หอสมุดกลาง จุฬาลงกรณ์มหาวิทยาลัย

บทที่ 2

ทฤษฎีทางสถิติ

2.1 สถิติเบื้องต้น (FUNDAMENTAL OF STATISTIC)

2.1.1 ตัวแปร (VARIABLE) คือค่าสังเกตุที่ได้จากการทดลองทั้งจากในสนามและทั้งจากห้องทดลองที่เรียกว่าข้อมูล (DATA) ตัวแปรจะมีคุณสมบัติประจำตัวคือมีความผันแปร (VARIATION) ตัวแปรนี้จำแนกได้เป็น 2 ชนิดคือ

2.1.1.1 ตัวแปรต่อเนื่อง (CONTINUOUS VARIABLE) เป็นตัวแปรใด ๆ ในช่วงพิสัย (RANGE) ที่กำหนดให้ เช่น ข้อมูลจำพวกส่วนสูง น้ำหนัก เวลา เป็นต้น

2.1.1.2 ตัวแปรไม่ต่อเนื่อง (DISCRETE หรือ DISCONTINUOUS-VARIABLE) เป็นตัวแปรที่มิอาจมีได้ทุกค่าในพิสัยนั้น เช่น ข้อมูลจำพวกจำนวนคน จำนวนสิ่งของ เป็นต้น

2.1.2 การกระจาย (DISTRIBUTION) เป็นการปรากฏมากน้อยของตัวแปรในตัวอย่างเรียกว่า ความถี่ (FREQUENCY) ซึ่งทางสถิติคิดเป็นค่าความน่าจะเป็น (PROBABILITY) การปรากฏมากน้อยของตัวแปรเป็นความถี่นี้เรียกว่า การกระจายของตัวแปร

2.1.3 วิธีวัดแนวโน้มหาศูนย์ (CENTRAL TENDENCY) เป็นการหาค่าเฉลี่ยของตัวอย่างซึ่งสุ่มมาจากประชากร การหาค่าเฉลี่ยแบ่งออกได้ 3 วิธีตามความเหมาะสมกับตัวอย่างที่เก็บมา

2.1.3.1 ตัวกลางเลขคณิต (MEAN) เป็นค่าเฉลี่ยของตัวอย่างทางเลขคณิต เป็นค่าที่ใช้บ่อยที่สุด ค่าของความแตกต่างระหว่างตัวแปรกับค่าเฉลี่ยเลขคณิตนี้เรียกว่า ส่วนเบี่ยงเบน (DEVIATION)

$$\bar{X} = \sum X_i / N \qquad (2.1)$$

$$d = X_i - \bar{X} \qquad (2.2)$$

เมื่อ $\bar{X}$ คือตัวกลางเลขคณิต

X_i คือตัวแปรค่าหนึ่ง ๆ

N คือจำนวนของตัวแปรในตัวอย่าง

d คือส่วนเบี่ยงเบน

2.1.3.2 ตัวกลางเรขาคณิต (MEDIAN) เป็นค่าเฉลี่ยอีกแบบหนึ่ง ซึ่งได้จากการเรียงค่าของตัวแปรจากต่ำไปหาสูง โดยค่าของตัวกลางเรขาคณิตคือค่าที่อยู่ตรงกลางสำหรับข้อมูลหรือตัวแปรจำนวนคี่ สำหรับตัวแปรจำนวนคู่นั้น ค่าของตัวกลางเรขาคณิตคือ ค่าเฉลี่ยของตัวแปรตรงคู่กลาง วิธีนี้เหมาะสำหรับตัวแปรที่มีค่ากระจายเท่า ๆ กัน ทั้งทางมากและน้อย

2.1.3.3 MODE เป็นค่าเฉลี่ยที่ได้จากค่าของตัวแปรที่เกิดขึ้นบ่อยที่สุด

2.1.4 วิธีวัดความแตกต่าง (DISPERSION) ของตัวอย่างในการทดลองที่มีการเปรียบเทียบข้อมูลของตัวอย่างแต่ละชุดหลังการทดลองมักจะอาศัยความแปรปรวน (VARIANCE) หรือใช้รากที่สองของความแปรปรวนที่เรียกว่า ส่วนเบี่ยงเบนมาตรฐาน (STANDARD DEVIATION) เป็นเครื่องวัด

$$\sigma^2 = \sum(X_i - \bar{X})^2 / (N - 1) \qquad (2.3)$$

σ^2 คือความแปรปรวน

เมื่อ $X_i - \bar{X}$ คือส่วนเบี่ยงเบน

N คือจำนวนตัวแปรหรือข้อมูล

σ คือส่วนเบี่ยงเบนมาตรฐานหรือ SD

2.2 การหาช่วงเชื่อมั่น (CONFIDENCE INTERVAL) ของค่าเฉลี่ย

ในการคำนวณหาช่วงเชื่อมั่นของตัวกลางหรือค่าเฉลี่ยของตัวอย่างนี้มักจะทำภายใต้ความเชื่อมั่น 95% หรือยอมให้เกิดความคลาดเคลื่อนได้ 5% ($\alpha = 0.05$) และถ้าการกระจายของตัวอย่างเป็นแบบปกติ (NORMAL DISTRBUTION) แล้ว สมการของการกระจายแบบปกติเขียนได้เป็น

$$Z = (X - \bar{X}) / (S/\sqrt{N}) \qquad (2.4)$$

เมื่อ Z คือการกระจายปกติมีค่าเฉลี่ยเป็น 0 และมีความแปรปรวนเป็น 1

S คือส่วนเบี่ยงเบนมาตรฐานของตัวอย่าง

ถ้าจะสร้างช่วงเชื่อมั่น 95% ของค่าเฉลี่ย ($\bar{X}$) ทำได้โดยกำหนดให้ Z เป็น $Z_{0.025}$ ซึ่งมีค่าเท่ากับ 1.960 (N มากกว่า 30) เมื่อเป็นการทดสอบสองทาง (TWO TAILS) แล้ว สูตรความน่าจะเป็นเขียนได้เป็น

$$P\ (\bar{X} - Z_{0.025}\ S/\sqrt{N} \ < X < \bar{X} + Z_{0.025}\ S/\sqrt{N})$$

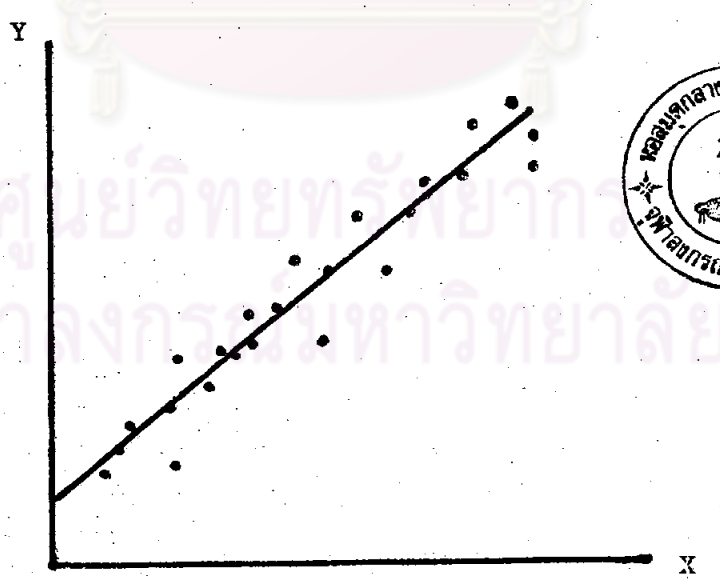
ดังนั้น ช่วงเชื่อมั่น 95% ของ $\bar{X}$ เขียนได้เป็น

$$X = \bar{X} \pm Z_{0.025} S/\sqrt{N} \qquad (2.5)$$

2.3 <u>วิธีการหาเส้นตัวแทนของคู่ความสัมพันธ์</u>

ในการหาความสัมพันธ์ของข้อมูลสองชุดที่มีความสัมพันธ์กันนั้น ในเชิงสถิติสามารถทำได้หลายวิธีแล้ว แต่ความเหมาะสมและความต้องการของผู้วิเคราะห์ วิธีต่าง ๆ ที่นิยมในสามารถแบ่งได้เป็น 4 วิธีคือ

2.3.1 การลากเส้นตัวแทนด้วยวิธีประมาณ (FREE HAND METHOD) วิธีนี้เป็นวิธีที่ง่ายที่สุด คือเป็นการลากเส้นโดยการกะให้ผ่านจุดต่าง ๆ มากที่สุดลงในไดอะแกรมการกระจาย (SCATTER DIAGRAM) ดังแสดงไว้ในรูปที่ 2.1

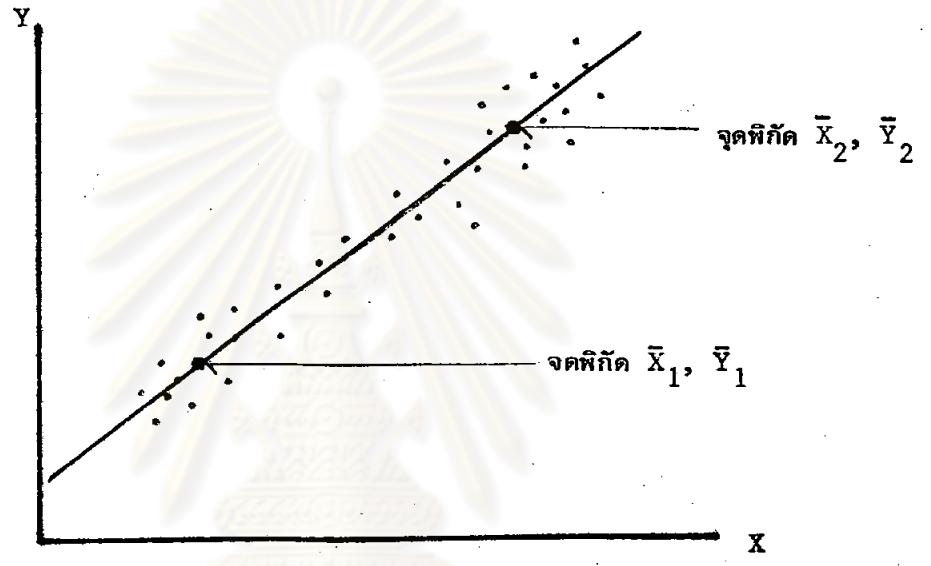


หอสมุดกลาง สถาบันวิทยบริการ จุฬาลงกรณ์มหาวิทยาลัย

รูปที่ 2.1 แสดงการลากเส้นตัวแทนแบบประมาณ

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

2.3.2 การลากเส้นตัวแทนแบบกึ่งเฉลี่ย (SEMI - AVERAGE METHOD) วิธีนี้ต้องเรียงคู่ของข้อมูลจากน้อยไปมาก หรือจากมากไปน้อยแล้วแต่ลักษณะของความสัมพันธ์ จากนั้นแบ่งข้อมูลเป็น 2 ชุด แล้วทำการหาค่าเฉลี่ยของข้อมูลทั้ง 2 ชุด แล้วจึงกำหนดจุดพิกัด (CO - ORDINATED) ทั้งสองจุดลงในไดอะแกรมการกระจาย เส้นตรงที่ลากผ่านจุดทั้งสองคือ เส้นตัวแทนของคู่ความสัมพันธ์ของข้อมูล (ดูรูปที่ 2.2 ประกอบ)



รูปที่ 2.2 แสดงการลากเส้นตัวแทนแบบกึ่งเฉลี่ย

2.3.3 วิธีการลากเส้นตัวแทนแบบค่าเฉลี่ยเคลื่อนที่ (MOVING AVERAGE METHOD) วิธีนี้เหมือนวิธีที่ 2.3.2 แต่แบ่งคู่ของข้อมูลเป็นชุด ๆ ละ 3 คู่ หาค่าเฉลี่ยของทุก ๆ 3 คู่ จากนั้นจึงนำค่าเฉลี่ยทั้งหมดไปลงไว้ในไดอะแกรมการกระจาย แล้วจึงลากเส้นตัวแทนผ่านจุดดังกล่าว แต่วิธีนี้เส้นตัวแทนที่ได้มักจะไม่เป็นเส้นตรง ต้องทำซ้ำหลายครั้งจนได้จุดที่พอจะลากเส้นตรงได้ วิธีนี้จึงไม่ค่อยเป็นที่นิยมและเสียเวลามาก

2.3.4 วิธีการลากเส้นตัวแทนโดยหลักกำลังสองน้อยที่สุด (LEAST - SQUARE METHOD) วิธีนี้เป็นการสร้างสมการเชิงเส้น โดยคำนวณจากข้อมูล หลักเกณฑ์ของวิธีกำลังสองน้อยที่สุดมีดังนี้

2.3.4.1 ผลรวมของความเบี่ยงเบนมาตรฐานของข้อมูลตามแนวดิ่งมีค่าเท่ากับ 0

2.3.4.2 ผลบวกกำลังสองของส่วนเบี่ยงเบน $(Y-\bar{Y})^2$ มีค่าน้อยที่สุด

เมื่อวิเคราะห์ด้วยหลักเกณฑ์ดังกล่าวแล้วสามารถเขียนสมการเส้นตรงได้ดังนี้

$$Y = a + bX$$

เมื่อ a คือจุดตัดบนแกนตั้งมีค่าเท่ากับ $\sum Y/n$

b คือความชันของเส้นตรงมีค่าเท่ากับ $\sum XY/\sum X^2$

จากที่กล่าวมาพบว่า วิธีกำลังสองน้อยที่สุดเป็นวิธีที่เหมาะสมสำหรับการวิเคราะห์มากกว่า 3 วิธีแรก แต่มีข้อจำกัดอยู่ที่ค่าของเลขในแกน X เป็นเลขจำนวนเต็มเรียงลำดับ สำหรับข้อมูลที่มีค่าในแกน X เป็นจำพวกค่าต่อเนื่อง (CONTINUOUS) นั้นจะในการวิเคราะห์ที่เรียกว่า "การถดถอย" (REGRESSION)

2.4 การถดถอยและเส้นถดถอย (REGRESSION AND REGRESSION LINE)

2.4.1 การถดถอย (REGRESSION) เป็นวิธีการหนึ่งซึ่งประยุกต์จากวิธีกำลังสองน้อยที่สุด เพื่อใช้วิเคราะห์หาความสัมพันธ์ตามธรรมชาติของตัวแปรตั้งแต่ 2 ตัวขึ้นไป การถดถอยสามารถจำแนกได้เป็นชนิดตามจำนวนของตัวแปรดังนี้

2.4.1.1 การถดถอยอย่างง่าย (SIMPLY REGRESSION) คือการถดถอยที่มีตัวแปรเพียงตัวเดียว (Y ON X)

2.4.1.2 การถดถอยเชิงซ้อน (COMPLEX REGRESSION) เป็นการถดถอยที่มีตัวแปรตั้งแต่ 2 ตัวขึ้นไป (Y ON X_1, X_2, X_3.....)

2.4.2 เส้นถดถอย (REGRESSION LINE) คือเส้นที่แสดงความสัมพันธ์ของตัวแปรและตัวแปรตาม เป็นเส้นที่จะบอกลักษณะของความสัมพันธ์ โดยส่วนใหญ่แล้วเส้นถดถอยถูกแบ่งออกตามลักษณะของลักษณะความสัมพันธ์ได้เป็น

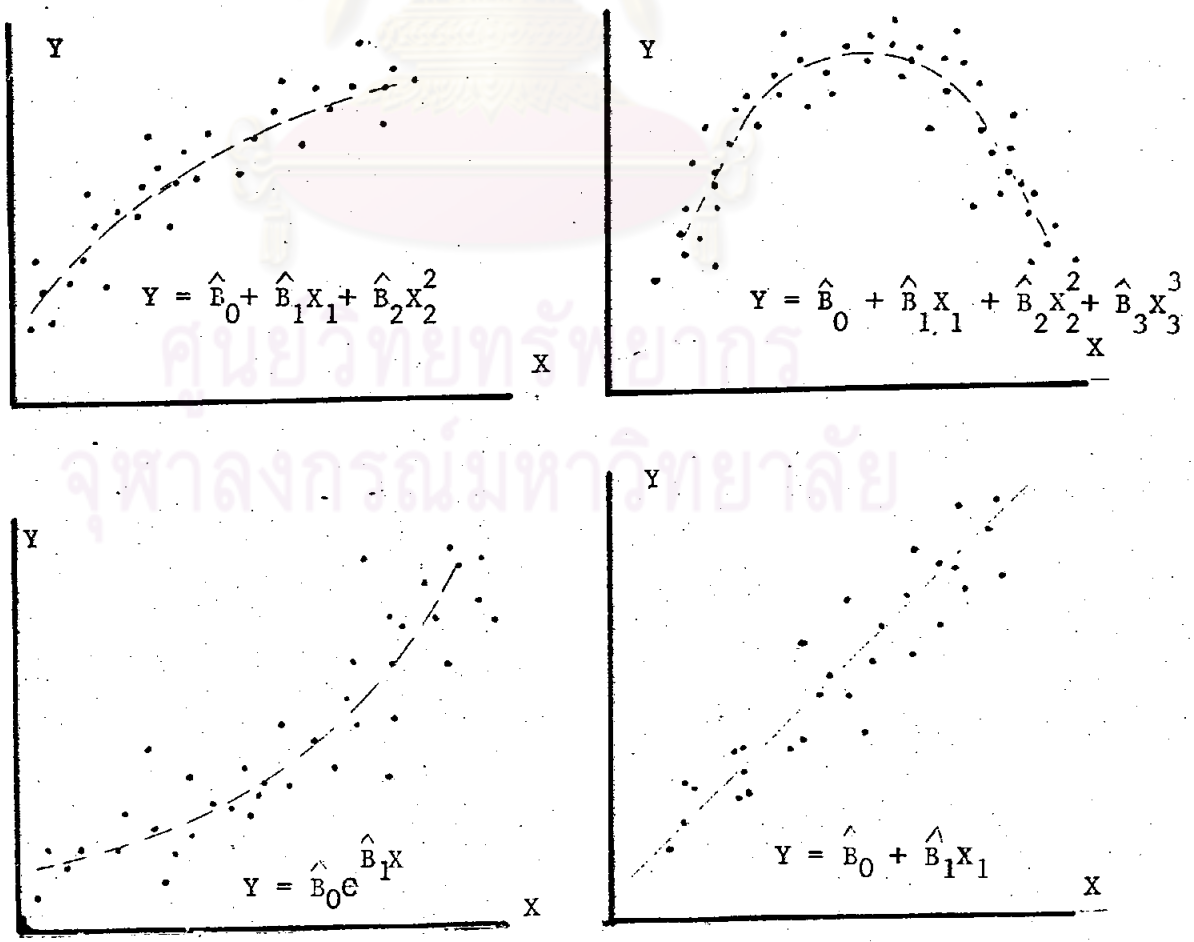
2.4.2.1 เส้นถดถอยที่เป็นเส้นตรง (LINEAR REGRESSION LINE) เป็นเส้นถดถอยซึ่งมีพฤติกรรมเป็นเส้นตรงคือ อัตราเปลี่ยนแปลงในแนวดิ่งต่อ-

แนวนอนจะมีค่าคงที่

2.4.2.2 เส้นถดถอยที่ไม่เป็นเส้นตรง (NON - LINEAR-REGRESSION LINE) เป็นเส้นถดถอยที่มีพฤติกรรมไม่เป็นเส้นตรง คืออัตราการเปลี่ยนแปลงในแนวดิ่งต่อแนวนอนมีค่าไม่เท่ากันตลอดทั้งเส้น ตัวอย่างของการถดถอยชนิดนี้คือ พวกเอ็กโพเนนเชียล (EXPONENTIAL) และพวกโพลีโนเมียล (POLYNOMIAL) เป็นต้น

2.5 การเลือกใช้แบบของเส้นถดถอยให้เหมาะสมกับข้อมูล

ในการวิเคราะห์คู่ของความสัมพันธ์ใด ๆ นั้น ผลที่ได้จะมีค่าถูกต้องหรือสมบูรณ์ขึ้นถ้าการเลือกใช้แบบของเส้นถดถอยเพื่อเป็นตัวแทนของข้อมูลนั้น นับว่ามีส่วนสำคัญมากที่สุด ในการวิเคราะห์นั้นการทำไดอะแกรมการกระจายของข้อมูล (SCATTER DIAGRAM) เป็นสิ่งแรกที่ต้องดำเนินการ ต่อไปจึงศึกษาลักษณะของจุดต่าง ๆ ในไดอะแกรมแล้วจึงถึงขั้นการกำหนดแบบ (MODEL) ของเส้นถดถอยที่ผู้วิจัยคิดว่าเหมาะสม พิจารณารูปที่ 2.3 ซึ่งแสดงการใช้แบบของเส้นถดถอยแบบต่าง ๆ กับข้อมูลที่มีการกระจายต่าง ๆ กัน



รูปที่ 2.3 แสดงการใช้แบบ (MODEL) ให้เหมาะสมกับข้อมูล

2.6 เส้นถดถอยที่เป็นเส้นตรง (LINEAR REGRESSION)

สมการของเส้นถดถอยที่เป็นเส้นตรงสามารถเขียนได้เป็น

$$Y = B_o + B_1X \qquad (2.6)$$

สมการ (2.6) นี้เป็นสมการถดถอยของประชากรทั้งหมดของตัวอย่างที่ได้เก็บมาโดยที่

B_o คือจุดตัดบนแกน Y

B_1 คือความชันของเส้นถดถอย

$B_o + B_1X$ คือค่าคาดหวังของ Y ซึ่งแปรตาม X

ในการวิจัยไม่สามารถทราบเส้นถดถอยของประชากรได้ เส้นถดถอยซึ่งวิเคราะห์จากตัวอย่างจำนวน n ที่เก็บมานั้นเป็นเพียงเส้นถดถอยของตัวอย่างซึ่งใช้ประมาณเส้นถดถอยของประชากร ด้วยเหตุผลดังกล่าวความคลาดเคลื่อนย่อมเกิดขึ้นได้ดังสมการ (2.7)

$$Y = \hat{B}_o + \hat{B}_1X + e \qquad (2.7)$$

เมื่อ $\hat{B}_o$ ค่าประมาณของ B_o ในสมการ (2.6)

$\hat{B}_1$ ค่าประมาณของ B_1 ในสมการ (2.6)

e ความคลาดเคลื่อนที่เกิดขึ้น

พิจารณาสมการ (2.7) ถ้าให้ $\hat{Y} = Y - e$ เป็นค่าประมาณของ Y แล้วจะสามารถเขียนสมการการถดถอยเชิงเส้นตรงของตัวอย่างได้เป็น

$$\hat{Y} = \hat{B}_o + \hat{B}_1X \qquad (2.8)$$

ค่าของ $\hat{B}_o$ และ $\hat{B}_1$ มีค่าเป็น + หรือ - ก็ได้แล้วแต่แนวโน้มของข้อมูล ค่าของ $\hat{B}_o$ และ $\hat{B}_1$ สามารถหาได้โดยอาศัยหลักกำลังสองน้อยที่สุด ดังนี้

2.6.1 ค่าของ $\bar{X}$ และ $\bar{Y}$ มีค่าอยู่บนเส้นถดถอย

2.6.2 ผลบวกของส่วนเบี่ยงเบนจากเส้นถดถอยในแนวตั้งของข้อมูลรวมกันมีค่าเป็น 0.

$$\sum(Y - Y) = 0$$

2.6.3 ผลบวกของส่วนเบี่ยงเบนจากเส้นถดถอยในแนวตั้งกำลังสองเมื่อรวมกันมีค่าน้อยที่สุด

$$\sum(Y - \bar{Y})^2 \text{ มีค่าน้อยที่สุด}$$

เมื่อเราหาอนุพันธ์ (DIFFERENTIAT) $\sum(Y - \bar{Y})^2$ เทียบกับ $\hat{B}_o$ และ $\hat{B}_1$ แล้วจัดให้เท่ากับ o จะได้

$$\hat{B}_1 = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} \qquad \text{(2.9)}$$

หรือ

$$\hat{B}_1 = \frac{\sum(X-\bar{X}).(Y-\bar{Y})}{\sum(X-\bar{X})^2} \qquad \text{(2.10)}$$

และ $$\hat{B}_o = \bar{Y} - \hat{B}_1\bar{X} \qquad \text{(2.11)}$$

ในการวิเคราะห์เมื่อสามารถหาค่าของ $\hat{B}_o$ และ $\hat{B}_1$ ได้แล้วก็สามารถลากเส้นถดถอยลงใน SCATTER DIAGRAM ได้ สำหรับอัตราของความสัมพันธ์ของเส้นถดถอยนั้นจะกล่าวถึงในหัวข้อต่อไป

2.7 สหสัมพันธ์ (CORRELATION)

สหสัมพันธ์คือ ตัวแสดงค่าแห่งความสัมพันธ์หรืองศาแห่งความสัมพันธ์ (DEGREE OF RELATIONSHIP) ระหว่างตัวแปรตั้งแต่สองตัวแปรขึ้นไป โดยจะบอกถึงความความสัมพันธ์ว่ามากน้อยเพียงใด

2.7.1 สัมประสิทธิ์ของสหสัมพันธ์ (r) คือค่าที่บ่งบอกสหสัมพันธ์ของเซทข้อมูล ถ้า X เป็นตัวแปร,Y เป็นตัวตาม,n เป็นจำนวนข้อมูลแล้ว

$$r = \sqrt{\sum(Y - \hat{Y})^2 / \sum(Y - \bar{Y})^2} \qquad \text{(2.12)}$$

แต่การหาค่าสัมประสิทธิ์จากสมการที่ (2.12) ทำได้ไม่สะดวกเพราะจะต้องหาค่า $\hat{Y}$ จากเส้นถดถอยก่อนทุกค่าของ X ซึ่งทำให้เสียเวลามาก ในกรณีที่ข้อมูลมีจำนวน-

มาก ๆ เราสามารถหาค่า r ได้จากสมการที่ (2.15) ซึ่งสะดวกกว่าเพราะสามารถหาค่า r ได้โดยตรงจากข้อมูล

$$r = \frac{n\sum XY - \sum X\sum Y}{\sqrt{\left[n\sum X^2 - (\sum X)^2\right]\left[n\sum Y^2 - (\sum Y)^2\right]}} \qquad (2.15)$$

เครื่องหมายของ r จะเหมือนกับเครื่องหมายของ $\hat{B}_1$ กล่าวคือ ถ้าความชันของเส้นถดถอยเป็นลบ สัมประสิทธิ์สหสัมพันธ์ก็เป็นลบ แต่ถ้าความชันเป็นบวก สัมประสิทธิ์สหสัมพันธ์ก็จะเป็นบวกเช่นกัน

2.7.2 การตีความหมายของ r ค่าของ r จะมีค่าอยู่ระหว่าง -1 ถึง +1 สามารถตีความหมายได้ดังนี้ ถ้า r มีค่าเท่ากับ ± 1 หรือหมายความว่าจุดทุกจุดสามารถเขียนเส้นตรงผ่านได้หรือข้อมูลทั้งหมดไม่มีส่วนเบี่ยงเบนเลย ถ้า r มีค่าเป็นศูนย์ หมายความว่าการกำหนดเส้นถดถอยนั้นใช้ไม่ได้เลยหรือทำนายไม่ได้เลย ถ้า r มีค่าใกล้หรือเข้าใกล้ศูนย์แสดงว่าความสัมพันธ์มีน้อยมาก (FIT IS POOR) ถ้า r มีค่าใกล้หรือเข้าใกล้ ± 1 แสดงว่าความสัมพันธ์มีมาก (FIT IS GOOD) แต่ถ้า r มีค่าอยู่ระหว่าง -1 ถึง +1 แล้วก็ขึ้นอยู่กับการตัดสินใจของผู้ที่จะนำไปใช้ว่าจะยอมรับเส้นถดถอยนั้นมากน้อยเพียงไร

2.7.3 กำลังสองของสัมประสิทธิ์สหสัมพันธ์ $(r)^2$ กำลังสองของ r จะเป็นตัวบ่งบอกถึงอัตราความสัมพันธ์ของตัวแปรที่มีต่อตัวแปรตามว่ามีมากน้อยเพียงใด โดยพิจารณาจากส่วนเบี่ยงเบนมักจะนิยมบอกเป็นเปอร์เซ็นต์ตัวอย่างเช่นถ้าคำนวณค่า r ได้เท่ากับ 0.94 กำลังสองของ r จะเป็น

$$r^2 = (0.94)^2 \text{ หรือเท่ากับ } 88\%$$

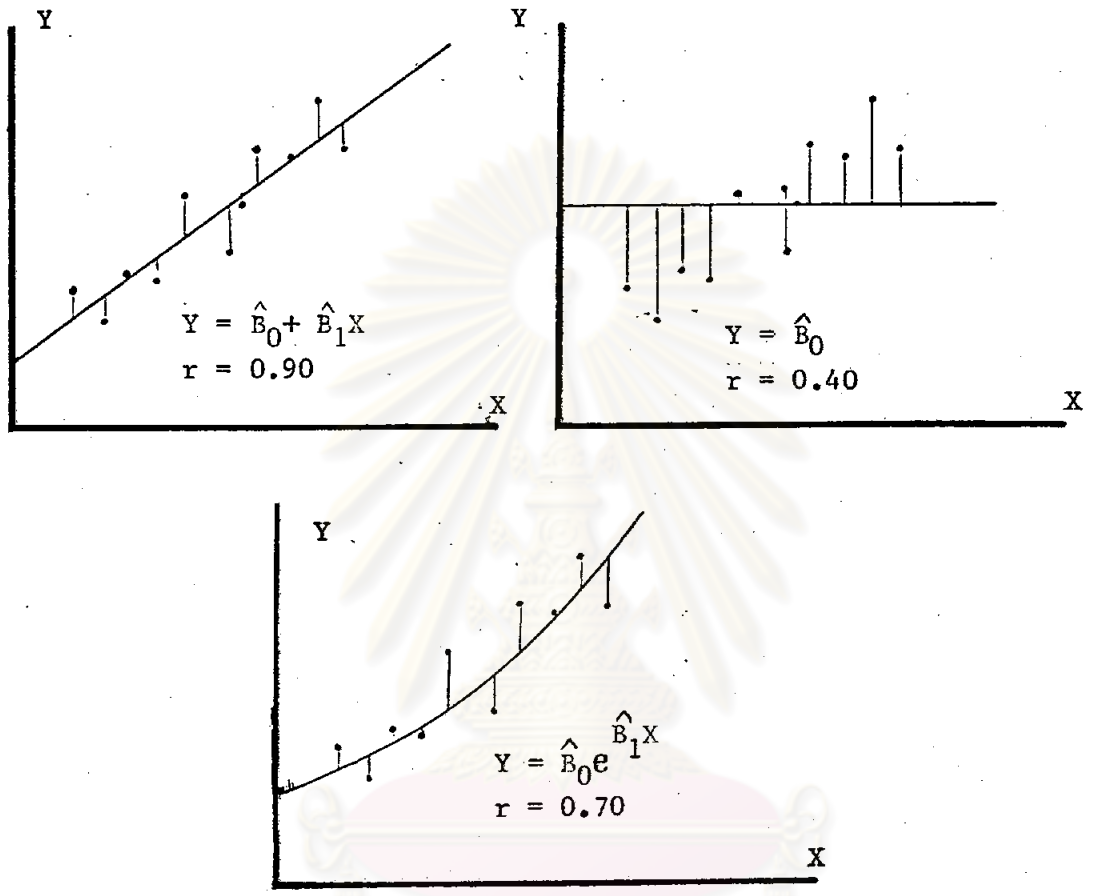
อธิบายได้ว่าตัวแปรตามจะแปรผันตามตัวแปรด้วยอัตรา 88% ส่วนอีก 22% จะเนื่องจากสาเหตุอื่น

2.8 <u>ข้อควรระวังในการวิเคราะห์โดยการถดถอยและสหสัมพันธ์</u>

จากที่กล่าวมาแล้วเบื้องต้นได้อะแกรมการกระจาย (SCATTER DIAGRAM) แล้ว ความผิดพลาดในการวิเคราะห์อาจเกิดขึ้นได้ถ้าผู้วิเคราะห์ตัดสินใจใช้ตัวแบบ (MODEL)

013414

ของเส้นถดถอยที่ไม่ใช่เส้นที่เหมาะสมที่สุดมาเป็นตัวแทน ดังแสดงไว้ในรูปที่ 2.4



รูปที่ 2.4 แสดงการเปรียบเทียบการใช้ตัวแบบต่าง ๆ กับข้อมูลชุดเดียวกัน

จากรูปที่ 2.4 แสดงการใช้ตัวแบบ (MODEL) ของเส้นถดถอยต่าง ๆ กันในข้อมูลชุดเดียวกัน ถ้าพิจารณาจากค่า r ของแต่ละแบบแล้วพบว่าแบบ (MODEL) การถดถอยแบบเส้นตรงจะเป็นตัวแทนที่ดีที่สุดคือมีค่า r เท่ากับ 0.90