บทที่ 3

วิธีการวิจัย

การวิจัยครั้งนี้เพื่อเปรียบเทียบประสิทธิภาพในการคัดเลือกตัวแบบ ของตัวแบบสมการถดถอยเชิงเส้นพหุคูณ 5 วิธี คือ

1) การคัดเลือกตัวแบบโดยใช้เกณฑ์ข้อสนเทศของอาไคเคะที่ปรับค่าความเสี่ยงโดยใช้ตัวประมาณของข้อสนเทศคูลส์แบล็ค-ไลท์เบอร์ (Kullback Leibler Information) หรือ AIC_{ri}

2) การคัดเลือกตัวแบบที่พัฒนาเกณฑ์ AIC_C ที่ลดความเอนเอียงโดยใช้การแยกออกของความสมมาตรของข้อสนเทศคูลส์แบล็ค-ไลท์เบอร์ (Kullback's Symmetric Divergence) หรือ KIC_C

3) การคัดเลือกตัวแบบที่พัฒนาเกณฑ์ AIC_I ที่ลดความเอนเอียงโดยใช้การแยกออกของความสมมาตรของข้อสนเทศคูลส์แบล็ค-ไลท์เบอร์ (Kullback's Symmetric Divergence) หรือ KIC_I

4) การคัดเลือกตัวแบบที่พัฒนา Cross-Validation (C_p) ที่ใช้สำหรับตัวอย่างขนาดเล็ก หรือ G_n^{CV} ที่ใช้ $C_n^{(R)}$

5) การคัดเลือกตัวแบบโดยใช้สถิติทดสอบเอฟบางส่วน (The Partial F-test Statistic) ด้วยวิธีการถดถอยแบบขั้นบันได (Stepwise Regression Method)

มีขั้นตอนการวิจัยดังนี้

ขั้นตอนการวิจัย

การวิจัยครั้งนี้เป็นการศึกษาจากการจำลองค่า (Simulation Study) โดยใช้โปรแกรม S-plus 2000 ในการประมวลผล มีขั้นตอนการดำเนินงานดังนี้

1 กำหนดค่าพารามิเตอร์ของสมการถดถอยเชิงเส้น

กรณีตัวแปรอิสระ 3 ตัว $(\beta_0,\beta_1,\beta_2,\beta_3)$ เป็น $(1,1,1,1)$

กรณีตัวแปรอิสระ 6 ตัว $(\beta_0,\beta_1,...,\beta_6)$ เป็น $(1,1,1,1,1,1,1)$

กรณีตัวแปรอิสระ 9 ตัว $(\beta_0,\beta_1,...,\beta_9)$ เป็น $(1,1,1,1,1,1,1,1,1,1)$

กรณีตัวแปรอิสระ 12 ตัว $(\beta_0,\beta_1,...,\beta_{12})$ เป็น $(1,1,1,1,1,1,1,1,1,1,1,1,1)$

2 กำหนดขนาดตัวอย่าง (n) ขึ้นอยู่กับตัวแปรอิสระ ดังนี้

กรณีจำนวนตัวแปรอิสระ 3 ตัว ขนาดตัวอย่างเป็น 10 , 20 , 30 และ 40

กรณีจำนวนตัวแปรอิสระ 6 ตัว ขนาดตัวอย่างเป็น 10 , 20 , 30 และ 40

กรณีจำนวนตัวแปรอิสระ 9 ตัว ขนาดตัวอย่างเป็น 15 , 25 , 35 และ 45

กรณีจำนวนตัวแปรอิสระ 12 ตัว ขนาดตัวอย่างเป็น 20 , 30 , 40 และ 50

3. จำลองค่าเมตริกซ์ $\mathbf{X} = \left[1 \quad X_{ij}\right]_{nxp}$ เมื่อ i=1,2,...,p-1 , j=1,2,...,n

4. จำลองค่าความคลาดเคลื่อนสุ่ม มีการแจกแจงแบบปกติมีค่าเฉลี่ยเท่ากับ 0 และ ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 1 , 5 และ 10

5. คำนวณตัวแปรตามจากความสัมพันธ์ $\mathbf{Y} = \mathbf{X\beta} + \mathbf{e}$

6. คัดเลือกตัวแบบจากวิธีต่างๆ ดังนี้

- เกณฑ์การคัดเลือกตัวแบบโดยใช้เกณฑ์ข้อสนเทศของอาไคเคะที่ปรับค่าความเสี่ยง โดยใช้ตัวประมาณของ ข้อสนเทศคูลส์แบล็ค-ไลท์เบอร์ (Kullback Leibler Information)

1. กำหนดตัวแบบที่เป็นไปได้ทั้งหมด

2. คำนวณค่าความแปรปรวนของแต่ละตัวแบบ ดังนี้

$$\hat{\sigma}^2 = (\mathbf{Y} - \mathbf{X\hat{\beta}})'(\mathbf{Y} - \mathbf{X\hat{\beta}})/n$$

3. คำนวณ AIC_{ri} แต่ละตัวแบบ ดังนี้

$$AIC_{ri} = n \ln \hat{\sigma}^2 + \frac{n(n+p)}{n-p-2} - \frac{2np}{(n-p-2)(n-p)} n^{\alpha-\varepsilon}$$

กำหนดให้ $\alpha = 3/4$ และ $\varepsilon = 2/5$

4. เปรียบเทียบค่า AIC_{ri} แต่ละตัวแบบ โดยเลือกตัวแบบที่ให้ค่า AIC_{ri} ต่ำที่สุด เป็นตัวแบบที่ดีที่สุด

- เกณฑ์การคัดเลือกตัวแบบที่พัฒนา AIC_C ที่ลดความเอนเอียงโดยใช้การแยกออกของความสมมาตรของข้อสนเทศคูลส์แบล็ค-ไลท์เบอร์ (Kullback's Symmetric Divergence)

1. กำหนดตัวแบบที่เป็นไปได้ทั้งหมด

2. คำนวณค่าความแปรปรวนของแต่ละตัวแบบ ดังนี้

$$\hat{\sigma}^2 = (\mathbf{Y} - \mathbf{X\hat{\beta}})'(\mathbf{Y} - \mathbf{X\hat{\beta}})/n$$

3. คำนวณ KIC_C แต่ละตัวแบบ ดังนี้

$$KIC_C = n \ln \hat{\sigma}^2 + n \ln \frac{n}{n-p} + \frac{n[(n+p)(n-p)+(n-p-2)]}{(n-p-2)(n-p)}$$

4. เปรียบเทียบค่า KIC_C แต่ละตัวแบบ โดยเลือกตัวแบบที่ให้ค่า KIC_C ต่ำที่สุด เป็นตัวแบบที่ดีที่สุด

- เกณฑ์การคัดเลือกตัวแบบที่พัฒนา AIC_I ที่ลดความเอนเอียงโดยใช้การแยกออกของความสมมาตรของข้อสนเทศคูลส์แบล็ค-ไลท์เบอร์ (Kullback's Symmetric Divergence)

1. กำหนดตัวแบบที่เป็นไปได้ทั้งหมด

2. คำนวณค่าความแปรปรวนของแต่ละตัวแบบ ดังนี้

$$\hat{\sigma}^2 = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta})/n$$

3. จำลองข้อมูลด้วยเทคนิคมอนติคาร์โล ของเทอมปรับค่าความเอนเอียง ด้วยการกำหนดค่า $\beta^+ = (0,0,...,0)$ และ $\sigma^2 = 1$ ดังนั้น เทอมที่นำมาจำลองค่า คือ

$$Bias_{adj} = \frac{1}{R}\sum_{j=1}^{R}\{[-n\ln\hat{\sigma}^2(j) + n/\hat{\sigma}^2(j)] + [(\mathbf{X}\hat{\beta}(j))'(\mathbf{X}\hat{\beta}(j))/\hat{\sigma}^2(j)]$$
$$+ [n\hat{\sigma}^2(j) + (\mathbf{X}\hat{\beta}(j))'(\mathbf{X}\hat{\beta}(j))] - 2n\}$$

เมื่อ R เป็นจำนวนการสุ่มซ้ำ และกำหนดให้ R=100

4. คำนวณ KIC_I แต่ละตัวแบบ ดังนี้

$$KIC_I = n\ln\hat{\sigma}^2 + n + Bias_{adj}$$

5. เปรียบเทียบค่า KIC_I แต่ละตัวแบบ โดยเลือกตัวแบบที่ให้ค่า KIC_I ต่ำที่สุด เป็นตัวแบบที่ดีที่สุด

- เกณฑ์การคัดเลือกตัวแบบ Cross-Validation (C_p) ที่ใช้สำหรับตัวอย่างขนาดเล็ก หรือ G_n^{CV} ที่ใช้ $C_n^{(R)}$

1. กำหนดตัวแบบที่เป็นไปได้ทั้งหมด

2. คำนวณตัวประมาณสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยสุด ของตัวแบบเต็มรูป (Full Model) ซึ่งประกอบไปด้วยตัวแปรอิสระทุกตัว ดังนี้

$$\tilde{\beta} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$$

3. คำนวณค่าความคลาดเคลื่อน ดังนี้

$$\hat{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\tilde{\beta}$$

4. สร้างสัมประสิทธิ์การถดถอยของตัวแบบใหม่ คือ

$$\mathbf{u} = \mathbf{X}\bar{\beta} + \hat{\mathbf{e}}$$

โดยที่ $\bar{\beta} = (\bar{\beta}_0,...,\bar{\beta}_{p-1})'$

$$\text{และ } \bar{\beta}_i = \begin{cases} \tilde{\beta}_i & \text{if } |\tilde{\beta}_i| \geq k \\ k(\text{sign}\tilde{\beta}_i) & \text{if } 0 < |\tilde{\beta}_i| < k \\ k & \text{if } |\tilde{\beta}_i| = 0 \end{cases} \quad \text{เมื่อ } i = 0,\ldots,p\text{-}1$$

กำหนดให้ $k = 0.01$

5. คำนวณ $D_{n0} = \sum_{i=1}^{n}(u_i - \hat{u}_{-i})^2$, i=1,2,...n

เมื่อ $\hat{u}_{-i}$ เป็นค่าพยากรณ์ของตัวแบบใหม่ จากเมตริกซ์ตัวแปรอิสระ $\mathbf{X}$ ตัดแถวที่ i

6. คำนวณ $D_n(h) = \sum_{i=1}^{n}(u_i - \hat{u}_{-i,-h})^2$ เมื่อ i=1,2,...n , h = 1,2,...,p-1

มื่อ $\hat{u}_{-i,-h}$ เป็นค่าพยากรณ์ของตัวแบบใหม่ที่ตัดข้อมูลชุดที่ i และ ตัวแปรอิสระตัวที่ h

7. คำนวณ $\Delta_n(h) = D_n(h) - D_{n0}$

8. เลือกฟังก์ชัน C_n จาก $C_n^{(R)}$ ดังนี้

$C_n^{(R)}$ = ค่ามากที่สุดของ {ค่าต่ำสุดของ $\Delta_n(h), h = 1,2,...,p-1$, τ}

$/(1+\sqrt{\lfloor 0.000 \ln \rfloor})$

กำหนดให้ $\tau = 10$

9. คำนวณหา $\hat{y}_{-i}$ เมื่อ i=1,2,...,n แต่ละตัวแบบ

หรือ $\hat{y}_{-i} = X_i\hat{\beta}_{-i}$ และ $\hat{\beta}_{-i} = (X_i'X_i)^{-1}(X_i'y_{n,-i})$

เมื่อ $\hat{y}_{-i}$ แทน ค่าพยากรณ์ของตัวแบบของเมตริกซ์ตัวแปรอิสระ $\mathbf{X}$ ที่ตัดแถวที่ i

และ $\hat{\beta}_{-i}$ แทน ตัวประมาณกำลังสองน้อยที่สุด จากข้อมูล

$\{ (y_1,X_1) , \ldots , (y_{i-1},X_{i-1}) , (y_{i+1},X_{i+1}) , \ldots , (y_n,X_n) \}$

10. คำนวณ $G_n^{CV} = \sum_{j=1}^{n}(y_i - \hat{y}_{-i})^2 + pC_n^{(R)}$ แต่ละตัวแบบ

11. เปรียบเทียบค่า G_n^{CV} ที่ใช้ $C_n^{(R)}$ แต่ละตัวแบบ โดยเลือกที่ให้ค่า G_n^{CV} ที่ใช้ $C_n^{(R)}$ ต่ำสุด เป็นตัวแบบที่ดีที่สุด

- เกณฑ์การคัดเลือกตัวแบบโดยใช้สถิติทดสอบเอฟบางส่วน (The Partial F-test Statistic) ด้วยวิธีการถดถอยแบบขั้นบันได (Stepwise Regression Method)

1. คำนวณ $r_{x_i,y} = 1 - \frac{SSE}{SST}$ เมื่อ i=1,2,..,p เพื่อหาความสัมพันธ์ของตัวแปรอิสระแต่ละตัวกับตัวแปรตามที่มากที่สุด และนำตัวแปรอิสระนั้นเข้าสู่ตัวแบบเป็นตัวแรก

2. สร้างสมการถดถอยของตัวแปรอิสระที่ได้จากขั้นตอนที่ 1 กับตัวแปรอิสระแต่ละตัวที่เหลืออยู่ แล้วเลือกตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอยสูงที่สุด และมีนัยสำคัญทางสถิติเข้าสู่ตัวแบบ

3. ทำการทดสอบสถิติเอฟบางส่วนของตัวแบบที่มี 2 ตัวแปรอิสระ ที่ได้จากขั้นตอนที่ 1 และ 2 เพื่อตรวจสอบตัวแปรอิสระที่มีค่าสถิติทดสอบเอฟบางส่วนน้อยที่สุด และไม่มีนัยสำคัญทางสถิติ ออกจากตัวแบบ

4. ทำซ้ำขั้นตอนที่ 1-3 จนกว่าพบเงื่อนไขข้อใดข้อหนึ่ง ดังต่อไปนี้

    1) ไม่มีตัวแปรอิสระที่สามารถเข้าหรือออกจากตัวแบบในขั้นถัดกัน

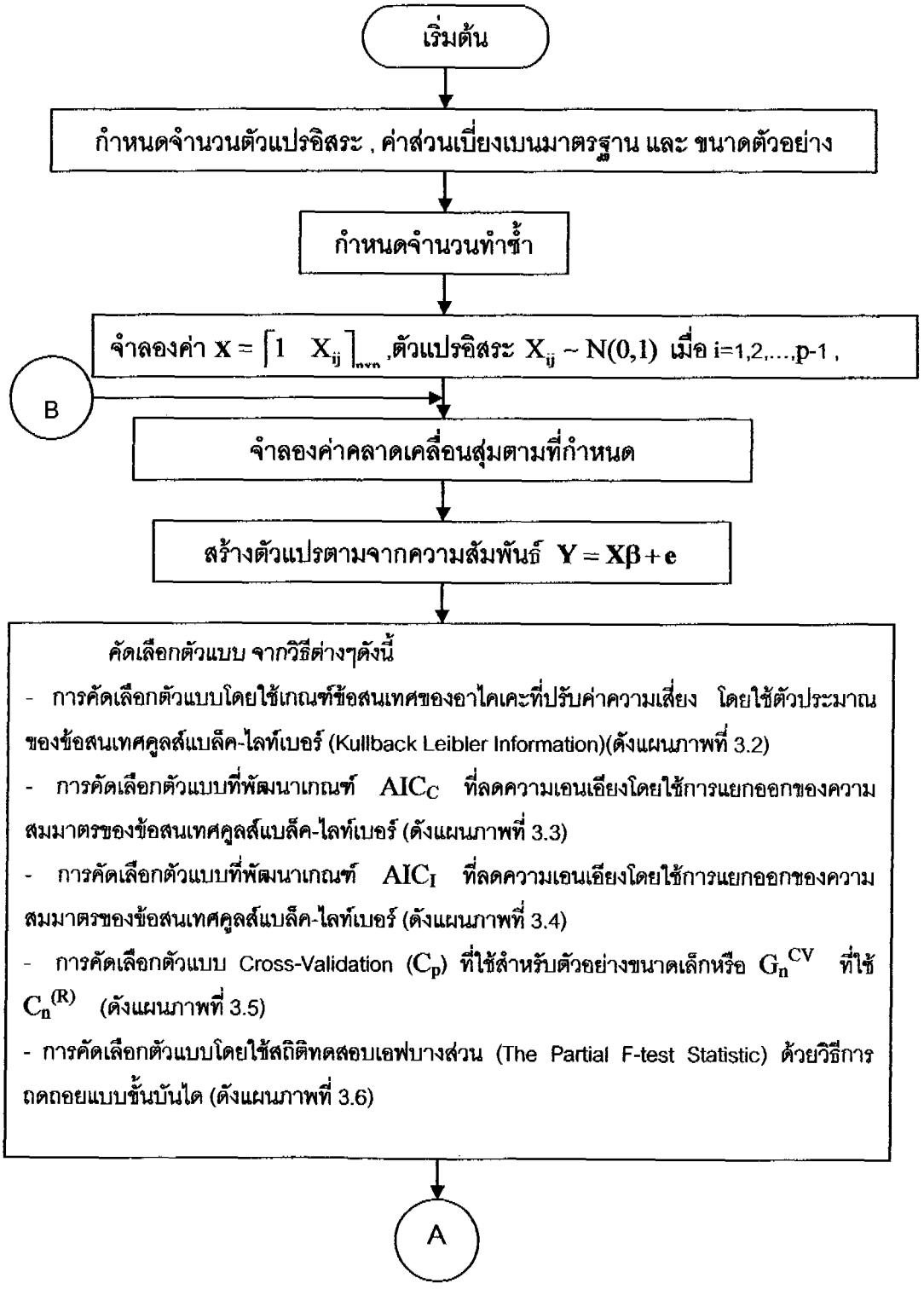
    2) ตัวแปรอิสระที่เข้าและออกจากตัวแบบในขั้นถัดกันเป็นตัวแปรเดียวกัน

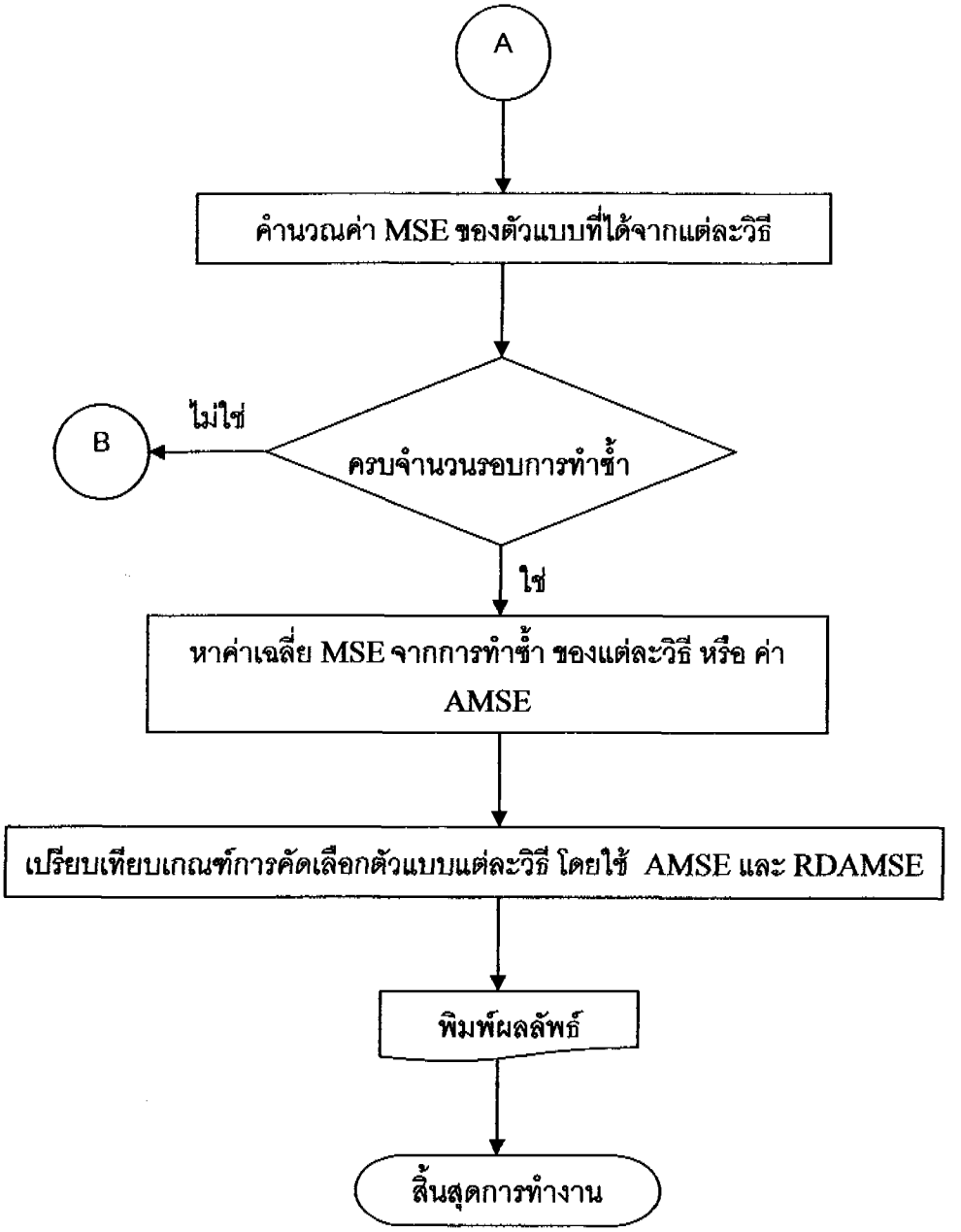
    3) ในขั้นตอนการออกไม่มีตัวแปรอิสระเหลืออยู่ในตัวแบบ

    4) ในขั้นตอนการเข้าไม่เหลือตัวแปรอิสระที่ไม่อยู่ในสมการถดถอย

7. คำนวณค่า MSE ของตัวแบบที่ได้จากแต่ละวิธี
8. ทำซ้ำจำนวน 500 ครั้ง
9. หาค่าเฉลี่ย MSE จากการทำซ้ำ 500 ครั้งของแต่ละวิธี หรือ AMSE และ RDAMSE
10. เปรียบเทียบเกณฑ์การคัดเลือกตัวแบบ แต่ละวิธี โดยใช้ AMSE และ RDAMSE
11. สรุปผล

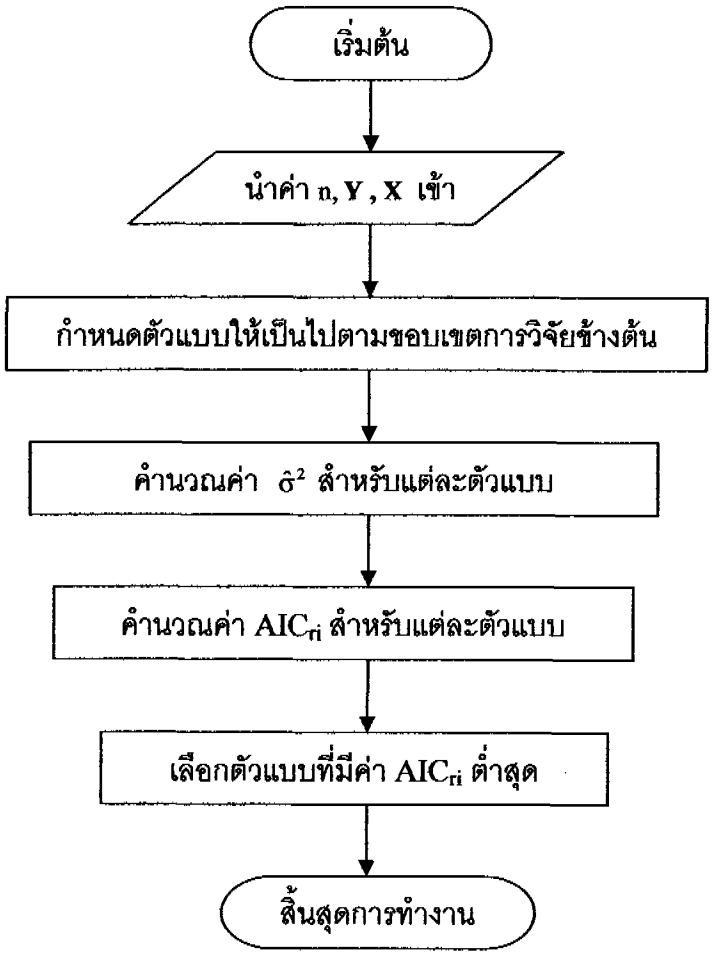
แผนภาพที่ 3.1
แสดงขั้นตอนการดำเนินการวิจัย




A
คำนวณค่า MSE ของตัวแบบที่ได้จากแต่ละวิธี
ครบจำนวนรอบการทำซ้ำ
ไม่ใช่
B
ใช่
หาค่าเฉลี่ย MSE จากการทำซ้ำ ของแต่ละวิธี หรือ ค่า AMSE
เปรียบเทียบเกณฑ์การคัดเลือกตัวแบบแต่ละวิธี โดยใช้ AMSE และ RDAMSE
พิมพ์ผลลัพธ์
สิ้นสุดการทำงาน

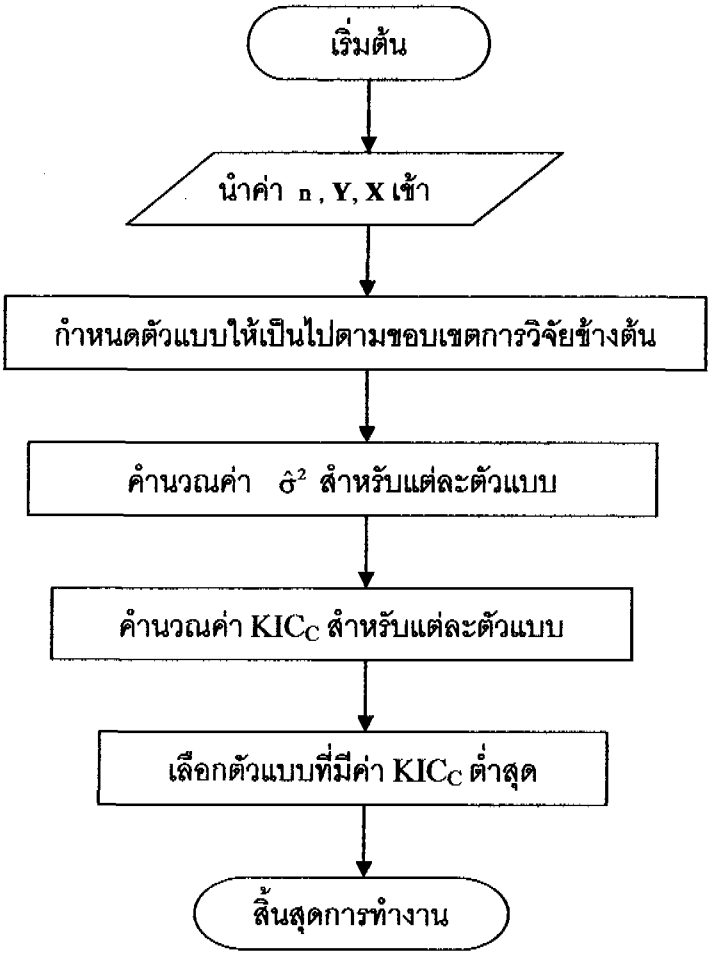
แผนภาพที่ 3.2

แสดงขั้นตอนวิธีการคัดเลือกตัวแบบโดยใช้เกณฑ์ข้อสนเทศของอาไคเคะ ที่ปรับค่าความเสี่ยงโดยใช้ตัวประมาณข้อสนเทศคูลแบล็ค-ไลท์เบอร์ หรือ AIC_{ri}



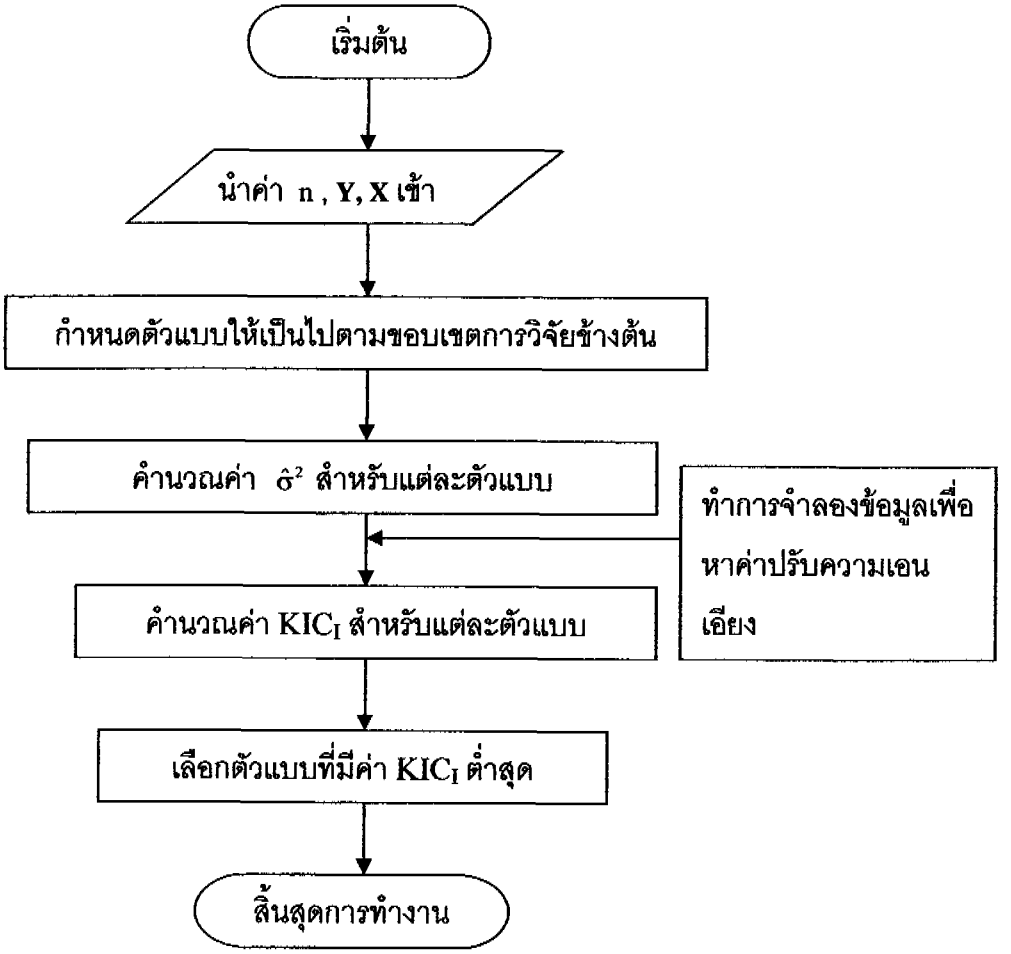
แผนภาพที่ 3.3

แสดงขั้นตอนวิธีการคัดเลือกตัวแบบที่พัฒนาเกณฑ์ AIC_C ที่ลดความเอนเอียงโดยใช้ การแยกออกของความสมมาตรของคูลส์แบล็ค-ไลท์เบอร์ (Kullback's symmetric divergence) หรือ KIC_C



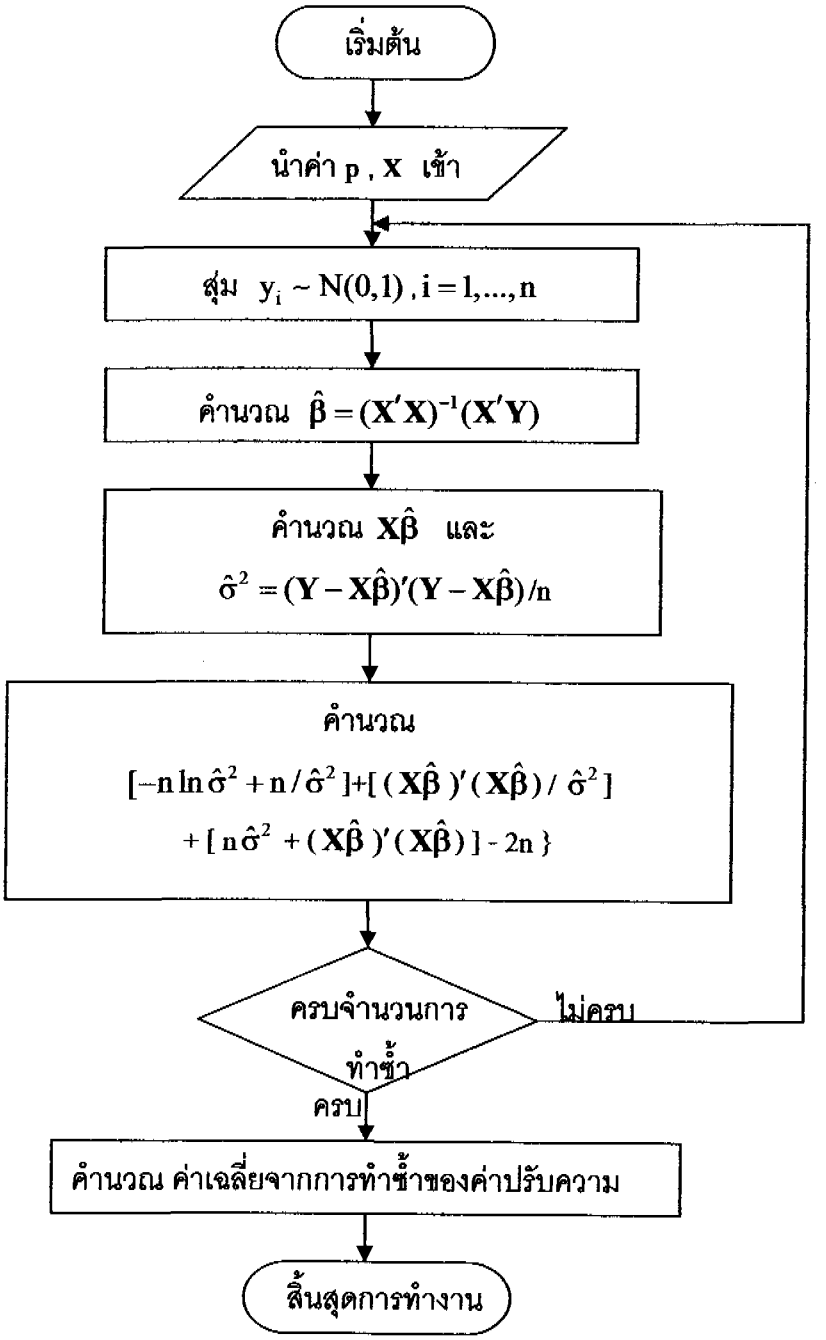
แผนภาพที่ 3.4

แสดงขั้นตอนวิธีการคัดเลือกตัวแบบที่พัฒนาเกณฑ์ AIC_I ที่ลดความเอนเอียงโดยใช้ การแยกออกของความสมมาตรของคูลส์แบล็ค-ไลท์เบอร์ (Kullback's symmetric divergence) หรือ KIC_I



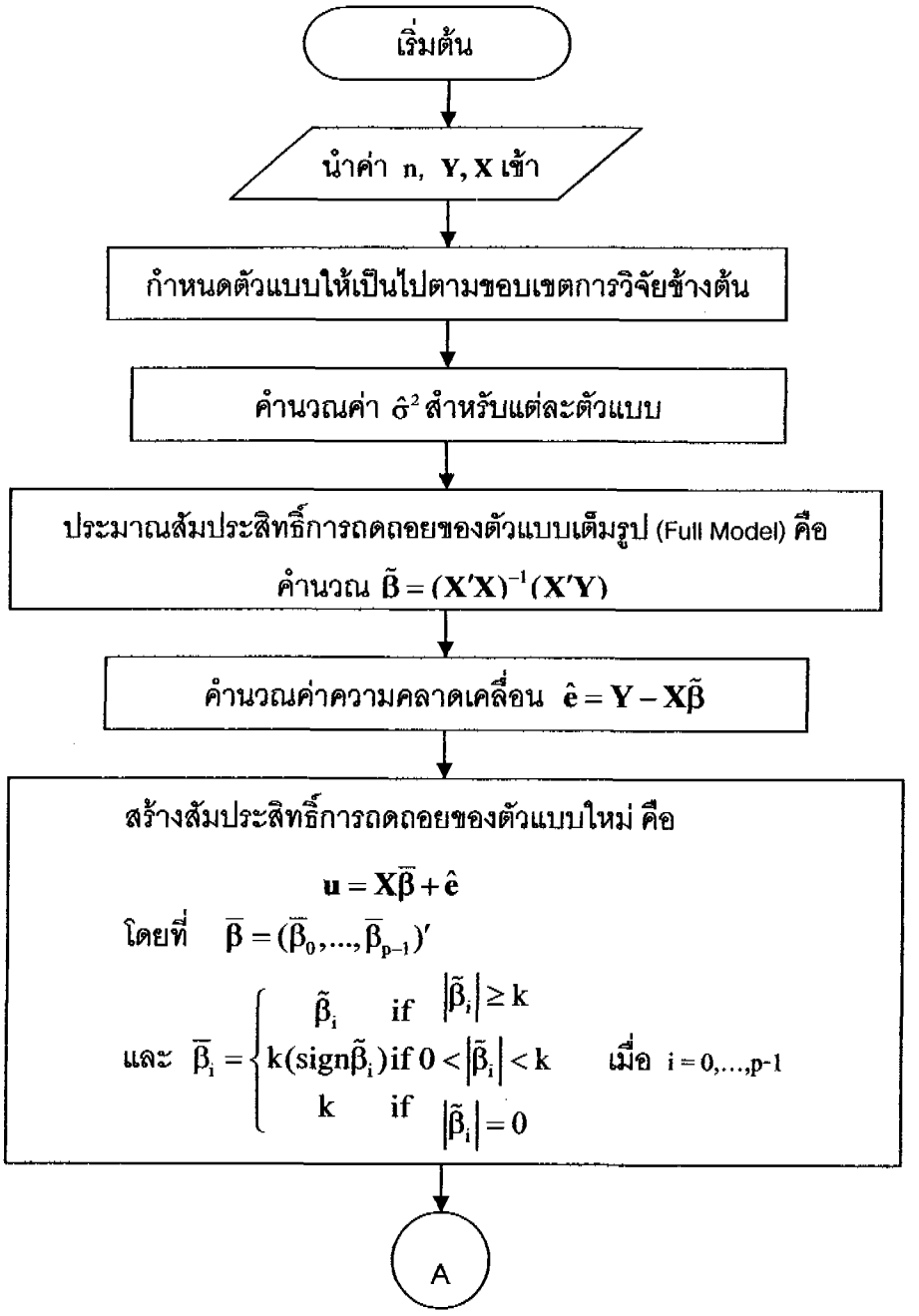
แผนภาพที่ 3.4.1

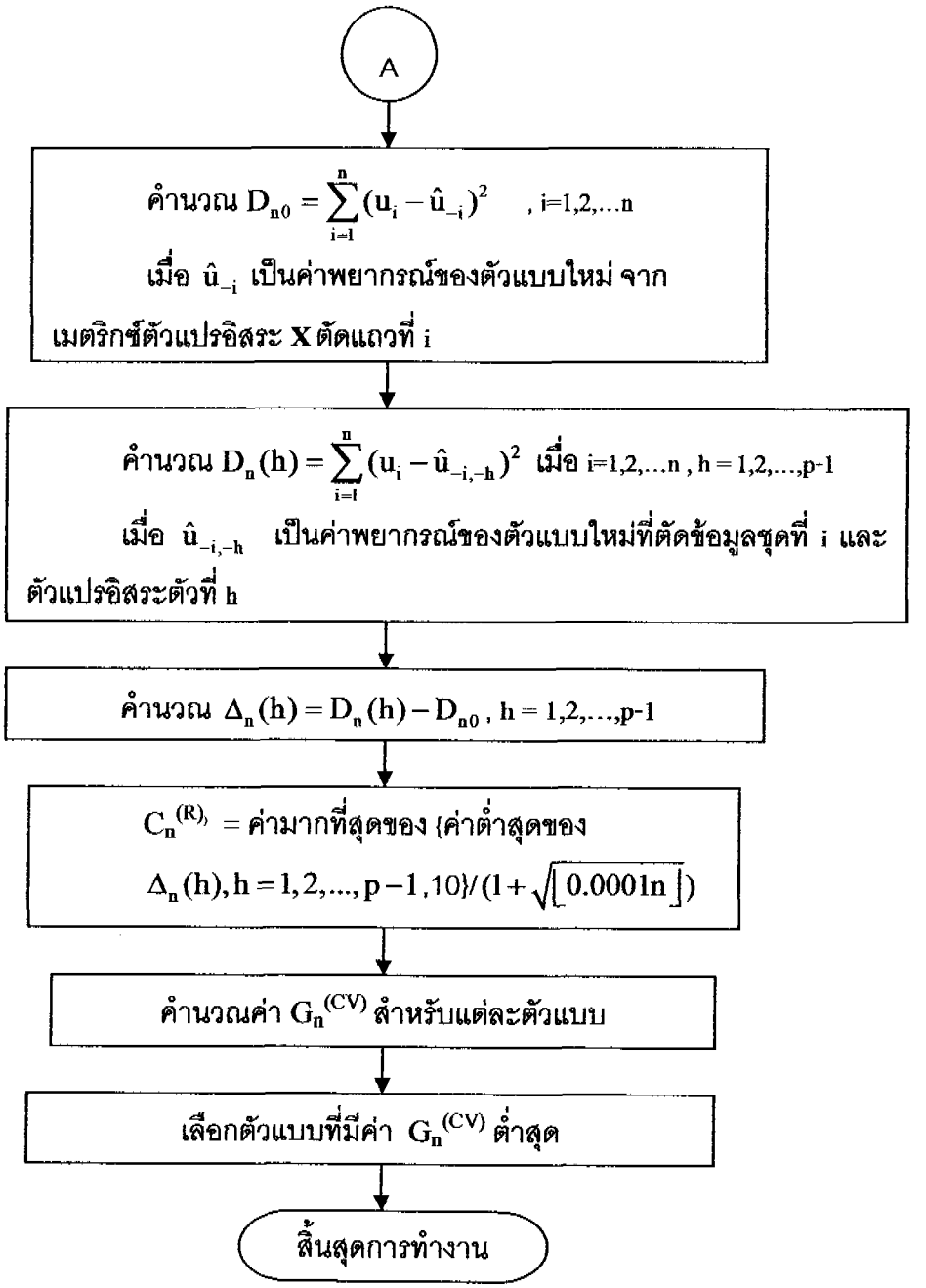
แสดงขั้นตอนการจำลองข้อมูลเพื่อหาค่าปรับความเอนเอียงของเกณฑ์ KIC_I



แผนภาพที่ 3.5

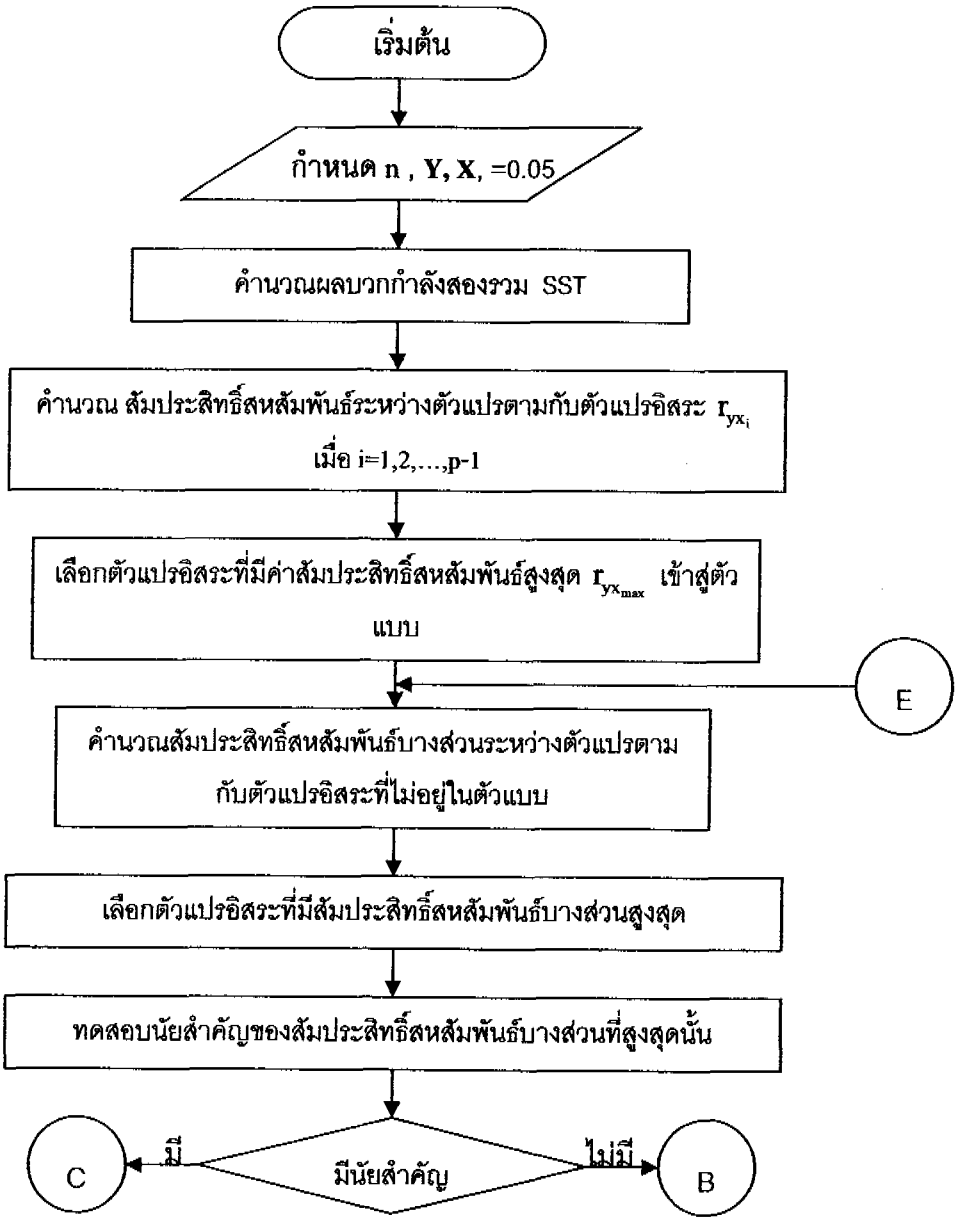
แสดงขั้นตอนวิธีการคัดเลือกตัวแบบ Cross-Validation (C_p) ที่ใช้สำหรับขนาดตัวอย่างเล็ก หรือ G_n^{CV} ที่ใช้ $C_n^{(R)}$

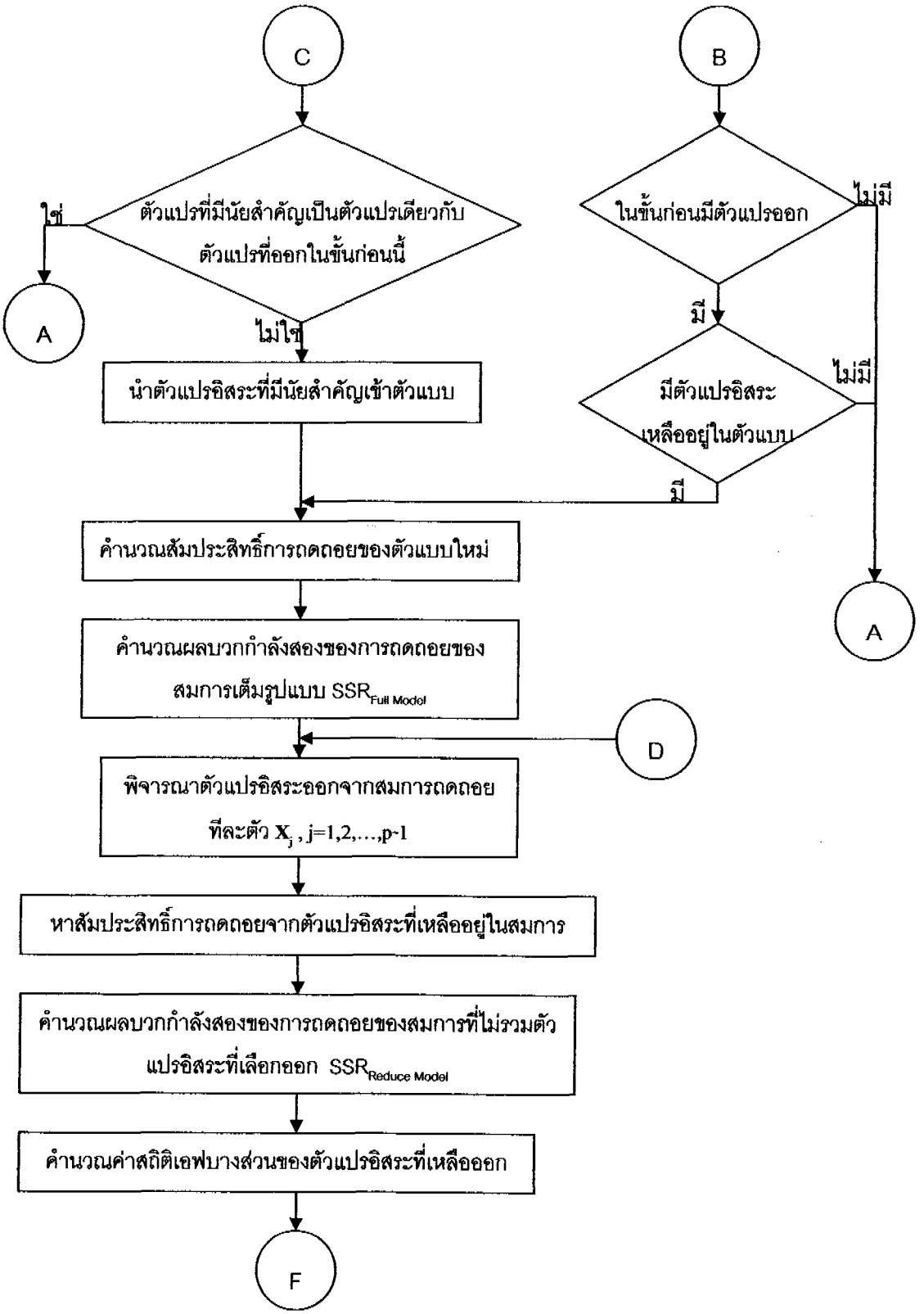


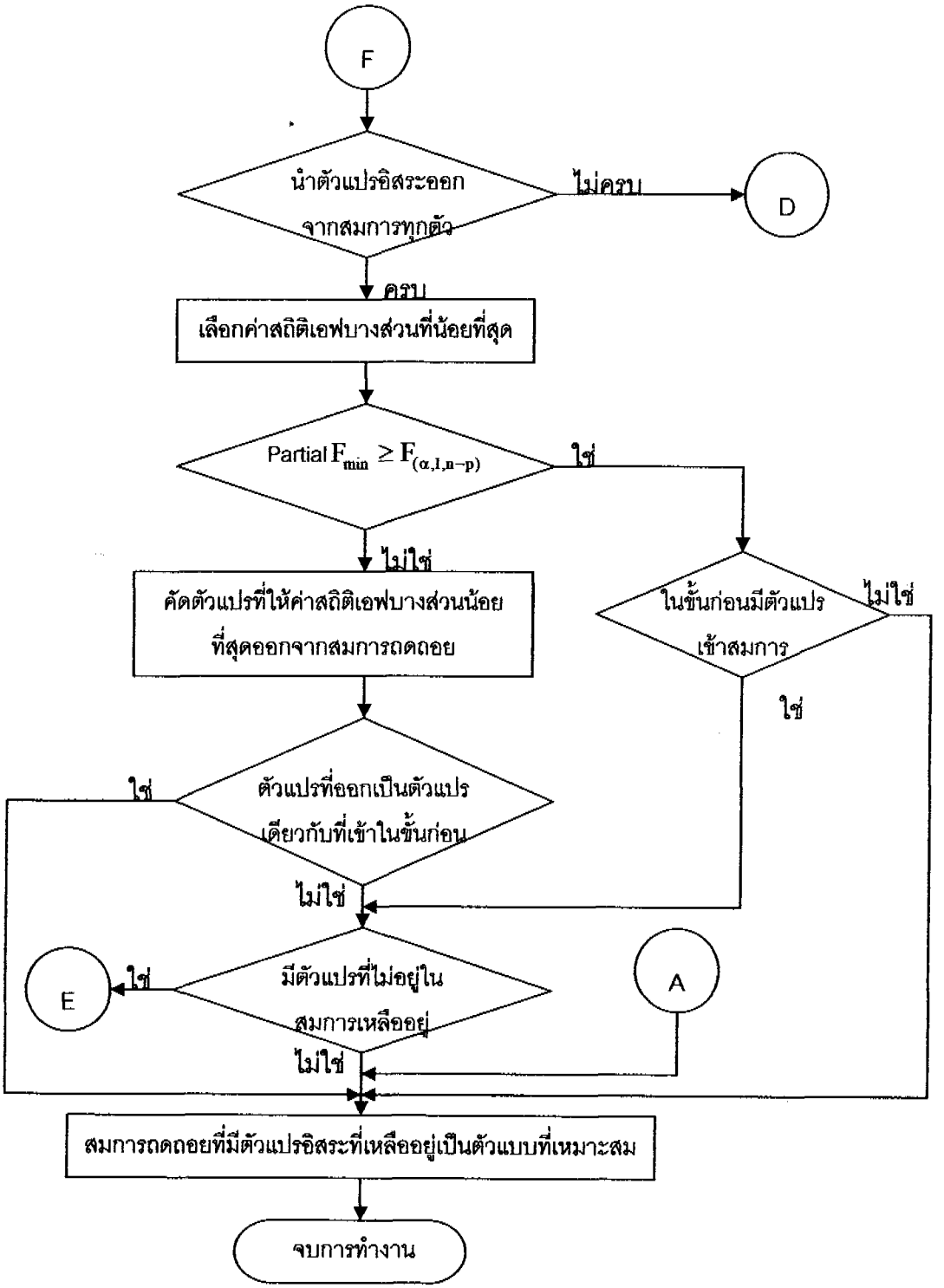

A
คำนวณ $D_{n0} = \sum_{i=1}^{n}(u_i - \hat{u}_{-i})^2$, i=1,2,...n
เมื่อ $\hat{u}_{-i}$ เป็นค่าพยากรณ์ของตัวแบบใหม่ จาก เมตริกซ์ตัวแปรอิสระ X ตัดแถวที่ i
คำนวณ $D_n(h) = \sum_{i=1}^{n}(u_i - \hat{u}_{-i,-h})^2$ เมื่อ i=1,2,...n , h = 1,2,...,p-1
เมื่อ $\hat{u}_{-i,-h}$ เป็นค่าพยากรณ์ของตัวแบบใหม่ที่ตัดข้อมูลชุดที่ i และ ตัวแปรอิสระตัวที่ h
คำนวณ $\Delta_n(h) = D_n(h) - D_{n0}$, h = 1,2,...,p-1
$C_n^{(R)}$ = ค่ามากที่สุดของ {ค่าต่ำสุดของ $\Delta_n(h), h = 1,2,...,p-1,10\}/(1+\sqrt{\lfloor 0.0001n \rfloor})$
คำนวณค่า $G_n^{(CV)}$ สำหรับแต่ละตัวแบบ
เลือกตัวแบบที่มีค่า $G_n^{(CV)}$ ต่ำสุด
สิ้นสุดการทำงาน

แผนภาพที่ 3.6

แสดงขั้นตอนการคัดเลือกตัวแบบวิธีการถดถอยแบบขั้นบันได




C
B
ตัวแปรที่มีนัยสำคัญเป็นตัวแปรเดียวกับ
ตัวแปรที่ออกในขั้นก่อนนี้
ใช่
ไม่ใช่
A
ในขั้นก่อนมีตัวแปรออก
ไม่มี
มี
มีตัวแปรอิสระ
เหลืออยู่ในตัวแบบ
ไม่มี
มี
A
นำตัวแปรอิสระที่มีนัยสำคัญเข้าตัวแบบ
คำนวณสัมประสิทธิ์การถดถอยของตัวแบบใหม่
คำนวณผลบวกกำลังสองของการถดถอยของ
สมการเต็มรูปแบบ $SSR_{Full\ Model}$
D
พิจารณาตัวแปรอิสระออกจากสมการถดถอย
ทีละตัว X_j , j=1,2,...,p-1
หาสัมประสิทธิ์การถดถอยจากตัวแปรอิสระที่เหลืออยู่ในสมการ
คำนวณผลบวกกำลังสองของการถดถอยของสมการที่ไม่รวมตัว
แปรอิสระที่เลือกออก $SSR_{Reduce\ Model}$
คำนวณค่าสถิติเอฟบางส่วนของตัวแปรอิสระที่เหลือออก
F


F
นำตัวแปรอิสระออก
จากสมการทุกตัว
ไม่ครบ
D
ครบ
เลือกค่าสถิติเอฟบางส่วนที่น้อยที่สุด
Partial $F_{min} \geq F_{(\alpha,1,n-p)}$
ใช่
ไม่ใช่
คัดตัวแปรที่ให้ค่าสถิติเอฟบางส่วนน้อย
ที่สุดออกจากสมการถดถอย
ในขั้นก่อนมีตัวแปร
เข้าสมการ
ไม่ใช่
ใช่
ใช่
ตัวแปรที่ออกเป็นตัวแปร
เดียวกับที่เข้าในขั้นก่อน
ไม่ใช่
E
ใช่
มีตัวแปรที่ไม่อยู่ใน
สมการเหลืออยู่
A
ไม่ใช่
สมการถดถอยที่มีตัวแปรอิสระที่เหลืออยู่เป็นตัวแบบที่เหมาะสม
จบการทำงาน