บทที่ 8
สหสัมพันธ์และการวิเคราะห์การถดถอย

การหาความสัมพันธ์ระหว่างข้อมูลเชิงปริมาณ 2 ชุด หรือมากกว่านั้นถ้าใช้เทคนิคการทดสอบไคสแควร์จะต้องแปลงข้อมูลเชิงปริมาณเป็นข้อมูลเชิงกลุ่มเสียก่อนแล้วจึงนับความถี่ ซึ่งการทำเช่นนี้จะสูญเสียรายละเอียดของข้อมูลไป เทคนิคการวิเคราะห์สหสัมพันธ์เป็นเทคนิคการหาความสัมพันธ์ระหว่างข้อมูลเชิงปริมาณตั้งแต่ 2 ชุดขึ้นไป หรือตัวแปรแบบต่อเนื่องตั้งแต่ 2 ตัวขึ้นไป และสามารถบอกได้ว่าข้อมูลที่มีความสัมพันธ์กันนั้นมีความสัมพันธ์กันในทิศทางใด ในขณะเดียวกันหากข้อมูลตั้งแต่ 2 ชุดมีความสัมพันธ์โดยที่ข้อมูลชุดหนึ่ง (ตัวแปรตาม) ขึ้นอยู่กับข้อมูลชุดอื่น ๆ (ตัวแปรอิสระ) เทคนิคการวิเคราะห์การถดถอยจะช่วยพยากรณ์ค่าของตัวแปรตามเมื่อกำหนดค่าของตัวแปรอิสระเหล่านั้น

สหสัมพันธ์ (Correlation)

สหสัมพันธ์เป็นการศึกษาความสัมพันธ์ระหว่างข้อมูลหรือตัวแปรตั้งแต่2 ตัวขึ้นไป ว่ามีความสัมพันธ์กันในระดับใด และมีความสัมพันธ์ในทิศทางใด เช่นความสูงกับน้ำหนักของคนมีความสัมพันธ์กันมากหรือน้อย และมีความสัมพันธ์ในทิศทางเดียวกันหรือตรงกันข้าม

ถ้าให้ X เป็นตัวแปรแทน ความสูง (cm)

Y เป็นตัวแปรแทนน้ำหนัก (kg)

ลักษณะของข้อมูลที่เก็บรวมรวมมาจะต้องเป็นข้อมูลที่มาจากหน่วยตัวอย่างเดียวกัน เช่นคนเดียวกัน ดังนี้

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

จากตารางจะเห็นว่าค่าของตัวแปร Y และ X ได้มาจากคน ๆ เดียวกัน ซึ่งอาจจะมีหน่วยเดียวกันหรือหน่วยต่างกันก็ได้

การหาความสัมพันธ์ระหว่างข้อมูลหรือตัวแปร 2 ตัว เรียกว่าสหสัมพันธ์อย่างง่าย (simple correlation) การหาความสัมพันธ์ระหว่างข้อมูลหรือตัวแปรมากกว่า 2 ตัว เรียกว่าสหสัมพันธ์เชิงพหุ (multiple correlation) ในเอกสารฉบับนี้จะกล่าวถึงสหสัมพันธ์อย่างง่ายเท่านั้น

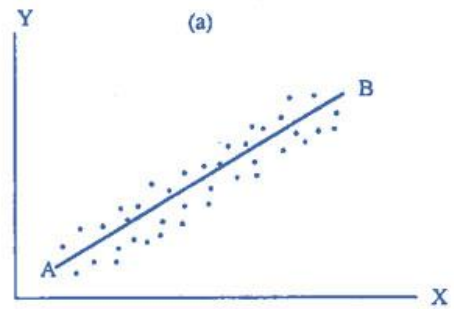
วิธีการตรวจสอบลักษณะความสัมพันธ์

การตรวจสอบลักษณะความสัมพันธ์ของตัวแปรสามารถทำได้หลายวิธี ดังนี้

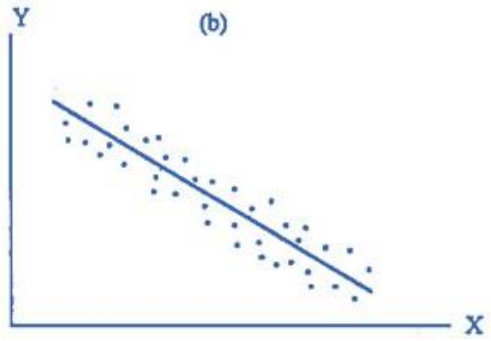
1. แผนภาพการกระจาย (scatter diagram)
2. ค่าสัมประสิทธิ์สหสัมพันธ์ (correlation coefficient)

1. แผนภาพการกระจาย

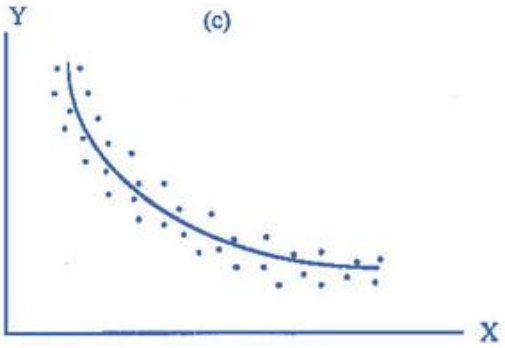
แผนภาพการกระจายเป็นวิธีการดูลักษณะความสัมพันธ์ของข้อมูลอย่างคร่าว ๆ โดยดูจากลักษณะการกระจาย หรือแนวโน้มของจุดเมื่อเทียบกับเส้นตรง ดังนี้



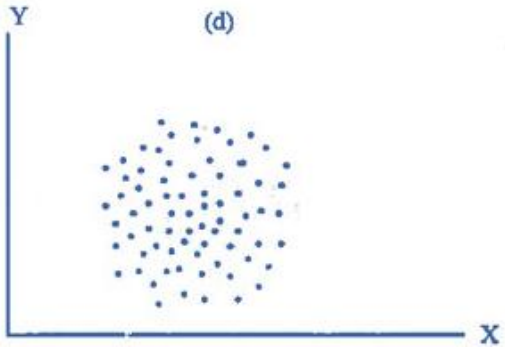
ในรูป (a) แนวโน้มของจุดชี้ขึ้นด้านขวาตามแนวเส้นตรง เมื่อ x มีค่ามาก y มีค่ามาก เมื่อ x มีค่าน้อย y มีค่าน้อย เรียกว่ามีความสัมพันธ์เชิงเส้นเชิงบวก (positive and linear correlation) หรือความสัมพันธ์แบบแปรตามกัน



ในรูป (b) แนวโน้มของจุดชี้ลงด้านขวาตามแนวเส้นตรง เมื่อ x มีค่ามาก y มีค่าน้อย เมื่อ x มีค่าน้อย y มีค่ามาก เรียกว่ามีความสัมพันธ์เชิงเส้นเชิงลบ (negative and linear correlation) หรือความสัมพันธ์แบบแปรผกผัน



ในรูป (c) แนวโน้มของจุดชี้ลงด้านขวาตามแนวเส้นโค้ง เรียกว่ามีความสัมพันธ์ไม่เชิงเส้นเชิงลบ (negative and nonlinear correlation)



ในรูป (d) แนวโน้มของจุดกระจายออกไม่มีแนวเส้นตรง เรียกว่าไม่มีความสัมพันธ์ (no correlation)

2. ค่าสัมประสิทธิ์สหสัมพันธ์

สำหรับตัวสถิติที่ใช้วัดค่าสหสัมพันธ์อย่างง่ายว่ามีความสัมพันธ์มากหรือน้อยเพียงใดคือสัมประสิทธิ์สหสัมพันธ์ (Correlation coefficient) ซึ่งในกรณีของสหสัมพันธ์อย่างง่ายตัวสถิตินี้เรียกว่าสัมประสิทธิ์สหสัมพันธ์อย่างง่าย (Simple Correlation Coefficient) เขียนแทนด้วยสัญลักษณ์ ρ หรือ ρ_{XY} ในกรณีที่เป็นค่าพารามิเตอร์ และ r หรือ r_{xy} ในกรณีที่เป็นค่าสถิติโดยที่ ρ และ r จะไม่มีหน่วยและมีค่าตั้งแต่ -1 ถึง 1

2.1 ความหมายของค่าสัมประสิทธิ์สหสัมพันธ์ ค่าสัมประสิทธิ์สหสัมพันธ์ที่คำนวณได้มีความหมายดังนี้

1. ถ้า ρ หรือ r มีค่าเป็นบวก แสดงว่าตัวแปร X และ Y มีความสัมพันธ์เชิงบวก หมายความว่าถ้าตัวแปร X มีค่าเพิ่มขึ้นตัวแปร Y จะมีค่าเพิ่มขึ้นหรือถ้าตัวแปร X มีค่าลดลงตัวแปร Y จะมีค่าลดลง

2. ถ้า ρ หรือ r มีค่าเป็นลบ แสดงว่าตัวแปร X และ Y มีความสัมพันธ์เชิงลบ หมายความว่าถ้าตัวแปร X มีค่าเพิ่มขึ้นตัวแปร Y จะมีค่าลดลง หรือถ้าตัวแปร X มีค่าลดลง ตัวแปร Y จะมีค่าเพิ่มขึ้น

3. ถ้า ρ หรือ r มีค่าเท่ากับ 1 แสดงว่าตัวแปร X และ Y มีความสัมพันธ์เชิงบวกอย่างสมบูรณ์ (perfect positive correlation)

4. ถ้า ρ หรือ r มีค่าเท่ากับ -1 แสดงว่าตัวแปร X และ Y มีความสัมพันธ์เชิงลบอย่างสมบูรณ์ (perfect negative correlation)

5. ถ้า ρ หรือ r มีค่าเข้าใกล้ 1 แสดงว่าตัวแปร X และ Y มีความสัมพันธ์เชิงบวกและมีความสัมพันธ์มาก

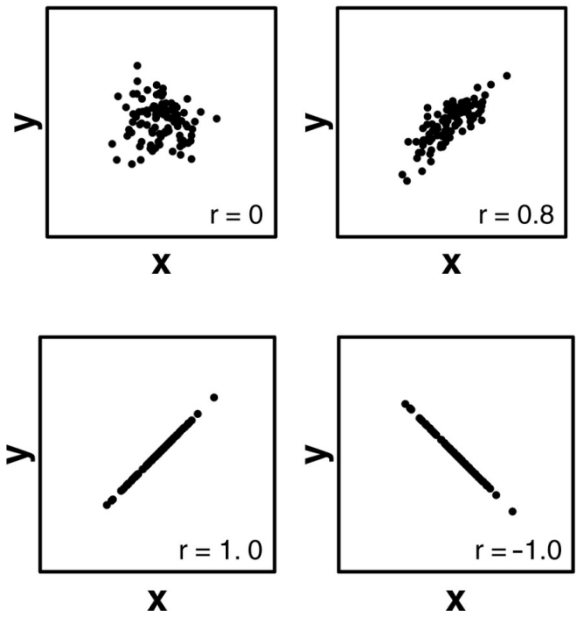
6. ถ้า ρ หรือ r มีค่าเข้าใกล้ -1 แสดงว่าตัวแปร X และ Y มีความสัมพันธ์เชิงลบและมีความสัมพันธ์มาก

7. ถ้า ρ หรือ r มีค่าเข้าใกล้ 0 แสดงว่าตัวแปร X และ Y มี ความสัมพันธ์กันน้อย

8. ถ้า ρ หรือ r มีค่าเท่ากับ 0 แสดงว่าตัวแปร X และ Y ไม่มีความสัมพันธ์เชิงเส้น

ตัวอย่างเช่นถ้าค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างความสูงกับน้ำหนักเท่ากับ 0.85 หมายความว่าความสูงมีความสัมพันธ์เชิงบวกอย่างมากกับน้ำหนัก ถ้าความสูงมากน้ำหนักจะมากด้วย และถ้าความสูงน้อยน้ำหนักจะน้อยด้วย เป็นต้น

ความหมายของ ρ หรือ r สามารถแสดงด้วยแผนภาพการกระจาย ดังนี้



2.2 วิธีการคำนวณหาค่าสัมประสิทธิ์สหสัมพันธ์ วิธีการคำนวณค่าสัมประสิทธิ์สหสัมพันธ์มีหลายวิธีขึ้นอยู่กับประเภทของข้อมูลหรือตัวแปร เช่นสัมประสิทธิ์สหสัมพันธ์ฟี สัมประสิทธิ์สหสัมพันธ์สเปียร์แมน สัมประสิทธิ์สหสัมพันธ์เพียร์สัน เป็นต้น

ในเอกสารฉบับนี้จะกล่าวเฉพาะการคำนวณหาค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน ซึ่งใช้หาความสัมพันธ์ระหว่างข้อมูลเชิงปริมาณ หรือตัวแปรแบบต่อเนื่อง สัมประสิทธิ์สหสัมพันธ์เพียร์สันเขียนแทนด้วยสัญลักษณ์ ρ ซึ่งคิดค้นโดย คาร์ล เพียรสั้น (Karl Pearson) บางครั้งอาจเรียกว่า Pearson Product moment correlation Coefficient ซึ่งสามารถคำนวณได้จากสูตร ดังนี้

$$\rho = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} \quad \text{หรือ} \quad \rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$= \frac{\sum_{i=1}^{N}(x_i - \mu_x)\,(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_x)^2 \sum_{i=1}^{N}(y_i - \mu_y)^2}}$$

โดยที่ ρ แทนสัมประสิทธิ์สหสัมพันธ์เพียร์สัน

cov(x,y) หรือ σ_{xy} แทนความแปรปรวนร่วมของตัวแปร X และ Y

σ_x, σ_y แทนส่วนเบี่ยงเบนมาตรฐานของตัวแปรX และ Y

μ_x, μ_y แทนค่าเฉลี่ยของตัวแปร X และ Y

ในการปฏิบัติเราเก็บรวบรวมข้อมูลจากตัวอย่าง ดังนั้นจะประมาณ ρ ด้วย r โดยที่

$$r = \frac{\text{cov}(x,y)}{s_x s_y}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$= \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n}(x_i)^2 - (\sum_{i=1}^{n} x_i)^2\right]\left[n\sum_{i=1}^{n}(y_i)^2 - (\sum_{i=1}^{n} y_i)^2\right]}}$$

โดยที่ r แทนสัมประสิทธิ์สหสัมพันธ์เพียร์สัน

cov(x,y) แทนความแปรปรวนร่วมของตัวแปร X และ Y

s_x , s_y แทนส่วนเบี่ยงเบนมาตรฐานของตัวแปร X และ Y

$\bar{x}$, $\bar{y}$ แทนค่าเฉลี่ยของตัวแปร X และ Y

n แทนจำนวนตัวอย่าง

2.3 ข้อตกลงเบื้องต้น ของสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน เป็นดังนี้

1. ตัวแปร 2 ตัว เป็นตัวแปรแบบต่อเนื่อง หรือเป็นข้อมูลเชิงปริมาณ
2. ความสัมพันธ์ระหว่าง 2 ตัวแปรเป็นเส้นตรง (Linear Relationship)

ตัวอย่าง 8.1 จากคะแนนสอบของนักศึกษาเอกคณิตศาสตร์ที่สอบวิชาแคลคูลัส 1 และแคลคูลัส2 จำนวน 12 คน เป็นดังตารางด้านล่างอยากทราบว่าคะแนนสอบทั้ง 2 วิชามีความสัมพันธ์กันเพียงใด

คนที่.	คะแนนสอบวิชา แคลคลัส 1 (Y)	คะแนนสอบวิชา แคลคูลัส 2 (X)
1	51	74
2	68	70
3	72	88
4	97	93
5	55	67
6	73	73
7	95	99
8	74	73
9	20	33
10	91	91
11	74	80
12	80	86
	$\sum y = 850$	$\sum x = 927$

วิธีทำ จากข้อมูล จะได้

y^2	x^2	xy
2601	5476	3774
4624	4900	4760
5184	7744	6336
9409	8649	9021
3025	7789	3685
5329	5329	5329
9025	9801	9045
5476	5329	5402
400	1089	660
8281	8281	8281
5476	6400	5920
6400	7396	6880
$\sum y^2 = 65230$	$\sum x^2 = 74883$	$\sum xy = 69453$

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n}(x_i)^2 - (\sum_{i=1}^{n} x_i)^2\right]\left[n\sum_{i=1}^{n}(y_i)^2 - (\sum_{i=1}^{n} y_i)^2\right]}}$$

$$= \frac{12(69453) - (927)(850)}{\sqrt{(12(74883) - (927)^2)\,(12(65230) - (850)^2)}}$$

$$= \frac{833436 - 787950}{\sqrt{(39267)(60260)}}$$

$$= \frac{45486}{\sqrt{2366229420}}$$

$$= 0.935$$

$r = 0.935$ หมายความว่าคะแนนสอบวิชาแคลคูลัส 1 และแคลคูลัส 2 ของนักศึกษา 12 คนมีความสัมพันธ์อย่างมากในเชิงบวก นั่นคือถ้าคะแนนสอบวิชาแคลคูลัส 1 เพิ่มขึ้นคะแนนสอบวิชาแคลคูลัส 2 จะเพิ่มขึ้น หรือถ้าคะแนนสอบวิชาแคลคูลัส 1 ลดลงคะแนนสอบวิชาแคลคูลัส 2 จะลดลงด้วย

การประมาณค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (ρ_{xy})

1. การประมาณค่าแบบจุดของสัมประสิทธิ์สหสัมพันธ์

ตัวประมาณแบบจุดของ ρ_{xy} คือ

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n}(x_i)^2 - (\sum_{i=1}^{n} x_i)^2\right]\left[n\sum_{i=1}^{n}(y_i)^2 - (\sum_{i=1}^{n} y_i)^2\right]}}$$

2. การประมาณค่าแบบช่วงของสัมประสิทธิ์สหสัมพันธ์

เนื่องจากค่าสัมประสิทธิ์สหสัมพันธ์มีค่าไม่เท่ากับ 0 ดังนั้นการแจกแจงของสัมประสิทธิ์สหสัมพันธ์จึงไม่เป็นสมมาตร เพื่อให้การแจกแจงของสัมประสิทธิ์สหสัมพันธ์เป็นสมมาตร จึงต้องเปลี่ยนค่า r_{xy} เป็น Z_r ตามวิธีการของ Fisher ดังนี้

$$Z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

จะได้ Z_r มีการแจกแจงแบบปกติด้วยค่าเฉลี่ย $\mu_{Zr} = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$ ความแปรปรวน $\sigma^2_{Zr} = \frac{1}{n-3}$

นั่นคือ $Z_r \sim N\left(\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{n-3}\right)$

ดังนั้นจะได้ $Z = \frac{Z_r - \mu_{Zr}}{\sigma_{Zr}}$

$= (Z_r - \mu_{Zr})\sqrt{n-3}$

จะได้ $(1-\alpha)100\%$ ของ μ_{Zr} คือ $Zr \pm z_{\frac{\alpha}{2}} \Big/ \sqrt{n-3}$

$Z_L = Zr - z_{\frac{\alpha}{2}} \Big/ \sqrt{n-3}$ และ $Z_u = Zr + z_{\frac{\alpha}{2}} \Big/ \sqrt{n-3}$

ในการประมาณค่าแบบช่วงของ ρ ต้องเปลี่ยนขอบเขตบนและขอบเขตล่างของการประมาณค่าแบบช่วงเป็นขอบเขตบนของ ρ (ρ_U) และขอบเขตล่างของ ρ (ρ_L) เสียก่อน โดยที่

$$\rho_U = \frac{e^{2Zu}-1}{e^{2Zu}+1} \quad \text{และ} \quad \rho_L = \frac{e^{2Z_L}-1}{e^{2Z_L}+1}$$

ดังนั้นช่วงความเชื่อมั่น $(1-\alpha)$ 100% ของ ρ คือ $\rho_L \le \rho \le \rho_U$

ตัวอย่าง 8.2 จากตัวอย่าง 8.1 คะแนนสอบวิชาแคลคูลัส 1 และแคลคูลัส 2 ของนักศึกษาเอกคณิตศาสตร์มีความสัมพันธ์กันในระดับใด ที่ความเชื่อมั่น 95 %

จาก
$$\begin{aligned} Z_r &= \frac{1}{2} ln\left(\frac{1+r}{1-r}\right) \\ &= \frac{1}{2} ln\left(\frac{1+0.935}{1-0.935}\right) \\ &= \frac{1}{2} ln\left(29.77\right) \\ &= \frac{1}{2}(3.39) \\ &= 1.67 \end{aligned}$$

จะได้ $(1-\alpha)$ 100% ของ μ_{Zr} คือ

$$Zr \pm z_{\frac{\alpha}{2}} \Big/ \sqrt{n-3} = (1.67) \pm \frac{1.96}{\sqrt{12-3}}$$

ดังนั้น $Zu = 2.32$ และ $Z_L = 1.02$

จาก
$$\rho_U = \frac{e^{2Zu}-1}{e^{2Zu}+1} \quad \text{และ} \quad \rho_L = \frac{e^{2Z_L}-1}{e^{2Z_L}+1}$$

$$\begin{aligned} \rho_U &= \frac{e^{2(2.32)}-1}{e^{2(2.32)}+1} & \rho_L &= \frac{e^{2(1.02)}-1}{e^{2(1.02)}+1} \\ &= \frac{102.54}{104.54} & &= \frac{6.69}{8.69} \\ &= 0.98 & &= 0.77 \end{aligned}$$

หมายความว่าที่ระดับความเชื่อมั่น 95 % สัมประสิทธิ์สหสัมพันธ์ของคะแนนสอบวิชาแคลคูลัส 1 และแคลคูลัส 2 มีค่าในช่วง 0.77 ถึง 0.98

การทดสอบสมมติฐานเกี่ยวกับสัมประสิทธิ์สหสัมพันธ์ (ρ)

เป็นการทดสอบว่าตัวแปร X และตัวแปร Y มีความสัมพันธ์กันหรือไม่

สมมติฐานเชิงสถิติ $H_0 : \rho = 0$ หรือ $H_0 :$ X และ Y ไม่มีความสัมพันธ์เชิงเส้นตรง

$H_1 : \rho \neq 0$ หรือ $H_1 :$ X และ Y มีความสัมพันธ์เชิงเส้นตรง

เนื่องจาก $\rho = 0$ จะได้ว่า r มีการแจกแจงแบบปกติ และมีส่วนเบี่ยงเบนมาตรฐาน คือ $\sqrt{(1-r^2)/(n-2)}$

ตัวสถิติทดสอบ

$$T = \frac{r-0}{\sqrt{(1-r^2)/n-2}}$$

ค่าวิกฤต $-t_{\frac{\alpha}{2},\,n-2}$, $t_{\frac{\alpha}{2},\,n-2}$

ในบางกรณีความต้องการทดสอบสมมติฐานว่า ส.ป.ส สหสัมพันธ์มีค่าเท่ากับค่าใดค่าหนึ่งหรือไม่ (หมายความว่า $\rho_0 \neq 0$) จะต้องเปลี่ยนค่า r ให้เป็น Zr_r เสียก่อนโดยที่

$$Zr = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right) \qquad \text{และ} \quad Zr \sim N\left(\frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right) , \frac{1}{n-3}\right)$$

ดังนั้น ตัวสถิติทดสอบ $Z = \dfrac{Zr - Z\rho_0}{\sqrt{1/(n-3)}}$

โดยที่ $Z\rho_0 = \dfrac{1}{2} ln\left(\dfrac{1+\rho_0}{1-\rho_0}\right)$

ค่าวิกฤต $-z_{\frac{\alpha}{2}}$, $z_{\frac{\alpha}{2}}$

ตัวอย่าง 8.3 จากตัวอย่าง 8.1 ต้องการทดสอบว่าคะแนนสอบวิชาแคลคูลัส 1 และแคลคูลัส 2 มีความสัมพันธ์ในรูปเชิงเส้นตรงหรือไม่ ที่ระดับนัยสำคัญ 0.05

H_0: คะแนนวิชาแคลคูลัส 1 และแคลคูลัส 2 ไม่มีความสัมพันธ์เชิงเส้นตรง

H_1: คะแนนวิชาแคลคูลัส 1 และแคลคูลัส 2 มีความสัมพันธ์เชิงเส้นตรง

หรือ $H_0 : \rho = 0$

$H_1 : \rho \neq 0$

ตัวสถิติทดสอบ $T = \dfrac{r - 0}{\sqrt{(1-r^2)/n-2}}$

$= \dfrac{0.935}{\sqrt{(1-0.935^2)/12-2}}$

$= \dfrac{0.935}{\sqrt{0.126/10}}$

$= 8.33$

ค่าวิกฤต $-t_{\frac{\alpha}{2}, n-2} = -t_{0.025,10} = -2.228$

$t_{\frac{\alpha}{2}, n-2} = t_{0.025,10} = 2.228$

เนื่องจาก T = 8.33 มากกว่า $t_{0.025,10} = 2.228$ มีค่าอยู่ในบริเวณปฏิเสธ H_0 หมายความว่า คะแนนวิชาแคลคูลัส 1 และแคลคูลัส 2 มีความสัมพันธ์เชิงเส้นตรง อย่างมีนัยสำคัญที่ระดับ 0.05

การวิเคราะห์การถดถอย (Regression Analysis)

เป็นการศึกษาถึงความสัมพันธ์ของตัวแปรตั้งแต่ 2 ตัวขึ้นไป โดยมีวัตถุประสงค์ที่จะประมาณหรือพยากรณ์ค่าของตัวแปรหนึ่ง จากตัวแปรอื่น ๆ ที่มีความสัมพันธ์กับตัวแปรที่ต้องการพยากรณ์ โดยจะต้องมีการกำหนดหรือทราบค่าของตัวแปรอื่น ๆ ล่วงหน้า ในที่นี้เรียกว่าตัวแปรอิสระ (Independence Variable) จึงจะทำให้ทราบค่าของตัวแปรอีกตัวหนึ่ง ซึ่งในที่นี้เรียกว่าตัวแปรตาม (Dependence Variable) เช่น เมื่อทราบว่าราคาที่ดินกับขนาดของที่ดินมีความสัมพันธ์กันแล้ว และคิดว่าราคาที่ดินขึ้นอยู่กับขนาดของที่ดินหรือขนาดของที่ดินมีผลต่อราคาที่ดิน นั่นคือขนาดที่ดินเป็นตัวแปรอิสระ และราคาที่ดินเป็นตัวแปรตาม หมายความว่าเมื่อกำหนดขนาดของที่ดินจะทำให้ประมาณหรือพยากรณ์ราคาที่ดินนั้นได้ และสามารถศึกษาการเปลี่ยนแปลงของราคาที่ดินเมื่อขนาดของที่ดินเปลี่ยนแปลงไป โดยอาศัยหลักการของการวิเคราะห์การถดถอย ดังนั้นวัตถุประสงค์ของ การวิเคราะห์การถดถอย คือพยากรณ์ค่าตัวแปรตามในอนาคต เมื่อกำหนดค่าตัวแปรอิสระ

หมายเหตุ โดยทั่วไปมักให้ตัวแปรอิสระแทนด้วย X และตัวแปรตามแทนด้วย Y

การที่จะนำตัวแปร X มาพยากรณ์ Y ได้นั้น ตัวแปร X และ Y จะต้องมีความสัมพันธ์กัน (จากแผนภาพการกระจาย) ถ้าตัวแปรX และ Y มีความสัมพันธ์เชิงเส้นตรงและ ค่าของ Y ขึ้นอยู่กับค่าของ X จะเรียกการถดถอยนั้นว่าการวิเคราะห์การถดถอยเชิงเส้น (Linear Regression Analysis)

ถ้าตัวแปร X และ Y มีความสัมพันธ์เชิงเส้นตรง และตัวแปร Y ถูกพยากรณ์ด้วยตัวแปร X เพียง 1 ตัวจะเรียกว่า การถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression)

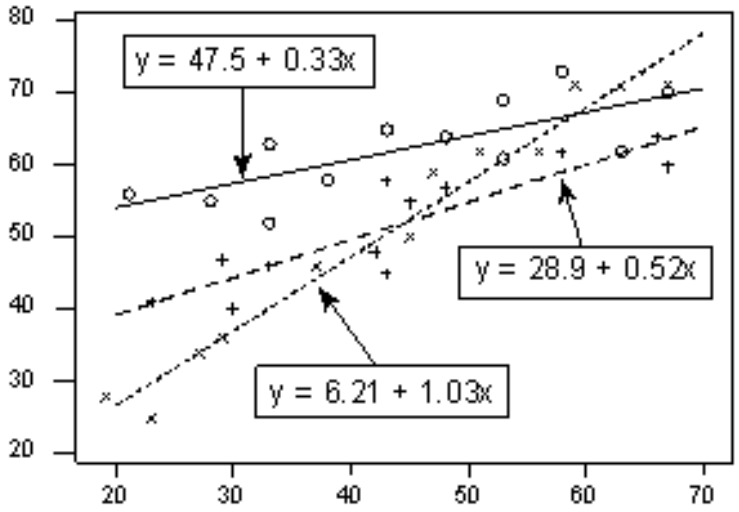
ถ้าตัวแปร X และ Y มีความสัมพันธ์เชิงเส้นตรง และตัวแปร Y ถูกพยากรณ์ด้วยตัวแปร X มากกว่า 1 ตัวจะเรียกว่าการถดถอยเชิงเส้นพหุคูณ (Multiple Linear Regression)

ซึ่งในเอกสารฉบับนี้จะศึกษาเฉพาะ Simple Linear Regression เท่านั้น

การวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression Analysis)

หลักการของการวิเคราะห์การถดถอยคือใช้ข้อมูลเมื่อกำหนดค่าตัวแปรอิสระ (X) แล้วทำให้เกิดค่าตัวแปรตาม (Y) ในอดีตนำมาสร้างสมการเชิงเส้นที่เหมาะสมที่สุด เพื่อพยากรณ์ค่า Y ในอนาคต

แต่ในบางครั้งค่าของ x และ y ที่เกิดขึ้นนั้นอาจสร้างสมการเชิงเส้นตรงได้หลายเส้น คู่อันดับ (x,y) อาจไม่อยู่ในแนวเส้นตรงเดียวกันทั้งหมด ดังรูป



ตัวอย่าง 8.4 รายจ่ายมักขึ้นอยู่กับรายได้เสมอ หากต้องการพยากรณ์ค่าของรายจ่ายในอนาคตเมื่อกำหนดรายได้จึงทำการบันทึกข้อมูลรายได้ และรายจ่ายของครอบครัว โดยที่

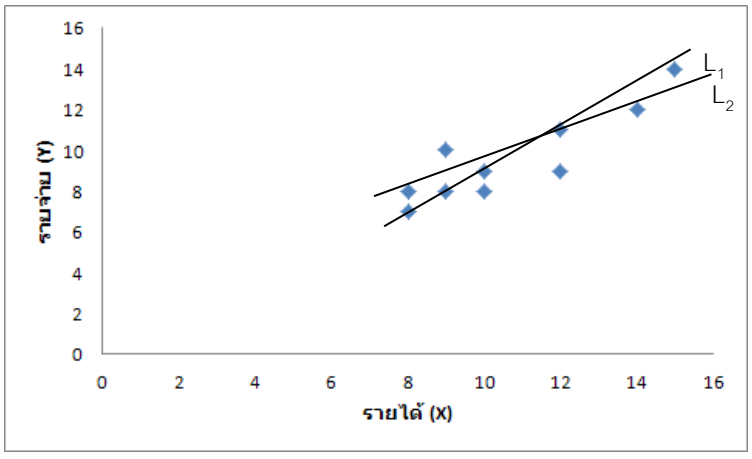
X เป็นตัวแปรอิสระแทนเงินเดือน (พันบาท)

Y เป็นตัวแปรตามแทนรายจ่าย (พันบาท)

ได้ข้อมูลดังนี้

ครอบครัวที่	รายได้ X (พันบาท)	รายจ่าย Y (พันบาท)
1	8	7
2	8	8
3	9	8
4	9	10
5	10	8
6	10	9
7	12	9
8	12	11
9	14	12
10	15	14

เมื่อนำข้อมูลนี้มาเขียน scatter diagram จะได้ดังรูป



จะเห็นว่าจากข้อมูลข้างต้นอาจสามารถสร้างสมการเชิงเส้นตรงได้ 2 เส้นคือ L_1 และ L_2 ถ้ากำหนดให้สมการทั่วไปของ L_1 และ L_2 คือ

$L_1 : Y = A_1 + B_1 x$

$L_2 : Y = A_2 + B_2 x$

การพิจารณาด้วยสายตาว่าควรใช้ L_1 หรือ L_2 ในการพยากรณ์ค่า Y นั้นทำให้ตัดสินใจลำบาก จึงมีแนวคิดทางคณิตศาสตร์มาช่วยในการตัดสินใจเลือกสมการเชิงเส้นตรงที่เหมาะสมที่สุด วิธีนั้นคือวิธีกำลังสองน้อยที่สุด (Least Square Method)

1. วิธีกำลังสองน้อยที่สุด (Least Square Method)

หลักการของวิธีกำลังสองน้อยที่สุด คือเลือกสมการถดถอยเชิงเส้นที่ให้ค่าความคลาดเคลื่อนระหว่างค่าสังเกตของ y (ค่าจริง) กับค่าประมาณของ y หรือค่าจากสมการ $(\hat{y})$ น้อยที่สุด พิจารณาข้อมูลต่อไปนี้

ลำดับที่	x	y
1	1	10
2	2	30
3	3	50

สามารถสร้างสมการเชิงเส้นได้ 3 สมการ

$$L_1 : \hat{y}_1 = 10 + 10x$$

$$L_2 : \hat{y}_2 = 15x$$

$$L_3 : \hat{y}_3 = -10 + 20x$$

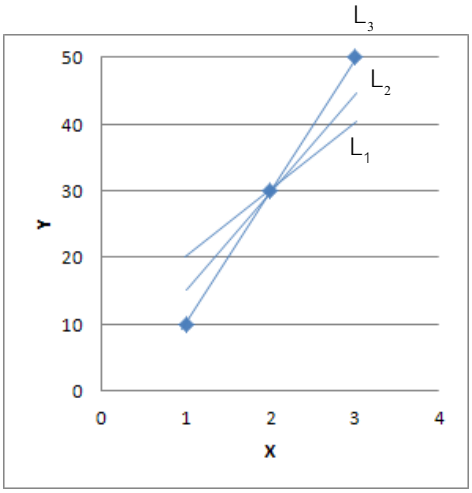
คำนวณหาความคลาดเคลื่อนระหว่าง y กับ $\hat{y}$

					ความคลาดเคลื่อน		
x	y	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$	$y-\hat{y}_1$	$y-\hat{y}_2$	$y-\hat{y}_3$
1	10	20	15	10	-10	5	0
2	30	30	30	30	0	0	0
3	50	40	45	50	10	5	0
	ผลรวมความคลาดเคลื่อน				0	0	0

เมื่อนำมาคำนวณกำลัง 2 ของความคลาดเคลื่อนจะได้ ดังนี้

	$(y-\hat{y}_1)^2$	$(y-\hat{y}_2)^2$	$(y-\hat{y}_3)^2$
	100	25	0
	0	0	0
	100	25	0
ผลรวม	200	50	0

จะเห็นว่าสมการเชิงเส้น L_3 ยังให้ค่าผลรวมกำลังสองต่ำสุด และเมื่อพิจารณาจากกราฟ



จะเห็นว่า L_3 เป็นเส้นกราฟดีที่สุด โดยปกติเรานิยามว่าเส้นตรงที่ดีควรมีค่าความคลาดเคลื่อนระหว่างเส้นตรงกับจุดข้อมูลน้อยที่สุด และเรียกเส้นตรงนั้นว่า สมการถดถอยเชิงเส้นอย่างง่าย วิธีการกำลังสองน้อยที่สุดจะเลือกสมการที่ให้ความคลาดเคลื่อนระหว่าง y กับ $\hat{y}$ น้อยที่สุดเช่นกัน

สรุปสมการถดถอยเชิงเส้นอย่างง่ายที่จะเป็นตัวแทนแสดงความสัมพันธ์เชิงเส้นระหว่างตัวแปรอิสระ (X) และตัวแปรตาม (Y) จึงหมายถึงสมการเส้นตรงที่ทำให้ผลรวมกำลังสองของความคลาดเคลื่อน $(\sum(y-\hat{y})^2)$ มีค่าต่ำสุด โดยอาศัยวิธีการกำลังสองน้อยที่สุด

2. ประโยชน์ของสมการถดถอยเชิงเส้น

สมการถดถอยเชิงเส้นสามารถใช้ประโยชน์ในการแสดงความสัมพันธ์ของตัวแปร X กับ Y ใน 2 ประการ ดังนี้

1. ใช้ในการประมาณค่า (estimation) เช่นในการศึกษาค่าสัมพันธ์ระหว่างปริมาณหมึกที่ใช้กับความเร็วที่พิมพ์

ถ้าให้ Y แทนปริมาณหมึกที่ใช้

X แทนความเร็วในการพิมพ์

เราอาจทำการทดลองหาปริมาณหมึกที่ใช้ในแต่ละระดับของความเร็วในการพิมพ์แล้วนำมาใช้ในการหาสมการประมาณค่าปริมาณหมึกที่ใช้ในระดับความเร็วต่าง ๆ

2. ใช้ในการพยากรณ์ (forecasting) เช่น ถ้า X คือปี พ.ศ. Y คือปริมาณการส่งออกของผลไม้ไทย การบันทึกปริมาณการส่งออกของผลไม้ไทยในปีต่าง ๆ ที่ผ่านมา ทำให้เราสามารถใช้สมการถดถอยพยากรณ์ปริมาณการส่งออกของผลไม้ไทยใน 5 หรือ 10 ปีข้างหน้าได้

3. ตัวแบบ (Model) ข้อสมมติ (Assumptions) และวิธีการหาสมการถดถอยเชิงเส้นอย่างง่าย

3.1 ตัวแบบ

โดยทั่วไปสมการถดถอยเชิงเส้นอย่างง่ายจะมีตัวแบบดังนี้

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

โดยที่ β_0 เป็นจุดตัดแกน Y หรือค่าของ y เมื่อ x=0

β_1 คืออัตราการเปลี่ยนแปลงของ Y เมื่อ X เปลี่ยนไป 1 หน่วย หรือเรียกว่า สัมประสิทธิ์ถดถอย (regression coefficient)

ε คือ ค่าคลาดเคลื่อนเชิงสุ่ม (error)

3.2 ข้อสมมติของสมการถดถอยเชิงเส้น

จากหลักการของการถดถอยเชิงเส้น คือกำหนดค่าของตัวแปรอิสระ X ทำให้เกิดตัวแปรตาม Y จะเห็นว่าตัวแปร X เป็นตัวแปรที่ใช้ควบคุม (Fixed Variable) ถูกกำหนดค่า หรือถูกสังเกต หรือมีผลต่อตัวแปร Y ดังนั้น Y จึงเป็นตัวแปรสุ่ม เช่น

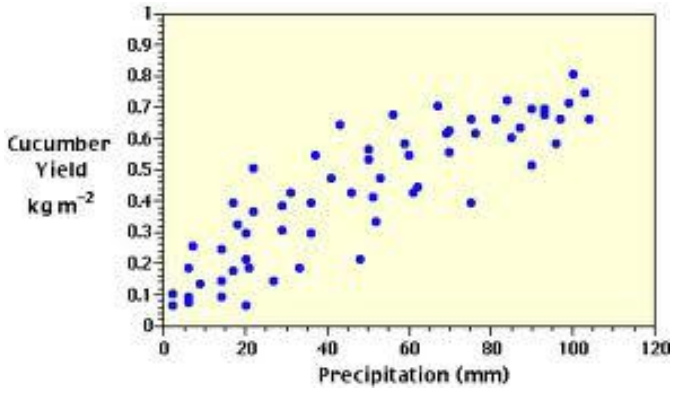
ปริมาณปุ๋ยมีผลต่อปริมาณผลผลิตข้าว

X	Y
(ควบคุม)	(ตัวแปรตาม)

คะแนนสอบเข้ามีผลต่อเกรดเฉลี่ยชั้นปีที่ 1

X	Y
(สังเกต)	(ตัวแปรตาม)

จะเห็นว่าตัวแปรอิสระ 1 ค่า สามารถมีค่า Y ที่เป็นไปได้มากมายหลายค่า ดังรูป



ซึ่งการใช้สมการถดถอยทำการประมาณ ค่า Y นั้น จึงทำให้เกิดความคลาดเคลื่อนขึ้น (ε_i) ซึ่งความคลาดเคลื่อนนี้ประกอบด้วยความคลาดเคลื่อน 2 ส่วนใหญ่ ๆ คือ

ส่วนที่ 1 ความคลาดเคลื่อนสุ่มซึ่งเกิดจากการที่มีปัจจัยอื่น ๆ ที่มีผลต่อตัวแปรตามแต่ไม่ได้นำมาร่วมกัน และความผิดพลาดในการวัดค่า (พิจารณาเฉพาะความคลาดเคลื่อนในตัวแปรตาม)

ส่วนที่ 2 ความคลาดเคลื่อนเนื่องจากใช้ตัวแบบไม่เหมาะสม (Lack of fit) เช่น ความสัมพันธ์ระหว่าง x กับ y จริง ๆ เป็นเส้นโค้ง แต่ตัวแบบที่ใช้ในการวิเคราะห์เป็นเส้นตรง เป็นต้น (ความคลาดเคลื่อนประเภทนี้จะไม่กล่าวถึงรายละเอียด ดังนั้นสำหรับการวิเคราะห์การถดถอยต่อไปนี้จะสมมติว่าตัวแบบที่ใช้เหมาะสม)

ด้วยเหตุผลนี้ในการหาสมการถดถอยเชิงเส้นโดยวิธีกำลังสองน้อยที่สุดจึงมีข้อสมมติเกี่ยวกับความคลาดเคลื่อน (ε_i) ดังต่อไปนี้

1. ε_i ของตัวแปรอิสระแต่ละค่าไม่สัมพันธ์กัน

2. ε_i ของตัวแปรแต่ละค่าเป็นตัวแปรสุ่มและ $\varepsilon_i \sim N(0,\sigma^2)$ เนื่องจาก $Y \sim N(0,\sigma^2)$ ภายใต้เงื่อนไขข้อนี้ จะต้องนำไปใช้ในการทดสอบสมมติฐานและการหาช่วงเชื่อมั่นต่อไป

ข้อสมมติในข้อ 2 เป็นส่วนสำคัญมาก นั่นคือถ้า $\varepsilon_i \sim N(0,\sigma^2)$ ไม่เป็นจริงวิธีการที่ใช้ในการประมาณค่า β_0, β_1 จะไม่เป็นไปตามที่ควรเป็น ดังนั้นควรตรวจสอบเสมอว่า $\varepsilon_i \sim N(0,\sigma^2)$ หรือไม่

3.3 การสร้างสมการถดถอยเชิงเส้นอย่างง่าย

จากตัวแบบสมการถดถอยเชิงเส้น $Y = \beta_0 + \beta_1 X + \varepsilon$ เป้าหมายคือต้องการหาค่า β_0, β_1 ที่ทำให้ผลรวมของความคลาดเคลื่อนเป็นศูนย์ ($\sum \varepsilon_i = 0$) และผลรวมกำลังสองของความคลาดเคลื่อน ($\sum \varepsilon_i^2$) น้อยที่สุด แต่ในความเป็นจริงเราไม่สามารถหาค่าพารามิเตอร์ β_0, β_1 จึงต้องอาศัยค่าจากตัวอย่างเพื่อหาค่าประมาณของ β_0, β_1 ในตัวแบบสมการถดถอยเชิงเส้นจากตัวอย่าง ดังนี้

$$\hat{y} = b_0 + b_1 \times + e_i$$

เมื่อ $\hat{y}$ แทนค่าประมาณของ y

b_0 แทนค่าประมาณของ β_0

b_1 แทนค่าประมาณของ β_1

e_i แทนค่าประมาณของ ε_i หรือเรียกว่าส่วนเหลือ (residual)

3.4 การประมาณค่าพารามิเตอร์ β_0 , β_1

การหาค่า b_0 และ b_1 มีด้วยกันหลายวิธี แต่ค่าประมาณที่ได้จากวิธีกำลังสองน้อยที่สุดจะทำให้ผลรวมของกำลังสองของผลต่างระหว่าง y กับ $\hat{y}$ มีค่าน้อยที่สุด จากวิธีกำลังสองน้อยที่สุดจะได้

$$b_1 = \frac{n\sum x_i y_i - \sum y_i \sum x_i}{n\sum x_i^2 - (\sum x_i)^2}$$

และ $b_0 = \bar{y} - b_1\bar{x}$

ตัวอย่าง 8.5 จากตัวอย่าง 8.4 ข้อมูลรายจ่ายกับรายได้ จงหาสมการถดถอยเชิงเส้นอย่างง่าย

ตัวแบบสมการถดถอยเชิงเส้นอย่างง่าย $\hat{y} = b_0 + b_1 \times + e_i$

$$b_1 = = \frac{n\sum x_i y_i - \sum y_i \sum x_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$= \frac{10(1070) - (96)(107)}{10(1199) - (107)^2}$$

$$= 0.791$$

$$b_0 = \bar{y} - b_1\bar{x}$$

$$= (9.6)-(0.791)(10.7)$$

$$= 1.1363$$

ดังนั้นสมการถดถอยเชิงเส้น คือ $\hat{y}$ = 1.1363 + 0.791X หมายความว่าถ้ารายได้เพิ่มขึ้น 1 พันบาท รายจ่ายจะเพิ่มขึ้น 0.791 พันบาท หรือถ้ารายได้ลดลง 1 พันบาท รายจ่ายจะลดลง 0.791 พันบาท

การประมาณค่าและการทดสอบสมมติฐาน

1. การประมาณค่า β_0

ช่วงความเชื่อมั่น 100(1-α)% ของ β_0 คือ

$$b_0 \pm t_{\alpha/2,\, n-2} \quad s\sqrt{\frac{\sum x_i^2}{n\sum (x_i - \bar{x})^2}}$$

2. การทดสอบสมมติฐานของ β_0 (ต้องการทราบว่าเส้นถดถอยนี้ผ่านจุดกำเนิดหรือไม่)

สมมติฐานเชิงสถิติ $H_0 : \beta_0 = \beta_{00}$ (ค่าคงที่)

$H_1 : \beta_0 \neq \beta_{00}$

$H_1 : \beta_0 > \beta_{00}$

$H_1 : \beta_0 < \beta_{00}$

ตัวสถิติทดสอบ $$t = \frac{b_0 - \beta_{00}}{s\sqrt{\frac{\sum x_i^2}{n\sum (x_i - \bar{x})^2}}}$$

ค่าวิกฤติ $\pm t_{\frac{\alpha}{2},n-2}$ หรือ $t_{\alpha,n-2}$ หรือ $-t_{\alpha,n-2}$

3. การประมาณค่า β_1

ช่วงความเชื่อมั่น 100(1-α)% ของ β_1 คือ

$$b_1 \pm t_{\frac{\alpha}{2},n-2}\ s / \sqrt{\sum (x_i - \bar{x})^2}$$

4. การทดสอบสมมติฐานของ β_1 (ต้องการทดสอบว่าตัวแปร X มีผลต่อตัวแปร Y หรือไม่)

สมมติฐาน $H_0 : \beta_1 = \beta_{10}$ (ค่าคงที่)

$H_1 : \beta_1 > \beta_{10}$ (มีความสัมพันธ์ในทางบวก)

$H_1 : \beta_1 < \beta_{10}$ (มีความสัมพันธ์ในทางลบ)

$H_1 : \beta_1 \neq \beta_{10}$ (มีความสัมพันธ์กัน)

ตัวสถิติทดสอบ $$t = \frac{b_1 - \beta_{10}}{s / \sqrt{\sum (x_i - \bar{x})^2}}$$

ค่าวิกฤติ $t_{\alpha,n-2}$ หรือ $-t_{\alpha,n-2}$ หรือ $\pm t_{\frac{\alpha}{2},n-2}$

ตัวอย่าง 8.6 จากตารางแสดงจำนวนชั่วโมงเฉลี่ยในการทำการบ้านต่อสัปดาห์ของนักศึกษา 6 คน กับเกรดเฉลี่ยที่นักศึกษาได้ในภาคเรียนนั้น เป็นดังนี้

จำนวนชม.เฉลี่ยในการทำการบ้าน (X)	15	28	13	20	4	10
เกรดเฉลี่ย (Y)	2	2.7	1.3	1.9	0.9	1.7

จาก $s_{xy} = \sum xy - \frac{1}{n}(\sum x)(\sum y) = 23.6$

$s_{xx} = \sum x^2 - \frac{1}{n}(\sum x^2) = 344$

$s_{yy} = \sum y^2 - \frac{1}{n}(\sum y^2) = 1.915$

$r_{xy} = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}$

หรือ $r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}}$

$= 0.919453$

r = 0.919 หมายความว่าจำนวนชั่วโมงเฉลี่ยในการทำการบ้านและเกรดเฉลี่ยมีความสัมพันธ์เชิงเส้นเชิงบวกกันค่อนข้างมาก นั่นคือถ้าจำนวนชั่วโมงเฉลี่ยในการทำการบ้านเพิ่มขึ้นเกรดเฉลี่ยจะเพิ่มขึ้นด้วย และถ้าจำนวนชั่วโมงเฉลี่ยในการทำการบ้านลดลงเกรดเฉลี่ยจะลดลงด้วย

ประมาณค่า ρ โดย แปลง r เป็น $Z_r = \frac{1}{2} ln \frac{(1+r)}{(1-r)} = 1.59$

ที่ระดับความเชื่อมั่น 95% ช่วงความเชื่อมั่นของ μ_{Z_r} คือ

$$Zr \pm \frac{z_{\alpha/2}}{\sqrt{n-3}} = 1.59 \pm \frac{(1.96)}{\sqrt{6-3}}$$
$$= 1.59 \pm 1.13$$

$\therefore$ จะได้ช่วงเชื่อมั่น 95% μ_{Z_r} คือ $0.46 \le \mu_{Z_r} \le 2.72$

ที่ระดับความเชื่อมั่น 95% ช่วงความเชื่อมั่นของ ρ คือ $\rho_L \le \rho \le \rho_U$ โดยที่

$$\rho_U = \frac{e^{2z_u} - 1}{e^{2z_u} + 1} \qquad \text{และ} \qquad \rho_L = \frac{e^{2z_L} - 1}{e^{2z_L} + 1}$$
$$= \frac{229.44}{231.44} \qquad\qquad = \frac{1.51}{3.51}$$
$$= 0.99 \qquad\qquad = 0.43$$

$0.43 \le \rho \le 0.99$ หมายความว่าที่ระดับความเชื่อมั่น 95% สัมประสิทธิ์สหสัมพันธ์ของจำนวนชั่วโมงเฉลี่ยในการทำการบ้านกับเกรดเฉลี่ยมีค่าประมาณ 0.43 ถึง 0.99

หาค่า b_0 และ b_1

จาก
$$b_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$
$$= 0.06860$$

$$b_0 = \bar{y} - b_1\bar{x}$$
$$= 0.72093$$

$\therefore$ สมการถดถอยเชิงเส้นคือ $\hat{y} = 0.72093 + 0.0686x$ หมายความว่าถ้าไม่มีชั่วโมงในการทำการบ้านเลยจะได้เกรดเฉลี่ย 0.72093 และถ้าเพิ่มชั่วโมงในการทำการบ้าน 1 ชั่วโมง เกรดเฉลี่ยจะเพิ่มขึ้น 0.0686

ทดสอบสมมติฐาน $H_0 : \beta_1 = 0$ $H_1 : \beta_1 \ne 0$ เมื่อ $\alpha = 0.05$

ตัวสถิติทดสอบ

$$t = \frac{b_1 - \beta_1}{s / \sqrt{\sum (x_i - \bar{x})^2}} \qquad \text{โดยที่} \quad s = \sqrt{\frac{SSE}{n-2}} = 0.272$$
$$= \frac{0.0686 - 0}{0.272 / \sqrt{344}}$$
$$= 4.678$$

ค่าวิกฤต $t_{0.025,24} = 2.776$

$\therefore$ เนื่องจาก t = 4.678 อยู่ในบริเวณปฏิเสธ H_0 แสดงว่า $\beta_1 \ne 0$ นั่นคือ x และ y มีความสัมพันธ์เชิงเส้น ที่ระดับนัยสำคัญ 0.05

ประมาณค่า β_1

ที่ระดับความเชื่อมั่น 95% ช่วงความเชื่อมั่นของ β_1คือ

$$b_1 \pm t_{\frac{\alpha}{2},n-2} \Big/ \sqrt{s_{xx}} = b_1 \pm 2.776(0.014665)$$

$$= 0.06860 \pm 0.0471$$

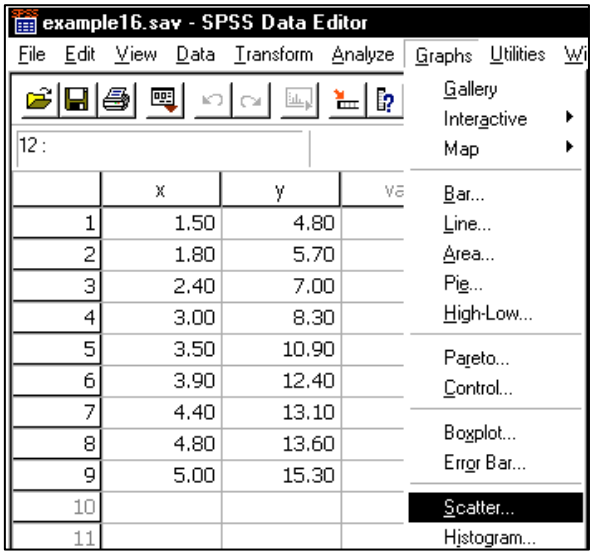
ดังนั้นที่ระดับความเชื่อมั่น 95% ช่วงความเชื่อมั่นของ β_1คือ $0.0215 \leq \beta_1 \leq 0.1157$

การใช้โปรแกรมสำเร็จรูป SPSS

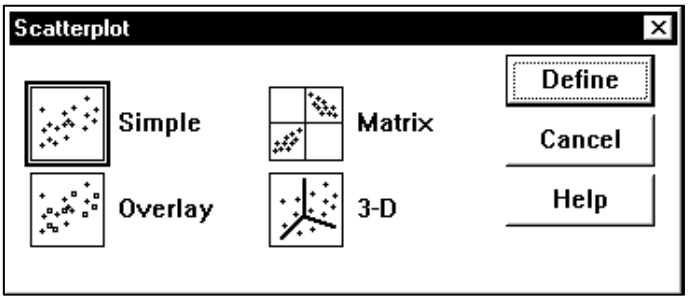
1. การตรวจสอบความสัมพันธ์ด้วยแผนภาพการกระจาย

ตัวอย่าง 8.7 จงตรวจสอบความสัมพันธ์ของตัวแปร X และ Y ในแฟ้มข้อมูล example16.sav

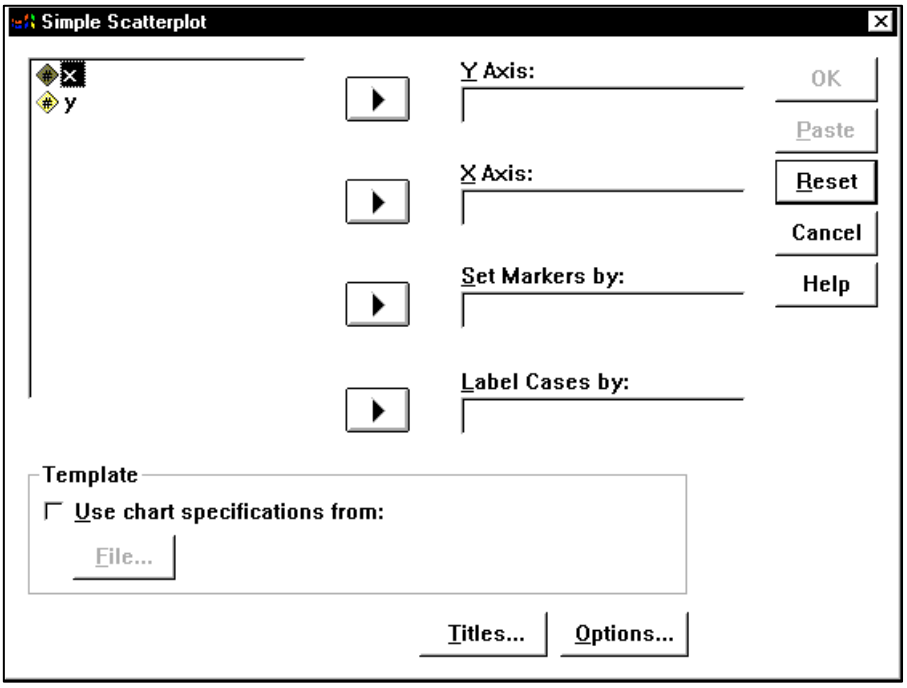
ขั้นที่ 1 เปิดแฟ้มข้อมูล example16.sav เลือกคำสั่ง Graphs / Scatter…



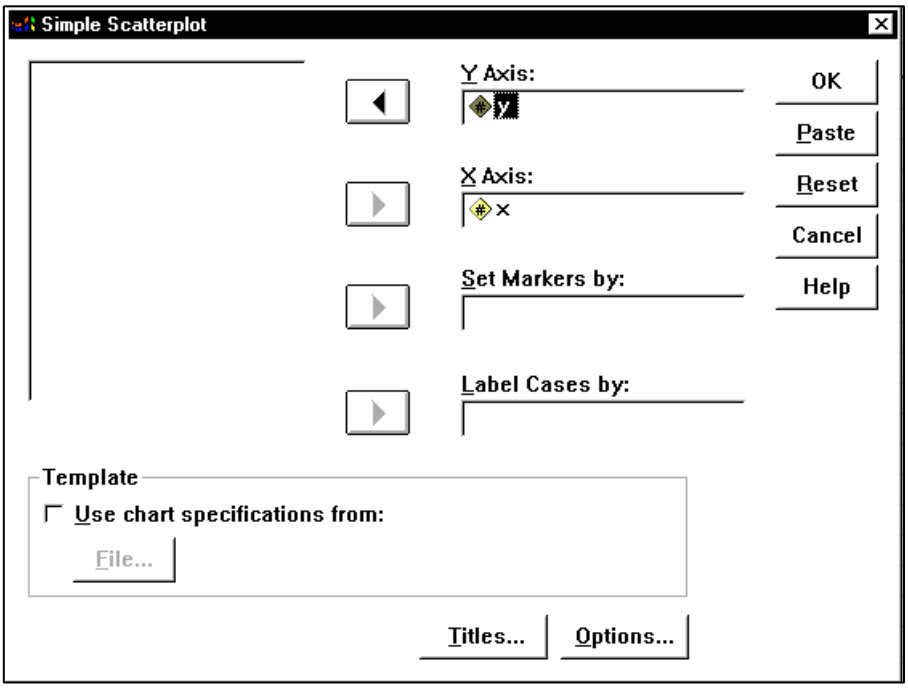
ขั้นที่ 2 เลือกคำสั่ง Scatter จะได้เมนูย่อยเป็นดังนี้



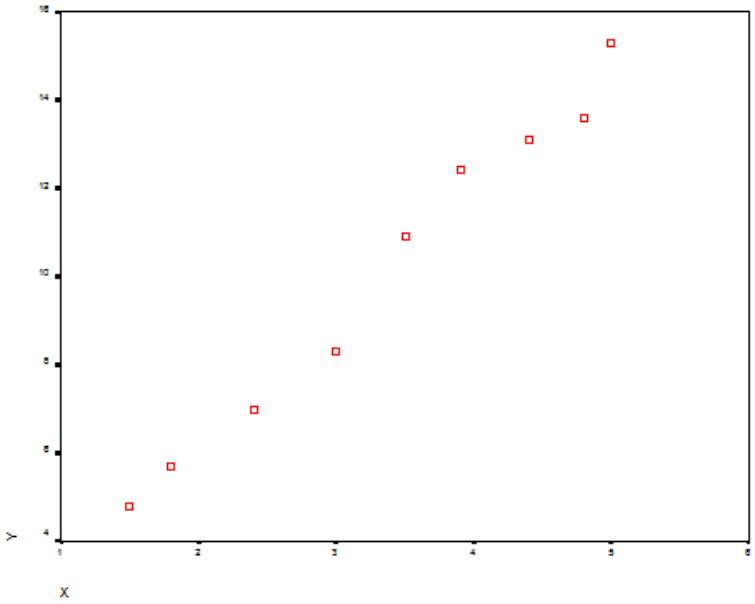
ขั้นที่ 3 เลือกรูปแบบกราฟเป็น Simple เสร็จแล้วคลิกปุ่ม Define จะได้เมนูย่อยเป็น



ขั้นที่ 4 เลือกตัวแปร x ไว้ที่ X Axis และตัวแปร y ไว้ที่ Y Axis



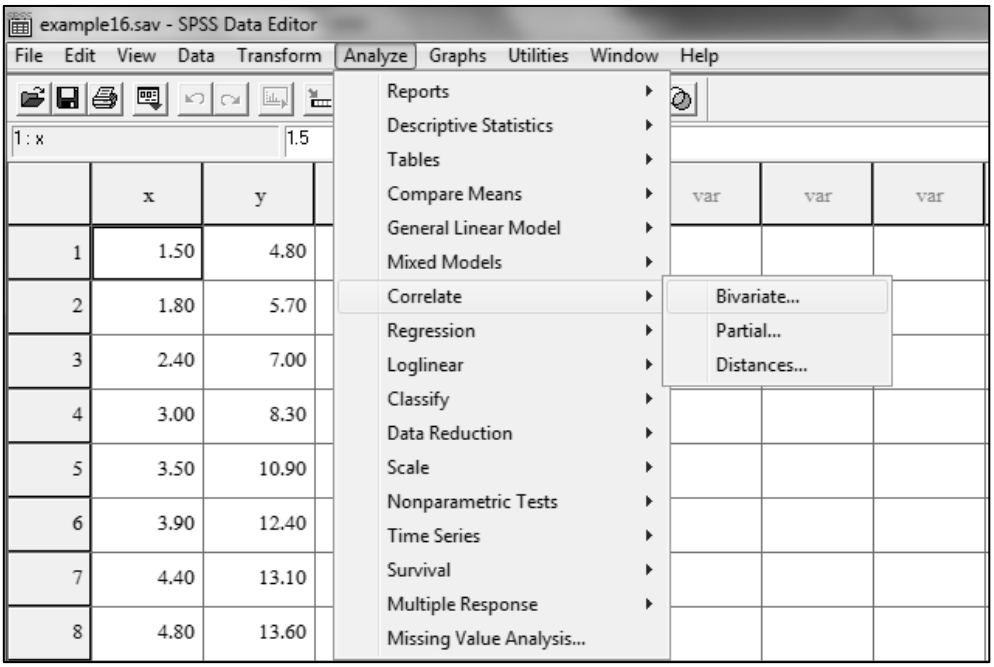
ขั้นที่ 5. คลิก OK จะได้กราฟของแผนภาพการกระจายที่ SPSS Viewer ดังนี้



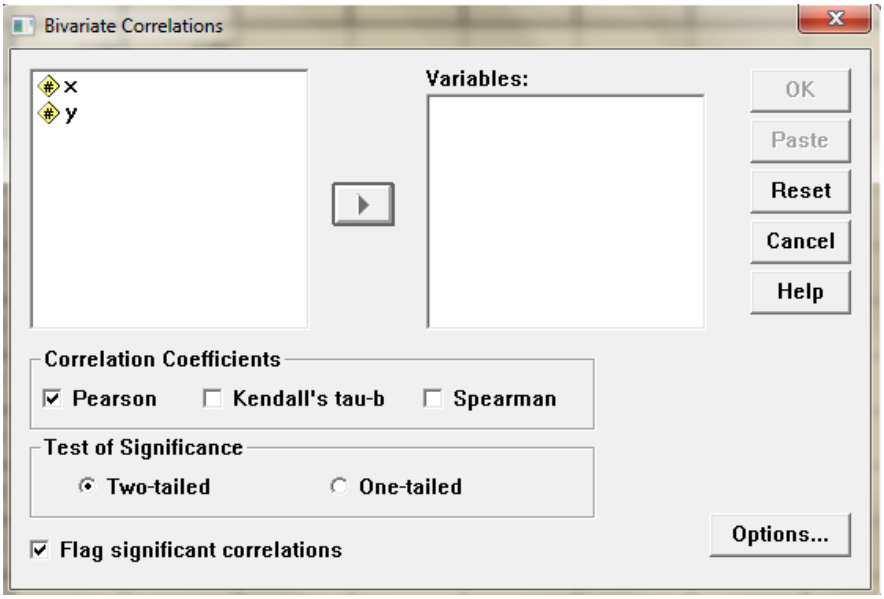
2. การหาค่าสัมประสิทธิ์สหสัมพันธ์

ตัวอย่าง 8.8 จงหาค่าสัมประสิทธิ์สหสัมพันธ์ของตัวแปร X และ Y ในแฟ้มข้อมูล example16.sav

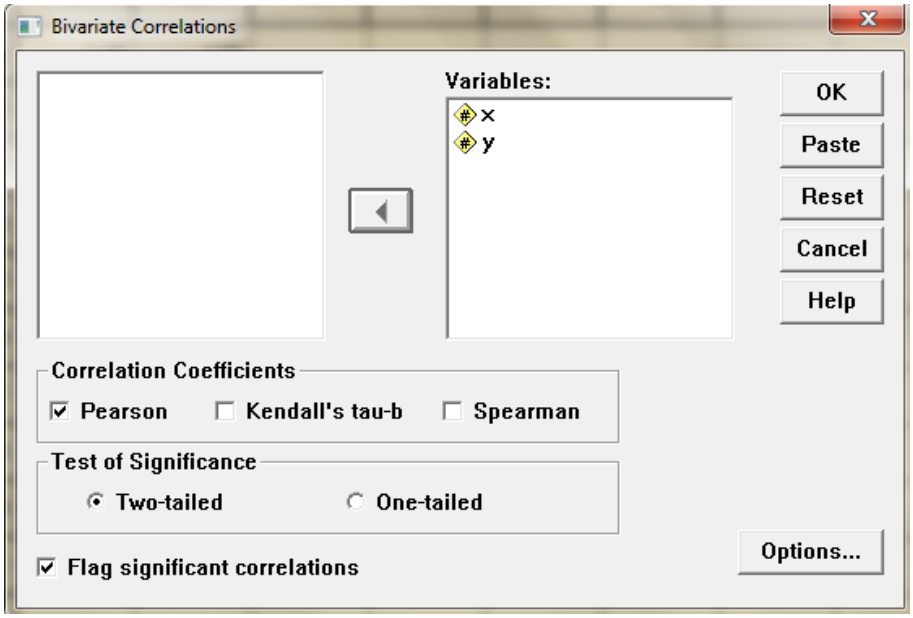
ขั้นที่ 1 เปิดแฟ้มข้อมูล example16.sav เลือกคำสั่ง Analyze / Correlate / Bivariate…



ขั้นที่ 2 เลือกคำสั่ง Bivariate จะได้เมนูย่อยดังนี้



ขั้นที่ 3 เลือกตัวแปร X และ Y ไว้ที่ช่อง Variables: และเลือก Pearson ในส่วน Correlation Coefficient



ขั้นที่ 4 เลือก OK ได้ผลลัพธ์ดังนี้

Correlations

		X	Y
X	Pearson Correlation	1	.991**
	Sig. (2-tailed)	.	.000
	N	9	9
Y	Pearson Correlation	.991**	1
	Sig. (2-tailed)	.000	.
	N	9	9

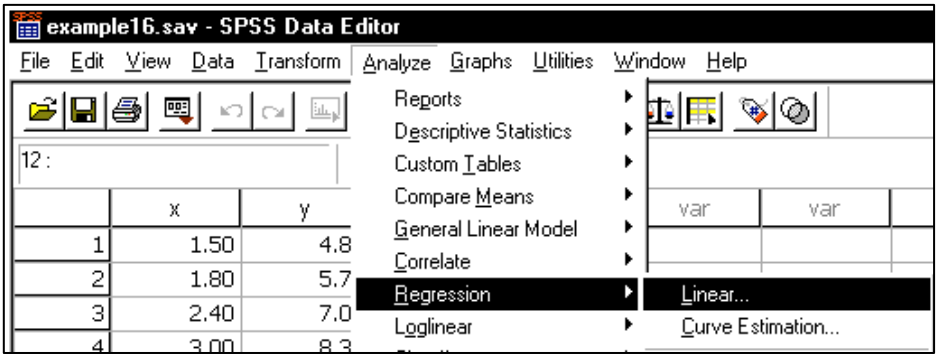
**. Correlation is significant at the 0.01 level

ตัวแปร X และ Y มีความสัมพันธ์กันค่อนข้างมากอย่างมีนัยสำคัญทางสถิติที่ระดับ 0.01 (p_value = 0.000, r = 0.991)

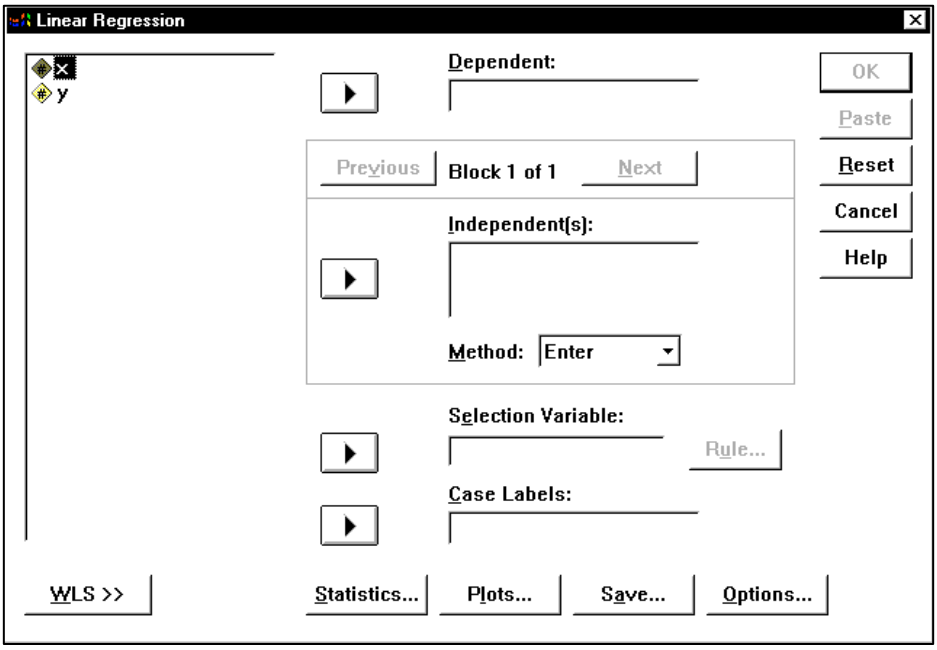
3. การสร้างสมการถดถอย

ตัวอย่าง 8.8 จงสร้างสมการถดถอยเชิงเส้นอย่างง่ายของตัวแปร Y ในแฟ้มข้อมูล example16.sav

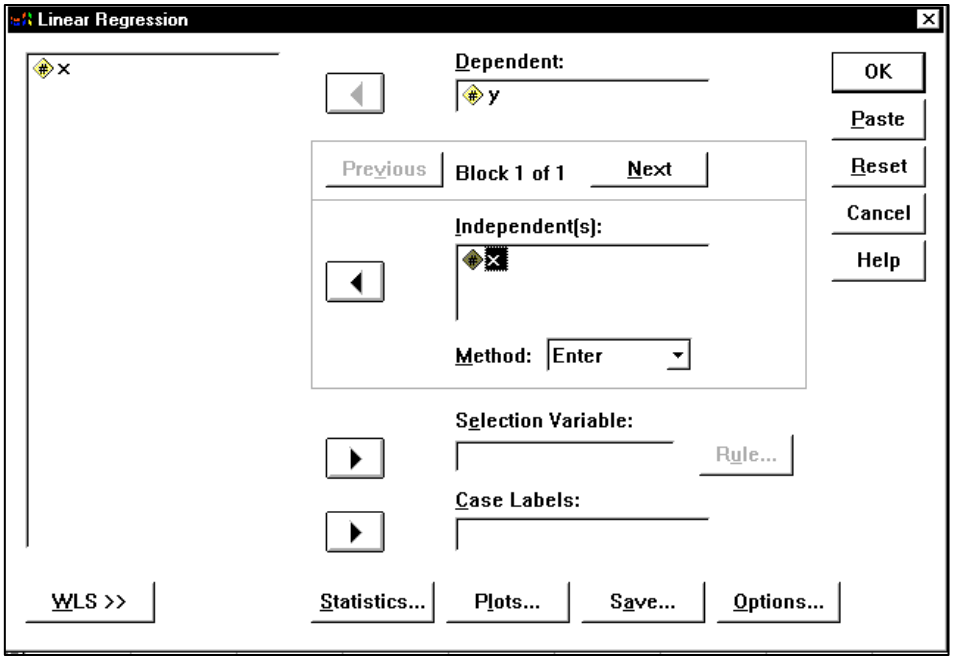
ขั้นที่ 1 เปิดแฟ้มข้อมูล example16.sav เลือกคำสั่ง เลือกคำสั่ง Analyze / Regression / Linear



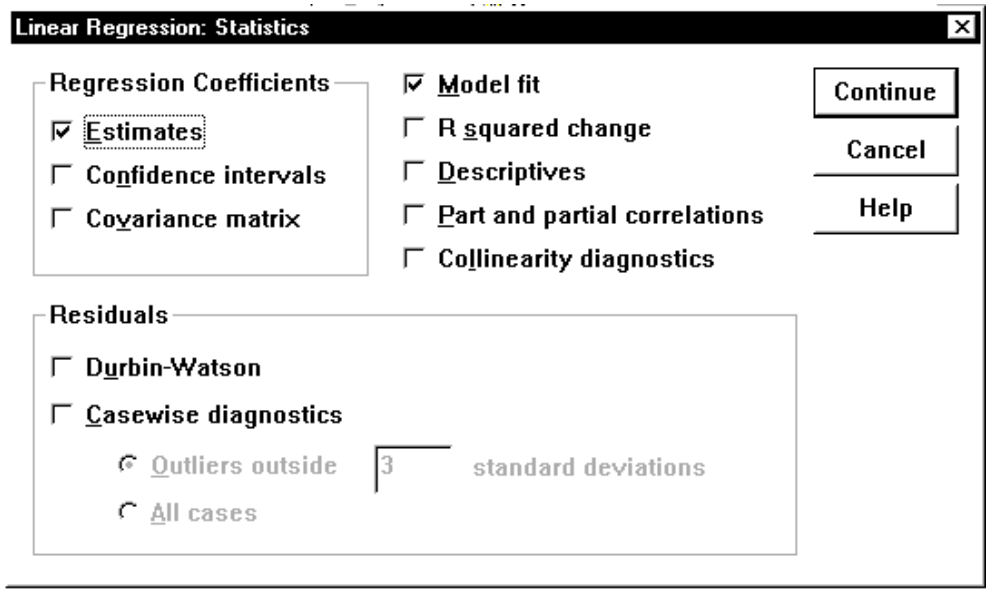
ขั้นที่ 2. คลิกที่ Linear จะได้เมนูของคำสั่งดังนี้



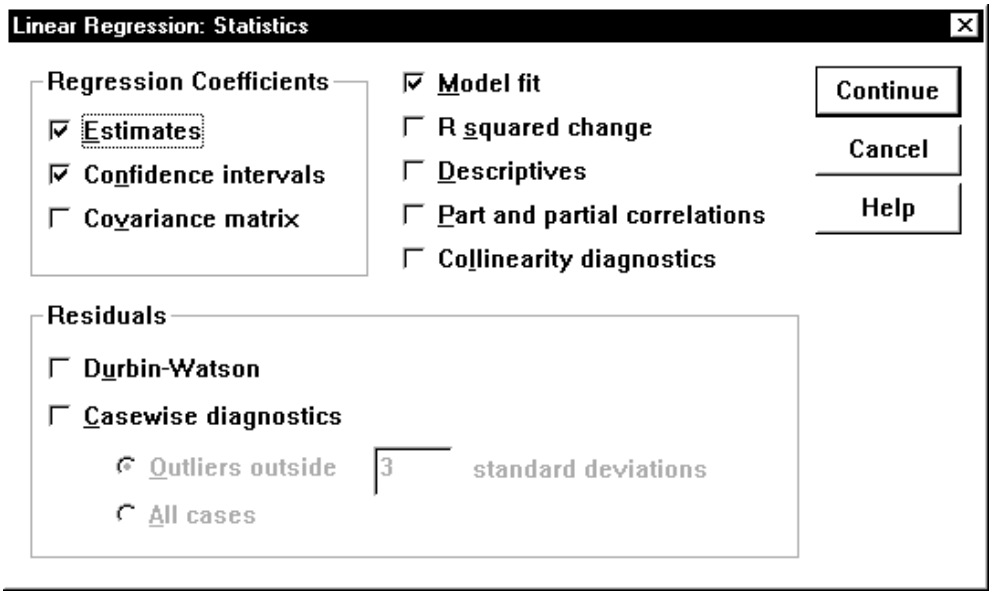
ขั้นที่ 3.เลือกตัวแปร x เป็นตัวแปรอิสระนำไปไว้ที่ช่อง Independent(s) เลือกตัวแปร y เป็นตัวแปรตามนำไปไว้ที่ช่อง Dependent



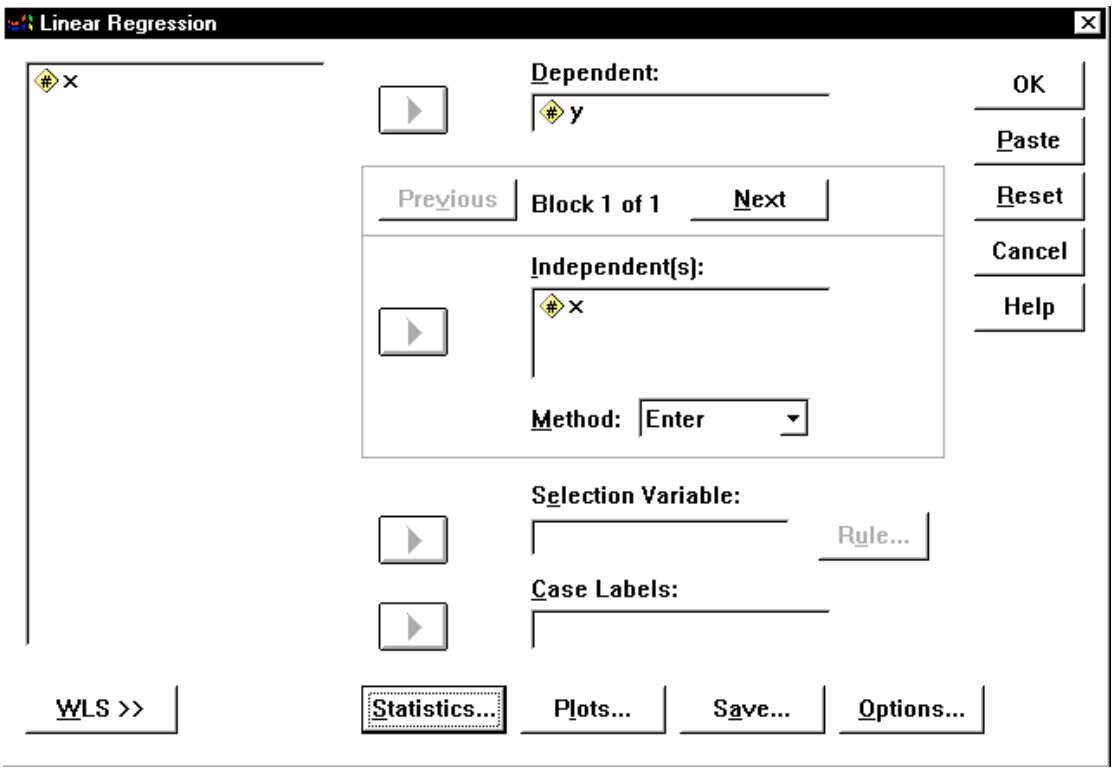
หมายเหตุ หากต้องการเฉพาะค่า b_0, b_1 และ r ให้คลิก OK จะได้ผลการวิเคราะห์ทันที แต่ถ้าต้องการให้มีการเขียนกราฟให้คลิก Plots หรือ ต้องการหาช่วงความเชื่อมั่นของค่าพารามิเตอร์ β_0 และ β_1 ให้คลิกที่ปุ่ม Statistics จะได้เมนูย่อย ดังนี้



ขั้นที่ 4. เลือก Confidence Intervals เพื่อหาช่วงความเชื่อมั่นของค่าพารามิเตอร์ β_0 และ β_1



ขั้นที่ 5. คลิก Continue เพื่อกลับไปที่เมนู Linear Regression



ขั้นที่ 7. คลิก OK จะได้ผลการคำนวณเป็นดังนี้

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	X[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Y

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.991[a]	.982	.980	.53877

a. Predictors: (Constant), X

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	112.484	1	112.484	387.516	.000[a]
	Residual	2.032	7	.290		
	Total	114.516	8			

a. Predictors: (Constant), X

b. Dependent Variable: Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.257	.532		.483	.644	-1.002	1.516
	X	2.930	.149	.991	19.685	.000	2.578	3.282

a. Dependent Variable: Y

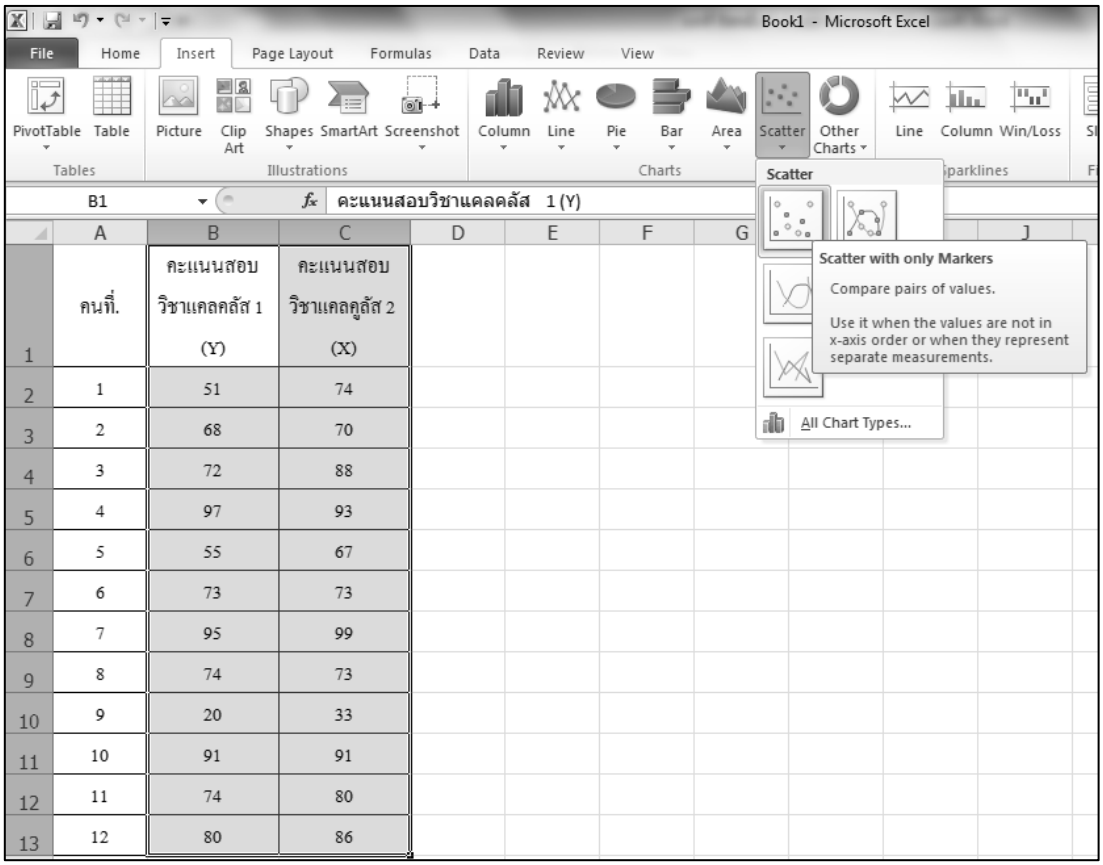
การใช้โปรแกรมสำเร็จรูป MS EXCEL

ใช้ข้อมูลจากตัวอย่างที่ 8.1 ดังนั้นการใส่ข้อมูลใน Microsoft Excel เป็นดังนี้

	A	B	C
1	คนที่.	คะแนนสอบวิชาแคลคลัส 1 (Y)	คะแนนสอบวิชาแคลคูลัส 2 (X)
2	1	51	74
3	2	68	70
4	3	72	88
5	4	97	93
6	5	55	67
7	6	73	73
8	7	95	99
9	8	74	73
10	9	20	33
11	10	91	91
12	11	74	80
13	12	80	86

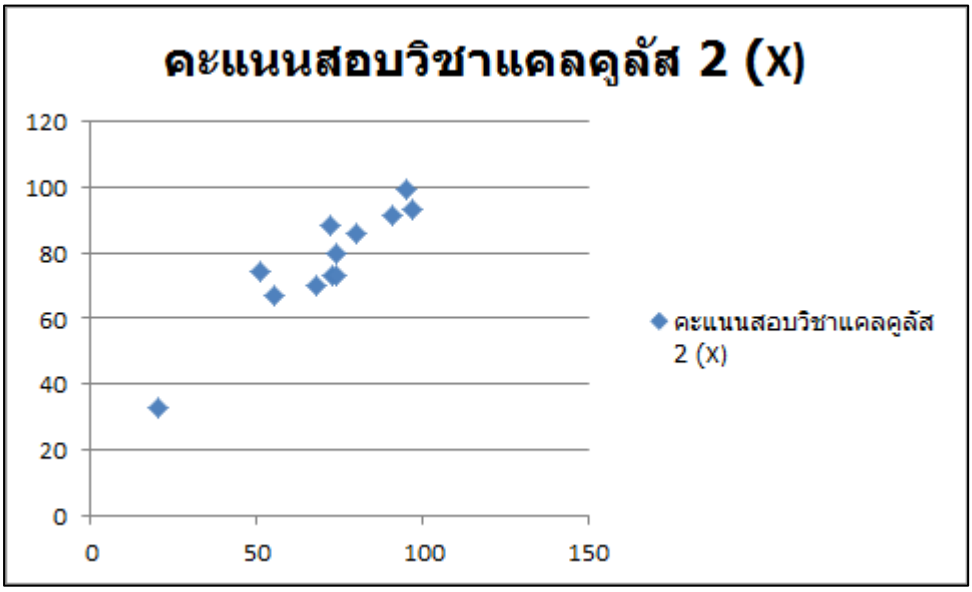
1. การตรวจสอบความสัมพันธ์ด้วยแผนภาพการกระจาย

ขั้นที่ 1 เลือก cell ข้อมูล เลือกเมนู Insert เลือก Chart เลือก Scatter ดังรูป



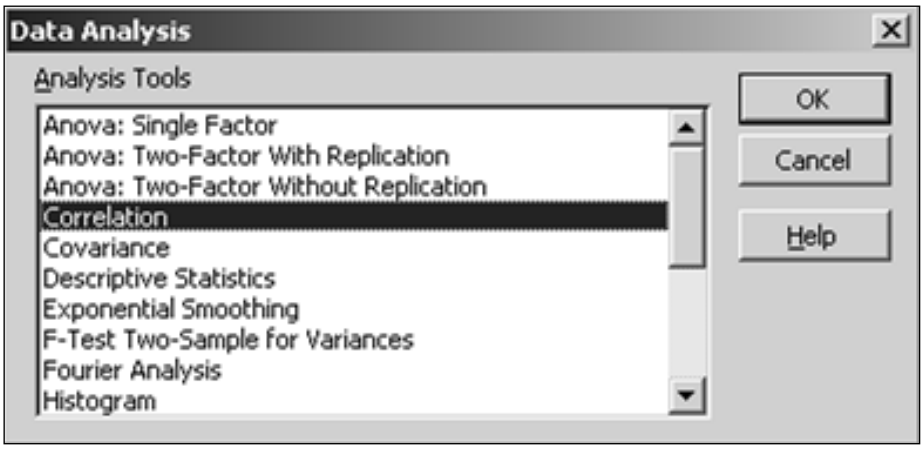
	A	B	C
1	คนที่.	คะแนนสอบวิชาแคลคลัส 1 (Y)	คะแนนสอบวิชาแคลคูลัส 2 (X)
2	1	51	74
3	2	68	70
4	3	72	88
5	4	97	93
6	5	55	67
7	6	73	73
8	7	95	99
9	8	74	73
10	9	20	33
11	10	91	91
12	11	74	80
13	12	80	86

ขั้นที่ 2 จะได้ผลลัพธ์ ดังรูป



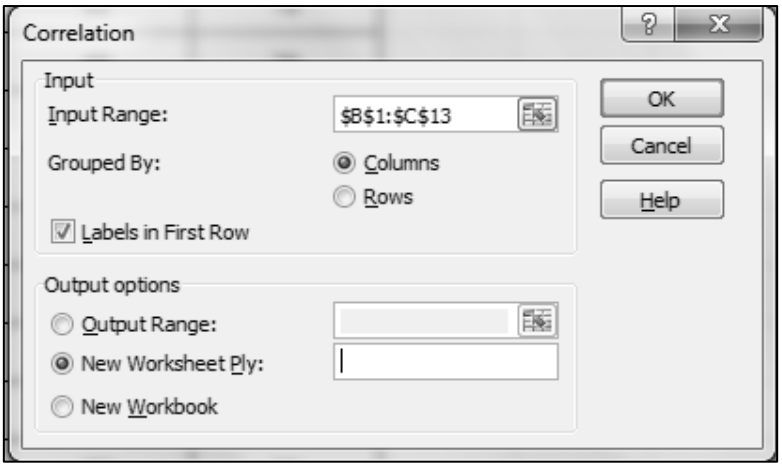
2. การหาค่าสัมประสิทธิ์สหสัมพันธ์

ขั้นที่ 1 เมนู Tools เลือก Data Analysis เลือก Correlation ดังรูป



ขั้นที่ 2 ในส่วน Input Range ให้ใส่ cell ที่เป็นข้อมูลทั้งหมด

ในส่วน Group By เป็นการเลือกการแบ่งกลุ่มของข้อมูล



ขั้นตอนที่ 3 จะได้ผลลัพธ์ ดังรูป

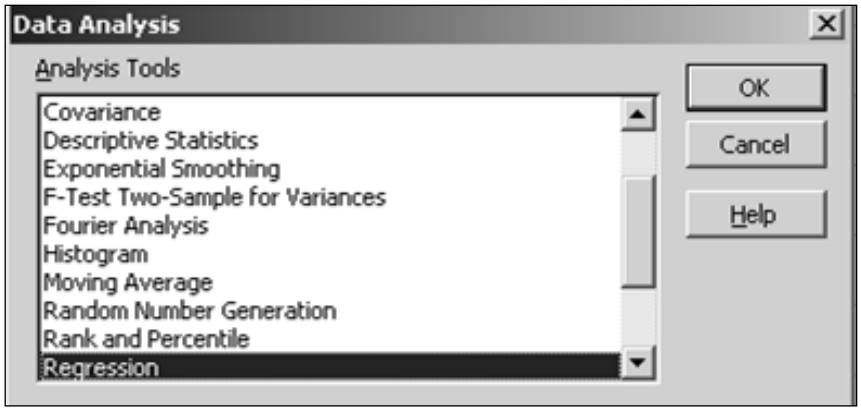
	A	B	C
1		*Economics (Y)*	*Anthropology (X)*
2	Economics (Y)	1	
3	Anthropology (X)	0.935081193	1
4			

$r = 0.935$ หมายความว่าคะแนนสอบวิชาเศรษฐศาสตร์ และวิชามานุษยวิทยา ของนักศึกษา 12 คนมีความสัมพันธ์อย่างมากในเชิงบวก

3. การสร้างสมการถดถอย

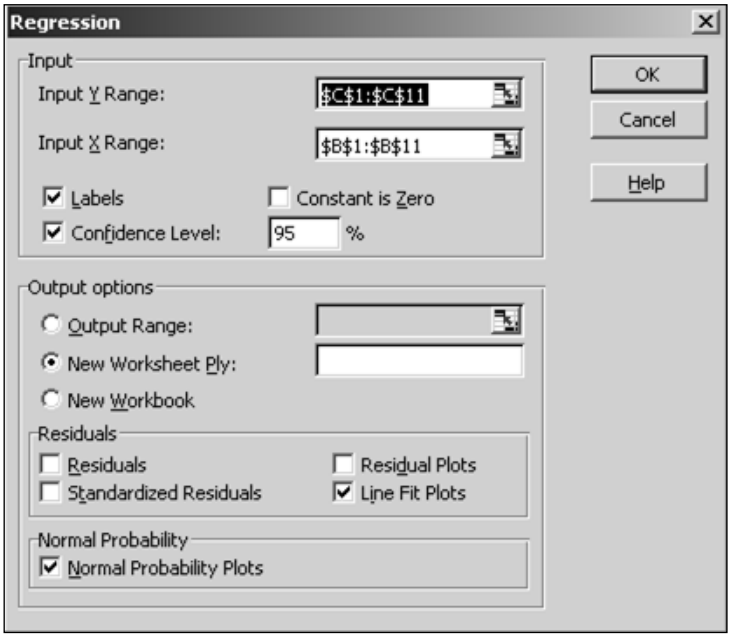
ขั้นตอนที่ 1 ใส่ข้อมูลดิบ เลือกเมนู Tools เลือก Data Analysis เลือก regression ดังรูป

	A	B	C	D
1	ครอบครัวที่	รายได้ X (พันบาท)	รายจ่าย Y (พันบาท)	
2	1	8	7	
3	2	8	8	
4	3	9	8	
5	4	9	10	
6	5	10	8	
7	6	10	9	
8	7	12	9	
9	8	12	11	
10	9	14	12	
11	10	15	14	
12				



ขั้นตอนที่ 2 ในส่วน Input Y Range ให้ใส่ cell ที่เป็นข้อมูลตัวแปรตาม

ในส่วน Input X Range ให้ใส่ cell ที่เป็นข้อมูลตัวแปรอิสระ



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.89363919							
5	R Square	0.79859101							
6	Adjusted R Square	0.77341489							
7	Standard Error	1.03318326							
8	Observations	10							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	33.86025878	33.86026	31.72017	0.00049155			
13	Residual	8	8.53974122	1.067468					
14	Total	9	42.4						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	1.1349353	1.538113276	0.737875	0.481671	-2.41196027	4.68183088	-2.411960267	4.681830877
18	รายได้ X (พันบาท)	0.79112754	0.14046843	5.632067	0.000492	0.467206761	1.11504832	0.467206761	1.115048322

ดังนั้นสมการถดถอยเชิงเส้น คือ $\hat{y} = 1.1363 + 0.791X$ หมายความว่าถ้ารายได้เพิ่มขึ้น 1 พันบาทรายจ่ายจะเพิ่มขึ้น 0.791 พันบาท หรือถ้ารายได้ลดลง 1 พันบาท รายจ่ายจะลดลง 0.791 พันบาท

สรุปท้ายบท

สหสัมพันธ์และการวิเคราะห์การถดถอยเป็นเทคนิคการวิเคราะห์ข้อมูลในทางสถิติ เพื่อหาความสัมพันธ์ของข้อมูล หรือตัวแปรตั้งแต่ 2 ตัว ถ้าข้อมูลทั้ง 2 ชุดเป็นข้อมูลเชิงปริมาณ หรือตัวแปร 2 ตัวแปรเป็นตัวแปรแบบต่อเนื่องการหาความสัมพันธ์จะเรียกว่าสหสัมพันธ์อย่างง่าย และวัดระดับควาสัมพันธ์โดยใช้สัมประสิทธิ์สหสัมพันธ์เพียร์สัน และถ้าข้อมูล หรือตัวแปรตัวหนึ่งเป็นตัวแปรที่มีค่าตามที่กำหนดหรือมีผลต่อตัวแปรอีกตัวหนึ่งจะเรียกตัวแปรนี้ว่าตัวแปรอิสระ ส่วนตัวแปรที่มีค่าเปลี่ยนไปตามค่าของตัวแปรอิสระเรียกว่าตัวแปรตาม เมื่อตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กัน และต้องการพยากรณ์ค่าของตัวแปรตามเมื่อกำหนดค่าตัวแปรอิสระ โดยอาศัยหลักการของการวิเคราะห์การถดถอย ดังนั้นวัตถุประสงค์ของ การวิเคราะห์การถดถอย คือพยากรณ์ค่าตัวแปรตามในอนาคต เมื่อกำหนดค่าตัวแปรอิสระ

แบบฝึกหัดท้ายบท

1. นายแพทย์ผู้หนึ่งต้องการศึกษาเกี่ยวกับความสัมพันธ์ระหว่างน้ำหนัก(Y) และความดันโลหิต(X) ของชายอายุ 25-30 ปี จึงทำการสุ่มตัวอย่างชายอายุ 25-30 ปีจำนวน 26 คน แล้ววัดน้ำหนักกับความดันโลหิต ได้ผลดังนี้

คนที่	น้ำหนัก	ความดันโลหิต
1	165	130
2	167	133
3	180	150
4	155	128
5	212	151
6	175	146
7	190	150
8	210	140
9	200	148
10	149	125
11	158	133
12	169	135
13	170	150
14	172	153
15	159	128

คนที่	น้ำหนัก	ความดันโลหิต
16	168	132
17	174	149
18	183	158
19	215	150
20	195	163
21	180	156
22	143	124
23	240	170
24	256	165
25	192	160
26	187	159

นายแพทย์ผู้นี้ต้องการทราบว่า

ก. น้ำหนักกับความดันโลหิตของชายอายุ 25-30 ปี จะมีความสัมพันธ์กับในระดับใด ที่ช่วงความชื่อมั่น 95%

ข. น้ำหนักกับความดันโลหิตจะมีความสัมพันธ์กันในระดับสูง ด้วยสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.9 หรือไม่ ที่ระดับนัยสำคัญ 0.05

2. ข้อมูลต่อไปนี้ เป็นปริมาณน้ำ (นิ้ว) และผลผลิตข้าว (ต้น/10 ดอร์) ที่สำรวจไว้ในไร่ทดลอง

ปริมาณน้ำ (X)	12	18	24	30	36	42	48
ผลผลิตข้าว (Y)	5.27	5.68	6.25	7.21	8.02	8.71	8.42

จงหา ก. สมการถดถอยโดยวิธีกำลังสองน้อยที่สุด

ข. หาค่า 90% ช่องความเชื่อมั่นและทดสอบสมมติฐานของ b_0 , b_1

ค. ด้วยความเชื่อมั่น 90% จงคาดคะเนค่า y ณ จุดที่ x = 40

ง. หา 90% ช่องความเชื่อมั่นของ E(y|x) ณ จุดที่ x = 40