา (X) 140 และเราใช้สมการ $\hat{Y} = -5.502 + 0.0657X$ ประมาณค่า Y คือ $\hat{Y}$ จะมีค่าเท่ากับ $-5.502 + 0.0657(140) = 3.696$ หน่วย อาจจะมีความคลาดเคลื่อนไปจากความจริงที่ควรเป็น เพราะเกรดเฉลี่ยของนักศึกษาไม่ได้มีผลมาจากเชาวน์ปัญญาอย่างเดียว ยังมีสาเหตุที่มีผลต่อเกรดเฉลี่ยอีกหลายสาเหตุ ทั้งนี้ต้องเข้าใจว่าตราบใดที่เราใช้สมการถดถอย $\hat{Y} = -5.502 + 0.0657X$ เป็นหลักในการคาดคะเนเกรดเฉลี่ย เราคำนึงถึงอิทธิพลของเชาวน์ปัญญาเพียงอย่างเดียว ส่วนค่าคลาดเคลื่อนจากสาเหตุอื่นนั้น ให้นักศึกษาดูตารางคอลัมน์ที่ 7



ถ้าเราพลอตกราฟเพื่อพิจารณาเส้นถดถอยของตัวอย่างนี้ จะได้รูปข้างต้น จากรูป $\hat{\beta}_0 = -5.502$ แสดงถึงเส้นถดถอยตัดแกน Y ที่จุด -5.502

12.5 ความแปรปรวนของ Y

จากสมการที่แสดงถึงความสัมพันธ์ที่แท้จริงของการถดถอยเชิงเส้นของประชากร $Y = \beta_0 + \beta_1 X + \varepsilon$ โดยที่ ε เป็นค่าคลาดเคลื่อน มีมัชฌิมเลขคณิตศูนย์และความแปรปรวน $\sigma^2_{y/x}$ (ไม่ทราบค่า) เราจึงใช้ตัวประมาณค่า $s^2_{y/x}$ แทนค่า $\sigma^2_{y/x}$ และคำนวณหาได้จากสูตร

$$s^2_{y/x} = \frac{\Sigma(Y - \hat{Y})^2}{n - 2}$$

หรือ

$$s^2_{y/x} = \frac{1}{n-2}\left[\Sigma Y^2 - n\bar{Y}^2 - \frac{(\Sigma XY - nX\bar{Y})^2}{\Sigma X^2 - n\bar{X}^2}\right]$$

12.6 มัชฌิมเลขคณิตและความแปรปรวนของ $\hat{\beta}_0$ และ $\hat{\beta}_1$

เราทราบแล้วว่า $\hat{\beta}_0$ เป็นตัวประมาณค่าของ β_0 และ $\hat{\beta}_1$ เป็นตัวประมาณค่าของ β_1 ทั้ง $\hat{\beta}_0$ และ $\hat{\beta}_1$ ต่างก็เป็นตัวประมาณค่าที่ดีและปราศจากความเอนเอียง นั่นคือ

$$E(\hat{\beta}_0) = \beta_0$$

$$V(\hat{\beta}_0) = \frac{\Sigma X^2 \sigma^2_{y/x}}{n\Sigma(X - \bar{X})^2}$$

$$E(\hat{\beta}_1) = \hat{\beta}_1$$

$$V(\hat{\beta}_1) = \frac{\sigma^2_{y/x}}{\Sigma(X - \bar{X})^2}$$

เรามาพิสูจน์ค่าคาดหวังและค่าความแปรปรวนของ $\hat{\beta}_1$ เท่านั้น จากสูตร

$$\hat{\beta}_1 = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(X - \bar{X})^2}$$

$$= \frac{\Sigma(X_i - \bar{X})Y_i}{\Sigma(X_i - \bar{X})^2}$$

เนื่องจากเทอม $\Sigma(X_i - \bar{X})\bar{Y} = \bar{Y}\Sigma(X - \bar{X}) = 0$

ดังนั้น $\hat{\beta}_1 = \{(X_1 - \bar{X})Y_1 + (X_2 - \bar{X})Y_2 + , .. + (X_n - \bar{X})Y_n\} / \Sigma(X - \bar{X})^2$

ให้ $a_i = (X_i - \bar{X}) / \Sigma(X_i - \bar{X})^2$ และเป็นตัวคงที่

$$\hat{\beta}_1 = a_1Y_1 + a_2Y_2 + \ldots + a_nY_n$$

$$= \Sigma a_iY_i$$

ค่าคาดหวังของ $\hat{\beta}_1 = \Sigma a_iY_i$

$$E(\hat{\beta}_1) = \Sigma a_iE(Y_i)$$

$$= \Sigma a_i(\beta_0 + \beta_1X_i)$$

$$= \beta_0\Sigma a_i + \beta_1\Sigma a_iX_i$$

$$= \beta_0\Sigma\frac{(X_i - \bar{X})}{\Sigma(X_i - \bar{X})^2} + \beta_1\Sigma\frac{(X_i - \bar{X})X_i}{\Sigma(X_i - \bar{X})^2}$$

เนื่องจากว่า $\Sigma(X_i - \bar{X}) = 0$

$$E(\hat{\beta}_1) = 0 + \beta_1\Sigma\frac{(X_i - \bar{X})(X_i - \bar{X})}{\Sigma(X_i - \bar{X})^2}$$

$$= 0 + \beta_1\frac{\Sigma(X_i - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}$$

$$= \beta_1$$

สำหรับความแปรปรวนของ $\hat{\beta}_1$ เราได้

$$V(\hat{\beta}_1) = \Sigma a_i^2V(Y_i)$$

$$= \Sigma a_i^2\sigma^2_{y/x}$$

$$= \Sigma\frac{(X_i - \bar{X})^2}{[\Sigma(X_i - \bar{X})^2]^2}\sigma^2_{y/x}$$

$$= \frac{\Sigma(X_i - \bar{X})^2\sigma^2_{y/x}}{[\Sigma(X_i - \bar{X})^2]^2}$$

$$= \frac{\sigma^2_{y/x}}{\Sigma(X_i - \bar{X})^2}$$

เมื่อเราทราบค่าคาดหวังและความแปรปรวนของ $\hat{\beta}_0$, $\hat{\beta}_1$ ก็เท่ากับว่าเราทราบการแจกแจงของ $\hat{\beta}_0$, $\hat{\beta}_1$

12.7 การทดสอบสมมติฐานและช่วงความเชื่อมั่นของ β_1

เมื่อไรเราทราบการแจกแจงของ $\hat{\beta}_1$ ว่ามีการแจกแจงปกติ มีมัชฌิมเลขคณิต β_1 ความแปรปรวน $\sigma^2_{y/x}/\Sigma(X_i - \bar{X})^2$ การทดสอบเกี่ยวกับ β_1 ก็สามารถใช้การทดสอบแบบ Z ได้

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\Sigma(X_i - \bar{X})^2}}}$$

แต่โดยทั่ว ๆ ไปแล้วเราไม่ทราบความแปรปรวน $\sigma^2_{y/x}$ จึงจำเป็นต้องประมาณด้วย $s^2_{y/x}$ ซึ่งเป็นตัวประมาณค่าที่ปราศจากความเอนเอียงของ $\sigma^2_{y/x}$ เมื่อไรที่ประมาณค่า $\sigma^2_{y/x}$ ด้วย $s^2_{y/x}$ แล้ว ค่าประมาณ $\hat{\beta}_1$ จะมีการแจกแจงแบบ t นั่นคือ

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{s^2_{y/x}}{\Sigma(X_i - \bar{X})^2}}}$$

มีองศาแห่งความอิสระเป็น $n - 2$

12.8 ช่วงความเชื่อมั่นของ β_1

จากตารางเราทราบว่าพื้นที่ใต้โค้งการแจกแจงของ t เท่ากับ 0.95 ระหว่างค่า $-t_{.025}$ กับ $+t_{.025}$ ที่องศาแห่งความอิสระ $n - 2$ นั่นคือ

$$P(-t_{.025} < t < +t_{.025}) = \mathbf{0.95}$$

ถ้าหากว่าเราเขียน t ในรูป $\frac{\hat{\beta}_1 - \beta_1}{s_{y/x}/\sqrt{\Sigma(X - \bar{X})^2}}$ เพื่อหาค่าช่วงของค่าอะไรที่ครอบคลุมค่า β_1 ด้วยความน่าจะเป็น 0.95 เราจะได้

$$P\left(-t_{.025} < \frac{(\hat{\beta}_1 - \beta_1)\sqrt{\Sigma(X - \bar{X})^2}}{S_{y/x}} < +t_{.025}\right) = \mathbf{0.95}$$

สมการไม่เท่ากันในวงเล็บแสดงได้เป็น

$$P\left(\hat{\beta}_1 - t_{.025}\frac{s_{y/x}}{\sqrt{\Sigma(X - \bar{X})^2}} < \beta_1 < \hat{\beta}_1 + t_{.025}\frac{s_{y/x}}{\sqrt{\Sigma(X - \bar{X})^2}}\right) = 0.95$$

ซึ่งให้ 95% ช่วงความเชื่อมั่นสำหรับ β_1 กล่าวคือ

$$\beta_1 = \hat{\beta}_1 \pm t_{.025}\frac{S_{y/x}}{\sqrt{\Sigma(X - \bar{X})^2}}$$

ในเมื่อ $\hat{\beta}_1 - t_{.025}\frac{S_{y/x}}{\sqrt{\Sigma(X - \bar{X})^2}}$ เป็นขีดจำกัดล่าง

$\hat{\beta}_1 + t_{.025}\frac{S_{y/x}}{\sqrt{\Sigma(X - \bar{X})^2}}$ เป็นขีดจำกัดบน

จากตัวอย่างก่อนระหว่างเกรดเฉลี่ยกับเชาวน์ปัญญา หากเราต้องการทดสอบว่าเชาวน์ปัญญาไม่มีอิทธิพลต่อเกรดเฉลี่ยโดยเปรียบเทียบกับ alternative สองข้าง เราจะได้ว่า

$$Ho : \beta_1 = 0$$

$$Ha : \beta_1 \neq 0$$

ไม่ยอมรับ Ho ที่ระดับนัยสำคัญ 5%

$$t = \frac{\hat{\beta}_1 - \beta_1}{S_{y/x}/\sqrt{\Sigma(X-\bar{X})^2}}$$

ในเมื่อ $S^2_{y/x} = \frac{\Sigma(Y - \hat{Y})^2}{n - 2} = \frac{2.0651}{10}$

$$= 0.20651$$

$$\Sigma(X - \bar{X})^2 = 476.917$$

$$\therefore t = \frac{0.0657}{\sqrt{\frac{0.20651}{476.917}}}$$

$$= \frac{0.0657}{0.0208} = 3.159$$

เราจะไม่ยอมรับ Ho เมื่อ $t > t_{.025\,;\,10} = 2.228$ หรือ $t < -t_{.025\,:\,10} = -2.228$

$\therefore$ เราจึงไม่สามารถยอมรับ Ho ที่ระดับนัยสำคัญ 0.05 จึงสรุปได้ว่าเชาวน์ปัญญามีอิทธิพลต่อเกรดเฉลี่ยหรือเชาวน์ปัญญามีความสัมพันธ์เชิงเส้นกับเกรดเฉลี่ย (Ho : $\beta_1 = 0$ หมายความว่าเชาวน์ปัญญากับเกรดเฉลี่ยไม่มีความสัมพันธ์เชิงเส้น)

12.9 ความคลาดเคลื่อนมาตรฐานของ $\hat{Y}$

เราทราบแล้วว่าสมการถดถอย คือ

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

ในเมื่อ $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

$$\therefore \hat{Y} = \bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 X$$
$$= \bar{Y} + \hat{\beta}_1 (X - \bar{X})$$

ในเมื่อทั้ง $\bar{Y}$ และ $\hat{\beta}_1$ มีส่วนทำให้ $\hat{Y}$ มีความคลาดเคลื่อนและมีความอิสระต่อกันด้วย ดังนั้น ความแปรปรวนของค่าคาดหวังของ Y คือ $\hat{Y}_k$ เมื่อกำหนดค่า X_k ของ X คือ

$$V(\hat{Y}) = V(\bar{Y}) + V[\hat{\beta}_1 (X_k - \bar{X})] + Cov(\bar{Y}, \hat{\beta}_1 (X_k - \bar{X}))$$

$$Cov(\bar{Y}, \hat{\beta}_1 (X_k - \bar{X})) = 0$$

$$V(\hat{Y}_k) = V(\bar{Y}) + V[\hat{\beta}_1 (X_k - \bar{X})]$$
$$= V(\bar{Y}) + (X_k - \bar{X})^2 V[\hat{\beta}_1)$$
$$= \frac{\sigma^2_{y/x}}{n} + \frac{(X_k - \bar{X})^2 \sigma^2_{y/x}}{\Sigma(X - \bar{X})^2}$$

ตัวค่าประมาณของ $V(\hat{Y}_k)$ คือ

$$V(\hat{Y}_k) = S^2_{y/x} \left[\frac{1}{n} + \frac{(X_k - \bar{X})^2}{\Sigma(X - \bar{X})^2}\right]$$

ความคลาดเคลื่อนมาตรฐานของ $\hat{Y}_k$ คือ

$$\sqrt{V(\hat{Y}_k)} = S_{y/x} \left[\frac{1}{n} + \frac{(X_k - \bar{X})^2}{\Sigma(X - \bar{X})^2}\right]^{1/2}$$

ค่านี้จะมีค่าน้อยที่สุดเมื่อ $X_k = \bar{X}$ และจะเพิ่มขึ้นเมื่อไรที่ค่า X_k เคลื่อนออกจาก $\bar{X}$ ไม่ว่าในทิศทางใดทางหนึ่ง ดังนั้น ความคาดหวังในการทำนายของเราที่ดีที่สุดก็คือพยายามทำนายค่ากลาง ๆ ของช่วงค่าสังเกตของ X

ตัวอย่าง เราใช้โจทย์ของตัวอย่างก่อนในการหา $V(\hat{Y}_k)$

$$n = 12,\ \Sigma(X - \bar{X})^2 = 476{,}917,\ S^2_{y/x} = 0.2065 : \bar{X} = 124.583$$

$$V(\hat{Y}_k) = s^2_{y/x}\left[\frac{1}{n} + \frac{(X_k - \bar{X})^2}{\Sigma(X - \bar{X})^2}\right]$$

$$= 0.2065\left[\frac{1}{12} + \frac{(X_k - 124.583)^2}{476.917}\right]$$

หาก $X_k = X$

$$V(\hat{Y}_k) = 0.2065\ (\frac{1}{12}) = 0.0172$$

$$\sqrt{V(\hat{Y}_k)} = \sqrt{0.0172} = 0.131$$

ถ้าหาก $X_k = 120$

$$V(\hat{Y}_k) = 0.2065\left[\frac{1}{12} + \frac{(120 - 124.583)^2}{476.917}\right]$$

$$= 0.1839$$

$$\sqrt{V(\hat{Y}_k)} = 0.429$$

ต้องการคำนวณหาช่วงความเชื่อมั่นที่ 90% สำหรับเกรดเฉลี่ยที่คาดหวังโดยกำหนดคะแนนเชาวน์ปัญญาเท่ากับ 120 คือ ($t_{.05\ ;\ 10} = 1.812$)

$$\hat{Y}_k = 5.5021 + 0.0657\ (120)$$

$$= -5.5021 + 7.884$$

$$= 2.3819$$

$$\mathbf{E\ (Y/x = 120)} = \hat{Y}_k \pm t_{\ 05\ \ 10}\ S_{y/x}\sqrt{\frac{1}{n} + \frac{(X_k - \bar{X})^2}{\Sigma(X - \bar{X})^2}}$$

$$= \mathbf{2.38} \pm (1.812)\ (.454)\sqrt{\frac{1}{12} + \frac{(120 - 124.583)^2}{476.9167}}$$

$$= \mathbf{2.38} \pm (1.812)\ (.1384)$$

$$= \mathbf{2.38} \pm .251$$

นั่นคือ $\mathbf{2.129 < E\ (Y/x = 120) < 2.831}$

ความแปรปรวนและความคลาดเคลื่อนมาตรฐานที่ได้กล่าวข้างต้นใช้สำหรับหาค่าคาดหวังของ Y กำหนด X_k แต่เนื่องจากว่าค่าสังเกตจริงของ Y แปรร่วมค่าคาดหวังจริงด้วยความ

แปรปรวน $\sigma^2_{y/x}$ ซึ่งมีความอิสระกับ V $(\hat{Y})$ ค่าที่ถูกทำนายของค่าสังเกตแต่ละค่าก็ยังกำหนดได้โดย Y แต่จะมีความแปรปรวน

$$\sigma^2_{y/x} + V(\hat{Y}_k) = \sigma^2_{y/x} + \frac{\sigma^2_{y/x}}{n} + \frac{(X_k - \bar{X})^2}{\Sigma(X - \bar{X})^2}\sigma^2_{y/x}$$

$$V(Y_p) = \sigma^2_{y/x}\left[1 + \frac{1}{n} + \frac{(X_k - \bar{X})^2}{\Sigma(X - \bar{X})^2}\right]$$

ใช้ $s^2_{y/x}$ แทนค่า $\sigma^2_{y/x}$ เราได้

$$V(Y_p) = s^2_{y/x}\left[1 + \frac{1}{n} + \frac{(X_k - \bar{X})^2}{\Sigma(X - \bar{X})^2}\right]$$

ค่าความเชื่อมั่นก็สามารถหาได้ด้วยวิธีเดียวกันกับที่ได้กล่าวมาก่อน นั่นคือ เราสามารถคำนวณ 90% ช่วงความเชื่อมั่น สำหรับค่าสังเกตใหม่ซึ่งจะขึ้นอยู่กับตัวประมาณค่าของค่าความแปรปรวนใหม่นี้

$$Y_p = \hat{Y}_k \pm t_{.05;n-2}\left\{1 + \frac{1}{n} + \frac{(X_k - \bar{X})^2}{\Sigma(X - \bar{X})^2}\right\}^{\frac{1}{2}} s_{y/x}$$

เพราะฉะนั้น 90% ช่วงความเชื่อมั่นของการทำนายเกรดเฉลี่ยของนักศึกษาโดยกำหนดคะแนนเชาวน์ปัญญาเท่ากับ 120 เป็น

$$Y_p = \hat{Y}_k \pm t_{.05;n-2}\, s_{y/x}\sqrt{1 + \frac{1}{n} + \frac{(X_k - \bar{X})^2}{\Sigma(X - \bar{X})^2}}$$

$$= 2.38 \pm (1.812)(.454)\sqrt{1 + \frac{1}{12} + \frac{(120 - 124.583)^2}{476.9161}}$$

$$= 2.38 \pm (1.812)(.454)(1.062)$$

$$= 2.38 \pm .8735$$

$$1.5065 < Y_p < 3.2535$$

12.10 สหสัมพันธ์

สหสัมพันธ์เป็นการแสดงถึงองศาความสัมพันธ์ระหว่างตัวแปร 2 ตัว หรือมากกว่า ว่ามีความสัมพันธ์มากน้อยแค่ไหน หากความสัมพันธ์นั้นมีเพียง 2 ตัวแปร ก็เป็นสหสัมพันธ์อย่างง่าย

ถ้าหากว่าความสัมพันธ์นั้นมีตัวแปรตั้งแต่สามตัวขึ้นไปก็เป็นสหสัมพันธ์เชิงซ้อน ในหัวข้อนี้จะกล่าวแต่สหสัมพันธ์อย่างง่ายและเป็นเชิงเส้นเท่านั้น นักสถิติใช้สัญลักษณ์ r แสดงค่าสัมประสิทธิ์สหสัมพันธ์อย่างง่ายของตัวอย่างและ ρ (rho) เป็นค่าสัมประสิทธิ์สหสัมพันธ์อย่างง่ายของประชากร

ในการศึกษาสหสัมพันธ์อย่างง่าย เราใช้ค่า r อนุมานค่า ρ เพราะในทางปฏิบัติเราไม่สามารถเอาค่าสังเกตทั้งหมดจากประชากรได้ r จึงเป็นค่าแสดงองศาแห่งความสัมพันธ์ระหว่างข้อมูล 2 ชุด r จะมีค่าอยู่ระหว่าง -1 กับ $+1 (-1 < r < +1)$ หาก r มีค่า $= +1$ แสดงว่าข้อมูลทั้งสองมีความสัมพันธ์กันอย่างสมบูรณ์ในทิศทางเดียวกัน นั่นคือ ข้อมูล X เปลี่ยนไปข้อมูล Y จะเปลี่ยนตามไปในทางเดียวกัน กล่าวคือ X เพิ่ม Y เพิ่ม X ลด Y ก็ลดตาม แต่ถ้า r มีค่าเท่ากับ -1 ก็แสดงว่าข้อมูล X กับ Y มีความสัมพันธ์กันอย่างสมบูรณ์ในทิศทางตรงข้าม นั่นคือ ข้อมูล X เพิ่มขึ้นจะมีผลให้อีกข้อมูล Y ลดลง หรือข้อมูล X ลดลงจะมีผลให้ข้อมูล Y เพิ่มขึ้น นอกจากนี้ค่า r ยังใช้วัด goodness of fit ของเส้นกำลังสองน้อยที่สุดหรือเส้นถดถอยอีกด้วย กล่าวคือ เราสามารถใช้ข้อมูล X ไปทำนายค่า Y ในเส้นถดถอยที่คำนวณได้ ได้ถูกต้องร้อยเปอร์เซ็นต์ หากเราหาค่า $r = \pm 1$ หาก r มีค่าเข้าใกล้ศูนย์ก็แสดงว่าข้อมูล X กับข้อมูล Y ไม่มีความสัมพันธ์กันเลย (หรือน้อยมาก) หรือ fit is poor

วิธีการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์อย่างง่ายที่นิยมทำมีดังนี้

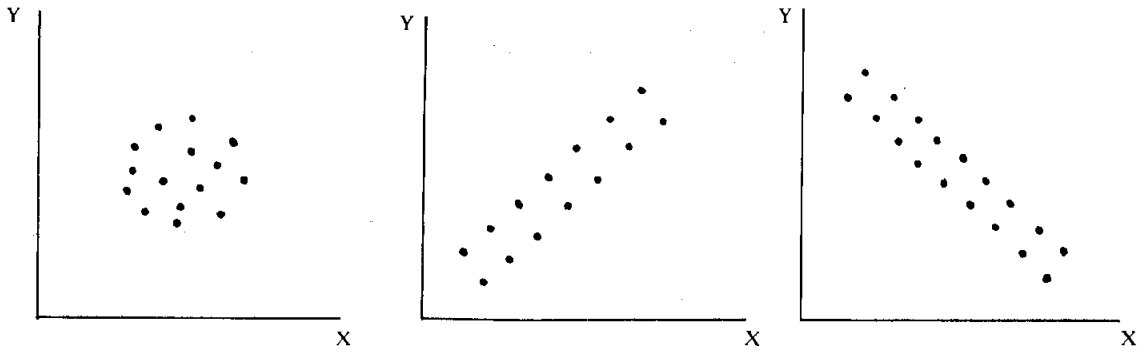
วิธีที่ 1 พิจารณาลักษณะของการกระจายของข้อมูลจากแผนภาพกระจาย

วิธีที่ 2 พิจารณาลักษณะของเส้นถดถอยโดยพิจารณาความชันของเส้นและทิศทางของเส้น

วิธีที่3 โดยใช้สูตร

วิธีที่ 1 พิจารณาลักษณะของการกระจายของข้อมูล

วิธีการนี้เป็นวิธีการอ่านความสัมพันธ์ของข้อมูล 2 ชุด อย่างง่าย ๆ โดยการนำค่าข้อมูลทั้งสองมาพลอตกราฟ หากจุดของข้อมูลในกราฟมีการกระจายกันมากหรือจับกันมีลักษณะเป็นกลุ่ม ก็แสดงว่า X และ Y มีองศาแห่งความสัมพันธ์น้อยหรือไม่มีเลย ดังรูป (ก)



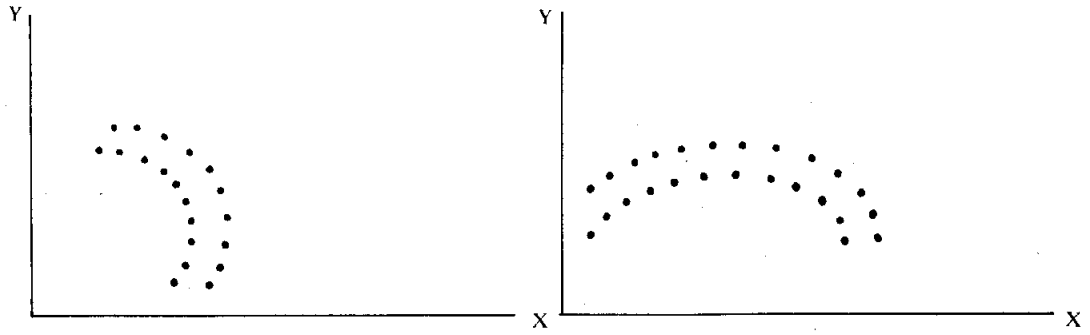
สหสัมพันธ์น้อยหรือศูนย์ (ก)

สหสัมพันธ์มากทางบวก (ข)

สหสัมพันธ์มากทางลบ (ค)

แต่ถ้าจุดของข้อมูลกระจายเป็นแนวเดียวกันก็แสดงว่าข้อมูล 2 ชุด นั้นมีสหสัมพันธ์สูง ดังรูป (ข) (ค)

ในบางครั้งข้อมูลมีความสัมพันธ์กันแต่เป็นไปในรูปของเส้นโค้ง เราเรียกข้อมูลทั้งสองมีสหสัมพันธ์แบบไม่เชิงเส้น

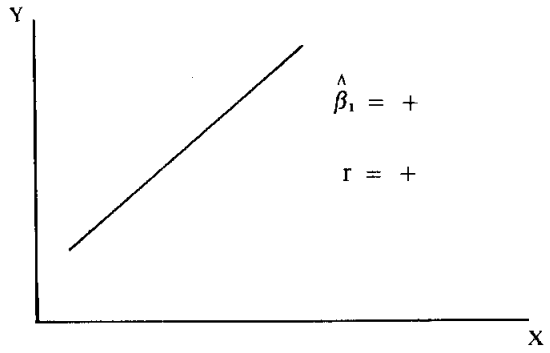


สหสัมพันธ์แบบไม่เชิงเส้น

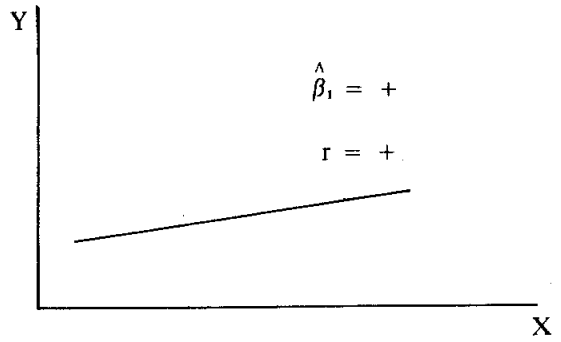
วิธีที่ 2 พิจารณาลักษณะของเส้นถดถอย

เราอ่านค่าสัมประสิทธิ์สหสัมพันธ์ได้จากการพิจารณาลักษณะความชันและทิศทางของเส้นถดถอยโดยประมาณอย่างหยาบ ๆ

ก. เส้นถดถอยที่มีความชันเป็นบวก ค่าสัมประสิทธิ์สหสัมพันธ์ก็จะมีค่าเป็นบวก ดังรูป

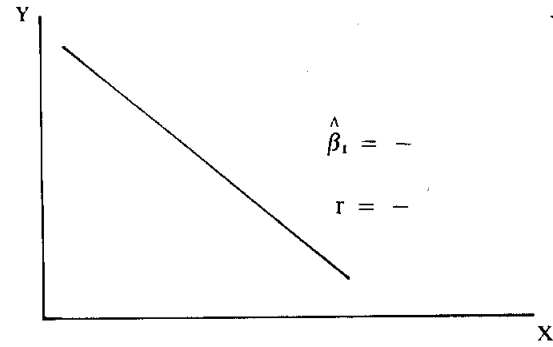


สหสัมพันธ์ทางบวกมาก

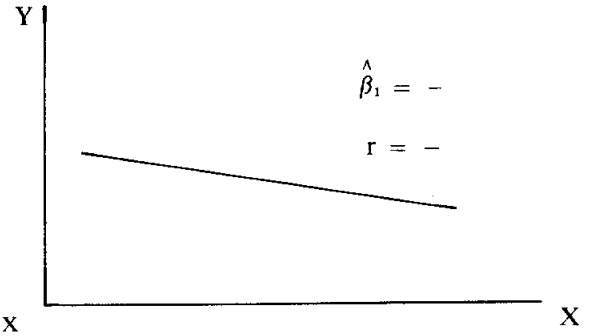


สหสัมพันธ์ทางบวกน้อย

ข. เส้นถดถอยที่มีความชันเป็นลบ ค่าสัมประสิทธิ์สหสัมพันธ์ก็จะมีค่าเป็นลบ ดังรูป

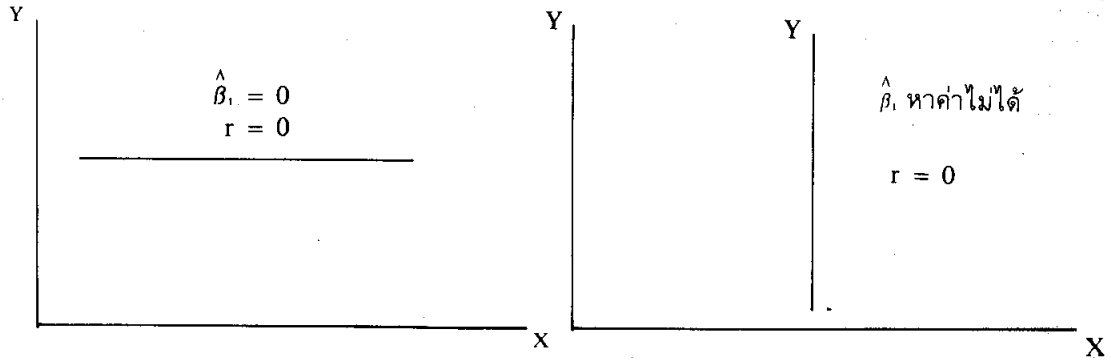


สหสัมพันธ์ทางลบมาก



สหสัมพันธ์ทางลบน้อย

ค. เส้นถดถอยที่มีความชันเป็นศูนย์ ค่าสัมประสิทธิ์สหสัมพันธ์จะมีค่าเป็นศูนย์ หรือไม่มีความสัมพันธ์กันเลย



สหสัมพันธ์เป็นศูนย์

วิธีที่ 3 โดยสูตร

สูตรที่ใช้คำนวณหาค่าสัมประสิทธิ์สหสัมพันธ์ที่นิยมใช้กันมี

1. $$r = \sqrt{\frac{\Sigma(\hat{Y} - \bar{Y})^2}{\Sigma(Y - \bar{Y})^2}}$$

สูตรนี้ใช้คำนวณสัมประสิทธิ์สหสัมพันธ์ของตัวแปรตั้งแต่สองตัวขึ้นไปทั้งที่มีสหสัมพันธ์เชิงเส้นและไม่เชิงเส้น ส่วนทิศทางของความสัมพันธ์นั้น จะต้องดูความชันของเส้นถดถอย

2. $$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{(\Sigma(X - \bar{X})^2)(\Sigma(Y - \bar{Y})^2)}}$$

หรือ $$r = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{(n\Sigma X^2 - (\Sigma X)^2)(n\Sigma Y^2 - (\Sigma Y)^2)}}$$

สูตรนี้ใช้คำนวณสัมประสิทธิ์สหสัมพันธ์ของตัวแปรเพียงสองตัวเท่านั้น ในกรณีตัวแปรเพียงสองตัว สูตรทั้งสองจะให้ค่า r เท่ากันหรือใกล้เคียงกัน

ตัวอย่าง คำนวณหาสัมประสิทธิ์สหสัมพันธ์ r ระหว่างคะแนนกับจำนวนเงินที่ขายได้ในตารางต่อไปนี้

X	Y	Y^2	X^2	XY	$\hat{Y}$	$(\hat{Y} - \bar{Y})^2$
48	312	97,344	2,304	14,976	307.04	5,218.62
32	164	26,896	1,024	5,248	196.96	1431.87
40	280	78,400	1,600	11,200	252.00	295.84
34	196	38,416	1,156	6,664	210.72	579.85
30	200	40,000	900	6,000	138.20	2,662.56
50	288	82,944	2,500	14,400	320.80	7,396.–
26	146	21,316	676	3,796	155.68	6,259.97
50	361	130,321	2,500	18,050	320.80	7,396.–
22	149	22,201	484	3,278	128.16	11,372.09
43	252	63,504	1,849	10,836	272.64	1,431.87
375	2,348	601,342	14,993	94,448	2,348.–	39,405.89

ใช้สูตรที่ 1 $r = \sqrt{\frac{\Sigma(\hat{Y} - \bar{Y})^2}{\Sigma(Y - \bar{Y})^2}}$

ในที่นี้ $\Sigma(Y - \bar{Y})^2 = \Sigma Y^2 - n\bar{Y}^2$

$= 601,342 - 10\left(\frac{2,348}{10}\right)^2$

$= 50,032$

$r = \sqrt{\frac{39,405.89}{50,032}}$

$= \sqrt{0.7876}$

$= 0.89$

ใช้สูตรที่ 2

$$r = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{(n\Sigma X^2 - (\Sigma X)^2)(n\Sigma Y^2 - (\Sigma Y)^2)}}$$

ในเมื่อ

$$n = 10\ ;\ \Sigma X = 375,\ \Sigma Y = 2,348\ ;\ \Sigma X^2 = 14,993$$

$$\Sigma Y^2 = 601,342\ ,\ \Sigma XY = 94,448$$

$$r = \frac{10(94,448) - (375)\,(2,348)}{\sqrt{[\,10(14,993) - (375)^2\,]\,[\,10(601,342) - (2,348)^2\,]}}$$

$$= \frac{944,480 - 880,500}{\sqrt{(9,305)\,(500,316)}} = \frac{63,980}{68,230.773} = .94$$

12.11 สัมประสิทธิ์ของการตัดสินใจ (Coefficient of Determination)

ค่าสัมประสิทธิ์สหสัมพันธ์ r เมื่อยกกำลังสองก็จะได้ค่าสัมประสิทธิ์ของการตัดสินใจ r^2 ซึ่งเป็นตัวใช้อธิบายการกระจายที่เกิดขึ้นทั้งหมดในค่าสังเกต Y ได้มากน้อยเพียงใด หรือกล่าวอีกนัยหนึ่ง r^2 เป็นตัวชี้ให้เห็นว่าเส้นถดถอยที่เราคำนวณหาได้นั้นเหมาะสมกับข้อมูลเพียงใด โดยทั่ว ๆ ไปเราอ่านค่า r^2 ในรูปเปอร์เซ็นต์ อย่างเช่น หากเราคำนวณค่า $100r^2 = 0$ หมายความว่าการกระจายในค่าสังเกต Y ที่เกิดขึ้นไม่ได้ถูกอธิบายได้โดยเส้นถดถอยเลย ซึ่งเป็นผลให้การกระจายที่อธิบายไม่ได้ $\Sigma(Y - \hat{Y})^2$ เท่ากับการกระจายทั้งหมด $\Sigma(Y - \bar{Y})^2$ ถ้าหากว่า $100r^2 = 100\%$ ก็หมายความว่าเส้นถดถอยที่เราคำนวณได้นั้นอธิบายการกระจายทั้งหมดของค่าสังเกต Y ได้ ซึ่งมีผลให้การกระจายที่อธิบายไม่ได้ไม่มีเลย $(\Sigma(Y - \hat{Y})^2 = 0)$

12.12 ความสัมพันธ์ระหว่าง r กับ β_1

กำหนดให้ข้อมูล 2 ชุด โดยชุดที่หนึ่ง X เป็นตัวแปรอิสระ อีกชุดหนึ่ง Y เป็นตัวแปรตาม สมการถดถอยก็เขียนได้เป็น

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

ในทำนองเดียวกัน หาก X เป็นตัวแปรตาม Y เป็นตัวแปรอิสระ สมการถดถอยก็เขียนได้เป็น (ข้อชุดเดียวกัน)

$$\hat{X} = \hat{\beta}_0' + \hat{\beta}_1' Y$$

จากสมการถดถอยแรกเราคำนวณหาค่า $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{\Sigma(X - \bar{X})\,(Y - \bar{Y})}{\Sigma(X - \bar{X})^2}$$

สมการถดถอยที่สอง

$$\hat{\beta}_1' = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(Y - \bar{Y})^2}$$

นิยาม สัมประสิทธิ์สหสัมพันธ์ r คือ รูทสองของผลคูณ $\hat{\beta}_1$ กับ $\hat{\beta}_1'$ ซึ่งเป็นความชันของสมการถดถอยสองเส้น นั่นคือ

$$r = \sqrt{\hat{\beta}_1 \hat{\beta}_1'}$$

$$r = \sqrt{\left[\frac{\Sigma(X - X)(Y - \bar{Y})}{\Sigma(X - \bar{X})^2}\right]\left[\frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(Y - \bar{Y})^2}\right]}$$

$$= \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2}}$$

ตัวอย่าง

X	2	6	4	7
Y	−15	41	13	55

(ก) จงคำนวณหาสมการถดถอย Y on X

(ข) จงคำนวณหาสมการถดถอย X on Y

(ค) จงคำนวณหาสัมประสิทธิ์สหสัมพันธ์

X	Y	X^2	Y^2	XY
2	−15	4	225	−30
6	41	36	1681	246
4	13	16	169	52
7	55	49	3025	385
19	94	105	5100	653

(ก) สมการถดถอย $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

$$\hat{\beta}_1 = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2}$$

$$= \frac{4(653) - (19)(94)}{4(105) - (19)^2}$$

$$= \frac{826}{59} = 14$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{94}{4} - \frac{(14)(19)}{4}$$

$$= 23.5 - 66.5 = -43$$

$$\hat{Y} = -43 + 14X$$

(ข) สมการถดถอย

$$\hat{X} = \hat{\beta}_0' + \hat{\beta}_1' Y$$

$$\hat{\beta}_1' = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma Y^2 - (\Sigma Y)^2} = \frac{4(653) - (19)(94)}{4(5100) - (94)^2}$$

$$= \frac{826}{11564} = 0.07143$$

$$\hat{\beta}_0' = \bar{X} - \hat{\beta}_1' \bar{Y}$$

$$= \frac{(19)}{4} - \frac{(0.07143)(94)}{4}$$

$$= 4.75 - 1.6786$$

$$= 3.0714$$

$$\hat{X} = 3.0714 - 0.07143\,Y$$

(ค)
$$r = \sqrt{\hat{\beta}_1 \hat{\beta}_1'}$$

$$= \sqrt{(14)(0.07143)}$$

$$= 1$$

ในกรณีที่เราทราบค่าส่วนเบี่ยงเบนมาตรฐานของตัวแปรค่าทั้งสองและความชันของเส้นถดถอย เราสามารถคำนวณหาสัมประสิทธิ์ของสหสัมพันธ์ได้ นั่นคือ

$$r = \hat{\beta}_1 \frac{S_x}{S_y}$$

หรือ
$$\hat{\beta}_1 = r \frac{S_y}{S_x}$$

ในเมื่อสมการถดถอยเป็น $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

หากสมการถดถอยเป็น $\hat{X} = \hat{\beta}_0' + \hat{\beta}'_1 Y$

เราได้ $$r = \hat{\beta}_1' \frac{S_y}{S_x}$$

หรือ $$\hat{\beta}_1' = r \frac{S_x}{S_y}$$

ตัวอย่าง มีนักศึกษา 100 คน สอบวิชาคณิตศาสตร์กับวิชาสถิติ ซึ่งมีความชัน $\hat{\beta}_1$ ของสมการเส้นถดถอยของวิชาสถิติเมื่อกำหนดวิชาคณิตศาสตร์เท่ากับ 0.8 และส่วนเฉลี่ยเลขคณิตและส่วนเบี่ยงเบนมาตรฐานเป็น ดังนี้

	คณิตศาสตร์	สถิติ
ส่วนเฉลี่ยเลขคณิต	50	55
ส่วนเบี่ยงเบนมาตรฐาน	5	6

$$S_x = 5;\ S_y = 6;\ \hat{\beta}_1 = 0.8$$

เนื่องจากว่า $r = \hat{\beta}_1 \frac{S_x}{S_y}$ (เป็นสมการถดถอย Y on X)

$$= (0.8)\ \frac{5}{6}$$

$$= \frac{4}{6} = \frac{2}{3}$$

เพราะฉะนั้นสัมประสิทธิ์สหสัมพันธ์ $r = \frac{2}{3}$

หมายเหตุ ค่าสัมประสิทธิ์สหสัมพันธ์จะมีค่าคงที่เสมอจากข้อมูล 2 ชุด ไม่ว่าข้อมูล 2 ชุด หรือชุดหนึ่งบวกหรือลบด้วยตัวคงที่ทุก ๆ ข้อมูล หรือข้อมูล 2 ชุด หรือชุดหนึ่งคูณหรือหารด้วยตัวคงที่ทุก ๆ ข้อมูล และค่าสัมประสิทธิ์สหสัมพันธ์ของข้อมูลคู่หนึ่งจะมีค่าเท่ากับศูนย์ 2 คู่ จะมีค่าเท่ากับ $+1$ หรือ -1 ขึ้นอยู่กับข้อคู่นั้นมีทิศทางไปทางเดียวกันหรือทิศทางตรงกันข้ามกัน

ตัวอย่าง มีข้อมูล 2 ชุด X และ Y มีมัชฌิมเลขคณิต $\bar{X}$, $\bar{Y}$ ค่าสัมประสิทธิ์สหสัมพันธ์ r_{xy} คำนวณได้จากสูตร

$$r_{xy} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - x)^2\ \Sigma(Y - \bar{Y})^2}}$$

หากข้อมูล 2 ชุด X และ Y บวกหรือลบด้วยตัวคงที่ C นั่นคือ $Z = X \pm C; U = Y \pm C$ เพราะฉะนั้น มัชฌิมเลขคณิตของ Z และ U คือ $\bar{Z} = \bar{X} \pm C; \bar{U} = \bar{Y} \pm C$ ค่าสัมประสิทธิ์สหสัมพันธ์ของ Z กับ U คือ

$$r_{zu} = \frac{\Sigma(Z - \bar{Z})(U - \bar{U})}{\sqrt{\Sigma(Z - \bar{Z})^2 \Sigma(U - \bar{U})^2}}$$

$$= \frac{\Sigma(X \pm C - (\bar{X} \pm C))(Y \pm C - (\bar{Y} \pm C))}{\sqrt{\Sigma(X \pm C - (\bar{X} \pm C))^2 \Sigma(Y \pm C - (\bar{Y} \pm C))^2}}$$

$$= \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2}}$$

$$= r_{xy}$$

ในกรณี X และ Y คูณด้วยตัวคงที่ C นั่นคือ $Z = XC$; Y มัชฌิมเลขคณิต Z และ Y คือ $\bar{X}C$ และ $\bar{Y}$

$$r_{zy} = \frac{\Sigma(Z - \bar{Z})(Y - \bar{Y})}{\sqrt{\Sigma(Z - \bar{Z})^2 \Sigma(Y - \bar{Y})^2}}$$

$$r_{zy} = \frac{\Sigma(XC - \bar{X}C)(Y - \bar{Y})}{\sqrt{\Sigma(XC - \bar{X}C)^2 \Sigma(Y - \bar{Y})^2}}$$

$$= \frac{C\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{C^2 \Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2}}$$

$$= \frac{C\Sigma(X - \bar{X})(Y - \bar{Y})}{C\sqrt{\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2}}$$

$$= r_{xy}$$

ในทำนองเดียวกันหากข้อมูล 2 ชุด จะหารด้วยตัวคงที่ก็จะได้ค่า r_{xy} เหมือนเดิมเช่นเดียวกัน

การใช้สมการถดถอยไปประมาณค่าหรือทำนายค่านั้น กระทำได้เพียงบางค่าหรือบางช่วงเท่านั้น นอกเหนือจากนั้นจะให้ค่าผิดความเป็นจริงหรือเป็นไปไม่ได้ดังตัวอย่างระหว่างราคารถที่ใช้แล้วกับอายุของรถยนต์ยี่ห้อหนึ่งจากนายหน้าขายรถ

x อายุ (ปี)	1	2	2	3	4	5	6	8	9	10
Y ราคา ($100)	**3.8**	**3.4**	**3.3**	**3.0**	**2.5**	**2.0**	**2.2**	**1.0**	**1.0**	**0.8**

เนื่องจากว่าอายุของรถเป็นตัวกำหนดราคารถยนต์ สมการถดถอยก็เขียนได้

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

	อายุ X	ราคารถยนต์ Y	XY	X^2
	1	3.8	3.8	1
	2	3.4	6.8	4
	2	3.3	6.6	4
	3	3.0	9.0	9
	4	2.5	10.0	16
	5	2.0	10.0	25
	6	2.2	13.2	36
	8	1.0	8.0	64
	9	1.0	9.0	81
	10	0.8	8.0	100
รวม	50	23.0	84.4	340
	$\bar{X} = 5$	$\bar{Y} = 2.3$		

แทนค่า $\Sigma X = 50$; $\Sigma Y = 23.0$; $\Sigma XY = 84.4$; $\Sigma X^2 = 340$

ลงในสูตร

$$\hat{\beta}_1 = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma x^2 - (\Sigma X)^2}$$

$$= \frac{10(84.4) - (50)(23)}{10(340) - (50)^2} = \frac{-306}{900}$$

$$= -0.34$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$= 2.3 \quad (-.34)5 = 4.0$$

สมการถดถอย คือ $\hat{Y} = 4.0 \quad .34X$

เมื่อไร X = 0 จุดตัดที่แกน Y คือ 4.0 (4000) แต่เนื่องจากราคารถยนต์ลดลงอย่างรวดเร็วสำหรับปีแรกและเนื่องจากว่าไม่มีรถใหม่ในตัวอย่าง 4.0 (4000) ไม่อาจเป็นตัวประมาณค่าที่ดีสำหรับราคาเฉลี่ยจริงเมื่อไร X = 0 เส้นถดถอยจะให้ค่าประมาณอย่างสมเหตุสมผลสำหรับค่า X จากหนึ่งถึงสิบปี แต่จะให้ค่าไม่เป็นจริงสำหรับ X < 1 หรือ X > 10 ปี

12.13 คำและประโยคที่ควรจำ

การถดถอย
สมการถดถอย
ตัวแปรอิสระ
การถดถอย Y on x
จุดตัดแกน Y
ค่าคลาดเคลื่อนของตัวประมาณค่า
สัมประสิทธิ์สหสัมพันธ์
การทำนาย
ค่าบวกของ r
การถดถอยเชิงเส้นอย่างง่าย
ส่วนเบี่ยงเบนมาตรฐานของ $\hat{\beta}_1$
สัมประสิทธิ์ของการตัดสินใจ
ช่วงความเชื่อมั่นของค่าทำนาย
สหสัมพันธ์
เกณฑ์กำลังสองน้อยที่สุด
ตัวแปรตาม
การถดถอย X on Y
ความชันของเส้น
แผนภาพกระจาย
ค่าลบของ r
r = ± 1
การถดถอยเชิงซ้อน
ช่วงความเชื่อมั่นของ-
ค่าคาดหวังเฉลี่ย

แบบฝึกหัดที่ 12.2

1. **สำหรับข้อมูลต่อไปนี้เกี่ยวกับความสูง (x) และน้ำหนัก (y) ของนักศึกษา 12 คน**
   (ก) พลอตจุด (ข) เดาค่าของ r (ค) คำนวณค่าของ r (.934)

x	65	73	70	68	66	69	75	70	64	72	65	71
Y	124	184	161	164	140	154	210	164	126	172	133	150

2. คำนวณค่าของ r สำหรับข้อมูลต่อไปนี้เกี่ยวกับคะแนนทดสอบสติปัญญากับ G.P.A. พลอตจุด และเดาค่าของ r (0.5497)

I.T.	295	152	214	171	131	178	225	141	116	173	230
G.P.A.	3.4	1.6	1.2	1.0	2.0	1.6	2.0	1.4	1.0	3.6	3.6
I.T.	195	174	236	198	217	143	135	146	227		
G.P.A.	1.0	2.8	2.8	1.8	2.0	1.2	2.4	2.2	2.4		

3. ค่าของ r ควรจะเป็นอะไรของตัวแปรค่าสำหรับคู่ต่อไปนี้
   (ก) จำนวนชั่วโมงของการทำงานของผู้ชายกับจำนวนหน่วยของผลิตภัณฑ์ที่ผลิตในโรงงานอุตสาหกรรมที่กำหนดให้ (+r)
   (ข) ขนาดของนครหลวงกับอัตราอาชญากรรม (+r)
   (ค) ราคาต่อหน่วยของการผลิตวัตถุกับจำนวนของหน่วยผลิตภัณฑ์ (−r)
   (ง) คะแนนคณิตศาสตร์กับคะแนนภาษาต่างประเทศ (−r)
   (จ) การบริโภคเนยกับราคาของเนย
   (ฉ) จำนวนของฝนตกในฤดูใบไม้ผลิกับอุณหภูมิเฉลี่ย (0)

4. หมายความว่าอะไรถ้าท่านได้รับการบอกเล่าว่าสหสัมพันธ์ระหว่างจำนวนของอุบัติเหตุรถยนต์ต่อปีกับอายุของผู้ขับขี่ r = −.06 ถ้าหากว่าได้พิจารณาผู้ขับขี่ที่มีอุบัติเหตุอย่างน้อยหนึ่งครั้ง

5. ท่านควรจะให้คำอธิบายอย่างไรถ้าหากว่าสหสัมพันธ์ระหว่างการเพิ่มปุ๋ยกับกำไรของผักที่เพิ่มขึ้นในฟาร์มทดลองเท่ากับ 0.20

6. สำหรับข้อมูลของปัญหาข้อที่ 2 ขีดฆ่าคะแนนทดสอบสติปัญญาที่น้อยกว่า 150 และมากกว่า 225 แล้วคำนวณค่า r เปรียบเทียบกับค่าที่คำนวณได้ในข้อ 2 (−0.1554)

7. ความสูงของบุตร (นิ้ว) 68 66 72 73 66
   ความสูงของบิดา (นิ้ว) 64 66 71 70 69
   จากตัวอย่างสุ่มข้างต้นของความสูงของบุตรและบิดา 5 คน จงหาสหสัมพันธ์ r ของตัวอย่าง (0.62)