บทที่ 2

ทฤษฎีและผลงานวิจัยที่เกี่ยวข้อง

ในการวิจัยครั้งนี้สิ่งที่สนใจศึกษาคือ การเปรียบเทียบวิธีการคัดเลือกตัวแปรอิสระเพื่อใช้ในตัวแบบถดถอยเชิงเส้นซึ่งประกอบด้วยการกำจัดตัวแปรแบบถอยหลัง การเลือกตัวแปรแบบไปข้างหน้า การถดถอยแบบขั้นบันได การถดถอยแบบขั้นตอน และการกำจัดตัวแปรโดยใช้สัมประสิทธิ์สหสัมพันธ์ ซึ่งในบทนี้จะกล่าวถึงรายละเอียดของแต่ละวิธีการ ส่วนในตอนท้ายของบทนี้ จะนำเสนอผลงานวิจัยที่เกี่ยวข้องโดยมีรายละเอียดต่าง ๆ ดังนี้

2.1 การกำจัดตัวแปรแบบถอยหลัง

2.1.1 ตั้งสมการที่รวมเอาตัวแปรอิสระที่ควรพิจารณาทั้งหมด สมมติว่ามี k ตัว สมการจะเป็นดังนี้

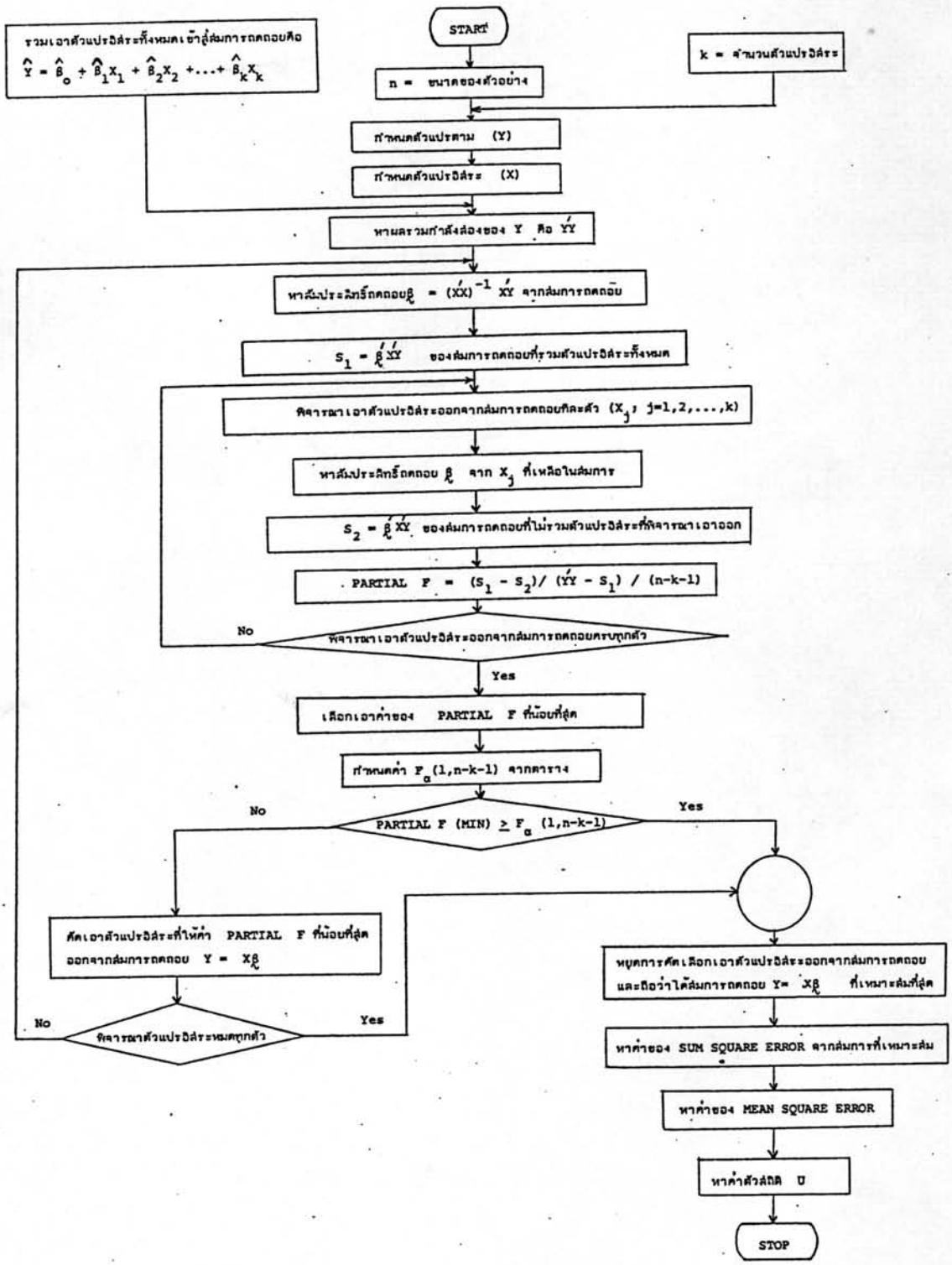
$$\hat{Y} = \hat{\beta}_o + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k \quad ; k = \text{จำนวนตัวแปรอิสระ}$$

2.1.2 คำนวณพาเชียลเอฟ (Partial F) ของตัวแปรอิสระแต่ละตัว ในจำนวนพาเชียลเอฟนี้เลือกค่าที่น้อยที่สุด สมมติได้ F_j และนำ F_j เปรียบเทียบ กับค่า F จากตาราง คือ $F_\alpha(1,n-k-1)$ ถ้า F_j น้อยกว่า F จากตาราง ให้กำจัดตัวแปรอิสระ X_j ออกจากสมการถดถอย

2.1.3 ตั้งสมการใหม่โดยไม่รวม X_j ในสมการแล้วทำตามข้อ (2.1.2) อีก โดยที่ค่า k จะเปลี่ยนเป็น k-1 ทำเช่นนี้จนในที่สุดพาเชียลเอฟ ทุก ๆ ตัวมีค่ามากกว่าเอฟที่กำหนดจากตาราง

2.1.4 หาค่าของผลรวมกำลังสองของความคลาดเคลื่อน ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง และหาค่าของตัวสถิติ U

รูปที่ 2.1 แสดงผังงานของวิธีการกำจัดตัวแปรแบบถอยหลัง



2.2 การเลือกตัวแปรแบบไปข้างหน้า

2.2.1 พิจารณาสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระแต่ละตัวกับตัวแปรตาม (r_{xy}) เลือกเอาตัวแปรอิสระที่ให้ค่าสัมประสิทธิ์สหสัมพันธ์สูงที่สุด สมมติได้ X_j สมการจะเป็น

$$\hat{Y} = \hat{\beta}_o + \hat{\beta}_j X_j \quad ; \quad j = 1,2,3,\ldots, k$$

2.2.2 หาค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนระหว่างตัวแปรตามกับตัวแปรอิสระ แต่ละตัวที่ยังไม่เข้าอยู่ในสมการถดถอย โดยถือว่าได้รวมเอาตัวแปรอิสระ X_j เข้าไว้ในสมการถดถอย แล้วนั่นคือหาค่า $r_{y\ell.j}$; $\ell = 1,2,3, \ldots, j-1, j+1,\ldots,k$

2.2.3 เลือกค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนที่สูงที่สุด สมมติว่าเป็น $r_{y\ell.j}$ จึงควรรวม X_ℓ เข้าไว้ในสมการถดถอยเป็นตัวแปรใหม่

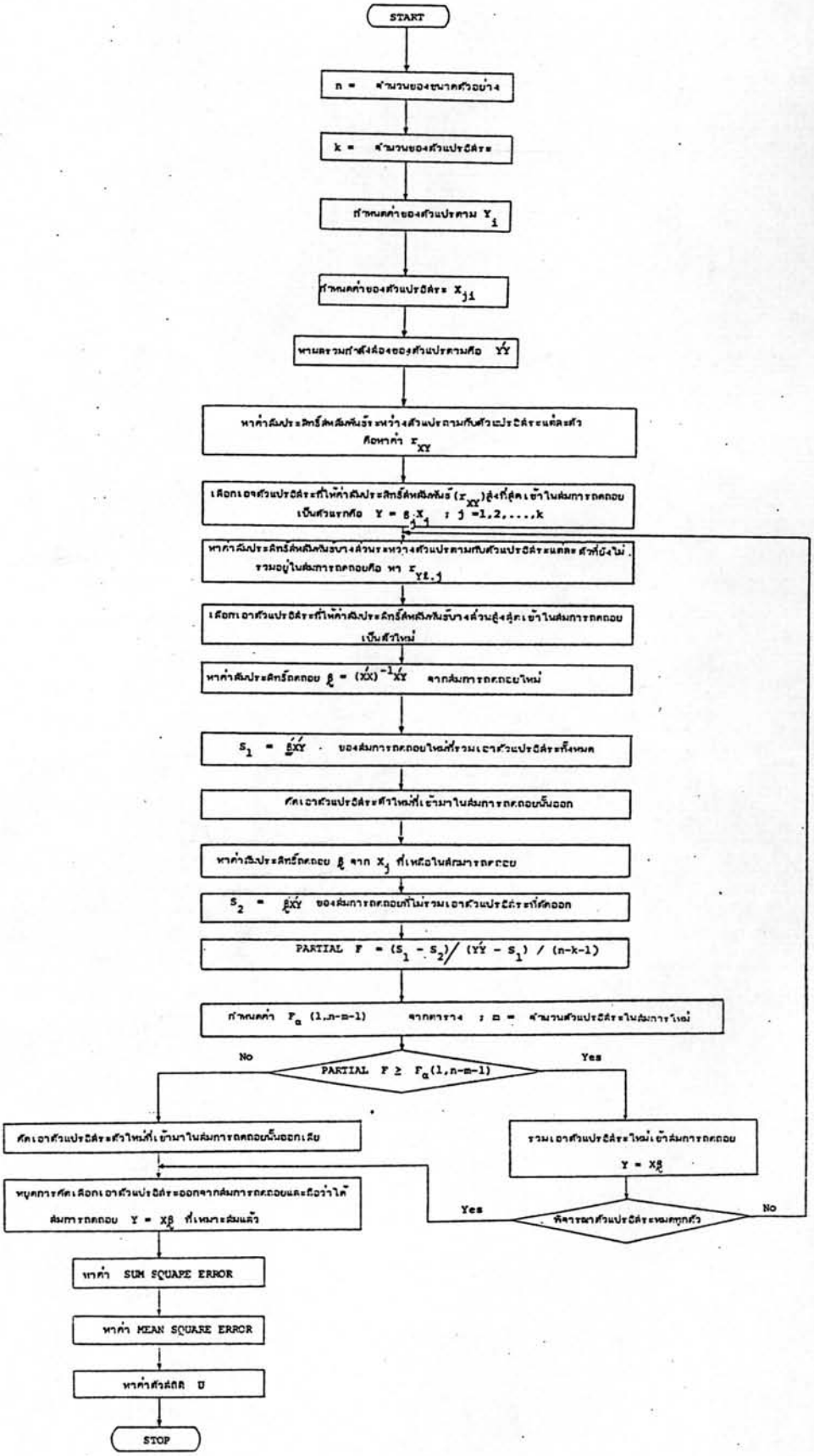
2.2.4 พิจารณาพาเชียลเอฟของตัวแปรอิสระใหม่ X_ℓ ในข้อ (2.2.3) ถ้ามีค่าสูงกว่า F_α (1,n-m-1) แสดงว่าเป็นการสมควรที่จะรวม X_ℓ ไว้ในสมการถดถอย ในที่นี้ m คือ จำนวนตัวแปรอิสระ ในสมการใหม่และ n คือ จำนวนค่าสังเกต (Observation)

2.2.5 ทำตามข้อ (2.2.2) (2.2.3) และ (2.2.4) อีกโดยถือว่า สมการได้รวมตัวแปรอิสระไว้แล้ว 2 ตัว 3 ตัว ฯลฯ ตามลำดับ จนกระทั่งพาเชียลเอฟที่ได้จากตัวแปรอิสระใหม่มีค่าน้อยกว่า F_α (1,n-m-1) คือ เอฟที่กำหนดจากตารางจึงถือได้ว่าตัวแปรอิสระใหม่ไม่ควรรวมอยู่ในสมการ ซึ่งแสดงว่าได้สมการถดถอยที่เหมาะสมที่สุดแล้ว

ในกรณีที่รวมตัวแปรอิสระไว้แล้ว 2 ตัว สมมติเป็นตัวที่ 1 และที่ 2 จะหาสัมประสิทธิ์สหสัมพันธ์บางส่วนของตัวแปรตามและตัวแปรอิสระตัวที่ 3 คือ $r_{y3.12}$

2.2.6 หาค่าผลรวมกำลังสองของความคลาดเคลื่อน ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง และหาค่าของตัวสถิติ U

รูปที่ 2.2 แสดงผังงานของวิธีการเลือกตัวแปรแบบไปข้างหน้า



2.3 <u>การถดถอยแบบขั้นบันได</u>

2.3.1 พิจารณาสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระและตัว (r_{XY}) เลือกตัวแปรอิสระที่ให้ค่าสัมประสิทธิ์สหสัมพันธ์สูงที่สุด สมมติได้ X_j สมการจะเป็น

$$\hat{Y} = \hat{\beta}_o + \hat{\beta}_j X_j \quad ; j = 1,2,\ldots,k$$

2.3.2 หาค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนระหว่างตัวแปรตามกับตัวแปรอิสระแต่ละตัวที่ยังไม่อยู่ในสมการ โดยถือว่าได้รวมตัวแปรอิสระ X_j ไว้ในสมการแล้วและเลือกตัวแปรอิสระที่ให้ค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนสูงสุด สมมติให้ X_ℓ สมการก็จะเป็น

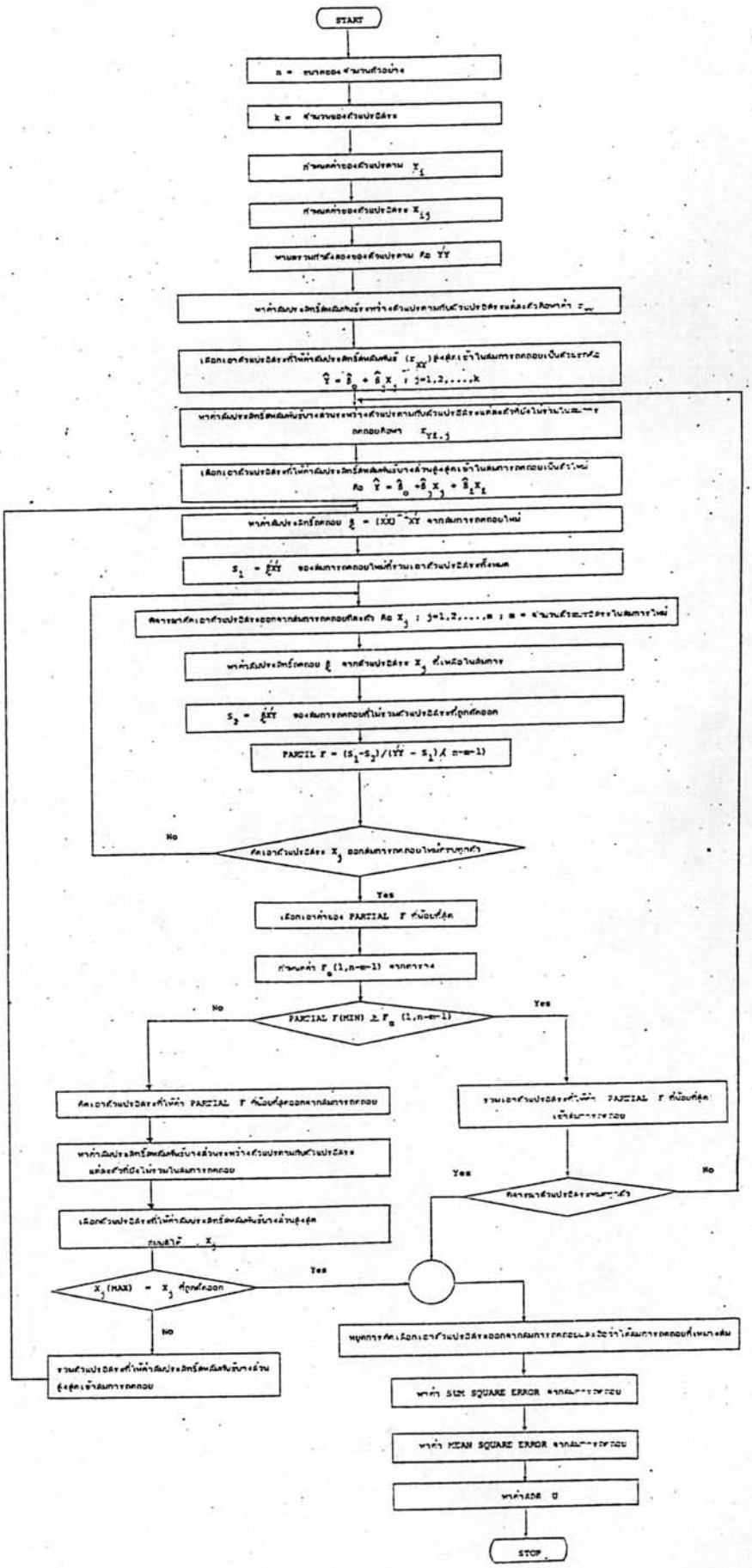
$$\hat{Y} = \hat{\beta}_o + \hat{\beta}_j X_j + \hat{\beta}_\ell X_\ell \quad ; \ell = 1,2,\ldots, j-1, j+1,\ldots,k$$

2.3.3 พิจารณาพาเชียลเอฟของทั้ง X_ℓ และ X_j ถ้ามีค่ามากกว่า F_α (1,n-3) ทั้งสองตัวก็รวม X_ℓ และ X_j ไว้ในสมการ

2.3.4 ทำตามข้อ (2.3.2) และ (2.3.3) โดยที่จะมีตัวแปรอิสระรวมอยู่ในสมการแล้ว 2 ตัว 3 ตัว ฯลฯ ตามลำดับในแต่ละขั้นต้องพิจารณาค่าพาเชียลเอฟของตัวแปรอิสระทุกตัว ถ้าตัวใดมีค่าน้อยกว่า F_α (1,n-m-1) ก็จะตัดตัวแปรอิสระนั้นออกจากสมการ

2.3.5 หาค่า ผลรวมกำลังสองของความคลาดเคลื่อน ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง และหาค่าของตัวสถิติ U

รูปที่ 2.3 แสดงผังงานของวิธีการถดถอยแบบขั้นบันได



2.4 การถดถอยแบบขั้นตอน

2.4.1 เลือกตัวแปรอิสระที่มีค่าสัมประสิทธิ์สหสัมพันธ์สูงที่สุด สมมติได้ X_j สมการเป็น

$$\hat{Y} = \hat{\beta}_o + \hat{\beta}_j X_j \quad ; \quad j = 1,2,3, \ldots, k$$

2.4.2 หาค่าความแตกต่างระหว่าง Y_i และ $\hat{Y}_i$ (Residual) โดยการแทนค่า X_{ji} ในสมการ เมื่อ i=1,2,3,...,n. และกำหนดให้ความแตกต่างนี้เรียกว่า Z_i นั่นคือ

$$Z_i = Y_i - \hat{Y}_i$$

2.4.3 ให้ Z เป็นตัวแปรตามตัวใหม่แทน Y แล้วหาสัมประสิทธิ์สหสัมพันธ์บางส่วนระหว่าง Z กับตัวแปรอิสระแต่ละตัวที่ยังไม่รวมในสมการ เลือกเอาตัวแปรอิสระที่ให้ค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนสูงสุด สมมติให้ X_ℓ ตั้งสมการใหม่

$$\hat{Z}_i = \hat{\beta}_o + \hat{\beta}_\ell X_\ell \quad ; \quad \ell = 1,2,3,\ldots, j-1, j+1,\ldots,k$$

หาค่า Z_i ตัวใหม่ซึ่งก็คือ ผลต่างระหว่าง Z_i ในข้อ (2.4.2) และ $\hat{Z}_i$ ในข้อ (2.4.3)

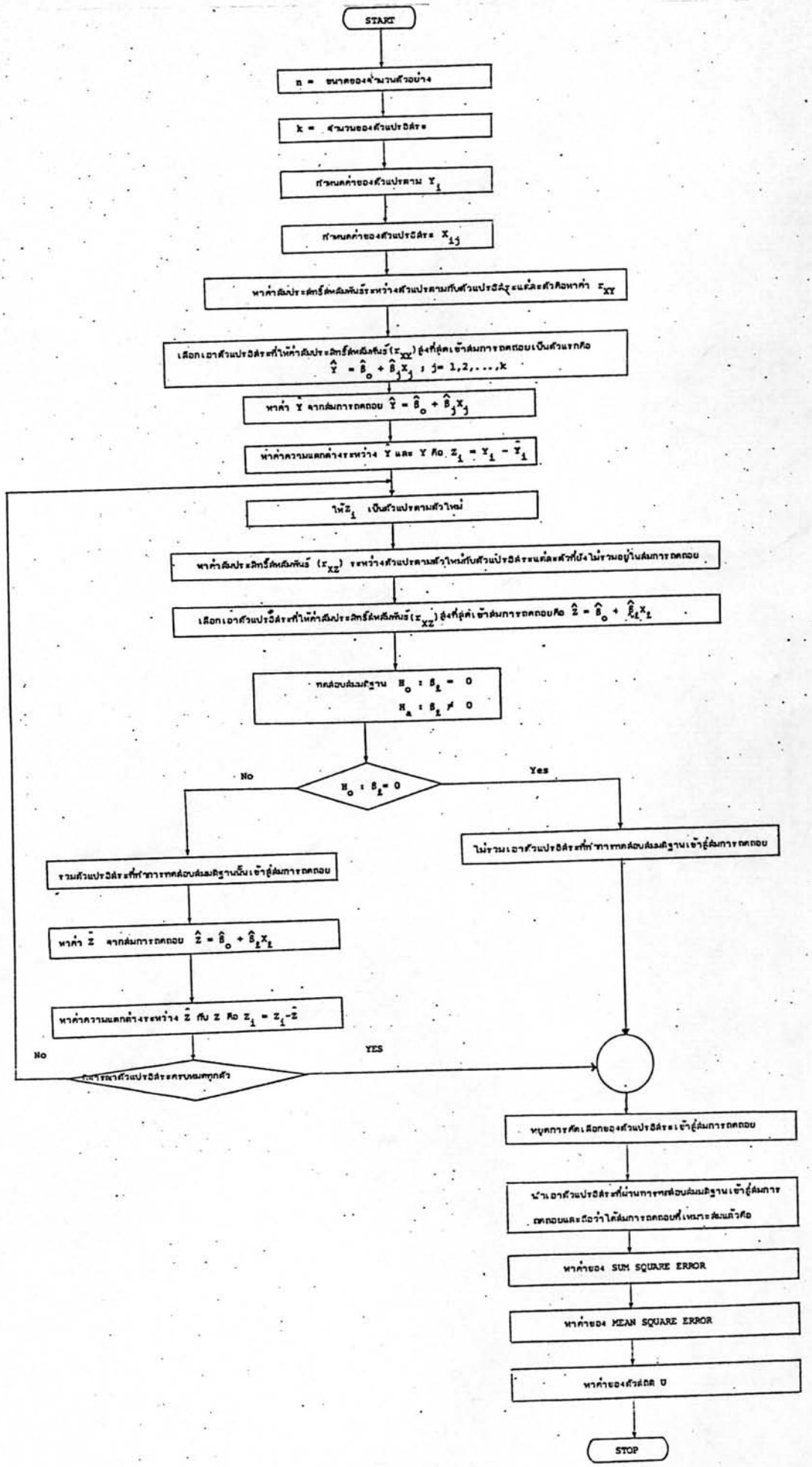
2.4.4 ใช้ Z_i ตัวใหม่นี้เป็นตัวแปรตามแล้วทำตามข้อ (2.4.3) จนสมการถดถอย $\hat{Z} = \hat{\beta}_o + \hat{\beta}_t X_t$ ไม่เป็นที่ยอมรับ นั่นคือ ยอมรับสมมติฐาน

$$H_o : \beta_t = 0$$

2.4.5 นำสมการในข้อ (2.4.1) และหลาย ๆ สมการในข้อ (2.4.3) มารวมกันก็จะได้สมการถดถอยที่แท้จริง

2.4.6 หาค่าผลรวมกำลังสองของความคลาดเคลื่อน ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง และหาค่าตัวสถิติ U

รูปที่ 2.4 แสดงผังงานของวิธีการถดถอยแบบขั้นตอน



2.5 <u>การกำจัดตัวแปรโดยใช้สัมประสิทธิ์สหสัมพันธ์</u>

2.5.1 หาสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระทุก ๆ ตัว (r_{XX}) แล้วเลือกเอาคู่ของตัวแปรอิสระที่ให้ค่าสัมประสิทธิ์สหสัมพันธ์สูงสุด สมมติได้คู่ X_i และ X_j

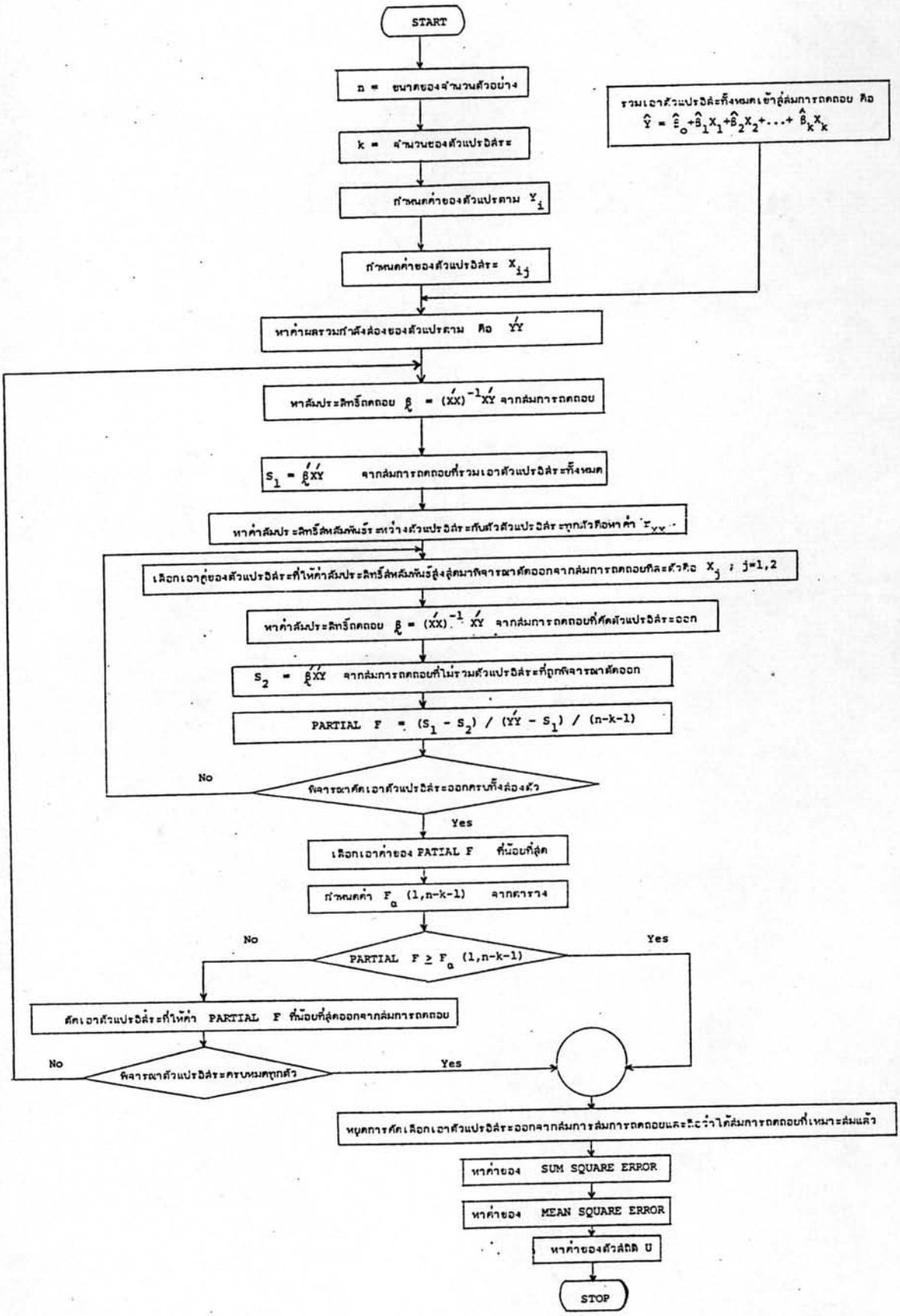
2.5.2 หาค่าพาเชียลเอฟระหว่างตัวแปรอิสระ X_i และ X_j สมมติได้ F_i และ F_j

2.5.3 เลือกค่าพาเชียลเอฟที่น้อยที่สุดระหว่าง F_i และ F_j สมมติได้ F_j แล้วนำเอาค่า F_j ไปเปรียบเทียบกัน F_α (1,n-k-1) ที่กำหนดจากตาราง ถ้า F_j มีค่าน้อยกว่า F_α ก็หมายความว่า X_j ไม่สมควรรวมอยู่ในสมการถดถอย

2.5.4 ตั้งสมการขึ้นใหม่โดยไม่รวมเอาตัวแปรอิสระ X_j แล้วกลับขึ้นไปทำข้อ (2.5.1) ใหม่ จนกว่าค่าของพาเชียลเอฟที่หาได้ในข้อ (2.5.3) จะมีค่ามากกว่าเอฟที่กำหนด จากตารางจึงจะหยุดการคัดเลือกเอาตัวแปรอิสระออกจากสมการถดถอย และถือว่าได้สมการถดถอยที่เหมาะสมแล้ว

2.5.5 หาค่าผลรวมกำลังสองของความคลาดเคลื่อน ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง และหาค่าตัวสถิติ U

รูปที่ 2.5 แสดงผังงานของวิธีการกำจัดตัวแปรโดยใช้สัมประสิทธิ์สหสัมพันธ์



2.6 การแปลงข้อมูลให้เป็นมาตรฐานเดียวกัน (Transformed Data)

ในกรณีที่ข้อมูลมีค่าอยู่ในช่วงค่อนข้างกว้างหรือ ในกรณีที่ต้องการให้ข้อมูลมีค่าอยู่ในระบบเดียวกัน เช่นการแปลงข้อมูลของตัวแปรบางตัวที่วัดค่าออกมาได้ในรูปของปริมาตร เพื่อกลับมาให้เป็นข้อมูลเชิงเส้น หรือเชิงปริมาณ เช่นเดียวกับตัวแปรอื่น ๆ ในบางครั้งการแปลงข้อมูลนี้ก็ช่วยในด้านการคำนวณได้มาก เพราะว่าบางครั้งข้อมูลที่เราได้มานั้นมีค่ามากในแต่ละค่า ถ้านำเอาข้อมูลเหล่านี้มาคำนวณหาสิ่งที่ต้องการ จะทำให้การคำนวณเกิดความยุ่งยาก การคำนวณที่ได้ออกมานั้นอาจจะไม่ถูกต้องและเสียเวลาเปล่า ไม่ว่าจะเป็นการคำนวณด้วยมือหรือคำนวณโดยใช้เครื่องคอมพิวเตอร์ ดังนั้นการแปลงข้อมูลจึงนับได้ว่ามีประโยชน์ต่อการปรับข้อมูลและลดเวลาในการคำนวณลงมาก โดยมีสูตรดังนี้

$$Log(Y_i) \quad ; \quad i = 1,2,3,\ldots, n$$

$$Log(X_{ij}) \quad ; \quad j = 1,2,3,\ldots, k$$

n = จำนวนข้อมูล

k = จำนวนตัวแปรอิสระ

2.7 เอ็กซ์ตราซัมสแควร์ และพาเชียลเอฟ (Extra Sum Square and Partial F)

ให้

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_q X_q + \mu \quad \ldots\ldots(1)$$

คือ สมการถดถอยแสดงความสัมพันธ์ $Y = f(X's)$

จากสมการ (1) ซึ่งสามารถหาค่าประมาณของสัมประสิทธิ์ถดถอย $_1\hat{\beta}$ โดยวิธีกำลังสองน้อยที่สุด (Ordinary Least Square Estimator) ได้ดังนี้

$$_1\hat{\beta}' = (\hat{\beta}_1 \quad \hat{\beta}_2 \quad \hat{\beta}_3 \cdots \quad \hat{\beta}_q)$$

พร้อมทั้งสามารถหาผลรวมกำลังสอง (Sum Square) ที่เกี่ยวข้องได้ดังนี้ คือ

(1) $_1\hat{\beta} = (_1X'_1X)^{-1}{}_1X'Y$ เมื่อ $_1X$ คือเมตริกซ์ขนาด nxq

(2) $SSR_1 = {}_1\hat{\beta}'\,{}_1X'Y$

(3) $SSE_1 = Y'Y - {}_1\hat{\beta}'_1X'Y$

$$MSE_1 = \hat{\sigma}_1^2 = \frac{1}{n-q}(Y'Y - {}_1\hat{\beta}'\,{}_1X'Y)$$

q = จำนวนพารามิเตอร์สัมประสิทธิ์ถดถอยของสมการ 1

$_1\hat{\beta}$ = สัมประสิทธิ์ถดถอยของสมการ 1

SSR_1 = ผลรวมกำลังสองของรีเกรชชัน (Sum Square Regression) ของสมการ 1

MSE_1 = ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (Mean Square Error) ของสมการ 1

ให้

$$Y = \beta_1 + \beta_2X_2 + \beta_3X_3 + \ldots + \beta_qX_q + \beta_{q+1}X_{q+1} + \ldots + \beta_pX_p + \mu \ldots (2)$$

คือสมการแสดงความสัมพันธ์ $Y = f(X's)$ โดยที่ $p>q$

จากสมการ (2) สามารถหาค่าประมาณของสัมประสิทธิ์ถดถอย $_2\hat{\beta}$ คือ

$$_2\hat{\beta}' = (\hat{\beta}_1 \hat{\beta}_2 \hat{\beta}_3 \cdots \hat{\beta}_q \hat{\beta}_{q+1} \cdots \hat{\beta}_p)$$

พร้อมทั้งสามารถคำนวณหาผลรวมกำลังสองที่เกี่ยวข้องได้ดังนี้คือ

(1) $_2\hat{\beta} = (_2X'_2X)^{-1}{}_2X'Y$ เมื่อ $_2X$ คือเมตริกซ์ขนาด nxp

(2) $SSR_2 = {}_2\hat{\beta}'_2X'Y$

(3) $SSE_2 = Y'Y - {}_2\hat{\beta}'_2X'Y$ และ $MSE_2 = \hat{\sigma}^2_2$ หรือ

$$s^2 = \frac{1}{n-p} (Y'Y - {}_2\hat{\beta}'_2X'Y)$$

p = จำนวนพารามิเตอร์สัมประสิทธิ์ถดถอยของสมการ 2

$_2\hat{\beta}$ = สัมประสิทธิ์ถดถอยของสมการ 2

SSR_2 = ผลรวมกำลังสองของรีเกรชชั่นของสมการ 2

SSE_2 = ผลรวมกำลังสองของความคลาดเคลื่อนของสมการ 2

MSE_2 = ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองของสมการ 2

ลองผลลัพธ์ข้างต้นจะพบว่า Extra Sum Square Regression

$$ESSR = SSR_2 - SSR_1$$

$$= {}_2\hat{\beta}'_2X'Y - {}_1\hat{\beta}'_1X'Y$$

ซึ่งเป็นผลรวมกำลังสองของตัวแปรอิสระ X_{q+1} , X_{q+2},..., X_p

ที่เพิ่มขึ้นมาจากสมการ (1) โดยที่สัมประสิทธิ์ความถดถอยคือ $\hat{\beta}_{q+1}$, $\hat{\beta}_{q+2}$, ..., $\hat{\beta}_p$

ซึ่ง * $\frac{SSR_1}{\sigma^2} \sim \chi^2_q$ มีการกระจายแบบไคสแควร์กำลังสอง และมีองศาแห่งความอิสระ

(Dregree of freedom) = q

* $\frac{SSR_2}{\sigma^2} \sim \chi^2_p$ มีการกระจายแบบไคสแควร์กำลังสอง และมีองศาแห่งความอิสระเท่ากับ p

และ * $\frac{ESSR}{\sigma^2} \sim \chi^2_{p-q}$ มีการกระจายแบบไคสแควร์กำลังสองและมีองศาแห่งความอิสระเท่ากับ p-q

ดังนั้น $\frac{ESSR/(p-q)\ \sigma^2}{(\acute{Y}Y - SSR_2)/(n-p)\ \sigma^2} = \frac{ESSR/(p-q)}{\hat{\sigma}^2_2} \sim F_{(p-q,n-p)}$ **

จะมีการกระจายแบบเอฟและมีองศาแห่งความอิสระเท่ากับ p-q และ n-p

ดังนั้นจึงสามารถปฏิเสธสมมติฐาน

$$H_o : \beta_{q+1} = \beta_{q+2} = \dots = \beta_p = 0$$

$$H_1 : \beta_j \text{ ไม่เท่ากับ } 0 \text{ ทั้งหมด ; } j = q+1, q+2, \dots, p$$

ระดับนัยสำคัญ α เมื่อ $F_c > F_{1-\alpha}(p-q, n-p)$ โดยที่

$$F_c = \frac{(SSR_2 - SSR_1)/(p-q)}{\hat{\sigma}^2_2}$$

---

* ดูการพิสูจน์ในหนังสือ Theory and Application of the Linear Model Chapter 4 Distribution of Quadratic Forms, P. 124-141 ของ FRANKLIN A. GRAYBILL

** จากนิยาม $U = (\chi^2_1/v_1)(\chi^2_2/v_2) \sim F(v_1, v_2)$

จากความรู้ Extra Sum Square และการทดล่อบล่มมติฐาน $H_o : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0$ นั้นอกจากนำไปใช้สำหรับทดล่อบทำเซ็ตย่อยที่ดีที่สุด (Best Subset) แล้ว ยังเป็นกรณีทั่วไปของ พาเยียลเอฟเทล่ท์ คือ

จากล่มการ $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p + \mu$ ซึ่งล่ามารถหาค่าเฉลี่ยของความคลาดเคลื่อนกำลังล่อง และผลรวมกำลังล่องของความถดถอยของ เฉพาะ β_j ได้ดังนี้

$$SS(\beta_j / \beta_1, \beta_2 \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p) = SS(\beta_1, \beta_2, \ldots, \beta_p) - SS(\beta_1, \beta_2 \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p) \quad ; \quad 2 \le j \le p$$

เมื่อ $SS(\beta_j / \beta_1, \beta_2, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p)$ คือ ผลรวมกำลังล่องของ X_j เมื่อ X's อื่น ๆ เข้าสู่แบบจำลองแล้วดังนั้นพาเยียลเอฟคือ

$$F_c = \frac{SS(\beta_1, \beta_2, \ldots, \beta_p) - SS(\beta_1, \beta_2, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p)}{\hat{\sigma}^2}$$

โดยที่จะปฏิเล่ธล่มมติฐาน $H_o : \beta_j = 0 \quad ; \quad 2 \le j \le p$ ณ ระดับนัยล่าคัญ

$$H_1 : \beta_j \neq 0$$

คือ $F_c \geq F_{1-\alpha}(1, n-p)$

พาเยียลเอฟ ใช้ล่าหรับตรวจล่อบนัยล่าคัญของ β_j เพื่อตัดสินใจว่าตัวแปรอิล่ระใดควรคงไว้ ตัวแปรอิล่ระใดควรตัดทิ้งจากล่มการถดถอยโดยที่ β_j ปรากฏอยู่ ณ ตำแหน่งในในล่มการถดถอยก็ได้ แต่ในทางปฏิบัติจะต้องยึดถือหลักเกณฑ์ของ Extra Sum Square ไว้เป็นแนวทางเล่มอกล่าวคือ จะต้องคำนวณหาค่า $SS(\beta_1, \beta_2, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p)$ ได้เฉพาะเมื่อจัดให้ X_j เข้าสู่ล่มการเป็นตัวสุดท้าย นั้นคือ $SS(\beta_j / \beta_1, \beta_2, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p)$ คำนวนหาได้จากผลต่างระหว่างผลรวมกำลังล่องของความถดถอย

จากสมการ

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1} + \ldots + \beta_p X_p + \mu$$

และ

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_{j-1} X_{j+1} + \beta_{j+1} X_{j+1} + \ldots + \beta_p X_p + \beta_j X_j + \mu$$

การหาและการนำค่าของพาเชียลเอฟที่ได้กล่าวมาแล้วในข้างต้นดูค่อนข้างยุ่งยาก และสลับซับซ้อน ซึ่งพอที่จะสรุปได้ดังนี้

$$\text{Partial} \quad F = \frac{S_1 - S_2}{(Y'Y - S_1)/(n-k-1)}$$

เมื่อ $S_1 = \hat{\beta}'_{XY}$ ของสมการถดถอยที่รวม X_j อยู่ด้วย

$S_2 = \hat{\beta}'_{XY}$ ของสมการถดถอยที่ไม่รวม X_j อยู่ด้วย

n = ขนาดตัวอย่าง

k = จำนวนตัวแปรอิสระ

2.8 สหสัมพันธ์ (Correlation)

สหสัมพันธ์ (Correlation) จะกล่าวถึงความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระแต่ละตัว ว่ามีความสัมพันธ์กันมากน้อยขนาดไหน ซึ่งค่าของสัมประสิทธิ์สหสัมพันธ์นั้นจะมีค่าอยู่ระหว่าง -1 และ 1 ถ้าค่าสัมประสิทธิ์สหสัมพันธ์เข้าใกล้ -1 หรือ 1 แสดงว่าตัวแปรตามและตัวแปรอิสระมีความสัมพันธ์กันมากในทางตรงกันข้ามกัน ถ้าค่าของสัมประสิทธิ์สหสัมพันธ์มีค่าเป็นลบ และจะมีความสัมพันธ์กันมากในทางเดียวกัน ถ้าค่าสัมประสิทธิ์สหสัมพันธ์มีค่าเป็นบวก ถ้าค่าของสัมประสิทธิ์สหสัมพันธ์มีค่าเป็นศูนย์หรือเข้าใกล้ศูนย์จะถือว่า ตัวแปรอิสระและตัวแปรตามไม่มีความสัมพันธ์หรือมีความสัมพันธ์กันน้อย ซึ่งสหสัมพันธ์มีรูปแบบดังนี้

2.8.1 สหสัมพันธ์อย่างง่าย (Simple Correlation)

เป็นการหาสหสัมพันธ์ระหว่างตัวแปรตามทั้งตัวแปรอิสระแต่ละตัว ซึ่งมีรูป มีรูปแบบ

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

n = ขนาดตัวอย่าง

Y = ตัวแปรตาม

X = ตัวแปรอิสระ

2.8.2 สหสัมพันธ์บางส่วน (Partial Correlation)

เป็นการหาค่าสหสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระแต่ละตัวที่ยังไม่รวมอยู่ในสมการถดถอยและเมื่อมีตัวแปรอิสระบางตัวเข้าอยู่ในสมการถดถอยแล้ว

$$R = \begin{bmatrix} 1 & r_{12} & r_{13} \cdots & r_{1k} \\ r_{21} & 1 & r_{23} \cdots & r_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ r_{k1} & r_{k2} & r_{k3} \cdots & 1 \end{bmatrix}$$

R = เมตริกซ์สหสัมพันธ์ของสัมประสิทธิ์สหสัมพันธ์ที่มีขนาด kxk

$$R^{-1} = C = \begin{pmatrix} C_{11} & C_{12} & C_{13} \cdots & C_{1k} \\ C_{21} & C_{22} & C_{23} \cdots & C_{2k} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ C_{k1} & C_{k2} & C_{k3} \cdots & C_{kk} \end{pmatrix}$$

R^{-1} หรือ C คือ ส่วนกลับ (Inverse) ของเมตริกซ์สัมประสิทธิ์สหสัมพันธ์ มีขนาด kxk ดังนั้น สัมประสิทธิ์สหสัมพันธ์บางส่วนระหว่างตัวแปรตามกับตัวแปรอิสระแต่ละตัวที่ยังไม่รวมอยู่ในสมการถดถอยและมีตัวแปรอิสระบางตัวเข้าอยู่ในสมการถดถอยแล้วคือ

$$r_{ij.1,2,3,\ldots,i-1,i+1,\ldots,j-1,j+1,\ldots,k} = \frac{-C_{ij}}{\sqrt{C_{ii}\, C_{jj}}}$$

2.9 การประมาณค่าสัมประสิทธิ์ถดถอยเชิงเส้นโดยวิธีกำลังสองน้อยที่สุด

วิธีการประมาณค่าสัมประสิทธิ์นี้มีรากฐานมาจากทฤษฎี การประมาณเชิงเส้น ที่คิดขึ้นโดยคาร์ล เฟรดริก เกาส์ (Karl Friedrick Gauss) ในปี ค.ศ.1777-1855 และ อังเดร แอนดรีวิช มาร์คอฟ (Andrei Andreevich Markov) ในปี ค.ศ. 1855-1922 โดยมีหลักในการประมาณค่าสัมประสิทธิ์ คือ ทำให้ผลรวมกำลังสองของความคลาดเคลื่อนมีค่าน้อยที่สุด ซึ่งแสดงรายละเอียดดังนี้

นิยาม 2.9.1 จากสมการ $\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$ เมื่อ $\underset{\sim}{\varepsilon} \sim N(0,\sigma^2 I)$ ตัวประมาณกำลังสองน้อยที่สุดของ $\underset{\sim}{\beta}$ คือ $\underset{\sim}{\hat{\beta}}$ ที่ทำให้ผลรวมกำลังสองของความคลาดเคลื่อนมีค่าน้อยที่สุด

ดังนั้นจากนิยาม 2.9.1 จะหาตัวประมาณกำลังสองน้อยที่สุดได้ดังนี้

เนื่องจาก

$$SSE = \underset{\sim}{e}' \underset{\sim}{e}$$

$$= (\underset{\sim}{Y} - X\hat{\underset{\sim}{\beta}})' (Y - X\hat{\underset{\sim}{\beta}})$$

$$= (\underset{\sim}{Y}'Y - 2\hat{\underset{\sim}{\beta}}'XY + \hat{\underset{\sim}{\beta}}'X'X\hat{\underset{\sim}{\beta}})$$

การหาค่าน้อยที่สุดของผลรวมกำลังสองของความคลาดเคลื่อนทำได้โดยการดิฟเฟอเรน-ชิเอท (Differentiate) เทียบกับ $\hat{\underset{\sim}{\beta}}$ แล้วกำหนดให้เท่ากับ 0 ดังนั้น

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

$$-2X'\underset{\sim}{Y} + 2X'X\hat{\underset{\sim}{\beta}} = 0$$

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

ดังนั้นในการศึกษาครั้งนี้จะทำการประมาณค่าสัมประสิทธิ์ถดถอยเชิงเส้นจาก

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

เมื่อ

$\underset{\sim}{Y}$ เป็นเวกเตอร์ของตัวแปรตามที่มีขนาด (nx1)

X เป็นเมตริกซ์ของตัวแปรอิสระที่มีขนาด (nxp)

n เป็นจำนวนข้อมูลทั้งหมด

p เป็นจำนวนพารามิเตอร์สัมประสิทธิ์ถดถอยเชิงเส้น

2.10 การทดสอบสมมติฐาน

ในการคำนวณหาค่าประมาณของพารามิเตอร์และได้สมการถดถอย

$$\hat{Y} = \hat{\beta}_o + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k$$

การทดสอบสมมติฐานที่มั่นใจคือ การทดสอบว่าค่าที่แท้จริงของ β_1 , β_2 , β_3, ..., β_k มีค่าเท่ากับศูนย์หรือไม่ ถ้าไม่เท่ากับศูนย์แสดงว่า Y และ X_1 , X_2 , X_3, ... X_k มีความสัมพันธ์กัน และสมการถดถอยนั้น จะเป็นสมการที่สามารถใช้พยากรณ์ค่าเฉลี่ยของ Y เมื่อกำหนด X_1 , X_2..., X_k แต่ถ้าค่า β_1 , β_2, ... , β_k มีค่าเท่ากับศูนย์หรือมีบางค่าเท่ากับศูนย์ จะทำสมการถดถอยใช้พยากรณ์ค่าเฉลี่ยของ Y ได้ไม่ดีเท่าที่ควร จึงจำเป็นต้องมีการคัดเลือกตัวแปรอิสระที่ให้ค่าสัมประสิทธิ์ถดถอยเป็นศูนย์ ออกจากสมการถดถอย โดยใช้การทดสอบสมมติฐาน ซึ่งมีรายละเอียดต่อไปนี้

สมมติฐาน H_o : $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

H_1 : β_j บางตัวไม่เท่ากับศูนย์ ; j = 1,2,3,..., k

ค่าสถิติที่ใช้ในการทดสอบคือ

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

ค่าของ MSR และ MSE สามารถคำนวณได้จากตารางวิเคราะห์ความแปรปรวน ดังตารางที่ 2.1

ตารางที่ 2.1 แสดงการวิเคราะห์ความแปรปรวน

Source	Sum of Square	Degree of freedom	mean Square	F
Regression	SSR	k	$MSR = \frac{SSR}{k}$	$\frac{MSR}{MSE}$
Eror	SSE	n-k-1	$MSE = \frac{SSE}{n-k-1}$	
Total	SST	n-1		

ถ้าสถิติ F ที่คำนวณได้มีค่าน้อยกว่า F จากตารางที่องค่าแห่งความอิสระ k และ n-k-1 ด้วยระดับนัยสำคัญ α และจะยอมรับสมมติฐาน

$H_o : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ ซึ่งหมายความว่าสมการถดถอยเป็นสมการที่ไม่เหมาะสมที่จะใช้พยากรณ์ค่า Y เนื่องจาก Y และ $X_1, X_2, X_3, \ldots, X_k$ ไม่มีความสัมพันธ์กันเลย

ในกรณีที่ค่าสถิติ F ที่ได้จากการคำนวณมีค่ามากกว่าF จากตารางก็จะเป็นการปฏิเสธสมมติฐาน H_o: ซึ่งหมายความว่าสัมประสิทธิ์ถดถอยบางตัวมีค่าไม่เท่ากับศูนย์ จึงควรทดสอบดูว่าสัมประสิทธิ์ถดถอยตัวใดมีค่าเป็นศูนย์บ้าง โดยการใช้พาเชียลเอฟ ดังที่กล่าวมาแล้วในตอนต้นของบทนี้

2.11 ตัวสถิติที่ใช้เป็นเกณฑ์ในการเปรียบเทียบ

ในการศึกษาวิจัยครั้งนี้ผู้วิจัยได้เปรียบเทียบสมการถดถอยที่ได้จากวิธีการเลือกเอาตัวแปรอิสระเข้าสู่สมการถดถอย ซึ่งมีทั้งหมด 5 วิธีคือ วิธีการกำจัดตัวแปรแบบถอยหลัง การเลือกตัวแปรแบบไปข้างหน้า การถดถอยแบบขั้นบันได การถดถอยแบบขั้นตอน การกำจัดตัวแปรโดยใช้สัมประสิทธิ์สหสัมพันธ์ ในการคัดเลือกตัวแปรอิสระเข้าสู่สมการถดถอยทั้ง 5 วิธีดังกล่าวนั้น จะมีวิธีการคัดเลือกแตกต่างกันออกไป ดังนั้นผลที่ได้ออกมาในขั้นสุดท้ายนั้นย่อมให้ผลที่แตกต่างกันออกไป หรืออาจจะให้ผลเหมือนกันก็ได้ในบางโอกาส ดังนั้นการที่จะทราบว่าวิธีไหนจะให้ผลดีกว่า นั้นจะต้องมีกฎเกณฑ์อะไรสักอย่าง เพื่อใช้เป็นเกณฑ์ในการเปรียบเทียบ ซึ่งในการวิจัยครั้งนี้ ได้ใช้ตัวสถิติ 3 ตัวเป็นเกณฑ์ในการเปรียบเทียบคือ

2.11.1 ผลรวมกำลังสองของความคลาดเคลื่อน

ผลรวมกำลังสองของความคลาดเคลื่อน (SSE) เป็นตัวสถิติตัวหนึ่งที่ใช้วัดค่าความแตกต่างระหว่างค่าจริงกับค่าประมาณหรือใช้วัดอำนาจการพยากรณ์ของสมการถดถอยว่าสูงต่ำเพียงไหน ซึ่งมีรูปแบบดังนี้

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

หรือ $$SSE = \underset{\sim}{Y}'\underset{\sim}{Y} - \underset{\sim}{\beta}'\underset{\sim}{X}'\underset{\sim}{Y}$$

Y_i = ค่าจริงของตัวแปรตาม

$\hat{Y}_i$ = ค่าประมาณของตัวแปรตาม

ค่าผลรวมกำลังสองของความคลาดเคลื่อน จะมีค่ามากกว่าหรือเท่ากับศูนย์เสมอ ถ้าค่าจริง Y_i และค่าประมาณของ Y_i มีค่าเดียวกันหรือมีค่าไม่แตกต่างกันเลย ดังนั้นค่าของ SSE จะเป็นศูนย์ ซึ่งถือได้ว่าสมการถดถอยที่ได้นั้นจะมีอำนาจการพยากรณ์สูงสุด และอำนาจการพยากรณ์ของสมการถดถอยจะลดลงไปเรื่อย ๆ ถ้าค่าของ SSE เพิ่มขึ้นเรื่อย ๆ ดังนั้นในการเปรียบเทียบ วิธีการคัดเลือกตัวแปรอิสระเข้าสู่สมการถดถอยนั้น จะถือว่าวิธีไหนให้ค่าของ SSE ต่ำที่สุด จะถือว่าเป็นวิธีที่ดีที่สุด

2.11.2 ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง

ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (MSE) เป็นตัวสถิติที่ได้มาจากตัวสถิติ SSE เพียงแต่ว่านำเอาค่าของ SSE มาเฉลี่ยด้วยองศาแห่งความอิสระ (Degree of Freedom) ซึ่งมีรูปแบบดังนี้

$$MSE = \sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i)^2}{n-k-1}$$

หรือ

$$MSE = \frac{\underset{\sim}{Y}'\underset{\sim}{Y} - \underset{\sim}{\beta}'\underset{\sim}{X}'\underset{\sim}{Y}}{n-k-1}$$

Y_i = ค่าจริงของตัวแปรตาม

$\hat{Y}_i$ = ค่าประมาณของตัวแปรตาม

n = จำนวนตัวอย่างของข้อมูล

k = จำนวนตัวแปรอิสระ

ค่าของ MSE จะมีค่ามากกว่าหรือเท่ากับศูนย์เสมอ และลักษณะที่ใช้MSE เป็นเกณฑ์ ในการเปรียบเทียบว่าวิธีไหนจะให้ผลดีที่สุดนั้นก็เหมือนกับค่าของ SSE ที่ใช้เป็นเกณฑ์

2.11.3 ตัวสถิติธีล(Theil's U Statistic)

ในการทดสอบว่าสมการถดถอย $\hat{Y} = X\hat{\beta}$ มีอำนาจการพยากรณ์สูงต่ำเพียงใดนั้น นอกจากจะใช้ SSE หรือ MSE แล้วยังสามารถใช้ค่าของตัวสถิติ U ซึ่งมีรูปแบบคือ

$$U = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n}(p_t - a_t)^2}{\frac{1}{n}\sum_{i=1}^{n}a_t^2}}$$

ซึ่ง P_t = ค่าประมาณของ Y ณ วาระที่ t

A_t = ค่าจริงของ Y ณ วาระที่ t

ให้ $p_t = \frac{P_t - A_t}{A_{t-1}}$ คืออัตราการเปลี่ยนแปลงของ Y ที่พยากรณ์ใช้สำหรับวาระที่ t

เมื่อเปรียบเทียบกับค่าจริงของ Y ในวาระก่อน

ให้ $a_t = \frac{A_t - A_{t-1}}{A_{t-1}}$ คือ อัตราการเปลี่ยนแปลงค่าจริงของ Y ที่ปรากฏขึ้นสำหรับวาระที่ t

เมื่อเปรียบเทียบกับค่าจริงของ Y ในวาระก่อน

ฉะนั้น $p_t - a_t$ คือ อัตราการความคลาดเคลื่อนระหว่างอัตราการเปลี่ยน-แปลงของค่าพยากรณ์กับ อัตราการเปลี่ยนแปลงที่เกิดขึ้นจริงของตัวแปรตาม Y ในการพิจารณาค่าของ U จะพบว่า

2.11.3.1 เมื่อ $p_t = a_t$ จะพบว่า U=0 แสดงว่าสมการถดถอย $Y = X\hat{\beta}$ มีอำนาจการพยากรณ์สูงสุด (Perfect Forecast)

2.11.3.2 ถ้า $p_t = 0$ จะพบว่า U = 1 เมื่อพิจารณาสมการจะพบว่า $\frac{P_t - A_{t-1}}{A_{t-1}} = 0$ หรือ $P_t = A_{t-1}$ หรือ $P_{t-1} = A_t$ แต่ P_{t+1} ก็คือ ค่าพยากรณ์ของ Y ณ วาระที่ t+1 และ A_t คือ ค่าจริงของ Y ณ วาระที่ t ดังนั้น ถ้า U=1 แสดงว่าสมการถดถอย $\hat{Y} = X\hat{\beta}$ พยากรณ์ค่า Y ณ วาระที่ t+1 คือ ค่า Y_{t+1} ด้วย Y_t หรืออีกนัยหนึ่ง u=1 แสดงว่าสมการถดถอย $\hat{Y} = X\hat{\beta}$ พยากรณ์ค่า Y ในวาระที่ t+1 ว่ามีค่าไม่แตกต่างไปจากเดิมในวาระที่ t

2.11.3.3 ถ้า p_t และ a_t มีค่าต่างไปจากข้อ (2.11.3.1) และ (2.11.3.2) คือค่าของ U จะสูงหรือ $U \to \infty$ ดังนั้นสามารถสรุปได้ว่า $0 \leq U < \infty$ โดย U = 0 สมการถดถอย $\hat{Y} = X\hat{\beta}$ จะมีอำนาจการพยากรณ์สูงสุด ถ้า 0<U< 1 สมการถดถอย $\hat{Y} = X\hat{\beta}$ จะมีอำนาจการพยากรณ์อยู่ในเกณฑ์ดี ถ้า u=1

สมการถดถอย $\hat{Y} = X\hat{\beta}$ จะพยากรณ์ Y_{t+1} ไม่แตกต่างไปจากค่า Y_t ถ้า U > 1

สมการถดถอย $\hat{Y} = X\hat{\beta}$ จะมีอำนาจการพยากรณ์ต่ำ ถ้าค่าU มีค่ามากกว่า 1 มากเพียงใด

สมการถดถอย $\hat{Y} = X\hat{\beta}$ จะมีอำนาจการพยากรณ์ต่ำเพียงนั้น