บทที่ 2

แนวคิด ทฤษฎีและทบทวนวรรณกรรม

การศึกษาค้นคว้าแบบอิสระเรื่องการพัฒนาโปรแกรมแบบจำลองพื้นฐานเพื่อการวิเคราะห์ปัจจัยเชิงกายภาพ ที่เกี่ยวข้องกับการจัดการสวนลำไยในจังหวัดเชียงใหม่ ผู้ศึกษาได้รวบรวมเอาแนวคิด ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องมาทำการศึกษาค้นคว้า และได้ประมวลความรู้ โดยครอบคลุมเรื่องดังต่อไปนี้

2.1 การวิเคราะห์ปัจจัย (Factor Analysis)

2.2 การวิเคราะห์ความแปรปรวน (ANOVA : Analysis of Variance)

2.3 การวิเคราะห์ความถดถอย (Regression Analysis)

2.4 เทคนิคดาต้าไมน์นิ่ง (Data Mining)

2.1 การวิเคราะห์ปัจจัย (Factor Analysis)

รองศาสตราจารย์ ดร. กัลยา วานิชย์บัญชา กล่าวถึงการวิเคราะห์ปัจจัย ไว้ดังนี้

2.1.1 ความหมาย

เทคนิคการวิเคราะห์ปัจจัย เป็นเทคนิคการแบ่งกลุ่มตัวแปรหรือรวมตัวแปรที่มีความสัมพันธ์กันไว้ในกลุ่มเดียวกัน ตัวแปรที่ไม่สัมพันธ์กันจะอยู่ต่างกลุ่มกันโดยที่ 1 กลุ่มจะเรียกว่า 1 ปัจจัย โดยที่ตัวแปรที่อยู่ในปัจจัย หรือกลุ่มเดียวกันจะมีความสัมพันธ์กันในทิศทางบวก หรือลบก็ได้

ดังนั้น เทคนิคการวิเคราะห์ปัจจัยจึงเป็นเทคนิคที่ไม่มีการแบ่งว่าตัวแปรใดจะเป็นตัวแปรตาม หรือตัวแปรใดเป็นตัวแปรอิสระ

2.1.2 วัตถุประสงค์ของการวิเคราะห์ปัจจัย

1. เพื่อลดจำนวนตัวแปร กรณีที่ผู้วิจัยมีตัวแปรจำนวนมาก และตัวแปรเหล่านั้นมีความสัมพันธ์กัน จะจัดกลุ่มตัวแปรที่มีความสัมพันธ์กันไว้ด้วยกัน แล้วเรียกว่าปัจจัย เช่น มีตัวแปร 20 ตัว ($x_1, x_2, \ldots, x_{20}$) ถ้าใช้เทคนิคการวิเคราะห์ปัจจัยแล้วเหลือ 4 ปัจจัย ดังนี้

ปัจจัยที่ 1 (Factor 1) ประกอบด้วยตัวแปร 7 ตัว คือ $x_3, x_7, x_8, x_{10}, x_{15}, x_{19}, x_{20}$

ปัจจัยที่ 2 (Factor 2) ประกอบด้วยตัวแปร 3 ตัว คือ x_1, x_6, x_{18}

ปัจจัยที่ 3 (Factor 3) ประกอบด้วยตัวแปร 8 ตัว คือ $x_2, x_5, x_{11}, x_{12}, x_{13}, x_{14}, x_{16}, x_{17}$

ปัจจัยที่ 4 (Factor 4) ประกอบด้วยตัวแปร 2 ตัว คือ x_4, x_9

ตัวแปรที่อยู่ในปัจจัยเดียวกันมีความสัมพันธ์กันมาก โดยสามารถวัดความสัมพันธ์ด้วยสัมประสิทธิ สหสัมพันธ์ (Correlation) ดังนั้น จากตัวแปร 20 ตัว อาจกลายเป็นตัวแปรใหม่ (ปัจจัย) 4 ตัวแปร ซึ่งมีรายละเอียดของตัวแปรเดิม (X 's) อยู่ในแต่ละปัจจัย ดังนั้นจะต้องมีการตั้งชื่อปัจจัยเพื่อสื่อให้เห็นความหมายของตัวแปร X 's ต่าง ๆ ที่อยู่ในปัจจัยนั้น ๆ

2. เพื่อนำตัวแปรหรือปัจจัยที่สร้างขึ้นมาใหม่สำหรับการวิเคราะห์ทางสถิติต่อไป เช่น - นำปัจจัยที่สร้างใหม่ไปใช้แก้ปัญหาการวิเคราะห์ความถดถอยเชิงพหุ ซึ่งมีตัวแปรอิสระหลาย ๆ ตัวที่คาดว่าส่งผลต่อตัวแปรตาม เช่น คาดว่ามีตัวแปรอิสระ 20 ตัว จะได้สมการความถดถอยดังนี้

$$\hat{Y} = a + b_1 x_1 + b_2 x_2 \ldots + b_{20} x_{20}$$

แต่เมื่อตรวจสอบเงื่อนไขของการวิเคราะห์ความถดถอยเชิงซ้อน(จากตัวแปรอิสระ 20 ตัว) พบว่าตัวแปรอิสระ X 's มีความสัมพันธ์กัน จะเกิดปัญหาที่เรียกว่า Multicollinearity วิธีการหนึ่งที่สามารถแก้ปัญหาดังกล่าวคือ การใช้เทคนิคการวิเคราะห์ปัจจัย โดยรวมกลุ่มตัวแปรอิสระ X 's ที่สัมพันธ์กันให้อยู่ในปัจจัยเดียวกัน เช่น จากตัวอย่างข้างต้นซึ่งมีตัวแปรอิสระ 20 ตัว อาจจะสามารถรวมกลุ่มตัวแปรมาสร้างเป็นตัวแปรใหม่ได้ 4 ตัวแปร หรือ 4 ปัจจัย จะทำให้สมการความถดถอยกลายเป็น

$$\hat{Y} = a + b_1 F_1 + b_2 F_2 + b_3 F_3 + b_4 F_4$$

โดย F_1, F_2, F_3 และ F_4 เป็นปัจจัยที่ 1-4 ตามลำดับ และปัจจัยต่าง ๆ จะไม่มีความสัมพันธ์กันหรือสัมพันธ์กันน้อยมาก

- นำปัจจัยที่ได้เป็นตัวแปรเพื่อนำไปใช้เทคนิค t – test , ANOVA , Z – test หรือเทคนิคอื่น ๆ ต่อไป

3. เพื่อตรวจสอบความถูกต้อง (Confirmnatory) เกี่ยวกับการที่ผู้วิจัยจะต้องกำหนดความสำคัญหรือน้ำหนักของตัวแปร เช่น ต้องการสร้างดัชนีชี้วัดประสิทธิภาพการทำงานของพนักงาน ซึ่งจะพัฒนาจากตัวแปรหลาย ๆ ตัว เช่น ผลงาน (x_1) ระยะเวลาทำงาน (x_2) จำนวนวันลา (x_3) การริเริ่มสร้างสรรค์ (x_4) คะแนนการปรับตัวเข้ากับสิ่งแวดล้อมหรือเพื่อร่วมงาน (x_5)

โดยใช้สมการ $I = w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5$

I = ดัชนีวัดประสิทธิภาพการทำงาน

w_1, w_2, w_3, w_4 และ w_5 เป็นความสัมพันธ์หรือน้ำหนักของตัวแปร x_1, x_2, x_3, x_4 และ x_5 ตามลำดับ

ดังนั้น ถ้าผู้วิจัยกำหนดค่าน้ำหนัก w_1, w_2, w_3, w_4, w_5 เอง อาจจะต้องตรวจสอบโดยใช้เทคนิคการวิเคราะห์ปัจจัยได้

4. เพื่อศึกษาโครงสร้างความสัมพันธ์ของตัวแปรที่อยู่ในกลุ่มเดียวกันหรือปัจจัยเดียวกัน

2.1.3 ชนิดของตัวแปรหรือสเกลของข้อมูลที่ใช้ในการวิเคราะห์ปัจจัย

ตัวแปรที่นำมาใช้ในการวิเคราะห์ปัจจัยจะต้องเป็นข้อมูลเชิงปริมาณ หรือเป็นสเกลแบ่งช่วง และสเกลอัตราส่วน กรณีที่มีตัวแปรบางตัวเป็นตัวแปรเชิงกลุ่ม คือ เป็นสเกลแบ่งกลุ่ม (Nominal) หรือสเกลอันดับ (Ordinal scale) จะต้องเปลี่ยนตัวแปรเชิงกลุ่มให้อยู่ในรูปตัวแปรเทียม (Dummy หรือ Indicator Variable) ก่อน นั่นคือ ตัวแปรเชิงกลุ่มที่จะนำมาใช้ในการวิเคราะห์ปัจจัยจะต้องมีค่าได้เพียง 2 ค่า คือ 0 กับ 1 เท่านั้น

เช่น ตัวแปรการเป็นเจ้าของบ้าน (x_1) จะต้องกำหนด

$$x_1 = \begin{cases} 1 & \text{ถ้าผู้ตอบเป็นเจ้าของบ้าน} \\ 0 & \text{ถ้าผู้ตอบไม่ได้เจ้าของบ้าน} \end{cases}$$

2.1.4 เงื่อนไขของเทคนิคการวิเคราะห์ปัจจัย

1. ความสัมพันธ์ระหว่างปัจจัย (Factor) กับตัวแปร (x's) ต้องอยู่ในรูปเชิงเส้น
2. ปัจจัย (Factor) และค่าคลาดเคลื่อน (e) ของตัวแปรเป็นอิสระกัน
3. จำนวนขอมูลจะต้องมากกว่าจำนวนตัวแปร

2.1.5 ขั้นตอนการวิเคราะห์ปัจจัย

ขั้นที่ 1 การตรวจสอบว่าตัวแปรต่าง ๆ (x_1,..., x_p) ที่จะนำมาแบ่งกลุ่มนั้นมีความสัมพันธ์กันหรือไม่ วิธีการตรวจสอบความสัมพันธ์ทำได้หลายวิธีดังนี้

1. ใช้ค่าสัมประสิทธิสหสัมพันธ์ของตัวแปรแต่ละคู่ ถ้ามีตัวแปรบางตัวที่ไม่มีความสัมพันธ์กันกับตัวแปรอื่น ๆ (สัมประสิทธิสหสัมพันธ์มีค่าใกล้ศูนย์) ให้ตัดตัวแปรนั้นออก
2. ใช้การทดสอบสมมติฐาน โดยตั้งสมมติฐานเพื่อการตรวจสอบดังนี้

H_0 : ตัวแปร $(x_1,\ldots,x_p)$ ไม่มีความสัมพันธ์กัน

H_1 : ตัวแปร $(x_1,\ldots,x_p)$ มีความสัมพันธ์กัน

สถิติทดสอบ

1. KMO (Kaiser – Meger – Olkin)

$$KMO = \frac{\sum r_i^2}{\sum r_i^2 + \sum (\vec{partial\ correlation})^2}$$

โดยที่ r = สัมประสิทธิสหสัมพันธ์

$0 \leq KMO \geq 1$

- ถ้า KMO มีค่ามาก (เข้าสู่ 2) แสดงว่าสามารถใช้เทคนิคการวิเคราะห์ปัจจัยได้ในการแบ่งกลุ่มตัวแปรได้
- ถ้า KMO มีค่าน้อย (เข้าสู่ ศูนย์) แสดงว่าไม่สมควรนำเทคนิคการวิเคราะห์ปัจจัยมาใช้

2. Barlett's Test of sphericity

เป็นสถิติที่มีการแจกแจงโดยประมาณแบบ ไคสแควร์ (chi-square) ถ้าค่าไคสแควร์มีค่ามาก หรือค่า Significance การทดสอบต่ำกว่าระดับนัยสำคัญ (α) ที่กำหนด จะปฏิเสธ H_0 (หรือยอมรับ H_1) นั่นคือตัวแปร $(x_1, x_2, \ldots, x_p)$ มีความสัมพันธ์กันจึงสามารถใช้เทคนิคการวิเคราะห์ปัจจัยได้

สรุปขั้นที่ 1

- ถ้าในขั้นที่ 1 ผลการทดสอบเป็นยอมรับ H_0 แสดงว่าตัวแปร X's ต่างไม่มีความสัมพันธ์กันไม่สมควรใช้เทคนิคการวิเคราะห์ปัจจัย จึงหยุดไม่ต้องทำต่อไปในขั้นที่ 2

- ถ้าในขั้นที่ 1 ผลการทดสอบเป็นปฏิเสธ H_0 หรือยอมรับ H_1 แสดงว่าตัวแปร X's มีความสัมพันธ์กัน จึงทำต่อในขั้นที่ 2 ซึ่งเป็นการนำเทคนิคการวิเคราะห์ปัจจัย มาจัดกลุ่ม ตัวแปรหรือลดจำนวนตัวแปร

ขั้นที่ 2 ทำการวิเคราะห์ปัจจัยโดยการสกัดปัจจัย ซึ่งหมายถึง การดึงข้อมูลจากตัวแปรมาใส่ในปัจจัยซึ่งสามารถเขียนสมการได้ดังนี้

$$F_1 = W_{11}X_1 + W_{12}X_2 + \ldots + W_{1p}X_p + e_1$$
$$F_2 = W_{21}X_1 + W_{22}X_2 + \ldots + W_{2p}X_p + e_2$$
$$\vdots$$
$$F_m = W_{m1}X_1 + W_{m2}X_2 + \ldots + W_{mp}X_p + e_m$$

M = จำนวนปัจจัยโดยที่ $m \leq p$

E = Unique Factor

$F_1, ..., F_m$ = Common Factor

วิธีการสกัดปัจจัยมีหลายวิธีดังนี้

- Princial component Analysis (PCA)
- Unweighted Least Squares
- Generallized Least Squares
- Maximum Likelihood
- ฯลฯ

วิธีที่นิยมใช้มากที่สุดคือ PCA (รายละเอียดของวิธี PCA ศึกษาได้จากหนังสือการวิเคราะห์สถิติขั้นสูงด้วย SPSS for Window บทที่ 2 ของ ดร. กัลยา วานิชย์บัญชา นอกจากนั้นยังสามารถเขียนตัวแปรแต่ละตัวให้เป็น linear combination ของปัจจัย ($F_1, \ldots, F_m$) ดังนี้

$$Z_1 = I_{11} F_1 + I_{12} F_2 + \ldots + I_{1m} F_m + e_1$$

$$Z_2 = I_{21} F_1 + I_{22} F_2 + \ldots + I_{2m} F_m + e_2$$

;

$$Z_p = I_{p1} F_1 + I_{p2} F_2 + \ldots + I_{pm} F_m + e_p$$

โดยที่ Z_i = ตัวแปร X_i ที่ทำการ Standardized แล้ว $i = 1, 2, \ldots, p$

I_{ij} = Factor loading

ขั้นที่ 3 การจัดตัวแปรให้อยู่ในปัจจัยต่าง ๆ

หลังจากสามารถหาค่า factor loading I_{ij} แล้วพิจารณาจากค่า I_{ij} ว่าตัวแปรใดจะอยู่ในปัจจัยใดบ้าง ถ้าค่า factor loading ของตัวแปรใดมีค่ามาก (เข้าสู่ +1 หรือ -1) ควรจัดตัวแปรนั้นอยู่ในปัจจัยดังกล่าว กรณีที่ค่า factor loading มีค่ากลาง ๆ เช่น 4 หรือ 5 ทำให้ไม่สามารถตัดสินใจว่าควรจัดตัวแปรนั้นอยู่ในปัจจัยใด ให้ทำต่อในขั้นที่ 4

ขั้นที่ 4 การหมุนแกนปัจจัย (Factor Rotation)

เมื่อไม่สามารถจัดตัวแปรว่าควรอยู่ในปัจจัยใด จะต้องทำการหมุนแกนเพื่อทำค่า factor loading ของตัวแปรมีค่ามากขึ้นหรือลดลง ซึ่งจะทำให้สามารถจัดตัวแปรว่าควรอยู่ในปัจจัยใดหรือไม่ควรอยู่ในปัจจัยใด

วิธีการหมุนแกนแบ่งเป็น 2 ประเภทใหญ่ ๆ คือ

1. Orthogonal Rotation

เป็นการหมุนแล้วยังคงทำให้ปัจจัยยังคงตั้งฉากกันหรือเป็นอิสระกันโดยมีเทคนิคย่อยหลายวิธี เช่น Varimax , Quartimax และ Equamax ส่วนใหญ่นิยมใช้วิธี Varimax

2. Oblique Rotation

เป็นการหมุนแกนปัจจัยที่เมื่อหมุนแล้ว แกนปัจจัยอาจไม่ตั้งฉากกัน หรือไม่เป็นอิสระกันซึ่งประกอบด้วยเทคนิค Direct Oblique และ Promax ซึ่งเมื่อหมุนแบบไม่ตั้งฉากกันแล้วอาจจะจัดตัวแปรให้แก่ปัจจัยได้ชัดเจนขึ้น

ขั้นที่ 5 การสร้างตัวแปรใหม่หรือปัจจัยใหม่

เมื่อจัดได้แล้วว่าในแต่ละปัจจัยประกอบด้วยตัวแปรใดบ้าง จะต้องสร้างตัวแปรใหม่โดยการ save ตัวแปร หรือ ปัจจัยที่สร้างขึ้น

ขั้นที่ 6 นำปัจจัยหรือตัวแปรใหม่ไปทำการวิเคราะห์ทางสถิติต่อไป เช่น ใช้เทคนิคการวิเคราะห์ความถดถอย การวิเคราะห์ความแปรปรวน t – test , ฯลฯ

2.2 การวิเคราะห์ความแปรปรวน (ANOVA : Analysis of Variance)

2.2.1 การทดสอบสมมติฐานโดยใช้การวิเคราะห์ความแปรปรวน

การทดสอบสมมติฐานเกี่ยวกับค่าเฉลี่ยของประชากรสองกลุ่มสามารถทดสอบสมมติฐานโดยใช้ตัวสถิติ Z, t อย่างไรก็ตาม ในทางปฏิบัติอาจจะมีประชากรมากกว่าสองกลุ่มดังนั้น การวิเคราะห์ความแปรปรวนสามารถถูกนำมาใช้ในการทดสอบความมีนัยสำคัญของความแตกต่างระหว่างค่าเฉลี่ยของประชากรตั้งแต่ 2 กลุ่มขึ้นไปโดยที่สมมติฐานหลักที่ว่า ค่าเฉลี่ยของประชากรทุกค่าเท่ากัน

$H_o : \mu_1 = \mu_2 = \mu_3 \ldots = \mu_b$

$H_a : \mu_1 \neq \mu_2 \neq \mu_3 \ldots \neq \mu_b$ b = จำนวนประชากร

มีลักษณะของข้อมูลดังตาราง 2.1

ตาราง 2.1 ตัวอย่างลักษณะข้อมูล

x_{11}	x_{21}	x_{31}	. . .	x_{b1}
x_{12}	x_{22}	x_{32}		x_{b2}
x_{1n1}	x_{2n2}	x_{3n3}		x_{bnn}

ถ้าข้อมูลตัวอย่างจากประชากร 3 กลุ่ม

	กลุ่มที่ 1	กลุ่มที่ 2	กลุ่มที่ 3
	2	4	6
	2	4	6
	2	4	6
		4	6
			6
ค่าเฉลี่ย	2	4	6
ความแปรปรวน	0	0	0

จากข้อมูลข้างต้นจะเห็นว่าในแต่ละกลุ่มนั้นไม่มีความแตกต่างกันภายในกลุ่ม (Within variance =0) แต่มีความแตกต่างกันระหว่างกลุ่ม (Between variance $\neq$0)

ดังนั้น อาจสามารถสรุปได้ว่าถ้าความแปรปรวนภายในกลุ่มมีค่าน้อย และความแปรปรวนระหว่างกลุ่มมีค่ามาก นั้น ค่าเฉลี่ยระหว่างกลุ่มต่างๆ มีค่าแตกต่างกัน อย่างไรก็ตาม จำเป็นจะต้องใช้วิธีทางสถิติในการทดสอบสมมติฐานดังกล่าวเพื่อความแน่ใจ ด้วยเหตุนี้ การวิเคราะห์ความแปรปรวนจะถูกใช้ในการทดสอบสมมติฐานที่เกี่ยวกับค่าเฉลี่ยประชากรที่มีจำนวนประชากรมากกว่า 2 ประชากร

2.2.2 ขั้นตอนในการวิเคราะห์ความแปรปรวน

1) ตั้งสมมติฐานในการทดสอบ

$H_o : \mu_1 = \mu_2 = \mu_3 \ldots = \mu_b$

$H_a : \mu_1 \neq \mu_2 \neq \mu_3 \ldots \neq \mu_b$

2) คำนวณค่าสถิติในการทดสอบ F โดยใช้ตารางการวิเคราะห์ความแปรปรวน

ตาราง 2.2 แสดงสูตรของAnova

แหล่งความแปรปรวน	df	ผลรวมกำลังสอง (Sum Square : SS)	ค่าเฉลี่ผลรวมกำลังสอง (Mean of Sum Square : MS)	สถิติ F
ระหว่างกลุ่ม	b-1	SSb	MSb = SSb/ (b-1)	F=MSb/MSe
ภายในกลุ่ม	(n-1)-(b-1)	SSe = SST-SSb	MSe = Sse/ (n-b)	
ผลรวม	n-1	SST		

เมื่อ n = จำนวนข้อมูลทั้งหมด

b = จำนวนกลุ่มข้อมูล

SSb = ผลรวมกำลังสองของความแตกต่างข้อมูลระหว่างกลุ่ม

$$SSb = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + \cdots + n_b(\bar{x}_b - \bar{\bar{x}})^2$$

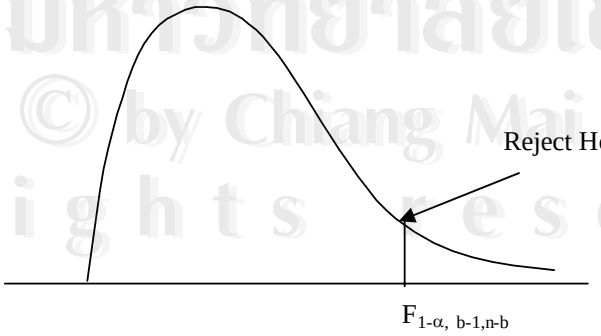
$$SST = \sum_{i=1}^{b}\sum_{j=1}^{n_i}(x_{ij} - \bar{\bar{x}})^2$$

SST = ผลรวมกำลังของความแตกต่างข้อมูลทั้งหมด

SSe = SST-SSb หรือ

$$SSe = \sum_{i=1}^{b}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$$

3) คำนวณค่าวิกฤต (Critical Value) $F_{1-\alpha,\ b-1,n-b}$



รูป 2.1แสดงกราฟคำนวนค่าวิกฤต

4) เปรียบเทียบค่าสถิติ F และค่าวิกฤต F

5) สรุปผลการทดสอบสมมติฐาน

2.3 การวิเคราะห์ความถดถอย (Regresssion Analysis)

เป็นการศึกษาถึงความสัมพันธ์ของตัวแปรตั้งแต่สองตัวขึ้นไป โดยมีวัตถุประสงค์ที่จะประมาณการหรือพยากรณ์ค่าของตัวแปรตัวหนึ่งจากตัวแปรตัวอื่นๆ ที่มีความสัมพันธ์กับตัวแปรที่ต้องการพยากรณ์ โดยจะต้องมีการกำหนดหรือทราบค่าตัวแปรอื่นๆล่วงหน้า เช่น ถ้าทราบความสัมพันธ์ระหว่างยอดขายกับค่าโฆษณาแล้ว จะทำให้สามารถประมาณ / พยากรณ์ยอดขายเมื่อกำหนดหรือทราบงบประมาณในการโฆษณา และจะศึกษาถึงการเปลี่ยนแปลงของยอดขายเมื่องบประมาณมรการโฆษณาเปลี่ยนแปลงไป โดยอาศัยหลักการของการวิเคราะห์ความถดถอย การวิเคราะห์ความถดถอยแบ่งเป็น 2 ประเภทคือ

1. การวิเคราะห์ความถดถอยอย่างง่าย
2. การวิเคราะห์ความถดถอยพหุคูณหรือความถดถอยเชิงซ้อน

2.3.1 การวิเคราะห์ความถดถอยอย่างง่าย (Simple Regression)

เป็นการศึกษาถึงความสัมพัธ์ระหว่างตัวแปร2 ตัว หรือลักษณะที่สนใจศึกษา 2 ลักษณะ โดยที่ต้องทราบค่าของตัวแปรตัวหนึ่งหรือต้องกำหนดค่าของตัวแปรตัวหนึ่งไว้ล่วงหน้า เช่น ถ้าศึกษาถึงความสัมพันธ์ระหว่างรายจ่ายกับรายได้ ยอดขายกับโฆษณา ฯลฯ ซึ่งจะต้องทราบหรือกำหนดรายได้ และค่าโฆษณาไว้ล่วงหน้า เช่น ทราบว่าเงินเดือนพนักงานทำความสะอาดของบริษัทแห่งหนึ่ง เป็น 2,000 , 2,500 , 3,000 และ 4,000 บาท ผู้วิเคราะห์จะต้องสอบถามพนักงานทำความสะอาดที่มีเงินเดือนดังกล่าวถึงรายจ่ายต่อเดือน จึงจะสามารถหาความสัมพันธ์ระหว่างรายได้กับรายจ่ายได้ หรือในการหาความสัมพันธ์ระหว่างยอดขาย กับค่าโฆษณาตะต้องทราบถึงงบประมาณในการโฆษณาที่บริษัทกำหนดไว้หรือใช้ไปจริง แล้วจึงจะทราบยอดขาย โดยจะเรียกรายได้ และ ค่าโฆษณา ซึ่งเป็นตัวแปรที่ต้องกำหนดค่าไว้ล่วงหน้าว่า ตัวแปรอิสระ (Independent Variable) และมักจะใช้สัญลักษณ์ X ส่วนยอดขายกับรายจ่ายจะเรียกว่า ตัวแปรตาม (Dependent Variable) และใช้สัญลักษณ์ Y ซึ่งหมายถึงยอดขายเป็นตัวแปรที่ขึ้นอยู่กับค่าโฆษณา และรายจ่ายเป็นตัวแปรที่ขึ้นอยู่กับรายได้

2.3.1.1 วัตถุประสงค์ของการวิเคราะห์ความถดถอยและสหสัมพันธ์

1. เพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรว่ามีความสัมพันธ์กันมากน้อยเพียงใด ถ้า X และ Y มีความสัมพันธ์กันมาก แสดงว่า ถ้า X มีการเปลี่ยนแปลงไปจะมีผลกระทบต่อค่าของ Y เป็นอย่างมาก

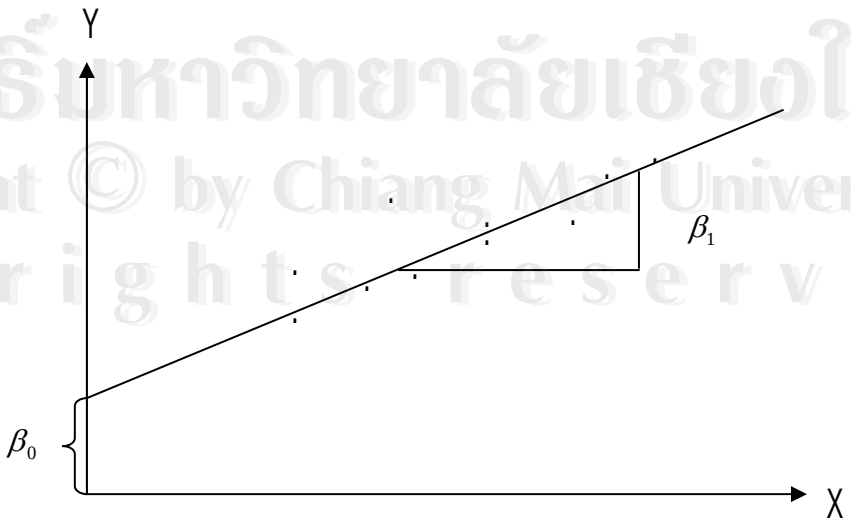
2. ใช้ความสัมพันธ์ที่วิเคราะห์ได้มาประมาณค่าหรือพยากรณ์ค่า Y ในอนาคต เมื่อกำหนดค่า X

สำหรับการหารูปแบบความสัมพันธ์ระหว่างตัวแปร X และ Y นั้น ในขั้นแรกจะนำเอาข้อมูลของตัวแปรทั้งสองมาเขียนกราฟแสดงความสัมพันธ์ ซึ่งจะเรียกกราฟนี้ว่า แผนภาพการกระจาย (Scatter Diagram) ผู้วิเคราะห์จะต้องพิจารณาจากแผนภาพการกระจายว่าความสัมพันธ์ของตัวแปรทั้งสองจะอยู่ในรูปแบบใด เช่น เส้นตรง พาราโพลา เส้นโค้ง หรือ อื่นๆ ฯลฯ โดยที่จะต้องสามารถเขียนความสัมพันธ์ให้อยู่ในรูปแบบทางคณิตศาสตร์ได้ ในที่นี้จะศึกษาเฉพาะความสัมพันธ์ของตัวแปร X และ Y ในรูปเชิงเส้นหรือเส้นตรงเท่านั้น จึงเรียกการวิเคราะห์ความถดถอยอย่างง่ายที่ความสัมพันธ์ของตัวแปรอยู่ในรูปเชิงเส้นว่า การวิเคราะห์ความถดถอยเชิงเส้นอย่างง่าย (Simple linear Regression Analysis)

2.3.1.2 การวิเคราะห์ความถดถอยเชิงเส้นอย่างง่าย (Simple linear Regression Analysis)

เป็นการศึกษาถึงความสัมพันธ์ระหว่างตัวแปร 2 ตัว ที่ความสัมพันธ์อยู่ในรูปเชิงเส้น ซึ่งสามารถแสดงความสัมพันธ์ในรูปสมการเชิงเส้นดังนี้

$$Y_i = \beta_0 + \beta_1 X_i + e_i \quad ; i = 1, 2, \ldots, N \quad \ldots\ldots\ldots\ldots\ldots\ldots (2.1)$$



รูป 2.2 แสดงสมการเชิงเส้น

โดยที่ Y คือ ตัวแปรตาม (Dependent Variable) เนื่องจากค่าของ Y ขึ้นอยู่กับค่าของ X

X คือ ตัวแปรอิสระ (Independent Variable)

β_0 = ส่วนตัดแกน Y หรือ คือค่าของ Y เมื่อ X มีค่าเป็นศูนย์

e = ความคลาดเคลื่ออย่างสุ่ม (random error)

β_1 = ความชัน (slope) ของเส้นตรง ซึ่งเป็นค่าที่แสดงถึงอัตราการเปลี่ยนแปลงของ Y เมื่อ X เปลี่ยนไป 1 หน่วย และจะเรียก β_1 ว่าสัมประสิทธ์ความถดถอย (Regression Coefficient)

ค่าของ β_1 อาจจะเป็น

- $\beta_1 > 0$ แสดงว่า X และ Y มีความสัมพันธ์ในทางเดียวกัน คือ ถ้า X เพิ่ม Y จะเพิ่มด้วย แต่ถ้า X ลดลง Y ก็จะลดลงด้วย
- $\beta_1 < 0$ แสดงว่าค่า X และ Y มีความสัมพันธ์ในทางตรงกันข้าม คือถ้า X เพิ่ม Y จะลดลง แต่ถ้า X ลดลง Y ก็จะกลับเพิ่มขึ้น
- β_1 มีค่าเข้าใกล้ศูนย์ แสดงว่าค่า X และ Y มีความสัมพันธ์กันน้อย
- $\beta_1 = 0$ แสดงว่า X และ Y ไม่มีความสัมพันธ์กันเลย

2.3.1.3 สมมุติฐานของการวิเคราะห์ความถดถอย

1. ค่า X จะต้องเป็นค่าที่กำหนดไว้ล่วงหน้าหรือทราบค่า

2. ความคลาดเคลื่อน e_i เป็นตัวแปรที่มี ค่าเฉลี่ย = 0 หรือ $E(e_i) = 0$

ค่าแปรปรวนของ e_i มีค่าเท่ากันทุกค่าของ i และมีค่าเท่ากับค่าแปรปรวนของ Y

$V(e_i) = V(Y) = \sigma^2_{y.x} = \sigma^2$

3. e_i และ e_j เป็นอิสระกัน นั้นคือ $Cov(e_i, e_j) = E(e_i, e_j) = 0 ; i \neq j$

4. e_i มีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเป็นศูนย์และค่าแปรปรวน = σ^2 นั้นคือ

$e_i \sim normal(0, \sigma^2)$

จากข้อสมมุติข้างต้น จะได้ว่า

$Y_i \sim normal(E(Y_i), \sigma^2)$

โดยที่ $E(Y_i) = E(\beta_0 + \beta_1 X_i + e_i)$

$= \beta_0 + \beta_1 X_i + E(e_i)$

$= \beta_0 + \beta_1 X_i$

2.3.1.4 การประมาณค่าพารามิเตอร์ของสมการถดถอย

เมื่อพิจารณาจากแผนภาพการกระจาย ซึ่งแสดงความสัมพันธ์ระหว่าง X และ Y แล้วพบว่า X และ Y สัมพันธ์กันในรูปเส้นตรง จะต้องคำนวนหาค่า β_0 และ β_1 ซึ่งจะทำให้ทราบถึงความสัมพันธ์ระหว่าง X และ Y ว่ามีความสัมพันธ์ตามกันหรือตรงข้ามกันและความสัมพันธ์นั้นมากหรือน้อยเพียงใด ถ้า β_1 มีค่ามากแสดงว่า Y มีความสัมพันธ์กับ X มากด้วย

การที่จะหาค่า β_0 และ β_1 ได้จำเป็นต้องทราบค่า X และ Y ทุกค่าที่ได้เกิดขึ้นแล้วในอดีต เช่นถ้า X = รายได้ของคนกรุงเทพฯ Y = รายจ่ายของคนกรุงเทพฯ การหาค่าของ β_0 และ β_1 จะต้องทราบถึงรายได้ และรายจ่ายของคนกรุงเทพฯทุกคน ซึ่งเป็นไปได้ยากในทางปฏิบัติเราจึงใช้ข้อมูลตัวอย่างขนาด n ในการประมาณค่า β_0 และ β_1 ดังนั้นค่าประมาณของ Y คือ

$$Y_i = \beta_0 + \beta_1 X_i$$

หรือ $\hat{Y} = a + bX_i$ .....(2.2)

โดยที่ $\beta_0 = a$, $\beta_1 = b$

2.3.1.5 การประมาณค่า β_0 และ β_1 โดยวิธีกำลังสองน้อยที่สุด

การประมาณค่า β_0 และ β_1 ด้วย a และ b ตามลำดับนั้น มีเป้าหมายเพื่อให้ความคลาดเคลื่อนในการประมาณค่า Y_i ด้วย Y_i ค่าต่ำสุด โดยที่ใช้วิธีที่เรียกว่า วิธีกำลังสองน้อยที่สุด (Least Square Method) ซึ่งเป็นวิธีที่ต้องการหาค่า a และ b ที่ทำให้ผลบวกของค่าคลาดเคลื่อนยกกำลังสองมีค่าน้อยที่สุด

เนื่องจาก $Y_i = \beta_0 + \beta_1 X_1 + e_i$

และ $\hat{Y} = a + bX_i$

$\therefore Y_i - \hat{Y}_i = e_i$

ผลบวกของค่าคลาดเคลื่อนยกกำลังสอง = $\sum_{i=1}^{n} e_i^2 = \sum (Y_i - \hat{Y}_i)^2$

ดังนั้นวิธีกำลังสองน้อยที่สุดคือการหาค่า a และ b ที่ทำให้ $\sum_{i=1}^{n} e_i^2$ มีค่าต่ำที่สุด

การที่ต้องใช้ผลบวกของค่าคลาดเคลื่อนยกกำลังสองเนื่องจากค่า e_i อาจจะเป็นค่าบวกเมื่อ Y_i มากกว่า $\hat{Y}_i$ และจะมีค่าลบ ถ้า Y_i น้อยกว่า $\hat{Y}_i$ ซึ่งอาจมีผลทำให้ $\sum e_i$ เป็นศูนย์หรือมค่าน้อยกว่าที่เป็นจริง

การหาค่า a และ b ที่ทำให้ $\sum e_i^2$ มีค่าต่ำสุดจะทำได้โดยการใช้อนุพันธ์เชิงส่วน (partial derivative) เทียบกับ a และ b แล้วให้เท่ากับศูนย์

$$\frac{\partial}{\partial a}\left[\sum_{i}^{n}\left(Y_i-\hat{Y}_i\right)^2\right] = \frac{\partial}{\partial a}\left[\sum_{i=1}^{n}\left(Y_i-a-bX_i\right)^2\right] = 0$$

หรือ $-2\sum(Y_i-a-bX_i) = 0$

$-2\sum Y_i+2na+2b\sum X_i = 0$

$$an+b\sum_{1}^{n}X_i = \sum_{1}^{n}Y_i \qquad \text{.....(2.3)}$$

และ $\frac{\partial}{\partial a}\left[\sum_{i}^{n}\left(Y_i-\hat{Y}_i\right)^2\right] = -2\sum(Y_i-a-bX_i)(X_i) = 0$

หรือ $$a\sum_{1}^{n}X_i + b\sum_{1}^{n}X_i^2 = \sum_{1}^{n}X_i\,Y_i \qquad \text{.....(2.4)}$$

และเรียกสมการที่ (2.3) และ (2.4) ว่าสมการปกติ (normal equation) แก้สมการที่ (2.3) และ (2.4) เพื่อหาค่า a และ b ดังนี้

นำ $(\sum X_i)$ คูณสมการที่ (2.3) ได้สมการที่ (2.5)

$$an(\sum X_i)+b(\sum X_i)^2 = (\sum X_i)(\sum Y_i) \qquad \text{.....(2.5)}$$

นำ n คูณสมการที่ (2.4) ได้สมการที่ (2.6)

$$an(\sum X_i)+b(\sum X_i)^2 = n(\sum X_iY_i) \qquad \text{.....(2.6)}$$

(2.6) – (2.5) ได้

$$bn(\sum X_i^2) - b(\sum X_i)^2 = n(\sum X_iY_i) - (\sum X_i)(\sum Y_i)$$

$$b[n(\sum X_i^2) - (\sum X_i)^2] = n(\sum X_iY_i) - (\sum X_i)(\sum Y_i)$$

$$b = \frac{n(\sum X_iY_i)-(\sum X_i)(\sum Y_i)}{n(\sum X_i^2)-(\sum X_i)^2}$$

$$= \frac{\sum X_iY_i - \left(\frac{(\sum X_i)(\sum Y_i)}{n}\right)}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}$$

หรือ b = $\frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X}^2)}$

จากสมการที่ (2.3) จะได้ว่า

an + b ($\sum X_i$) = $\sum Y_i$

an = $\sum Y_i$ - b ($\sum X_i$)

หรือ a = $\frac{\sum Y_i}{n} - b\frac{\sum X_i}{n}$

a = $\overline{y} - b\overline{X}$

หรือเขียนได้ว่า

$$\hat{\beta}_1 = b = \frac{\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}$$

$$\hat{\beta}_0 = a = \overline{y} - b\overline{x} \qquad \text{.....(2.7)}$$

หรือ

b = $\frac{xy}{xx}$

a = $\overline{y} - b\overline{x}$ .....(2.8)

โดยที่ $SS_{xx} = \sum(X_i - \bar{x})^2 = \sum_1^n X_i^2 - \frac{(\sum_1^n X_i)^2}{n}$

$$SS_{xy} = \sum(X_i - \bar{x})\ (Y_i - \bar{y}) = \sum_1^n X_i Y_i - \frac{(\sum_1^n X_i)(\sum_1^n X_i)}{n}$$

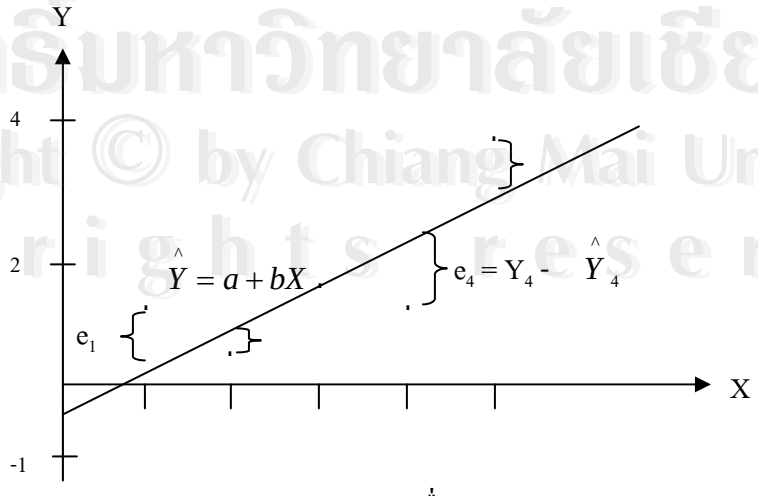
$$SS_{yy} = \sum(Y_i - \bar{y})^2 = \sum_1^n Y_i^2 - \frac{(\sum_1^n Y_i)^2}{n}$$

การที่ประมาณค่า β_0 และ β_1 ด้วยค่า a และ b โดยใช้วิธีกำลังสองน้อยที่สุดจะทำให้

1. ผลรวมของค่าคลาดเคลื่อนในการประมาณค่า Y_i ด้วย $\hat{Y}$ เป็นศูนย์ คือ
   $(\sum Y_i - \hat{Y}_i) = \sum e_i = 0$
2. จุด $(\bar{x}, \bar{y})$ เป็นจุดที่อยู่บนเส้นความถดถอย
3. $\sum(Y_i - \bar{Y})^2$ มีค่าต่ำสุด

2.3.1.6 การประมาณค่าแปรปรวนของความคลาดเคลื่อน

ในการประมาณค่า $Y = \beta_0 + \beta_0$ และ $\beta_1 X + e$ ด้วย $\hat{Y} = a + bX$ ซึ่งจะเกิดค่าคลาดเคลื่อนในการประมาณ Y_i ด้วย $\hat{Y}_i$ เป็น $e = Y_i - \hat{Y}_i$ ดังแสดงในรูปที่ 2.3



รูป 2.3 แสดงค่าคลาดเคลื่อน e_i

จากหัวข้อที่ 2.3.1.3 ได้ว่า $e_i \sim \text{normal}(0, \sigma^2)$

และ $Y_i \sim \text{normal}(E(Y_i), \sigma^2)$

$$\sigma^2 = V(Y_i) = V(e_i)^2$$
$$= E(e_i - E(e_i))^2$$
$$= E(e_i - 0)^2$$
$$= E(e_i)^2$$
$$= E(Y_i - \hat{Y}_i)^2$$
$$= \sigma^2_{Y.X}$$

การที่ใช้สัญลักษณ์ $\sigma^2_{Y.X}$ หมายถึงค่าแปรปรวนของ Y ที่เกิดขึ้นเนื่องจากอิทธิพลของ X โดยทั่วไปจะใช้สัญลักษณ์ σ^2 แทน $\sigma^2_{Y.X}$

ค่าประมาณของ $\sigma^2_{Y.X}$ คือ $S^2_{Y.X}$ หรือ S^2 โดยที่

$$S^2 = \frac{[\sum (Y_i - \hat{Y}_i)^2]}{n-2} \text{ หรือ } S^2 = \frac{SSE}{n-2}$$

โดยที่ $SSE = \sum (Y_i - \hat{Y})^2$

$$= \sum [Y_i - (a + bX_i)]^2$$

$$= \sum [Y_i - (\bar{y} - b\bar{x}) - bX_i]^2$$

$$= \sum [(Y_i - \bar{y}) - b(X_i - \bar{x})]^2$$

$$= \sum [(Y_i - \bar{y})^2 - 2b(X_i - \bar{x})(Y_i - \bar{y}) + b^2 (X_i - \bar{x})^2]$$

หรือ $SSE = SS_{YY} - 2b\, SS_{XY} + b^2 SS_{XX}$ ……(2.9)

แต่เนื่องจาก $b = \frac{SS_{XY}}{SS_{XX}}$

หรือ $SS_{XX} = \frac{SS_{XY}}{b}$

นำ $SS_{XX} = \frac{SS_{XY}}{b}$ แทนในสมการที่(2.9)ได้ SSE $= SS_{YY} - 2bSS_{XY} + bSS_{XY}$

$= SS_{YY} - bSS_{XY}$

หรือ $SSE = SS_{YY} - \frac{(SS_{XY})^2}{SS_{XX}}$

ดังนั้น $S^2 = \frac{SSE}{\quad}$ ……(2.10)

2.3.1.7 การประมาณค่าแบบช่วงและการทดสอบสมมุติฐานเกี่ยวกับสัมประสิทธิ์ความถดถอย($\beta 1$)และ(β_0)

2.3.1.7.1 การประมาณค่า β_1 แบบช่วง

การประมาณค่าสัมประสิทธ์ความถดถอย(β_1) แบบช่วงทำได้ดังนี้

เราทราบว่า b เป็นค่าประมาณแบบจุดของ β_1

ค่าประมาณของค่าแปรปรวนของ b = $\hat{V}(b) = s_b^2$

$$\hat{V}(b) = s_b^2 = \hat{V}(SS_{XY} / SS_{XX})$$

$$= \hat{V}\left[\frac{\sum(X-\bar{x})Y}{\sum(X-\bar{x})^2}\right]$$

$$= \frac{\sum(X-\bar{x})^2 \hat{V}(Y)}{\left[\sum(X-\bar{x})^2\right]^2}$$

$$= \frac{\hat{V}(Y)}{\sum(X-\bar{x})^2}$$

$$= \frac{S^2}{SS_{XX}}$$

$$\therefore \quad S_b^2 = \frac{S^2}{SS_{XX}} \qquad \text{หรือ} \qquad S_b = \frac{S}{\sqrt{SS_{XX}}}$$

ดังนั้นช่วงความเชื่อมั่น $(1-\alpha)100\%$ ของ β_1 คือ

$$b - t_{1-\alpha/2}\frac{S}{\sqrt{SS_{XX}}} < \beta_1 < b + t_{1-\alpha/2} - S_b$$

หรือ $b - t1\text{-}\alpha/2\frac{S}{\sqrt{SS_{XX}}} < \beta_1 < b + t_{1-\alpha/2}\frac{S}{\sqrt{SS_{XX}}}$ โดยที่ t มีองศาอิสระ n-2 .....(2.11)

หรือค่าประมาณแบบช่วงของ β_1 ที่ระดับความเชื่อมั่น $(1-\alpha)100\%$ คือ

$$b \pm t_{1-\alpha/2}\frac{S}{\sqrt{SS_{XX}}}$$

กรณีที่ขนาดตัวอย่างใหญ่ (n>30) จะใช้สถิติ z แทน t ซึ่งทำให้ค่าประมาณช่วงของ β_1 กลายเป็น

$$b - Z_{1-\alpha/2}\frac{S}{\sqrt{SS_{XX}}} < \beta_1 < b + Z_{1-\alpha/2}\frac{S}{\sqrt{SS_{XX}}} \qquad \text{.....(2.12)}$$

2.3.1.7.2 การทดสอบสมมุติฐานเกี่ยวกับ β_1 แบบสองข้าง

การทดสอบสมมุติฐายเกี่ยวกับค่า β_1 เป็นการทดสอบว่าตัวแปร X และ Y มีความสัมพันธ์ในลักษณะเชิงเส้นหรือไม่ โดยเป็นการทดสอบสมมุติฐานแบบ 2 ข้าง

จากสมการความถดถอย $Y_i = \beta_0 + \beta_1 X_i + e_i$

ถ้า $\beta_1 = 0$ แสดงว่า X และ Y ไม่มีความสัมพันธ์กันในลักษณะเชิงเส้น ดังนั้นจึงทดสอบ

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

หรือ H_0 : Y และ X ไม่มีความสัมพันธ์ในรูปเชิงเส้น

H_1 : Y และ X มีความสัมพันธ์ในรูปเชิงเส้น

สถิติทดสอบ :

$$t = \frac{b-0}{S_b} = \frac{b}{S\big/\sqrt{SS_{XX}}}$$

เขตปฏิเสธ : จะปฏิเสธสมมุติฐาน H_0 ถ้า $t < -t_{1-\alpha/2;n-2}$ หรือ $t < t_{1-\alpha/2;n-2}$

หรือจะปฏิเสธ H_0 ถ้า $|t| > t_{1-\alpha/2;n-2}$

ถ้าปฏิเสธ H_0 แสดงว่า X และ Y สัมพันธ์กันในลักษณะเชิงเส้นหรือการเปลี่ยนแปลงของ X จะมีอิทธิพลต่อค่าของ Y ในรูปแบบเชิงเส้นนั่นเอง

2.3.1.7.3 การทดสอบสมมุติฐานเกี่ยวกับ β_1 แบบข้างเดียว

ถ้าต้องการทราบว่าความสัมพันธ์ของตัวแปร Y และ X อยู่ในทิศทางเดียวกัน หรือทิศทางตรงกันข้ามกัน จะต้องทดสอบค่า β_1 ว่าค่า β_1 จะมากกว่าศูนย์ หรือน้อยกว่าศูนย์ ซึ่งสรุปได้ดังนี้

H_0 : $\beta_1 \leq 0$

H_1 : $\beta_1 > 0$

สถิติทดสอบ

$$t = \frac{b}{S_b}$$

เขตปฏิเสธ จะปฏิเสธ H_0 ถ้า $t > t_{1-\alpha;n-2}$

H_0 : $\beta_1 \geq 0$

H_1 : $\beta_1 < 0$

สถิติทดสอบ

$$t = \frac{b}{S_b}$$

เขตปฏิเสธ จะปฏิเสธ H_0 ถ้า $t < -t_{1-\alpha;n-2}$

2.3.1.7.4 การประมาณค่า β_0 แบบช่วง

$$S_a^2 = \hat{V}(a) = \hat{V}(\bar{y} - b\bar{x})$$
$$= \hat{V}(\bar{y}) + \bar{x}^2 \hat{V}(b)$$
$$= \frac{S^2}{n} + \overline{X}^2 \frac{S^2}{SS_{XX}}$$
$$\therefore S_a^2 = S^2\left[1/n + \bar{x}^2 / SS_{XX}\right]$$

ดังนั้นค่าประมาณแบบช่วงของ β_0 ที่ระดับความเชื่อมั่น $(1-\alpha)100\%$ คือ

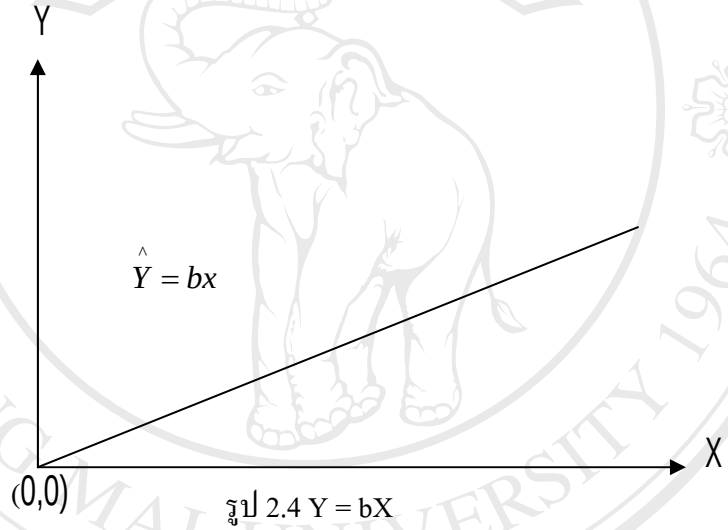
a $\pm t_{1-\alpha/2;n-2} S\sqrt{1/n+\bar{x}^2/SS_X}$ เมื่อขนาดตัวอย่าง n > 30

แต่ถ้า n ≥ 30 ค่าประมาณแบบช่วงของ β_0 ที่ระดับความเชื่อมั่น $(1-\alpha)100\%$ คือ

a $\pm Z_{1-\alpha/2;} S\sqrt{1/n+\bar{x}^2/SS_{XX}}$

2.3.1.7.5 การทดสอบสมมุติฐานเกี่ยวกับ β_0

ถ้า β_0 = 0 แสดงว่า ถ้า X=0 จะทำให้ Y =0 หรือกราฟเส้นตรงผ่านจุดกำเนิด(จุด(X,Y) = (0,0)) ดังแสดงในรูปที่ 2.4



รูป 2.4 Y = bX

กรณีที่ n < 30

สมมุติฐาน $H_0: \beta_0 = 0$

$H_0: \beta_0 \neq 0$

สถิติทดสอบ

$$t = \frac{a-0}{S_a}$$

เขตปฏิเสธ จะปฏิเสธ H_0 ถ้า $|t| > t_{1-\alpha;n-2}$

กรณีที่ n ≥ 30

สมมุติฐาน $H_0: \beta_0 = 0$

$H_0: \beta_0 \neq 0$

สถิติทดสอบ

$$Z = \frac{a}{S_a}$$

เขตปฏิเสธ จะปฏิเสธ H_0 ถ้า $|Z| > Z_{1-\alpha/2}$

2.3.1.7.6 การทดสอบสัมประสิทธ์ความถดถอยโดยใช้การวิเคราะห์ความแปรปรวน

การทดสอบสัมประสิทธิ์ความถดถอย β_1 ซึ่งแสดงถึงความสัมพันธ์ของ X และ Y นอกจากจะใช้สถิติทดสอบ t หรือ Z ดังหัวข้อ 2.3.1.7 แล้วยังสามารถใช้หลักการของการวิเคาะห์ความแปรปรวนมาทดสอบเกี่ยวกับ β_1 ได้ด้วย นั่นคือจะพิจารณาว่าการที่ค่าของ Y มีค่าไม่คงที่อาจมีสาเหตุมาจาก

1. เนื่องจาก Y มีความสัมพันธ์กับ X เมื่อค่าของ X เปลี่ยนไป จะมีผลทำให้ค่าของ Y เปลี่ยนไปด้วย เช่น ถ้าโฆษณามากจะทำให้ยอดขายเพิ่มขึ้น มากกว่ายอดขายของสินค้าที่ไม่ได้โฆษณา
2. ค่าของ Y เปลี่ยนแปลงเนื่องจากอิทธิพลของปัจจัยอื่นๆนอกจาก X เช่น สินค้าที่มีค่าโฆษณาเท่ากัน แต่มียอดขายไม่เท่ากัน อาจมีปัจจัยอื่นๆที่มีอิทธิพลต่อยอดขาย เช่น ราคาขายต่อหน่วย คุณภาพของสินค้า ราคาขายต่อหน่วยของคู่แข่ง เป็นต้น

ส่วนแผนการทดลองแบบสุ่มโดยสมบูรณ์เป็นผลการทดลองที่แบ่งความแปรปรวนของ Y เป็น 2 ส่วนคือ

1. ค่าแปรปรวนของ Y ที่เกิดจากการที่ X เปลี่ยนแปลงไป
2. ค่าแปรปรวนของ Y ที่เกิดจากปัจจัย(ตัวแปร)อื่นๆที่สัมพันธ์กับ Y

ค่าแปรปรวนของ Y คือ SST (Sum Square of Total) ซึ่งคือผลบวกของผลต่างระหว่าง Y_i กับ y^2

$$\text{SST} = \sum\left(Y_i - \bar{y}\right)^2 = \sum Y^2 - \left(\sum Y\right)^2 / n \quad \text{.....(2.13)}$$

หรือ

$$\text{SST} = \sum\left(Y_i - \bar{y}\right)^2 = \sum\left(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{y}\right)^2$$

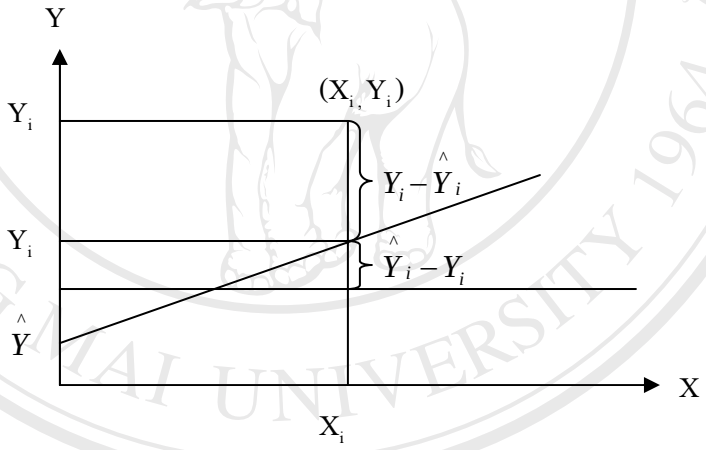
$$= \sum\left(Y_i - \hat{Y}_i\right)^2 + \sum\left(\hat{Y}_i - \bar{y}\right)^2 + 2\sum\left(Y_i - \hat{Y}_i\right)\left(Y_i - \bar{y}\right)$$

โดยที่ $2\sum\left(Y_i-\hat{Y}_i\right)\left(\hat{Y}_i-\bar{y}\right)$ = $2\sum e_i\left(a+bx_i-\bar{y}\right)$

= $2\sum e_i\left[\left(\bar{y}-b\bar{x}\right)+bx_i-\bar{y}\right]$ = 0

$\therefore$ SST = $\sum\left(\hat{Y}_i-\bar{y}\right)^2+\sum\left(Y_i-\hat{Y}\right)^2$

ดังนั้น ค่าแปรปรวนของ Y ทั้งหมด = ค่าแปรปรวนเนื่องจากอิทธิพลของ X + ค่าแปรปรวนเนื่องจากอิทธิพลของปัจจัยอื่นๆ(หรือเรียกว่า ค่าแปรปรวนของความคลาดเคลื่อน)



รูป2.5 กราฟความแปรปรวน

โดยที่ SST (Sum Square of Total) = ค่าแปรปรวนของ Y = $\sum\left(Y_i-\bar{y}\right)^2$

SSR(Sum Square of Regression) = เป็นความแปรปรวนของ Y เนื่องจากอิทธิพลของ X หรือเป็นค่าแปรปรวนของ Y ซึ่งสามารถอธิบายได้โดยตัวแปร X เนื่องจาก

SSR = $\sum\left(\hat{Y}_i-\bar{y}\right)^2$

และ $\hat{Y}$ เป็นค่าประมาณของ Y ซึ่งขึ้นอยู่กับค่า X

$\therefore$ SSR = $\sum\left(\hat{Y}_i-\bar{y}\right)^2 = b^2\sum\left(X_i-\bar{x}\right)^2 = bSS_{XY}$

SSR = $\frac{(SS_{XY})^2}{SS_{XX}}$ ……(2.14)

และ SEE(Sum Square of Error) = ค่าแปรปรวนของ Y เนื่องจากอิทธิพลของปัจจัยอื่นๆ

SSE = SST – SSR = $\sum\left(Y_i - \hat{Y}_i\right)^2$

ดังนั้นการวิเคราะห์ความแปรปรวนแบบทางเดียว(1-WAY ANOVA) เพื่อทดสอบความสัมพันธ์ระหว่าง Y กับ X เป็นดังนี้

ตาราง 2.3 แสดงสูตร 1-WAY ANOVA

แหล่งความแปรปรวน	องศาอิสระ (df)	SS	MS = SS/df	F
ความถดถอย (Regression)	1	SSR	MSR	$\frac{MSR}{MSE}$
ความคลาดเคลื่อน	n-2	SSE	MSE (S^2)	
รวม	n-1	SST		

จากตารางที่ 2.3 จะพบว่า $S^2_{Y.X}$ = S^2 = MSE

เนื่องจาก $\sigma^2 = V(e_i)$ $\therefore S^2 = \hat{V}(e_i)$ = MSE

สมมุติฐานเพื่อการทดสอบ

$H_0 : \beta_1 = 0$ หรือ H_0 : Y และ X ไม่มีความสัมพันธ์ในรูปเส้นตรง

$H_0 : \beta_1 \neq 0$ หรือ H_0 : Y และ X สัมพันธ์ในรูปเส้นตรง

สถิติทดสอบ

F = $\frac{MSR}{MSE}$

เขตปฏิเสธ จะปฏิเสธ H_0 ถ้า $F > F_{1-\alpha;n-2}$

2.3.2 การวิเคราะห์ความถดถอยเชิงซ้อน

ในหัวข้อการวิเคราะห์ความถดถอยอย่างง่าย ได้ศึกษาถึงความสัมพันธ์ของตัวแปร Y กับ ตัวแปรอิสระ X เพียงตัวเดียว นั่นคือ สนใจศึกษาปัจจัยที่มีอิทธิพลหรือผลต่อตัวแปรตาม Y เพียงปัจจัยเดียว แต่โดยทั่วๆไป ปัจจัยที่มีอิทธิพลต่อ Y จะมีหลายปัจจัยหรือกล่าวได้ว่ามีตัวแปรอิสระหลายตัวที่มีอิทธิพลต่อ Y

2.3.2.1 รูปแบบของสมการความถดถอยเชิงซ้อน

ถ้ามีตัวแปรอิสระ k ตัว ($X_1, X_2, \ldots X_k$) ที่มีความสัมพันธ์กับตัวแปรตาม Y โดยที่ความสัมพันธ์อยู่ในรูปเชิงเส้น จะได้สมการความถดถอยเชิงซ้อน ซึ่งแสดงความสัมพันธ์ระหว่าง Y และ $X_1, X_2, \ldots X_k$ ดังนี้

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + e$$

โดยที่ β_0 = ส่วนตัดแกน Y เมื่อกำหนดให้ $X_1 = X_2 = \ldots = X_k = 0$

$\beta_1, \beta_2, \ldots, \beta_k$ เป็นสัมประสิทธิ์ความถดถอยเชิงส่วน (Partial Regression Coeffient) โดยที่ค่า β_i เป็นค่าที่แสดงถึงการเปลี่ยนแปลงของตัวแปรตาม Y เมื่อตัวแปรอิสระ X_i เปลี่ยนไป 1 หน่วย โดยที่ตัวแปรอิสระ X ตัวอื่นๆมีค่าคงที่

2.3.2.2 การประมาณค่าพารามิเตอร์ของสมการความถดถอยเชิงซ้อน

จากสมการความถดถอยเชิงซ้อน ซึ่งมีพารามิเตอร์ k+1 ตัว คือ $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ จะต้องใช้ข้อมูลตัวอย่างของตัวแปร Y , X_1 , X_2 , … , X_k โดยใช้ตัวอย่างขนาด n จากสมการความถดถอยเชิงซ้อน $Y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + e_i$

จะประมาณค่า Y หรือประมาณสมการได้เป็น

$$\hat{Y} = a + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_k X_{ki}$$

การประมาณค่า $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ ด้วยค่า a, b_1 , b_2 , … , b_k ตามลำดับนั้นยังคงมีเป้าหมายเหมือนกับความถดถอยเชิงเส้นอย่างง่าย คือ เพื่อทำให้ผลบวกของค่าคลาดเคลื่อนยกกำลังสองมีค่าน้อยที่สุด โดยใช้วิธีกำลังสองน้อยที่สุด

2.3.2.3 ความหมายของสัมประสิทธิ์ความถดถอยเชิงส่วน

ถ้ามีตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตาม (Y) 2 ตัว คือ X_1 , X_2 สมการความถดถอยเชิงซ้อน คือ $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e_i$

ค่าประมาณของ Y คือ

$\hat{Y} = a + b_1 X_1 + b_2 X_2$

a คือ ส่วนหรือระยะตัดแกน Y ซึ่งหมายถึงเมื่อกำหนดให้ $X_1 = X_2 = 0$

b_1 , b_2 เป็นค่าประมาณของสัมประสิทธิ์การตัดสินใจเชิงส่วน ซึ่งมีหน่วยเหมือน Y และมีความหมายดังนี้

b_1 เป็นค่าซึ่งแสดงถึงความสัมพันธ์ระหว่าง Y และ X_1 หมายถึง ถ้า X_1 เพิ่มขึ้น 1 หน่วย จะทำให้ Y เปลี่ยนไป b_1 หน่วย(ขึ้นอยู่กับเครื่องหมายของ b_1) โดยที่กำหนดให้ตัวแปรอิสระอื่นๆ คือ X_2 มีค่าคงที่

b_2 เป็นค่าซึ่งแสดงถึงความสัมพันธ์ระหว่าง Y และ X_2 หมายถึง ถ้า X_2 เพิ่มขึ้น 1 หน่วย จะทำให้ Y เปลี่ยนไป b_2 หน่วยโดยที่กำหนดให้ตัวแปรอิสระอื่นๆ คือ X_1 มีค่าคงที่

2.3.2.4 สัมประสิทธิ์การตัดสินใจเชิงซ้อน (Multiple Coefficient of Determination : r^2)

สัมประสิทธิ์การตัดสินใจเชิงซ้อน เป็นสัดส่วนหรือเปอร์เซ็นต์ที่ตัวแปรอิสระ (X_1 , X_2 , …, X_k) สามารถอธิบายการเปลี่ยนแปลงของ Y ได้ หรือกล่าวได้ว่าสัมประสิทธิ์การตัดสินใจเชิงซ้อนเป็นสัดส่วนหรือเปอร์เซ็นต์ของความผันแปร Y ที่มีสามาเหตุเนื่องจากความผันแปรของ X_1 , X_2 , …, X_k โดยที่สัมประสิทธิ์การตัดสินใจเชิงซ้อนจะใช้สัญลักษณ์ r^2

$$r^2 = R^2 = \frac{\text{ความผันแปรของ Y เนื่องจากอิทธิพลของ } X_1, X_2 \ldots X_k}{\text{ความผันแปรทั้งหมด}}$$

โดยที่ $0 \le R^2, r^2 \le 1$

ถ้าค่า R^2 ที่ใกล้ 1 จะหมายถึง X_1 , X_2 , …, X_k มีความสัมพันธ์กับ Y มาก แต่ถ้า R^2 เข้าใกล้ศูนย์ หมายถึง ค่า X_1 , X_2 , …, X_k มีความสัมพันธ์กันน้อย

2.3.2.5 สัมประสิทธิ์สหสัมพันธ์เชิงซ้อน (Multiple Coefficient of Correlation : r)

ค่าของสัมประสิทธิ์สหสัมพันธ์เชิงซ้อน ได้จากการถอดรากที่สองของสัมประสิทธิ์การตัดสินใจเชิงซ้อน สัมประสิทธิ์สหสัมพันธ์เชิงซ้อน = r = $\sqrt{r^2}$

โดยที่ $0 \le r \le 1$ แสดงความสัมพันธ์ระหว่าง Y กับ $X_1, X_2, \ldots, X_k$ ดังนี้

- r มีค่าเข้าใกล้ศูนย์ แสดงว่า Y มีความสัมพันธ์กับ $X_1, X_2, \ldots, X_k$ น้อยมาก และถ้า r= 0 แสดงว่า Y ไม่มีความสัมพันธ์กับ $X_1, X_2, \ldots, X_k$ เลย

- r มีค่าเข้าใกล้ 1 แสดงว่า Y มีความสัมพันธ์กับตัวแปรอิสระทั้ง k ตัวมาก

2.3.2.6 สัมประสิทธิ์สหสัมพันธ์เชิงส่วน (Coefficient of Partial Correlation)

สัมประสิทธิ์สหสัมพันธ์เชิงส่วนเป็นค่าที่แสดงความสัมพันธ์ระหว่าง Y กับ X ตัวใดตัวหนึ่ง โดยให้ X ตัวอื่นๆมีค่าคงที่ เช่นถ้า X เป็นความสัมพันธ์กับตัวแปรอิสระ 3 ตัว(X_1,X_2,X_k) สัมประสิทธิ์สหสัมพันธ์เชิงส่วนระหว่าง Y กับ X_i จริงๆโดยกำจัดอิทธิพลของ Xj และ Xk ที่มีต่อ Y

สัญลักษณ์ของสัมประสิทธิ์สหสัมพันธ์เชิงส่วนที่ใช้คือ

$r_{Y1.2}$ = สัญลักษณ์ของสัมประสิทธิ์สหสัมพันธ์เชิงส่วนระหว่าง Y กับ X_1 โดยกำหนดให้ X_2 มีค่าคงที่ เป็นค่าที่แสดงความสัมพันธ์ระหว่าง Y กับ X_1 เท่านั้น (ไม่ได้เป็นความสัมพันธ์กับ X_2)

$r_{Y2.1}$ = สัญลักษณ์ของสัมประสิทธิ์สหสัมพันธ์เชิงส่วนระหว่าง Y กับ X_2 โดยกำหนดให้ X_1 มีค่าคงที่ เป็นค่าที่แสดงความสัมพันธ์ระหว่าง Y กับ X_2 เท่านั้น

โดยที่ $-1 \le r_{Yi;jk} \le 1$

การคำนวณหาสัญลักษณ์ของสัมประสิทธิ์สหสัมพันธ์เชิงส่วนทำได้โดยการใช้สัญลักษณ์ของสัมประสิทธิ์สหสัมพันธ์อย่างง่ายซึ่งเป็นค่าที่แสดงความสัมพันธ์ระหว่างตัวแปร2 ตัว

2.3.2.7 สัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient)

ถ้าผู้วิเคราะห์ใช้ Y คนละหน่วย จะได้ค่า b แตกต่างกัน และค่า b ไม่มีขีดจำกัด(คือไม่ทราบค่าสูงสุดของb) นักสถิติจึงคิดตัวสถิติที่สามารถแสดงความสัมพันธ์ระหว่าง X และ Y ที่สามารถแสดงว่าความสัมพันธ์นั้นมากหรือน้อย

สำหรับสถิติที่ใชวัดความสัมพันธ์ระหว่าง X และ Y ว่ามากหรือน้อยนั้นจะเรียกว่าสัมประสิทธิ์สหสัมพันธ์ (ρ) ซึ่งในกรณีที่ค่าของ Y ขึ้นกับค่า X เพียงตัวเดียวจะเรียกว่าสัมประสิทธิ์สหสัมพันธ์อย่างง่าย โดยที่ ρ จะไม่มีหน่วย จึงสามารถใช้วัดความสัมพันธ์ระหว่าง Y และ X ได้ว่า มีความสัมพันธ์มากหรือน้อยเพียงใด เนื่องจากค่า ρ จะมีค่าสูงสุดเป็น 1 และต่ำสุดเป็น -1 เนื่องจากเราใช้ข้อมูลตัวอย่าง จึงประมาณค่า ρ ด้วยค่า r

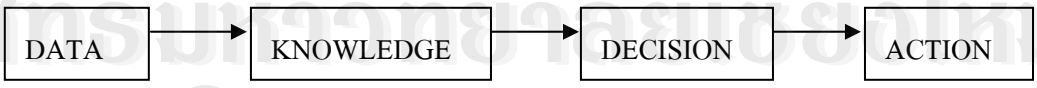
ความหมายของค่า r

1. ค่า r เป็นลบ แสดงว่า X และ Y มีความสัมพันธ์ในทิศทางตรงกันข้าม คือ ถ้า X เพิ่ม Y จะลด แต่ถ้า X ลด Y จะเพิ่ม
2. ค่า r เป็นบวก แสดงว่า X และ Y มีความสัมพันธ์ในทิศทางเดียวกัน คือ ถ้า X เพิ่ม Y จะเพิ่มด้วย แต่ถ้า X ลด Y ก็จะลดด้วย
3. ถ้า r มีค่าเข้าใกล้ 1 หมายถึง X และ Y สัมพันธ์ในทิศทางเดียวกันและมีความสัมพันธ์กันมาก
4. ถ้า r มีค่าเข้าใกล้ -1 หมายถึง X และ Y สัมพันธ์ในทิศทางตรงกันข้ามและมีความสัมพันธ์กันมาก
5. ถ้า r มีค่าเข้าใกล้ 0 แสดงว่า X และ Y ไม่มีความสัมพันธ์กัน
6. ถ้า r เข้าใกล้ 0 แสดงว่า X และ Y มีความสัมพันธ์กันน้อย

2.4 เทคนิคดาต้าไมน์นิ่ง (Data Mining)

Data Mining คือ เทคนิควิธีการวิเคราะห์ข้อมูลที่ได้ถูกออกแบบและพัฒนาขึ้นเป็นชุดซอฟต์แวร์เพื่อเป็นระบบสนับสนุนการตัดสินใจของผู้ใช้ เป็นซอฟต์แวร์ที่สมบูรณ์ทั้งเรื่องการค้นหา การทำรายงาน แบะโปรแกรมในการจัดการ ซึ่งเราคุ้นเคยดีกับคำว่า Executive Information System (EIS) หรือระบบข้อมูลสำหรับการตัดสินใจในการบริหาร ซึ่งเป็นเครื่องมือชิ้นใหม่ที่สามารถค้นหาข้อมูลในฐานข้อมูลขนาดใหญ่หรือข้อมูลที่เป็นประโยชน์ในการบริหาร ซึ่งเป็นการเพิ่มคุณค่าให้กับฐานข้อมูลที่มีอยู่

ระบบสนับสนุนการตัดสินใจ(Decision Support System) คือการทำให้ข้อมูลที่เรามีอยู่กลายเป็นความรู้อันมีคุณค่าได้และสามารถสร้างคำตอบในอนาคตได้ดังรูป



รูป 2.6 Decision Support System

ปัจจุบันระบบสนับสนุนการตัดสินใจได้เข้ามามีอิทธิพลในการรวบรวมข้อมูลและปรับค่าข้อมูลในคลังสินค้า ซึ่งฐานข้อมูลขนาดใหญ่นี้จะประกอบไปด้วยข้อมูลเป็นพันๆล้านไบต์ ยากแก่การค้นหาโดยวิธี DBMS (Database Management System) โดยทั่วไปข้อมูลที่เป็นที่สนใจของผู้บริหารธุรกิจวันนี้สามารถจะค้นหาได้ง่ายขึ้นแล้ว โดยอาศัย Data Mining ด้วยเทคนิค

เดียวกันนี้เราสามารถใช้ค้นข้อมูลสำคัญที่ปะปนกับข้อมูลอื่นๆในฐานข้อมูลที่ไม่ใช่แค่การสุ่มหา บางคนเรียกว่า KKD (Knowledge Discovery in Database) หรือการค้นหาข้อมูลด้วยความรู้

2.4.1 ประเภทข้อมูลที่สามารถทำ Data Mining

1. Relational Database เป็นฐานข้อมูลที่จัดเก็บอยู่ในรูปแบบของตาราง โดยในแต่ละตารางจะประกอบไปด้วยแถวและคอลัมน์ ความสัมพันธ์ของข้อมูลทั้งหมดสามารถแสดงได้โดย Entity-relationship (ER) model

2. Data Warehouse เป็นการเก็บรวบรวมข้อมูลจากหลายแหล่งมาเก็บไว้ในรูปแบบเดียวกันและรวบรวมไว้ในที่ๆเดียวกัน

3. Trasactional Database ประกอบด้วยข้อมูลที่แต่ละ Trasaction แทนด้วยเหตุการณ์ในขณะใดขณะหนึ่ง เช่น ใบเสร็จรับเงิน จะเก็บข้อมูลในรูป ชื่อลูกค้าและรายการสินค้าที่ลูกค้ารายนั้นซื้อ

4. Advanced Database เป็นฐานข้อมูลที่จัดเก็บในรูปแบบอื่นๆ เช่น ข้อมูลแบบ Object-oriented , ข้อมูล multimedia

2.4.2 ลักษณะเฉพาะของข้อมูลที่สามารถทำ Data Mining

1. ข้อมูลขนาดใหญ่ เกินกว่าจะพิจารณาความสัมพันธ์ที่ซ่อนอยู่ภายในข้อมูลได้ด้วยตาเปล่าหรือโดยการใช้ Database Management System (DBMS) ในการจัดการฐานข้อมูล

2. ข้อมูลที่มาจากหลายแหล่ง โดยอาจรวบรวมมาจากหลายระบบปฏิบัติการ หรือหลาย DBMS เช่น Oracle , DB2 , MS SQL , MS Access เป็นต้น

3. ข้อมูลที่ไม่มีการเปลี่ยนแปลงตลอดช่วงเวลาที่ทำการ Mining หากข้อมูลที่มีอยู่นั้นเป็นข้อมูลที่เปลี่ยนแปลงตลอดเวลาจะต้องแก้ปัญหานี้ก่อน โดยบันทึกฐานข้อมูลนั้นไว้และนำฐานข้อมูลที่บันทึกไว้มาทำ Mining แต่เนื่องจากข้อมูลนั้นมีการเปลี่ยนแปลงอยู่ตลอดเวลาจึงทำให้ผลลัพธ์ที่ได้จากการทำ Mining สมเหตุสมผลในช่วงเวลาหนึ่งเท่านั้น ดังนั้นเพื่อให้ได้ผลลัพธ์ที่มีความถูกต้องเหมาะสมอยู่ตลอดเวลาจึงต้องทำ Mining ใหม่ทุกครั้งในช่วงเวลาที่เหมาะสม

4. ข้อมูลที่มีโครงสร้างซับซ้อน เช่น ข้อมูลรูปภาพ ข้อมูลมัลติมีเดีย ข้อมูลเหล่านี้สามารถนำมาทำ Mining ได้เช่นกันแต่ต้องใช้เทคนิคการทำ Data Mining ขั้นสูง

2.4.3 เทคนิคต่างๆของ Data Mining

2.4.3.1 Association rule discovery

เป็นเทคนิคหนึ่งของ Data Mining ที่สำคัญ และสามารถนำไปประยุกต์ใช้ได้จริงกับงานต่างๆหลักทำงานการของวิธีนี้คือ การค้นหาความสัมพันธ์ของข้อมูลจากข้อมูลขนาดใหญ่ที่มีอยู่เพื่อนำไปใช้ในการวิเคราะห์หรือทำนายปรากฏการณ์ต่างๆหรือมาจากการวิเคราะห์การซื้อสินค้าของลูกค้าเรียกว่า “Market Basket Analysis” ซึ่งประเมินจากข้อมูลในตารางที่รวบรวมไว้ ผลการวิเคราะห์ที่ได้จะเป็นคำตอบของปัญหา ซึ่งการวิเคราะห์แบบนี้เป็นการใช้ “กฎความสัมพันธ์” (Association Rule) เพื่อหาความสัมพันธ์ของข้อมูล

ตัวอย่างการนำเทคนิคนี้ไปประยุกต์ใช้กับงานจริง ได้แก่ ระบบแนะนำหนังสือให้กับลูกค้าแบบอัตโนมัติของ Amazon ข้อมูลการสั่งซื้อทั้งหมดของ Amazon ซึ่งมีขนาดใหญ่มากจะถูกนำมาประมวลผลเพื่อหาความสัมพันธ์ของข้อมูล หรือ ลูกค้าที่ซื้อหนังสือเล่มหนึ่งๆ มักจะซื้อหนังสือเล่มใดพร้อมกันด้วยเสมอ ความสัมพันธ์ที่ได้จากกระบวนการนี้จะสามารถนำไปใช้คาดเดาได้ว่าควรแนะนำหนังสือเล่มใดเพิ่มเติมให้กับลูกค้าที่เพิ่งซื้อหนังสือจากร้าน

2.4.3.2 Classification

เป็นกระบวนการสร้าง model จัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ ตัวอย่างเช่น จัดกลุ่มนักเรียนว่า ดีมาก ดี ปานกลาง ไม่ดี โดยพิจารณาจากประวัติและผลการเรียน หรือแบ่งประเภทของลูกค้าว่าเชื่อถือได้ หรือไม่โดยพิจารณาจากข้อมูลที่มีอยู่ กระบวนการ Classification นี้แบ่งออกเป็น 3 ขั้นตอน

1) Model Construction (Learning) เป็นขั้นการสร้าง model โดยการเรียนรู้จากข้อมูลที่ได้กำหนดคลาสไว้เรียบร้อยแล้ว (training data) ซึ่ง model ที่ได้อาจแสดงในรูปของแบบต้นไม้ (Decision Tree) หรือในรูปแบบของ นิวรอลเน็ต (Neural Net)
2) Model Evaluation (Accuracy) เป็นขั้นการประมาณความถูกต้องโดยอาศัยข้อมูลที่ใช้ทดสอบ (testing data) ซึ่งคลาสที่แท้จริงของข้อมูลที่ใช้ทดสอบนี้จะถูกนำมาเปรียบเทียบกับคลาสที่หามาได้จาก model เพื่อทดสอบความถูกต้อง
3) Model Usage (Classification) เป็น Model สำหรับใช้ข้อมูลที่ไม่เคยเห็นมาก่อน (unseen data) โดยจะทำการกำหนด คลาสให้กับ object ใหม่ที่ได้มา หรือ ทำนายค่าออกมาตามที่ต้องการ

2.4.3.3 Prediction

เป็นการทำนายหาที่ต้องการจากข้อมูลที่มีอยู่ ตัวอย่างเช่น หายอดขายของเดือนถัดไปจากข้อมูลที่มีอยู่ หรือทำนายโรคจากอาการของคนไข้ในอดีตเป็นต้น

2.4.3.4 Database Clustering หรือ Segmentation

เป็นเทคนิคการลดขนาดของข้อมูลด้วยการรวมกลุ่มตัวแปรที่มีลักษณะเดียวกันไว้ด้วยกัน ตัวอย่างเช่น บริษัทจำหน่ายรถยนต์ได้แยกกลุ่มลูกค้าออกเป็น 3 กลุ่ม

- กลุ่มมีรายได้สูง (> 80,000)
- กลุ่มมีรายได้ปานกลาง (25,000 – 80,000)
- กลุ่มมีรายได้ต่ำ (< 25,000)

และภายในแต่ละกลุ่มยังแบ่งออกเป็น

- Have Children
- Married
- Last car is a used car
- Own cars

จากข้อมูลข้างต้นทำให้ทางบริษัทรู้ว่าเมื่อมีลูกค้าเข้ามาที่บริษัทควรจะเสนอขายรถประเภทใด เช่น ถ้าเป็นกลุ่มรายได้ค่อนข้างต่ำ ควรเสนอรถมือสอง หรือรถขนาดค่อนข้างเล็ก

2.4.3.5 Deviation Detection

เป็นกรรมวิธีในการหาค่าเฉลี่ยที่แตกต่างไปจากค่ามาตรฐาน หรือค่าที่คาดไว้ว่าต่างไปมากน้อยเพียงใด โดยทั่วไปมักใช้วิธีการทางสถิติ หรือการแสดงให้เห็นภาพ (Virsualization) สำหรับเทคนิคนี้ใช้ในการตรวจสอบ ลายเซ็นปลอม หรือ บัตรเครดิตปลอม รวมทั้งการตรวจหาจุดบกพร่องของชิ้นงานในโรงงานอุตสาหกรรม

2.4.3.6 Link Analysis

จุดมุ่งหมายของ Link Analysis คือ การสร้าง link ที่เรียกว่า “association” ระหว่าง record เดียว หรือ กลุ่มของ record ในฐานข้อมูลLink Analysis สามารถแบ่งออกเป็น 3 ชนิด คือ

- association discovery
- sequential pattern discovery
- similar time sequence discovery