สถิติทดสอบเพื่อคัดเลือกตัวแบบการถดถอยเชิงเส้นพหุคูณ

Test Statistics for Selecting Multiple Linear Regression Models

พลากร สีน้อย* และ จิราวัลย์ จิตรถเวช
คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์
Palakorn Seenoi* and Jirawan Jitthavech
School of Applied Statistics, National Institute of Development Administration

บทคัดย่อ

งานวิจัยนี้ได้เสนอสถิติทดสอบเพื่อการคัดเลือกตัวแบบการถดถอยเชิงเส้นพหุคูณ โดยสร้างสถิติทดสอบจากเกณฑ์ซีพีและเกณฑ์เอไอซี สถิติทดสอบที่สร้างขึ้นนี้เรียกว่าสถิติทดสอบซีพี และสถิติทดสอบเอไอซี ตามลำดับ การคัดเลือกตัวแบบใช้วิธีการกำจัดตัวแปรแบบถอยหลัง โดยใช้เกณฑ์ซีพี เกณฑ์เอไอซี สถิติทดสอบซีพี สถิติทดสอบเอไอซี และสถิติทดสอบเอฟบางส่วน การศึกษาได้ใช้วิธีการเปรียบเทียบเกณฑ์และสถิติทดสอบทั้ง 5 โดยวิธีการจำลองข้อมูล และพิจารณาร้อยละของจำนวนครั้งที่คัดเลือกตัวแบบได้ถูกต้อง พบว่าเมื่อขนาดตัวอย่างเท่ากับ 10 15 และ 20 สถิติทดสอบเอฟบางส่วน คัดเลือกตัวแบบได้ถูกต้องมากกว่าเกณฑ์และสถิติทดสอบอื่น เมื่อขนาดตัวอย่างเท่ากับ 25 และ 30 สถิติทดสอบซีพี สถิติทดสอบเอไอซี และสถิติทดสอบเอฟบางส่วนคัดเลือกตัวแบบได้ถูกต้องใกล้เคียงกันและคัดเลือกตัวแบบได้ถูกต้องมากกว่าเกณฑ์อื่น และเมื่อขนาดตัวอย่างเท่ากับ 50 100 และ 150 สถิติทดสอบซีพี และสถิติทดสอบเอไอซี คัดเลือกตัวแบบได้ถูกต้องใกล้เคียงกันและคัดเลือกตัวแบบได้ถูกต้องมากกว่าเกณฑ์และสถิติทดสอบอื่น และสถิติทดสอบซีพี ได้ตัวแบบ Over-fit น้อยกว่าสถิติทดสอบเอไอซี

คำสำคัญ : การคัดเลือกตัวแบบ การกำจัดตัวแปรแบบถอยหลัง เกณฑ์ซีพี เกณฑ์เอไอซี สถิติทดสอบเอฟบางส่วน

Abstract

This research has proposed test statistics for model selection in multiple linear regressions. The test statistics created by C_P and AIC criteria are called C_P-test and AIC-test statistics respectively. Model selection is conducted by backward elimination method using C_P, AIC criteria and C_P-test, AIC-test and Partial F-test statistics. Verification of making correct model selection under the five criteria based on the percent correctness is achieved through simulation technique. The results indicate that the Partial F-test statistic gives greater correct model selection than other criteria and test statistics when sample sizes equal to 10, 15 and 20. In sample sizes equal to 25 and 30, three of the C_P-test, the AIC-test and the Partial F-test statistics give similar correct percentage which is greater than other criteria. For sample sizes equal to 50, 100 and 150, the C_P-test and the AIC-test statistics give similar correct percentage which is greater than other criteria and test statistics, and the C_P-test is less likely to yield an over-fit model than the AIC-test.

Keywords : Model Selection, Backward Elimination, C_P, AIC, Partial F-test

Corresponding author. Email: seenoi_p@hotmail.com

บทนำ

การวิเคราะห์การถดถอยเชิงเส้นพหุคูณ (Multiple Linear Regression Analysis) เป็นการหารูปแบบความสัมพันธ์ระหว่างตัวแปรตาม (Dependent Variable) กับตัวแปรอิสระ (Independent Variables) เพื่อนำไปพยากรณ์ตัวแปรตาม เมื่อจำนวนตัวแปรอิสระมีตั้งแต่สองตัวขึ้นไป ตัวแบบที่เหมาะสมในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณควรมีจำนวนตัวแปรอิสระน้อยที่สุด แต่ทำให้ได้สมการถดถอยที่สามารถพยากรณ์ได้ถูกต้องและแม่นยำที่สุด เพราะหากมีจำนวนตัวแปรอิสระมากเกินไป ความคลาดเคลื่อนของค่าพยากรณ์มีค่าสูงขึ้น นอกจากนี้ตัวแปรอิสระบางตัวอาจมีความสัมพันธ์กันเอง เกิดปัญหาพหุสัมพันธ์ (Multicollinearity) ส่งผลให้ความแปรปรวนของค่าประมาณพารามิเตอร์ของตัวแบบมีค่าสูง ทำให้การทดสอบพารามิเตอร์ในตัวแบบไม่มีนัยสำคัญ ดังนั้นการคัดเลือกตัวแปรอิสระหรือการคัดเลือกตัวแบบที่เหมาะสม ในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณจึงมีความสำคัญเป็นอย่างยิ่ง

เกณฑ์การคัดเลือกตัวแบบที่นิยมใช้เป็นอันดับแรก และพบในหนังสือการถดถอยในทุกวันนี้ คือ สัมประสิทธิ์ตัวกำหนดปรับปรุง (Adjusted Coefficient of Determination) (McQuarrie & Tsai, 1998) หรือ R^2_{adj} นักวิจัยเข้าใจว่าตัวแบบใดให้ R^2_{adj} ค่าสูงที่สุด แสดงว่าตัวแบบนั้นเหมาะสมที่สุด แต่ความเป็นจริงตัวแบบที่ให้ค่า R^2_{adj} สูงที่สุด จะทำให้ตัวแบบที่ได้รับการคัดเลือก มีจำนวนตัวแปรอิสระในตัวแบบมากเกินไป (Over-fit) (Sheather, 2009)

ด้วยเหตุนี้จึงมีนักสถิติหลายท่านได้สร้างเกณฑ์การคัดเลือกตัวแบบ เช่น มอลโลว์ (Mallows, 1973) ได้เสนอเกณฑ์ซีพี (C_P) แต่เมื่อตัวอย่างมีขนาดเล็กจะทำให้ตัวแบบที่ได้รับการคัดเลือกมีจำนวนตัวแปรอิสระในตัวแบบมากเกินไป (Over-fit) ด้วยเหตุนี้ ราว และ วู (Rao & Wu, 2005 อ้างถึงใน รจนา คุณเจริญจิต, 2549) ได้เสนอเกณฑ์การคัดเลือกตัวแบบขึ้นมาใหม่ โดยลดโอกาสที่จะทำให้ตัวแบบที่ได้รับการคัดเลือกมีจำนวนตัวแปรอิสระมากกว่าหรือน้อยกว่าความเป็นจริง จึงทำให้เกณฑ์ที่เสนอขึ้นใหม่นี้คัดเลือกตัวแบบได้ถูกต้องมากขึ้นเมื่อตัวอย่างขนาดเล็ก

อาไคเคะ (Akaike, 1973) ได้เสนอเกณฑ์ข้อสนเทศของอาไคเคะ (Akaike's Information Criterion: AIC) หรือเกณฑ์เอไอซี โดยตัวแบบที่ให้ค่า AIC น้อยที่สุด แสดงว่าตัวแบบนั้นเหมาะสมที่สุด ซึ่งเกณฑ์นี้ใช้ได้ดีเมื่อตัวอย่างมีขนาดใหญ่ แต่เมื่อตัวอย่างมีขนาดเล็กเมื่อเทียบกับจำนวนพารามิเตอร์ที่ต้องการประมาณ จะทำให้ตัวแบบที่ได้รับการคัดเลือกมีจำนวนตัวแปรอิสระในตัวแบบมากเกินไป (Over-fit) ด้วยเหตุนี้ เฮอร์วิช และไซ (Hurvich & Tsai, 1989) ได้เสนอเกณฑ์เอไอซีซี (AIC_C) เพื่อลดการเกิดปัญหาดังกล่าว แต่เมื่อขนาดตัวอย่างเพิ่มขึ้นเกณฑ์เอไอซีซี จะทำให้ตัวแบบที่ได้รับการคัดเลือก มีจำนวนตัวแปรอิสระในตัวแบบมากเกินไป (Over-fit) ดังนั้น แม็คควาร์รี่ และคณะ (McQuarrie *et al*., 1997) ได้เสนอเกณฑ์เอไอซียู (AICu) ซึ่งเมื่อตัวอย่างมีขนาดใหญ่จะคัดเลือกตัวแบบได้ดีกว่าเกณฑ์เอไอซีซี

จากเกณฑ์การคัดเลือกตัวแบบที่นักสถิติได้เสนอในข้างต้น จะทำการคัดเลือกตัวแบบที่เหมาะสมโดยพิจารณาจากตัวแบบที่ให้ค่าของเกณฑ์นั้นๆ ต่ำที่สุด เช่น หากพิจารณาจากเกณฑ์ซีพี ก็จะคำนวณค่า C_P ทุกตัวแบบที่เป็นไปได้ ตัวแบบใดให้ค่า C_P ต่ำสุด ถือว่าตัวแบบนั้นเหมาะสมที่สุด แต่เนื่องจากต้องพิจารณาทุกตัวแบบที่เป็นไปได้ เมื่อตัวแปรอิสระมีจำนวนมาก การใช้เกณฑ์ที่แตกต่างกันในการพิจารณาอาจได้ตัวแบบที่แตกต่างกันได้ นอกจากนี้ผู้วิจัยจะต้องใช้วิจารณญาณในการพิจารณาว่าควรเลือกตัวแบบใด เพราะตัวแบบที่พิจารณาอาจจะให้ค่าของเกณฑ์ในการคัดเลือกตัวแบบใกล้เคียงกัน แต่ค่าของเกณฑ์ที่ลดลงอาจจะไม่มีนัยสำคัญ จึงควรหยุดกระบวนการในการคัดเลือก ผู้วิจัยจึงเสนอวิธีการสร้างสถิติทดสอบสำหรับการคัดเลือกตัวแบบในการวิเคราะห์การถดถอย โดยพิจารณาจากการแจกแจงของเกณฑ์การคัดเลือกตัวแบบ

ในการวิจัยครั้งนี้ผู้วิจัยได้เสนอสถิติทดสอบ สำหรับการคัดเลือกตัวแบบในการวิเคราะห์การถดถอย โดยได้เสนอสถิติทดสอบสำหรับการคัดเลือกตัวแบบ คือ สถิติทดสอบซีพี (C_P-test) และสถิติทดสอบเอไอซี (AIC-test) และได้ทำการเปรียบเทียบการทดสอบกับสถิติทดสอบเอฟบางส่วน (Partial F-test) และเปรียบเทียบกับเกณฑ์การคัดเลือกตัวแบบที่ใช้เกณฑ์ซีพี (C_P) และเกณฑ์เอไอซี (AIC) คัดเลือกตัวแบบด้วยเกณฑ์ต่ำสุด

ทฤษฎีที่ใช้

ตัวแบบการถดถอยที่เกี่ยวข้องกับงานวิจัยนี้ มีดังนี้

ตัวแบบการถดถอยเชิงเส้นพหุคูณที่แท้จริง (True Multiple Linear Regression Model)

$$\boldsymbol{y} = X^*\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^*$$

เมื่อ $\boldsymbol{y} = (y_1, ..., y_n)'$ เป็นเวกเตอร์สุ่มของค่าสังเกตของตัวแปรตาม n เป็นขนาดตัวอย่าง $\boldsymbol{y} = (y_1, ..., y_n)'$ เป็นเมตริกซ์ของตัวแปรอิสระ ที่มีขนาด $n \times p^*$ k^* เป็นจำนวนตัวแปรอิสระ

$p^* = k^* + 1$ เป็นจำนวนพารามิเตอร์ที่ไม่ทราบค่า $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*, ..., \beta_{k^*}^*)'$ เป็นเวกเตอร์สัมประสิทธิ์การถดถอยที่ไม่ทราบค่า $\boldsymbol{\varepsilon}^* = (\varepsilon_1^*, ..., \varepsilon_n^*)'$ เป็นเวกเตอร์ของความคลาดเคลื่อนสุ่ม ที่มีการแจกแจง $N_n(\mathbf{0}, \sigma^{*^2} I_n)$

ตัวแบบการถดถอยเชิงเส้นพหุคูณที่เหมาะสม (Fitted or Candidate Multiple Linear Regression Model)

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

เมื่อ $X = (\mathbf{1}, \boldsymbol{x}_1, ..., \boldsymbol{x}_k)$ เป็นเมตริกซ์ของตัวแปรอิสระที่มีขนาด $n \times p$ k เป็นจำนวนตัวแปรอิสระ $p = k+1$ เป็นจำนวนพารามิเตอร์ที่ไม่ทราบค่า $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)'$ เป็นเวกเตอร์สัมประสิทธิ์การถดถอยที่ไม่ทราบค่า $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)'$ เป็นเวกเตอร์ของความคลาดเคลื่อนสุ่ม ที่มีการแจกแจง $N_n(\mathbf{0}, \sigma^{*^2} I_n)$

ตัวแบบการถดถอยเชิงเส้นพหุคูณเต็มรูป (Full Multiple Linear Regression Model)

$$\boldsymbol{y} = X_{full}\boldsymbol{\beta}_{full} + \boldsymbol{\varepsilon}$$

เมื่อ $X_{full} = (\mathbf{1}, \boldsymbol{x}_1, ..., \boldsymbol{x}_{k_{full}})$ เป็นเมตริกซ์ของตัวแปรอิสระที่มีขนาด $n \times p_{full}$ k_{full} เป็นจำนวนตัวแปรอิสระทั้งหมด $p_{full} = k_{full} + 1$ เป็นจำนวนพารามิเตอร์ที่ไม่ทราบค่า $\boldsymbol{\beta}_{full} = (\beta_0, \beta_1, ..., \beta_{k_{full}})'$ เป็นเวกเตอร์สัมประสิทธิ์การถดถอยที่ไม่ทราบค่า

เนื่องจากตัวแบบที่ศึกษาเป็นตัวแบบการถดถอยเชิงเส้นพหุคูณ ที่ค่าความคลาดเคลื่อนมีการแจกแจงปกติ ทำให้การประมาณค่าพารามิเตอร์ด้วยวิธีกำลังสองน้อยที่สุด (Ordinary Least Square Method) ได้ตัวประมาณค่า คือ

$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}$ เป็นตัวประมาณไม่เอนเอียงของ $\boldsymbol{\beta}$

$MSE = \dfrac{SSE}{n-p}$ เป็นตัวประมาณไม่เอนเอียงของ σ^2

เมื่อ $SSE = (\boldsymbol{y} - X\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})$ เป็นผลบวกกำลังสองของค่าคลาดเคลื่อน

เกณฑ์การคัดเลือกตัวแบบการถดถอยเชิงเส้นพหุคูณที่ใช้ในการศึกษาครั้งนี้ มีดังนี้

1. เกณฑ์ซีพี (C_p)

มอลโลว์ (Mallows, 1973) ได้เสนอเกณฑ์การคัดเลือกตัวแบบที่เกี่ยวข้องกับค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error) ของค่าประมาณ $\hat{\boldsymbol{y}}$ โดยนิยามให้

$$\Gamma_p = \frac{1}{\sigma^2} E\left[\left(\hat{\boldsymbol{y}} - E(\boldsymbol{y})\right)' \left(\hat{\boldsymbol{y}} - E(\boldsymbol{y})\right) \right]$$

เป็นค่าคาดหมายของค่าคลาดเคลื่อนกำลังสองของค่าพยากรณ์กับค่าคาดหมายของ $\boldsymbol{y}$ หารด้วย σ^2 จึงได้

$$\begin{aligned}
\Gamma_p &= \frac{1}{\sigma^2}\left\{ E\left[\left(\hat{\boldsymbol{y}} - E(\hat{\boldsymbol{y}})\right)' \left(\hat{\boldsymbol{y}} - E(\hat{\boldsymbol{y}})\right) \right] + \right. \\
&\quad \left. \left[\left(E(\boldsymbol{y}) - E(\hat{\boldsymbol{y}})\right)' \left(E(\boldsymbol{y}) - E(\hat{\boldsymbol{y}})\right) \right] \right\} \\
&= \frac{1}{\sigma^2}\left\{ p\sigma^2 + \left[\left(E(\boldsymbol{y})\right)' \left(I - H\right) \left(E(\boldsymbol{y})\right) \right] \right\} ; \\
&\quad H = X(X'X)^{-1}X' \\
&= \frac{1}{\sigma^2}\left\{ p\sigma^2 - tr\left[(I-H)\sigma^2 \right] + E\left[\boldsymbol{y}'(I-H)\,\boldsymbol{y} \right] \right\} \\
&= \frac{1}{\sigma^2}\left\{ p\sigma^2 + \left[E(SSE) - (n-p)\sigma^2 \right] \right\} \\
&= \frac{E(SSE)}{\sigma^2} - (n-2p) \qquad \text{..... (1)}
\end{aligned}$$

เมื่อ $E(\boldsymbol{y})$ เป็นค่าคาดหมายของตัวแปรตามจากตัวแบบการถดถอยที่แท้จริง

$E(\hat{\boldsymbol{y}})$ เป็นค่าคาดหมายของค่าพยากรณ์จากตัวแบบที่มีพารามิเตอร์ p ตัว

$E\left[\left(\hat{\boldsymbol{y}} - E(\hat{\boldsymbol{y}})\right)' \left(\hat{\boldsymbol{y}} - E(\hat{\boldsymbol{y}})\right) \right]$ เป็นความแปรปรวนของ $\hat{\boldsymbol{y}}$

$\left[\left(E(\boldsymbol{y}) - E(\hat{\boldsymbol{y}})\right)' \left(E(\boldsymbol{y}) - E(\hat{\boldsymbol{y}})\right) \right]$ เป็นความเอนเอียงกำลังสอง (Square Bias)

ให้ C_p เป็นตัวประมาณของ Γ_p ดังนั้น

$$C_p = \frac{SSE}{MSE_{full}} - n + 2p \qquad \text{..... (2)}$$

ตัวแบบการถดถอยที่เหมาะสมเป็นตัวแบบที่มีค่า C_p ต่ำ และมีค่าใกล้กับ p ถ้า C_p มีค่าเท่ากับ p ตัวประมาณ C_p จะไม่มีความเอนเอียง ถ้า C_p มีค่ามากกว่า p ตัวประมาณ C_p จะมีความเอนเอียงมาก แต่ถ้า C_p มีค่าน้อยว่า p ตัวประมาณ C_p จะมีความเอนเอียงน้อย (ทรงศิริ, 2548)

2. เกณฑ์ข้อสนเทศอาไคเคะ (AIC)

อาไคเคะ (Akaike, 1973) ได้เสนอเกณฑ์การคัดเลือกตัวแบบคือเกณฑ์เอไอซี ซึ่งสร้างจากการประมาณความแปรปรวนของข้อสนเทศคูลส์แบล็ค-ไลท์เบอร์ (Kullback Leibler Information) ระหว่างตัวแบบที่แท้จริง (True Model) กับตัวแบบที่เหมาะสมที่มีคุณสมบัติไม่เอนเอียง เมื่อขนาดตัวอย่างใหญ่ การคัดเลือกตัวแบบโดยใช้เกณฑ์เอไอซี เลือกตัวแบบที่ให้ค่าเอไอซีต่ำสุดเป็นตัวแบบที่ดีที่สุด และเกณฑ์เอไอซี คัดเลือกตัวแบบได้ดีเมื่อ

ตัวอย่างมีขนาดใหญ่ เนื่องจากเกิดความผิดพลาดในการคัดเลือกตัวแบบสูงเมื่อตัวอย่างมีขนาดเล็ก (McQuarrie *et al*., 1997) เกณฑ์เอไอซี นิยามโดย

$$\begin{aligned} AIC_p &= -2\left[\log\left(L(\hat{\beta}_0,\hat{\beta}_1,...,\hat{\beta}_k,\hat{\sigma}^2 \mid y)\right)+(p+1)\right] \\ &= -2\left[-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log\left(\frac{SSE}{n}\right)-\frac{n}{2}+(p+1)\right] \\ &= n\log(2\pi)+n\log\left(\frac{SSE}{n}\right)+n+2(p+1) \\ &= n\log\left(2\pi\frac{SSE}{n}\right)+n+2(p+1) \\ &= n\log(2\pi\hat{\sigma}^2)+n+2(p+1) \end{aligned} \quad \text{..... (3)}$$

เมื่อ $\hat{\sigma}^2 = \frac{1}{n}\boldsymbol{y}'(I_n - P_X)\boldsymbol{y}$ เป็นตัวประมาณด้วยวิธีภาวะน่าจะเป็นสูงสุดของ σ^2

$P_X = X(X'X)^{-1}X'$ เป็นเมตริกซ์ภาพฉาย

3. สถิติทดสอบซีพี (C_P-test)

พิจารณาเกณฑ์การคัดเลือกตัวแบบ C_P จากสมการ (2)

ให้ $\Upsilon = C_p + n - 2p$ จะได้ $\Upsilon = \frac{SSE}{MSE_{full}}$ หรือ $\Upsilon = \frac{SSE}{E(MSE)}$ มีการแจกแจงไคกำลังสอง ที่มีองศาความเสรี $n-p$ ดังนั้น $E(\Upsilon) = n-p$ และ $Var(\Upsilon) = 2(n-p)$

เมื่อขนาดตัวอย่างใหญ่พอ $\frac{\Upsilon - E(\Upsilon)}{\sqrt{Var(\Upsilon)}}$ ประมาณด้วยการแจกแจงปกติมาตรฐานเชิงกำกับ (Asymptotic Standard Normal Distribution) (Canal, 2005)

สร้างสถิติทดสอบสมมติฐาน $H_0 : \Gamma_p = \Gamma_{p_{full}}$ ได้ดังนี้

$$\begin{aligned} T_{C_p} &= \frac{(C_p + n - 2p) - (n - p_{full})}{\sqrt{2(n-p_{full})}} \\ &= \frac{C_p - 2p + p_{full}}{\sqrt{2(n-p_{full})}} \rightarrow N(0,1) \end{aligned} \quad \text{..... (4)}$$

เมื่อ $\Gamma_{p_{full}}$ เป็นค่าคาดหมายของค่าคลาดเคลื่อนกำลังสองของค่าพยากรณ์กับค่าคาดหมายของ $\boldsymbol{y}$ หารด้วย σ^2 จากตัวแบบการถดถอยเต็มรูป

จะปฏิเสธ H_0 เมื่อ $\left|T_{C_p}\right|$ มีค่ามากกว่า $z_{\frac{\alpha}{2}}$ ซึ่งหมายถึงไม่สามารถนำตัวแปรอิสระออกจากตัวแบบได้

4. สถิติทดสอบเอไอซี (AIC-test)

ยานากิฮาร่า และ โอโมโต๊ะ (Yanagihara & Ohmoto, 2005) ได้พิจารณาการแจกแจงเชิงกำกับ (Asymptotic Distribution) ของเกณฑ์เอไอซี และเสนอคุณลักษณะเฉพาะ (Characteristics) ของเกณฑ์เอไอซี ในตัวแบบการถดถอยเชิงเส้น เช่น ความเบ้ (Skewness) ความโด่ง (Kurtosis) ซึ่งจะบอกถึงความแม่นยำของการประมาณด้วยการแจกแจงปกติ เมื่อกำหนดให้จากสมการ (3)

ค่าคาดหมายของ AIC_p เท่ากับ

$$E(AIC_p) = n(\log(2\pi\sigma^{*2}\delta^2)+1) + 2(p+1) - p\delta^{-2} - \frac{1}{2}\tau^2\delta^{-4}$$

และความแปรปรวนของ AIC_p เท่ากับ

$$\psi_p^2 = n\tau^2\delta^{-4} + h^2\delta^{-8}$$

เมื่อ $\delta^2 = 1+\theta^2$

$$\theta^2 = \frac{1}{\sigma^{*2}n}\boldsymbol{\beta}^{*\prime}X^{*\prime}(I_n - P_X)X^*\boldsymbol{\beta}^*$$

$$\tau^2 = 2(1+2\theta^2)$$

$$h^2 = 2(p+1) + 8(p+1)\theta^2 + 2(3p+8)\theta^4$$

ยานากิฮาร่า และ โอโมโต๊ะ พบว่าเอไอซี มีความเบ้ซ้าย แต่ค่ามาตรฐานของเอไอซี ไม่มีความเบ้ และมีการแจกแจงปกติมาตรฐานเชิงกำกับ (Asymptotic Standard Normal Distribution) (Yanagihara & Ohmoto, 2005) นั่นคือ

$$\frac{AIC_p - E(AIC_p)}{\sqrt{\psi_p^2}} \rightarrow N(0,1)$$

สร้างสถิติทดสอบสมมติฐาน $H_0 : E(AIC_p) = E(AIC_{full})$ ได้ดังนี้

$$T_{AIC} = \frac{AIC_p - E(AIC_{full})}{\sqrt{\psi_{full}^2}}$$

เมื่อประมาณ $E(AIC_{full})$ ด้วย $\widehat{E(AIC_{full})}$ และประมาณ ψ_{full}^2 ด้วย $\hat{\psi}_{full}^2$ จึงได้ค่าประมาณของสถิติ T_{AIC} คือ

$$t_{AIC} = \frac{AIC_p - \widehat{E(AIC_{full})}}{\sqrt{\hat{\psi}_{full}^2}} \rightarrow N(0,1) \quad \text{..... (5)}$$

เมื่อ $E(AIC_{full})$ เป็นค่าคาดหมายของข้อสนเทศอาไคเคะ จากตัวแบบการถดถอยเต็มรูป

$\widehat{E(AIC_{full})}$ เป็นค่าประมาณค่าคาดหมายของข้อสนเทศอาไคเคะ จากตัวแบบการถดถอยเต็มรูป

ψ^2_{full} เป็นความแปรปรวนของข้อสนเทศอาไคเคะจากตัวแบบการถดถอยเต็มรูป

และ $\hat{\psi}^2_{full}$ เป็นค่าประมาณความแปรปรวนของข้อสนเทศอาไคเคะ จากตัวแบบการถดถอยเต็มรูป

จะปฏิเสธ H_o เมื่อ $|t_{AIC}|$ มีค่ามากกว่า $z_{\frac{\alpha}{2}}$ ซึ่งหมายถึงไม่สามารถนำตัวแปรอิสระออกจากตัวแบบได้

5. สถิติทดสอบเอฟบางส่วน (Partial F-test)

การคัดเลือกตัวแบบอาจใช้สถิติทดสอบเอฟบางส่วน เป็นการทดสอบสัมประสิทธิ์การถดถอยของตัวแปรอิสระที่เราสนใจว่ามีส่วนในการอธิบายความผันแปรของตัวแปรตามหรือไม่ อาจทดสอบทีละตัวหรือทดสอบหลายตัวพร้อมกันก็ได้

สมมติให้ตัวแบบเต็มรูป (Full Model) คือ $\boldsymbol{y} = X_{full}\boldsymbol{\beta}_{full} + \boldsymbol{\varepsilon} = X_r\boldsymbol{\beta}_r + X_q\boldsymbol{\beta}_q + \boldsymbol{\varepsilon}$ โดยที่ $q + r = p_{full}$ เมื่อ $X_{full} = (X_q, X_r)$ และ $\boldsymbol{\beta}_{full} = (\boldsymbol{\beta}_q, \boldsymbol{\beta}_r)'$

หากต้องการทดสอบ $H_0 : \boldsymbol{\beta}_q = 0$ โดยที่ $\boldsymbol{\beta}_q = (\beta_1, \beta_2, ..., \beta_q)'$ เมื่อ $q < p_{full}$

จะได้ตัวแบบลดรูป (Reduced Model) คือ $\boldsymbol{y} = X_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}$ เมื่อ $X_r = (\mathbf{1}, \boldsymbol{x}_{q+1}, ..., \boldsymbol{x}_{q+r-1})$ และ $\boldsymbol{\beta}_r = (\beta_0, \beta_{q+1}, ..., \beta_{q+r-1})'$

จากนั้นจะใช้ผลบวกค่าคลาดเคลื่อนกำลังสองของตัวแบบลดรูปและตัวแบบเต็มรูป ในการคำนวณสถิติทดสอบจะได้สถิติทดสอบเอฟบางส่วน คือ

$$F = \frac{(\hat{\boldsymbol{\beta}}'_r X'_r \boldsymbol{y} - \hat{\boldsymbol{\beta}}'_{q+r} X'_{q+r} \boldsymbol{y}) / q}{\hat{\boldsymbol{\beta}}'_{q+r} X'_{q+r} \boldsymbol{y} / (n-q-r)} \quad \text{..... (6)}$$

จะปฏิเสธ H_o เมื่อ $F > F_{\alpha,q,(n-q-r)}$ ซึ่งหมายถึงตัวแปรอิสระ X_q มีส่วนในการอธิบายความผันแปรของตัวแปรตาม

วิธีดำเนินการและขั้นตอนการคัดเลือกตัวแบบ

1. วิธีดำเนินการ

งานวิจัยนี้ได้ศึกษากรณีที่ตัวแปรอิสระมีการแจกแจงเอกรูป และการแจกแจงปกติ ดังนี้

กรณีตัวแปรอิสระมีการแจกแจงเอกรูป

สร้างตัวแปรอิสระให้มีการแจกแจงเอกรูป (Uniform Distribution) ดังนี้ $\boldsymbol{x}_1 \sim U(10,150)$ $\boldsymbol{x}_2 \sim U(0,1000)$ $\boldsymbol{x}_3 \sim U(8000,15000)$ $\boldsymbol{x}_4 \sim U(-200,100)$ $\boldsymbol{x}_5 \sim U(-300,-500)$ $\boldsymbol{x}_6 \sim U(5,15)$ และ $\boldsymbol{x}_7 \sim U(500,2000)$

สร้างค่าความคลาดเคลื่อนในตัวแบบการถดถอยเชิงเส้นพหุคูณที่แท้จริงให้มีการแจกแจงปกติ (Normal Distribution) ที่มีค่าเฉลี่ยเท่ากับ 0 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 50 หรือ $\boldsymbol{\varepsilon}^* \sim N(0,50^2)$ ให้มีขนาดเท่ากับขนาดประชากร โดยกำหนดให้ประชากรมีขนาด 100000

กำหนดค่าพารามิเตอร์ หรือค่าสัมประสิทธิ์การถดถอยในตัวแบบการถดถอยเชิงเส้นพหุคูณที่แท้จริง คือ $\boldsymbol{\beta}^* = (-20, 60, 150, 35, -100)$

กรณีตัวแปรอิสระมีการแจกแจงปกติ

สร้างตัวแปรอิสระให้มีการแจกแจงปกติ (Normal Distribution) ดังนี้ $\boldsymbol{x}_1 \sim N(45,30^2)$ $\boldsymbol{x}_2 \sim N(-60,100^2)$ $\boldsymbol{x}_3 \sim N(500,80^2)$ $\boldsymbol{x}_4 \sim N(1000,150^2)$ $\boldsymbol{x}_5 \sim N(10,5^2)$ $\boldsymbol{x}_6 \sim N(-200,130^2)$ และ $\boldsymbol{x}_7 \sim N(3500,500^2)$

สร้างค่าความคลาดเคลื่อนในตัวแบบการถดถอยเชิงเส้นพหุคูณที่แท้จริงให้มีการแจกแจงปกติ (Normal Distribution) ที่มีค่าเฉลี่ยเท่ากับ 0 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 64 หรือ $\boldsymbol{\varepsilon}^* \sim N(0,64^2)$ ให้มีขนาดเท่ากับขนาดประชากร

กำหนดค่าพารามิเตอร์ หรือค่าสัมประสิทธิ์การถดถอยในตัวแบบการถดถอยเชิงเส้นพหุคูณที่แท้จริง คือ $\boldsymbol{\beta}^* = (8, 20, -15, 70, 40)$

สร้างตัวแปรตาม $\boldsymbol{y}$ จากตัวแบบการถดถอยเชิงเส้นพหุคูณที่แท้จริง $\boldsymbol{y} = X^*\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^*$ เมื่อ $X^* = (\mathbf{1}, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4)$ ให้มีขนาดเท่ากับขนาดประชากร

เมื่อได้ประชากรตามขนาดที่ต้องการแล้ว ให้สุ่มตัวอย่างขนาด 10 15 20 25 30 50 100 และ 150 ในการสุ่มตัวอย่างแต่ละขนาดให้สุ่มซ้ำจำนวน 1000 ครั้ง โดยให้ถือว่าตัวอย่างที่สุ่มได้เป็นตัวอย่างที่สุ่มมาจากตัวแบบการถดถอยเชิงเส้นพหุคูณเต็มรูป จะได้ $\boldsymbol{y} = X_{full}\boldsymbol{\beta}_{full} + \boldsymbol{\varepsilon}$ เมื่อ $X_{full} = (\mathbf{1}, \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_7)$

ทำการคัดเลือกตัวแบบด้วยเกณฑ์ซีพี เกณฑ์เอไอซี สถิติทดสอบซีพี สถิติทดสอบเอไอซี และสถิติทดสอบเอฟบางส่วน โดยกำหนดระดับนัยสำคัญในการทดสอบ 0.05 หากเกณฑ์หรือสถิติทดสอบใด คัดเลือกตัวแบบได้ตัวแบบการถดถอยเชิงเส้นพหุคูณที่แท้จริง หรือได้ตัวแปรอิสระในตัวแบบเฉพาะ $X^* = (\mathbf{1}, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4)$ เท่านั้น แสดงว่าคัดเลือกตัวแบบได้ถูกต้อง หากเกณฑ์หรือสถิติทดสอบใดคัดเลือกตัวแบบแล้วมีจำนวนตัวแปรอิสระมากกว่าจำนวนตัวแปรอิสระในตัวแบบการถดถอยเชิงเส้นพหุคูณที่แท้จริง แสดงว่าตัวแบบที่ได้ Over-fit และหากเกณฑ์หรือสถิติทดสอบใดคัดเลือกตัวแบบแล้วได้จำนวนตัวแปรอิสระน้อยกว่าจำนวนตัวแปรอิสระในตัวแบบการถดถอย

เชิงเส้นพหุคูณที่แท้จริง แสดงว่าตัวแบบที่ได้ Under-fit

ทำการทดลองซ้ำ 1000 ครั้ง ในแต่ละขนาดตัวอย่าง

พิจารณาร้อยละที่แต่ละเกณฑ์และสถิติทดสอบ สามารถคัดเลือกตัวแบบได้ถูกต้อง

2. ขั้นตอนการคัดเลือกตัวแบบ

งานวิจัยนี้ได้ทำการคัดเลือกตัวแบบด้วยการกำจัดตัวแปรแบบถอยหลัง (Backward Elimination) โดยทั่วไปการกำจัดตัวแปรอิสระแบบถอยหลังจะเป็นวิธีการคัดเลือกตัวแบบการถดถอยที่ดีที่สุด (ทรงศิริ แต้สมบัติ, 2548; Montgomery *et al*., 2006) ทำได้โดยกำหนดให้สมการถดถอยประกอบด้วยตัวแปรอิสระทั้งหมดก่อน แล้วคัดเลือกตัวแปรอิสระ ที่ไม่มีส่วนในการอธิบายความผันแปรของตัวแปรตาม ออกจากตัวแบบทีละตัว โดยมีขั้นตอนการคัดเลือกตัวแบบของแต่ละเกณฑ์ ดังนี้

ขั้นตอนการคัดเลือกตัวแบบสำหรับเกณฑ์ซีพี และเกณฑ์เอไอซี

ขั้นที่ 1 คำนวณค่า C_P และ AIC_P จากตัวแบบที่มีตัวแปรอิสระทุกตัว สมมติให้เป็น $C_{P_{full}}$ และ $AIC_{P_{full}}$ ตามลำดับ

ขั้นที่ 2 คำนวณค่า C_P และ AIC_P ของทุกตัวแบบที่เป็นไปได้ เมื่อนำตัวแปรอิสระออกจากตัวแบบ 1 ตัว

ขั้นที่ 3 พิจารณาตัวแบบที่มีค่า C_P และ AIC_P น้อยกว่าตัวแบบอื่นๆ สมมติให้เป็น $C_{P_{\min}}$ และ $AIC_{\min}$ ตามลำดับ

ขั้นที่ 4 เปรียบเทียบค่า C_P และ AIC_P ถ้า $C_{P_{\min}}$ และ $AIC_{\min}$ มีค่าน้อยกว่าหรือเท่ากับ $C_{P_{full}}$ และ AIC_{full} ตามลำดับ แสดงว่าสามารถนำตัวแปรอิสระที่ทำให้ $C_{P_{\min}}$ และ $AIC_{\min}$ ออกจากตัวแบบได้และจะไม่นำตัวแปรอิสระนี้มาพิจารณาอีก แต่ถ้า $C_{P_{\min}}$ และ $AIC_{\min}$ มีค่ามากกว่า $C_{P_{full}}$ และ AIC_{full} ตามลำดับ แสดงว่าไม่สามารถนำตัวแปรอิสระออกจากตัวแบบได้ และสิ้นสุดการคัดเลือกตัวแบบ

ขั้นที่ 5 ให้ทำซ้ำในขั้นที่ 2 และขั้นที่ 3 ส่วนในขั้นที่ 4 ให้เปรียบเทียบค่า $C_{P_{\min}}$ และ $AIC_{\min}$ ในรอบปัจจุบัน กับค่า $C_{P_{\min}}$ และ $AIC_{\min}$ ในรอบก่อนหน้า ถ้า $C_{P_{\min}}$ และ $AIC_{\min}$ ในรอบปัจจุบันมีค่าน้อยกว่าหรือเท่ากับ $C_{P_{\min}}$ และ $AIC_{\min}$ ในรอบก่อนหน้า แสดงว่าสามารถนำตัวแปรอิสระตัวที่ทำให้ $C_{P_{\min}}$ และ $AIC_{\min}$ ออกจากตัวแบบได้ และทำซ้ำไปเรื่อยๆ จนกว่าจะไม่สามารถนำตัวแปรอิสระออกจากตัวแบบได้ และสิ้นสุดการคัดเลือกตัวแบบ

ขั้นตอนการคัดเลือกตัวแบบสำหรับสถิติทดสอบซีพีและสถิติทดสอบเอไอซี

ขั้นที่ 1 คำนวณค่า C_P และ AIC_P ของทุกตัวแบบที่เป็นไปได้ เมื่อนำตัวแปรอิสระออกจากตัวแบบ 1 ตัว

ขั้นที่ 2 พิจารณาตัวแบบที่มีค่า C_P และ AIC_P น้อยกว่าตัวแบบอื่นๆ สมมติให้เป็น $C_{P_{\min}}$ และ $AIC_{\min}$ ตามลำดับ

ขั้นที่ 3 คำนวณสถิติทดสอบ $T_{C_{P_{\min}}}$ และ $T_{AIC_{\min}}$ ถ้า $\left|T_{C_{P_{\min}}}\right|$ และ $\left|t_{AIC_{\min}}\right|$ มีค่าน้อยกว่าหรือเท่ากับ $z_{\frac{\alpha}{2}}$ แสดงว่าสามารถนำตัวแปรอิสระออกจากตัวแบบได้ และจะไม่นำตัวแปรอิสระนี้มาพิจารณาอีก แต่ถ้า $\left|T_{C_{P_{\min}}}\right|$ และ $\left|t_{AIC_{\min}}\right|$ มีค่ามากกว่า $z_{\frac{\alpha}{2}}$ แสดงว่าไม่สามารถนำตัวแปรอิสระออกจากตัวแบบได้ และสิ้นสุดการคัดเลือกตัวแบบ

ขั้นที่ 4 ให้ทำซ้ำในขั้นที่ 1 ขั้นที่ 2 และขั้นที่ 3 จนกระทั่งไม่สามารถนำตัวแปรอิสระออกจากตัวแบบได้ และจบการคัดเลือกตัวแบบ

ขั้นตอนการคัดเลือกตัวแบบสำหรับสถิติทดสอบเอฟบางส่วน

ขั้นที่ 1 ให้ตัวแบบเต็มรูปประกอบด้วยตัวแปรอิสระทั้งหมด

ขั้นที่ 2 คำนวณค่าสถิติเอฟบางส่วนของทุกตัวแบบที่เป็นไปได้ เมื่อนำตัวแปรอิสระออกจากตัวแบบ 1 ตัว

ขั้นที่ 3 พิจารณาตัวแบบที่มีค่าสถิติเอฟบางส่วนน้อยกว่าตัวแบบอื่นๆ สมมติให้เป็น $F_{\min}$

ขั้นที่ 4 ทดสอบนัยสำคัญของ $F_{\min}$ ถ้า $F_{\min}$ มีนัยสำคัญแสดงว่าไม่สามารถนำตัวแปรอิสระออกจากตัวแบบได้ แต่ถ้าพบว่า $F_{\min}$ ไม่มีนัยสำคัญก็จะนำตัวแปรอิสระนี้ออกจากตัวแบบ แล้วคำนวณค่าสถิติเอฟบางส่วนของทุกตัวแบบที่เป็นไปได้ จากตัวแปรอิสระที่เหลือใหม่อีกครั้ง และทำซ้ำขั้นที่ 2 ขั้นที่ 3 และขั้นที่ 4 จนกระทั่งไม่สามารถนำตัวแปรอิสระออกจากตัวแบบได้ และจบการคัดเลือกตัวแบบ

ผลการศึกษา

ผลการคัดเลือกตัวแบบด้วยเกณฑ์ซีพี เกณฑ์เอไอซี สถิติทดสอบซีพี สถิติทดสอบเอไอซี และสถิติทดสอบเอฟบางส่วน เมื่อจำแนกตามการแจกแจงของตัวแปรอิสระ ที่ระดับนัยสำคัญ 0.05 ได้ผลดังตารางที่ 1 และตารางที่ 2

จากตารางที่ 1 พบว่า เมื่อขนาดตัวอย่างเท่ากับ 10 15 และ 20 สถิติทดสอบเอฟบางส่วน คัดเลือกตัวแบบได้ถูกต้องมากกว่าเกณฑ์และสถิติทดสอบอื่น คิดเป็นร้อยละ 94.3 95.4 และ 95.1 ตามลำดับ

เมื่อขนาดตัวอย่างเท่ากับ 25 และ 30 สถิติทดสอบซีพี สถิติทดสอบเอไอซี และสถิติทดสอบเอฟบางส่วน คัดเลือก

ตารางที่ 1 ร้อยละของตัวแบบที่คัดเลือกได้ตามเกณฑ์และสถิติทดสอบ จำแนกตามผลการคัดเลือก และขนาดตัวอย่าง เมื่อตัวแปรอิสระมีการแจกแจงเอกรูป

ขนาดตัวอย่าง	ผลการคัดเลือก	เกณฑ์และสถิติทดสอบ				
		CP	AIC	CP-test	AIC-test	Partial F-test
10	Over-fit	53.4	86.9	49.9	63.4	5.7
	Fit	46.6	13.1	50.1	36.6	94.3
	Under-fit	0.0	0.0	0.0	0.0	0.0
15	Over-fit	45.7	68.6	19.9	25.2	4.6
	Fit	54.3	31.4	80.1	74.8	95.4
	Under-fit	0.0	0.0	0.0	0.0	0.0
20	Over-fit	43.1	60.9	9.9	12.6	4.9
	Fit	56.9	39.1	90.1	87.4	95.1
	Under-fit	0.0	0.0	0.0	0.0	0.0
25	Over-fit	43.5	54.4	3.8	4.6	3.8
	Fit	56.5	45.6	96.2	95.4	96.2
	Under-fit	0.0	0.0	0.0	0.0	0.0
30	Over-fit	44.0	54.6	3.2	3.8	5.5
	Fit	56.0	45.4	96.8	96.2	94.5
	Under-fit	0.0	0.0	0.0	0.0	0.0
50	Over-fit	43.2	49.1	0.3	0.6	5.8
	Fit	56.8	50.9	99.7	99.4	94.2
	Under-fit	0.0	0.0	0.0	0.0	0.0
100	Over-fit	39.8	43.0	0.0	0.0	3.9
	Fit	60.2	57.0	100.0	100.0	96.1
	Under-fit	0.0	0.0	0.0	0.0	0.0
150	Over-fit	39.5	41.3	0.0	0.0	4.9
	Fit	60.5	58.7	100.0	100.0	95.1
	Under-fit	0.0	0.0	0.0	0.0	0.0

หมายเหตุ Over-fit คือร้อยละของตัวแบบที่ตัวแปรอิสระมีมากกว่าความเป็นจริง
Fit คือร้อยละของตัวแบบที่คัดเลือกได้ถูกต้อง
Under-fit คือร้อยละของตัวแบบที่ตัวแปรอิสระมีน้อยกว่าความเป็นจริง

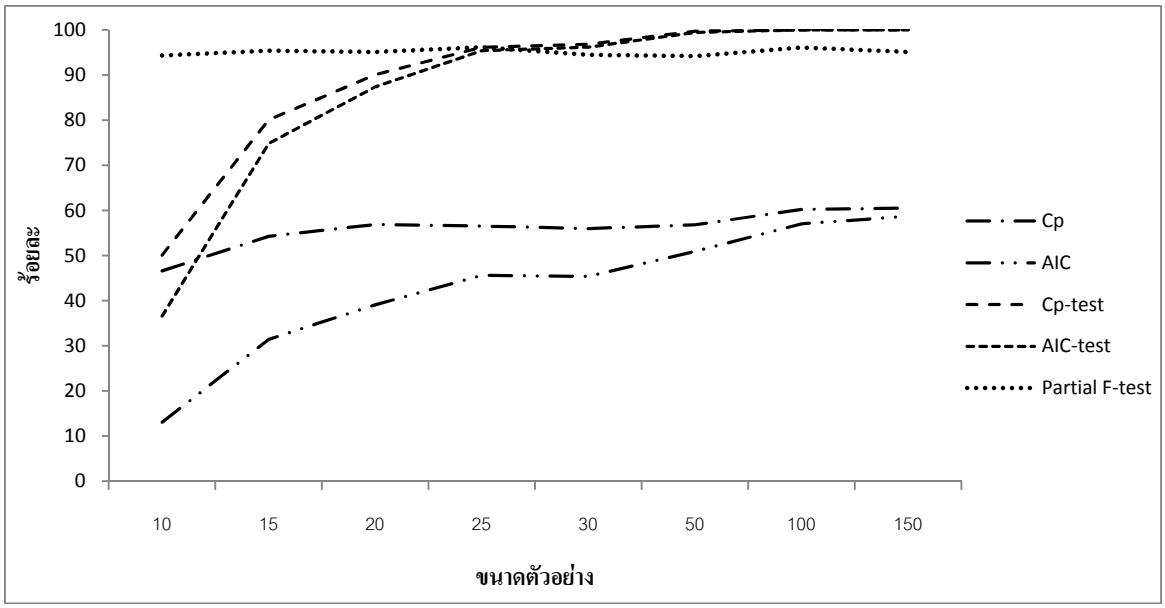
ตัวแบบได้ถูกต้องใกล้เคียงกันและคัดเลือกตัวแบบได้ถูกต้องมากกว่าเกณฑ์อื่น ขนาดตัวอย่างเท่ากับ 25 คัดเลือกตัวแบบได้ถูกต้องร้อยละ 96.2 95.4 และ 96.2 ตามลำดับ ขนาดตัวอย่างเท่ากับ 30 คัดเลือกตัวแบบได้ถูกต้องร้อยละ 96.8 96.2 และ 95.4 ตามลำดับ

เมื่อขนาดตัวอย่างเท่ากับ 50 100 และ 150 สถิติทดสอบซีพีและสถิติทดสอบเอไอซี คัดเลือกตัวแบบได้ถูกต้องใกล้เคียงกันและคัดเลือกตัวแบบได้ถูกต้องมากกว่าเกณฑ์และสถิติทดสอบอื่น ขนาดตัวอย่างเท่ากับ 50 คัดเลือกตัวแบบได้ถูกต้องร้อยละ 99.7 และ 99.4 ตามลำดับ ขนาดตัวอย่างเท่ากับ 100 และ 150 สถิติทดสอบทั้งสองคัดเลือกตัวแบบได้ถูกต้องร้อยละ 100.0

ทุกขนาดตัวอย่าง เกณฑ์เอไอซี มีร้อยละที่ตัวแปรอิสระในตัวแบบมีมากกว่าความเป็นจริง มากกว่าเกณฑ์และสถิติทดสอบอื่น

ทุกขนาดตัวอย่าง ไม่มีเกณฑ์และสถิติทดสอบใด ที่ทำให้ร้อยละของตัวแบบที่ตัวแปรอิสระมีน้อยกว่าความเป็นจริง

กราฟแสดงความสัมพันธ์ระหว่างการคัดเลือกตัวแบบด้วยเกณฑ์และสถิติทดสอบต่างๆ กับขนาดตัวอย่าง เมื่อตัวแปรอิสระมีการแจกแจงเอกรูป ได้แสดงร้อยละของตัวแบบที่คัดเลือกได้ถูกต้อง ในภาพที่ 1



ภาพที่ 1 ร้อยละของตัวแบบที่คัดเลือกได้ถูกต้อง เมื่อตัวแปรอิสระมีการแจกแจงเอกรูป

จากภาพที่ 1 พบว่า เมื่อขนาดตัวอย่างเพิ่มขึ้น ร้อยละของตัวแบบที่คัดเลือกได้ถูกต้อง ด้วยเกณฑ์และสถิติทดสอบต่างๆ จะเพิ่มขึ้น ยกเว้นสถิติทดสอบเอฟบางส่วน เมื่อขนาดตัวอย่างเพิ่มขึ้น ร้อยละของตัวแบบที่คัดเลือกได้ถูกต้องจะคงที่ ซึ่งคัดเลือกตัวแบบถูกต้องประมาณร้อยละ 95 เมื่อขนาดตัวอย่างใหญ่ เกณฑ์ซีพีและเกณฑ์เอไอซี มีแนวโน้มคัดเลือกตัวแบบได้ถูกต้องใกล้เคียงกัน สถิติทดสอบซีพีและสถิติทดสอบเอไอซี คัดเลือกตัวแบบได้ดีเมื่อขนาดตัวอย่างใหญ่พอ เนื่องจากสถิติทดสอบทั้งสองเป็นสถิติทดสอบเชิงกำกับ ซึ่งต่างจากสถิติทดสอบเอฟบางส่วนที่คัดเลือกตัวแบบได้ดีทุกขนาดตัวอย่าง

จากตารางที่ 2 พบว่า เมื่อขนาดตัวอย่างเท่ากับ 10 15 และ 20 สถิติทดสอบเอฟบางส่วน คัดเลือกตัวแบบได้ถูกต้องมากกว่าเกณฑ์และสถิติทดสอบอื่น คิดเป็นร้อยละ 92.5 95.3 และ 93.2 ตามลำดับ

เมื่อขนาดตัวอย่างเท่ากับ 25 และ 30 สถิติทดสอบซีพี สถิติทดสอบเอไอซี และสถิติทดสอบเอฟบางส่วน คัดเลือกตัวแบบได้ถูกต้องใกล้เคียงกันและคัดเลือกตัวแบบได้ถูกต้องมากกว่าเกณฑ์อื่น ขนาดตัวอย่างเท่ากับ 25 คัดเลือกตัวแบบได้ถูกต้องร้อยละ 95.1 94.2 และ 95.1 ตามลำดับ ขนาดตัวอย่างเท่ากับ 30 คัดเลือกตัวแบบได้ถูกต้องร้อยละ 96.8 95.8 และ 94.7 ตามลำดับ

เมื่อขนาดตัวอย่างเท่ากับ 50 100 และ 150 สถิติทดสอบซีพีและสถิติทดสอบเอไอซี คัดเลือกตัวแบบได้ถูกต้องใกล้เคียงกันและคัดเลือกตัวแบบได้ถูกต้องมากกว่าเกณฑ์และสถิติทดสอบอื่น ขนาดตัวอย่างเท่ากับ 50 คัดเลือกตัวแบบได้ถูกต้องร้อยละ 99.1 และ 99.0 ตามลำดับ ขนาดตัวอย่างเท่ากับ 100 สถิติทดสอบทั้งสองคัดเลือกตัวแบบได้ถูกต้องร้อยละ 99.9 ขนาดตัวอย่างเท่ากับ 150 สถิติทดสอบทั้งสองคัดเลือกตัวแบบได้ถูกต้องร้อยละ 100.0

ทุกขนาดตัวอย่าง เกณฑ์เอไอซี มีร้อยละที่ตัวแปรอิสระในตัวแบบมีมากกว่าความเป็นจริง มากกว่าเกณฑ์และสถิติทดสอบอื่น

กราฟแสดงความสัมพันธ์ระหว่างการคัดเลือกตัวแบบด้วยเกณฑ์และสถิติทดสอบต่างๆ กับขนาดตัวอย่าง เมื่อตัวแปรอิสระมีการแจกแจงปกติ ได้แสดงร้อยละของตัวแบบที่คัดเลือกได้ถูกต้อง ในภาพที่ 2

ตารางที่ 2 ร้อยละของตัวแบบที่คัดเลือกได้ตามเกณฑ์และสถิติทดสอบ จำแนกตามผลการคัดเลือก และขนาดตัวอย่าง เมื่อตัวแปรอิสระมีการแจกแจงปกติ

ขนาดตัวอย่าง	ผลการคัดเลือก	เกณฑ์และสถิติทดสอบ				
		CP	AIC	CP-test	AIC-test	Partial F-test
10	Over-fit	51.2	86.2	47.6	59.7	3.8
	Fit	48.6	13.7	52.2	40.2	92.5
	Under-fit	0.2	0.1	0.2	0.1	3.7
15	Over-fit	45.7	66.6	21.0	25.9	4.7
	Fit	54.3	33.4	79.0	74.1	95.3
	Under-fit	0.0	0.0	0.0	0.0	0.0
20	Over-fit	45.5	61.8	11.5	13.6	6.8
	Fit	54.5	38.2	88.5	86.4	93.2
	Under-fit	0.0	0.0	0.0	0.0	0.0
25	Over-fit	41.7	52.2	4.9	5.8	4.9
	Fit	58.3	47.8	95.1	94.2	95.1
	Under-fit	0.0	0.0	0.0	0.0	0.0
30	Over-fit	41.6	51.6	3.2	4.2	5.3
	Fit	58.4	48.4	96.8	95.8	94.7
	Under-fit	0.0	0.0	0.0	0.0	0.0
50	Over-fit	40.7	45.8	0.9	1.0	6.0
	Fit	59.3	54.2	99.1	99.0	94.0
	Under-fit	0.0	0.0	0.0	0.0	0.0
100	Over-fit	41.5	45.3	0.1	0.1	5.7
	Fit	58.5	54.7	99.9	99.9	94.3
	Under-fit	0.0	0.0	0.0	0.0	0.0
150	Over-fit	42.1	44.4	0.0	0.0	4.1
	Fit	57.9	55.6	100.0	100.0	95.9
	Under-fit	0.0	0.0	0.0	0.0	0.0

จากภาพที่ 2 พบว่า ร้อยละของตัวแบบที่คัดเลือกได้ถูกต้อง เมื่อตัวแปรอิสระมีการแจกแจงปกติ คล้ายกับกรณีที่ตัวแปรอิสระมีการแจกแจงเอกรูป

สรุปผลและวิจารณ์ผล

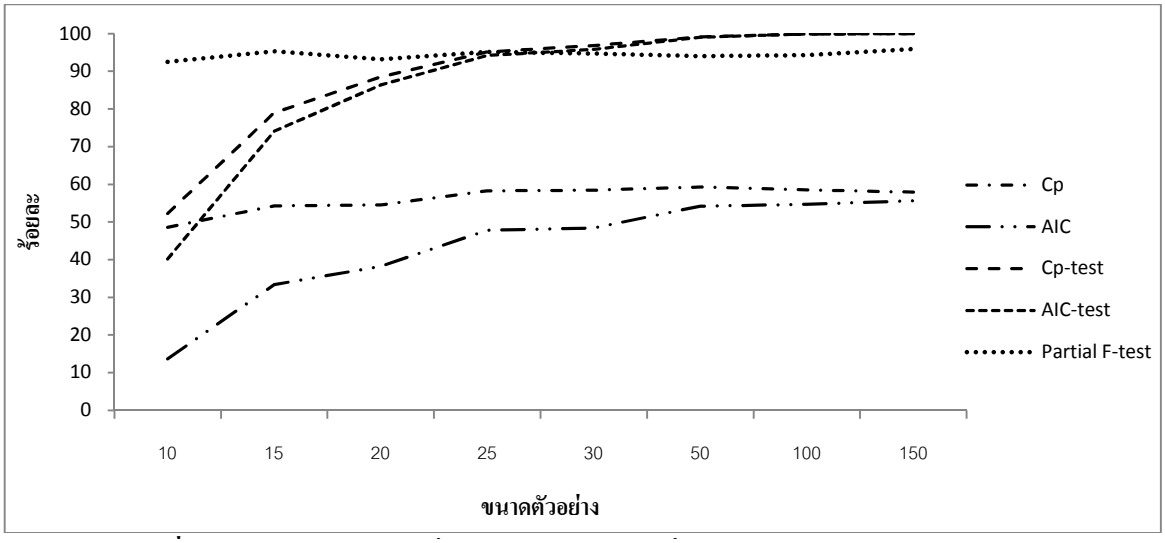
จากการจำลองข้อมูลเมื่อตัวแปรอิสระมีการแจกแจงเอกรูปและมีการแจกแจงปกติ ได้ผลการคัดเลือกตัวแบบสอดคล้องกัน ดังนี้

เมื่อขนาดตัวอย่างเล็ก (ขนาดตัวอย่างเท่ากับ 10 15 และ 20) สถิติทดสอบเอฟบางส่วนคัดเลือกตัวแบบได้ถูกต้องมากกว่าเกณฑ์และสถิติทดสอบอื่น

เมื่อขนาดตัวอย่างปานกลาง (ขนาดตัวอย่างเท่ากับ 25 และ 30) สถิติทดสอบซีพี สถิติทดสอบเอไอซี และสถิติทดสอบเอฟบางส่วน คัดเลือกตัวแบบได้ถูกต้องใกล้เคียงกันและคัดเลือกตัวแบบได้ถูกต้องมากกว่าเกณฑ์อื่น

เมื่อขนาดตัวอย่างใหญ่ (ขนาดตัวอย่างเท่ากับ 50 100 และ 150) สถิติทดสอบซีพีและสถิติทดสอบเอไอซี คัดเลือกตัวแบบได้ถูกต้องใกล้เคียงกันและคัดเลือกตัวแบบได้ถูกต้องมากกว่าเกณฑ์และสถิติทดสอบอื่น

เมื่อขนาดตัวอย่างเล็ก เกณฑ์เอไอซี มีร้อยละของตัวแบบ



ภาพที่ 2 ร้อยละของตัวแบบที่คัดเลือกได้ถูกต้อง เมื่อตัวแปรอิสระมีการแจกแจงปกติ

ที่ตัวแปรอิสระมีมากกว่าความเป็นจริง มากกว่าจำนวนร้อยละของตัวแบบที่คัดเลือกได้ถูกต้อง ซึ่งเป็นไปตามทฤษฎีของเกณฑ์เอไอซี ที่คัดเลือกตัวแบบได้ดีเมื่อตัวอย่างมีขนาดใหญ่เท่านั้น

เมื่อขนาดตัวอย่างเพิ่มขึ้น เกณฑ์และสถิติทดสอบต่างๆ มีร้อยละของตัวแบบที่คัดเลือกได้ถูกต้องเพิ่มขึ้น ยกเว้นสถิติทดสอบเอฟบางส่วน ที่ร้อยละของตัวแบบที่คัดเลือกได้ถูกต้องคงที่ ซึ่งคัดเลือกตัวแบบได้ถูกต้องประมาณร้อยละ 95 ทุกขนาดตัวอย่าง

สถิติทดสอบซีพีและสถิติทดสอบเอไอซี คัดเลือกตัวแบบได้ดีเมื่อขนาดตัวอย่างใหญ่พอ เนื่องจากสถิติทดสอบทั้งสองเป็นสถิติทดสอบเชิงกำกับ

เอกสารอ้างอิง

ทรงศิริ แต้สมบัติ. (2548). *การวิเคราะห์การถดถอย*. (พิมพ์ครั้งที่ 3). กรุงเทพฯ : มหาวิทยาลัยเกษตรศาสตร์.

รจนา คุณเจริญจิต. (2549). *การเปรียบเทียบเกณฑ์การคัดเลือกตัวแบบในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ*. วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์.

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. *In 2nd International Symposium on Information Theory*. B.N. Petrov and F. Csaki, eds. Akademiai Kiado, Budapest, 267-281.

Canal, L. (2005). A Normal Approximation for the Chi-square Distribution. *Computational Statistics & Data Analysis, 48*, 803-808.

Hurvich, C.M., & Tsai, C.L. (1989). Regression and Time Series Model Selection in Small Sample. *Biometrika Trust, 76*(2), 297-307.

Mallows, C.L. (1973). Some Comments on CP. *Technometrics, 15*(4), 661-675.

McQuarrie, A., Robert, S., & Tsai, C.L. (1997). The Model Selection Criterion AICu. *Statistics and Probability Letters, 34*, 285-292.

McQuarrie, A.D.R., & Tsai, C.L. (1998). *Regression and Time Series Model Selection*. Singapore : World Scientific.

Montgomery, D.C., Peak, E.A., & Vining, G.G. (2006). *Introduction to Linear Regression Analysis*. (4th ed). New York : Wiley.

Sheather, S.J. (2009). *A Modern Approach to Regression with R*. New York : Springer.

Yanagihara, H., & Ohmoto, C. (2005). On Distribution of AIC in Linear Regression Models. *Journal of Statistical Planning and Inference, 133*, 417-433.