บทที่ 2

เอกสารและผลงานวิจัยที่เกี่ยวข้อง

การวิเคราะห์สหสัมพันธ์และการถดถอย

(Simple Correlation and Regression analysis)

การวิเคราะห์สหสัมพันธ์และถดถอยเป็นวิธีการทางสถิติที่ใช้ในการประมาณว่าตัวแปร (Variable) สองตัวหรือมากกว่านั้น มีความเกี่ยวพันใกล้ชิดกันหรือไม่เพียงใด เทคนิคการวิเคราะห์ทั้งสองนี้ มีความใกล้เคียงกันมาก ซึ่งโดยปกติแล้วความสัมพันธ์ระหว่างตัวแปรจะช่วยให้สามารถอธิบายหรือพยากรณ์ล่วงหน้าได้ แต่ในปัจจุบันพบว่าได้มีการเน้นในเรื่องการวิเคราะห์การถดถอยมากกว่า ดังนั้นจึงสามารถแยกให้เห็นความแตกต่างระหว่างเทคนิคการวิเคราะห์ทั้งสองวิธี ดังนี้ (Samuel 1982:193-203)

การวิเคราะห์การถดถอย (Regression analysis) ใช้ในการพิจารณาถึงรูปแบบที่เป็นไปได้ของความสัมพันธ์ระหว่างตัวแปร มีวัตถุประสงค์เพื่อใช้ประโยชน์ในการทำนาย (Predict) หรือประมาณ (Estimate) ค่า ๆ หนึ่งที่สัมพันธ์กับค่าที่กำหนดให้อีกค่าหนึ่ง นักวิทยาศาสตร์ชาวอังกฤษชื่อ Sir Francis Galton เป็นคนแรกที่ศึกษาถึงความสัมพันธ์ระหว่างตัวแปร ซึ่งนำไปสู่ทางคิดค้นเกี่ยวกับการวิเคราะห์การถดถอย โดยการได้ศึกษาถึงแนวโน้มของลักษณะพันธุกรรมที่บุตรหลานสืบทอดจากบิดามารดาในรายงานการวิจัยเกี่ยวกับพันธุกรรมของเขาเมื่อปี ค.ศ. 1899

การวิเคราะห์สหสัมพันธ์ (Correlation analysis) จะเกี่ยวกับการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร การคำนวณค่าสหสัมพันธ์ชุดใดชุดหนึ่งคือการคำนวณดูว่าตัวแปรที่ได้จากข้อมูลชุดนั้นมีความสัมพันธ์กันมากน้อยเพียงใด โดยแนวคิดและศัพท์ต่าง ๆ ในการวิเคราะห์สหสัมพันธ์นี้ได้มาจาก Galton เช่นกัน

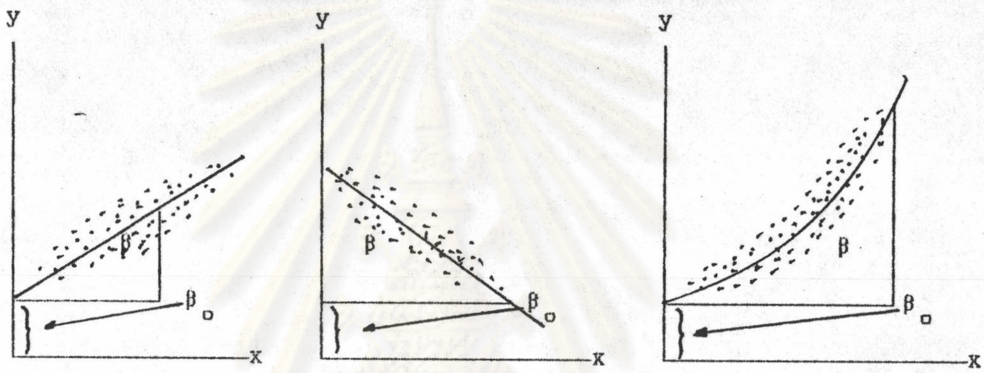
การวิเคราะห์สหสัมพันธ์และการถดถอยนี้ ถ้าเป็นการวิเคราะห์เกี่ยวกับตัวแปรเพียงสองตัว (Bivariate) เรียกว่า สหสัมพันธ์หรือถดถอยอย่างง่าย (Simple Correlation or Regression) ส่วนสหสัมพันธ์หรือการถดถอยพหุคูณ (Multiple Correlation or Regression) หมายถึงการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรตั้งแต่ 3 ตัวขึ้นไป

การวิเคราะห์การถดถอยอย่างง่าย

(Simple Regression Analysis)

ในกรณีที่มีตัวแปรเพียง 2 ตัว คือ Y และ X นั้น สามารถดูลักษณะการถดถอยของ Y ที่มีต่อ X ได้ จากแผนภาพการกระจาย (Scatter diagram) ซึ่งมีลักษณะต่าง ๆ ดังนี้

แผนภาพที่ 1 การถดถอยของตัวแปร X และ Y



ก ข ค

จากแผนภาพที่ 1 (ก,ข) แสดงให้เห็นถึงการถดถอยเชิงเส้นอย่างง่ายของตัวแปรเกณฑ์ Y ที่มีต่อตัวแปรพยากรณ์ X เพียงสองตัว ซึ่งสามารถแสดงความสัมพันธ์ในรูปของตัวแบบสมการทางคณิตศาสตร์ ดังนี้

$$Y_i = \beta_0 + \beta X_i + \varepsilon_i \qquad (2.1)$$

เมื่อ β_0 คือค่าที่เส้นตรงตัดแกน Y และ β คือพารามิเตอร์ที่แสดงความชันของเส้นตรงเรียกว่าสัมประสิทธิ์การถดถอย (Regression Coefficient) ซึ่งเป็นค่าที่แสดงอัตราการเปลี่ยนของค่า Y เมื่อ X เปลี่ยนไป 1 หน่วย โดยจะมีค่ามากกว่า 0 เมื่อ Y มีการถดถอยไปทางเดียวกับ X มีค่าน้อยกว่า 0 เมื่อ Y มีการถดถอยไปทางตรงข้ามกับ X และมีค่าเท่ากับ 0 เมื่อการเปลี่ยนแปลงของค่า Y ไม่ขึ้นอยู่กับการเปลี่ยนแปลงของ X เลย

ส่วน ε_i เป็นค่าความคลาดเคลื่อนที่มีลักษณะเป็นตัวแปรสุ่ม ซึ่งเป็นอิสระจาก X และ Y ภายใต้ข้อสมมติดังนี้

1. ค่า X_i ต้องเป็นค่าที่วัดได้โดยไม่มีความผิดพลาดเลยและเป็นค่าที่กำหนดให้คงที่

2. ε_i มีค่าเฉลี่ยเท่ากับ 0 หรือ ($E(\varepsilon_i) = 0$)

3. ค่าความแปรปรวนของ ε_i มีค่าคงที่และเท่ากับความแปรปรวนของ Y นั่นคือ $V(\varepsilon_i) = V(Y_i) = \sigma^2$ และค่า σ^2 นี้จะเท่ากับ $\sigma^2_{Y.X}$ ซึ่งเป็นค่าความแปรปรวนของ Y เมื่อกำหนดให้ X คงที่ด้วย

4. ε_i และ ε_j เป็นอิสระต่อกัน นั่นคือ $Cov(\varepsilon_i, \varepsilon_j) = 0$ เมื่อ $i \neq j$

Y_i หมายถึง Y ที่ได้จากหน่วยตัวอย่างซึ่งมีค่า $X = X_i$ จึงอาจเขียนค่า Y_i ในรูปของ Y/X_i ได้จากข้อสมมติข้างต้นดังนี้ Y_i จะมีค่าเฉลี่ยดังนี้

$$E(Y_i) = E(Y/X_i) = \mu_{Y.Xi} = \beta_0 + \beta X_i \tag{2.2}$$

นั่นคือ $Y_i = \mu_{Y.Xi} + \varepsilon_i$

ฉะนั้นค่าประมาณของ Y_i จึงหมายถึง $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}X_i = \hat{\mu}_{Yi.Xi}$ นั่นเอง

<u>ค่าประมาณของพารามิเตอร์ β_0 และ β</u>

ในทางปฏิบัติผู้วิจัยจะไม่สามารถทราบค่าพารามิเตอร์ (β_0, β, σ^2) ที่แท้จริงของประชากรได้ แต่จะประมาณได้จากข้อมูลกลุ่มตัวอย่างที่ศึกษา (Y_i, X_i) จำนวน n คู่ ซึ่งจะได้ค่าประมาณของ Y_i ดังนี้

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}X_i = b_0 + bX_i = \hat{\mu}_{Yi.Xi} \tag{2.3}$$

ค่า b_0 และ b นี้จะหาได้โดยวิธีกำลังสองน้อยที่สุด (Least Squares Method) ซึ่งจะกำหนดโดยการหาค่าต่ำสุดของผลรวมของความคลาดเคลื่อน (ε_i) ยกกำลังสอง ($\sum_{i=1}^{n} \varepsilon_i^2$) โดยใช้อนุพันธ์เชิงส่วน (Partial Derivative) ดังนี้

$$\sum_{t=1}^{n} \hat{\varepsilon}_i^2 = \sum_{t=1}^{n} (Y_i - \hat{Y})^2 = \sum_{t=1}^{n} (Y_i - b_0 - bX_i)^2 \tag{2.4}$$

$$\frac{\partial}{\partial \beta_0} \left(\sum_{t=1}^{n} \varepsilon_i^2 \right) = -2 \sum_{t=1}^{n} Y_i + 2an + 2b \sum_{t=1}^{n} X_i \tag{2.5}$$

$$\frac{\partial}{\partial \beta}\left(\sum_{i=1}^{n} \varepsilon_i^2\right) = -2 \sum_{i=1}^{n} X_i Y_i + 2a \sum_{i=1}^{n} + 2b \sum_{i=1}^{n} X_i^2 \tag{2.6}$$

จากผลข้างต้นนี้จะทำให้ได้ ชุดสมการปกติ (Normal equations) ดังนี้ (Lindeman 1980 : 99)

$$nb_0 + b \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \tag{2.7}$$

$$b_0 \sum_{i=1}^{n} X_i + b \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i \tag{2.8}$$

ซึ่งให้ค่าประมาณของ β_0 และ β ดังนี้

$$\hat{\beta}_0 = b_0 = \bar{y} - b\bar{x} \tag{2.9}$$

$$\hat{\beta} = b = \frac{\sum_{i=1}^{n} (X_i - \bar{x})(Y_i - \bar{y})}{\sum_{i=1}^{n} (X_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{x}) Y_i}{\sum_{i=1}^{n} (X_i - \bar{x})^2} \tag{2.10}$$

<u>สหสัมพันธ์อย่างง่าย</u>

(Simple Correlation)

สหสัมพันธ์อย่างง่ายหมายถึงความสัมพันธ์ระหว่างตัวแปรใด ๆ 2 ตัว โดยไม่คำนึงว่าตัวแปรใดเป็นตัวแปรอิสระหรือตัวแปรตาม ความสัมพันธ์ระหว่างตัวแปรสองตัวนี้อาจจะเป็นไปในทางเดียวกันหรือตรงข้ามกันก็ได้ โดยมีค่าที่ใช้วัดระดับความสัมพันธ์เรียกว่า สัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) ซึ่งค่าดังกล่าวเป็นค่าที่ได้จากการเปรียบเทียบความแปรปรวน

ร่วม (Covariance) ระหว่างตัวแปรสองตัว (X,Y) กับผลคูณของความเบี่ยงเบนมาตรฐาน (Standard Deveation) ของ X และ Y ดังนี้

$$\rho = \rho_{XY} = \rho_{YX} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{2.11}$$

$$\rho_{XY} = \frac{\sum_{t=1}^{n} (X_t-\mu_X)(Y_t-\mu_Y)}{\sqrt{\sum_{t=1}^{n} (X_t-\mu_X)^2 \sum_{t=1}^{n} (Y_t-\mu_Y)^2}} \tag{2.12}$$

เมื่อ n คือ จำนวนข้อมูลทั้งหมดในประชากรของ X และ Y และ ρ คือสัมประสิทธิ์สหสัมพันธ์อย่างง่าย มีค่าอยู่ระหว่าง -1 ถึง +1 และเป็นค่าที่ไม่มีหน่วย แต่จะบอกถึงระดับความสัมพันธ์ระหว่างตัวแปรว่ามีมากน้อยเพียงใด ซึ่งสามารถประมาณค่าได้จากข้อมูลกลุ่มตัวอย่างดังนี้

$$\rho = \hat{\rho}_{XY} = r = r_{XY} = r_{YX} \tag{2.13}$$

$$r_{XY} = \frac{\sum_{t=1}^{n} (X_t-\bar{x})(Y_t-\bar{y})}{\sqrt{\sum_{t=1}^{n} (X_t-\bar{x})^2 \sum_{t=1}^{n} (Y_t-\bar{y})^2}} \tag{2.14}$$

<u>การวิเคราะห์สหสัมพันธ์และการถดถอยพหุคูณ</u>

(Multiple correlation and Regression analysis)

การวิเคราะห์สหสัมพันธ์และการถดถอยพหุคูณ เป็นแนวคิดและเทคนิคที่ขยายมาจากการวิเคราะห์สหสัมพันธ์และการถดถอยอย่างง่าย (Simple Correlation and Regression) เป็นการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรพยากรณ์หลาย ๆ ตัวกับตัวแปรเกณฑ์ ซึ่งโดยปกติแล้วจะทำให้สามารถวิเคราะห์และอธิบายตัวแปรเกณฑ์ได้มากกว่า เพราะในการวิเคราะห์ปัญหาบางอย่างที่จำเป็นต้องใช้การถดถอยนั้น บางครั้งการศึกษาการถดถอยอย่างง่ายอาจจะไม่เพียงพอ ทั้งนี้เพราะการประมาณค่าของตัวแปรเกณฑ์เพื่อให้ใกล้เคียงที่สุดนั้น เรามักจะต้องพิจารณาตัวแปรพยากรณ์ที่มีอิทธิพลหรือมีความสัมพันธ์ต่อตัวแปรเกณฑ์มากกว่า 1 ตัวขึ้นไป โดยมีสมการถดถอยเป็นตัวชี้ให้เห็นถึงความสัมพันธ์ถัวเฉลี่ยของตัวแปรเหล่านั้นดังนี้ (Lindeman 1980: 94)

$$Y_i = b_o + b_1X_1 + b_2X_2 + ...b_pX_p + e_i \qquad (2.15)$$

โดยที่ Y_i คือตัวแปรเกณฑ์ (Criterion variable or Dependent valiable)

X_i คือตัวแปรพยากรณ์ (Predictor variable or Independent variable) ; $i = 1, 2,..........p$

b_o คือค่าคงที่ ซึ่งจะมีค่าเท่ากับ Y เมื่อ X_i ทั้งหมดมีค่าเท่ากับศูนย์ ค่า b_o เป็นค่าประมาณของ β_o

b_i คือสัมประสิทธิ์การถดถอยของประชากรบางส่วน (Population Partial Regression Coefficient) ของ X_i เมื่อให้ X_{i+1} ....X_p เป็นค่าคงที่ นั่นคือเมื่อ X_i มีค่าเปลี่ยนแปลงไป 1 หน่วย จะมีผลทำให้ Y เปลี่ยนแปลงไป b_i หน่วย เมื่อตัวแปรพยากรณ์อื่น ๆ คงที่ และเป็นค่าประมาณของ β_i

e_i คือความคลาดเคลื่อนที่แสดงถึงความแตกต่างระหว่างสมการถดถอยกับค่าจริง มีลักษณะเป็นตัวแปรสุ่มที่ไม่ทราบค่า และเป็นค่าประมาณของ ε_i $\varepsilon \sim N(0,\sigma^2 I)$

$$E(\varepsilon_i) = 0 \;;\; i = 1, 2,......n$$

$$E(\varepsilon_i \varepsilon_j) = \sigma^2 \;;\; i = j = 1, 2,..n$$

$$= 0 \;;\; i \neq j$$

จากสมการ (2.2) สามารถแสดงในรูปของเมตริกซ์ได้ดังนี้

$$Y = X\beta + \varepsilon$$

เมื่อ

$$Y = \begin{bmatrix} Y_n \\ Y_n \\ \cdot \\ \cdot \\ Y_n \end{bmatrix}_{n \times 1} \qquad X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots\cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots\cdots & X_{2p} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 1 & X_{n1} & X_{n2} & \cdots\cdots & X_{np} \end{bmatrix}_{n \times (k+1)}$$

$$\beta = \begin{bmatrix} b_0 \\ b_1 \\ \cdot \\ \cdot \\ b_p \end{bmatrix}_{(k+1) \times 1} \qquad \varepsilon = \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ e_n \end{bmatrix}_{n \times 1}$$

จากตัวแบบ (2.2) ผลรวมของความคลาดเคลื่อนกำลังสองจะมีค่าเป็น (วิชิต หล่อจีระชุณห์กุล 2524 : 124)

$$\begin{aligned} \sum_{t=1}^{n} e_t^2 &= \underset{\sim}{\varepsilon}' \underset{\sim}{\varepsilon} = (\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\beta}')(\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\beta}) \\ &= \underset{\sim}{Y}'\underset{\sim}{Y} - \beta'\underset{\sim}{X}'\underset{\sim}{Y} - \underset{\sim}{Y}'\underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{\beta}'\,\underset{\sim}{X}'X\,\beta \\ &= \underset{\sim}{Y}'\underset{\sim}{Y} - 2\underset{\sim}{\beta}'\,\underset{\sim}{X}\,\underset{\sim}{Y} + \underset{\sim}{\beta}'\,\underset{\sim}{X}'X\,\beta \qquad (2.16) \end{aligned}$$

<u>การประมาณค่าพารามิเตอร์ในการวิเคราะห์การถดถอย</u>

(Estimation of Parameters in Multiple Regression)

เนื่องจากการวิจัยในทางปฏิบัตินั้น ผู้วิจัยจะไม่สามารถศึกษาจากกลุ่มประชากรทั้งหมดได้ จึงไม่ทราบค่าพารามิเตอร์ β ที่แท้จริง โดยทั่วไปจะประมาณค่าพารามิเตอร์ (Parameter) จากข้อมูลกลุ่มตัวอย่างที่นำมาศึกษา วิธีการประมาณตัวประมาณค่าที่นิยมใช้กันมากคือวิธีกำลังสองน้อยที่สุด (Least Square Estimation Method) ซึ่งจะได้ค่าประมาณสัมประสิทธิ์การถดถอยพหุคูณ

จากการอนุพันธ์ (Differentiate) สมการ (2.6) เทียบกับ β แล้วได้ค่าเท่ากับศูนย์

$$\frac{\partial(\varepsilon'\varepsilon)}{\partial\beta} = -2\,\underset{\sim}{X}'\underset{\sim}{Y} + 2\,\underset{\sim}{X}'\underset{\sim}{X}\,\underset{\sim}{\beta} = 0$$

$$\underset{\sim}{X}'\underset{\sim}{X}\,\underset{\sim}{\beta} = \underset{\sim}{X}'\underset{\sim}{Y}$$

$$\underset{\sim}{\beta} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\,\underset{\sim}{X}'\underset{\sim}{Y} \qquad (2.17)$$

ซึ่งจะได้ตัวประมาณค่าที่ไม่เอนเอียงของ (β) ซึ่งมีค่าเฉลี่ยความคลาดเคลื่อนกำลังสองน้อยที่สุด ในบรรดาตัวประมาณค่าที่ไม่เอนเอียงทั้งหลาย

การเลือกสมการถดถอยที่ดีที่สุด

การวิจัยที่ใช้วิธีการพยากรณ์นั้น โดยพื้นฐานแล้วนักวิจัยไม่มีจุดมุ่งหมายในการทดสอบสมมติฐานในการเปรียบเทียบว่า ตัวแปรพยากรณ์ตัวใดมีความสำคัญต่อการเปลี่ยนแปลงของตัวแปรเกณฑ์มากกว่ากัน ความสำคัญของการวิจัยประเภทนี้มักจะอยู่ที่การค้นหาตัวแปรพยากรณ์ที่สามารถพยากรณ์ตัวแปรเกณฑ์ที่สนใจได้ถูกต้องแม่นยำที่สุดเท่าที่ความรู้เกี่ยวกับตัวพยากรณ์จะมีอยู่ ดังนั้นหน้าที่สำคัญของนักวิจัยก็คือการค้นหาสมการหรือการประมาณค่าสัมประสิทธิ์การถดถอยของตัวแปรในสมการพยากรณ์ เพื่อให้มีความคลาดเคลื่อนในการพยากรณ์ต่ำสุด

จากสมการ (2.17) ค่าสัมประสิทธิ์การถดถอยที่ได้เป็นค่าแสดงการเปลี่ยนแปลงค่าเฉลี่ยของ Y เมื่อ X_i เปลี่ยนไป 1 หน่วย ขณะที่ตัวแปรพยากรณ์อื่น ๆ คงที่ และค่าสัมประสิทธิ์การถดถอยของคะแนนดิบ (Unstandardized Coefficience) นี้ เป็นค่าซึ่งใช้ในการประมาณค่า Y เท่านั้น ถ้าต้องการเปรียบเทียบความสำคัญของตัวแปรพยากรณ์ที่มีต่อตัวแปรเกณฑ์จะทำได้โดยการแปลงค่าสัมประสิทธิ์การถดถอยคะแนนดิบ (b_i) ให้เป็นสัมประสิทธิ์คะแนนมาตรฐาน (Standardize Coefficience)

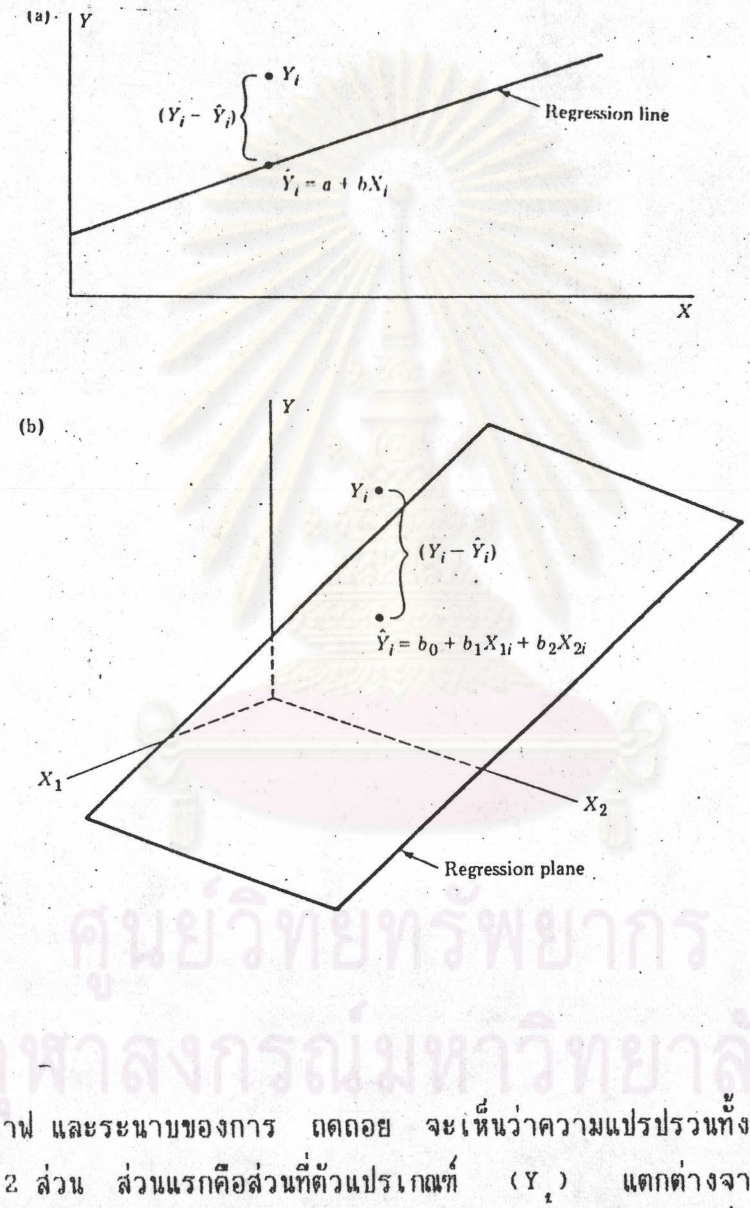
$$\text{โดย} \quad \beta_i = b_i \frac{S_{x_i}}{S_y} \qquad (2.18)$$

เมื่อ β_i = ค่าสัมประสิทธิ์คะแนนมาตรฐาน (Standardized Beta Weight)

S_{x_i} = ส่วนเบี่ยงเบนมาตรฐานของ X_i

S_y = ส่วนเบี่ยงเบนมาตรฐานของ Y

จากข้อมูลกลุ่มตัวอย่าง สามารถพิจารณาลักษณะความแปรปรวนของตัวแปรเกณฑ์ (Y) จากค่าเฉลี่ย Y ได้ ดังรูป (Lindeman 1980 : 100)



ซึ่งแสดงกราฟ และระนาบของการ ถดถอย จะเห็นว่าความแปรปรวนทั้งหมดประกอบด้วยความแปรปรวน 2 ส่วน ส่วนแรกคือส่วนที่ตัวแปรเกณฑ์ (Y_i) แตกต่างจากค่าประมาณที่ได้จากเส้นถดถอย หรือระนาบการถดถอย ($\hat{Y}_i$) เรียกว่า ความแปรปรวนที่ไม่สามารถอธิบายได้ (Unexplained variation) ส่วนที่สองคือส่วนที่ตัวแปรเกณฑ์ที่ประมาณค่าได้จากการประมาณค่าจากเส้นถดถอยหรือระนาบการถดถอย ($\hat{Y}_i$) แตกต่างจากค่าเฉลี่ยของตัวแปรเกณฑ์ ซึ่งเรียกว่า ความแปรปรวนที่สามารถอธิบายได้ (Explained variation) นั่นคือ

$$Y_i-\bar{Y} = (Y_i-\hat{Y}_i) + (\hat{Y}_i-\bar{Y}) \tag{2.19}$$

เมื่อนำ (2.19) มายกกำลังสองจะได้

$$\sum_{t=1}^{n}(Y-\bar{Y})^2 = \sum_{t=1}^{n}(Y-\hat{Y})^2 + \sum_{t=1}^{n}(\hat{Y}-\bar{Y})^2 + 2\sum_{t=1}^{n}(Y-\hat{Y})(\hat{Y}-\bar{Y}) \tag{2.20}$$

$$\sum_{t=1}^{n}(Y-\bar{Y})^2 = \sum_{t=1}^{n}(Y_i-\hat{Y}_i)^2 + \sum_{t=1}^{n}(\hat{Y}_i-\bar{Y})^2 \tag{2.21}$$

ให้ $$SST = \sum_{t=2}^{n}(Y_i-\bar{Y})^2 \tag{2.22}$$

$$SSE = \sum_{t=1}^{n}(Y_i-\hat{Y}_i)^2 \tag{2.23}$$

และ $$SSR = \sum_{t=1}^{n}(\hat{Y}_i-\bar{Y})^2 \tag{2.24}$$

ซึ่งสามารถสรุปเป็นตารางวิเคราะห์ความแปรปรวน (Analysis of Variance) ดังตารางที่ 1

ตารางที่ 1 แสดงแหล่งความแปรปรวนในการวิเคราะห์การถดถอยพหุคูณ

แหล่งความแปรปรวน	ระดับความเป็นอิสระ	ผลบวกกำลังสอง	ผลบวกกำลังสองเฉลี่ย	F
การถดถอย	p	$B'X'Y - n\bar{Y}^2 = SSR$	$\frac{SSR}{p} = MSR$	$\frac{MSR}{MSE}$
ความคลาดเคลื่อน	$n-p-1$	$Y'Y - B'X'Y = SSE$	$\frac{SSE}{n-p-1} = MSE$	
ยอดรวม	$n-1$	$Y'Y - n\bar{Y}^2 = SST$		

ดังนั้นจะได้

$$\hat{Y} = a + b_1X_1 + b_2X_2 + \ldots\ldots + b_pX_p \tag{2.25}$$

นำสัมประสิทธิ์การถดถอย b_1 , b_2 ,......b_p ที่ได้จากสมการ (2.4) มาทดสอบความมีนัยสำคัญ โดยทดสอบสมมติฐาน

$$H_0 : \beta_1 = \beta_2 = \ldots\ldots = \beta_p = 0 \tag{2.26}$$

ค่าสถิติที่ใช้ในการทดสอบคือ

$$F = \frac{MSR}{MSE} \tag{2.27}$$

ดังนั้นเมื่อสร้างสมการพยากรณ์ได้แล้ว ก่อนที่จะมีการนำเอาสมการไปใช้ ต้องคำนึงถึงว่าสมการนั้นน่าเชื่อถือหรือไม่ เกณฑ์อันหนึ่งที่นิยมใช้กันมากในการตัดสินใจเกี่ยวกับการศึกษาเรื่องการถดถอยเชิงเส้นคือสัมประสิทธิ์การตัดสินใจ (Coefficient of detertmination)

$$R^2 = \frac{SSR}{SST} \qquad (2.28)$$

R^2 นี้เรียกว่าสัมประสิทธิ์การตัดสินใจ (Coefficient of Determination) ซึ่งจะมีค่าอยู่ระหว่าง 0 กับ 1

$R^2 \times 100$ หมายถึงร้อยละของความแปรปรวนทั้งหมดของค่าที่สังเกตได้ (Y_i) ที่ถูกอธิบายได้โดยสมการพยากรณ์ หรืออาจกล่าวได้อีกว่า R^2 ก็คือ Goodness of fit ของพื้นผิวของระนาบการถดถอยนั่นเอง

ในการวิเคราะห์การถดถอยพหุคูณ ค่า R^2 ที่สูงขึ้นย่อมเป็นสิ่งที่ต้องการ เพราะนั่นหมายถึงว่าตัวแปรพยากรณ์ (Xs) สามารถใช้พยากรณ์ตัวแปรเกณฑ์ได้ดีขึ้น อย่างไรก็ตามการคัดเลือกสมการพยากรณ์ด้วยวิธีนี้ก็มีข้อบกพร่อง (Herzberg 1967 : 1) เนื่องจากในค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณ (R) ซึ่งเป็นดัชนีที่ชี้ให้เห็นถึงระดับความสัมพันธ์พหุคูณระหว่างตัวแปรเกณฑ์กับผลรวมเชิงเส้นตรงของตัวแปรพยากรณ์นี้จัดเป็นตัวประมาณค่าที่เอนเอียงของพารามิเตอร์ (ρ) (Murihead 1982 : 179) และมักจะมีค่าสูงกว่าความเป็นจริงเสมอ ทำให้เกิดปัญหาการลดลงของค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสอง (Shrinkage) เมื่อนำเอาสมการพยากรณ์ที่สร้างจากกลุ่มตัวอย่างหนึ่งไปใช้กับอีกกลุ่มตัวอย่างหนึ่งที่สุ่มมาจากประชากรเดียวกัน (Pedhazur 1982 : 147-143) เนื่องจากในการคำนวณค่าสัมประสิทธิ์การถดถอย (b) เพื่อให้ได้สมการพยากรณ์ที่มีค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณสูงสุดและมีความคลาดเคลื่อนในการพยากรณ์ต่ำสุดนั้น ถือว่าค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองทุกตัวมีความคลาดเคลื่อนเป็นอิสระต่อกัน (Error Free) ซึ่งในความเป็นจริงไม่ได้เป็นเช่นนั้น จึงทำให้ค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองที่คำนวณได้ครั้งแรกเป็นค่าที่ไม่ถูกต้องตามความเป็นจริงนัก เหตุที่ทำให้ค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองมีค่าสูงกว่าปกติคือ ขนาดของกลุ่มตัวอย่าง ซึ่งพบว่าถ้าหากว่ากลุ่มตัวที่มีขนาดเล็กแล้ว จะทำให้ค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองมีค่าสูงกว่าความเป็นจริงมาก

<u>การแจกแจงแบบปกติหลายตัวแปร</u>

(Multivariate Normal Distribution)

เมื่อ X_{ij} เป็นตัวแปรสุ่ม (random variable)

$$X = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ X_{n1} & \cdots & X_{np} \end{bmatrix}$$

X_{ij} จะมีการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution) เมื่อมี p.d.f. (Probability Density Function) ดังนี้ (Morrison 1967 : 98-99)

$$Fx(X) = \frac{1}{(2\pi)^{p/2}\,|\Sigma|^{1/2}} \exp -1/2\,(X_i-\mu)\,\Sigma^{-1}(X-\mu) \qquad (2.29)$$

โดยที่ $-\infty < X_i < \infty$; $i = 1,2....p$

เขียนได้เป็น $X \sim N(\underset{\sim}{\mu}, \Sigma)$

เมื่อ $\underset{\sim}{\mu} = (\mu_1, \mu_2, ...\mu_p)$

$$\Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \sigma_{n1} & \cdots & \sigma_{np} \end{bmatrix}$$

Σ เป็น pxp positive definite และสมมาตร (Symmetric) ซึ่งคือเมตริกซ์ความแปรปรวนร่วม (Covariance Matrix)

การแจกแจงของค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสอง

(Distribution of the Multiple Correlation Coeeficeint Square)

ค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณ ($\rho_{y.12..p}$) หมายถึงความสัมพันธ์เชิงเส้นระหว่างตัวแปรเกณฑ์ (Y) กับผลรวมเชิงเส้นของตัวแปรพยากรณ์ (Xs) (Lindeman 1982 : 108) หรือกล่าวได้ว่าเป็นสัมประสิทธิ์สหสัมพันธ์โพรดัคโมเมนต์ (Product moment) ของตัวแปรเกณฑ์ที่สังเกตได้กับตัวแปรเกณฑ์ที่พยากรณ์ได้จากสมการถดถอย (Y)

$$\rho_{y.12..p} = \rho_{Y\hat{Y}} = \sqrt{1 - \frac{\sigma^2_{(Y-\hat{Y})}}{\sigma^2_y}} \qquad (2.30)$$

เมื่อศึกษากับกลุ่มตัวอย่างสามารถประมาณค่าโดย

$$R_{y.12..p} = \sqrt{1 - \frac{MSE}{S^2_y}} \qquad (2.31)$$

$$= \sqrt{\beta_1 r_{y1} + \beta_2 r_{y2} + \ldots\ldots + \beta_p r_{yp}} \qquad (2.32)$$

เมื่อ r_{iy} คือสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรเกณฑ์กับตัวแปรพยากรณ์ (X_i) แต่ละตัว และ β คือสัมประสิทธิ์การถดถอยมาตรฐานที่จะทำให้ค่าสหสัมพันธ์พหุคูณมีค่าสูงสุด ซึ่งมีค่าอยู่ระหว่าง 0 ถึง 1 และจะมีค่าเพิ่มขึ้นเมื่อจำนวนตัวแปรพยากรณ์เพิ่มขึ้น

การประมาณค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองของประชากรจะมีลักษณะเช่นเดียวกับในการศึกษาสหสัมพันธ์อย่างง่าย โดยคาดว่าค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองที่คำนวณได้จากกลุ่มตัวอย่าง ($R^2_{y.1.2..p}$) จะกระจายอยู่รอบ ๆ ค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองของประชากร แต่เนื่องจากค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองนี้มีค่าอยู่ระหว่าง 0 ถึง1 และจัด

เป็นตัวประมาณค่าที่เอนเอียง จึงทำให้การแจกแจงมีความเอนเอียงไปทางบวกเสมอ ซึ่งไม่สามารถคาดคะเนลักษณะการแจกแจงที่แน่นอนได้ จากการศึกษาพบว่า ลักษณะการแจกแจงของค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองนี้ จะขึ้นอยู่กับอิทธิพลของขนาดของกลุ่มตัวอย่าง จำนวนตัวแปรพยากรณ์ และขนาดความสัมพันธ์ในประชากร (ρ)(Muirhead 1982: 171) ดังนี้

เมื่อ $\rho = 0$

$$E(R^2) = \frac{p}{n-1} \tag{2.33}$$

และ

$$Var(R^2) = \frac{2(n-p)(p-1)}{(n^2-1)(n-1)} \tag{2.34}$$

จาก (2.33) และ (2.34) เมื่อขนาดความสัมพันธ์ของประชากรมีค่าเป็นศูนย์ หรือไม่มีความสัมพันธ์กันเลยลักษณะการแจกแจงของค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองจะขึ้นอยู่กับขนาดของกลุ่มตัวอย่าง และจำนวนตัวแปรพยากรณ์เท่านั้น เมื่อจำนวนตัวแปรพยากรณ์เพิ่มขึ้นจะทำให้ค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองสูงขึ้นด้วย ($p \rightarrow n : R \rightarrow 1$) แต่ถ้าขนาดความสัมพันธ์ในประชากรไม่เท่ากับศูนย์ ($\rho \neq 0$) แล้ว การคาดคะเนลักษณะการแจกแจงของค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองจะทำได้ยากมากดังนี้

เมื่อ $\rho \neq 0$

$$E(R^2) = 1 - \frac{n-p-1}{n-1}(1-\rho^2)\,F(1,1,(n+1)/2,\rho^2) \tag{2.35}$$

เมื่อ F (a, b, c, x) เป็นฟังก์ชันไฮเปอร์จีออเมตริกซ์ (Hypergeometrix) ซึ่งถ้าใช้เพียงสองเทอมแรกของสมการจะได้

$$E(R^2) = 1 - \frac{n-p-1}{n-1}(1-\rho^2) - \frac{n-p-1}{n-1}\,\frac{2}{n+1}\,\rho^2(1-\rho^2) \tag{2.36}$$

และ

$$\mathrm{Var}(R^2) = \frac{n-p+1}{n^2(n+2)}(1-\rho^2)^2 \left\{ 2(p-1)+4\rho^2 \left[\frac{4(p-1)+n(n+P+1)}{n+4} \right] + O(n^{-2}) \right\} \quad (2.37)$$

ซึ่ง วิชาร์ด (Wishart 1931 : 353-367) ได้ทำการศึกษาและได้เสนอสูตรการคำนวณค่าที่คาดหวัง (Expected) ของค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองไว้ดังนี้

$$E(R^2) = \rho^2 + \frac{(1-\rho^2)(a-\rho^2)}{a+b+1/2}$$

$$= \frac{a+(b-1/2)\rho^2+\rho^4}{a+b+1/2} \quad (2.38)$$

และ

$$\mathrm{Var}(R^2) = \frac{4\rho^2(1-\rho^2)}{n}$$

$$= \frac{2\rho^2(1-\rho^2)^2}{(a+b+1/2)} \quad (2.39)$$

เมื่อ a คือ 1/2 ของชั้นแห่งความเป็นอิสระ (Degree of Freedom) อันเนื่องมาจากฟังก์ชั่นการถดถอย (SSR)

b คือ 1/2 ของชั้นแห่งความเป็นอิสระ (Degree of Freedom) อันเนื่องมาจากความคลาดเคลื่อน (SSE)

จะเห็นว่าลักษณะการแจกแจงของค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองนอกจากจะขึ้นอยู่กับขนาดของกลุ่มตัวอย่าง และจำนวนตัวแปรพยากรณ์แล้วยังขึ้นอยู่กับขนาดของความสัมพันธ์ในประชากร ซึ่งไม่ทราบค่าอีกด้วย

คุณสมบัติของตัวประมาณค่าที่ดี

เนื่องจากค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองที่คำนวณได้จากกลุ่มตัวอย่าง ($R^2_{Y.123}$) เป็นตัวประมาณค่าที่เอนเอียงไปทางบวก ซึ่งมีค่าสูงกว่าความเป็นจริงเสมอ ทำให้เกิดปัญหาในการทดสอบสมมติฐานเพื่ออ้างอิงไปสู่ประชากรการทดสอบสมมติฐานโดยทั่วไปผู้วิจัยย่อมต้องการตัวประมาณค่าที่ดีที่สุด

ในการประมาณค่าพารามิเตอร์หรือลักษณะต่าง ๆ ของประชากรที่ใช้ในการวิเคราะห์ข้อมูล วิธีที่ใช้ในการประมาณค่ามี 2 แบบ คือ การประมาณค่าแบบจุด (Point estimation) และการประมาณค่าแบบช่วง (Interval estimation) สำหรับการประมาณค่าแบบจุดเป็นการประมาณค่าพารามิเตอร์ของประชากรที่สนใจศึกษาด้วยค่าเพียงค่าเดียวเท่านั้น เช่น ใช้ $\bar{x}$ ประมาณค่า μ หรือใช้ s^2 ประมาณค่า σ^2 เป็นต้น ส่วนวิธีการประมาณค่าแบบช่วงเป็นวิธีประมาณค่าพารามิเตอร์ของประชากรที่สนใจศึกษาด้วยช่วงค่าช่วงหนึ่ง ซึ่งมีคุณสมบัติว่า ค่าของประชากรที่แท้จริงจะตกอยู่ในช่วงค่าที่ประมาณนี้ ด้วยความเชื่อมั่นระดับหนึ่ง โดยจะต้องอาศัยการประมาณค่าแบบจุด และการแจกแจงค่าความน่าจะเป็นของตัวประมาณเป็นพื้นฐานในการคำนวณการประมาณค่าทั้งสองแบบจะเหมาะสมกับการใช้งานในกรณีที่ต่างกัน กรณีที่มีตัวประมาณค่า (Estimators) อยู่หลายตัวที่ สามารถนำมาใช้ในการประมาณค่าพารามิเตอร์ตัวใดตัวหนึ่งได้ จึงมีการกำหนดคุณสมบัติของวิธีการประมาณค่าที่ดี ควรมีคุณสมบัติครบ 4 ประการดังนี้ (Hay 1963: 196-201; Yamane 1967: 239-245 ; Wilks 1962 : 256-261)

1. ความไม่เอนเอียง (Unbiasness) หมายถึง ถ้า $\hat{\theta}$ เป็นตัวประมาณค่าที่ไม่เอนเอียงของพารามิเตอร์ θ แล้ว จะได้ว่า $E(\hat{\theta}) = \theta$ นั่นคือ ค่าที่คาดหวัง (Expected Value) ของตัวประมาณค่า $\hat{\theta}$ มีค่าเท่ากับค่าของพารามิเตอร์ และตัวประมาณค่าที่ไม่เอนเอียงนั้นมีคุณสมบัติที่ดีอยู่ว่า ถ้ามีชุดของตัวประมาณค่าที่ไม่เอนเอียงที่เป็นอิสระต่อกันอยู่แล้ว ค่าเฉลี่ยของค่าเหล่านั้น ย่อมไม่เอนเอียงด้วย และในทางตรงกันข้ามค่าเฉลี่ยของตัวประมาณค่าที่เอนเอียงย่อมเอนเอียงด้วยไม่ว่าจะเฉลี่ยมาจากกี่ค่าก็ตาม

2. ความสอดคล้อง (Consistency) หมายถึงถ้าประมาณค่าพารามิเตอร์ θ ด้วย $\hat{\theta}$ เมื่อขนาดของกลุ่มตัวอย่างใหญ่ขึ้น ตัวประมาณค่า $\hat{\theta}$ ที่มีคุณสมบัติที่ดีนี้ จะประมาณค่าเข้าใกล้ค่าพารามิเตอร์มากขึ้นด้วย ($p(\theta-\hat{\theta}) \rightarrow 1:\ n \rightarrow \infty$)

3. ความมีประสิทธิภาพ (Efficiency) หมายถึงตัวประมาณค่าหนึ่ง ๆ สามารถประมาณค่าพารามิเตอร์ได้ถูกต้องแม่นยำ (Accuracy) เพียงใด ซึ่งเกณฑ์ที่ใช้พิจารณาความมีประสิทธิภาพของตัวประมาณค่า ก็คือ ค่าความแปรปรวนของตัวประมาณค่าที่เปรียบเทียบกับกลุ่มของ

ตัวประมาณค่าที่ไม่เอนเอียงด้วยกัน กล่าวคือ ถ้าความแปรปรวนของ $\hat{\theta}_i$ หรือ Var($\hat{\theta}_i$) น้อยกว่าความแปรปรวนของ $\hat{\theta}_j$ หรือ Var($\hat{\theta}_j$) เมื่อทั้ง $\hat{\theta}_i$ และ $\hat{\theta}_j$ เป็นตัวประมาณค่าที่ไม่เอนเอียงแล้ว จะได้ $\hat{\theta}_i$ เป็นตัวประมาณค่าที่มีประสิทธิภาพของ θ

4. ความพอเพียง (Sufficiency) หมายถึงตัวประมาณค่า θ จะเป็นตัวประมาณค่าที่มีความพอเพียง ถ้ามันให้สารสนเทศที่ก่อให้เกิดประโยชน์ได้ทั้งหมดที่ต้องการเกี่ยวกับพารามิเตอร์ที่ต้องการประมาณ เช่น $\bar{x}$ เป็นตัวประมาณที่มีความพอเพียงของ μ ก็หมายความว่าไม่มีตัวประมาณค่าของ μ ตัวอื่น เช่น มัธยฐาน (Median) ที่จะสามารถให้ข่าวสารเกี่ยวกับ μ เพิ่มขึ้นได้อีก

สำหรับคุณสมบัติของตัวประมาณค่าที่ดีทั้ง 4 ประการดังกล่าว เป็นเกณฑ์ที่ใช้ในการตัดสินใจเลือกวิธีประมาณค่าทางทฤษฎีที่มีวิธีประมาณค่าอยู่หลายวิธี และแต่ละวิธีนั้นจะมีคุณสมบัติข้อใดบ้างสามารถใช้หลักฐานพิสูจน์ให้เห็นจริงได้

การปรับแก้ค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองด้วยวิธีของเวอร์รี่

(Adjusted Value of the Multiple R-Square by Wherry's Method)

เวอร์รี่ได้เสนอให้มีการปรับแก้ค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสอง โดยได้พัฒนาสูตรในการปรับแก้มาจากแนวคิดของลาร์สัน (Larson) ถ้าให้ ρ^2 เป็นค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองของประชากร σ_y เป็นค่าเบี่ยงเบนมาตรฐานของ Y และ $\sigma_{(Y-\hat{Y})}$ เป็นค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนในการพยากรณ์ (Y-Ŷ) และ Ŷ เป็นฟังก์ชั่นการถดถอยแล้ว (อ้างถึงใน Herzberg 1969 : 3)

$$\rho^2 = 1 - \frac{\sigma^2_{(Y-\hat{Y})}}{\sigma^2_y} \qquad (2.40)$$

$$\text{และ} \quad \bar{R}^2 = 1 - \frac{MSE}{s^2_y} \qquad (2.41)$$

เมื่อศึกษากับกลุ่มตัวอย่าง จาก (2.40) ลาร์สันเสนอให้ปรับแก้ค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองโดยการประมาณค่า (Y-Ŷ) ด้วย (MSE) n/n-p ดังนี้

$$R^2_L = 1 - \frac{(MSE)n/n-p}{s^2_y} \qquad (2.42)$$

$$= 1 - \frac{(MSE)n}{s^2_y(n-p)} \qquad (2.43)$$

แต่เนื่องจากใน (2.41) $MSE/s^2_y = 1-R^2$ ดังนั้น

$$R^2_L = 1 - \frac{n}{n-p}(1-R^2) \qquad (2.44)$$

เมื่อ n เป็นขนาดของกลุ่มตัวอย่างและ p เป็นจำนวนตัวแปรพยากรณ์ ต่อมาเวอร์รี่ได้แสดงให้เห็นว่าสูตรของลาร์สันยังประมาณค่าพารามิเตอร์ได้ไม่ดีนัก เนื่องจากประมาณค่า σ^2_y ด้วย S^2_y จึงเสนอให้ประมาณค่า σ^2_y ด้วย $S^2_y\, n/(n-1)$ ดังนี้

$$R^2_{w1} = 1 - \frac{n-1}{n-p}(1-R^2) \qquad (2.45)$$

แต่การนำสูตรดังกล่าวนี้ไปใช้ปรับแก้ค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองนั้นมีข้อจำกัดบางอย่าง คือค่าคงที่ (β_0) ในตัวแบบสมการถดถอยต้องมีค่าเป็นศูนย์ มิเช่นนั้น ค่าประมาณที่ไม่เอนเอียงของ $(Y-\hat{Y})$ จะต้องใช้ (MSE) n/n-p-1 จึงได้ปรับปรุงสูตร (2.44) ใหม่เป็น

$$R^2_{w2} = 1 - \frac{n-1}{n-p-1}(1-R^2) \qquad (2.46)$$

สูตรดังกล่าวนี้เป็นที่นิยมใช้กันอย่างกว้างขวางในหมู่นักวิจัยและสถิติ ดังจะเห็นได้จาก

การนำไปบรรจุไว้ในโปรแกรมคอมพิวเตอร์สำเร็จรูปที่ใช้ในการวิเคราะห์ทางสถิติต่าง ๆ เช่น โปรแกรมสำเร็จรูป SPPS* (Norusis 1983:141) และ SAS (SAS institute 1985 : 690) เป็นต้น

การปรับแก้ค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองด้วยวิธีของโอลกินกับแพรตต์
(Adjusted Value of the Multiple R-Square by Olkin & Pratt's Method)

การนำสูตรการปรับแก้ค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองของเวอร์รี่ไปใช้นั้น พบว่ายังไม่สามารถที่จะประมาณค่าพารามิเตอร์ได้อย่างถูกต้องนัก เนื่องจากอัตราส่วนของตัวประมาณค่าที่ไม่เอนเอียงทั้งสอง คือ (MSE) n/(n-p-1) และ (S^2_y) n/(n-1) จัดเป็นตัวประมาณค่าที่เอนเอียง ทำให้ขาดคุณสมบัติที่ดีของตัวประมาณค่าด้านความไม่เอนเอียง (Unbiasness) ไป ต่อมาในปี ค.ศ.1958 โอลกินกับแพรตต์ (Olkin & Pratt 1958: 201-211) จึงได้เสนอสูตรการประมาณค่าที่ไม่เอนเอียงเพื่อใช้ปรับแก้ค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองขึ้น โดยพัฒนามาจากแนวคิดเรื่องการแจกแจงของค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองของวิชาร์ด (Wishart) ดังนี้

จาก (2.36) และ (2.37)

$$E(R^2) = 1 - \frac{n-p-1}{n-1}(1-\rho^2)\ F(1,1,(n+1)/2,\ \rho^2) \qquad (2.47)$$

$$E(R^2) = 1 - \frac{n-p+1}{n-1}(1-\rho^2) - \frac{n-p-1}{n-1}\frac{2}{n+1}(1-\rho^2) \qquad (2.48)$$

โอลกินกับแพรตต์ได้เสนอการประมาณค่าที่ไม่เอนเอียงของ ρ^2 ไว้ดังนี้ คือ ถ้ามีจำนวนตัวแปรเกณฑ์ที่มีการแจกแจงแบบปกติหลายตัวแปร p+1 ตัว มีค่าเฉลี่ย $\underline{\mu}$ และความแปรปรวน Σ จำนวน N ค่า ตัวประมาณค่าที่ไม่เอนเอียงของค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองจะมีค่าเท่ากับ

$$\rho^2 = \rho^2_{0.(12..p)} = 1 - P / P_{00} \qquad (2.49)$$

เมื่อ P คือดีเทอร์มิแนนท์ (Determinant) ของเมตริกซ์สหสัมพันธ์และ P_{oo} เป็นแฟคเตอร์ร่วมอันดับแรก (First Cofactor) ของมัน ซึ่งเมื่อศึกษากับกลุ่มตัวอย่างจะได้

$$R^2 = R^2_{0.(12..p)} = 1 - \mathcal{R} / \mathcal{R}_{oo} \qquad (2.50)$$

เมื่อ $\mathcal{R}$ คือดีเทอร์มิแนนท์ (Determinant) ของเมตริกซ์สหสัมพันธ์ของกลุ่มตัวอย่างและ $\mathcal{R}_{oo}$ เป็นแฟคเตอร์ร่วมอันดับแรก (First Cofactor) ของมัน ซึ่งมีสภาวะ (Condition) ดังนี้

$$\sum_{k=0}^{\infty} \frac{\delta^2((n/2 + k)}{\delta\ ((p/2)+ k)} \frac{\rho^{2k}}{K!} \int_0^1 R^2_{OP}\ (R^2)^{((p-2)/2)+k}\ (1-R^2)^{(n-p-1)/2}\ dR^2$$

$$= \delta\ (n-p/2)\ \delta(n/2)\ (1- \rho^2)^{-n/2}\ \rho^2 \qquad (2.51)$$

ซึ่งจะได้

$$R^2_{OP} = 1 - \frac{n - 2}{n - p} (1-R^2)\ F\ (1,1;\ n-p+2/2;\ 1-R^2) \qquad (2.52)$$

จาก (2.52) ลบด้วยค่าคงที่ 1 จะได้

$$R^2_{OP} = 1 - \frac{n-3}{n-p-1} (1-R^2)\ F(1,1,(n-p+1)/2\ ,\ 1-R^2) \qquad (2.53)$$

เมื่อ F (a, b, c, x) เป็นฟังก์ชันไฮเปอร์จีออเมตริกซ์ (Hypergeometrix) ซึ่งถ้าใช้เพียงสองเทอมแรกของสมการจะได้

$$R^2_{OP} = 1 - \frac{n-3}{n-p-1} (1- R^2) - \frac{n-3}{n-p-1} \frac{2}{n-p+1} (1-R^2)^2 \qquad (2.54)$$

ซึ่งจะเห็นว่าในสองเทอมแรกของ (2.46) ที่เสนอโดยโอลกินกับแพรตต์ จะคล้ายกับ

(2.50) ที่เสนอโดยเวอร์รี่มาก ค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณที่ปรับแก้ มีค่าเปลี่ยนแปลงขึ้นอยู่กับจำนวนตัวแปรพยากรณ์ ขนาดของกลุ่มตัวอย่างและค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณที่คำนวณได้ครั้งแรก (R) ค่าสหสัมพันธ์พหุคูณที่ปรับแก้ด้วยวิธีทั้งสองนี้จะมีค่าเท่ากันเมื่อ $R^2 = 1$ แต่ถ้าในกรณีที่ $R^2 \neq 1$ แล้วจะพบว่าทั้งสองวิธีจะให้ค่าประมาณที่แตกต่างกันอย่างชัดเจน โดยเฉพาะเมื่อขนาดของกลุ่มตัวอย่างที่นำมาศึกษามีขนาดเล็กเมื่อเทียบกับจำนวนตัวแปรพยากรณ์ เช่น ถ้ามีจำนวนตัวแปรพยากรณ์เท่ากับ 2 ขนาดของกลุ่มตัวอย่างเท่ากับ 5 และ $R^2 = 0.58$ แล้ว ค่าสหสัมพันธ์พหุคูณที่คำนวณด้วยวิธีการปรับแก้ของ Wherry จะมีค่าเท่ากับ 0.16 ส่วนวิธีของ Olkin & Pratt จะมีค่าเท่ากับ 0.49 และเมื่อ R^2 มีค่าเท่ากับ 0.75 ค่าสหสัมพันธ์พหุคูณที่คำนวณได้จากวิธีการปรับแก้ทั้งสองจะมีค่าเป็น 0.50 และ 0.72 ตามลำดับ ที่ค่า $R^2 = 0.75$ นี้เมื่อจำนวนตัวแปรพยากรณ์เพิ่มขึ้นเท่ากับ 4 และขนาดของกลุ่มตัวอย่างเท่ากับ 7 ค่าสหสัมพันธ์พหุคูณที่คำนวณด้วยของ Wherry จะมีค่าเท่ากับ 0.25 ส่วนวิธีของ Olkin & Pratt มีค่าเท่ากับ 0.44 และเมื่อ R^2 มีค่าเท่ากับ 0.80 ค่าสหสัมพันธ์พหุคูณที่คำนวณด้วยวิธีการปรับแก้ทั้งสองจะมีค่าเท่ากับ 0.40 และ 0.56 ตามลำดับ จากลักษณะความแตกต่างที่เกิดขึ้นดังกล่าวนี้เอง ทำให้เกิดปัญหาแก่ผู้วิจัยเสมอมาในการเลือกใช้วิธีการปรับแก้ค่าสหสัมพันธ์พหุคูณว่า ควรจะเลือกใช้วิธีการใดและในสถานการณ์ใดจึงจะได้ตัวประมาณที่มีค่าใกล้เคียงกับค่าสหสัมพันธ์พหุคูณของประชากรมากที่สุด

งานวิจัยที่เกี่ยวข้อง

กรรณิการ์ เลียงเจริญสิทธิ์ (2527 : 48-49) ได้ใช้เทคนิคมอนติคาร์โลซิมูเลชั่น ทำการศึกษาการแจกแจงของค่าสหสัมพันธ์แบบปกติสองตัวแปร ณ ระดับความสัมพันธ์ในประชากร (ρ) ต่าง ๆ ตั้งแต่ ρ = 0.1, 0.2,......,0.9 เพื่อนำไปใช้ประโยชน์กรณีที่ต้องการสุ่มตัวอย่างที่มีคุณสมบัติตามที่ต้องการ ผลการศึกษาพบว่ายืนยันลักษณะการแจกแจงของข้อมูลว่ามีความเบ้ในกรณีที่ $\rho = 0$ แล้วขนาดของกลุ่มตัวอย่างมีน้อยกว่า 25 แต่เมื่อขนาดของกลุ่มตัวอย่างมีค่าเท่ากับหรือมากกว่า 25 การแจกแจงของค่าสหสัมพันธ์จะมีลักษณะเป็นปกติโดยประมาณ และยืนยันว่าเมื่อแปลงค่าสหสัมพันธ์โดยวิธี Fisher's tranformation แล้ว Z_F จะมีลักษณะการแจกแจงเป็นปกติโดยประมาณ ข้อสรุปที่สำคัญที่ได้จากการศึกษา คือในการทดสอบสมมติฐานกรณีที่ ρ มีค่าอื่น ๆ ที่ไม่เท่ากับ 0 ณ ระดับ $\alpha = 0.01$ ขนาดของกลุ่มตัวอย่างที่เหมาะสมควรใช้ตั้งแต่ 9 ขึ้นไปที่ระดับ $\alpha = 0.05$ และที่ระดับ $\alpha = 0.10$ ควรใช้ตั้งแต่ 5 ขึ้นไป

ฮาลินสกีและเฟลด์ (Halinske and Feldt 1970: 151-158) ได้ทำการศึกษา

โดยใช้เทคนิคมอนติการ์โลซิมูลชั่น เกี่ยวกับขนาดของกลุ่มตัวอย่างที่เหมาะสมในการวิเคราะห์การถดถอยพหุคูณ พบว่า ควรใช้อัตราส่วนระหว่างขนาดของกลุ่มตัวอย่างกับจำนวนตัวแปรอย่างน้อยที่สุดเท่ากับ 10:1

มิลเลอร์และคันซ์ (Miller and Kunce 1978: 157-163) ได้ทำการศึกษาแบบครอสแวลิเดชั่น (Cross-Validation) เกี่ยวกับขนาดของกลุ่มตัวอย่างที่เหมาะสมในการวิเคราะห์การถดถอยพหุคูณ พบว่าควรใช้อัตราส่วนระหว่างขนาดของกลุ่มตัวอย่างกับจำนวนตัวแปรอย่างน้อยที่สุดเท่ากับ 10:1

นอร์แมนและเทอร์รี่ (Norman and Terry 1970 : 481-489) ได้ทำการศึกษาเปรียบเทียบสูตรการปรับแก้ค่าสหสัมพันธ์พหุคูณ 3 คือวิธีของลอร์ด (Lord) และ เวอร์รี่ (Wherry) ดังนี้

$$R^2_{LORD} = 1 - \frac{n-p+1}{n-p-1}(1-R^2) \qquad (2.55)$$

$$R^2_{w1} = 1 - \frac{n-1}{n-p}(1-R^2) \qquad (2.56)$$

$$R^2_{w2} = 1 - \frac{n-1}{n-p-1}(1-R^2) \qquad (2.57)$$

โดยใช้ข้อมูลที่เก็บได้จริงจากแบบสอบ Army Classification Battery และแบบสอบ Navy General Classification จำนวน 975 คน พบว่า วิธีของลอร์ดมีความคงที่ในการประมาณค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองของประชากรได้ถูกต้องมากกว่าวิธีของเวอร์รี่ทั้งสองวิธี และเมื่อเปรียบเทียบคะแนนมาตรฐานของความแตกต่างระหว่างค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองที่คำนวณได้กับค่าจริง พบว่าวิธีของลอร์ดมีค่าต่ำกว่าวิธีของเวอรี่เช่นเดียวกัน ที่เป็นเช่นนี้ผู้วิจัยได้ให้ความเห็นว่าเป็นเพราะวิธีของลอร์ดเป็นวิธีที่พัฒนาอยู่บนพื้นฐานของการศึกษาค่า R^2 แบบครอส-แวลิเดชั่น ส่วนวิธีของเวอร์รี่นั้นเป็นสูตรที่พัฒนาจากการคาดหวัง (Expected) ค่า R^2 จากจักรวาล (Universe) ของ R^2