บทที่ 2

เอกสารและผลงานวิจัยที่เกี่ยวข้อง

ในบทนี้ ผู้วิจัยได้กำหนดเนื้อหาออกเป็น 6 ตอน ซึ่งในแต่ละตอนประกอบด้วยหัวข้อต่าง ๆ ดังนี้

ตอนที่ 1 สหสัมพันธ์และความถดถอย สหสัมพันธ์อย่างง่าย การวิเคราะห์การถดถอยอย่างง่าย และค่าประมาณของพารามิเตอร์ β_0 และ β

ตอนที่ 2 การวิเคราะห์สหสัมพันธ์และการถดถอยพหุคูณ และการประมาณค่าพารามิเตอร์ในการวิเคราะห์การถดถอย

ตอนที่ 3 การเลือกสมการถดถอยที่ดีที่สุด และวิธีคัดเลือกตัวแปรเพื่อการพยากรณ์

ตอนที่ 4 การซิมูเลชั่น (Simulation) และการสร้างเลขสุ่ม (Random Number)

ตอนที่ 5 การแจกแจงแบบปกติหลายตัวแปร และ การแจกแจงค่าสหสัมพันธ์พหุคูณกำลังสอง

ตอนที่ 6 งานวิจัยที่เกี่ยวข้อง

ซึ่งแต่ละหัวข้อมีรายละเอียดต่างๆ ดังนี้

<u>**ตอนที่ 1**</u>

<u>สหสัมพันธ์และความถดถอย (Correlation and Regression)</u>

การวิเคราะห์สหสัมพันธ์ (Correlation Analysis) เป็นระเบียบวิธีที่ใช้วัดความสัมพันธ์ระหว่างตัวแปรว่ามีความสัมพันธ์กันมากน้อยเพียงใดโดยไม่คำนึงถึงว่าตัวแปรใดจะเป็นตัวแปรอิสระ และตัวแปรใดจะเป็นตัวแปรตาม

ส่วนการวิเคราะห์ความถดถอย (Regression Analysis) เป็นวิธีการที่ได้ชื่อมาจาก Sir Francis Galton เมื่อปี ค.ศ.1899 ซึ่งเป็นการศึกษาในเชิงเปรียบเทียบเกี่ยวกับแนวโน้มของลักษณะความสูงของบุตรหลานที่มีบิดามารดาสูงและเตี้ย ปรากฏว่าความสูงของบุตรหลานจะถดถอยเข้าหาส่วนสูงเฉลี่ยของประชากร ดังนั้นการวิเคราะห์ถดถอยจึงเป็นวิธีการที่ใช้ในการพิจารณาถึงรูปแบบที่เป็นไปได้ของความสัมพันธ์ระหว่างตัวแปร มีวัตถุประสงค์เพื่อใช้เป็นประโยชน์ในการพยากรณ์ (Predict) หรือการประมาณ (Estimate) ค่าของตัวแปรตัวหนึ่งซึ่งเรียกว่าตัวแปรตามหรือตัวแปรเกณฑ์ (Dependent Variable) โดยใช้ค่าของตัวแปร

อีกตัวหนึ่งหรืออีกชุดหนึ่งที่กำหนดให้ ซึ่งเรียกว่า ตัวแปรอิสระหรือตัวแปรทำนาย (Independent Variable) เป็นตัวประมาณ

การวิเคราะห์สหสัมพันธ์และการถดถอยนี้ ถ้าเป็นการวิเคราะห์เกี่ยวกับตัวแปรเพียงสองตัว (Bivariate) เรียกว่า สหสัมพันธ์อย่างง่าย (Simple Correlation) ถ้าประกอบด้วยตัวแปรเกณฑ์ 1 ตัว และตัวแปรทำนายอีก 1 ตัว เรียกว่า การถดถอยอย่างง่าย (Simple Regression) ส่วนสหสัมพันธ์พหุคูณ (Multiple Correlation) หมายถึง การวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรมากกว่า 2 ตัวขึ้นไป แต่ถ้าในการวิเคราะห์นั้นเป็นการพยากรณ์ตัวแปรเกณฑ์ตัวหนึ่งด้วยตัวแปรทำนายมากกว่า 1 ตัวขึ้นไป เราเรียกว่า การถดถอยพหุคูณ (Multiple Regression)

<u>สหสัมพันธ์อย่างง่าย (Simple Correlation)</u>

สหสัมพันธ์อย่างง่าย หมายถึง ความสัมพันธ์ระหว่างตัวแปรใด ๆ 2 ตัว โดยไม่คำนึงถึงว่าตัวแปรใดเป็นตัวแปรอิสระหรือตัวแปรตาม เพราะตัวแปรทั้งสองนี้เป็นตัวแปรสุ่ม ซึ่งได้มาจากการสุ่มกลุ่มตัวอย่าง และวัดค่าของตัวแปรทั้งสองพร้อม ๆ กัน และนำค่าที่ได้มาหาความสัมพันธ์ ค่าที่ใช้วัดความสัมพันธ์นี้เรียกว่า สัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) ใช้สัญลักษณ์เป็นตัวอักษรกรีก คือ ρ (อ่านว่า โร) ถ้ามีตัวแปรเพียง 2 ตัว เรียกว่า สัมประสิทธิ์สหสัมพันธ์อย่างง่าย (Simple Correlation Coefficient) ซึ่งค่าดังกล่าวเป็นค่าที่ได้จากการเปรียบเทียบความแปรปรวน ร่วม (Covariance) ระหว่างตัวแปร 2 ตัว (x, y) กับผลคูณของความเบี่ยงเบนมาตรฐาน (Standard Deviation) ของ x และ y ดังนี้

$$\rho = \rho_{xy} = \rho_{yx} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \qquad (2.1)$$

018221

$$\rho_{xy} = \frac{\sum_{i=1}^{n}(X_i - \mu_x)(Y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(X_i - \mu_x)^2 \sum_{i=1}^{n}(Y_i - \mu_y)^2}} \qquad (2.2)$$

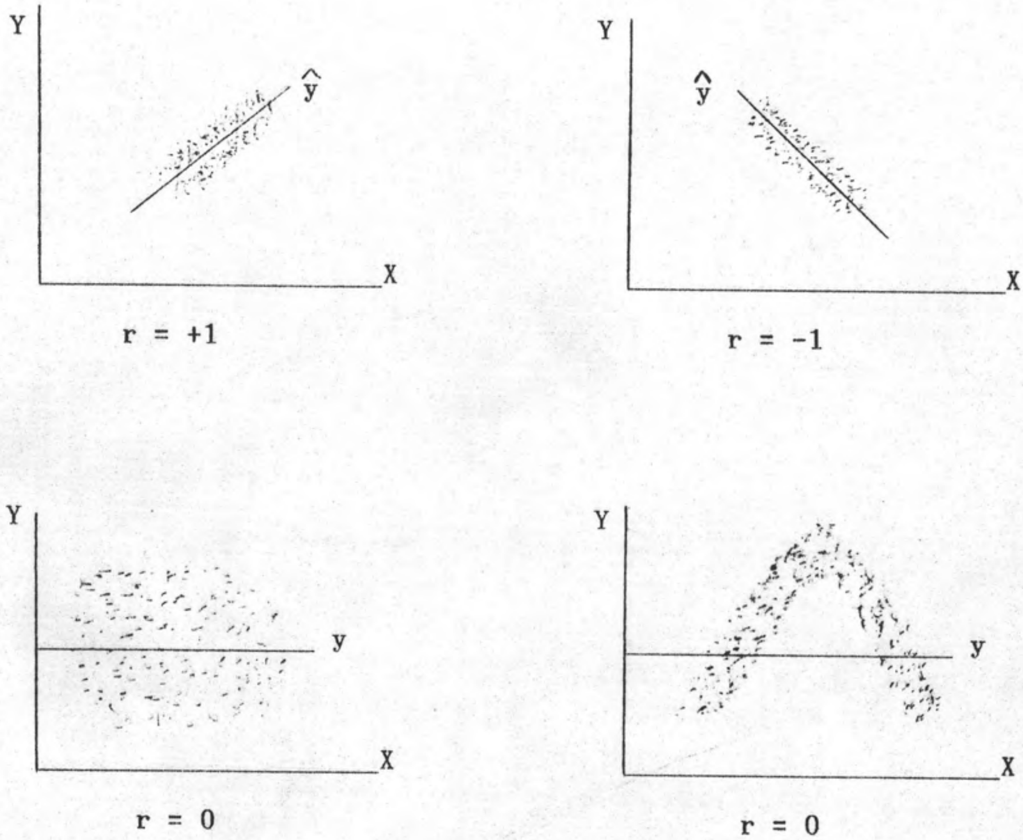
เมื่อ n คือจำนวนข้อมูลทั้งหมด และสัมประสิทธิ์สหสัมพันธ์อย่างง่าย (ρ) มีค่าอยู่ระหว่าง -1 หรือ +1 เป็นค่าที่ไม่มีหน่วย แต่จะบอกถึงระดับความสัมพันธ์ระหว่างตัวแปรว่ามีมากน้อยเพียงใด ในการคำนวณค่า ρ นี้จะต้องทราบข้อมูลทั้งหมดในประชากร ซึ่งในทางปฏิบัติแล้ว เราไม่อาจทราบค่าเหล่านี้ได้ จึงประมาณค่า ρ ด้วยค่าสหสัมพันธ์ (r) จากข้อมูลกลุ่มตัวอย่างดังนี้

$$\rho = \rho_{xy} = r = r_{xy} = r_{yx} \qquad (2.3)$$

$$r_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \qquad (2.4)$$

ซึ่งค่าสหสัมพันธ์ (r) นี้ อาจจะมีความสัมพันธ์ไปในทางเดียวกัน หรือตรงกันข้ามก็ได้ ถ้า x และ y มีความสัมพันธ์ไปในทางเดียวกันอย่างสมบูรณ์ จะให้ค่า r = +1 ถ้า x และ y มีความสัมพันธ์กันอย่างสมบูรณ์ แต่เป็นไปในทางตรงกันข้าม จะให้ค่า r = -1 และถ้า x, y มีการกระจายมาก จะให้ค่า r = 0 ตามแผนภาพ ดังนี้

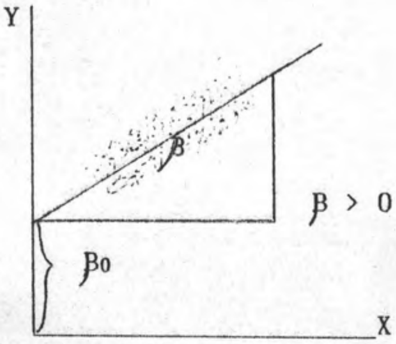


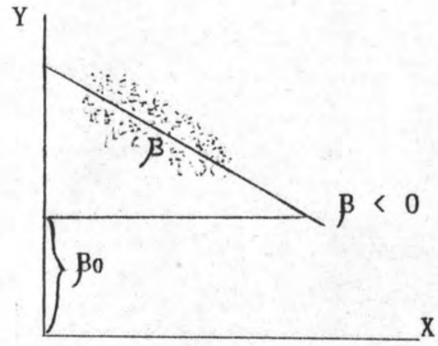
แผนภาพที่ 2 แสดงการกระจายสหสัมพันธ์ระหว่าง x และ y

<u>**การวิเคราะห์การถดถอยอย่างง่าย**</u> (Simple Regression Analysis)

ในกรณีที่มีตัวแปรเพียง 2 ตัว คือ ตัวแปรเกณฑ์ (y) และตัวแปรทำนาย (x) นั้น สามารถดูลักษณะการถดถอยของ y ที่มีต่อ x ได้จากแผนภาพการกระจาย (Scatter Diagram) ซึ่งมีลักษณะดังนี้

แผนภาพที่ 3 แสดงการถดถอยเชิงเส้นตรงของตัวแปร x และ y





ความสัมพันธ์ระหว่าง x กับ y
เป็นเส้นตรง และไปในทางเดียวกัน
x เพิ่ม y เพิ่ม
x ลด y ลด

ความสัมพันธ์ระหว่าง x กับ y
เป็นเส้นตรง แต่เป็นไปในทางตรงกันข้าม
x เพิ่ม y ลด
x ลด y เพิ่ม

จากแผนภาพที่ 3 แสดงให้เห็นการถดถอยเชิงเส้นตรงอย่างง่ายของตัวแปรเกณฑ์ y ที่มีตัวแปรทำนาย x เพียงตัวเดียว ซึ่งสามารถแสดงความสัมพันธ์ในรูปของตัวแบบสมการทางคณิตศาสตร์ ดังนี้

$$Y_i = \beta_0 + \beta X_i + \varepsilon_i \qquad (2.5)$$

เมื่อ β_0 คือ ค่า Y-intercept คือค่าที่เส้นตรงตัดแกน Y หรือค่า y เมื่อ x = 0

β ค่าความชัน (Slope) ของเส้นตรง คือค่าที่แสดงให้ทราบว่า เมื่อ x มีค่าเปลี่ยนไป 1 หน่วย y จะเปลี่ยนไปโดยเฉลี่ยเท่าไร เรียกค่านี้ว่า สัมประสิทธิ์ความถดถอย (Regression Coefficient) ซึ่งจะมีค่าเป็นบวก เมื่อ x และ y มีความสัมพันธ์ไปในทางเดียวกัน

และมีค่าเป็นลบ เมื่อ x และ y มีความสัมพันธ์ไปในทางตรงกันข้าม และมีค่าเป็น 0 เมื่อการเปลี่ยนแปลงของ y ไม่ขึ้นกับการเปลี่ยนแปลงของ x

ε_i คือ ค่าความคลาดเคลื่อนที่มีลักษณะเป็นตัวแปรสุ่ม ซึ่งเป็นอิสระจาก x และ y ภายใต้ข้อสมมุติดังนี้

1. ค่า X_i ต้องเป็นค่าที่วัดได้โดยไม่มีความผิดพลาด และเป็นค่าที่กำหนดให้คงที่
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

$$E(Y_i) = E(Y/X_i) = \mu_{y.xi} = \beta_0 + \beta X_i \qquad (2.6)$$

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 ซึ่งจะได้ค่าประมาณของ Y_i ดังนี้

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta} X_i = b_0 + bX_i = \hat{\mu}_{Yi.xi} \qquad (2.7)$$

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

$$\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \sum_{i=1}^{n}(Y_i - \hat{Y})^2 = \sum_{i=1}^{n}(Y_i - b_0 - bX_i)^2 \qquad (2.8)$$

$$\frac{\partial}{\partial \beta_0}\left(\sum_{i=1}^{n}\varepsilon_i^2\right) = -2\sum_{i=1}^{n}(Y_i - b_0 - bX_i) = 0$$

$$= -2\sum_{i=1}^{n}(Y_i - b_0 - bX_i)(X_i) = 0 \qquad (2.9)$$

จากผลข้างต้นนี้ จะทำให้ได้ชุดสมการปกติ (Normal equations) ดังนี้

$$nb_0 + b\sum_{i=1}^{n}X_i = \sum_{i=1}^{n}Y_i \qquad (2.10)$$

$$b_0\sum_{i=1}^{n}X_i + b\sum_{i=1}^{n}X_i^2 = \sum_{i=1}^{n}X_iY_i \qquad (2.11)$$

ซึ่งให้ค่าประมาณของ β_0 และ β ดังนี้

$$\hat{\beta}_0 = b_0 = \bar{y} - b\bar{x} \qquad (2.12)$$

$$\hat{\beta} = b = \frac{\sum_{i=1}^{n}(X_i-\bar{x})(Y_i-\bar{y})}{\sum_{i=1}^{n}(X_i-\bar{x})^2} = \frac{\sum_{i=1}^{n}(X_i-\bar{x})Y_i}{\sum_{i=1}^{n}(X_i-\bar{x})^2} \qquad (2.13)$$

ตอนที่ 2

การวิเคราะห์สหสัมพันธ์และการถดถอยพหุคูณ

(Multiple correlation and Regression analysis)

การวิเคราะห์สหสัมพันธ์และการถดถอยพหุคูณ เป็นแนวคิดและเทคนิคที่ขยายมาจากการวิเคราะห์สหสัมพันธ์และการถดถอยอย่างง่าย (Simple Correlation and Regression) เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรตาม (Dependent Variable) และกลุ่มของตัวแปรอิสระ (Set of Independent Variables) ภายใต้รูปแบบที่กำหนด ซึ่งอาจเป็นเส้นตรงหรือเส้นโค้ง จากการวิเคราะห์การถดถอยจะได้สมการถดถอยพหุคูณ ซึ่งโดยปกติแล้วจะร่วมกันอธิบายตัวแปรเกณฑ์ได้มากกว่า เพราะในการวิเคราะห์ปัญหาบางอย่างที่จำเป็นต้องใช้การถดถอยนั้น บางครั้งการศึกษาการถดถอยอย่างง่ายอาจจะไม่เพียงพอ ทั้งนี้เพราะการประมาณค่าของตัวแปรเกณฑ์เพื่อให้ใกล้เคียงที่สุดนั้น เรามักจะต้องพิจารณาตัวแปรพยากรณ์ที่มีอิทธิพลหรือมีความสัมพันธ์ต่อตัวแปรเกณฑ์มากกว่า 1 ตัวขึ้นไป โดยมีสมการถดถอยเป็นตัวชี้ให้เห็นถึงความสัมพันธ์เฉลี่ยของตัวแปรเหล่านั้นดังนี้ (Lindeman 1980: 94)

$$Y_i = b_0 + b_1X_1 + b_2X_2 + \ldots\ldots b_pX_p + e_1 \qquad (2.14)$$

โดยที่ Y_i คือ ตัวแปรเกณฑ์ (Criterion variables or Dependent variable)

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

b_0 คือ ค่าคงที่ ซึ่งจะมีค่าเท่ากับ Y เมื่อ X_i ทั้งหมดมีค่าเท่ากับศูนย์ ค่า b_0 เป็นค่าประมาณของ β_0

b_i คือ สัมประสิทธิการถดถอยของประชากรบางส่วน (Population Partial Regression Coefficient) ของ X_i เมื่อให้ $X_i = X_1 \ldots . X_p$ เป็นค่าคงที่ นั่นคือ เมื่อ X_i มีค่าเปลี่ยนแปลงไป 1 หน่วย จะมีผลทำให้ Y เปลี่ยนแปลงไป b_i หน่วย เมื่อตัวแปรพยากรณ์อื่น ๆ คงที่ และเป็นค่าประมาณของ β_i

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

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

$$E(\varepsilon_i \varepsilon_j) = \sigma^2 \quad ; i = j$$

$$= 0 \quad ; i = j$$

จากสมการ (2.14) สามารถแสดงในรูปของเมตริกซ์ได้ดังนี้

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

เมื่อ

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{bmatrix}_{nx1} \qquad X = \begin{bmatrix} 1 & X_{11} & X_{12} & \ldots\ldots\ldots & X_{1p} \\ 1 & X_{21} & X_{22} & \ldots\ldots\ldots & X_{2p} \\ . & . & . & & . \\ . & . & . & & . \\ 1 & X_{n1} & X_{n2} & \ldots\ldots\ldots & X_{np} \end{bmatrix}_{nx(k+1)}$$

$$\beta = \begin{bmatrix} b_0 \\ b_1 \\ . \\ . \\ b_p \end{bmatrix}_{(k+1)x1} \qquad \varepsilon = \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ e_n \end{bmatrix}_{nx1}$$

จากตัวแบบ (2.14) ผลรวมของความคลาดเคลื่อนกำลังสองจะมีค่าเป็น (วิชิต หล่อธีระชุณสกุล 2524: 124)

$$
\begin{aligned}
\sum_{i=1}^{n} e_i^2 &= \underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon} = (\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\beta})'(\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\beta}) \\
&= \underset{\sim}{Y}'\underset{\sim}{Y} - \underset{\sim}{\beta}'\underset{\sim}{X}'\underset{\sim}{Y} - \underset{\sim}{Y}'\underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\beta}'\underset{\sim}{X}'X\beta \\
&= \underset{\sim}{Y}'\underset{\sim}{Y} - 2\underset{\sim}{\beta}'\underset{\sim}{X}'\underset{\sim}{Y} + \underset{\sim}{\beta}'\underset{\sim}{X}'X\beta \qquad (2.15)
\end{aligned}
$$

<u>การประมาณค่าพารามิเตอร์ในการวิเคราะห์การถดถอย</u>

(Estimation of Parameters in Multiple Regression)

เนื่องจากการวิจัยในทางปฏิบัตินั้น ผู้วิจัยจะไม่สามารถศึกษาจากกลุ่มประชากรทั้งหมดได้ จึงไม่ทราบค่าพารามิเตอร์ β ที่แท้จริง โดยทั่วไปจะประมาณค่าพารามิเตอร์ (Parameter) จากข้อมูลกลุ่มตัวอย่างที่นำมาศึกษา วิธีการประมาณตัวประมาณค่าที่นิยมใช้กันมากคือ วิธีกำลังสองน้อยที่สุด (Least Square Estimation Method) ซึ่งจะได้ค่าประมาณสัมประสิทธิ์การถดถอยพหุคูณจากการอนุพันธ์ (differentiate) สมการ (2.15) เทียบกับ β แล้วได้ค่าเท่ากับศูนย์

$$
\begin{aligned}
\frac{\partial(\varepsilon'\varepsilon)}{\partial\beta} &= -2\underset{\sim}{X}'\underset{\sim}{Y} + 2\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{\beta} = 0 \\
\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{\beta} &= \underset{\sim}{X}'\underset{\sim}{Y} \\
\underset{\sim}{\beta} &= (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y} \qquad (2.16)
\end{aligned}
$$

ซึ่งจะได้ตัวประมาณค่าที่ไม่เอนเอียงของ (β) ซึ่งมีค่าเฉลี่ยความคลาดเคลื่อนกำลังสองน้อยที่สุด ในบรรดาตัวประมาณค่าที่ไม่เอนเอียงทั้งหลาย

ตอนที่ 3

การเลือกสมการถดถอยที่ดีที่สุด

การวิจัยที่ใช้วิธีการพยากรณ์นั้น โดยพื้นฐานแล้ว นักวิจัยไม่มีจุดมุ่งหมายในการทดสอบสมมติฐานในการเปรียบเทียบว่า ตัวแปรพยากรณ์ตัวใดมีความสำคัญต่อการเปลี่ยนแปลงของตัวแปรเกณฑ์มากกว่ากัน ความสำคัญของการวิจัยประเภทนี้มักจะอยู่ที่การค้นหาตัวแปรพยากรณ์ที่สามารถพยากรณ์ตัวแปรที่สนใจได้ถูกต้องแม่นยำที่สุดเท่าที่ความรู้เกี่ยวกับตัวพยากรณ์จะมีอยู่ ดังนั้นหน้าที่สำคัญของนักวิจัยก็คือ การค้นหาสมการหรือการประมาณค่าสัมประสิทธิ์การถดถอยของตัวแปรในสมการพยากรณ์ เพื่อให้มีความคลาดเคลื่อนในการพยากรณ์ต่ำสุด

จากสมการ (2.16) ค่าสัมประสิทธิ์การถดถอยที่ได้เป็นค่าการแสดงการเปลี่ยนแปลงค่าเฉลี่ยของ Y เมื่อ X_i เปลี่ยนไป 1 หน่วย ขณะที่ตัวแปรพยากรณ์อื่น ๆ คงที่ และค่าสัมประสิทธิ์การถดถอยของคะแนนดิบ (Unstandardized Coefficience) นี้ เป็นค่าซึ่งใช้ในการประมาณค่า Y เท่านั้น ถ้าต้องการเปรียบเทียบความสำคัญของตัวแปรพยากรณ์ที่มีต่อตัวแปรเกณฑ์จะทำได้โดยการแปลงค่าสัมประสิทธิ์การถดถอยคะแนนดิบ (b_i) ให้เป็นสัมประสิทธิ์คะแนนมาตรฐาน (Standardized Coefficience)

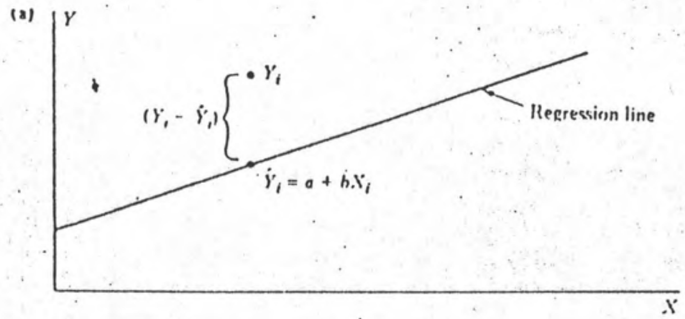
โดย $$\beta_i = b_i \frac{S_{xi}}{S_y} \qquad (2.17)$$

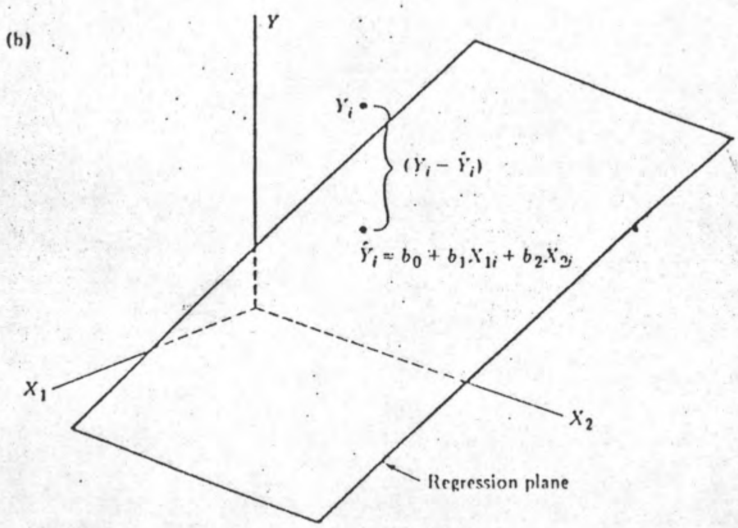
เมื่อ β_i = ค่าสัมประสิทธิ์คะแนนมาตรฐาน (Standardized Beta Weight)

S_{xi} = ส่วนเบี่ยงเบนมาตรฐานของ X_i

S_y = ส่วนเบี่ยงเบนมาตรฐานของ Y

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





ซึ่งแสดงกราฟและระนาบของการถดถอย จะเห็นว่าความแปรปรวนทั้งหมดประกอบด้วยความแปรปรวน 2 ส่วน ส่วนแรกคือส่วนที่ตัวแปรเกณฑ์ (Y_i) แตกต่างจากค่าประมาณที่ได้จากเส้นถดถอยหรือระนาบการถดถอย ($\hat{Y}_i$) เรียกว่า ความแปรปรวนที่ไม่สามารถอธิบายได้ (Unexplained variation) ส่วนที่สอง คือ ส่วนที่ตัวแปรเกณฑ์ที่ประมาณค่าได้จากการประมาณค่าจากเส้นถดถอยหรือระนาบการถดถอย ($\hat{Y}_i$) แตกต่างจากค่าเฉลี่ยของตัวแปรเกณฑ์ ซึ่งเรียกว่าความแปรปรวนที่สามารถอธิบายได้ (Explained variation) นั่นคือ

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) \qquad (2.18)$$

เมื่อนำ (2.18) มายกกำลังสองจะได้

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2 + 2\sum_{i=1}^{n}(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y}_i) \quad (2.19)$$

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

ให้ $SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ (2.21)

$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ (2.22)

และ $SSR = \sum_{i=1}^{n}(\hat{Y}_i- \bar{Y})^2$ (2.23)

ซึ่งสามารถสรุปเป็นตารางวิเคราะห์ความแปรปรวน (Analysis of Variance) ดังตารางที่ 1

ตารางที่ 1 แสดงแหล่งความแปรปรวนในการวิเคราะห์การถดถอยพหุคูณ

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

เราจึงสร้างสมการพยากรณ์ได้ดังนี้

$$\hat{Y} = a + b_1X_1 + b_2X_2 + \ldots\ldots\ldots + b_pX_p \qquad (2.24)$$

นำสัมประสิทธิ์การถดถอย $b_1, b_2, \ldots\ldots b_p$ มาทดสอบความมีนัยสำคัญ โดยทดสอบสมมติฐาน

$$H_0 : \beta_1 = \beta_2 = \ldots\ldots\ldots = \beta_p = 0 \qquad (2.25)$$

ค่าสถิติที่ใช้ในการทดสอบคือ

$$F = \frac{MSR}{MSE} \qquad (2.26)$$

ดังนั้นเมื่อสร้างสมการพยากรณ์ได้แล้ว ก่อนที่จะมีการนำเอาสมการไปใช้ ต้องคำนึงถึงว่าสมการนั้นน่าเชื่อถือหรือไม่ เกณฑ์อันหนึ่งที่นิยมใช้กันมากในการตัดสินใจเกี่ยวกับการศึกษาเรื่องการถดถอยเชิงเส้นคือ สัมประสิทธิ์การทำนาย (Coefficient of Determination)

$$R^2 = \frac{SSR}{SST} \qquad (2.27)$$

R^2 นี้ เรียกว่า สัมประสิทธิ์การทำนาย (Coefficient of Determination) ซึ่งจะมีค่าอยู่ระหว่าง 0 กับ 1

R^2 x 100 หมายถึง ร้อยละของความแปรปรวนทั้งหมดของค่าที่สังเกตได้ (Y_i) ที่ถูกอธิบายได้โดยสมการพยากรณ์ หรืออาจกล่าวได้อีกว่า R^2 ก็คือ Goodness of Fit ของพื้นที่ผิวของระนาบการถดถอยนั่นเอง

ในการวิเคราะห์การถดถอยพหุคูณ ค่า R^2 ที่สูงขึ้นย่อมเป็นสิ่งที่ต้องการ เพราะนั่นหมายถึงว่า ตัวแปรพยากรณ์ (X_s) สามารถใช้พยากรณ์ตัวแปรเกณฑ์ได้ดี

วิธีคัดเลือกตัวแปรเพื่อการพยากรณ์

ในการวิเคราะห์การถดถอยพหุคูณนั้น มักจะมีตัวแปรทำนายหลายตัวที่ใช้ในการพยากรณ์ตัวแปรเกณฑ์ ปัญหาที่ผู้วิจัยต้องการคือจะเลือกตัวแปรทำนายเพียงจำนวนหนึ่งที่น้อยที่สุดและมีประสิทธิภาพมากที่สุด เพื่อให้ตัวแบบถดถอยเชิงเส้นที่ดีหรือเหมาะสมที่สุดที่ใช้ในการพยากรณ์ วิธีคัดเลือกตัวแปรทำนายเข้าสู่สมการถดถอยมีอยู่หลายวิธีด้วยกัน แต่วิธีที่ผู้วิจัยเสนอมีอยู่ 3 วิธี คือ

1. การคัดเลือกตัวแปรแบบไปข้างหน้า (forward Selection)
2. การกำจัดตัวแปรแบบถอยหลัง (Backward Elinination) (Draper and Smith 1966)
3. การถดถอยแบบขั้นบันได (Stepwise Regression) (Efroymson 1960)

ซึ่งทั้ง 3 วิธีนี้เป็นวิธีที่ใช้กันแพร่หลายมากในการคัดเลือกเอาตัวแปรทำนายเข้าสู่ตัวแบบถดถอยเชิงเส้น ซึ่งในแต่ละวิธีจะมีวิธีการคัดเลือกตัวแปรทำนายเป็นขั้นและมีข้อดีข้อเสียดังนี้

1. การเลือกตัวแปรแบบไปข้างหน้า (Forward Selection)

1.1 พิจารณาสัมประสิทธิสหสัมพันธ์บางส่วนระหว่างตัวแปรเกณฑ์กับตัวแปรทำนาย แต่ละตัว (r_{xy}) เลือกตัวแปรทำนายที่ให้ค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนสูงที่สุดสมมุติได้ X_j สมการจะเป็น $\hat{Y} = a + b_j X_j$

1.2 หาค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนระหว่างตัวแปรเกณฑ์กับตัวแปรทำนายแต่ละตัวที่ยังไม่อยู่ในสมการ โดยถือว่าได้รวมตัวแปรทำนาย X_j ไว้ในสมการแล้ว นั่นคือหาค่า

$$r_y\,(l_j) = \frac{r_{yl} - r_{yj}\, r_{lj}}{\sqrt{(1 - r^2{}_{lj})}}$$

เมื่อ $l = 1,2,\ldots\ldots\ldots, j-1, j+1,\ldots\ldots k$

k = จำนวนตัวแปรทำนายทั้งหมดที่จะพิจารณา

1.3 เลือกค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนที่สูงที่สุด สมมุติให้ $r_y(l.j)$ จึงควรรวม X_l ไว้ในสมการเป็นตัวแปรใหม่

1.4 พิจารณา partial F ของตัวแปรทำนายใหม่ X_1 ในข้อ 1.3 ถ้ามีค่าสูงกว่า $F_{\alpha}(1,n-m-1)$ แสดงว่าเป็นการสมควรที่จะรวม X_1 ไว้ในสมการ ในที่นี้ m คือจำนวนตัวแปรทำนายในสมการใหม่ และ n คือจำนวนค่าสังเกต (observation)

1.5 ทำตามข้อ 1.2, 1.3 และ 1.4 อีก โดยถือว่าสมการได้รวมตัวแปรทำนายไว้แล้ว 2 ตัว 3 ตัว ฯลฯ ตามลำดับ จน partial F ที่ได้จากตัวแปรทำนายมีค่าน้อยกว่า $F_{\alpha}(1,n-m-1)$ ตัวแปรทำนายใหม่จึงไม่ควรรวมอยู่ในสมการ ซึ่งแสดงว่าได้สมการที่เหมาะสมที่สุดแล้ว

ในกรณีที่รวมตัวแปรทำนายไว้แล้ว 2 ตัว สมมุติเป็นตัวที่ 1 และ 2 จะหาสัมประสิทธิ์สหสัมพันธ์บางส่วนของ Y และตัวแปรทำนายตัวที่ 3 ได้ดังนี้

$$r_y(3.12) = \frac{r_y(3.1) - r_y(2.1)\, r3(2.1)}{\sqrt{(1-r^2{}_{3(2.1)})}} \qquad (2.29)$$

และในกรณีที่รวมตัวแปรทำนายไว้แล้ว 3 ตัว สัมประสิทธิ์สหสัมพันธ์บางส่วนของ Y และตัวแปรทำนายตัวที่ 4 คือ

$$r_y(4.123) = \frac{r_y(4.12) - r_y(3.12)\, r4(3.12)}{\sqrt{(1-r^2{}_{4(3.12)})}} \qquad (2.30)$$

วิธีนี้ดีกว่าวิธีการกำจัดตัวแปรแบบถอยหลังในแง่ของการประหยัดเวลาในการคำนวณโดยที่ไม่ต้องพิจารณาตัวแปรทำนายทั้งหมดโดยไม่จำเป็น แต่มีข้อเสียตรงที่ไม่ได้มีการพิจารณาบทบาทของตัวแปรทำนายที่รวมอยู่ในสมการในขั้นก่อน เมื่อมีตัวแปรทำนายตัวใหม่เข้าไปอยู่ในสมการ บทบาทในที่นี้หมายถึง ความสามารถในการพยากรณ์ตัวแปรเกณฑ์ ซึ่งอาจจะเปลี่ยนแปลงได้ในแต่ละขั้น เนื่องจากความสัมพันธ์ระหว่างตัวแปรทำนายที่เข้าไปใหม่กับตัวแปรทำนายที่มีอยู่เดิมในสมการ ซึ่งอาจทำให้ไม่ได้สมการที่เหมาะสมที่สุด

2. การกำจัดตัวแปรแบบถอยหลัง (Backward Elimination)

2.1 ตั้งสมการที่รวมตัวแปรทำนายที่ควรจะพิจารณาทั้งหมด สมมุติมี K ตัวสมการจะเป็นดังนี้

$$\hat{Y} = a+b_1X_1+b_2X_2+.........+b_k X_k$$

2.2 คำนวน partial F ของตัวแปรทำนายแต่ละตัว ในจำนวน partial F นี้ ตัวเลือกที่น้อยที่สุด สมมุติได้ F_j แล้วนำ F_j เปรียบเทียบกับค่า F จากตารางคือ $F_{\infty}(1,n-k-1)$ ถ้า F_j น้อยกว่า F จากตารางให้กำจัดตัวแปรทำนาย X_j ออกจากสมการ

2.3 ตั้งสมการใหม่โดยไม่รวม X_j ในสมการ แล้วทำตามข้อ 2.2 อีก โดยที่ค่า K จะเปลี่ยนเป็น K-1 ทำเช่นนี้จนในที่สุด partial F ทุก ๆ ตัว มีค่ามากกว่า F ที่กำหนดจากตาราง

วิธีนี้ดีในแง่ที่ได้พิจารณาตัวแปรทำนายทุกตัวในตอนเริ่มแรก อย่างไรก็ตามยังเสียเวลาในการคำนวนมาก และสำคัญคือไม่ได้นำตัวแปรทำนายเดิมที่ออกจากสมการแล้วกลับมาพิจารณาอีก เมื่อคัดตัวแปรทำนายใหม่ออกจากสมการ

3. การถดถอยแบบขั้นบันได (Stepwise Regression)

วิธีนี้คล้ายกับวิธีการเลือกตัวแปรแบบไปข้างหน้า แตกต่างกันตรงที่ในแต่ละขั้นที่มีการเพิ่มตัวแปรเข้าไปในสมการทีละตัวจะมีการคำนวน partial F ของตัวแปรทำนายอื่น ๆ ที่มีอยู่ในสมการเดิม โดยที่ถือว่าตัวแปรทำนายนั้น ๆ เข้าไปอยู่ในสมการที่รวมตัวแปรใหม่ด้วยเป็นตัวสุดท้าย การที่ต้องตรวจสอบค่า partial F ของตัวแปรทำนายที่มีอยู่เดิมนั้น เนื่องจากความจริงที่ว่าตัวแปรทำนายเดิมนั้นอาจจะไม่เหมาะสมที่จะอยู่ในสมการใหม่ เพราะตัวแปรเดิมอาจมีความสัมพันธ์กับตัวแปรใหม่ ซึ่งการเลือกตัวแปรแบบไปข้างหน้าไม่ได้ตรวจสอบในเรื่องนี้ นั่นคือ เมื่อรวมตัวแปรทำนายใดไว้ในสมการแล้วจะไม่กำจัดออกในขั้นต่อไป และการกำจัดตัวแปรแบบถอยหลังก็เช่นเดียวกัน คือจะไม่นำตัวแปรทำนายที่เอาออกไปแล้วกลับเข้ามาพิจารณาอีก

การถดถอยแบบขั้นบันไดทำได้ดังนี้

3.1 พิจารณาสัมประสิทธิ์สหสัมพันธ์บางส่วนระหว่างตัวแปรเกณฑ์กับตัวแปรทำนายแต่ละตัว (r_{xy}) เลือกตัวแปรทำนายที่ให้ค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนสูงที่สุด สมมุติให้ X_j สมการจะเป็น $\hat{Y} = a + b_j X_j$

3.2 หาค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนระหว่างตัวแปรเกณฑ์กับตัวแปรทำนายแต่ละตัวที่ยังไม่อยู่ในสมการ โดยถือว่าได้รวมตัวแปรทำนาย X_j ไว้ในสมการแล้ว และเลือกตัวแปรทำนายที่ให้ค่าสัมประสิทธ์สหสัมพันธ์บางส่วนสูงสุด สมมุติให้ X_1 สมการจะเป็น

$$\hat{Y} = a + b_j X_j + b_1 X_1$$

3.3 พิจารณา partial F ของทั้ง X_1 และ X_j ถ้ามีค่ามากกว่า $F_{\infty}(1,n-3)$ ทั้ง 2 ตัวก็รวม X_1 และ X_j ไว้ในสมการ

3.4 ทำตามข้อ 3.2 และ 3.3 โดยที่จะมีตัวแปรทำนายรวมอยู่ในสมการแล้ว 2 ตัว 3 ตัว ฯลฯ ตามลำดับ ในแต่ละขั้นต้องพิจารณาค่า partial F ของตัวแปรทำนายทุกตัว ถ้า ตัวใดมีค่าน้อยกว่า F_{∞} (1,n-m-1) ก็จะตัดตัวแปรทำนายนั้นออกจากสมการ ทำเช่นนี้จนไม่มีตัวแปรต้นตัวใดที่จะรวมในสมการจะไม่มีตัวใดที่จะถูกตัดออกจากสมการ (วิชิต หล่อจีระชุนห์กุล 2524)

ตอนที่ 4

ซิมูเลชั่น (Simulation)

ซิมูเลชั่น (Simulation) คือ ขบวนการที่จัดสร้างระบบจำลองขึ้นมาแทนระบบจริง แล้วนำเอาระบบจำลองที่สร้างขึ้นมานี้ไปทำการทดลอง ซึ่งจุดมุ่งหมายของการทำซิมูเลชั่น ก็คือ

1. เพื่อศึกษาพฤติกรรมของระบบจริง เพื่อจะได้นำมาใช้ในการพยากรณ์พฤติกรรมของระบบจริงว่าในอนาคตจะเป็นอย่างไร
2. เพื่อประเมินผลหรือเพื่อควบคุมระบบจริง

รูปแบบของระบบจำลองมีอยู่ 2 ชนิด คือ

1. Continuos Model คือ รูปแบบจำลองที่มีการเปลี่ยนไปตามเงื่อนไขของเวลา
2. Discrete models คือ รูปแบบจำลองที่ให้กฎเกณฑ์เกี่ยวกับการเปลี่ยนแปลงของสถานภาพของรูปแบบในช่วงเวลาที่กำหนด ซึ่งการเปลี่ยนแปลงสถานภาพของรูปแบบนี้เป็นไปได้ 2 ลักษณะ คือ Deterministic และ Probabilistic การเปลี่ยนแปลงในลักษณะ Deterministic เป็นการเปลี่ยนแปลงตามกฎเกณฑ์ที่กำหนดขึ้นมา แต่การเปลี่ยนแปลงในลักษณะ Probabilistic เป็นการเปลี่ยนแปลงที่เกิดขึ้นตามโอกาส (chance)

การสร้างตัวเลขสุ่ม (Random Number)

จากการซิมูเลชั่น (Simulation) ในรูปแบบจำลองที่เรียกว่า probabilistic หรือ Stochastic ซึ่งเป็นการจำลองที่มีการเปลี่ยนแปลงตามโอกาส (chance) คือ มีโอกาสที่จะเปลี่ยนแปลงไปในรูปแบบต่าง ๆ ได้หลายอย่าง ทำให้พยากรณ์ล่วงหน้าได้ยาก มีลักษณะเหมือนการเล่นเกมส์หรือการพนัน การจำลองข้อมูลที่มีลักษณะไม่แน่นอนเช่นนี้ ทำได้โดยการสร้างตัวเลขขึ้นมา ซึ่งเรียกวิธีการเช่นนี้ว่า การสร้างเลขสุ่ม (Random Number)

การสร้างเลขสุ่ม (Random Number) นี้แจกแจงได้หลายลักษณะ แต่ละลักษณะทำได้หลายวิธี ซึ่งผู้วิจัยเสนอเฉพาะวิธีที่ใช้ในการทำวิจัยครั้งนี้เท่านั้น ดังนี้

1. การสร้างเลขสุ่มให้มีการแจกแจงอย่างสม่ำเสมอ (Uniformly Distributed Random Numbers) ซึ่งมีค่าเฉลี่ยเท่ากับ 0 ความแปรปรวนเท่ากับ 1 ซึ่งมีสูตรดังนี้

$$n_{i+1} = (an+c) \pmod M$$
$$r = n/m$$

r คือ เลขสุ่ม (Random Number) ที่ต้องการ

n คือ เลขจำนวนเต็ม ที่ใช้เป็นค่าเริ่มต้น (หรือเรียกว่า Seed)

m คือ เลขจำนวนเต็ม มีค่า 2^{31} หรือ 2^{33}

a คือ เลขจำนวนเต็ม ที่มีค่าไม่แน่นอน แต่ละสูตรมีค่าไม่เท่ากัน

c คือ เลขจำนวนเต็ม แต่ละสูตรมีค่าไม่เท่ากัน

วิธีที่ใช้ในการวิจัยครั้งนี้คือวิธี Prime Modulus Multiplicative ซึ่งมีสูตรดังนี้

R_{i+1} = $(aR_i + c) \pmod M$

M = 2147483647 หรือ $2^{31}-1$

a = 16807 หรือ 7^5

R_i = 973253 หรือเป็นเลขไพม์ที่มีค่าอยู่ระหว่าง $(\pm\ 2^{31} - 1)$

R_i+1 = เลขสุ่มที่ต้องการ

2. การสร้างเลขสุ่มให้มีการแจกแจงแบบปกติ (Normal Random Number Distribution) ที่มีค่าเฉลี่ยเท่ากับ 0 ความแปรปรวนเท่ากับ 1 ซึ่งมีอยู่ด้วยกันหลายวิธี วิธีที่ใช้ในการวิจัยครั้งนี้คือวิธีของ Box และ Muller (1975: 550-556) มีดังนี้

$$X_n = (-2\ln R_n)^{1/2} \ \text{Cos}\ 2\pi\ R_{n+1}$$

$$X_{n+1} = (-2\ln R_n)^{1/2} \ \text{Sin}\ 2\pi\ R_{n-1}$$

เมื่อ X_n, X_{n+1} คือ เลขสุ่มที่มีการแจกแจงแบบปกติ ตามที่ต้องการ

R_n, R_{n+1} คือ ค่าที่เรียงลำดับคู่คี่ของเลขสุ่มที่มีการแจกแจงแบบสม่ำเสมอ

3. การสร้างเลขสุ่มให้มีการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Random Distribution) ที่มีค่าเฉลี่ย เท่ากับ 0 ศวามแปรปรวนเท่ากับ 1 เราสามารถสร้างตัวแปรสุ่ม $X_i = (X_1, X_2 .... X_n)$ ได้ดังวิธีของ Scheuer ซึ่งผู้วิจัยใช้ในการทำวิจัยครั้งนี้ และได้เสนอไว้อย่างละเอียดในบทที่ 3

<u>ตอนที่ 5</u>

<u>การแจกแจงแบบปกติหลายตัวแปร</u>

(Multivariate Normal Distribution)

เมื่อ X_{ij} เป็นตัวแปรสุ่ม (random variable)

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

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

$$Fx(X) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp - 1/2\,(X_i - \mu)\Sigma^{-1}(X - \mu) \quad (2.31)$$

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

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

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

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

Σ เป็น pxp positive definite และสมมาตร (Symmetric) ซึ่งคือเมตริกซ์ความแปรปรวนและความแปรปรวนร่วม (Variance - Covariance Matrix)

<u>การแจกแจงของค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณกำลังสอง</u>

(Distribution of the Multiple Correlation Coefficient Square)

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

$$\rho_{y.12..p} = \rho_{yy} = \sqrt{1 - \frac{\sigma^2_{(Y-\hat{Y})}}{\sigma^2_y}} \qquad (2.32)$$

เมื่อศึกษากับกลุ่มตัวอย่างสามารถประมาณค่าโดย

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

$$= \sqrt{\beta_1 r_{y1} + \beta_2 r_{y2} + \ldots\ldots + \beta_p r_{yp}} \qquad (2.33)$$

เมื่อ r_{yi} คือ สัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรเกณฑ์กับตัวแปรทำนาย (X_i) แต่ละตัว และ β คือ สัมประสิทธิ์การถดถอยมาตรฐานที่จะทำให้ค่าสหสัมพันธ์พหุคูณมีค่าสูงสุด ซึ่งมีค่าอยู่ระหว่าง 0 ถึง 1 และจะมีค่าเพิ่มขึ้นเมื่อจำนวนตัวแปรทำนายเพิ่มขึ้น

การประมาณค่าสหสัมพันธ์พหุคูณกำลังสองของประชากรจะมีลักษณะเช่นเดียวกับ การศึกษาสหสัมพันธ์อย่างง่าย โดยคาดว่าค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณกำลังสองที่คำนวณได้จาก กลุ่มตัวอย่าง ($R^2_{y.1.2..p}$) จะกระจายอยู่รอบ ๆ ค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณกำลังสอง ของประชากร แต่เนื่องจากค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณกำลังสองนี้มีค่าอยู่ระหว่าง 0 ถึง 1 และจัดเป็นตัวประมาณค่าที่เอนเอียง จึงทำให้การแจกแจงมีความเอนเอียงไปทางบวกเสมอ ซึ่งไม่สามารถคาดคะเนลักษณะการแจกแจงที่แน่นอนได้ จากการศึกษาพบว่า ลักษณะการแจกแจงของค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณกำลังสองนี้ จะขึ้นอยู่กับอิทธิพลของขนาดของกลุ่มตัวอย่าง จำนวน ตัวแปรทำนายและระดับความสัมพันธ์ในประชากร (ρ) (Muirhead 1982 : 171) ดังนี้

เมื่อ $\rho = 0$

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

และ $$\mathrm{Var}(R^2) = \frac{2(n-p)(p-1)}{(n^2-1)(n-1)} \qquad (2.35)$$

จาก (2.34) และ (2.35) เมื่อระดับความสัมพันธ์ของประชากรมีค่าเป็นศูนย์ หรือไม่มีความสัมพันธ์กันเลย ลักษณะการแจกแจงของค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณกำลังสองจะขึ้นอยู่กับขนาดของกลุ่มตัวอย่างและจำนวนตัวแปรทำนายเท่านั้น เมื่อจำนวนตัวแปรทำนายเพิ่มขึ้นจะทำให้ค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณกำลังสองสูงขึ้นด้วย (p --> n : R --> 1) แต่ถ้าระดับความสัมพันธ์ในประชากรไม่เท่ากับศูนย์ ($\rho \neq 0$) แล้ว การคาดคะเนลักษณะการแจกแจงของค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณกำลังสองจะทำได้ยากมากดังนี้

เมื่อ $\rho \neq 0$

$$E(R^2) = 1 - \frac{n-p-1}{n-1}(1-\rho^2)\,F\left(1,1,(n+1)/2,\rho^2\right) \qquad (2.36)$$

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

$$E(R^2) = 1 - \frac{n-p-1}{n-1}(1-\rho^2) - \frac{n-p-1}{n-1}, \frac{2}{n+1}\rho^2)(1-\rho^2) \qquad (2.37)$$

และ

$$\mathrm{Var}(R^2) = \frac{n-p+1}{n^2(n+2)}(1-\rho^2)^2\left\{2(p-1)+4p^2\left[\frac{4(p-1)+n(n+p+1)}{n+4}\right] + 0(n^{-2})\right\} \qquad (2.38)$$

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

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

$$= \frac{a + (b-1/2)\rho^2 + \rho^4}{a+b+1/2} \qquad (2.39)$$

และ

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

$$= \frac{2\rho^2(1-\rho^2)^2}{(a+b+1/2)} \qquad (2.40)$$

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

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

จะเห็นว่าลักษณะการแจกแจงของค่าสัมประสิทธิ์สหสัมพันธ์พหุคูณกำลังสอง นอกจากจะขึ้นอยู่กับขนาดของกลุ่มตัวอย่าง และจำนวนตัวแปรทำนายแล้ว ยังขึ้นอยู่กับระดับของความสัมพันธ์ในประชากรซึ่งไม่ทราบค่าอีกด้วย

ตอนที่ 6

งานวิจัยที่เกี่ยวข้อง

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

จะเด็ด สวรรค์ตรานนท์ (2530) ได้เปรียบเทียบวิธีการคัดเลือกตัวแปรอิสระเข้าสู่สมการถดถอย 5 วิธีด้วยกัน คือ วิธีการกำจัดตัวแปรปรวนแบบถอยหลัง วิธีการเลือกตัวแปรแบบไปข้างหน้า วิธีการถดถอยแบบขั้นบันได วิธีการถดถอยแบบขั้นตอน และวิธีการกำจัดตัวแปรโดยใช้สัมประสิทธิ์สหสัมพันธ์ เมื่อความคลาดเคลื่อนมีการแจกแจงแบบปกติ โลจิสติด ดับเบิ้ลเอ็กซ์โปแนนเชียล และแบบปกติปลอมปน สำหรับรูปแบบของการแจกแจงแบบปกติปลอมปนนั้น ได้ทำการศึกษาเมื่อสเกลแฟคเตอร์มีค่าเท่ากับ 3 และ 10 การปลอมปนเท่ากับ 5%, 10% และ 25% ขนาดตัวอย่างเท่ากับ 15, 30, 50 และ 100 จำนวนตัวแปรอิสระเท่ากับ 3, 5, 7, และ 9 ใน ระดับนัยสำคัญ .01 และ .05 ตามลำดับ โดยการเปรียบเทียบวิธีการเลือกสมการถดถอยนั้น ได้พิจารณาค่าผลรวมของความคลาดเคลื่อนกำลังสอง ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง และค่าของตัวสถิติธีล ซึ่งสรุปผลได้ดังนี้

1) เมื่อความคลาดเคลื่อนมีการแจกแจงเป็นไปตามข้อตกลงเบื้องต้น นั้นวิธีของการถดถอยแบบขั้นบันไดให้ผลดีที่สุด ส่วนวิธีที่ให้ผลรองลงไปคือ วิธีการเลือกตัวแปรแบบไปข้างหน้า วิธีการถดถอยแบบขั้นตอน วิธีการกำจัดตัวแปรโดยใช้สัมประสิทธิ์สหสัมพันธ์และวิธีที่ให้ผลเป็นอันดับสุดท้าย คือวิธีการกำจัดตัวแปรแบบถอยหลัง

2) เมื่อความคลาดเคลื่อนมีการแจกแจงไม่เป็นไปตามข้อตกลงเบื้องต้นนั้น ในกรณีที่ความคลาดเคลื่อนมีการแจกแจงแบบโลจิสติค วิธีการถดถอยแบบขั้นบันไดให้ผลดีที่สุด ส่วนวิธีที่ให้ผลรองลงไปคือ วิธีการเลือกตัวแปรไปข้างหน้า วิธีการถดถอยแบบขั้นตอน และวิธีที่ให้ผลเป็นอันดับสุดท้ายคือ วิธีการกำจัดตัวแปรแบบถอยหลังและการกำจัดตัวแปรโดยใช้สัมประสิทธิ์สหสัมพันธ์ ในกรณีที่ความคลาดเคลื่อนมีการแจกแจงแบบดับเบิ้ลเอ็กซ์โปเนนเชียลและแบบปกติปลอมปนนั้น จะให้ผลเหมือนกันคือ วิธีการถดถอยแบบขั้นบันไดให้ผลดีที่สุด ส่วนวิธีที่ให้ผลรองลงไปตามลำดับ คือ วิธีการเลือกตัวแปรแบบไปข้างหน้า วิธีการกำจัดตัวแปรโดยใช้สัมประสิทธิ์สหสัมพันธ์การถดถอยแบบขั้นตอน และวิธีที่ให้ผลเป็นอันดับทุดท้าย คือ วิธีการกำจัดตัวแปรแบบถอยหลัง

ฮาลินสกีและเฟลด์ (Halinske and feldt 1970:151-158) ได้ทำการศึกษาโดยใช้เทคนิคมอนติการ์โลซิมูลชั่น เกี่ยวกับขนาดของกลุ่มตัวอย่างที่เหมาะสมในการวิเคราะห์การถดถอยพหุคูณ พบว่า ควรใช้อัตราส่วนระหว่างขนาดของกลุ่มตัวอย่างกับจำนวนตัวแปรอย่างน้อยที่สุดเท่ากับ 10 : 1

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