บทที่ 1

บทนำ

ที่มาและความสำคัญของปัญหา

ในปัจจุบันการพยากรณ์มีบทบาทในการวิจัยสาขาต่าง ๆ โดยส่วนใหญ่การวิจัยจะใช้วิธีการวิเคราะห์ความถดถอย (regression analysis) ซึ่งเป็นวิธีทางสถิติที่ใช้หารูปแบบความสัมพันธ์ระหว่างตัวแปรอิสระ 2 ตัวหรือมากกว่า เพื่อใช้ในการพยากรณ์ค่าจริง

ในการวิเคราะห์ความถดถอยที่มีตัวแปรอิสระตั้งแต่ 2 ตัวแปรขึ้นไปนั้น มีข้อสมมุติเกี่ยวกับความสัมพันธ์ระหว่างตัวแปรอิสระคือ ตัวแปรเหล่านี้จะต้องไม่มีความสัมพันธ์กัน แต่ในการวิจัยด้านเศรษฐมิติบางสถานการณ์ข้อสมมุตินี้อาจใช้ไม่ได้ถ้าตัวแปรเหล่านี้มีความสัมพันธ์ซึ่งกันและกัน และหากความสัมพันธ์ดังกล่าวมีมากการวิเคราะห์ความถดถอยก็อาจใช้ไม่ได้ ดังนั้นการคัดเลือกตัวแปรโดยวิธีของ เบส์เซียน (Bayesian Variable Selection) ซึ่งพิจารณาความเป็นไปได้ที่แต่ละตัวแบบจะเหมาะสมที่สุดในการพยากรณ์จึงเป็นวิธีหนึ่งที่เราอาจจะนำมาหารูปแบบความสัมพันธ์ที่เหมาะสม เพราะวิธีนี้จะไม่คำนึงถึงความสัมพันธ์ใด ๆ ระหว่างตัวแปรอิสระ

ตัวแบบของการถดถอยเชิงเส้นเมื่อพิจารณาเทอมพหุนามแบบลำดับขั้นของตัวแปรอิสระ $X_1, X_2, \ldots, X_p$ เป็นดังนี้

(1) $$y = \mathbf{Z(X)}\beta^{1} + \varepsilon$$

เมื่อ y แทนเวกเตอร์ของตัวแปรตามขนาด $n\times 1$ ที่มีการแจกแจงพหุแบบปกติ ด้วยค่าเฉลี่ย $\mathbf{Z(X)}\beta$ และความแปรปรวน σ^2 เราเขียนได้เป็น $y \sim N_k\left(\boldsymbol{Z}(\boldsymbol{X})\beta, \sigma^2 I_n\right)$

$\boldsymbol{X}$ แทนเมทริกซ์ของตัวแปรอิสระขนาด $n\times p$ ซึ่งเขียนได้เป็น $\boldsymbol{X} = \left(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p\right)$

$\boldsymbol{Z(X)}$ แทนเมทริกซ์ของพจน์พหุนามแบบลำดับขั้นของ $\boldsymbol{X}$ ซึ่งมีขนาด $n\times(k+1)$ และ $\boldsymbol{Z(X)}$ มีค่าลำดับชั้น (rank) เป็น k โดยที่ $k<n$

เมื่อ k เป็นจำนวนพจน์พหุนามแบบลำดับขั้นซึ่ง

k = จำนวนพจน์ของปัจจัยหลัก+จำนวนพจน์อันตรกิริยา+จำนวนพจน์พหุนามกำลังสอง

$$= p + \binom{p}{2} + p$$

$$= 2p + \binom{p}{2}$$

---

[1] อักษรเข้ม จะหมายถึง เวกตอร์

β เป็นเวกเตอร์ของพารามิเตอร์ที่ไม่ทราบค่าขนาด $(k+1)\times 1$ ซึ่งเขียนได้เป็น $\beta = (\beta_0, \beta_1, \ldots, \beta_k)'$

ε เป็นเวกเตอร์ของความคลาดเคลื่อนขนาด $n\times 1$

^{2} เป็นพารามิเตอร์ที่ไม่ทราบค่า

I_n เป็นเมทริกซ์เอกลักษณ์จากข้อมูลจำนวน n ชุด

ในการวิเคราะห์ความถดถอยที่มีตัวแปรอิสระหลายตัวสิ่งที่ต้องให้ความสนเป็นอย่างมากคือการคัดเลือกตัวแปรเพื่อให้ได้ตัวแบบที่เหมาะสมที่สุด โดยวิธีการที่ถูกนำมาใช้มากคือวิธีการถดถอยแบบขั้นบันได (stepwise regression) เมื่อเราพิจารณาพจน์อันตรกิริยา (interaction term) จะมีจำนวนพจน์ที่เข้าสู่ตัวแบบมากขึ้นวิธีการถดถอยแบบขั้นบันไดจะยอมรับเอาตัวแบบที่ง่ายที่สุด (parsimonious model) ซึ่งประกอบด้วยพจน์ที่มีอันดับสูง ๆ (high order term) ในขณะที่ดึงพจน์ที่อันดับต่ำกว่า (lower order term) ออกจากตัวแบบ ซึ่งขัดแย้งกับแนวคิดของตัวแบบที่มีหลักเกณฑ์ดี (Well-formulated model) ซึ่งมีความเชื่อว่าตัวแบบที่ปรากฏพจน์ที่มีอันดับสูง ๆ จะต้องมีพจน์ที่มีอันดับต่ำกว่าทั้งหมดซึ่งสอดคล้องกันอยู่ในตัวแบบด้วย เมื่อมีการแปลงค่าของตัวแปรอิสระจากมาตรวัดหนึ่งไปเป็นอีกมาตรวัดหนึ่ง เช่น การลบค่าของตัวแปรอิสระด้วยค่าเฉลี่ย ซึ่งมีผลกระทบต่อการวิเคราะห์ความถดถอยในกรณีที่มีพจน์พหุนาม (polynomial term) ในตัวแบบ หากตัวแบบนั้นเป็นตัวแบบที่มีหลักเกณฑ์ดีแล้วการแปลงดังกล่าวจะไม่ทำให้ปริภูมิการประมาณ (estimation space) เปลี่ยนแปลง

วัตถุประสงค์ของการวิจัย

เพื่อเปรียบเทียบความถูกต้องของการพยากรณ์จากตัวแบบที่คัดเลือกตัวแปรด้วยวิธีของเบส์เซียน วิธีการกำจัดตัวแปรแบบถอยหลัง และวิธีการถดถอยแบบขั้นบันได ในการวิเคราะห์ความถดถอยเชิงเส้นเมื่อตัวแปรอิสระมีความสัมพันธ์พหุนามแบบลำดับขั้น

ข้อตกลงเบื้องต้น

1. รูปแบบทั่วไปของสมการถดถอยพหุนามแบบลำดับขั้นมีรูปแบบดังสมการ (1)
2. ตัวแปรอิสระแต่ละตัวเป็นค่าคงที่
3. ความคลาดเคลื่อนสุ่มเป็นตัวแปรสุ่มที่มีการแจกแจง $N(0,\sigma^2)$ เหมือนกันและเป็นอิสระซึ่งกันและกัน
4. การประมาณค่าสัมประสิทธิ์ถดถอยเชิงเส้นจะใช้วิธีกำลังสองน้อยสุด (least square error method) ในการประมาณ

นิยามศัพท์

ตัวแปรตาม (dependence variable)

หมายถึง ตัวแปรที่ได้รับผลกระทบจากตัวแปรอื่น ๆ

ตัวแปรอิสระ (independence variable)

หมายถึง ตัวแปรที่มีผลกระทบต่อตัวแปรตาม

ตัวพยากรณ์ (predictors)

หมายถึง ตัวแปรที่มีผลกระทบต่อตัวแปรตามใช้ในการทำนายพฤติกรรมของตัวแปรตาม โดยในงานวิจัยนี้กำหนดให้ตัวพยากรณ์เป็นพจน์พหุนามของตัวแปรอิสระซึ่งมีอันดับสูงสุดไม่เกิน 2 ได้แก่ ปัจจัยหลัก (main effect) ปัจจัยพหุนามกำลังสอง (polynomial effect) และปัจจัยอันตรกิริยา (interaction effect)

สมมติฐานของการวิจัย

เมื่อตัวแปรอิสระมีความสัมพันธ์พหุนามแบบลำดับขั้น วิธีของเบส์เซียนจะให้ค่าพยากรณ์ที่มีความถูกต้องและเชื่อถือได้มากกว่าวิธีการกำจัดตัวแปรแบบถอยหลังและวิธีการถดถอยแบบขั้นบันได ทั้งนี้เพราะวิธีการกำจัดตัวแปรแบบถอยหลังและวิธีการถดถอยแบบขั้นบันไดอาจเกิดความผิดพลาดในขั้นตอนการคัดเลือกตัวแปร เนื่องจากวิธีดังกล่าวพิจารณาความสัมพันธ์ระหว่างกลุ่มของตัวแปรที่อยู่ในตัวแบบกับตัวแปรที่จะเข้าหรือออกจากตัวแบบ ในขณะที่วิธีการของเบส์เซียนไม่ได้พิจารณาความสัมพันธ์ดังกล่าวเลย แต่กลับพิจารณาความน่าจะเป็นที่แต่ละตัวแบบจะเป็นตัวแบบที่เหมาะสมที่สุดในการพยากรณ์

ขอบเขตของการวิจัย

1. ตัวแบบของการถดถอยพหุนามเชิงเส้นที่สนใจศึกษาเป็นดังนี้

1.1 เมื่อจำนวนตัวแปรอิสระเป็น 6 ตัว ตัวแบบอยู่ในรูปของ

$$
\begin{aligned}
y_i = \beta_0^* &+ \beta_{10}(x_{1i} - \bar{x}_1) + \beta_{20}(x_{2i} - \bar{x}_2) + \beta_{30}(x_{3i} - \bar{x}_3) \\
&+ \beta_{40}(x_{4i} - \bar{x}_4) + \beta_{50}(x_{5i} - \bar{x}_5) + \beta_{60}(x_{6i} - \bar{x}_6) \\
&+ \beta_{11}(x_{1i}^2 - c_1) + \beta_{22}(x_{2i}^2 - c_2) + \beta_{33}(x_{3i}^2 - c_3) \\
&+ \beta_{44}(x_{4i}^2 - c_4) + \beta_{55}(x_{5i}^2 - c_5)) + \beta_{66}(x_{6i}^2 - c_6) \\
&+ \beta_{12}(x_{1i}x_{2i} - d_{12}) + \beta_{13}(x_{1i}x_{3i} - d_{13}) + \beta_{14}(x_{1i}x_{4i} - d_{14}) \\
&+ \beta_{15}(x_{1i}x_{5i} - d_{15}) + \beta_{16}(x_{1i}x_{6i} - d_{16}) \\
&+ \beta_{23}(x_{2i}x_{3i} - d_{23}) + \beta_{24}(x_{2i}x_{4i} - d_{24}) + \beta_{25}(x_{2i}x_{5i} - d_{25}) \\
&+ \beta_{26}(x_{2i}x_{6i} - d_{26}) \\
&+ \beta_{34}(x_{3i}x_{4i} - d_{34}) + \beta_{35}(x_{3i}x_{5i} - d_{35}) + \beta_{36}(x_{3i}x_{6i} - d_{36}) \\
&+ \beta_{45}(x_{4i}x_{5i} - d_{45}) + \beta_{46}(x_{4i}x_{6i} - d_{46}) \\
&+ \beta_{56}(x_{5i}x_{6i} - d_{56}) + \varepsilon_i
\end{aligned}
$$

เมื่อ $\beta_0^* = \beta_0 + \sum_j^6 \beta_{j0}\bar{x}_{j.} + \sum_j^6 \beta_{jj}c_j + \sum_j^6\sum_k^6 \beta_{jk}d_{jk}$

โดยที่ $c_j = \frac{\sum_{i=1}^n x_{ji}^2}{n}$, $j = 1,2,....,6$

และ $d_{jk} = \frac{\sum_{i=1}^n x_{ji}x_{ki}}{n}$, $j = 1,2,...,6$, $k = 1,2,...,6$

ในกรณีนี้จำนวนตัวพยากรณ์ 27 ตัว

1.2 เมื่อจำนวนตัวแปรอิสระเป็น 5 ตัว ตัวแบบอยู่ในรูปของ

$$\begin{aligned} y_i = {} & \beta_0^* + \beta_{10}(x_{1i} - \bar{x}_{1.}) + \beta_{20}(x_{2i} - \bar{x}_{2.}) + \beta_{30}(x_{3i} - \bar{x}_{3.}) \\ & + \beta_{40}(x_{4i} - \bar{x}_{4.}) + \beta_{50}(x_{5i} - \bar{x}_{5.}) \\ & + \beta_{11}(x_{1i}^2 - c_1) + \beta_{22}(x_{2i}^2 - c_2) + \beta_{33}(x_{3i}^2 - c_3) \\ & + \beta_{44}(x_{4i}^2 - c_4) + \beta_{55}(x_{5i}^2 - c_5)) \\ & + \beta_{12}(x_{1i}x_{2i} - d_{12}) + \beta_{13}(x_{1i}x_{3i} - d_{13}) + \beta_{14}(x_{1i}x_{4i} - d_{14}) \\ & + \beta_{15}(x_{1i}x_{5i} - d_{15}) \\ & + \beta_{23}(x_{2i}x_{3i} - d_{23}) + \beta_{24}(x_{2i}x_{4i} - d_{24}) + \beta_{25}(x_{2i}x_{5i} - d_{25}) \\ & + \beta_{34}(x_{3i}x_{4i} - d_{34}) + \beta_{35}(x_{3i}x_{5i} - d_{35}) \\ & + \beta_{45}(x_{4i}x_{5i} - d_{45}) + \varepsilon_i \end{aligned}$$

เมื่อ $\beta_0^* = \beta_0 + \sum_j^5 \beta_{j0}\bar{x}_{j.} + \sum_j^5 \beta_{jj}c_j + \sum_j^5\sum_k^5 \beta_{jk}d_{jk}$

โดยที่ $c_j = \frac{\sum_{i=1}^n x_{ji}^2}{n}$, $j = 1,2,....,5$

และ $d_{jk} = \frac{\sum_{i=1}^n x_{ji}x_{ki}}{n}$, $j = 1,2,...,5$, $k = 1,2,...,5$

ในกรณีนี้จำนวนตัวพยากรณ์ 20 ตัว

1.3 เมื่อจำนวนตัวแปรอิสระเป็น 4 ตัว ตัวแบบอยู่ในรูปของ

$$\begin{aligned} y_i = {} & \beta_0^* + \beta_{10}(x_{1i} - \bar{x}_{1.}) + \beta_{20}(x_{2i} - \bar{x}_{2}) + \beta_{30}(x_{3i} - \bar{x}_{3.}) \\ & + \beta_{40}(x_{4i} - \bar{x}_{4.}) \\ & + \beta_{11}(x_{1i}^2 - c_1) + \beta_{22}(x_{2i}^2 - c_2) + \beta_{33}(x_{3i}^2 - c_3) + \beta_{44}(x_{4i}^2 - c_4) \\ & + \beta_{12}(x_{1i}x_{2i} - d_{12}) + \beta_{13}(x_{1i}x_{3i} - d_{13}) + \beta_{14}(x_{1i}x_{4i} - d_{14}) \\ & + \beta_{23}(x_{2i}x_{3i} - d_{23}) + \beta_{24}(x_{2i}x_{4i} - d_{24}) \\ & + \beta_{34}(x_{3i}x_{4i} - d_{34}) + \varepsilon_i \end{aligned}$$

เมื่อ $\beta_0^* = \beta_0 + \sum_j^4 \beta_{j0}\bar{x}_{j\cdot} + \sum_j^4 \beta_{jj}c_j + \sum_j^4\sum_k^4 \beta_{jk}d_{jk}$

โดยที่ $c_j = \frac{\sum_{i=1}^n x_{ji}^2}{n}$, $j = 1,2,....,4$

และ $d_{jk} = \frac{\sum_{i=1}^n x_{ji}x_{ki}}{n}$, $j = 1,2,...,4$, $k = 1,2,...,4$

ในกรณีนี้จำนวนตัวพยากรณ์ 14 ตัว

1.4 เมื่อจำนวนตัวแปรอิสระเป็น 3 ตัว ตัวแบบอยู่ในรูปของ

$$\begin{aligned} y_i = \beta_0^* &+ \beta_{10}(x_{1i} - \bar{x}_{1\cdot}) + \beta_{20}(x_{2i} - \bar{x}_{2\cdot}) + \beta_{30}(x_{3i} - \bar{x}_{3\cdot}) \\ &+ \beta_{11}(x_{1i}^2 - c_1) + \beta_{22}(x_{2i}^2 - c_2) + \beta_{33}(x_{3i}^2 - c_3) \\ &+ \beta_{12}(x_{1i}x_{2i} - d_{12}) + \beta_{13}(x_{1i}x_{3i} - d_{13}) \\ &+ \beta_{23}(x_{2i}x_{3i} - d_{23}) + \varepsilon_i \end{aligned}$$

เมื่อ $\beta_0^* = \beta_0 + \sum_j^3 \beta_{j0}\bar{x}_{j\cdot} + \sum_j^3 \beta_{jj}c_j + \sum_j^3\sum_k^3 \beta_{jk}d_{jk}$

โดยที่ $c_j = \frac{\sum_{i=1}^n x_{ji}^2}{n}$, $j = 1,2,3$

และ $d_{jk} = \frac{\sum_{i=1}^n x_{ji}x_{ki}}{n}$, $j = 1,2,3$, $k = 1,2,3$

ในกรณีนี้จำนวนตัวพยากรณ์ 9 ตัว

1.5 เมื่อจำนวนตัวแปรอิสระเป็น 2 ตัว ตัวแบบอยู่ในรูปของ

$$\begin{aligned} y_i = \beta_0^* &+ \beta_{10}(x_{1i} - \bar{x}_{1\cdot}) + \beta_{20}(x_{2i} - \bar{x}_{2\cdot}) \\ &+ \beta_{11}(x_{1i}^2 - c_1) + \beta_{22}(x_{2i}^2 - c_2) \\ &+ \beta_{12}(x_{1i}x_{2i} - d_{12}) + \varepsilon_i \end{aligned}$$

เมื่อ $\beta_0^* = \beta_0 + \sum_j^2 \beta_{j0}\bar{x}_{j\cdot} + \sum_j^2 \beta_{jj}c_j + \sum_j^2\sum_k^2 \beta_{jk}d_{jk}$

โดยที่ $c_j = \frac{\sum_{i=1}^n x_{ji}^2}{n}$, $j = 1,2$

และ $d_{jk} = \frac{\sum_{i=1}^n x_{ji}x_{ki}}{n}$, $j = 1,2$, $k = 1,2$

ในกรณีนี้จำนวนตัวพยากรณ์ 5 ตัว

1.6 เมื่อจำนวนตัวแปรอิสระเป็น 1 ตัว ตัวแบบอยู่ในรูปของ

$$y_i = \beta_0^* + \beta_{10}(x_{1i} - \bar{x}_1) + \beta_{11}(x_{1i}^2 - c_1) + \varepsilon_i$$

เมื่อ $\beta_0^* = \beta_0 + \beta_{10}\bar{x}_1 + \beta_{11}c_1$

และ $c_1 = \dfrac{\sum_{i=1}^{n} x_{1i}^2}{n}$

ในกรณีนี้จำนวนตัวพยากรณ์ 2 ตัว

2. จำนวนอันดับสูงสุดที่ศึกษาไม่เกิน 2
3. การวิจัยครั้งนี้กำหนดให้ $\beta' = (1, 1, ..., 1)_{1\times(k+1)}$ ในประชากรทุกรูปแบบที่ศึกษา โดยที่ k เป็นจำนวนตัวพยากรณ์ (จำนวนพจน์พหุนามแบบลำดับขั้นในตัวแบบโดยไม่นับพจน์ค่าคงที่)
4. ขนาดตัวอย่าง (n) ที่ศึกษาคือ 25[1] 50 75 และ 100
5. จำนวนตัวแปรอิสระที่ศึกษามี 6 ตัวแปร โดยสร้างจากการแจกแจงปกติที่มีค่าเฉลี่ย 0 และความแปรปรวนเป็น 1
6. ระดับนัยสำคัญของการทดสอบ (α) ที่ศึกษาคือ 0.05 และ 0.01
7. ในการวิจัยครั้งนี้จะศึกษาเมื่อค่าคลาดเคลื่อนสุ่มมีการแจกแจง $N_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$ โดยกำหนดให้ σ = 5 10 20 และ 25

ประโยชน์ของการวิจัย

เพื่อเป็นแนวทางในการเลือกตัวแบบที่เหมาะสมเพื่อใช้ในการพยากรณ์ค่าของตัวแปรตาม เมื่อตัวแปรอิสระมีความสัมพันธ์พหุนามแบบลำดับขั้น

เกณฑ์ที่ใช้ในการตัดสินใจ

เกณฑ์ที่ใช้ในการตัดสินใจว่าค่าพยากรณ์จากตัวแบบใดจะมีความถูกต้องมากที่สุดพิจารณาจาก เกณฑ์ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Avereage Mean Sum Square Error (AMSE)) และเกณฑ์ที่ใช้ประกอบการพิจารณาจะใช้เกณฑ์ค่าอัตราส่วนผลต่างของค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Ratio of Different Average Mean Squares Error (RDAMSE)) มีสูตรดังนี้

---

[1] ในกรณีที่จำนวนตัวแปรอิสระเป็น 6 ตัวแปร ตัวพยากรณ์ที่ใช้ในตัวแบบเริ่มต้นจะมีถึง 27 ตัว ซึ่งส่งผลกระทบต่อค่าระดับขั้นความเสรี ดังนั้นเมื่อจำนวนตัวแปรอิสระเป็น 6 ตัวแปร จะไม่ทำการศึกษาเมื่อขนาดตัวอย่างเป็น 25

$$MSE_i = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p}$$

$$AMSE = \frac{\sum_{j=1}^{200} MSE_j}{200}$$

เมื่อ y_i แทนค่าสังเกตที่ i

$\hat{y}_i$ แทนค่าพยากรณ์ที่ i

p แทนจำนวนของตัวพยากรณ์ในตัวแบบ

n แทนขนาดตัวอย่าง

MSE_j แทนค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการทำซ้ำรอบที่ j

$AMSE$ แทนค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยจากการทำซ้ำ 200 รอบ

$$RDAMSE_i = \frac{(AMSE_i - AMSE_{\min})}{AMSE_{\min}} \times 100\%$$

$AMSE_i$ แทนค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยจากวิธีที่ i

$AMSE_{\min}$ แทนค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยที่มีค่าต่ำสุดจากวิธีทั้ง 3 วิธี

โดยทั้งสองเกณฑ์นั้น วิธีใดที่มีค่าต่ำสุดจะเป็นวิธีที่ดีที่สุด