เกณฑ์และสถิติทดสอบในการคัดเลือกตัวแปรอิสระในตัวแบบการถดถอยเชิงเส้นแบบพหุ กรณีที่ไม่ใช้ตัวแบบเต็มรูป

นันทพร บุญสุข[1]
มหาวิทยาลัยราชภัฏราชนครินทร์
และ จิราวัลย์ จิตรเวช[2*]
สถาบันบัณฑิตพัฒนบริหารศาสตร์ แขวงคลองจั่น เขตบางกะปิ กรุงเทพฯ 10240

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเสนอการปรับเกณฑ์ซีพีในกรณีที่ไม่ใช้ตัวแบบเต็มรูปเนื่องจากตัวแปรอิสระทั้งหมดที่นำมาพิจารณาเพื่อสร้างตัวแบบเต็มรูปมีความสัมพันธ์เชิงเส้นต่อกันอย่างสมบูรณ์ จากนั้นเปรียบเทียบการคัดเลือกตัวแบบโดยใช้เกณฑ์ซีพีที่ปรับแล้วกับการคัดเลือกตัวแบบที่ใช้เกณฑ์เอไอซี และเกณฑ์บีไอซี นอกจากนี้ได้นำเสนอสถิติทดสอบซีพีที่ปรับแล้วในการทดสอบเพื่อค้นหากลุ่มของสมการถดถอยที่สามารถใช้ในการอธิบายตัวแปรตาม โดยใช้วิธีการจำลองข้อมูล กำหนดขนาดตัวอย่างเท่ากับ 25 50 และ 100 ผลการวิจัยพบว่าเกณฑ์ซีพีที่ปรับแล้วให้ผลการคัดเลือกตัวแบบสอดคล้องกับการใช้เกณฑ์เอไอซีและเกณฑ์บีไอซี กล่าวคือเมื่อขนาดตัวอย่างเพิ่มขึ้นร้อยละของการคัดเลือกตัวแบบได้ถูกต้องมีค่าสูงขึ้น จากสถิติทดสอบซีพีที่ปรับแล้วโดยใช้ระดับนัยสำคัญ 0.01 0.05 และ 0.1 พบว่าเมื่อขนาดตัวอย่างเพิ่มขึ้น ร้อยละของการคัดเลือกกลุ่มของตัวแบบที่เพียงพอในการอธิบายตัวแปรตามได้เป็นตัวแบบที่ถูกต้องมีค่าเพิ่มขึ้น และเมื่อระดับนัยสำคัญลดลงร้อยละการคัดเลือกกลุ่มของตัวแบบที่ถูกต้องมีค่าเพิ่มขึ้น

คำสำคัญ : การคัดเลือกตัวแบบเกณฑ์ซีพี / เกณฑ์เอไอซี / เกณฑ์บีไอซี

* *Corresponding author E-mail : jirawan@as.nida.ac.th*

1 *อาจารย์ สาขาคณิตศาสตร์และสถิติประยุกต์*

2 *รองศาสตราจารย์ คณะสถิติประยุกต์*

Criterion and Test Statistic for Selecting Multiple Linear Regression Models without Full Model

Nantaporn Boonsuk[1]

Rajabhat Rajanagarindra University

and Jirawan Jitthavech[2*]

National Institute of Development Administration, Klogjan, Bangkapi, Bangkok, 10240

Abstract

The objective of this research was to modify C_P criterion for a case wherea full model cannot be constructed due to the perfect multicollinearity problem. The modified C_P criterion was compared with Akaike information criterion and Bayesian information criterion via simulation with the sample sizes of 25, 50 and 100. From the simulation, it was found that the modified C_P criterion gave the consistent results with those of the Akaike information criterion and Bayesian information criterion, i.e., the percentage of selecting the correct models increased as the sample size increased. The test statistic for the modified C_P criterion was proposed to select a group of acceptance regression models with the significant levels of 0.01, 0.05 and 0.1. The percentage of selecting correct models also increased as the sample size increased. However, when the significant level of model selection decreased, the percentage of the correct models increased.

Keywords : Model selection / C_p / AIC / BIC

* *Corresponding author E-mail: jirawan@as.nida.ac.th*

1 *Lecturer, Department of Mathematics and Applied Statistic.*

2 *Associate Professor, Department of Applied Statistic.*

1. บทนำ

Mallow [6] ได้เสนอสถิติในการคัดเลือกตัวแปรอิสระ เรียกว่าสถิติซีพี (C_p) การสร้างเกณฑ์ในการคัดเลือกตัวแปรอิสระมีแนวคิดเพื่อให้เกณฑ์ที่สร้างขึ้นปราศจากหน่วยวัด ภายใต้ตัวแบบการถดถอยเชิงเส้นแบบพหุที่อยู่ในรูปเชิงเส้นของพารามิเตอร์และตัวแปรอิสระโดยนิยามให้

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

Γ_p เป็นค่าคาดหมายของค่าคลาดเคลื่อนกำลังสองของค่าพยากรณ์กับค่าคาดหมายของ y หารด้วย σ^2 ซึ่งเป็นค่าคาดหมายของค่ากำลังสองเฉลี่ยในตัวแบบจริง ตัวแปรอิสระที่อยู่ในตัวแบบที่ให้ค่า Γ_p ต่ำแสดงว่าค่าพยากรณ์ของ y มีค่าใกล้ค่าคาดหมายของ y

เกณฑ์ซีพีสามารถเขียนได้เป็น

$$\Gamma_p = \frac{E\,(\mathrm{SSE})}{\sigma^2} - (n - 2p) \quad \textbf{(2)}$$

เมื่อ SSE เป็นค่าความคลาดเคลื่อนกำลังสองของตัวแบบการถดถอยในทางปฏิบัติค่าคาดหมายของค่ากำลังสองเฉลี่ยในตัวแบบที่ถูกต้องไม่ทราบค่าจึงประมาณด้วยค่ากำลังสองเฉลี่ยของตัวแบบเต็มรูป ซึ่งตัวแบบเต็มรูปหมายถึง ตัวแบบที่ประกอบด้วยตัวแปรอิสระทุกตัวที่ต้องพิจารณา ในบางสถานการณ์โดยเฉพาะทางเศรษฐมิติอาจไม่สามารถสร้างตัวแบบเต็มรูปได้เนื่องจากตัวแปรอิสระที่ใช้ในการพิจารณามีความสัมพันธ์กัน ตัวอย่างเช่น ในประเทศสหรัฐอเมริกา ปี ค.ศ. 1960-1982 ได้ทำการศึกษาการบริโภคไก่ต่อคน (แหล่งข้อมูล:เก็บข้อมูลโดย Robert J. Fisher และเป็นต้นฉบับในกระทรวงพาณิชย์ของประเทศสหรัฐอเมริกา) โดยมีตัวแปร y คือ การบริโภคไก่ต่อคน (ปอนด์), X_1 คือ รายได้จริงหลังหักภาษีส่วนบุคคล (ดอลล่าร์), X_2 คือ ราคาขายปลีกของไก่ต่อปอนด์ (เซนต์), X_3 คือ ราคาขายปลีกของหมูต่อปอนด์ (เซนต์), X_4 คือ ราคาขายปลีกของเนื้อต่อปอนด์ (เซนต์) และ X_5 คือ ราคาขายเฉลี่ยของหมูและเนื้อต่อปอนด์ (เซนต์) พิจารณาตัวแบบได้ดังนี้

ตัวแบบที่ 1 $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon_1$ **(3)**

ตัวแบบที่ 2 $y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_5 + \varepsilon_2$ **(4)**

ตัวแบบเต็มรูป $y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 X_4 + \gamma_5 X_5 + \varepsilon$ **(5)**

หากต้องการเปรียบเทียบตัวแบบที่ 1 และตัวแบบที่ 2 โดยใช้เกณฑ์ซีพี ต้องหาค่า MSE ของตัวแบบเต็มรูป ในสมการ (5) ที่ประกอบด้วยตัวแปรอิสระ 5 ตัว แต่เนื่องจากตัวแปรอิสระเป็นตัวแปรที่เกิดจากตัวแปร ดังนั้นไม่สามารถประมาณค่าของตัวแบบเต็มรูปที่จะใช้ในการประมาณค่าของเกณฑ์ของตัวแบบที่ 1 และ ตัวแบบที่ 2 ได้ จึงต้องปรับเกณฑ์ โดยไม่ใช้ค่ากำลังสองเฉลี่ยของตัวแบบเต็มรูป

ในงานวิจัยครั้งนี้ผู้วิจัยได้เสนอการปรับเกณฑ์ซีพีในกรณีที่ไม่สามารถสร้างตัวแบบเต็มรูปได้เนื่องจากตัวแปรอิสระที่นำมาพิจารณามีความสัมพันธ์เชิงเส้นต่อกัน ดังตัวอย่างที่กล่าวแล้วข้างต้นและทำการเปรียบเทียบการคัดเลือกตัวแบบโดยใช้เกณฑ์ซีพีที่ปรับแล้วกับเกณฑ์เอไอซี และเกณฑ์บีไอซี และเสนอสถิติทดสอบเกณฑ์ซีพีที่ปรับแล้ว เพื่อค้นหากลุ่มของสมการถดถอยที่เพียงพอที่ใช้ในการอธิบายตัวแปรตามได้

นิยามคำศัพท์ที่เกี่ยวข้อง

1) ตัวแบบเต็มรูป (Full Model) หมายถึง ตัวแบบการถดถอยเชิงเส้นแบบพหุที่ประกอบด้วยตัวแปรอิสระทั้งหมดจากทุกตัวแบบทีทำการศึกษาทั้งตัวแปรที่เกี่ยวข้องและไม่เกี่ยวข้องกับตัวแปรตาม

2) ตัวแบบอ้างอิง (Reference Model) หมายถึง ตัวแบบที่มีค่า MSE ต่ำสุดในทุกตัวแบบที่พิจารณา

3) ตัวแบบที่เพียงพอในการอธิบายตัวแปรตาม หมายถึง ตัวแบบที่มีค่าของเกณฑ์ซีพีที่ปรับแล้วไม่แตกต่างกันกับตัวแบบอ้างอิง

4) ตัวแบบที่ over fit หมายถึง ตัวแบบที่มีตัวแปรอิสระที่เกี่ยวข้องกับตัวแปรตามครบทุกตัว และมีตัวแปรอิสระที่ไม่เกี่ยวข้องกับตัวแปรตามรวมอยู่ในตัวแบบ

5) ตัวแบบที่ถูกต้อง หมายถึง ตัวแบบที่มีเฉพาะตัวแปรอิสระที่เกี่ยวข้องกับตัวแปรตามอยู่ในตัวแบบ

6) ตัวแบบที่ under fit หมายถึง ตัวแบบที่มีตัวแปรอิสระที่เกี่ยวข้องกับตัวแปรตามไม่ครบทุกตัว และไม่มีตัวแปรอิสระที่ไม่เกี่ยวข้องกับตัวแปรตามรวมอยู่ในตัวแบบ

7) ตัวแบบที่ misspecified หมายถึง ตัวแบบที่มีตัวแปรอิสระที่เกี่ยวข้องกับตัวแปรตามไม่ครบทุกตัว และมีตัวแปรอิสระที่ไม่เกี่ยวข้องกับตัวแปรตามรวมอยู่ในตัวแบบ

8) ร้อยละของการคัดเลือกตัวแบบที่ถูกต้องหมายถึง อัตราส่วนของจำนวนครั้งที่คัดเลือกตัวแบบที่ถูกต้อง ต่อจำนวนครั้งทั้งหมดที่ใช้ในการจำลองข้อมูลแล้วคูณกับ 100

9) ตัวแบบการถดถอยเชิงเส้นแบบพหุกรณีที่ไม่สามารถสร้างตัวแบบเต็มรูป หมายถึง ตัวแบบเต็มรูปที่ประกอบด้วยตัวแปรอิสระทุกตัวที่มาจากตัวแบบต่างๆ ที่นำมาพิจารณา ซึ่งตัวแปรอิสระบางตัวมีความสัมพันธ์ในลักษณะเป็นผลรวมเชิงเส้น (Linear Combination) ต่อกัน ทำให้ตัวแบบเต็มรูปที่สร้างขึ้นไม่สามารถคำนวณหาเมทริกซ์ผกผัน ได้ดังเช่นสมการที่ (11)

2. ทฤษฎีที่ใช้ในการศึกษา

1) ศึกษาภายใต้ตัวแบบการถดถอยเชิงเส้นแบบพหุที่อยู่ในรูปเชิงเส้นของพารามิเตอร์และตัวแปรอิสระ โดยมีรูปแบบทั่วไปคือ

$$y = X\beta + \varepsilon \quad \textbf{(6)}$$

โดยที่ y เป็นเวกเตอร์ของตัวแปรตาม y ขนาด $n \times 1$, X เป็นเมทริกซ์ของตัวแปรอิสระขนาด $n \times (k+1)$, β เป็นเวกเตอร์ของพารามิเตอร์ของตัวแบบขนาด $(k+1) \times 1$, ε เป็นเวกเตอร์ของความคลาดเคลื่อน ขนาด $n \times 1$ ภายใต้ข้อกำหนด $\varepsilon \sim N(0, \sigma^2 I_n)$, I_n เป็นเมทริกซ์เอกลักษณ์ขนาด $n \times n$, n เป็นขนาดตัวอย่างและ k เป็นจำนวนตัวแปรอิสระในตัวแบบ

2) ตัวแปรอิสระในการศึกษาแบ่งออกเป็น 2 กลุ่ม คือ ตัวแปรอิสระที่เกี่ยวข้องกับตัวแปรตามโดยมีทฤษฎีหรือการศึกษาในอดีตสนับสนุนว่าตัวแปรอิสระเหล่านี้มีความสำคัญที่จะต้องอยู่ในตัวแบบ และตัวแปรอีกกลุ่มหนึ่งเป็นตัวแปรที่ไม่เกี่ยวข้อง ซึ่งอาจจะมีความสำคัญในการให้ข้อสนเทศเกี่ยวกับตัวแปรตามแต่ไม่มีทฤษฎีหรือการศึกษาในอดีตสนับสนุนจึงต้องมีการพิจารณาว่าตัวแปรเหล่านี้ควรอยู่ในตัวแบบหรือไม่

3) ตัวแบบที่ใช้ในการศึกษาเป็นตัวแบบการถดถอยเชิงเส้นแบบพหุที่มีตัวแปรอิสระอยู่ในตัวแบบ จำนวน 13 ตัวแปรได้แก่ $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}$, เมื่อ $X_{10} = X_2 + X_3 + X_{12} = X_3 + X_4$ และให้ X_5 เป็นตัวแปรอิสระที่อยู่ในกลุ่มของตัวแปรอิสระที่ไม่มีความเกี่ยวข้องกับตัวแปรตาม y ตัวแบบที่ทำการศึกษามีดังนี้

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon_1 \quad \textbf{(7)}$$

$$y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_7 + \alpha_3 X_8 + \alpha_4 X_5 + \varepsilon_2 \quad \textbf{(8)}$$

$$y = \gamma_0 + \gamma_1 X_9 + \gamma_2 X_{10} + \gamma_3 X_{11} + \gamma_4 X_5 + \varepsilon_3 \quad \textbf{(9)}$$

$$y = \theta_0 + \theta_1 X_{12} + \theta_2 X_{13} + \theta_4 X_5 + \varepsilon_4 \quad \textbf{(10)}$$

ตัวแบบเต็มรูป $$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_{12} X_{12} + \beta_{13} X_{13} + \varepsilon \quad \textbf{(11)}$$

4) Mallow [6] ได้เสนอเกณฑ์การคัดเลือกตัวแบบที่ใช้อัตราส่วนระหว่างค่าคาดหมายของค่าความคลาดเคลื่อนกำลังสองของค่าพยากรณ์ ($\hat{y}$) กับค่าคาดหมายของ y ($E(y)$) ซึ่งนิยามตาม (1) การใช้ตัวสถิติ C_p เป็นเกณฑ์ในการพิจารณาตัวแบบถ้าสมการถดถอยมีความเอนเอียงเพียงเล็กน้อยค่าซีพีจะเข้าใกล้ค่าพี แต่ถ้าสมการมีความเอนเอียงมากค่าของซีพีจะมากกว่าพี โดยปกติแล้วต้องการสมการที่มีค่าซีพีต่ำ [8]

5) Akaike [2] ได้เสนอเกณฑ์การคัดเลือกตัวแบบ คือ เกณฑ์เอไอซี ซึ่งสร้างจากการประมาณความแปรปรวนของข้อสนเทศคูลแบล็ค-ไลเบอร์ (Kullback Leibler Information) ที่มีสมบัติไม่เอนเอียง เมื่อขนาดตัวอย่างใหญ่การคัดเลือกตัวแบบโดยใช้เกณฑ์เอไอซีเลือกตัวแบบที่ให้ค่าเอไอซีต่ำสุด เป็นตัวแบบที่ใช้ในการอธิบายตัวแปรตามได้ และเกณฑ์เอไอซี คัดเลือกตัวแบบได้ดีเมื่อตัวอย่างมีขนาดใหญ่ เนื่องจากเกิดความผิดพลาดที่คัดเลือกตัวแบบ

ที่มีตัวแปรอิสระจำนวนมากเกินความจำเป็นมีความเป็นไปได้สูง [7] เกณฑ์เอไอซี นิยามโดย

$$\mathrm{AIC} = n\cdot\ln\left(\frac{\mathrm{SSE}}{n}\right)+2p \qquad \textbf{(12)}$$

เมื่อ n เป็นขนาดตัวอย่าง SSE เป็นค่าความคลาดเคลื่อนกำลังสองของตัวแบบการถดถอย p เป็นจำนวนพารามิเตอร์ในตัวแบบการถดถอย และ ln เป็นลอการิทึมฐานอี

6) Sawa [10] ได้พัฒนาเกณฑ์การคัดเลือกตัวแบบที่ได้มาจากการดัดแปลงแบบเบส์ของเกณฑ์ AIC เรียกว่า เกณฑ์ข้อสนเทศของเบส์ (Bayesian Information Criteria : BIC) ซึ่งเกณฑ์ BIC จะเลือกตัวแบบที่ให้ค่าบีไอซีต่ำสุด เป็นตัวแบบที่ถูกต้องโดยมีสูตรดังนี้ (Beal [4])

$$\mathrm{BIC} = n\cdot\ln\left(\frac{\mathrm{SSE}}{n}\right)+\frac{2(p+1)\,n\sigma^2}{\mathrm{SSE}}-\frac{2n^2\sigma^4}{\mathrm{SEE}^2} \qquad \textbf{(13)}$$

เมื่อ n เป็นขนาดตัวอย่าง SSE เป็นค่าความคลาดเคลื่อนกำลังสองของตัวแบบการถดถอย σ^2 เป็นค่าความคลาดเคลื่อนเฉลี่ยกำลังสองของตัวแบบการถดถอยและ p เป็นจำนวนพารามิเตอร์ในตัวแบบการถดถอย

3. การพัฒนาเกณฑ์การคัดเลือกตัวแบบ

งานวิจัยนี้เสนอการปรับเกณฑ์ซีพีและ เสนอสถิติทดสอบเกณฑ์ซีพีที่ปรับแล้ว กรณีต้องการเปรียบเทียบ 2 ตัวแบบ ที่ตัวแปรอิสระที่นำมาพิจารณามีความสัมพันธ์เชิงเส้นต่อกัน ทำให้ไม่สามารถประมาณค่า MSE ของตัวแบบเต็มรูปที่จะใช้ในการหาค่าของเกณฑ์ซีพีได้ ดังนี้

การพัฒนาเกณฑ์ซีพีโดยการปรับ

ทฤษฎีบทที่ 1 กำหนดให้

ตัวแบบที่ถูกต้อง: $y = X_{\mathrm{T}}\beta_{\mathrm{T}} + \varepsilon$ โดยที่ $\varepsilon \sim N(0, \sigma^2 I)$

ตัวแบบที่ 1: $y = X\beta + \varepsilon_1$ โดยที่ $\varepsilon_1 \sim N(0, \sigma_1^2 I)$

ตัวแบบที่ 2: $y = Z\alpha + \varepsilon_2$ โดยที่ $\varepsilon_2 \sim N(0, \sigma_2^2 I)$

ภายใต้ตัวแบบที่ 1 ถ้าให้ $\varphi = E[(\hat{y} - E(y))'(\hat{y} - E(y))]$ แล้วจะได้ว่า

$\varphi_1 = E(\mathrm{SSE}_1) - (\mathrm{n} - 2p_1)\,\sigma^2$ และภายใต้ตัวแบบที่ 2 จะได้ว่า

$\varphi_2 = E(\mathrm{SSE}_2) - (\mathrm{n} - 2p_2)\,\sigma^2$ เมื่อ SSE_1 เป็นค่าความคลาดเคลื่อนกำลังสองของตัวแบบที่ 1 และ SSE_2 เป็นค่าความคลาดเคลื่อนกำลังสองของตัวแบบที่ 2

พิสูจน์

ภายใต้ตัวแบบที่ 1

$$\begin{aligned}
\varphi_1 &= E\left[(\hat{y} - E(y))'(\hat{y} - E(y))\right] \\
&= E\left[(H_{\mathrm{x}}y - E(y))'(H_{\mathrm{x}}y - E(y))\right] ;\, H = X(X'X)^{-1}\,\mathrm{X}' \\
&= E\left\{y'H'_{\mathrm{x}}H_{\mathrm{x}}y - y'H_{\mathrm{x}}E(y) - (E(y))'H_{\mathrm{x}}y + (E(y))'\,E(y)\right\} \\
&= E(y'H'_{\mathrm{x}}y) - 2(E(y))'\,H_{\mathrm{x}}E(y) + (E(y))'\,E(y) \\
&= tr(\sigma^2 H_{\mathrm{x}}) + (E(y))'\,H_{\mathrm{x}}E(y) - 2(E(y))'\,H_{\mathrm{x}}E(y) + (E(y))'\,E(y) \\
&= \sigma^2 p_1 + (E(y))'\,(I - H_{\mathrm{x}})E(y) ;\, p_1 = k_1 + 1
\end{aligned}$$

จาก ทฤษฎีบทที่ 4.6.1 ใน [5] จะได้ว่า

$E(y'(I - H)\,y) = tr\,(I - H)\sigma^2 + (E(y))'\,(I - H)E(y)$ หรือ

$(E(y))'(I - H)E(y) = E(y'(I - H)y) - tr(I - H)\sigma^2$

ดังนั้น

$$\begin{aligned}
\varphi_1 &= \sigma^2 p_1 - \sigma^2\,(n - p_1) + E(y'(I - H)\,y) \\
&= E(\mathrm{SSE}_1) - (n - 2p_1)\sigma^2
\end{aligned}$$

ในทำนองเดียวกัน

$\varphi_2 = E\left[(\hat{y} - E(y))'(\hat{y} - E(y))\right] = E(\mathrm{SSE}_2) - (n - 2p_1)\sigma^2$ เมื่อ $p_2 = k_1 + 1$

จากทฤษฎีบทที่ 1 เกณฑ์ซีพีตาม Mallow [6] ได้ว่า

$$\Gamma_{\mathrm{p1}} = \frac{E(\mathrm{SSE}_1)}{\sigma^2} - (n - 2p_1)$$

และ

$$\Gamma_{\mathrm{p_2}} = \frac{E(\mathrm{SSE}_2)}{\sigma^2} - (n - 2p_2)$$

เกณฑ์ซีพีของสมการที่ 1 $C_{p_1} = \frac{SSE_1}{\hat{\sigma}^2} - (n - 2p_1)$

เกณฑ์ซีพีของสมการที่ 2 $C_{p_2} = \frac{SSE_2}{\hat{\sigma}^2} - (n - 2p_2)$

เกณฑ์ซีพีตัวประมาณค่าของ σ^2 ใช้ค่า MSE ของตัวแบบเต็มรูป แต่ในกรณีที่หาตัวแบบเต็มรูปไม่ได้ตามที่ได้กล่าวแล้ว จึงทำการปรับเกณฑ์ซีพีของมาล์โลวส์ดังนี้

$\Gamma_p^* = \frac{1}{\sigma_{ref}^2} E[(\hat{y} - E(y))'(\hat{y} - E(y))]$ การใช้ σ_{ref}^2 แทน σ^2

เนื่องจาก σ_{ref}^2 ที่นิยามให้ $\sigma_{ref}^2 = \min\left(\sigma_j^2\right), j = 1, 2, ..., m$ เมื่อ m เป็นจำนวนสมการที่ต้องการเปรียบเทียบ ดังนั้นตัวประมาณค่า $\hat{\sigma}_{ref}^2 = \min_{1 \le j \le m}\left(MSE_{p_j}\right)$

ให้ $C_{p_i}^*$ เป็นตัวประมาณค่าของ $\Gamma_{p_i}^*$ ดังนั้น

$$C_{p_1}^* = \frac{SSE_1}{\hat{\sigma}_{ref}^2} - (n - 2p_1)$$

และ

$$C_{p_2}^* = \frac{SSE_2}{\hat{\sigma}_{ref}^2} - (n - 2p_2)$$

4. การพัฒนาสถิติทดสอบ

การทดสอบค่าของเกณฑ์ซีพีในแต่ละตัวแบบว่าตัวแบบการถดถอยตัวแบบใดมีค่าเกณฑ์ซีพีไม่แตกต่างไปจากค่าเกณฑ์ซีพีของตัวแบบที่มีค่าเกณฑ์ซีพีต่ำสุด หมายถึง ตัวแบบเหล่านั้นเพียงพอในการอธิบายตัวแปรตาม ทดสอบสมมติฐาน $H_0 : \Gamma_{p_i}^* = \Gamma_{p_{ref}}^*$ หรือ $H_0 : \Gamma_{p_i}^* = p_{ref}$ เทียบกับ $H_1 : \Gamma_{p_i}^* = \Gamma_{p_{ref}}^*$ หรือ $H_1 : \Gamma_{p_i}^* = p_{ref}$ ได้ดังนี้

จากทฤษฎีบทที่ 1 พิจารณาเกณฑ์ของซีพีที่ปรับของตัวแบบที่ i

$$C_{p_i}^* = \frac{SSE_i}{\hat{\sigma}_{ref}^2} - (n - 2p_i) \quad \textbf{(14)}$$

เมื่อ SSE_i เป็นค่าความคลาดเคลื่อนกำลังสองของตัวแบบการถดถอยที่พิจารณา $\hat{\sigma}_{ref}^2$ เป็นค่าความลาดเคลื่อนเฉลี่ยกำลังสองต่ำสุดของตัวแบบการถดถอยทั้งหมดที่พิจารณา n เป็นขนาดตัวอย่าง p_i เป็นจำนวนพารามิเตอร์ในตัวแบบการถดถอยที่พิจารณามีค่าเท่ากับ $k_i + 1$

เกณฑ์ที่ปรับแล้วของตัวแบบอ้างอิง

$$C_{p_{ref}}^* = \frac{SSE_{ref}}{\hat{\sigma}_{ref}^2} - (n - 2p_{ref}) \quad \textbf{(15)}$$

เมื่อ SSE_{ref} เป็นค่าความคลาดเคลื่อนกำลังสองของตัวแบบอ้างอิง $\hat{\sigma}_{ref}^2$ เป็นค่าความคลาดเคลื่อนเฉลี่ยกำลังสองต่ำสุดของตัวแบบการถดถอยทั้งหมดที่พิจารณา n เป็นขนาดตัวอย่าง p_{ref} เป็นจำนวนพารามิเตอร์ในตัวแบบอ้างอิงมีค่าเท่ากับ $k_{ref} + 1$

จากสมการที่ (14) และ (15) จะได้

$$C_{p_i}^* + (n - 2p_i) = \frac{SSE_i}{\hat{\sigma}_{ref}^2} \quad \textbf{(16)}$$

$$C_{p_{ref}}^* + (n - 2p_{ref}) = \frac{SSE_{ref}}{\hat{\sigma}_{ref}^2} \quad \textbf{(17)}$$

นำสมการ (16) / (17) จะได้

$$\frac{C_{p_i}^* + (n - 2p_i)}{C_{p_{ref}}^* + (n - 2p_{ref})} = \frac{SSE_i}{SSE_{ref}}$$

$$\frac{n - p_{ref}}{n - p_i} \cdot \frac{C_{p_i}^* + (n - 2p_i)}{C_{p_{ref}}^* + (n - 2p_{ref})} = \frac{MSE_i}{MSE_{ref}}$$

ให้ $F_{cal} = \frac{n - p_{ref}}{n - p_i} \cdot \frac{C_{p_i}^* + (n - 2p_i)}{C_{p_{ref}}^* + (n - 2p_{ref})}$

ทฤษฎีบทที่ 2 ค่าเกณฑ์ซีพีที่ปรับแล้วของตัวแบบอ้างอิงมีค่าเท่ากับ

พิสูจน์

$$C_{p_{ref}}^* = \frac{SSE_{ref}}{MSE_{ref}} - (n - 2p_{ref})$$

$$= \frac{SSE_{ref}}{SSE_{ref} / n - 2p_{ref}} - (n - 2p_{ref}) = p_{ref}$$

โดยอาศัยทฤษฎีบทที่ 2 ตัวสถิติทดสอบเกณฑ์ซีพีที่ปรับจะเท่ากับ

$$F_{cal} = \frac{C_{p_i}^* + (n - 2p_i)}{n - p_i}$$

พิจารณาเกณฑ์ซีพีที่ปรับแล้ว

$$C^*_{p_i} = \frac{SSE_i}{\hat{\sigma}^2_{ref}} - (n - 2p_i)$$

$$\frac{C^*_{p_i}}{n - p_i} = \frac{SSE_i/(n - p_i)}{\hat{\sigma}^2_{ref}} - \frac{(n - 2p_i)}{n - p_i} = \frac{\hat{\sigma}^2_i}{\hat{\sigma}^2_{ref}} - \frac{n - 2p_i}{n - p_i}$$

จาก Anderson [3] $\frac{\hat{\sigma}^2_i}{\hat{\sigma}^2_0}$ มีการแจกแจงแบบเอฟด้วยองศาอิสระ df, df_0 เมื่อ $\hat{\sigma}^2_i$ เป็นตัวประมาณค่าของ σ^2_i ของตัวแบบที่ i และ $\hat{\sigma}^2_i$ เป็นตัวประมาณค่าของ σ^2_0 ของตัวแบบจริง ในที่นี้ใช้ตัวประมาณค่า $\hat{\sigma}^2_{ref}$ แทนตัวประมาณค่าของ $\hat{\sigma}^2_0$ ดังนั้นจึงสรุปได้ว่า $\frac{C^*_{p_i}}{n - p_i} = \frac{\hat{\sigma}^2_i}{\hat{\sigma}^2_{ref}} - \frac{n - 2p_i}{n - p_i}$ มีการแจกแจงเอฟโดยประมาณองศาเสรี (Degrees of Freedom) df_i, df_{ref} และด้วยพารามิเตอร์ไร้ศูนย์กลาง λ เนื่องจากค่าคาดหมายของสถิติทดสอบมีค่าไม่เท่ากับ 0 ดังนั้นค่าพารามิเตอร์ไร้ศูนย์กลาง λ มีค่าเท่ากับ $\frac{C^*_{p_i} - p_i}{n - p_i}$ [9] ภายใต้สมมุติฐานหลักเป็นจริง $\Gamma^*_{p_i} = \Gamma^*_{p_{ref}}$ หรือ $\sigma^2_i = \hat{\sigma}^2_{ref}$

$$E\left(\frac{C^*_{p_i}}{n - p_i}\right) = 1 - \frac{(n - 2p_i)}{n - p_i}$$

$$E\left(\frac{C^*_{p_i}}{n - p_i}\right) = \frac{p_i}{n - p_i}$$

เนื่องจากการแจกแจงแบบเอฟไร้ศูนย์กลาง (Non-central F Distribution) ประมาณค่ายากเพราะขึ้นกับพารามิเตอร์ 3 ตัวที่แตกต่างกัน (v_1, v_2, λ) จึงได้มีการประมาณค่าการแจกแจงเอฟไร้ศูนย์กลางด้วยการแจกแจงแบบเอฟ โดยมีความสัมพันธ์กันดังนี้ [11]

$$F_{\alpha(v_1, v_2, \lambda)} = F^*_{\alpha(g, v_2)} = \frac{F_{\alpha(v_1, v_2)} \cdot v_1}{v_1 + \lambda}$$

เมื่อ $g = \frac{(v_1 + \lambda)^2}{(v_1 + 2\lambda)} v_1$ เป็นองศาเสรีของตัวแบบที่พิจารณาและ v_2 เป็นองศาเสรีของตัวแบบอ้างอิงดังนั้นจะปฏิเสธ H_0 เมื่อ $F_{cal} > F^*_{\alpha(g, v_2)}$ ที่ระดับนัยสำคัญ α ซึ่งหมายถึง ตัวแบบที่พิจารณามีค่าของเกณฑ์ซีพีแตกต่างกันอย่างมีนัยสำคัญ α ตัวแบบที่มีเกณฑ์ซีพีสูงหมายถึงตัวแบบไม่เพียงพอที่จะอธิบายตัวแปรตามและ หากไม่ปฏิเสธสมมุติฐานหลักแสดงว่าตัวแบบมีค่าของเกณฑ์ซีพีไม่แตกต่างกันสามารถใช้ตัวแบบอธิบายตัวแปรตาม y ได้เช่นเดียวกับตัวแบบอ้างอิง

5. การจำลองข้อมูล

งานวิจัยนี้ศึกษากรณีตัวแปรอิสระมีการแจกแจงเอกรูป (Uniform distribution) และความคลาดเคลื่อนมีการแจกแจงปกติ (Normal distribution)ที่มีค่าเฉลี่ยเป็นศูนย์และความแปรปรวนคงที่โดยมีขั้นตอนในการจำลองข้อมูลดังนี้

1) กำหนดประชากร N = 100,000

2) สร้างตัวแปรอิสระและ ความคลาดเคลื่อนดังนี้, $X_1 \sim U(1,15)$, $X_2 \sim U(20,70)$, $X_3 \sim U(5,30)$, $X_4 \sim U(1,70)$, $X_5 \sim U(35,90)$, $X_6 \sim U(25,60)$, $X_7 \sim U(15,100)$, $X_9 \sim U(45,80)$, $\varepsilon_1 \sim N(0,20^2)$, $\varepsilon_2 \sim N(0,25^2)$, $\varepsilon_3 \sim U(0,18^2)$, $\varepsilon_4 \sim N(0,28^2)$,

3) สร้างตัวแปรอิสระที่มีความสัมพันธ์เชิงเส้นต่อกัน คือ $X_{10} = X_2 + X_3$ และ $X_{12} = X_3 + X_4$

4) กำหนดค่าสัมประสิทธิ์การถดถอย คือ β'_1 = (-20,1.5,3,2,1.2,0), β'_2 = (12,0.2,2,1.6,0), β'_3 = (-15,3,1.5,4,0), β'_4 = (-5,2,3,0) ในการสร้างตัวแปรตาม y หรืออีกนัยหนึ่งตัวแปรตาม y จะไม่ขึ้นอยู่กับตัวแปรอิสระ X_5

5) สร้างข้อมูลตัวแปรตามจากตัวแปรอิสระและค่าคลาดเคลื่อน โดยให้ตัวแปรตามมีลักษณะความสัมพันธ์เชิงเส้นในพารามิเตอร์กับตัวแปรอิสระซึ่งตัวแปรอิสระมีทั้งตัวแปรที่เกี่ยวข้องกับตัวแปรตามและที่ไม่เกี่ยวข้องกับตัวแปรตามและ ตัวแปรตามในแต่ละตัวแบบมีค่าเท่ากันและให้ตัวแปรอิสระ X_5 คือตัวแปรที่ไม่เกี่ยวข้องกับตัวแปรตาม y ดังตัวแบบการถดถอยดังนี้

สร้างตัวแบบที่ 1 $y = \beta_{01} + \beta_{11}X_1 + \beta_{21}X_2 + \beta_{31}X_3 + \beta_{41}X_4 + \beta_{51}X_5 + \varepsilon_1$

สร้างตัวแบบที่ 2 ให้ $z = \beta_{02} + \beta_{12}X_6 + \beta_{22}X_7 + \beta_{42}X_5 + \varepsilon_2,\ X_8 = \dfrac{(y - z)}{\beta_{32}}$
จะได้ตัวแบบที่ 2 คือ $y = \beta_{02} + \beta_{12}X_6 + \beta_{22}X_7 + \beta_{32}X_8 + \beta_{42}X_5 + \varepsilon_2$

สร้างตัวแบบที่ 3 ให้ $w = \beta_{03} + \beta_{13}X_9 + \beta_{23}X_{10} + \beta_{43}X_5 + \varepsilon_3,\ X_{11} = \dfrac{(y - w)}{\beta_{33}}$
จะได้ตัวแบบที่ 3 คือ $y = \beta_{03} + \beta_{13}X_9 + \beta_{23}X_{10} + \beta_{33}X_{11} + \beta_{43}X_5 + \varepsilon_3$

สร้างตัวแบบที่ 4 ให้ $t = \beta_{04} + \beta_{14}X_{12} + \beta_{34}X_5 + \varepsilon_4,\ X_{13} = \dfrac{(y - t)}{\beta_{24}}$
จะได้ตัวแบบที่ 4 คือ $y = \beta_{04} + \beta_{14}X_{12} + \beta_{24}X_{13} + \beta_{34}X_5 + \varepsilon_4$

6) สุ่มตัวอย่างขนาด 25, 50 และ 100 แต่ละขนาดสุ่มซ้ำ 100 ครั้ง โดยถือว่าตัวอย่างที่สุ่มได้เป็นตัวอย่างสุ่มมาจากตัวแบบทั้ง 4 ตัวแบบ

7) พิจารณา 20 ตัวแบบที่ประกอบด้วยตัวแปรอิสระดังตารางที่ 1 โดยตัวแปรอิสระแบ่งออกเป็น 4 กลุ่ม กลุ่มที่ 1 ประกอบด้วย X_1, X_2, X_3, X_4, X_5 กลุ่มที่ 2 ประกอบด้วย X_6, X_7, X_8, X_5 กลุ่มที่ 3 ประกอบด้วย X_9, X_{10}, X_{11}, X_5 และ กลุ่มที่ 4 ประกอบด้วย X_{12}, X_{13}, X_5

จากตารางที่ 1 ตัวแบบที่ 1.1-1.5 นำตัวแปรอิสระออก 1 ตัว จากตัวแบบที่ 1, ตัวแบบที่ 2.1-2.4 นำตัวแปรอิสระออก 1 ตัว จากตัวแบบที่ 2, ตัวแบบที่ 3.1-3.4 นำตัวแปรอิสระออก 1 ตัว จากตัวแบบที่ 3 และ ตัวแบบที่ 4.1-4.3 นำตัวแปรอิสระออก 1 ตัว จากตัวแบบที่ 4 ซึ่งไม่ได้พิจารณาทุกตัวแบบที่เป็นไปได้ ผลการศึกษาจึงไม่สามารถบอกร้อยละของตัวแบบที่ under fit

ตารางที่ 1 ตัวแปรอิสระที่อยู่ในแต่ละตัวแบบ

ตัวแบบที่	ตัวแปรอิสระ
1	X_1, X_2, X_3, X_4, X_5
1.1	X_2, X_3, X_4, X_5
1.2	X_1, X_3, X_4, X_5
1.3	X_1, X_2, X_4, X_5
1.4	X_1, X_2, X_3, X_5
1.5	X_1, X_2, X_3, X_4
2	X_6, X_7, X_8, X_5
2.1	X_7, X_8, X_5
2.2	X_6, X_8, X_5
2.3	X_6, X_7, X_5
2.4	X_6, X_7, X_8
3	X_9, X_{10}, X_{11}, X_5
3.1	X_{10}, X_{11}, X_5
3.2	X_9, X_{11}, X_5
3.3	X_9, X_{10}, X_5
3.4	X_9, X_{10}, X_{11}
4	X_{12}, X_{13}, X_5
4.1	X_{13}, X_5
4.2	X_{12}, X_5
4.3	X_{12}, X_{13}

การคัดเลือกตัวแบบ

คัดเลือกตัวแบบโดยใช้เกณฑ์ซีพีที่ปรับแล้วเกณฑ์เอไอซี และเกณฑ์บีไอซี หากตัวแบบใดให้ค่าของเกณฑ์ต่ำที่สุด แสดงว่า ตัวแบบที่พิจารณาสามารถอธิบายตัวแปรตามได้ คัดเลือกตัวแบบโดยใช้สถิติทดสอบซีพีที่ปรับแล้ว โดยคำนวณสถิติทดสอบ $F_{cal} = \frac{C_p^* + (n - 2p)}{n - p}$ ถ้า $F_{cal} > F^*_{\alpha\,(g,\,v_2)}$ เมื่อ $F^*_{\alpha\,(g,\,v_2)}$ เป็นค่าสถิติของการแจงแจกเอฟแบบไร้ศูนย์กลาง แสดงว่า สมการที่พิจารณาไม่สามารถอธิบายตัวแปรตามได้

6. ผลการศึกษา

ผลการวิจัยนำเสนอในรูปแบบของตาราง เพื่อความสะดวกในการนำเสนอ กำหนดสัญลักษณ์เพื่อใช้แทนความหมายต่างๆ ดังนี้

C_p^* หมายถึง เกณฑ์ซีพีที่ปรับแล้ว

AIC หมายถึง เกณฑ์เอไอซี

BIC หมายถึง เกณฑ์บีไอซี

F_{cal} หมายถึง ค่าสถิติทดสอบที่ได้จากการคำนวณ

F_{α} หมายถึง ค่าสถิติเอฟจากการเปิดตารางด้วยองศาอิสระ และ

λ หมายถึง ค่าพารามิเตอร์ไร้ศูนย์กลางของการแจกแจงเอฟแบบไร้ศูนย์กลาง

F^*_{α} หมายถึง ค่าสถิติของการแจกแจงเอฟแบบไร้ศูนย์กลางที่ระดับนัยสำคัญ

ตัวแบบที่ 1 ตัวแบบที่ 2 ตัวแบบที่ 3 และตัวแบบที่ 4 เป็นตัวแบบที่ over fit ตัวแบบที่ 1.1 ตัวแบบที่ 1.2 ตัวแบบที่ 1.3 ตัวแบบที่ 1.4 ตัวแบบที่ 2.1 ตัวแบบที่ 2.2 ตัวแบบที่ 2.3 ตัวแบบที่ 3.1 ตัวแบบที่ 3.2 ตัวแบบที่ 3.3 ตัวแบบที่ 4.1 และตัวแบบที่ 4.2 เป็นตัวแบบที่ misspecified ตัวแบบที่ 1.5 ตัวแบบที่ 2.4 ตัวแบบที่ 3.4 และ ตัวแบบที่ 4.3 เป็นตัวแบบที่ถูกต้อง

จากตารางที่ 2 พบว่า เกณฑ์ซีพีที่ปรับแล้ว เกณฑ์เอไอซีและ เกณฑ์บีไอซี เลือกตัวแบบที่ให้ค่าของเกณฑ์ต่ำสุดเป็นตัวแบบที่ใช้ในการอธิบายตัวแปรตามได้ คือ ตัวแบบที่ 3.4

ตารางที่ 2 ตัวอย่างของการคัดเลือกตัวแบบที่ใช้ได้ในการอธิบายตัวแปรตาม

ตัวแบบที่	C_p^*	AIC	BIC	F_{cal}	$F^*_{\alpha=0.01}$	$F^*_{\alpha=0.05}$	$F^*_{\alpha=0.1}$
1	45.89	612.72	610.06	1.42	1.09	0.94	0.88
1.1	48.52	614.12	611.74	1.46	1.07	0.93	0.86
1.2	567.66	769.88	761.66	6.92	0.23	0.20	0.19
1.3	108.44	650.07	645.61	2.09	0.75	0.66	0.61
1.4	265.35	708.33	701.54	3.74	0.43	0.37	0.34
1.5	46.24	612.46	610.19	1.43	1.09	0.94	0.88
2	66.23	626.15	623.01	1.64	0.95	0.83	0.77
2.1	65.47	624.94	622.48	1.64	0.96	0.83	0.77
2.2	342.04	726.33	720.53	4.52	0.35	0.31	0.29
2.3	881.82	807.14	800.14	10.14	0.16	0.14	0.13
2.4	66.01	625.28	622.81	1.65	0.96	0.83	0.77
3	5.85	577.30	577.64	1.01	**1.52**	**1.32**	**1.23**
3.1	97.62	643.52	640.24	1.98	0.80	0.69	0.64
3.2	149.74	667.80	663.60	2.52	0.63	0.55	0.51
3.3	266.58	707.23	701.86	3.74	0.43	0.37	0.34
3.4	**4.00**	**575.45**	**575.70**	1.00	**1.55**	**1.35**	**1.25**
4	100.15	644.84	641.50	2.00	0.79	0.69	0.64
4.1	641.29	777.04	772.09	7.58	0.21	0.18	0.17
4.2	568.85	766.67	761.83	6.83	0.23	0.20	0.19
4.3	98.28	642.91	640.53	1.98	0.80	0.70	0.65

สถิติทดสอบซีพีที่ปรับแล้ว α = 0.01 α = 0.05 และ α = 0.1 ตัวแปรอิสระที่อยู่ในแต่ละตัวแบบที่ใช้ในการอธิบายตัวแปรตามได้ คือ ตัวแบบที่ 3 และตัวแบบที่ 3.4

ตัวแบบที่ถูกต้องในการจำลอง คือ ตัวแบบที่ 1.5 ตัวแบบที่ 2.4 ตัวแบบที่ 3.4 และตัวแบบที่ 4.3 แต่ความแปรปรวนของความคลาดเคลื่อนในตัวแบบทั้ง 4 ไม่เท่ากัน มีค่าเท่ากับ 202, 252, 182 และ 282 ตามลำดับ ดังนั้น ตัวแบบที่เหมาะสมที่สุดในตัวแบบทั้ง 4 คือ ตัวแบบที่ 3.4 เพราะมีความแปรปรวนต่ำสุด

จากตารางที่ 3 พบว่าการคัดเลือกตัวแบบที่ใช้ในการอธิบายตัวแปรตามได้ โดยใช้เกณฑ์ C_p^* เกณฑ์ AIC และเกณฑ์ BIC เมื่อขนาดตัวอย่างเพิ่มขึ้นถึงระดับหนึ่งแล้ว ร้อยละการคัดเลือกตัวแบบที่ถูกต้องจากเกณฑ์ทั้งสามจะไม่เพิ่มขึ้นอีกต่อไป ผลการคัดเลือกจากเกณฑ์ทั้ง 3 เป็นไปตามที่คาดหวัง คือ คัดเลือกตัวแบบที่ 3.4 ในร้อยละสูงสุด

ตารางที่ 3 ร้อยละการคัดเลือกตัวแบบที่ใช้ได้ในการอธิบายตัวแปรตามโดยใช้เกณฑ์, และ ตามขนาดตัวอย่าง

ตัวแบบที่	$n = 25$			$n = 50$			$n = 100$		
	C_p^*	AIC	BIC	C_p^*	AIC	BIC	C_p^*	AIC	BIC
1	5	7	6	0	1	0	0	0	0
1.1	3	2	2	0	0	0	0	0	0
1.2	0	0	0	0	0	0	0	0	0
1.3	0	0	0	0	0	0	0	0	0
1.4	0	0	0	0	0	0	0	0	0
1.5	4	5	2	1	0	1	1	1	1
2	0	0	0	0	0	0	0	0	0
2.1	1	1	1	1	1	1	0	0	0
2.2	0	0	0	0	0	0	0	0	0
2.3	0	0	0	0	0	0	0	0	0
2.4	3	3	3	1	1	1	0	0	0
3	19	19	13	15	18	9	18	19	19
3.1	0	0	0	0	0	0	0	0	0
3.2	0	0	0	0	0	0	0	0	0
3.3	0	0	0	0	0	0	0	0	0
3.4	63	61	70	82	79	88	81	80	80
4	0	0	0	0	0	0	0	0	0
4.1	0	0	0	0	0	0	0	0	0
4.2	0	0	0	0	0	0	0	0	0
4.3	2	2	3	2	0	0	0	0	0

จากตารางที่ 4 พบว่าการคัดเลือกตัวแบบที่ใช้ในการอธิบายตัวแปรตามได้ โดยใช้สถิติทดสอบซีพีที่ปรับแล้วเมื่อทดสอบสมมุติฐาน H_0: $\Gamma^*_{p_i} = \Gamma^*_{p_{ref}}$, H_1: $\Gamma^*_{p_i} = \Gamma^*_{p_{ref}}$ เพื่อค้นหาตัวแบบอื่นที่อาจใช้แทนตัวแบบที่ 3.4 ซึ่งเป็นตัวแบบที่ถูกต้องได้ ในระดับนัยสำคัญ α = 0.01 α = 0.05 และ α = 0.1 ผลปรากฏว่า เมื่อขนาดตัวอย่างเท่ากับ 25 สถิติทดสอบซีพีที่ปรับแล้วโดยใช้ α = 0.01 α = 0.05 และ α = 0.1 ผลการทดสอบสมมุติฐานคัดเลือกตัวแบบที่ใช้ในการอธิบายตัวแปรตามได้แทนตัวแบบที่ 3.4 คือ ตัวแบบที่ 3 โดยมีตัวแปรอิสระที่ไม่เกี่ยวข้องกับตัวแปรตามเพิ่มขึ้น 1 ตัว ในอัตราที่ใกล้เคียงกับการคัดเลือกตัวแบบที่ถูกต้อง เมื่อขนาดตัวอย่างเท่ากับ 50 และ 100 ผลการทดสอบสมมุติฐานก็คัดเลือกตัวแบบที่ 3 ในอัตราที่เพิ่มขึ้นและอัตราเดียวกันกับตัวแบบที่ถูกต้อง อนึ่งการทดสอบสมมุติฐานก็ยังปฏิเสธสมมุติฐานหลักในกรณีตัวแบบที่ 1.5 ตัวแบบที่ 2.4 และตัวแบบที่ 4.3 เช่นกันผลการทดสอบสมมุติฐานโดยใช้สถิติทดสอบซีพีที่ปรับแล้วสรุปได้ว่าร้อยละการคัดเลือกตัวแบบที่ 1.5 ตัวแบบที่ 2.4 ตัวแบบที่ 3.4 และตัวแบบที่ 4.3 เรียงตามลำดับของค่าความแปรปรวนของความคลาดเคลื่อน และร้อยละของการคัดเลือกตัวแบบที่ 1.5 ตัวแบบที่ 2.4 และตัวแบบที่ 4.3 ตลอดจนร้อยละของการคัดเลือกตัวแบบที่ over fit และ misspecified ก็ลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น และเมื่อระดับนัยสำคัญเพิ่มขึ้นร้อยละการคัดเลือกตัวแบบที่ over fitและ misspecified ก็ลดลงเช่นกัน ทิศทางดังกล่าวเป็นทิศทางที่ควรจะเป็นตามที่ปรากฏในทฤษฎีการวิเคราะห์การถดถอย

ตารางที่ 4 ร้อยละของกลุ่มตัวแบบที่ใช้ได้ในการอธิบายตัวแปรตาม โดยใช้สถิติทดสอบซีพีที่ปรับแล้วตามขนาดตัวอย่าง

ตัวแบบที่	ร้อยละของตัวแบบที่ใช้ในการอธิบายตัวแปรตามได้ $n = 25$			ร้อยละของตัวแบบที่ใช้ในการอธิบายตัวแปรตามได้ $n = 50$			ร้อยละของตัวแบบที่ใช้ในการอธิบายตัวแปรตามได้ $n = 100$		
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
1	46	31	21	23	13	7	8	5	4
1.1	41	21	13	14	7	4	2	1	0
1.2	0	0	0	0	0	0	0	0	0
1.3	12	4	2	1	0	0	0	0	0
1.4	0	0	0	0	0	0	0	0	0
1.5	46	26	20	27	13	9	9	6	3
2	22	10	6	7	3	2	0	0	0
2.1	24	12	6	4	3	2	0	0	0
2.2	0	0	0	0	0	0	0	0	0
2.3	0	0	0	0	0	0	0	0	0
2.4	22	11	7	7	3	3	0	0	0
3	96	92	91	99	99	99	100	99	99
3.1	4	2	1	0	0	0	0	0	0
3.2	1	1	0	0	0	0	0	0	0
3.3	0	0	0	0	0	0	0	0	0
3.4	95	94	91	99	99	99	100	100	99
4	10	4	4	2	0	0	0	0	0
4.1	0	0	0	0	0	0	0	0	0
4.2	0	0	0	0	0	0	0	0	0
4.3	9	6	3	3	0	0	0	0	0

7. สรุปและอภิปราย

ผลการเปรียบเทียบร้อยละการคัดเลือกตัวแบบโดยเกณฑ์ซีพีที่ปรับแล้ว เกณฑ์เอไอซี และเกณฑ์บีไอซี จะพบว่า เมื่อขนาดตัวอย่างเพิ่มขึ้น เกณฑ์ซีพีที่ปรับแล้ว เกณฑ์เอไอซี และเกณฑ์บีไอซี มีร้อยละของการคัดเลือกตัวแบบได้ตัวแบบที่ถูกต้องสูงขึ้น ร้อยละของการคัดเลือกตัวแบบที่ misspecified ลดลง

สถิติทดสอบซีพีที่ปรับแล้ว สามารถคัดเลือกตัวแบบที่ใช้ในการอธิบายตัวแปรตามได้ถูกต้องมากขึ้นเมื่อขนาดตัวอย่างใหญ่พอ ซึ่งสอดคล้องกับสถิติทดสอบซีพี ของ พลากร สีน้อย [1] แต่ทดลองภายใต้สถานการณ์ที่แตกต่างกัน คือ สถิติทดสอบซีพี สร้างขึ้นจากเกณฑ์ซีพีปกติ ที่ตัวแปรอิสระไม่มีความสัมพันธ์กันและไม่มีปัญหาในการคำนวณหาเมทริกซ์ผกผัน และค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของตัวแบบเต็มรูปหาค่าได้

เมื่อพิจารณาความแตกต่างระหว่างเกณฑ์และสถิติทดสอบการคัดเลือกตัวแบบที่ใช้ในการอธิบายตัวแปรตามได้จะเห็นว่าสถิติทดสอบซีพีที่ปรับแล้วให้ผลการคัดเลือกตัวแบบที่ใช้ในการอธิบายตัวแปรตามได้เป็นตัวแบบที่ 3.4 มากกว่าการคัดเลือกตัวแบบด้วยเกณฑ์ซีพีที่ปรับแล้ว เกณฑ์เอไอซีและ เกณฑ์บีไอซีเพราะสถิติทดสอบที่ได้เสนอจะคัดเลือกตัวแบบที่ใช้ในการอธิบายตัวแปรตามได้โดยการทดสอบความมีนัยสำคัญในการคัดเลือกแต่เกณฑ์ซีพี เกณฑ์เอไอซีและ เกณฑ์บีไอซี จะคัดเลือกตัวตัวแบบที่ใช้ในการอธิบายตัวแปรตามได้โดยพิจารณาจากค่าต่ำสุด

8. กิตติกรรมประกาศ

ผู้วิจัยขอขอบพระคุณ รองศาสตราจารย์ ดร. วิชิต หล่อจีระชุณห์กุล ที่ช่วยให้คำแนะนำ และแนวคิดที่เป็นประโยชน์ ตลอดจนช่วยตรวจสอบ แก้ไขให้งานวิจัยชิ้นนี้สมบูรณ์มากยิ่งขึ้น

9. เอกสารอ้างอิง

1. Seenoi, P., 2010, “Test Statistics for Selecting Multiple Linear Regression Models”, *Journal of Burapa University*, vol.15 (2), pp. 47-56. (In Thai)

2. Akaike, H., 1973, *Information Theory and an Extension of the Maximum Likelihood Principle*, In B.N. Petrov and F. Csaki (Eds.), Second International Symposium On Information Theory,pp. 267-281.

3. Anderson T.W., 1957, *An Introduction to Multivariate Statistic Analysis*, New York, John Willey & Sons.

4. Beal, D J., 2007, *Information Criteria Methods in SAS for Multiple Linear Regression Models*, Paper SA05. Retrieved August 1, 2012 from http://analytics.ncsu.edu/sesug/2007/SA05.pdf

5. Graybill, F. A., 1976, *Theory and Application of the Linear Model*, Duxbury Press, North Scituate, Massachusetts.

6. Mallows, C.L., 1973, “Some Comments on C P”, *Journal of Technometrics*, Vol. 15, pp. 661-675.

7. McQuarrie, A. and Tsai, C.L., 1998, *Regression and Time Series Model Selection*, Singapore, World Scientific.

8. Montgomery, D. C., Peck, E. A. and Vining, G. G., 2006, *Introduction to Linear Regression Analysis*, New Jersey.

9. Pesaran, M.H. and Weeks, M., 1999, “Non-nested Hypothesis Testing: An Overview”, *Cambridge Working Papers in Economics 9981 Department of Applied Economics University of Cambridge*.

10. Sawa, T., 1978, “Information Criteria for Discriminating among Alternative Regression Model”, *Journal of Econometrica*, Vol.46, pp. 1273-1282.

11. Tiku, M.L. and Yip, D.Y.N., 1978, “A Four-Moment Approximation Based On The F Distribution”, *Australian Journal of Statistic*, Vol. 20, pp. 257-261.