บทที่3

หลักสถิติ

การพยากรณ์ คือ การประมาณหรือการคาดว่าอะไรจะเกิดขึ้นในอนาคต เพื่อนำมาข้อมูลมาใช้ในการวางแผน การพยากรณ์แบ่งได้เป็น 2 ประเภท คือ

1. การพยากรณ์เชิงคุณภาพ เป็นการพยากรณ์ที่ใช้ผู้ที่มีประสบการณ์ ความรู้ ความสามารถจึงตรวจสอบความแม่นยำของการพยากรณ์ได้ยาก แบ่งออกเป็น 5 เทคนิคย่อย คือ การคาดคะเน, การระดมความคิด, การพยากรณ์ยอดขาย, การพยากรณ์โดยการสำรวจ และการพยากรณ์ด้วยเทคนิคเดลไฟ

2. การพยากรณ์เชิงปริมาณ เป็นการพยากรณ์ที่ใช้ข้อมูลเชิงปริมาณ (ตัวเลข) ในอดีตเพื่อนำมาพยากรณ์ค่าในอนาคต โดยสร้างตัวแบบทางคณิตศาสตร์ แบ่งออกเป็น 2 เทคนิคย่อย คือ การพยากรณ์ความสัมพันธ์ และการพยากรณ์อนุกรมเวลา

ในการวิจัยครั้งนี้จะใช้การวิเคราะห์ความถดถอยและสหสัมพันธ์ซึ่งเป็นเทคนิคการพยากรณ์เชิงปริมาณเพื่อใช้หาความสัมพันธ์กับตัวแปรที่จะพยากรณ์ โดยเนื้อหาในบทนี้จะกล่าวถึงการหาค่าพารามิเตอร์ของสมการความถดถอยแบบง่ายและเชิงซ้อน, การหาค่าสัมประสิทธิ์การตัดสินใจ, การตรวจสอบสมมุติฐานของการวิเคราะห์ และข้อจำกัดในการพยากรณ์

3.1 การวิเคราะห์ถดถอย (Regression Methods)

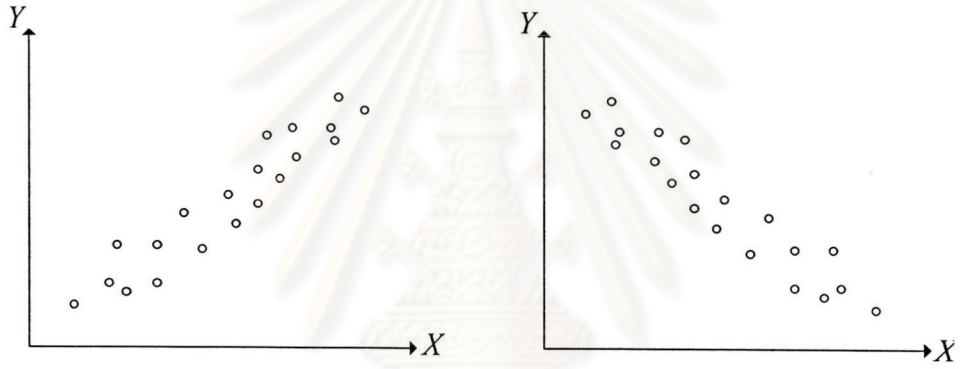
เป็นการศึกษาถึงความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามหรือลักษณะที่สนใจศึกษา โดยที่ต้องทราบค่าของตัวแปรต้นหรือต้องกำหนดค่าของตัวแปรต้นไว้ล่วงหน้า เช่นต้องการศึกษาถึงความสัมพันธ์ระหว่างกระแสรั่วไหลกับความต้านทานกระแสรั่วไหล หรือศึกษาความสัมพันธ์ของค่าแรงดันเสียสภาพฉับพลันของแกปทรงกลมกับระยะแกปและความดันอากาศ เป็นต้น โดยจะเรียกความต้านทานกระแสรั่วไหล ระยะแกปและความดันอากาศ ซึ่งเป็นตัวแปรที่ทราบหรือทำการกำหนดค่าไว้ล่วงหน้าว่า ตัวแปรอิสระ (Independent Variable) และมักจะใช้สัญลักษณ์ X ส่วนกระแสรั่วไหล หรือ ค่าแรงดันเสียสภาพฉับพลัน จะเรียกว่า ตัวแปรตาม (Dependent Variable) และใช้สัญลักษณ์ Y

3.1.1 วัตถุประสงค์ของการวิเคราะห์ความถดถอย

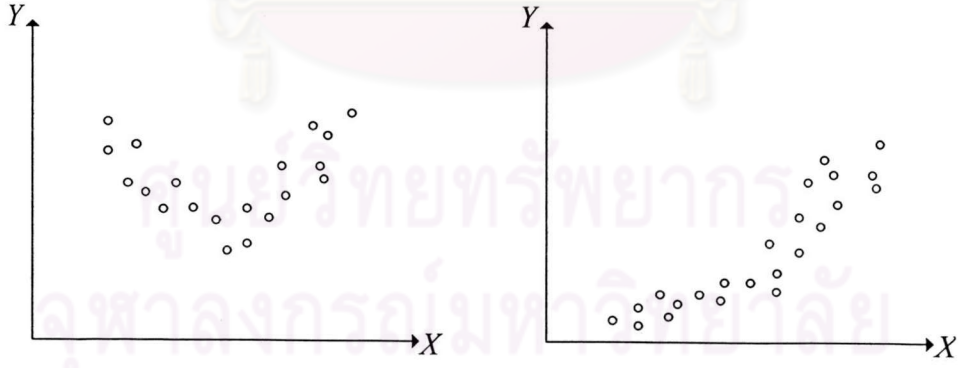
การวิเคราะห์ความสัมพันธ์ของตัวแปรต่างๆมีวัตถุประสงค์ดังนี้

- เพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรว่ามีความสัมพันธ์กันมากน้อยเพียงใด ถ้า X และ Y มีความสัมพันธ์กันมาก แสดงว่า เมื่อ X มีค่าเปลี่ยนแปลงไปจะมีผลกระทบต่อค่าของ Y เป็นอย่างมาก
- ใช้ความสัมพันธ์ที่วิเคราะห์ได้มาประมาณหรือพยากรณ์ค่า Y ในอนาคตเมื่อกำหนดค่า X

สำหรับการหารูปแบบความสัมพันธ์ระหว่างตัวแปร Y และ X นั้นในขั้นแรกจะนำเอาข้อมูลของตัวแปรทั้งหมดมาเขียนกราฟแสดงความสัมพันธ์ ซึ่งจะเรียกกราฟนี้ว่า แผนภาพการกระจาย (Scatter Diagram) ผู้วิเคราะห์ต้องพิจารณาจากแผนภาพการกระจายว่าความสัมพันธ์ของตัวแปรอิสระกับตัวแปรตามจะอยู่ในรูปแบบใด เช่น เส้นตรง พาราโบลาหรือเส้นโค้ง เป็นต้น

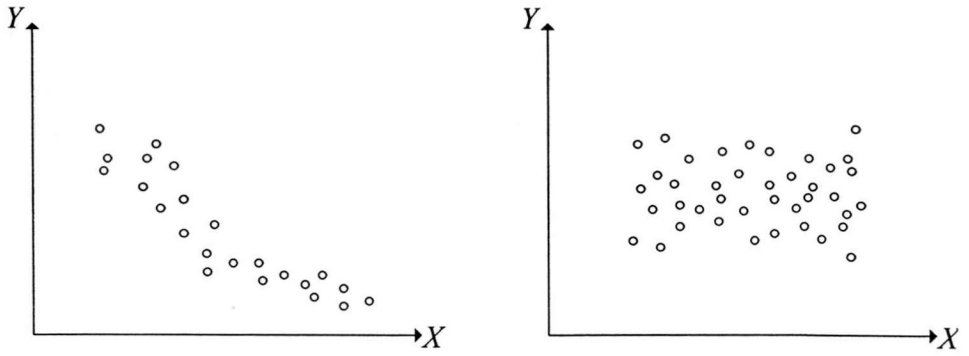


ก) ความสัมพันธ์อยู่ในรูปเส้นตรงและเป็นบวก ข) ความสัมพันธ์อยู่ในรูปเส้นตรงและเป็นลบ



ค) ความสัมพันธ์อยู่ในรูปพาราโบลา ง) ความสัมพันธ์อยู่ในรูปเอกซ์โพเนนเชียลและเป็นบวก

รูปที่ 3.1 ตัวอย่างแผนภาพการกระจาย



จ) ความสัมพันธ์อยู่ในรูปเอกซ์โพเนนเชียล และเป็นลบ

ฉ) ไม่มีความสัมพันธ์

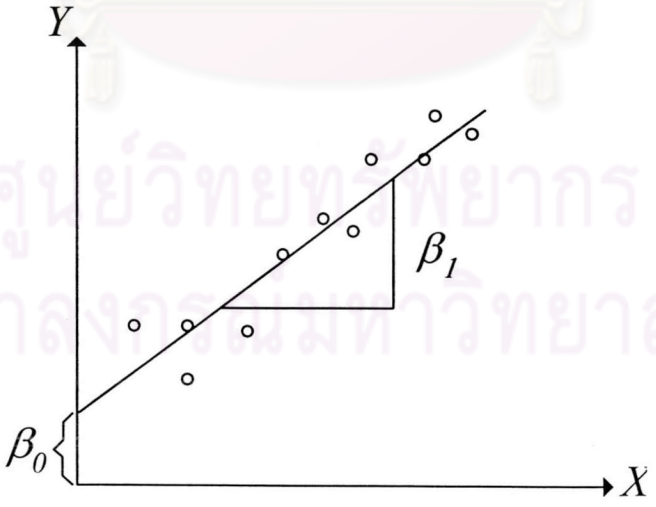
รูปที่ 3.1 (ต่อ) ตัวอย่างแผนภาพการกระจาย

โดยในการวิจัยครั้งนี้จะกล่าวถึงวิธีการวิเคราะห์ความถดถอยด้วยกัน 2 แบบ คือ การวิเคราะห์ความถดถอยเชิงเส้นอย่างง่าย และ การวิเคราะห์ความถดถอยเชิงซ้อน

3.1.2 การวิเคราะห์ความถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression Analysis)

เป็นการศึกษาถึงความสัมพันธ์ระหว่างตัวแปร 2 ตัว ที่มีความสัมพันธ์อยู่ในรูปเชิงเส้น ซึ่งสามารถแสดงความสัมพันธ์ในรูปสมการได้ดังนี้

$$Y_i = \beta_0 + \beta_1 X_i + e_i; i = 1, 2, ..., N \qquad (3.1)$$



รูปที่ 3.2

โดยที่ Y = ตัวแปรตาม (Dependent Variable)

X = ตัวแปรอิสระ (Independent Variable)

β_0 = ส่วนตัดแกน Y

e = ความคลาดเคลื่อนอย่างสุ่ม (random error)

β_1 = ความชัน (slope) ของเส้นตรง

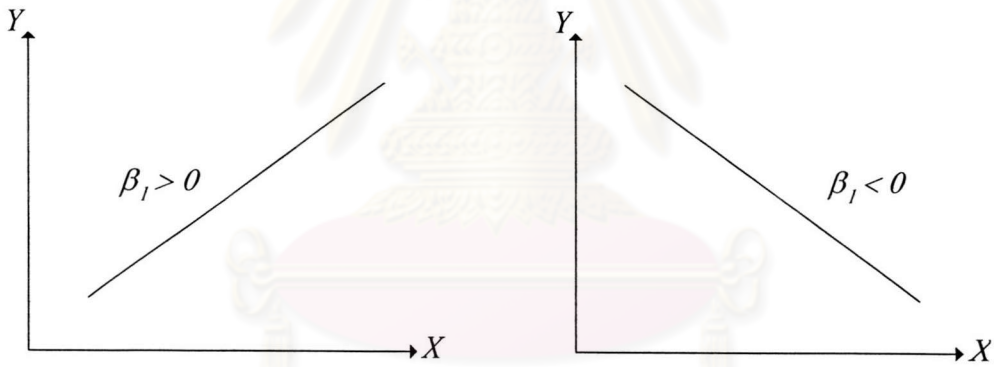
และจะเรียก β_1 ว่าสัมประสิทธิ์ความถดถอย (Regression Coefficient) ค่าของ β_1 อาจจะเป็น

ก. $\beta_1 > 0$ แสดงว่า X และ Y มีความสัมพันธ์ในทางเดียวกันคือถ้า X เพิ่ม Y จะเพิ่มด้วย แต่ถ้า x ลดลง Y จะลดลงด้วย

ข. $\beta_1 < 0$ แสดงว่า X และ Y มีความสัมพันธ์ในทางตรงข้ามกันคือถ้า X เพิ่ม Y จะลดลงแต่ถ้า X ลดลง Y จะเพิ่มขึ้น

ค. β_1 มีค่าเข้าใกล้ศูนย์ แสดงว่า X และ Y มีความสัมพันธ์กันน้อย

ง. $\beta_1 = 0$ แสดงว่า X และ Y ไม่มีความสัมพันธ์กันเลย



รูปที่ 3.3 แสดงค่า β_1 เมื่อ X และ Y มีความสัมพันธ์รูปเส้นตรง

ก) สมมุติฐานของการวิเคราะห์ความถดถอยเชิงเส้นอย่างง่าย

1. ค่า X จะต้องเป็นค่าที่กำหนดไว้ล่วงหน้าหรือทราบค่า
2. ความคลาดเคลื่อน e_i เป็นตัวแปรที่มีค่าเฉลี่ยเท่ากับศูนย์ หรือ $E(e_i) = 0$

   ค่าแปรปรวนของ e_i มีค่าเท่ากันทุกค่าของ i และมีค่าเท่ากับค่าแปรปรวนของ Y

$$V(e_i) = V(Y) = \sigma^2_{yx} = \sigma^2$$

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

4. e_i มีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเป็นศูนย์และค่าแปรปรวนเท่ากับ σ^2 นั้นคือ

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

จากข้อสมมุติข้างต้น จะได้ว่า

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

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

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

$$= \beta_0 + \beta_1 X_i$$

ข) การประมาณค่าพารามิเตอร์ของสมการความถดถอย

เมื่อพิจารณาจากแผนภาพการกระจาย ซึ่งแสดงความสัมพันธ์ระหว่าง X และ Y แล้วพบว่า X และ Y สัมพันธ์กันในรูปเส้นตรง จะต้องคำนวณหาค่า β_0 และ β_1 ซึ่งจะทำให้ทราบถึงความสัมพันธ์ระหว่าง X และ Y ว่ามีความสัมพันธ์ตามกันหรือตรงกันข้ามกันและความสัมพันธ์นั้นมากหรือน้อยเพียงใด ถ้า β_1 มีค่ามากแสดงว่า Y มีความสัมพันธ์กับ X มากด้วย

การที่จะหาค่า β_0 และ β_1 ได้จำเป็นจะต้องทราบค่า X และ Y ทุกค่าที่ได้เกิดขึ้นแล้วในอดีต เช่น ถ้า X คือระยะแกป และ Y คือค่าแรงดันเสียสภาพฉับพลันของแกปทรงกลม การหาค่า β_0 และ β_1 จะต้องทราบระยะแกปและค่าแรงดันเสียสภาพฉับพลันของแกปทรงกลมทุกค่า ซึ่งเป็นไปได้ยาก ในทางปฏิบัติเราจึงใช้ข้อมูลตัวอย่างขนาด n ในการประมาณค่า β_0 และ β_1 ดังนั้นค่าประมาณของ Y คือ

$$Y_i = \beta_0 + \beta_1 X_i \quad \text{หรือ} \quad \hat{Y}_i = a + bX_i \tag{3.2}$$

โดยที่ $\beta_0 = a, \; \beta_1 = b$

การประมาณค่า β_0 และ β_1 ด้วย a และ b โดยวิธีกำลังสองน้อยที่สุด (Least Square Method) เป็นวิธีการประมาณค่า Y_i ด้วย $\hat{Y}_i$ เพื่อให้ความคลาดเคลื่อนต่ำที่สุด

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

และ $\hat{Y}_i = a + bX_i$

จะได้ $Y_i - \hat{Y}_i = e_i$

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

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

การที่ต้องใช้ผลบวกของค่าคลาดเคลื่อนยกกำลังสองเนื่องจากค่า e_i อาจจะมีค่าบวกเมื่อ Y_i มากกว่า $\hat{Y}_i$ และจะเป็นค่าลบ ถ้า Y_i น้อยกว่า $\hat{Y}_i$ ซึ่งอาจมีผลทำให้ $\sum e_i$ เป็นศูนย์หรือมีค่าน้อยกว่าที่เป็นจริง

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

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

$$-2\sum \left(Y_i - a - bX_i \right) = 0$$

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

$$an + b\sum_{1}^{n} X_i = \sum_{1}^{n} Y_i \qquad (3.3)$$

และ

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

$$a\sum_{1}^{n} X_i + b\sum_{1}^{n} X_i^2 = \sum_{1}^{n} X_i Y_i \qquad (3.4)$$

และเรียกสมการ (3.3) และ (3.4) ว่าสมการปกติ (Normal Equations) แก้สมการที่ (3.3) และ (3.4) เพื่อหาค่า a และ b ดังนี้

$\sum X_i \times (3.3)$; $$an(\sum X_i) + b(\sum X_i)^2 = (\sum X_i)(\sum Y_i) \quad (3.5)$$

$n \times (3.4)$; $$an(\sum X_i) + bn(\sum X_i^2) = n(\sum X_i Y_i) \quad (3.6)$$

$(3.6) - (3.5)$; $$bn(\sum X_i^2) - b(\sum X_i)^2 = n(\sum X_i Y_i) - (\sum X_i)(\sum Y_i)$$

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

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

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

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

จาก (3.3); $$an + b(\sum X_i) = \sum Y_i$$

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

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

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

$\therefore$ $$\widehat{\beta}_1 = b = \frac{SS_{XY}}{SS_{XX}} \quad (3.7)$$

$$\widehat{\beta}_0 = a = \overline{Y} - b\overline{X} \quad (3.8)$$

โดยที่ $$SS_{XX} = \sum (X_i - \overline{X})^2 = \sum_1^n X_i^2 - \frac{(\sum_1^n X_i)^2}{n} \quad (3.9)$$

$$SS_{XY} = \sum (X_i - \overline{X})(Y_i - \overline{Y}) = \sum_1^n X_i Y_i - \frac{(\sum_1^n X_i)(\sum_1^n Y_i)}{n} \quad (3.10)$$

$$SS_{YY} = \sum (Y_i - \overline{Y})^2 = \sum_{1}^{n} Y_i^2 - \frac{(\sum_{1}^{n} Y_i)^2}{n} \qquad (3.11)$$

การประมาณค่า β_0 และ β_1 ด้วยค่า a และ b โดยใช้วิธีกำลังสองน้อยที่สุดจะทำให้

- ผลรวมของค่าคลาดเคลื่อนในการประมาณค่า Y_i ด้วย $\hat{Y}$ เป็นศูนย์

  คือ $\sum (Y_i - \hat{Y}_i) = \sum e_i = 0$
- จุด $(\overline{X}, \overline{Y})$ เป็นจุดที่อยู่บนเส้นความถดถอย
- $\sum (Y_i - \hat{Y}_i)^2$ มีค่าต่ำสุด

ค) สัมประสิทธิ์การตัดสินใจ (Coefficient of Determination: R^2)

สัมประสิทธิ์การตัดสินใจ หมายถึงสัดส่วนที่ตัวแปร X สามารถอธิบายการเปลี่ยนแปลงของตัวแปร Y ได้ ดังนั้นถ้า R^2 มีค่ามากแสดงว่า Y และ X มีความสัมพันธ์กันมากหรือ X สามารถอธิบายการเปลี่ยนแปลงค่าของ Y ได้โดยที่

$$R^2 = \frac{\text{ความแปรปรวนของ } Y \text{ ที่เกิดจาก } X}{\text{ความแปรปรวนของ } Y \text{ ทั้งหมด}} \qquad (3.12)$$

$$\therefore R^2 = \frac{SSR}{SST} = \frac{bSS_{XY}}{SS_{YY}} \qquad (3.13)$$

แต่เนื่องจาก $SST = SSR + SSE$

$$R^2 = 1 - \frac{SSE}{SST} \qquad (3.14)$$

โดยที่

SST = Sum Square of Total

$$= \sum (Y_i - \overline{Y})^2 = \sum (\hat{Y}_i - \overline{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 \qquad (3.15)$$

SSR = Sum Square of Regression

$$= \sum (\hat{Y}_i - \overline{Y})^2 = b^2 \sum (X_i - X)^2 = bSS_{XY} = \frac{(SS_{XY})^2}{SS_{XX}} \qquad (3.16)$$

SSE = Sum Square of Error

$$= \sum (Y_i - \hat{Y}_i)^2 = SS_{YY} - \frac{(SS_{XY})^2}{SS_{XX}} \quad (3.17)$$

คุณสมบัติของ R^2

- R^2 ไม่มีหน่วย
- ถ้า R^2 มีค่าเข้าใกล้ 1 แสดงว่าเปอร์เซ็นต์ที่ X สามารถอธิบายการเปลี่ยนแปลงของ Y มีค่ามาก หรือ X และ Y มีความสัมพันธ์กันมาก แต่ถ้า R^2 มีค่าเข้าใกล้ 0 แสดงว่าเปอร์เซ็นต์ที่ X สามารถอธิบายการเปลี่ยนแปลงของ Y มีค่าน้อย

ง) สัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient)

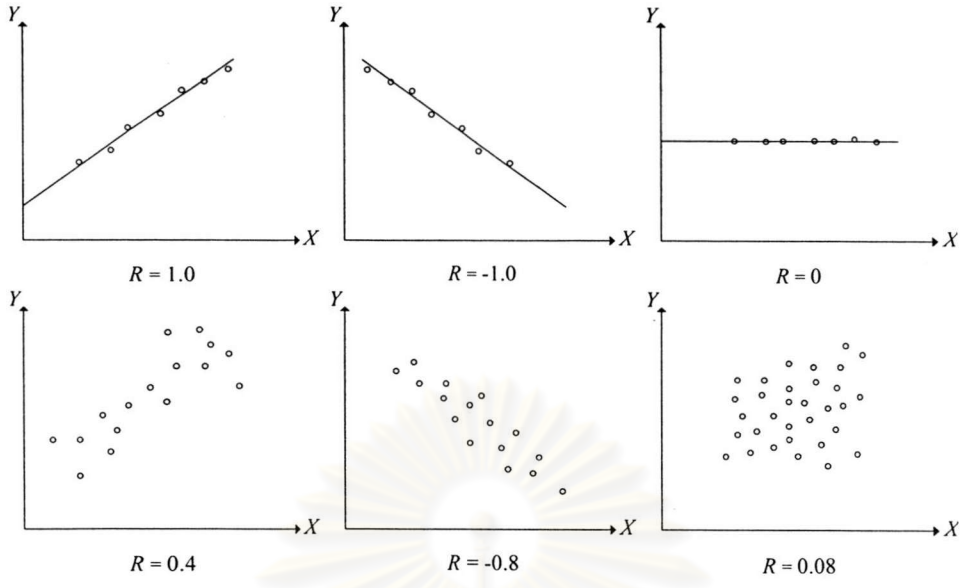
สัมประสิทธิ์สหสัมพันธ์ (ρ) เป็นค่าที่ใช้วัดความสัมพันธ์ระหว่าง X และ Y โดยที่ ρ จะไม่มีหน่วยและจะมีค่าสูงสุดเป็น 1 และต่ำสุดเป็น -1 เนื่องจากเราใช้ข้อมูลตัวอย่างจึงประมาณค่า ρ ด้วยค่า R โดยที่ R คือ สัมประสิทธิ์สหสัมพันธ์ตัวอย่าง

$$R = \sqrt{R^2} \quad (3.18)$$

นอกจากนั้น R และ b จะมีเครื่องหมายเดียวกัน คือ เป็นบวกเหมือนกันหรือเป็นลบเหมือนกัน เนื่องจากทั้ง R และ b เป็นค่าที่แสดงถึงความสัมพันธ์ระหว่าง X และ Y

ความหมายของค่า R

- ค่า R เป็นลบ แสดงว่า X และ Y มีความสัมพันธ์ในทิศทางตรงข้าม คือถ้า X เพิ่ม Y จะลด แต่ถ้า X ลด Y จะเพิ่ม
- ค่า R เป็นบวก แสดงว่า X และ Y มีความสัมพันธ์ในทิศทางเดียวกัน คือถ้า X เพิ่ม Y จะเพิ่มด้วย แต่ถ้า x ลด Y จะลดลงด้วย
- ถ้า R มีค่าเข้าใกล้ 1 หมายถึง X และ Y สัมพันธ์ในทิศทางเดียวกันและมีความสัมพันธ์กันมาก
- ถ้า R มีค่าเข้าใกล้ -1 หมายถึง X และ Y สัมพันธ์ในทิศทางตรงกันข้ามและมีความสัมพันธ์กันมาก
- ถ้า $R = 0$ แสดงว่า X และ Y ไม่มีความสัมพันธ์กัน
- ถ้า R มีค่าเข้าใกล้ 0 แสดงว่า X และ Y มีความสัมพันธ์กันน้อย



รูปที่ 3.4 แสดงค่าของ R ที่มีค่า $-1 < R < 1$

3.1.3 การวิเคราะห์ความถดถอยเชิงซ้อน (Multiple Regression Analysis)

ก) รูปแบบของสมการความถดถอยเชิงซ้อน

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

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + e \qquad (3.19)$$

โดยที่ β_0 คือ ส่วนตัดแกน Y เมื่อกำหนดให้ $X_1 = X_2 = ... = X_k = 0$

$\beta_1, \beta_2, ..., \beta_k$ เป็นสัมประสิทธิ์ความถดถอยเชิงส่วน (Partial Regression Coefficient) โดยที่ β_i เป็นค่าที่แสดงถึงการเปลี่ยนแปลงของตัวแปรตาม Y เมื่อตัวแปรอิสระ X_i เปลี่ยนไป 1 หน่วย โดยที่ตัวแปรอิสระ X ตัวอื่นๆมีค่าคงที่

ก) สมมุติฐานของการวิเคราะห์ความถดถอยเชิงซ้อน

สมมุติฐานของการวิเคราะห์ความถดถอยเชิงซ้อนจะเหมือนกับสมมุติฐานของการวิเคราะห์ความถดถอยเชิงเส้นอย่างง่าย โดยที่สมการความถดถอยเชิงซ้อนเป็น

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

1. ความคลาดเคลื่อน e เป็นตัวแปรที่มีการแจกแจงปกติ
2. ค่าเฉลี่ยของความคลาดเคลื่อนเป็นศูนย์ นั่นคือ $E(e) = 0$
3. ค่าแปรปรวนของความคลาดเคลื่อนเป็นค่าคงที่ที่ไม่ทราบค่า $V(e) = \sigma_e^2$
4. e_i และ e_j เป็นอิสระต่อกัน ; $i \neq j$ นั่นคือ $\text{cov}(e_i, e_j) = 0$

ข) การประมาณค่าพารามิเตอร์ของสมการความถดถอยเชิงซ้อน

จากสมการความถดถอยเชิงซ้อนซึ่งมีพารามิเตอร์ k+1 ตัวคือ $\beta_0, \beta_1, \beta_2, ..., \beta_k$ การประมาณค่า $\beta_0, \beta_1, \beta_2, ..., \beta_k$ จะต้องใช้ข้อมูลตัวอย่างของตัวแปร $Y, X_1, X_2, ..., X_k$ โดยใช้ตัวอย่างขนาด n จากสมการความถดถอยเชิงซ้อน

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + e_i \qquad (3.20)$$

จะประมาณค่า Y หรือประมาณสมการ (3.20) ด้วยสมการที่ (3.21)

$$\hat{Y}_i = \widehat{\beta_0} + \widehat{\beta_1} X_{1i} + \widehat{\beta_2} X_{2i} + ... + \widehat{\beta_k} X_{ki} \qquad (3.21)$$

หรือ

$$\hat{Y}_i = a + b_1 X_{1i} + b_2 X_{2i} + ... + b_k X_{ki} \qquad (3.22)$$

โดยที่ $\widehat{\beta_0} = a, \widehat{\beta_1} = b_1, \widehat{\beta_2} = b_2, ..., \widehat{\beta_k} = b_k$

ดังนั้นค่าคลาดเคลื่อนในการประมาณค่า Y_i ด้วย $\hat{Y}_i$ คือ $Y_i - \hat{Y}_i = e_i$

การประมาณค่า $\beta_0, \beta_1, \beta_2, ..., \beta_k$ ด้วยค่า $a, b_1, b_2, ..., b_k$ ตามลำดับนั้นยังคงมีเป้าหมายเหมือนกับความถดถอยเชิงเส้นอย่างง่าย คือ เพื่อทำให้ผลบวกของค่าคลาดเคลื่อนยกกำลังสองมีค่าน้อยที่สุด โดยใช้วิธีกำลังสองน้อยที่สุด นั่นคือหาค่า $a, b_1, b_2, ..., b_k$ ที่ทำให้ $\sum_{i=1}^{n} e_i^2 = \sum \left(Y_i - \hat{Y}_i \right)^2$ มีค่าต่ำสุด

ในกรณีที่มีตัวแปรอิสระ 2 ตัว (X_1, X_2) ที่มีความสัมพันธ์กับ Y สมการถดถอยคือ

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i \tag{3.23}$$

ค่าประมาณของ Y_i คือ

$$\widehat{Y}_i = a + b_1 X_{1i} + b_2 X_{2i} \tag{3.24}$$

และความคลาดเคลื่อน $e_i = Y_i - \widehat{Y}_i$

ต้องการ $min \sum_{i=1}^{n} e_i^2 = min \sum \left(Y_i - \hat{Y}_i \right)^2$ จึงใช้อนุพันธ์เชิงส่วนเทียบกับ a , b_1 และ b_2 แล้วให้เท่ากับศูนย์ดังนี้

$$\frac{d}{da}\left[\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2\right] = \frac{d}{da}\left[\sum_{i=1}^{n}\left(Y_i - a - b_i X_{1i} - b_2 X_{2i}\right)^2\right]$$

$$= -2\sum_{i=1}^{n}\left(Y_i - a - b_i X_{1i} - b_2 X_{2i}\right)$$

$$= 0$$

$$\Rightarrow \quad -2\sum_{i=1}^{n} Y_i + 2na + 2b_1\sum_{i=1}^{n} X_{1i} + 2b_2\sum_{i=1}^{n} X_{2i} = 0$$

$$na + b_1\sum_{i=1}^{n} X_{1i} + b_2\sum_{i=1}^{n} X_{2i} = \sum_{i=1}^{n} Y_i \tag{3.25}$$

$$\frac{d}{db_1}\left[\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2\right] = \frac{d}{db_1}\left[\sum_{i=1}^{n}\left(Y_i - a - b_i X_{1i} - b_2 X_{2i}\right)^2\right]$$

$$= -2\sum_{i=1}^{n}\left(Y_i - a - b_i X_{1i} - b_2 X_{2i}\right)X_{1i}$$

$$= 0$$

$$\Rightarrow \quad -2\sum_{i=1}^{n} X_{1i}Y_i + 2a\sum X_{1i} + 2b_1\sum_{i=1}^{n} X_{1i}^2 + 2b_2\sum_{i=1}^{n} X_{1i}X_{2i} = 0$$

$$a\sum X_{1i} + b_1\sum X_{1i}^2 + b_2\sum X_{1i}X_{2i} = \sum X_{1i}Y_i \tag{3.26}$$

$$\frac{d}{db_2}\left[\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)^2\right]=\frac{d}{db_2}\left[\sum_{i=1}^{n}\left(Y_i-a-b_iX_{1i}-b_2X_{2i}\right)^2\right]$$

$$\Rightarrow \quad -2\sum X_{2i}Y_i+2a\sum X_{2i}+2b_1\sum X_{1i}X_{2i}+2b_2\sum X_{2i}^2=0$$

$$a\sum X_{2i}+b_1\sum X_{1i}X_{2i}+b_2\sum X_{2i}^2=\sum X_{2i}Y_i \qquad (3.27)$$

เรียกสมการ (3.25) , (3.26) และ (3.27) ว่าชุดของสมการปกติ

การคำนวณหาค่า a , b_1 และ b_2 ทำได้ดังนี้

1. การหาค่า a , b_1 และ b_2 จากชุดสมการปกติ 3 สมการ

ทำได้โดยการแทนค่า $\sum X_{1i},\sum X_{2i},\sum X_{1i}^2,\sum X_{2i}^2,\sum X_{1i}X_{2i}$ ในชุดสมการปกติแล้วจึงหาค่า a , b_1 และ b_2 ได้โดยการแก้สมการปกติ 3 สมการ

2. ใช้เมตริกซ์

ในกรณี $k \geq 2$ การคำนวณหาค่า $a,b_1,b_2,....,b_k$ จากชุดสมการปกติ $k+1$ สมการสมการอาจยุ่งยาก ดังนั้นอาจเขียนชุดสมการปกติในรูปของเมตริกซ์ ในกรณี $k=2$ ได้ดังนี้

$$\begin{bmatrix} n & \sum X_1 & \sum X_2 \\ \sum X_1 & \sum X_1^2 & \sum X_1X_2 \\ \sum X_2 & \sum X_1X_2 & \sum X_2^2 \end{bmatrix}\begin{bmatrix} a \\ b_1 \\ b_2 \end{bmatrix}=\begin{bmatrix} \sum Y \\ \sum X_1Y \\ \sum X_2Y \end{bmatrix} \qquad (3.28)$$

หรือใช้สัญลักษณ์ $$X'X\underline{b}=X'\underline{Y} \qquad (3.29)$$

โดยที่

$$X=\begin{bmatrix} 1 & X_{11} & X_{21} \\ 1 & X_{12} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} \end{bmatrix} \qquad \underline{b}=\begin{bmatrix} a \\ b_1 \\ b_2 \end{bmatrix}$$

$$\underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad X'X = \begin{bmatrix} n & \sum X_1 & \sum X_2 \\ \sum X_1 & \sum X_1^2 & \sum X_1 X_2 \\ \sum X_2 & \sum X_1 X_2 & \sum X_2^2 \end{bmatrix}_{(k+1)\times(k+1)}$$

$$X'\underline{Y} = \begin{bmatrix} \sum Y \\ \sum X_1 Y \\ \sum X_2 Y \end{bmatrix}_{(k+1)\times 1}$$

$$\Rightarrow \qquad b = \begin{bmatrix} a \\ b_1 \\ b_2 \end{bmatrix} = (X'X)^{-1} X'\underline{Y}$$

ค) สัมประสิทธิ์การตัดสินใจเชิงซ้อน (Multiple Coefficients of Determination: R^2)

สัมประสิทธิ์สหสัมพันธ์เชิงซ้อนเป็นสัดส่วนหรือเปอร์เซ็นต์ที่ตัวแปรอิสระ $(X_1, X_2, ..., X_k)$ สามารถอธิบายการเปลี่ยนแปลงของ Y ได้หรือกล่าวได้ว่าสัมประสิทธิ์สหสัมพันธ์เชิงซ้อนเป็นสัดส่วนหรือเปอร์เซ็นต์ของความผันแปร Y ที่มีสาเหตุเนื่องจากความผันแปรของ $X_1, X_2, ...$ และ X_k โดยที่สัมประสิทธิ์สหสัมพันธ์เชิงซ้อนจะใช้สัญลักษณ์ $R^2_{Y.123...k}$ แต่โดยทั่วไปจะใช้ R^2

$$R^2 = \frac{\text{ความแปรปรวนของ } Y \text{ ที่เนื่องจากอิทธิพลของ } X_1, X_2, ..., X_k}{\text{ความแปรปรวนของ } Y \text{ ทั้งหมด}}$$

$$= \frac{SSR}{SST}$$

โดยที่ $0 \le R^2$

ถ้าค่า R^2 ใกล้ 1 จะหมายถึง $X_1, X_2, ..., X_k$ มีความสัมพันธ์กับ Y มาก แต่ถ้า R^2 เข้าใกล้ศูนย์ หมายถึง ค่า $X_1, X_2, ..., X_k$ มีความสัมพันธ์กับ Y น้อย

เนื่องจาก SSR จะเพิ่มขึ้นถ้าเพิ่มตัวแปรอิสระ เช่น เดิมมี X_1 และ X_2 ที่มีความสัมพันธ์กับ Y แต่ถ้าเพิ่มตัวแปรอิสระ X_3 เข้าในสมการถดถอยจะได้ว่า

$$SSR(X_1, X_2, X_3) > SSR(X_1, X_2)$$

โดยที่ $SSR(X_1, X_2, X_3)$ หมายถึง SSR ของสมการความถดถอยที่มีตัวแปรอิสระ X_1 , X_2 และ X_3

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e$$

และ $SSR(X_1, X_2)$ หมายถึง SSR ของสมการความถดถอยที่มีตัวแปรอิสระ X_1 และ X_2

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$

ดังนั้นเมื่อเพิ่มตัวแปรอิสระเข้าสมการความถดถอยจะทำให้ค่า R^2 มากขึ้นทั้งที่ตัวแปรอิสระ X ที่เพิ่มอาจจะไม่มีความสัมพันธ์กับ Y เลยก็ได้ จึงมีการปรับค่า R^2 ให้ถูกต้องขึ้น เรียกว่า $Adjusted\ R^2$ โดยที่

$$R_a^2 = Adjusted\ \ R^2$$

$$= 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} \qquad (3.30)$$

$$= 1 + \frac{(n-1)}{(n-k-1)}\left(R^2 - 1\right)$$

โดยที่

n = จำนวนตัวอย่าง

k = จำนวนตัวแปรอิสระ

ง) สัมประสิทธิ์สหสัมพันธ์เชิงซ้อน (Multiple Coefficient of Correlation)

ค่าของสัมประสิทธิ์สหสัมพันธ์เชิงซ้อนได้จากการถอดรากที่สองของสัมประสิทธิ์สหสัมพันธ์เชิงซ้อน

$$R_{Y.12...k} = R = \sqrt{R^2_{Y.12...k}}$$

โดยที่ $0 \le R \le 1$

สัมประสิทธิ์สหสัมพันธ์เชิงซ้อนแสดงถึงความสัมพันธ์ระหว่าง Y กับ $X_1, X_2, ..., X_k$ ดังนี้

- R มีค่าเข้าใกล้ศูนย์ แสดงว่า Y มีความสัมพันธ์กับ $X_1, X_2, ..., X_k$ น้อยมาก และถ้า $R = 0$ แสดงว่า Y ไม่มีความสัมพันธ์กับ $X_1, X_2, ..., X_k$ เลย
- R มีค่าเข้าใกล้ 1 แสดงว่า Y มีความสัมพันธ์กับตัวแปรอิสระทั้ง k ตัวมาก

จ) สัมประสิทธิ์สหสัมพันธ์เชิงส่วน (Coefficient of Partial Correlation)

สัมประสิทธิ์สหสัมพันธ์เชิงส่วนเป็นค่าที่แสดงความสัมพันธ์ระหว่าง Y กับ X ตัวใดตัวหนึ่งโดยให้ X ตัวอื่นๆมีค่าคงที่ เช่น ถ้า Y มีความสัมพันธ์กับตัวแปรอิสระ 3 ตัว (X_1, X_2, X_3)

สัมประสิทธิ์สหสัมพันธ์เชิงส่วนระหว่าง Y กับ X_i โดยกำหนดให้ X_j และ X_k คงที่ $(i \neq j \neq k)$ จะแสดงความสัมพันธ์ระหว่าง Y กับ X_i จริงๆ โดยกำจัดอิทธิพลของ X_j และ X_k ที่มีต่อ Y

สัญลักษณ์ของสัมประสิทธิ์สหสัมพันธ์เชิงส่วนที่ใช้คือ

$R_{Y1.23}$ = สัมประสิทธิ์สหสัมพันธ์เชิงส่วนระหว่าง Y กับ X_1 โดยกำหนดให้ X_2 และ X_3 คงที่ เป็นค่าที่แสดงความสัมพันธ์ระหว่าง Y กับ X_1 โดยให้ X_2 และ X_3 มีค่าคงที่ จึงเป็นค่าที่แสดงความสัมพันธ์ระหว่าง Y กับ X_1 เท่านั้น มิใช่ความสัมพันธ์ของ X_2 และ X_3 กับ Y

$R_{Y2.13}$ = สัมประสิทธิ์สหสัมพันธ์เชิงส่วนระหว่าง Y กับ X_2 โดยกำหนดให้ X_1 และ X_3 คงที่

$R_{Y3.12}$ = สัมประสิทธิ์สหสัมพันธ์เชิงส่วนระหว่าง Y กับ X_3 โดยกำหนดให้ X_1 และ X_2 คงที่

เมื่อ $-1 \leq R_{Yi;jk} \leq 1$

การคำนวณหาสัมประสิทธิ์สหสัมพันธ์เชิงส่วนทำได้โดยการใช้สัมประสิทธิ์สหสัมพันธ์อย่างง่ายซึ่งเป็นค่าที่แสดงความสัมพันธ์ระหว่างตัวแปร 2 ตัวดังที่ได้กล่าวไปแล้วในหัวข้อการวิเคราะห์ความถดถอยเชิงเส้นอย่างง่าย

- ถ้ามีตัวแปรอิสระ 2 ตัว คือ X_1 และ X_2 ซึ่งมีความสัมพันธ์กับตัวแปรตาม Y สูตรสำหรับการคำนวณหาค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรคู่ใดคู่หนึ่งเป็นดังนี้

$$R_{12} = \frac{\sum\left(X_1 - \overline{X_1}\right)\left(X_2 - \overline{X_2}\right)}{\sqrt{\sum\left(X_1 - \overline{X_1}\right)^2 \sum\left(X_2 - \overline{X_2}\right)^2}} = \frac{\sum x_1 x_2}{\sqrt{\left(\sum x_1^2\right)\left(\sum x_2^2\right)}}$$

$$R_{Y1} = \frac{\sum\left(Y - \overline{Y}\right)\left(X_1 - \overline{X_1}\right)}{\sqrt{\sum\left(Y - \overline{Y}\right)^2 \sum\left(X_1 - \overline{X_1}\right)^2}} = \frac{\sum y x_1}{\sqrt{\left(\sum y^2\right)\left(\sum x_1^2\right)}}$$

$$R_{Y2} = \frac{\sum\left(Y - \overline{Y}\right)\left(X_2 - \overline{X_2}\right)}{\sqrt{\sum\left(Y - \overline{Y}\right)^2 \sum\left(X_2 - \overline{X_2}\right)^2}} = \frac{\sum y x_2}{\sqrt{\left(\sum y^2\right)\left(\sum x_2^2\right)}}$$

โดยที่ $x_i = X_i - \overline{X_i} \quad ; i = 1,2$

$$y_i = Y_i - \overline{Y_i}$$

ดังนั้นจึงสามารถคำนวณหาสัมประสิทธิ์สหสัมพันธ์เชิงส่วนได้ดังนี้

$R_{Y1.2}$ = สัมประสิทธิ์สหสัมพันธ์เชิงส่วนระหว่าง Y กับ X_1 เมื่อกำหนดให้ X_2 คงที่

$$= \frac{R_{Y1} - R_{Y2} \cdot R_{12}}{\sqrt{\left(1 - R_{Y2}^2\right)\left(1 - R_{12}^2\right)}}$$

$R_{Y2.1}$ = สัมประสิทธิ์สหสัมพันธ์เชิงส่วนระหว่าง Y กับ X_2 เมื่อกำหนดให้ X_1 คงที่

$$= \frac{R_{Y2} - R_{Y1} \cdot R_{12}}{\sqrt{\left(1 - R_{Y1}^2\right)\left(1 - R_{12}^2\right)}}$$

- กรณีที่มีมีตัวแปรอิสระ 3 ตัว คือ X_1 , X_2 และ X_3 ซึ่งมีความสัมพันธ์กับตัวแปรตาม Y การคำนวณหาสัมประสิทธิ์สหสัมพันธ์ระหว่าง Y กับ X_1 , Y กับ X_2 , Y กับ X_3 , X_1 กับ X_2 , X_1 กับ X_3 และ X_2 กับ X_3 ทำได้โดยใช้สูตรสหสัมพันธ์อย่างง่าย

ส่วนการคำนวณหาสัมประสิทธิ์สหสัมพันธ์เชิงส่วน $R_{Y1.23}, R_{Y2.13}, R_{Y3.12}$ ทำได้ดังนี้

$$R_{Y1.23} = \frac{R_{Y1.3} - R_{12.3} \cdot R_{Y2.3}}{\sqrt{\left(1 - R_{12.3}^2\right)\left(1 - R_{Y2.3}^2\right)}}$$

$$R_{Y2.13} = \frac{R_{Y2.3} - R_{12.3} \cdot R_{Y1.3}}{\sqrt{\left(1 - R_{12.3}^2\right)\left(1 - R_{Y1.3}^2\right)}}$$

$$R_{Y3.12} = \frac{R_{Y3.2} - R_{31.2} \cdot R_{Y1.2}}{\sqrt{\left(1 - R_{31.2}^2\right)\left(1 - R_{Y1.2}^2\right)}}$$

ฉ) Standardized Regression Coefficient (β)

ค่าสัมประสิทธิ์ความถดถอย b_i ของตัวแปรอิสระที่ i หมายถึง ค่า Y ที่เปลี่ยนแปลงไปเมื่อ X_i เปลี่ยนไป 1 หน่วย กรณีที่มีตัวแปรอิสระหลายตัว (k ตัว) การเปรียบเทียบอิทธิพลหรือความสำคัญของ X_i ที่มีต่อ Y จะสามารถนำค่า b_i ของตัวแปรอิสระแต่ละตัวมาเปรียบเทียบกันได้ถ้าตัวแปรอิสระทุกตัวมีหน่วยเหมือนกัน เช่น หน่วยของตัวแปรอิสระทุกตัวเป็นเมตร แต่จะพบว่าในทางปฏิบัติตัวแปรมักจะมีหน่วยแตกต่างกัน การที่ X มีหน่วยต่างๆกันจะมีผลต่อค่า b ด้วย ถ้าต้องการนำค่า b_i มาเปรียบเทียบกันจะต้องทำให้อยู่ในรูปมาตรฐานเสียก่อน นั่นคือทำให้ b ไม่มีหน่วย ค่า β_i เป็นค่ามาตรฐานของ b_i ซึ่งไม่มีหน่วยจึงสามารถนำมาเปรียบเทียบกันได้

$$\beta_i = b_i \frac{S_i}{S_Y} \tag{3.31}$$

โดยที่ b_i = สัมประสิทธิ์ความถดถอยของตัวแปรอิสระที่ i ; $i = 1,2,....,k$

S_i = ค่าเบี่ยงเบนมาตรฐานของตัวแปรอิสระที่ i (X_i)

S_Y = ค่าเบี่ยงเบนมาตรฐานของตัวแปรตาม Y

ถ้าตัวแปรอิสระตัวใดมีค่า β มาก (อาจเป็นบวกหรือลบก็ได้) แสดงว่า ตัวแปรอิสระนั้นมีความสัมพันธ์กับตัวแปรตามมาก

3.2 การตรวจสอบสมมุติฐานของการวิเคราะห์ความถดถอย

การที่จะพยากรณ์ค่า Y โดยใช้สมการถดถอย โดยกำหนดค่า X ได้นั้น ค่าคลาดเคลื่อน (e) จะต้องมีคุณสมบัติตามสมมุติฐานดังที่ได้ที่กล่าวในข้างต้น ถ้าไม่เป็นไปตามสมมุติฐานที่กล่าวไว้ก็ไม่สามารถนำสมการความถดถอยไปพยากรณ์ค่า Y ได้ ดังนั้นก่อนที่จะประมาณหรือพยากรณ์ค่า Y ควรตรวจสอบคุณสมบัติทั้ง 4 ข้อ ดังนี้

3.2.1 ค่า X ต้องเป็นค่าที่กำหนดไว้ล่วงหน้า ในกรณีนี้สามารถทำได้โดยผู้วิเคราะห์จะต้องเลือกตัวแปรอิสระ X ที่ทราบค่าหรือกำหนดค่าได้

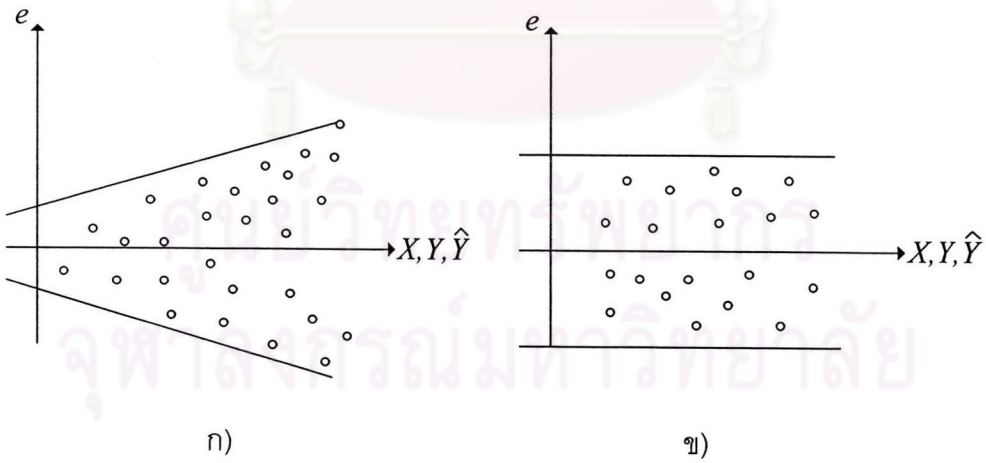
3.2.2 e เป็นตัวแปรที่มีค่าเฉลี่ยเป็นศูนย์ $E(e)=0$ และ $V(e)=\sigma^2$

ก) การตรวจสอบว่า $E(e)=0$

เนื่องจากเราใช้วิธีกำลังสองน้อยที่สุดในการประมาณ β_0 ด้วย a และ β_i ด้วย b_i ซึ่งจะทำให้ $\sum e_i = 0$ ดังนั้นจึงทำให้ $E(e)=0$ เพราะ $E(e)=\dfrac{\sum e_i}{n}$

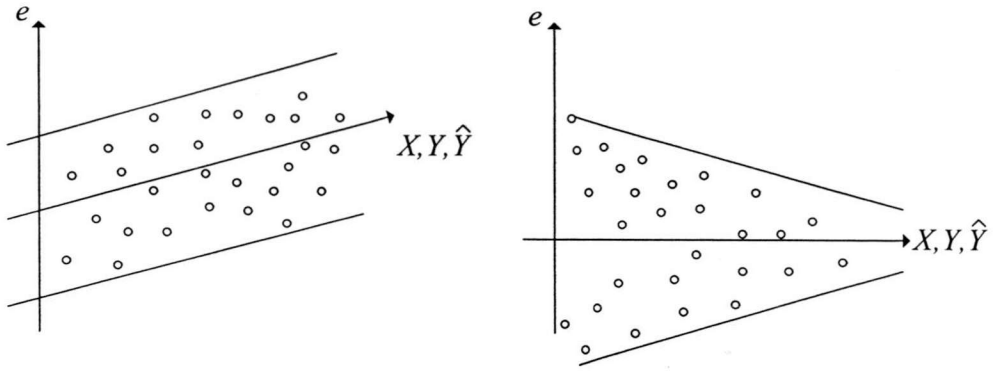
ข) การตรวจสอบว่า $V(e)=\sigma^2$

การตรวจสอบว่า $V(e)=V(y)=\sigma^2=$ ค่าคงที่ จะทำโดยการเขียนกราฟแสดงความสัมพันธ์ระหว่าง e กับ $X,Y,\hat{Y}$ ถ้า $V(e)$ ไม่เท่ากับค่าคงที่จะเรียกว่าเกิดปัญหา Heterocedastic ดังแสดงในรูปที่ 3.5(ก) แต่ถ้า $V(e)$ = ค่าคงที่ ตามสมมุติฐานที่ตั้งไว้ จะเรียกว่า Homoscedastic ดังแสดงในรูปที่ 3.5(ข) , 3.5(ค) และ 3.5(ง)



ก) ข)

รูปที่ 3.5 กราฟความสัมพันธ์ระหว่าง e กับ $\hat{Y}$



ค) ง)

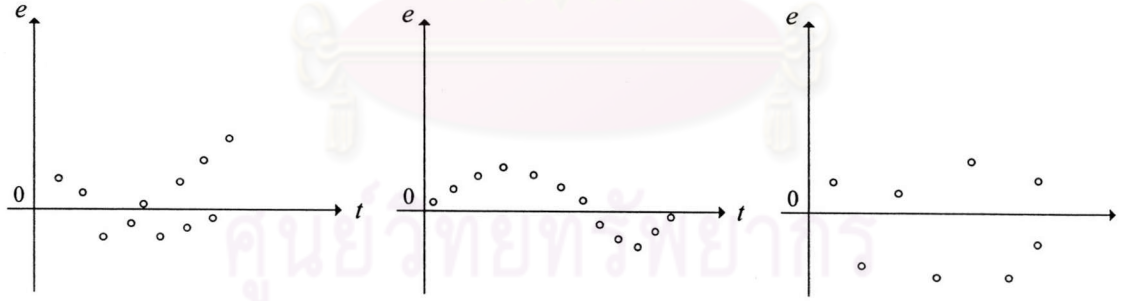
รูปที่ 3.5 (ต่อ) กราฟความสัมพันธ์ระหว่าง e กับ $\hat{Y}$

จากรูปที่ 3.5(ก) จะพบว่าค่า σ_e^2 จะมีค่าน้อยเมื่อ $\hat{Y}$ มีค่าน้อย และเมื่อ $\hat{Y}$ มีค่ามาก σ_e^2 จะมีค่ามากด้วย ในขณะที่รูปที่ 3.5(ข) และ 3.5(ค) ค่า σ_e^2 จะมีค่าคงที่เมื่อ $\hat{Y}$ เปลี่ยนไป ส่วนรูปที่ 3.5(ง) ค่า σ_e^2 จะมีค่าน้อยเมื่อ $\hat{Y}$ มีค่ามาก

3.2.3 การตรวจสอบ e_i และ e_j เป็นอิสระกัน

การตรวจสอบความเป็นอิสระกัน e_i และ e_j โดยที่ $e_i = Y_i - \hat{Y}_i$ และ $e_j = Y_j - \hat{Y}_j$ ทำได้ 2 วิธี คือ

ก) โดยการเขียนกราฟแสดงความสัมพันธ์ระหว่าง e_t กับ t



ก) ข) ค)

รูปที่ 3.6 กราฟแสดงความสัมพันธ์ระหว่าง e_t กับ t

ถ้า e_i และ e_j มีความสัมพันธ์จะเรียกว่าเกิด Autocorrelation ดังแสดงในรูปที่ 3.6(ข) และ 3.6(ค) โดยรูปที่ 3.6(ข) แสดงความสัมพันธ์ระหว่าง e_t และ e_{t+1} ในทางบวกซึ่งเรียกว่าเกิด Positive Autocorrelation ส่วนโดยรูปที่ 3.6(ค) แสดงความสัมพันธ์ระหว่าง e_t และ e_{t+1} ในทางลบซึ่งเรียกว่าเกิด Negative Autocorrelation ส่วนในรูปที่ 3.6(ก) แสดงความเป็นอิสระ กันของ e_t และ e_{t+1} แสดงว่าค่าคลาดเคลื่อนเป็นอิสระกัน

ข) ใช้สถิติทดสอบ Durbin – Watson

การทดสอบความเป็นอิสระกันของค่าคลาดเคลื่อนเมื่อใช้การทดสอบของ Durbin – Watson เป็นการทดสอบความสัมพันธ์ของ e_t และ e_{t-1} โดยที่ t เป็นช่วงเวลา

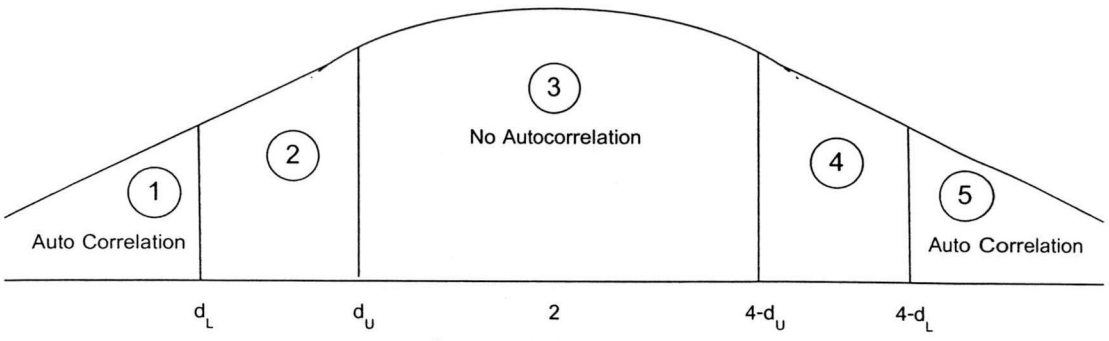
$$\text{สถิติทดสอบ Durbin – Watson} = d = \frac{\sum_{2}^{n}(e_t - e_{t-1})^2}{\sum_{1}^{n} e_t^2} \qquad (3.32)$$

โดยที่ $0 \le d \le 4$ และมีคุณสมบัติดังนี้

- ถ้าค่าคลาดเคลื่อน (e_t) เป็นอิสระกัน ค่า d จะมีค่าใกล้ 2
- ถ้า $d < 2$ จะแสดงถึงความสัมพันธ์ในทางบวกของค่าคลาดเคลื่อน และถ้า $d \approx 0$ ความสัมพันธ์จะมาก
- ถ้า $d > 2$ จะแสดงถึงความสัมพันธ์ในทางลบของค่าคลาดเคลื่อน และถ้า $d \approx 4$ ความสัมพันธ์จะมาก

การสรุปถึงความเป็นอิสระกันของค่าคลาดเคลื่อน จะทำโดยนำสถิติทดสอบ Durbin – Watson (d)เทียบกับค่าที่ได้จากตารางของ Durbin – Watson ซึ่งอยู่ในตารางที่ ค.4 ในภาคผนวก ค ค่าที่ได้จากตาราง คือ ค่า d_U และ d_L ซึ่งขึ้นอยู่กับขนาดตัวอย่าง ***(n)*** จำนวนตัวแปรอิสระ ***(k)*** และระดับนัยสำคัญ (α) จากรูปที่ 3.7 ถ้าเป็นการทดสอบแบบ 2 ข้างจะใช้ $\frac{\alpha}{2}$ แต่ถ้าทดสอบแบบข้างเดียวจะใช้ α การแจกแจงของ Durbin – Watson ซึ่งแบ่งเป็น 5 ช่วง ดังนี้

- น้อยกว่า d_L
- อยู่ระหว่าง d_L และ d_U
- อยู่ระหว่าง d_U และ 4 - d_U
- อยู่ระหว่าง 4 - d_U และ 4 - d_L
- มากกว่า 4 - d_L



รูปที่ 3.7 การแจกแจงของ α

จากรูปที่ 3.7 อธิบายได้ดังนี้

1. ถ้าสถิติทดสอบ d อยู่ในช่วงที่ 1 หรือ 5 แสดงว่าค่าคลาดเคลื่อน e จะมีความสัมพันธ์กัน
2. ถ้า d อยู่ในช่วงที่ 3 แสดงว่าค่าคลาดเคลื่อนไม่มีความสัมพันธ์กัน
3. ถ้า d อยู่ในช่วง 2 หรือ 4 แสดงว่ายังไม่สามารถสรุปได้ว่าค่าคลาดเคลื่อนมีความสัมพันธ์กันหรือไม่

3.2.4 การทดสอบว่า e_t มีการแจกแจงปกติ

ก) Chi-Square Test

สมมุติฐานเพื่อการทดสอบ

H_0 : ลักษณะทั้งสองลักษณะเป็นอิสระกัน

H_1 : ลักษณะทั้งสองลักษณะไม่เป็นอิสระกัน

สถิติทดสอบ :

$$\chi^2 = \frac{\sum\sum\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \tag{3.33}$$

โดยที่ O_{ij} = ความถี่ของข้อมูลที่มีลักษณะที่ 1 ในระดับที่ i และลักษณะที่ 2 ในระดับที่ j

E_{ij} = จำนวนตัวอย่าง/ข้อมูลที่คาดว่าจะอยู่ในระดับที่ i ของลักษณะที่1 และอยู่ในระดับที่ j ของลักษณะที่ 2

เขตปฏิเสธ : จะปฏิเสธ H_0 ถ้า $\chi^2 > \chi^2_{1-\alpha}$ ด้วยองศาอิสระ $(r-1)(c-1)$ เมื่อ r คือ ระดับของตัวแปรที่ 1 และ c ระดับของตัวแปรที่ 2

ข้อจำกัด : - ความถี่ที่คาดไว้ไม่ควรต่ำกว่า 5 นั่นคือ $E_{ij} \geq 5$ ทุกๆค่า i, j

- ถ้า $r = 2, c = 2$ ต้องมีการปรับค่าสถิติทดสอบ χ^2 ดังนี้

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(\left|O_{ij} - E_{ij}\right| - 0.5\right)^2}{E_{ij}} \tag{3.34}$$

ข) Kolmogorov–Smirnov Test

หลักเกณฑ์ของ Kolmogorov–Smirnov Test ในการทดสอบการแจกแจงของประชากรคือ การเปรียบเทียบความน่าจะเป็นสะสมของตัวอย่าง *(S(x))* กับความน่าจะเป็นสะสมภายใต้สมมุติฐานว่าง *(H_0(F(x))* ดังนี้

H_0 : ประชากรมีการแจกแจงตามที่คาดไว้

H_1 : ประชากรไม่มีการแจกแจงตามที่คาดไว้

ถ้าสมมุติฐานว่าง H_0 จริง $S(x)$ และ $F(x)$ จะมีค่าใกล้เคียงกันทุกค่าของ X แต่ถ้า H_0 ไม่จริง คือ ประชากรไม่ได้มีการแจกแจงตามที่คาดไว้ ค่า $S(x)$ และ $F(x)$ จะแตกต่างกันมาก สำหรับบางค่าของ X

โดยที่ $F(X) = P(X \leq x)$

$$= \sum_{0}^{x} P(X = x) \qquad \text{ถ้า } X \text{ เป็นตัวแปรสุ่มไม่ต่อเนื่อง} \tag{3.35}$$

หรือ $$F(X) = \int_{-\infty}^{x} f(x)dx \qquad \text{ถ้า } X \text{ เป็นตัวแปรสุ่มต่อเนื่อง} \tag{3.36}$$

สถิติทดสอบ : $$D = \max\left|F(x) - S(x)\right| \tag{3.37}$$

เขตปฏิเสธ : ถ้า D มีค่ามากแสดงว่า $F(x)$ และ $S(x)$ แตกต่างกันมากจึงปฏิเสธ H_0 นั่นคือถ้า D มากกว่าค่าวิกฤตที่ได้จากตาราง Kolmogorov–Smirnov Test ซึ่งเป็นตารางที่ ค.2 ในภาคผนวก ค ถ้าปฏิเสธสมมุติฐานว่าง H_0 จะสรุปได้ว่าประชากรไม่ได้มีการแจกแจงตามที่คาดไว้ แต่ถ้า D น้อยกว่าค่าวิกฤตที่ได้จากตารางจะต้องยอมรับ H_0 นั่นคือ ประชากรมีการแจกแจงตามที่คาดไว้

ค) Lilliefors Test

Lilliefors Test เป็นการทดสอบการแจกแจงของประชากรว่ามีการแจกแจงแบบปกติหรือไม่ โดยจะต่างจาก Kolmogorov–Smirnov Test คือ Kolmogorov–Smirnov Test จะต้องกำหนดค่าเฉลี่ย $\mu = \mu_0$ และค่าเบี่ยงเบนมาตรฐาน $\sigma = \sigma_0$ ไว้ในสมมุติฐาน H_0 แต่ Lilliefors Test จะไม่กำหนดค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐาน จึงต้องประมาณ μ ด้วย $\overline{X}$ และประมาณ σ ด้วย S

สมมุติฐานเพื่อการทดสอบ

H_0 : ประชากรมีการแจกแจงแบบปกติ

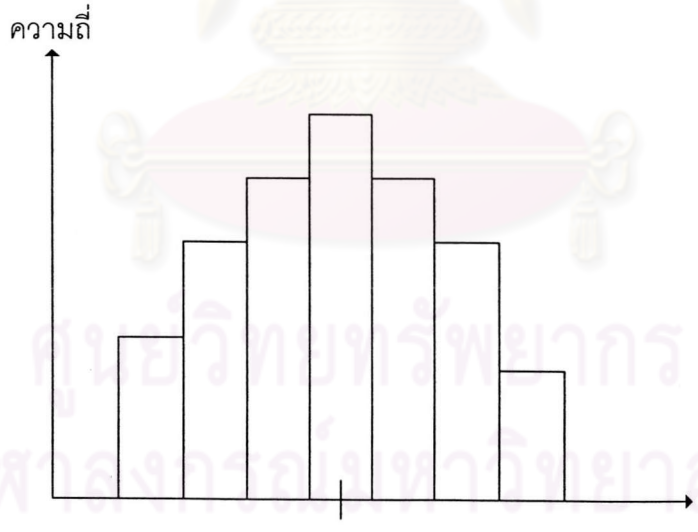
H_1 : ประชากรไม่มีการแจกแจงแบบปกติ

สถิติทดสอบ : $D = \max\left|F(x) - S(x)\right|$

โดยที่ $F(x) = P(X \le x) = P(Z < \dfrac{x - \bar{x}}{s})$ (3.38)

เขตปฏิเสธ : จะปฏิเสธ H_0 ถ้า D มากกว่าค่าวิกฤตที่ได้จากตาราง Lilliefors Test ซึ่งเป็นตารางที่ ค.3 ในภาคผนวก ค

ค) การเขียนฮิสโตแกรม



รูปที่ 3.8 ฮิสโตแกรมแสดงความสัมพันธ์ระหว่าง e_t กับความถี่ที่เกิด

จากรูปที่ 3.8 จะพบว่า e_t จะมีการแจกแจงโดยประมาณแบบปกติ

3.3 ข้อจำกัดในการพยากรณ์ Y เมื่อกำหนดค่า X

3.3.1 ค่า x ที่นำมาใช้ในการพยากรณ์ค่า Y ต้องอยู่ในช่วงของข้อมูลตัวอย่าง

เมื่อกำหนดค่า $X = x$ เราสามารถพยากรณ์ค่า Y ได้โดย

$$\hat{Y} = a + bx$$

แต่การใช้สมการดังกล่าวพยากรณ์ค่า Y ได้นั้นจะต้องให้ค่า X อยู่ในช่วงของข้อมูลตัวอย่างที่นำมาคำนวณหาค่าสัมประสิทธิ์เพื่อสร้างสมการความถดถอย ดังนั้นถ้าค่า X ที่กำหนดให้เพื่อพยากรณ์ค่า Y อยู่นอกช่วงของค่าตัวอย่างของ X ความสัมพันธ์ระหว่าง X และ Y อาจไม่มีความสัมพันธ์ในรูปแบบเดียวกันกับกรณีที่ X อยู่ในช่วงของข้อมูลตัวอย่าง

3.3.2 การเกิดปัญหา Multicollinearity

การวิเคราะห์ความถดถอยเชิงซ้อนซึ่งศึกษาถึงความสัมพันธ์ระหว่างตัวแปรตาม *(Y)* กับตัวแปรอิสระหลายตัว $(X_1, X_2, ..., X_k; k \geq 2)$ นั้น มีข้อกำหนดว่าตัวแปรอิสระเหล่านั้นจะต้องไม่มีความสัมพันธ์กัน แต่ในทางปฏิบัติจะพบว่าตัวแปรอิสระมักจะมีความสัมพันธ์กันเอง การที่ตัวแปร X มีความสัมพันธ์กันจะทำให้เกิดปัญหาที่เรียกว่า Multicollinearity การเกิดปัญหา Multicollinearity จะมากหรือน้อยจะขึ้นอยู่กับความสัมพันธ์ระหว่างตัวแปรอิสระ X ถ้าตัวแปรอิสระมีความสัมพันธ์กันมาก ปัญหา Multicollinearity จะมากด้วย ซึ่งทำให้ผลของการเกิดปัญหา Multicollinearity รุนแรงด้วย

ผลของการเกิดปัญหา Multicollinearity

ก. ทำให้ค่าเบี่ยงเบนมาตรฐานของสัมประสิทธิ์ความถดถอยมีค่าสูงมาก

ข. การที่ตัวแปรอิสระ X มีความสัมพันธ์กันจะทำให้เครื่องหมายของสัมประสิทธิ์ความถดถอย (β, b) ตรงข้ามกับที่ควรจะเป็น

ค. การที่ตัวแปรอิสระมีความสัมพันธ์กันจะทำให้ค่าสัมประสิทธิ์ความถดถอย (β, b) เปลี่ยนแปลงไป (ไม่คงที่) เมื่อมีตัวแปรอิสระเพิ่มขึ้น

<u>การป้องกันการเกิดปัญหา Multicollinearity</u>

1) คำนวณหาสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระ *(X)* ต่างๆ แล้วทำการทดสอบสมมุติฐานว่าสัมประสิทธิ์สหสัมพันธ์ ρ ของ X แต่ละคู่เป็นศูนย์หรือไม่ ถ้าผลการทดสอบยอมรับว่า ρ ของแต่ละคู่เป็นศูนย์แสดงว่าตัวแปรอิสระ *(X)* ต่างๆ ไม่มีความสัมพันธ์กัน แต่จะพบว่าในทางปฏิบัตินั้นการที่จะหาตัวแปร X ที่เป็นอิสระกันทุกคู่เป็นไปได้ยากเนื่องจากตัว

แปร X มักจะมีความสัมพันธ์กัน กรณีที่มี X บางคู่มีความสัมพันธ์กันจะต้องตัด X ตัวใดตัวหนึ่งออกจากสมการถดถอย

2) ใช้วิธี Stepwise ซึ่งเป็นวิธีการเลือกตัวแปรอิสระเข้าสมการถดถอย ซึ่งมีหลักเกณฑ์ว่า จะนำตัวแปรอิสระเข้าสมการความถดถอยครั้งละ 1 ตัว ถ้าตัวแปรอิสระที่นำเข้ามีความสัมพันธ์กับตัวแปรอิสระที่มีอยู่แล้วในสมการถดถอยจะตัดตัวแปรอิสระที่สัมพันธ์กันตัวใดตัวหนึ่งออกจากสมการความถดถอย