บทที่ 2

เอกสารและผลงานวิจัยที่เกี่ยวข้อง

การวิเคราะห์สหสัมพันธ์และการถดถอย

(Simple Correlation and Regression Analysis)

การวิเคราะห์สหสัมพันธ์และการถดถอย เป็นวิธีการทางสถิติที่ใช้ในการประมาณว่าตัวแปร (Variable) สองตัวหรือมากกว่านั้นมีความสัมพันธ์กันหรือไม่เพียงใด เทคนิคการวิเคราะห์ทั้งสองนี้ มีธรรมชาติการสรุปผลใกล้เคียงกันมาก โดยปกติแล้วความสัมพันธ์ระหว่างตัวแปรช่วยให้สามารถอธิบายหรือพยากรณ์ล่วงหน้าได้ แต่ในปัจจุบันพบว่าได้มีการเน้นในเรื่องการวิเคราะห์การถดถอยมากกว่า (Samuel 1982 : 193-203) ดังนั้น จึงสามารถแยกให้เห็นความแตกต่างระหว่างเทคนิคการวิเคราะห์ทั้งสองวิธีดังนี้

การวิเคราะห์การถดถอย (Regression analysis) ใช้ในการพิจารณาถึงรูปแบบที่เป็นไปได้ของความสัมพันธ์ระหว่างตัวแปร ในรูปแบบที่สามารถสร้างเป็นสมการได้ คือ $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + e_i$ เมื่อ Y_i เป็นตัวแปรเกณฑ์ (Criterion variable) และ X_i เป็นตัวแปรทำนาย (Predictor variable) ซึ่งการวิเคราะห์การถดถอยนี้มีวัตถุประสงค์เพื่อใช้ประโยชน์ในการทำนาย (Predict) หรือประมาณ (Estimate) ค่าๆ หนึ่งที่สัมพันธ์กับค่าที่กำหนดให้อีกค่าหนึ่ง ผู้ที่ริเริ่มศึกษาเรื่องนี้ได้แก่ Sir Francis Galton ซึ่งเป็นนักวิทยาศาสตร์ชาวอังกฤษ โดยการศึกษาถึงความสัมพันธ์ระหว่างตัวแปร ซึ่งนำไปสู่ทางคิดค้นเกี่ยวกับการวิเคราะห์การถดถอยในเวลาต่อมา โดยการได้ศึกษาถึงแนวโน้มของลักษณะพันธุกรรมที่บุตรหลานสืบทอดจากบิดามารดา ในรายงานการวิจัยเกี่ยวกับพันธุกรรมของเขาเมื่อปี ค.ศ.1899

การวิเคราะห์สหสัมพันธ์ (Correlation analysis) จะเกี่ยวกับการวิเคราะห์เชิงเส้นระหว่างตัวแปร ซึ่งรูปแบบการวิเคราะห์นั้นได้รับการคิดค้นโดย Galton ในระยะหลังจากการวิเคราะห์การถดถอยได้ใช้กันมาในระยะหนึ่งแล้ว

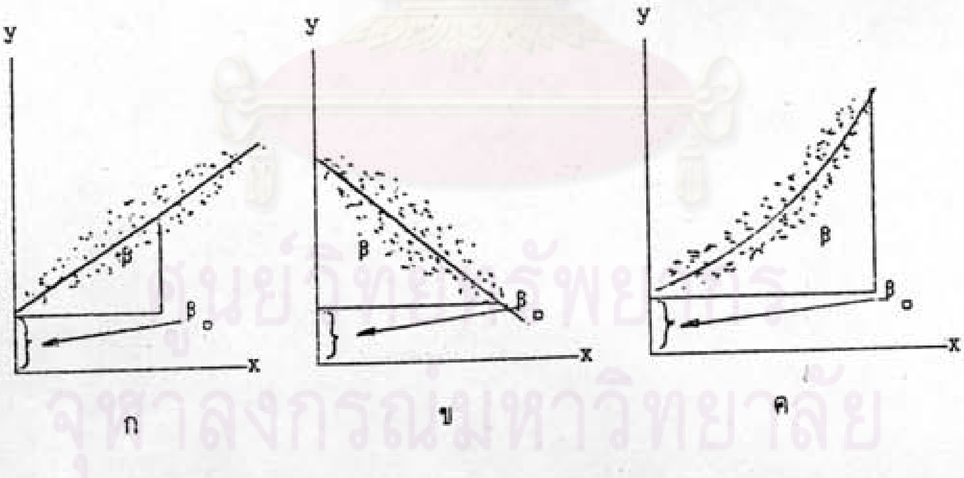
การวิเคราะห์สหสัมพันธ์และการถดถอยนี้ ถ้าเป็นการวิเคราะห์เกี่ยวกับตัวแปรเพียงสองตัว (Bivariate) เรียกว่า สหสัมพันธ์ หรือการถดถอยอย่างง่าย (Simple Correlation or Simple Regression) ส่วนสหสัมพันธ์หรือการถดถอยพหุคูณ (Multiple Correlation or Multiple Regression) หมายถึง การวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรตาม 1 ตัว กับตัวแปรพยากรณ์ตั้งแต่ 2 ตัวขึ้นไป

การวิเคราะห์การถดถอยอย่างง่าย

(Simple Regression Analysis)

การถดถอยเชิงเส้นตรงอย่างง่ายนี้ หมายถึง การถดถอยของ Y ที่มีต่อตัวแปรอิสระ X เพียงตัวเดียว และสามารถดูลักษณะการถดถอยของ Y ที่มีต่อ X ได้จากแผนภาพการกระจาย (Scatter diagram) ซึ่งมีลักษณะต่างๆ กันดังนี้

<u>แผนภาพที่ 1</u> การถดถอยของตัวแปร X และ Y



จากรูปที่ 1 (ก, ข) แสดงให้เห็นถึงการถดถอยเชิงเส้นอย่างง่ายของตัวแปรตาม Y ที่มีตัวแปรอิสระ X เพียงตัวเดียว ซึ่งสามารถแสดงความสัมพันธ์ในรูปของตัวแบบสมการทางคณิตศาสตร์ ดังนี้

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

เมื่อ β_0 คือ ค่าที่เส้นตรงตัดแกน y (y-intercept) และ β คือ พารามิเตอร์ที่แสดงความชันของเส้นตรง เรียกว่า สัมประสิทธิ์การถดถอย (Regression Coefficient) ซึ่งเป็นค่าที่แสดงอัตราการเปลี่ยนของค่า Y เมื่อ X เปลี่ยนไป 1 หน่วย โดยจะมีค่ามากกว่า 0 เมื่อ Y มีการถดถอยไปทางเดียวกับ X (รูป ก) คือ เมื่อค่า X เพิ่ม ค่า Y จะเพิ่ม เมื่อค่า X ลด ค่า Y ก็จะลดด้วย ส่วนรูป ข β จะมีค่าเป็นลบ เมื่อ Y มีการถดถอยไปทางตรงกันข้ามกับ X นั่นคือ เมื่อค่า X เพิ่ม ค่า Y จะลดลง และเมื่อค่า X ลด ค่า Y จะเพิ่ม และถ้า β มีค่าเท่ากับ 0 เมื่อการเปลี่ยนแปลงของค่า Y ไม่ขึ้นอยู่กับการเปลี่ยนแปลงของค่า X เลย

จากสมการการถดถอย $Y_i = \beta_0 + \beta_i X_i + \varepsilon_i$ นั้น ε_i เป็นค่าความคลาดเคลื่อนที่มีลักษณะเป็นตัวแปรสุ่ม ซึ่งเป็นอิสระจาก X และ Y ภายใต้ข้อสมมติดังนี้

1. ค่า X_i ต้องเป็นค่าที่วัดได้ และเป็นค่าที่กำหนดให้คงที่ (fixed variable)

2. ε_i มีค่าเฉลี่ยสำหรับแต่ละ Y_i เท่ากับ 0 หรือ $(E(\varepsilon_i) = 0)$

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

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 $(\beta_0, \beta, \sigma^2)$ ที่แท้จริงของประชากรได้ แต่จะประมาณได้จากข้อมูลกลุ่มตัวอย่างที่ศึกษา (Y_i, X_i) จำนวน n คู่ ซึ่งจะได้

ค่าประมาณของ Y_i ดังนี้

$$\hat{Y}_i = \hat{\beta}_o + \hat{\beta}X_i = b_o + bX_i = \hat{\mu}_{Yi.xi} \quad (2.3)$$

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

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y})^2 = (Y_i - b_o - bX_i)^2 \quad (2.4)$$

$$\frac{\partial}{\partial \beta_o} (\sum_{i=1}^{n} \hat{\varepsilon}_i^2) = -2 \sum_{i=1}^{n} Y_i + 2an + 2b \sum_{i=1}^{n} X_i) \quad (2.5)$$

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

จากผลข้างต้นนี้จะทำให้ได้ชุดสมการปกติ (Normal equations) ดังนี้ (Lindeman 1980 : 99)

$$nb_o + b \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \quad (2.7)$$

$$b_o \sum_{i=1}^{n} X_i + b \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i \quad (2.8)$$

ซึ่งให้ค่าประมาณของ $\hat{\beta}_o$ และ $\hat{\beta}$ ดังนี้

$$\hat{\beta}_o = b_o = \bar{y} - b\bar{x} \quad (2.9)$$

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

การวิเคราะห์สหสัมพันธ์ (Correlation Analysis)

การวิเคราะห์สหสัมพันธ์ หมายถึง กระบวนการวัดความสัมพันธ์ระหว่างตัวแปรใดๆ ตั้งแต่ 2 ตัวขึ้นไป โดยไม่ได้คำนึงว่า ตัวแปรใดเป็นตัวแปรอิสระ หรือตัวแปรตาม ซึ่ง Karl Pearson เป็นผู้คิดค้นกระบวนการวัดนี้ขึ้นมา

ค่าของความสัมพันธ์นี้จะเป็นจำนวนจริงอยู่ตั้งแต่ -1 ถึง +1 ถ้าระดับขั้นของความสัมพันธ์เป็น 0 แสดงว่าตัวแปรทั้งสองไม่มีความสัมพันธ์กันเลย แต่ถ้าระดับขั้นของความสัมพันธ์เป็น 1 แสดงว่าตัวแปรเหล่านั้นมีความสัมพันธ์กันอย่างสมบูรณ์เชิงบวก (positive perfect correlation) และถ้าระดับขั้นของความสัมพันธ์เป็นบวกอยู่ระหว่าง 0 ถึง +1 ก็จะแสดงระดับขั้นความสัมพันธ์เชิงบวกสูงต่ำตามลำดับ เช่น ความสัมพันธ์เป็น +0.5 ก็แสดงว่าตัวแปรเหล่านั้นมีความสัมพันธ์กันเชิงบวกปานกลาง ถ้ามากกว่า 0.5 ก็แสดงว่ามีความสัมพันธ์กันเชิงบวกค่อนข้างมาก เป็นต้น ส่วนในกรณีที่ระดับขั้นของความสัมพันธ์เป็น -1 ก็แสดงว่าตัวแปรเหล่านั้นมีความสัมพันธ์กันอย่างสมบูรณ์เชิงลบ (Negative perfect correlation) และถ้าระดับขั้นของความสัมพันธ์เป็นเลขอยู่ระหว่าง -1 ถึง 0 ก็จะแสดงระดับขั้นความสัมพันธ์เชิงลบสูงต่ำตามลำดับ

สหสัมพันธ์สามารถแยกประเภทของระดับขั้นความสัมพันธ์ได้เป็น 4 ประเภท ดังนี้ (คณิต ไข่มุกด์ 2529 : 268-269)

1. **สหสัมพันธ์เชิงเส้นอย่างง่าย** (Simple Linear Correlation) ก็คือ ระดับขั้นความสัมพันธ์ของตัวแปร 2 ตัว ในลักษณะเชิงเส้น

2. **สหสัมพันธ์ที่ไม่เป็นเชิงเส้นอย่างง่าย** (Simple Non-Linear Correlation) ก็คือระดับขั้นความสัมพันธ์ของตัวแปร 2 ตัว ในลักษณะที่ไม่เป็นเชิงเส้น

3. **สหสัมพันธ์เชิงเส้นพหุคูณ** (Multiple Linear Correlation) ก็คือระดับขั้นความสัมพันธ์ของตัวแปรตั้งแต่ 3 ตัวขึ้นไปในลักษณะเชิงเส้น

4. **สหสัมพันธ์ที่ไม่เป็นเชิงเส้นพหุคูณ** (Multiple Non-Linear Correlation) ก็คือ ระดับขั้นความสัมพันธ์ของตัวแปรตั้งแต่ 3 ตัวขึ้นไปในลักษณะที่ไม่เป็นเชิงเส้น

สหสัมพันธ์เชิงเส้นอย่างง่าย

(Simple Linear Correlation)

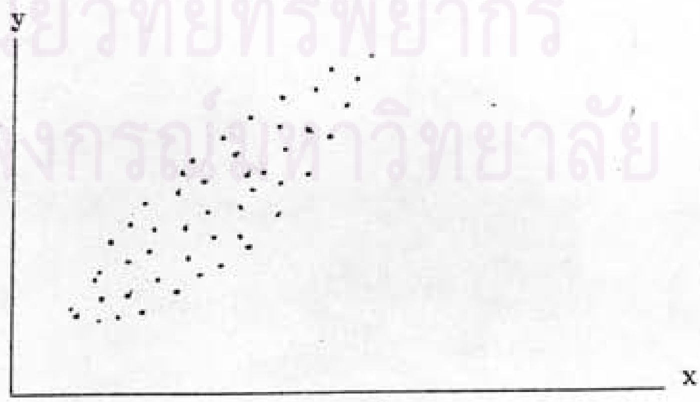
เป็นการวัดระดับขั้นความสัมพันธ์ของตัวแปรเพียง 2 ตัวในลักษณะเชิงเส้น ซึ่งก็มีค่าอยู่ตั้งแต่ -1 ถึง +1 ในการวัดสหสัมพันธ์โดยปกติในขั้นต้น เราจะนำข้อมูลมาทำแผนภาพกระจาย (Sactter Diagram) เมื่อดูแนวโน้มของความสัมพันธ์ว่าเป็นลักษณะเชิงเส้น หรือไม่เป็นเชิงเส้น หรือไม่มีความสัมพันธ์กันเลย แล้วอาจจะพิจารณาจากค่าความชัน เพราะถ้าค่าความชันเป็นบวก ค่าสหสัมพันธ์ก็เป็นบวกด้วย ถ้าค่าความชันเป็นลบ ค่าสหสัมพันธ์ก็เป็นลบด้วย และถ้าค่าความชันเป็น 0 หรือเป็นค่าอนันต์ ค่าสหสัมพันธ์ก็เป็น 0 ด้วย

ความสัมพันธ์ระหว่างตัวแปรสองตัวนี้อาจจะเป็นไปในทางเดียวกันหรือตรงข้ามกันก็ได้ โดยมีค่าที่ใช้วัดความสัมพันธ์เรียกว่า สัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) ซึ่งใช้สัญลักษณ์ " ρ " เป็นสัมประสิทธิ์สหสัมพันธ์ของประชากร และใช้ " r " เป็นสัมประสิทธิ์สหสัมพันธ์ของกลุ่มตัวอย่าง

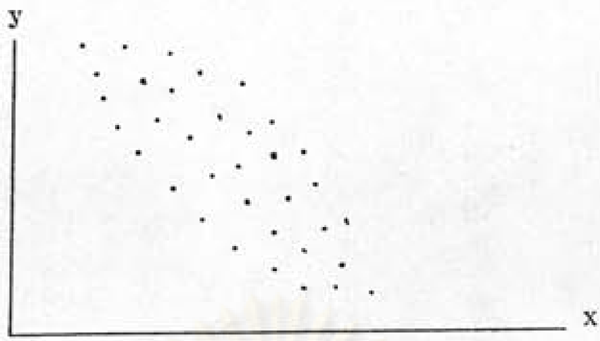
สำหรับการวิเคราะห์สหสัมพันธ์เชิงเส้นอย่างง่าย อาจจะพิจารณาได้จาก

1) แผนภาพการกระจาย (Scatter Diagram) ซึ่งได้จากการนำจุดพิกัดมาลงจุด (plot) แล้วดูแนวโน้มว่าจะมีความสัมพันธ์กันเชิงบวก เชิงลบ หรือไม่มีความสัมพันธ์กัน ดังนี้

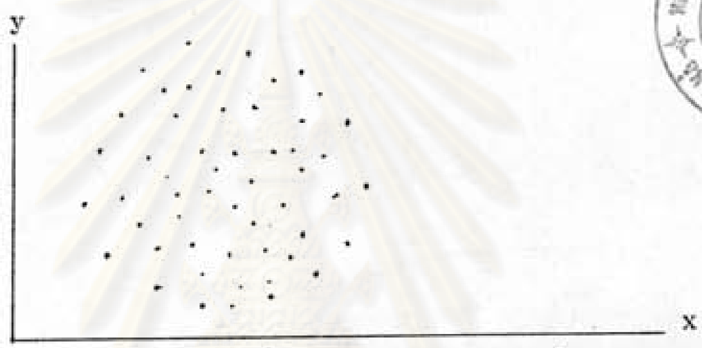
<u>แผนภาพที่ 2</u> แผนภาพการกระจายของค่าสหสัมพันธ์เชิงเส้นอย่างง่ายของ x และ y



รูป 2.ก แสดงความสัมพันธ์เชิงบวก



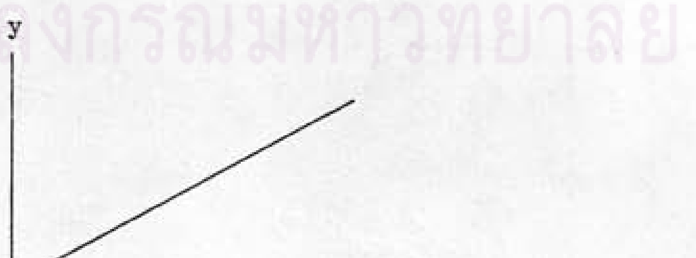
รูป 2.ข แสดงความสัมพันธ์เชิงลบ



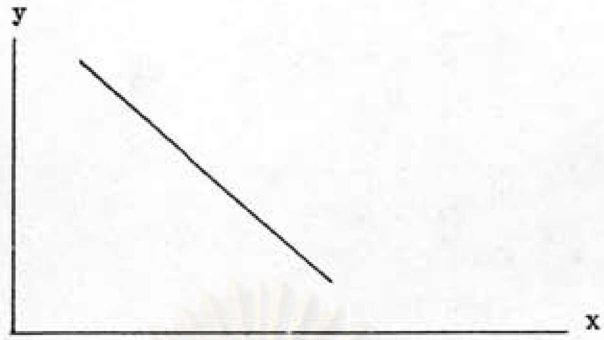
รูป 2.ค แสดงถึงการไม่มีความสัมพันธ์กันของตัวแปรทั้งสอง

2) พิจารณาจากค่าความสัมพันธ์ของสมการถดถอยเชิงเส้นอย่างง่าย $y = a + bx$ ดังนี้

<u>แผนภาพที่ 3</u> ความชันของสมการถดถอยเชิงเส้นอย่างง่ายของ x และ y

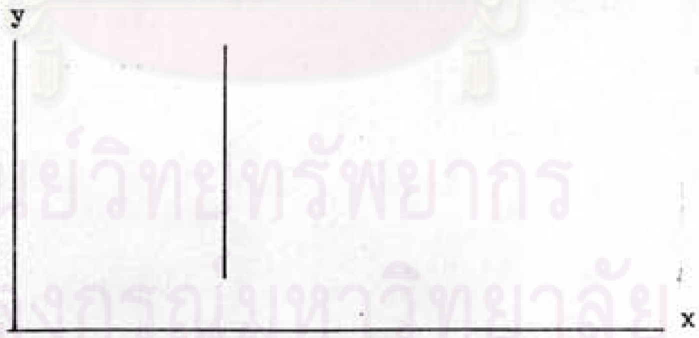


รูป 3.ก แสดงค่าความชันเป็นบวก สหสัมพันธ์ก็เป็นบวกด้วย



รูป 3.ข แสดงค่าความชันเป็นลบ สหสัมพันธ์ก็เป็นลบด้วย

รูป 3.ค แสดงค่าความชันเป็น 0 สหสัมพันธ์ก็เป็น 0 ด้วย



รูป 3.ง แสดงค่าความชันเป็นอนันต์ สหสัมพันธ์ก็เป็น 0

จะเห็นได้ว่า ถ้าเราพิจารณาจากค่าของความชันแล้ว สหสัมพันธ์จะมีค่าสูง เมื่อค่าความชันเท่ากับ 1 และจะมีค่าลดลงเมื่อความชันมีค่าสูงขึ้นหรือลดลง ดังนั้นการพิจารณาค่าสหสัมพันธ์จากค่าความชัน ก็พิจารณาได้ระดับหนึ่งเท่านั้น และอาจทำให้เกิดความผิดพลาดได้ จึงมีผู้คิดค้นสูตรเพื่อหาค่าสหสัมพันธ์ดังนี้

3) สูตรในการคำนวณหาค่าสัมประสิทธิ์สหสัมพันธ์ ได้จากการเปรียบเทียบความแปรปรวนร่วม (Covariance) ระหว่างตัวแปรสองตัว (x,y) กับผลคูณของความเบี่ยงเบนมาตรฐาน (Standard Deviation) ของ x และ y ดังนี้

$$\rho = \frac{\Sigma Z_x Z_y}{N}$$

$$\rho = \rho_{xy} = \rho_{yx} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \qquad (2.11)$$

$$\rho_{xy} = \frac{\sum_{i=1}^{N} (X_i - \mu_x)(Y_i - \mu_y)}{\sum_{i=1}^{N} (X_i - \mu_x)^2 \sum_{i=1}^{N} (Y_i - \mu_y)^2} \qquad (2.12)$$

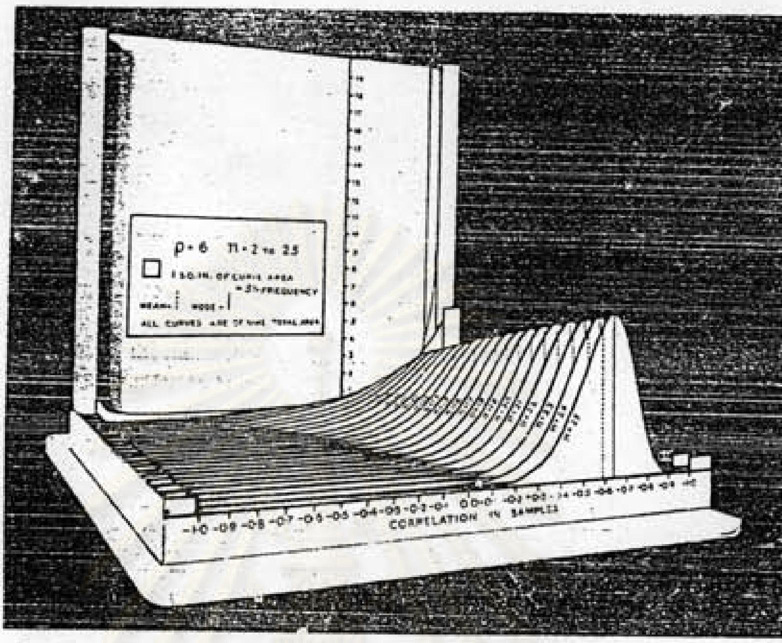
เมื่อ n คือจำนวนข้อมูลทั้งหมดในประชากรของ x และ y และ ρ คือสัมประสิทธิ์สหสัมพันธ์อย่างง่ายของประชากร ซึ่งเป็นค่าที่ไม่มีหน่วย แต่จะบอกถึงความสัมพันธ์ระหว่างตัวแปรว่ามีมากน้อยเพียงใด ค่า ρ นี้สามารถประมาณได้จากข้อมูลกลุ่มตัวอย่างดังนี้

$$\rho = \hat{\rho}_{xy} = r = r_{xy} = r_{yx} \qquad (2.13)$$

$$r_{xy} = \frac{N \Sigma XY - \Sigma X \Sigma Y}{\sqrt{[N \Sigma X^2 - (\Sigma X)^2][N \Sigma Y^2 - (\Sigma Y)^2]}} \qquad (2.14)$$

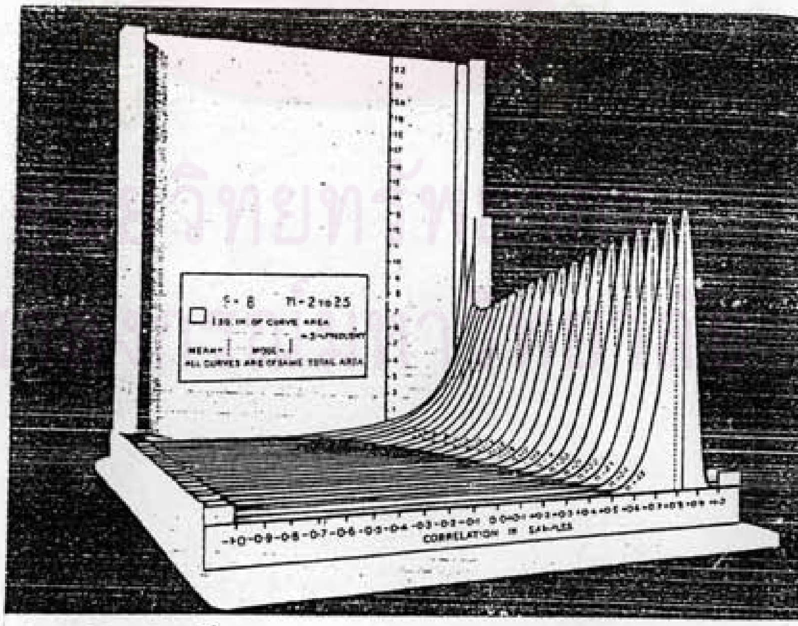
จากการศึกษาการแจกแจงของค่าสัมประสิทธิ์สหสัมพันธ์ (r_{xy}) ของ R.A. Fisher เมื่อปี ค.ศ.1915 (R.A. Fisher 1915 : 507-521) พบว่า การแจกแจงของค่า r_{xy} นั้นขึ้นอยู่กับค่าของ ρ และ n เท่านั้น ปีต่อมา H.E. Soper (H.E. Soper 1916 : 318-413) ก็ได้ศึกษาการแจกแจงของค่า r_{xy} เมื่อ n มีขนาดเล็ก และขณะที่ $\rho \neq 0$ ก็พบว่าการแจกแจงของค่า r_{xy} เบ้ซ้าย ดังแสดงไว้ในแผนภาพที่ 4 และ 5

<u>แผนภาพที่ 4</u> การแจกแจงของค่าสหสัมพันธ์อย่างง่าย เมื่อขนาดของกลุ่มตัวอย่างเป็น 2 ถึง 25 ขณะที่ค่า $\rho = 0.6$



Correlation in Small Samples. $\rho = 0{\cdot}6$. Frequency curves for samples of sizes two to twenty-five, showing the changes in type from a skew "cocked hat" to J- and U-forms. Model A.

<u>แผนภาพที่ 5</u> แสดงการแจกแจงของค่าสหสัมพันธ์อย่างง่าย เมื่อขนาดของกลุ่มตัวอย่างเป็น 2 ถึง 25 ขณะที่ค่า $\rho = 0.80$



Correlation in Small Samples. $\rho = 0{\cdot}8$. Frequency curves for samples of sizes two to twenty-five, showing the changes in type from a skew "cocked hat" to J- and U-forms. Model B.

การวิเคราะห์การถดถอยและสหสัมพันธ์เชิงพหุ

(Multiple Regression and Correlation Analysis)

การวิเคราะห์การถดถอยเชิงพหุ เป็นแนวคิดและเทคนิคที่ขยายมาจากการวิเคราะห์การถดถอยอย่างง่าย (Simple Correlation Regression) เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรตาม (Dependent Variable) และกลุ่มของตัวแปรอิสระ (Set of Independent Variables) ภายใต้รูปแบบที่กำหนด ซึ่งอาจเป็นเส้นตรงหรือเส้นโค้ง จากการวิเคราะห์การถดถอยจะได้สมการถดถอยเชิงพหุ ซึ่งโดยปกติแล้วจะทำให้สามารถวิเคราะห์และอธิบายตัวแปรเกณฑ์ได้มากกว่า เพราะในการวิเคราะห์ปัญหาบางอย่างที่จำเป็นต้องใช้การถดถอยนั้น บางครั้งการศึกษาการถดถอยอย่างง่ายอาจจะไม่เพียงพอ ทั้งนี้เพราะการประมาณค่าของตัวแปรเกณฑ์เพื่อให้ใกล้เคียงที่สุดนั้น เรามักจะต้องพิจารณาตัวแปรพยากรณ์ที่มีอิทธิพล หรือมีความสัมพันธ์ต่อตัวแปรเกณฑ์มากกว่า 1 ตัวขึ้นไป โดยมีสมการถดถอยเป็นตัวชี้ให้เห็นถึงความสัมพันธ์ถัวเฉลี่ยของตัวแปรเหล่านั้น ดังนี้ (Lindeman 1980 : 93)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i \quad (2.15)$$

โดยที่ Y_i คือ ตัวแปรเกณฑ์ (Criterion Variables or Dependent Variable)

X_i คือ ตัวแปรพยากรณ์ (Predictor Variable or Independent Variable) ; $i = 1, 2, \ldots, p$

β_0 คือ ค่าคงที่ ซึ่งจะมีค่าเท่ากับ Y เมื่อ X_{ip} ทั้งหมดมีค่าเท่ากับ 0

β_i คือ สัมประสิทธิ์การถดถอยของประชากรบางส่วน (Population Partial Regression Coefficient) ของ X_i เมื่อให้ $X_{i+1}, \ldots, X_p$ เป็นค่าคงที่ นั่นคือ เมื่อ X_i มีค่าเปลี่ยนแปลงไป b_i หน่วย เมื่อตัวแปรพยากรณ์อื่นๆ คงที่ และเป็นค่าประมาณของ β_i

ε_i คือ ความคลาดเคลื่อนที่แสดงถึงความแตกต่างระหว่างสมการถดถอย
กับค่าจริง มีลักษณะเป็นตัวแปรสุ่มที่ไม่ทราบค่า และเป็นค่า
ประมาณของ ε_i, $\varepsilon \sim N\ (0, \sigma^2 I)_{n \times n}$

$$
\begin{aligned}
E(\varepsilon_i) &= 0 \quad ; i = 1, 2, 3, \ldots, n \\
E(\varepsilon_i, \varepsilon_j) &= \sigma^2_{ij} \quad ; i = j \\
&= 0 \quad ; i \neq j
\end{aligned}
$$

จากสมการ (2.15) สามารถแสดงในรูปของเมตริกซ์ได้ดังนี้

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

เมื่อ

$$
\underset{\sim}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1} , \quad X = \begin{bmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1p} \\ 1 & X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \ldots & X_{np} \end{bmatrix}_{n \times (p+1)}
$$

$$
\underset{\sim}{\beta} = \begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_p \end{bmatrix}_{(p+1) \times 1} , \quad \underset{\sim}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n \times 1}
$$

จากตัวแบบ (model) (2.15) ผลรวมของความคลาดเคลื่อนกำลังสองจะมีค่าเป็น
(วิชิต หล่อธีระชุณห์กุล 2524 : 124)

$$
\begin{aligned}
\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\underset{\sim}{\varepsilon}}' \hat{\underset{\sim}{\varepsilon}} &= (\underset{\sim}{Y} - X\underset{\sim}{b})'(\underset{\sim}{Y} - X\underset{\sim}{b}) \\
&= \underset{\sim}{Y}'\underset{\sim}{Y} - \underset{\sim}{b}'X'\underset{\sim}{Y} - \underset{\sim}{Y}'X\underset{\sim}{b} + \underset{\sim}{b}'X'X\underset{\sim}{b}
\end{aligned}
$$

$$= \underset{\sim}{Y}'\underset{\sim}{Y} - 2\underset{\sim}{b}'X'\underset{\sim}{Y} + \underset{\sim}{b}'X'X\underset{\sim}{b} \qquad (2.16)$$

การประมาณค่าพารามิเตอร์ในการวิเคราะห์การถดถอย

(Estimation of Parameters in Multiple Regression)

เนื่องจากการวิจัยในทางปฏิบัตินั้น ผู้วิจัยจะไม่สามารถศึกษาจากกลุ่มประชากรทั้งหมดได้ จึงไม่ทราบค่าพารามิเตอร์ β ที่แท้จริง โดยทั่วไปจะประมาณค่าพารามิเตอร์ (Parameter) จากข้อมูลกลุ่มตัวอย่างที่นำมาศึกษา วิธีการประมาณตัวประมาณค่าที่นิยมใช้กันมาก คือ วิธีกำลังสองน้อยที่สุด (Least Square Estimation Method) ซึ่งจะได้ค่าประมาณสัมประสิทธิ์การถดถอยพหุคูณ จากอนุพันธ์ (Differentiate) ในสมการ (2.16) เทียบกับ b แล้วได้ค่าเท่ากับศูนย์

$$\frac{\partial(\hat{e}'\hat{e})}{\partial \underset{\sim}{b}} = -2X'\underset{\sim}{Y} + 2X'X\underset{\sim}{b} = 0$$

$$X'X\underset{\sim}{b} = X'\underset{\sim}{Y}$$

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

ซึ่งจะได้ตัวประมาณค่าที่ไม่เอนเอียงของ β ซึ่งมีค่าเฉลี่ยความคลาดเคลื่อนกำลังสองน้อยที่สุดในบรรดาตัวประมาณค่าที่ไม่เอนเอียงทั้งหลาย ดังจะเห็นได้จากสมการทั่วไป (Normal equations) ที่จะใช้หา b_i , $i = 0,1,2,\ldots,p$

จากวิธีกำลังสองน้อยที่สุด (Least Squares Method) ซึ่งกำหนดโดย

$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ (Lindeman 1980 : 98-99)

ให้ $\quad G = \sum_{i=1}^{n} (Y_i - Y_i)^2$

และ $\quad \hat{Y}_i = b_0 + b_1X_{i1} + b_2X_{i2}$

แทนค่า $\hat{Y}_i$ จะได้ $G = \sum_{i=1}^{n} (Y_i - b_0 - b_1X_{i1} - b_2X_{i2})^2$

$$= \sum_{i=1}^{n} Y_i + nb_0 + b^2 \sum_{i=1}^{n} X_{i1} + b^2 \sum_{i=1}^{n} X_{i2} - 2b_0 \sum_{i=1}^{n} Y_i$$

$$- 2b_1 \sum_{i=1}^{n} X_{i1}Y_i - 2b_2 \sum_{i=1}^{n} X_{i2}Y_i + 2b_0 b_1 \sum_{i=1}^{n} X_{i1}$$

$$+ 2b_0 b_2 \sum_{i=1}^{n} X_{i2} \quad 2b_1 b_2 \sum_{i=1}^{n} X_{i1}X_{i2}$$

จากนั้นหาอนุพันธ์ของ G ในสมการ 2.19 เทียบกับ b_0, b_1 และ b_2

$$\frac{\partial G}{\partial b_0} = 2nb_0 - 2\sum_{i=1}^{n} Y_i + 2b_1 \sum_{i=1}^{n} X_{i1} + 2b_2 \sum_{i=1}^{n} X_{i2}$$

$$\frac{\partial G}{\partial b_1} = 2b_1 \sum_{i=1}^{n} X_{i1}^2 - 2\sum_{i=1}^{n} X_{i1}Y_i + 2b_0 \sum_{i=1}^{n} X_{i1} + 2b_2 \sum_{i=1}^{n} X_{i1}X_{i2}$$

$$\frac{\partial G}{\partial b_2} = 2b_2 \sum_{i=1}^{n} X_{i2}^2 - 2\sum_{i=1}^{n} X_{i2}Y_i + 2b_0 \sum_{i=1}^{n} X_{i2} + 2b_1 \sum_{i=1}^{n} X_{i1}X_{i2}$$

ให้แต่ละสมการเหล่านี้เป็น 0 เราก็จะได้ normal equations เพื่อที่จะแก้สมการหาค่า b_0, b_1 และ b_2 ต่อไป

$$nb_0 + \left(\sum_{i=1}^{n} X_{i1}\right)b_1 + \left(\sum_{i=1}^{n} X_{i2}\right)b_2 = \sum_{i=1}^{n} Y_i$$

$$\left(\sum_{i=1}^{n} X_{i1}\right)b_0 + \left(\sum_{i=1}^{n} X_{i2}^2\right)b_1 + \left(\sum_{i=1}^{n} X_{i1}X_{i2}\right)b_2 = \sum_{i=1}^{n} X_{i1}Y_i$$

$$\left(\sum_{i=1}^{n} X_{i2}\right)b_0 + \left(\sum_{i=1}^{n} X_{i1}X_{i2}\right)b_1 + \left(\sum_{i=1}^{n} X_{i2}^2\right)b_2 = \sum_{i=1}^{n} X_{i2}Y_i$$

การเลือกสมการถดถอยที่ดีที่สุด

การวิจัยที่ใช้วิธีการพยากรณ์ โดยพื้นฐานแล้วนักวิจัยไม่มีจุดมุ่งหมายในการทดสอบสมมติฐานในการเปรียบเทียบว่า ตัวแปรพยากรณ์ตัวใดมีความสำคัญต่อการเปลี่ยนแปลงของตัวแปรเกณฑ์มากกว่ากัน ความสำคัญของการวิจัยประเภทนี้ มักจะอยู่ที่การค้นหาตัวแปรพยากรณ์ที่สามารถพยากรณ์ตัวแปรเกณฑ์ที่สนใจได้ถูกต้องแม่นยำที่สุดเท่าที่ความรู้เกี่ยวกับตัวพยากรณ์จะมีอยู่ ดังนั้นหน้าที่สำคัญของนักวิจัยก็คือ การค้นหาสมการหรือประมาณค่าสัมประสิทธิ์การถดถอยของตัวแปรในสมการพยากรณ์ เพื่อให้มีความคลาดเคลื่อนในการพยากรณ์ต่ำที่สุด

จากสมการ (2.17) ค่าสัมประสิทธิ์การถดถอยที่ได้ เป็นค่าแสดงการเปลี่ยนแปลงค่าเฉลี่ยของ Y เมื่อ X_i เปลี่ยนไป 1 หน่วย ขณะที่ตัวแปรพยากรณ์อื่นๆ คงที่ และค่าสัมประสิทธิ์การถดถอยของคะแนนดิบ (Unstandardized Coefficience) นี้ เป็นค่าซึ่งใช้ในการประมาณค่า Y เท่านั้น ถ้าต้องการเปรียบเทียบความสำคัญของตัวแปรพยากรณ์ที่มีต่อตัวแปรเกณฑ์ จะทำได้โดยการแปลงค่าสัมประสิทธิ์การถดถอยคะแนนดิบ (b_i) ให้เป็นสัมประสิทธิ์คะแนนมาตรฐาน (Standardized Coefficience)

โดย $$\beta = b_i \frac{S_{xi}}{S_y} \qquad (2.18)$$

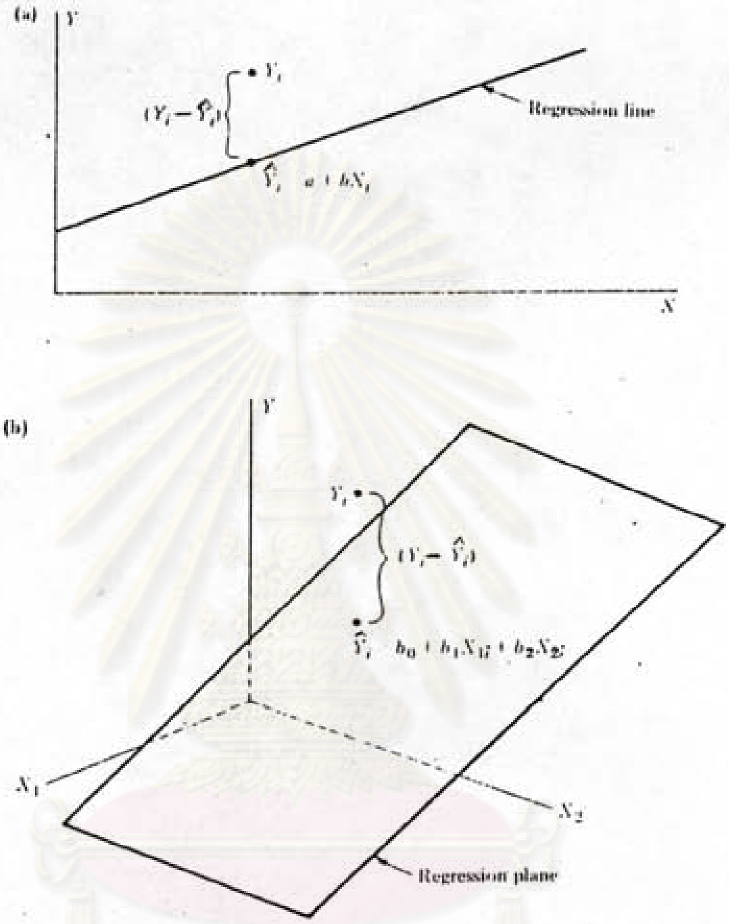
เมื่อ β_i = ค่าสัมประสิทธิ์คะแนนมาตรฐาน (Standardized Beta Weight)

S_{xi} = ส่วนเบี่ยงเบนมาตรฐานของ X_i

S_y = ส่วนเบี่ยงเบนมาตรฐานของ Y

จากข้อมูลกลุ่มตัวอย่าง สามารถพิจารณาลักษณะความแปรปรวนของตัวแปรเกณฑ์ (Y) จากค่าเฉลี่ย Y ได้ดังรูป (Lindeman 1980: 100)

<u>แผนภาพที่ 6</u> ลักษณะความแปรปรวนของตัวแปรเกณฑ์ Y จากค่าเฉลี่ย Y



Actual (Y_i) and Predicted ($\hat{Y}_i$) values and their discrepancy are shown in each case.

จากภาพ (a) เป็นการแสดงถึงเส้นการถดถอย (Regression line) ของตัวแปรเกณฑ์ (Y) และตัวแปรพยากรณ์ (X) เพียง 1 ตัวเท่านั้น และเส้นถดถอยจะตัดแกน Y ที่จุด (0,a) หรือค่า Y-intercept เท่ากับ a นั่นคือ ถ้าสัมประสิทธิ์การถดถอย (b) เท่ากับ 0 จะได้ค่าของ $\hat{Y}$ เท่ากับ a นั่นเอง

จากภาพ (b) เป็นการแสดงถึงระนาบการถดถอย (Regression plane) ของตัวแปรเกณฑ์ (Y) และตัวแปรพยากรณ์ (X) ตั้งแต่ 2 ตัวขึ้นไป นั่นคือค่าของ Y เปลี่ยนไปเนื่องจากการเปลี่ยนแปลงของตัวแปรพยากรณ์ (X) หลายตัวนั่นเอง

ซึ่งเมื่อมองภาพรวมของภาพ (a) และ (b) แล้วจะเห็นว่าความแปรปรวนทั้งหมดประกอบด้วยความแปรปรวน 2 ส่วน ส่วนแรกคือส่วนที่ตัวแปรเกณฑ์ (Y_i) แตกต่างจากค่าประมาณที่ได้จากเส้นถดถอยหรือระนาบการถดถอย (Y_i) เรียกว่า ความแปรปรวนที่ไม่สามารถอธิบายได้ (Unexplained variation) ส่วนที่สอง คือ ส่วนที่ตัวแปรเกณฑ์ที่ประมาณค่าได้จากการประมาณค่าจากเส้นถดถอยหรือระนาบการถดถอย (Y_i) แตกต่างจากค่าเฉลี่ยของตัวแปรเกณฑ์ ซึ่งเรียกว่าความแปรปรวนที่สามารถอธิบายได้ (Explained variation) นั่นคือ

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

เมื่อนำ (2.19) มายกกำลังสองจะได้

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2 + 2\sum_{i=1}^{n}(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y}) \qquad (2.20)$$

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

ให้ $$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \qquad (2.22)$$

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

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

ซึ่งสามารถสรุปเป็นตารางวิเคราะห์ความแปรปรวน (Analysis of Variance) ดังตารางที่ 1

ตารางที่ 1 แสดงแหล่งความแปรปรวนในการวิเคราะห์การถดถอยพหุคูณ

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

ดังนั้นจะได้

$$\hat{Y} = a + b_1X_1 + b_2X_2 + \ldots\ldots + b_pX_p \quad (2.25)$$

นำสัมประสิทธิ์การถดถอย $b_1, b_2, \ldots\ldots b_p$ ที่ได้จากสมการ (2.5) มาทดสอบความมีนัยสำคัญ โดยทดสอบสมมติฐาน

$$H_o : \beta_1 = \beta_2 = \ldots\ldots = \beta_p = 0 \quad (2.26)$$

ค่าสถิติที่ใช้ในการทดสอบคือ

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

ดังนั้นเมื่อสร้างสมการพยากรณ์ได้แล้ว ก่อนที่จะมีการนำเอาสมการไปใช้ ต้องคำนึงถึงว่าสมการนั้นน่าเชื่อถือหรือไม่ เกณฑ์อันหนึ่งที่นิยมใช้กันมากในการตัดสินใจเกี่ยวกับการศึกษาเรื่องการถดถอยเชิงเส้น คือ สัมประสิทธิ์การตัดสินใจ (Coeffficient of Determination)

$$R^2 = \frac{SSR}{SST} \tag{2.28}$$

R^2 นี้ เรียกว่า สัมประสิทธิ์การตัดสินใจ ซึ่งจะมีค่าอยู่ระหว่าง 0 กับ 1

$R^2 \times 100$ หมายถึง ร้อยละของความแปรปรวนทั้งหมดของค่าที่สังเกตได้ (Y_i) ที่ถูกอธิบายได้โดยสมการพยากรณ์ หรืออาจกล่าวได้อีกว่า R^2 ก็คือ Goodness of Fit ของพื้นผิวของระนาบการถดถอยนั่นเอง

ในการวิเคราะห์การถดถอยพหุคูณ ค่า R^2 ที่สูงขึ้นย่อมเป็นสิ่งที่ต้องการ เพราะนั่นหมายถึงว่าตัวแปรพยากรณ์ (X_j) สามารถใช้พยากรณ์ตัวแปรเกณฑ์ (Y) ได้ดีขึ้น อย่างไรก็ตาม การคัดเลือกสมการพยากรณ์ด้วยวิธีนี้ก็มีข้อบกพร่อง (Herzberg 1967: 1) เนื่องจากในค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณ (R) ซึ่งเป็นดัชนีที่ชี้ให้เห็นถึงระดับความสัมพันธ์พหุคูณระหว่างตัวแปรเกณฑ์กับผลรวมเชิงเส้นตรงของตัวแปรพยากรณ์นี้ จัดเป็นตัวประมาณค่าที่เอนเอียงของพารามิเตอร์ (ρ) (Murihead 1982: 179) และมักจะมีค่าสูงกว่าความเป็นจริงเสมอ ทำให้เกิดปัญหาการลดลงของค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสอง (Shrinkage) เมื่อนำเอาสมการพยากรณ์ที่สร้างจากกลุ่มตัวอย่างหนึ่งไปใช้กับอีกกลุ่มตัวอย่างหนึ่งที่สุ่มมาจากประชากรเดียวกัน (Pedhazur 1982: 147-153) เนื่องจากในการคำนวณค่าสัมประสิทธิ์การถดถอย (b) เพื่อให้ได้สมการพยากรณ์ที่มีค่าสหสัมพันธ์พหุคูณสูงสุด และมีความคลาดเคลื่อนในการพยากรณ์ต่ำสุดนั้น

ถือว่าค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองทุกตัวมีความคลาดเคลื่อนเป็นอิสระต่อกัน (Error Free) ซึ่งในความเป็นจริงไม่ได้เป็นเช่นนั้น จึงทำให้ค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองที่คำนวณได้เป็นค่าที่ไม่ถูกต้องตามความเป็นจริงนัก เหตุที่ทำให้ค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณมีค่าสูงกว่าปกติคือ ขนาดของกลุ่มตัวอย่าง ซึ่งพบว่า ถ้าหากว่ากลุ่มตัวอย่างที่มีขนาดเล็กแล้ว จะทำให้ค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสองมีค่าสูงกว่าความเป็นจริงมาก

<u>การแจกแจงแบบปกติหลายตัวแปร</u>

(Multivariate Normal Distribution)

เมื่อ X_{ij} เป็นตัวแปรสุ่ม (random variable)

$$X = \begin{bmatrix} X_{11} & \cdots\cdots\cdots & X_{1p} \\ . & & . \\ . & & . \\ X_{n1} & \cdots\cdots\cdots & X_{np} \end{bmatrix}_{n \times p}$$

X_{ij} จะมีการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution) เมื่อมี p.d.f. (Probability Density Function) ดังนี้ (Morrison 1967: 85)

$$F_X(X) = \frac{1}{(2\pi)^{\frac{1}{2}p}\ |\Sigma|^{1/2}} \exp\left[-\ 1/2\ (X_i - \mu)'\Sigma^{-1}(X_i - \mu)\right] \quad (2.29)$$

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

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

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

$$\Sigma = \begin{bmatrix} \sigma_{ii} & \cdots\cdots\cdots & \sigma_{ip} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \sigma_{ni} & \cdots\cdots\cdots & \sigma_{np} \end{bmatrix}$$

Σ^{-1} เป็น pxp positive definite และสมมาตร (Symmetric) ซึ่งคือเมตริกซ์ความแปรปรวนร่วม (Variance - Covariance Matrix)

<u>การแจกแจงของค่าสหสัมพันธ์พหุคูณ</u>

(Distribution of the Multiple Correlation)

ค่าสหสัมพันธ์พหุคูณ ($\rho_{y.12..p}$) หมายถึงความสัมพันธ์เชิงเส้นระหว่างตัวแปรเกณฑ์ (Y) กับผลรวมเชิงเส้นของตัวแปรพยากรณ์ (Xs) (Lindeman 1982 : 108) หรือกล่าวได้ว่าเป็นสัมประสิทธิ์สหสัมพันธ์โพรดัคโมเมนต์ (Product moment) ของตัวแปรเกณฑ์ที่สังเกตได้ (Y) กับตัวแปรเกณฑ์ที่พยากรณ์ได้จากสมการถดถอย ($\hat{Y}$)

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

เมื่อศึกษากับกลุ่มตัวอย่างสามารถประมาณค่าโดย

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

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

เมื่อ $r_{y.i}$ คือ สัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรเกณฑ์ (Y) กับตัวแปรพยากรณ์ (X_i) แต่ละตัว และ β_i คือ สัมประสิทธิ์การถดถอยมาตรฐานที่จะทำให้ค่าสหสัมพันธ์พหุคูณมีค่าสูงสุด ซึ่งมีค่าอยู่ระหว่าง 0 ถึง 1 และจะมีค่าเพิ่มขึ้นเมื่อจำนวนตัวแปรพยากรณ์เพิ่มขึ้น

การประมาณค่าสหสัมพันธ์พหุคูณของประชากรจะมีลักษณะเช่นเดียวกับในการศึกษาสหสัมพันธ์อย่างง่าย โดยคาดว่าค่าสหสัมพันธ์พหุคูณที่คำนวณได้จากกลุ่มตัวอย่าง ($R_{y.12...p}$) จะกระจายอยู่รอบๆ ค่าสหสัมพันธ์ของประชากร แต่เนื่องจากค่าสหสัมพันธ์พหุคูณนี้มีค่าอยู่ระหว่าง 0 ถึง 1 และจัดเป็นตัวประมาณค่าที่เอนเอียง จึงทำให้การแจกแจงมีความเอนเอียงไปทางบวกเสมอ ซึ่งไม่สามารถคาดคะเนลักษณะการแจกแจงที่แน่นอนได้ จากการศึกษาพบว่า ลักษณะการแจกแจงของค่าสหสัมพันธ์พหุคูณนี้ จะขึ้นอยู่กับอิทธิพลของขนาดของกลุ่มตัวอย่าง จำนวนตัวแปรพยากรณ์ และระดับความสัมพันธ์ในประชากร (ρ) (Muirhead 1982: 171) ดังนี้

เมื่อ $\rho = 0$

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

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

จาก (2.32) และ (2.34) เมื่อระดับความสัมพันธ์ของประชากรมีค่าเป็นศูนย์หรือไม่มีความสัมพันธ์กันเลย ลักษณะการแจกแจงของค่าสหสัมพันธ์พหุคูณจะขึ้นอยู่กับขนาดของกลุ่มตัวอย่างและจำนวนตัวแปรพยากรณ์เท่านั้น เมื่อจำนวนตัวแปรพยากรณ์เพิ่มขึ้นจะทำให้ค่าสหสัมพันธ์พหุคูณสูงขึ้นด้วย ($p \longrightarrow n : R \longrightarrow 1$) แต่ถ้าระดับความสัมพันธ์ในประชากรไม่เท่ากับศูนย์ ($\rho \neq 0$) แล้ว การคาดคะเนลักษณะการแจกแจงของค่าสหสัมพันธ์จะทำได้ยากมาก ดังนี้

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) เป็นฟังก์ชั่นไฮเปอร์จีออเมตริกซ์ (Hypergeometric) ซึ่งถ้าใช้เพียงสองเทอมแรกของสมการจะได้

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

และ

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

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

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

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

และ

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

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

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

จะเห็นว่าลักษณะการแจกแจงของค่าสหสัมพันธ์พหุคูณนอกจากจะขึ้นอยู่กับขนาดของกลุ่มตัวอย่าง และจำนวนตัวแปรพยากรณ์แล้ว ยังขึ้นอยู่กับระดับความสัมพันธ์ในประชากร ซึ่งไม่ทราบค่าอีกด้วย

งานวิจัยที่เกี่ยวข้อง

กรรณิกา เลียงเจริญสิทธิ์ (2527: 48-49) ได้ใช้เทคนิคมอนติคาร์โลซิมูเลชั่น ทำการศึกษาการแจกแจงของค่าสหสัมพันธ์แบบปกติสองตัวแปร ณ ระดับความสัมพันธ์ในประชากร (ρ) ต่าง ๆ ตั้งแต่ ρ = 0.1, 0.2,....., 0.9 เพื่อนำไปใช้ประโยชน์กรณีที่ต้องการสุ่มตัวอย่างที่มีคุณสมบัติตามต้องการ ผลการศึกษาพบว่า ลักษณะการแจกแจงของข้อมูลว่ามีความเบ้ ในกรณีที่ ρ = 0 แล้วขนาดของกลุ่มตัวอย่างมีน้อยกว่า 25 แต่เมื่อขนาดของกลุ่มตัวอย่างมีค่าเท่ากับหรือมากกว่า 25 การแจกแจงของค่าสหสัมพันธ์จะมีลักษณะเป็นปกติโดยประมาณ และยืนยันว่า เมื่อแปลงค่าสหสัมพันธ์โดยวิธี Fisher's Z-transformation แล้ว Z_r จะมีลักษณะการแจกแจงเป็นปกติโดยประมาณ ข้อสรุปที่สำคัญที่ได้จากการศึกษา คือ ในการทดสอบสมมติฐานกรณีที่ ρ มีค่าอื่นๆ ที่ไม่เท่ากับ 0 ณ ระดับ α = 0.01 ขนาดของกลุ่มตัวอย่างที่เหมาะสมควรใช้ตั้งแต่ 9 ขึ้นไปที่ระดับ α = 0.05 และที่ระดับ α = 0.10 ควรใช้ตั้งแต่ 5 ขึ้นไป

ฮาลินสกีและเฟลด์ (Halinske and Feldt 1970: 151-158) ได้ทำการศึกษาโดยใช้เทคนิคมอนติคาร์โลซิมูลชั่น เกี่ยวกับขนาดของกลุ่มตัวอย่างที่เหมาะสมในการวิเคราะห์การถดถอยพหุคูณ พบว่า ควรใช้อัตราส่วนระหว่างขนาดของกลุ่มตัวอย่างกับจำนวนตัวแปรอย่างน้อยที่สุดเท่ากับ 10 : 1

มิลเลอร์และคันซ์ (Miller and Kunce 1978: 157-163) ได้ทำการศึกษาแบบครอสแวลิเดชั่น (Cross-Validation) เกี่ยวกับขนาดของกลุ่มตัวอย่างที่เหมาะสมในการ

วิเคราะห์การถดถอยพหุคูณ พบว่าควรใช้อัตราส่วนระหว่างขนาดของกลุ่มตัวอย่างกับจำนวนตัวแปรอย่างน้อยที่สุดเท่ากับ 10 : 1

วันชัย นันคะเง้น (2532:105) ได้ใช้เทคนิคมอนติคาร์โลซิมูเลชั่นทำการศึกษาหาข้อสรุปเกี่ยวกับการแจกแจงและเปรียบเทียบค่าเฉลี่ยและความแปรปรวนของค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณยกกำลังสอง (R^2) ที่ยังไม่ได้ปรับแก้ และที่ปรับแก้ตามวิธีของเวอร์รี่ และวิธีของโอลกินกับแพรตต์ ทำการทดลองสถานการณ์ต่างๆ ในลักษณะที่ประชากรมีการแจกแจงแบบปกติหลายตัวแปร ทำการทดลองสถานการณ์ต่างๆ ในลักษณะที่ประชากรมีการแจกแจงแบบปกติและหลายตัวแปร ซึ่งมีค่า ρ = 0.20, 0.40, 0.60 และ 0.80 มีจำนวนตัวแปรพยากรณ์เท่ากับ 3, 5, 7 และ 9 ตัว ใช้กลุ่มตัวอย่างขนาด 2, 5, 10, 15, 20, 25 และ 30 เท่าของตัวแปรทั้งหมด พบว่า ควรจะใช้กลุ่มตัวอย่างประมาณ 20 เท่าของตัวแปรขึ้นไป หรืออย่างน้อยที่สุดไม่ควรต่ำกว่า 15 เท่าของตัวแปร จึงจะทำให้ได้ค่า R^2 ที่ใกล้เคียงกับ ρ^2