บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การวิเคราะห์ความถดถอย (regression analysis) มีบทบาทสำคัญมากด้านการพยากรณ์ในงานวิจัยสาขาต่าง ๆ ในปัจจุบัน ซึ่งการวิเคราะห์ความถดถอยเป็นวิธีทางสถิติที่ใช้หารูปแบบความสัมพันธ์ระหว่างตัวแปร 2 ตัว หรือมากกว่า เพื่อนำไปสู่การพยากรณ์ค่าจริงของตัวแปรตาม รูปแบบความสัมพันธ์ของตัวแปรตามและตัวแปรอิสระอาจสัมพันธ์กันในลักษณะเชิงเส้นตรงหรือเส้นโค้ง ในกรณีที่ความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระเป็นแบบเส้นโค้ง วิธีทางสถิติที่จะนำมาหารูปแบบความสัมพันธ์คือ การวิเคราะห์ความถดถอยพหุนาม (polynomial regression analysis) ในทางปฏิบัติวิธีการวิเคราะห์นี้เป็นวิธีที่ใช้ในกรณีที่ความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระมีลักษณะเป็นเส้นโค้ง เนื่องจากการวิเคราะห์ความถดถอยพหุนามเป็นวิธีที่อาจใช้แนวความคิดของวิธีการวิเคราะห์ความถดถอยเชิงเส้นโดยให้แต่ละพจน์ของพหุนามเป็นตัวแปรอิสระตัวหนึ่ง ดังนั้นการวิเคราะห์ความถดถอยพหุนามจึงเป็นกรณีพิเศษของการวิเคราะห์ความถดถอยเชิงเส้น (linear regression analysis)

ในการปฏิบัติงานจริงบางครั้งอาจเกิดความคลาดเคลื่อนในการวัดข้อมูล (measurement error) ขึ้นในตัวแปรอิสระและตัวแปรตาม เช่น ข้อมูลประเภทความสูง น้ำหนัก หรือข้อมูลเกี่ยวกับการทดลองทางด้านเคมี เป็นต้น ซึ่งวิธีการประมาณโดยทั่วไปไม่ได้คำนึงถึงความคลาดเคลื่อนในตัวแปรอิสระจึงเป็นสาเหตุให้ค่าพยากรณ์ที่ได้ไม่ใกล้เคียงกับความจริงดังนั้นผู้วิจัยจึงได้พิจารณากรณีที่เกิดความคลาดเคลื่อนในตัวแปรอิสระด้วย ตัวแบบการถดถอยพหุนามโดยทั่วไปซึ่งขึ้นกับตัวแปรอิสระจะมีรูปแบบดังนี้

(1.1) $$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + \varepsilon_i \qquad ; \; i = 1,2,\ldots,n$$

แต่หากเกิดความคลาดเคลื่อนในตัวแปรอิสระกล่าวคือ

$$x_i = x_i^* + \delta_i \qquad ; \; i = 1,2,\ldots,n$$

เมื่อ x_i^* เป็นตัวแปรแฝงในตัวแปรอิสระ

และ δ_i เป็นความคลาดเคลื่อนในตัวแปรอิสระ

จะได้ว่าเราสามารถเขียนสมการ (1.1) ให้มีความถูกต้องมากขึ้นโดยใช้ตัวแปรแฝง (x_i^*) ในตัวแปรอิสระ ได้ดังนี้

(1.2) $$y_i^* = \beta_0 + \beta_1 x_i^* + \beta_2 x_i^{*2} + \ldots + \beta_p x_i^{*p} + \varepsilon_i^{*}\ ^{1} \qquad ; i = 1,2,\ldots,n$$

หรือ

$$y_i^* = \beta_0 + \beta_1 (x_i - \delta_i) + \beta_2 (x_i - \delta_i)^2 + \ldots + \beta_p (x_i - \delta_i)^p + \varepsilon_i^*$$

จากสมการ (1.1) และ (1.2) จะได้ว่า y_i^* และ y_i มีความสัมพันธ์กันในรูปแบบดังนี้

$$y_i^* = y_i - \beta_1 \delta_i - \beta_2 \delta_i (2x_i - \delta_i) + \ldots + (-1)^p \beta_p \delta_i^p$$

เราสามารถเขียนสมการ (1.2) ในรูปของเมทริกซ์ได้ดังสมการ (1.3)

(1.3) $$\underset{\sim}{y}^* = X^* \underset{\sim}{\beta} + \underset{\sim}{\varepsilon}^*$$

เมื่อ $\underset{\sim}{y}^*$ เป็นเวกเตอร์ตัวแปรตามขนาด $n \times 1$ ที่ขึ้นอยู่กับตัวแปรแฝง X^* โดยที่ $\underset{\sim}{y}^*$ มีการแจกแจงแบบปกติหลายตัวแปร ด้วยค่าเฉลี่ย $X^* \underset{\sim}{\beta}$ และความแปรปรวน σ_ε^2 เขียนได้เป็น $\underset{\sim}{y}^* \sim N_n(X^* \underset{\sim}{\beta}, \sigma_\varepsilon^2 . I_n)$

X^* เป็นเมทริกซ์ของตัวแปรแฝงในตัวแปรอิสระ X (latent variable in X) ขนาด $n \times (p+1)$ และเขียนได้เป็น $X^* = [\underset{\sim 1}{x}^{*\prime}, \underset{\sim 2}{x}^{*\prime}, \ldots, \underset{\sim n}{x}^{*\prime}]'$ โดยที่ $\underset{\sim i}{x}^{*\prime} = (1, x_i^*, x_i^{*2}, \ldots, x_i^{*p})$ หรือ $\underset{\sim i}{x}^{*\prime} = [1, (x_i - \delta_i), (x_i - \delta_i)^2, \ldots, (x_i - \delta_i)^p]$ ซึ่ง x_i เป็นตัวแปรอิสระและ δ_i เป็นความคลาดเคลื่อนในตัวแปรอิสระ

$\underset{\sim}{\beta}$ เป็นเวกเตอร์ของพารามิเตอร์ที่ไม่ทราบค่าขนาด $(p+1)$ เขียนได้เป็น $\underset{\sim}{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$

$\underset{\sim}{\varepsilon}^*$ เป็นเวกเตอร์ความคลาดเคลื่อนสุ่มในตัวแปรตามขนาด $n \times 1$ โดยที่ $\varepsilon_i^* \sim N(0, \sigma_\varepsilon^2 .)$

$\sigma_\varepsilon^2 .$ เป็นพารามิเตอร์ที่ไม่ทราบค่า

I_n เป็นเมทริกซ์เอกลักษณ์จากข้อมูลจำนวน n ชุด

และ p เป็นจำนวนพจน์พหุนาม

---

[1] ในกรณีนี้ $\varepsilon_i = \varepsilon_i^*$ สาเหตุที่เปลี่ยนสัญลักษณ์เพื่อให้เกิดความสอดคล้องของเรื่องสัญลักษณ์ในสมการ (1.2)

ในการวิเคราะห์ความถดถอยที่มีตัวแปรอิสระหลายตัว สิ่งสำคัญเป็นอย่างมากคือ การสร้างตัวแบบให้มีเหมาะสมเพื่อให้ได้ค่าพยากรณ์ที่ใกล้เคียงกับค่าจริงมากที่สุด ซึ่งตัวแบบที่เหมาะสมนั้นควรสร้างมาจากวิธีการประมาณค่าพารามิเตอร์ที่เหมาะสม ผู้วิจัยจึงได้ทำการศึกษาวิธีการประมาณค่าพารามิเตอร์ที่เหมาะสมจากงานวิจัยต่าง ๆ ดังนี้

นพมาศ อัครจันทโชติ (พ.ศ. 2539) ได้ศึกษาเกี่ยวกับการเปรียบเทียบวิธีการสร้างตัวแบบในการวิเคราะห์ความถดถอยพหุนามกรณีที่มี 2 ตัวแปรอิสระ ซึ่งเกิดอันตรกิริยา (interaction) โดยทำการเปรียบเทียบวิธีการสร้างตัวแบบ 4 วิธี ได้แก่ วิธีการสร้างตัวแบบด้วยวิธีกำลังสองน้อยสุดสามัญ (Ordinary Least Squares Method (OLS)) การสร้างตัวแบบด้วยวิธีกำจัดตัวแปรอิสระย้อนหลัง (Backward Elimination Method (BE)) การสร้างตัวแบบด้วยวิธีการถดถอยขั้นบันได (Stepwise Regression Method (SW)) และการสร้างตัวแบบด้วยวิธีตัวแบบหลักเกณฑ์ดี (Well-Formulated Method (WF))

แชง และ ชนิวไวส์ (Chang and Schneewiess (ค.ศ. 1998)) ได้ทำการศึกษาเปรียบเทียบความเอนเอียงในการประมาณค่าพารามิเตอร์ในตัวแบบถดถอยพหุนามกรณีที่มีความคลาดเคลื่อนเกิดขึ้นในตัวแปรอิสระ โดยทำการเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ 2 วิธี ได้แก่ วิธี OLS และวิธีกำลังสองน้อยสุดปรับปรุง (Adjusted Least Squares Method (ALS)) ในกรณีที่ตัวแปรแฝง (x_i^*) มีการ แจกแจงแบบ $U(-1,1)$ และกำลังสูงสุดของตัวแปรอิสระที่ใช้สำหรับสร้างตัวแปรตามในตัวแบบถดถอยพหุนาม (MB) เป็น 2

ฮวง และ ฮูวาง (Huang and Huwang (ค.ศ. 2001)) ได้ทำการศึกษาเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ในตัวแบบถดถอยพหุนามกรณีที่มีความคลาดเคลื่อนเกิดขึ้นในตัวแปรอิสระ โดยทำการเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ 2 วิธี ได้แก่ วิธี OLS และวิธีกำลังสองน้อยสุดถ่วงน้ำหนัก (Weighted Least Squares Method (WLS)) ในกรณีที่กำลังสูงสุดของตัวแปรอิสระที่ใช้สำหรับสร้างตัวแปรตามในตัวแบบถดถอยพหุนาม (MB) เป็น 2

จากงานวิจัยดังกล่าวผู้วิจัยจึงได้ทำการเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ 3 วิธี ได้แก่ OLS ALS และ WLS ในกรณีที่กำลังสูงสุดของตัวแปรอิสระที่ใช้สำหรับสร้างตัวแปรตามในตัวแบบถดถอยพหุนาม (MB) เป็น 2, 3, 4, 5 และ 6 และตัวแปรแฝง (x_i^*) มีการแจกแจงตามที่ได้กล่าวไว้ข้างต้น

1.2 วัตถุประสงค์การวิจัย

เพื่อเปรียบเทียบความถูกต้องของค่าพยากรณ์ของตัวแบบที่ได้จากวิธีการประมาณค่าพารามิเตอร์ที่เหมาะสม วิธีต่าง ๆ ดังนี้

1. การสร้างตัวประมาณพารามิเตอร์ โดยวิธีกำลังสองน้อยสุดสามัญ (Ordinary Least Squares Method (OLS))
2. การสร้างตัวประมาณพารามิเตอร์ โดยวิธีกำลังสองน้อยสุดปรับปรุง (Adjusted Least Squares Method (ALS))
3. การสร้างตัวประมาณพารามิเตอร์ โดยวิธีกำลังสองน้อยสุดถ่วงน้ำหนัก (Weighted Least Squares Method (WLS))

1.3 ข้อตกลงเบื้องต้น

1. รูปแบบทั่วไปของสมการถดถอยพหุนามมีรูปแบบดังสมการที่ (1.1)
2. ความคลาดเคลื่อนสุ่มในตัวแปรตาม และความคลาดเคลื่อนในตัวแปรอิสระ เป็นตัวแปรสุ่มที่มีการ

แจกแจงเหมือนกันและเป็นอิสระกันแบบ $N(0,\sigma^2_{\varepsilon^*})$ และ $N(0,\ \sigma^2_{\delta})$ ตามลำดับ

3. ตัวแปรแฝง(ตัวแปรอิสระที่ปราศจากความคลาดเคลื่อน) แต่ละตัวเป็นค่าคงที่

1.4 สมมติฐานการวิจัย

การประมาณค่าพารามิเตอร์ของตัวแบบการถดถอยพหุนามจะมีความถูกต้องในการพยากรณ์ตัวแปรตาม เมื่อใช้วิธีการประมาณค่าพารามิเตอร์โดยวิธีกำลังสองน้อยสุดถ่วงน้ำหนัก (Weighted Least Squares Method (WLS))

1.5 ขอบเขตของการวิจัย

1. ตัวแบบของการถดถอยพหุนามที่สนใจศึกษาเป็นดังนี้

   1.1 ตัวแบบที่กำลังสูงสุดของตัวแปรแฝงเป็น 6 มีรูปแบบดังนี้

$$y_i^* = \beta_0 + \beta_1 x_i^* + \beta_2 x_i^{*2} + \beta_3 x_i^{*3} + \beta_4 x_i^{*4} + \beta_5 x_i^{*5} + \beta_6 x_i^{*6} + \varepsilon_i^*$$

หรือ

$$y_i^* = \beta_0 + \beta_1 (x_i - \delta_i) + \beta_2 (x_i - \delta_i)^2 + \beta_3 (x_i - \delta_i)^3 + \beta_4 (x_i - \delta_i)^4 + \beta_5 (x_i - \delta_i)^5 + \beta_6 (x_i - \delta_i)^6 + \varepsilon_i^* \quad ; \ i = 1,\dots,n$$

1.2 ตัวแบบที่กำลังสูงสุดของตัวแปรแฝงเป็น 5 มีรูปแบบดังนี้

$$y_i^* = \beta_0 + \beta_1 x_i^* + \beta_2 x_i^{*2} + \beta_3 x_i^{*3} + \beta_4 x_i^{*4} + \beta_5 x_i^{*5} + \varepsilon_i^*$$

หรือ $$y_i^* = \beta_0 + \beta_1(x_i - \delta_i) + \beta_2(x_i - \delta_i)^2 + \beta_3(x_i - \delta_i)^3 + \beta_4(x_i - \delta_i)^4 + \beta_5(x_i - \delta_i)^5 + \varepsilon_i^* \quad ; \ i = 1,\ldots,n$$

1.3 ตัวแบบที่กำลังสูงสุดของตัวแปรแฝงเป็น 4 มีรูปแบบดังนี้

$$y_i^* = \beta_0 + \beta_1 x_i^* + \beta_2 x_i^{*2} + \beta_3 x_i^{*3} + \beta_4 x_i^{*4} + \varepsilon_i^*$$

หรือ $$y_i^* = \beta_0 + \beta_1(x_i - \delta_i) + \beta_2(x_i - \delta_i)^2 + \beta_3(x_i - \delta_i)^3 + \beta_4(x_i - \delta_i)^4 + \varepsilon_i^* \quad ; \ i = 1,\ldots,n$$

1.4 ตัวแบบที่กำลังสูงสุดของตัวแปรแฝงเป็น 3 มีรูปแบบดังนี้

$$y_i^* = \beta_0 + \beta_1 x_i^* + \beta_2 x_i^{*2} + \beta_3 x_i^{*3} + \varepsilon_i^*$$

หรือ $$y_i^* = \beta_0 + \beta_1(x_i - \delta_i) + \beta_2(x_i - \delta_i)^2 + \beta_3(x_i - \delta_i)^3 + \varepsilon_i^* \quad ; \ i = 1,\ldots,n$$

1.5 ตัวแบบที่กำลังสูงสุดของตัวแปรแฝงเป็น 2 มีรูปแบบดังนี้

$$y_i^* = \beta_0 + \beta_1 x_i^* + \beta_2 x_i^{*2} + \varepsilon_i^*$$

หรือ $$y_i^* = \beta_0 + \beta_1(x_i - \delta_i) + \beta_2(x_i - \delta_i)^2 + \varepsilon_i^* \quad ; \ i = 1,\ldots,n$$

2. การวิจัยครั้งนี้กำหนดให้ $\underset{\sim}{\beta}' = (1 \ 1 \ \ldots \ 1)_{1\times(p+1)}$ ในประชากรทุกรูปแบบที่ศึกษา โดยที่ p เป็นจำนวนตัวพยากรณ์ (จำนวนพจน์พหุนามในตัวแบบโดยไม่นับพจน์ค่าคงที่)

3. ขนาดตัวอย่าง (n) ที่ศึกษาคือ 15, 30, 50, 100 และ 200

4. จำนวนตัวแปรอิสระเริ่มต้นที่ศึกษาคือ 1 ตัว สร้างจากตัวแปรแฝง (x^*) และความคลาดเคลื่อนในตัวแปรอิสระ (δ) โดยที่ตัวแปรแฝงเริ่มต้นสร้างจาก $x_i^* \sim N(10,1)$ [1] และ

---

[1] เมื่อกำลังของตัวแปรอิสระมีค่าสูง (>6) และค่า x มีค่ามากจะมีผลทำให้ค่าของ x ที่ยกกำลังต่าง ๆ ในเมทริกซ์ X มีค่าเพิ่มสูงขึ้น ดังนั้นการหาค่าผกผันของเมทริกซ์ $X'X$ จะเกิดความคลาดเคลื่อนสูง ทำให้ตัวประมาณของพารามิเตอร์ที่ได้มีความคลาดเคลื่อนสูงด้วย นอกจากนั้นกรณีที่ $x \in (0,1)$ จะมีผลทำให้ค่าของ x ที่ยกกำลังต่าง ๆ ในเมทริกซ์ X มีค่าลดต่ำลง ซึ่งทำให้เกิดปัญหาเช่นเดียวกับกรณีข้างต้น (จาก Seber,G.A.F(1977).Linear Regression Analysis.New York.)

$\delta_i \sim N(0,\sigma_{\delta}^2)$ เนื่องจากต้องการศึกษาอิทธิพลของ σ_{δ} จึงกำหนดให้ $\sigma_{\delta} = 0.1, 0.3, 0.5$ และ 0.7 ซึ่งคิดเป็น 10%, 30%, 50% และ 70% [2] ของ σ_x ที่เท่ากับ 1 ตามลำดับ

5. ในการวิจัยครั้งนี้จะศึกษาเมื่อความคลาดเคลื่อนสุ่มในตัวแปรตามมีการแจกแจง $N(0,\sigma_{\varepsilon}^2)$ โดยกำหนด $\sigma_{\varepsilon} = 0.1, 0.3, 0.5, 0.7$ และ 1.0

6. ตัวแปรตาม (y^*) ที่ใช้สำหรับคำนวณค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยสัมพัทธ์ สร้างจากตัวแปรแฝง (x^*) และความคลาดเคลื่อน (ε^*) โดยที่ตัวแปรตาม (y^*) มีการแจกแจงแบบ $N_n(X^* \underset{\sim}{\beta}, \sigma_{\varepsilon}^2.I_n)$

7. ข้อมูลที่ใช้ในการศึกษาวิจัยครั้งนี้ใช้โปรแกรมคอมพิวเตอร์จำลองด้วยเทคนิคมอนติคาร์โล กระทำซ้ำ 1000 รอบ ในแต่ละสถานการณ์

1.6 ประโยชน์ของการวิจัย

เพื่อเป็นแนวทางในการประมาณค่าพารามิเตอร์ของตัวแบบถดถอยพหุนามกรณีที่เกิดความคลาดเคลื่อนในตัวแปรอิสระที่เหมาะสมในการนำไปใช้พยากรณ์ค่าตัวแปรตาม

1.7 เกณฑ์การตัดสินใจ

เกณฑ์การตัดสินใจว่าตัวแบบใดมีความถูกต้องมากที่สุด จะพิจารณาจากเกณฑ์ค่าเฉลี่ยรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยสัมพัทธ์ [3] (Average Relative Root Mean Squares Error ($ARRMSE$)) และเกณฑ์ที่ใช้ในการประกอบการตัดสินใจจะใช้เกณฑ์ค่าอัตราส่วนผลต่างของรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยสัมพัทธ์ (Ratio of Different Average Relative Root Mean Squares Error ($RDARRMSE$)) มีสูตรดังนี้

$$RRMSE = \frac{\sqrt{\dfrac{\sum_{i=1}^{n}\left(y_i^* - \hat{y}_i\right)^2}{n}}}{\bar{y}^*} \times 100\%$$

$$ARRMSE = \frac{\sum_{j=1}^{1000} RRMSE_j}{1000}$$

---

[2] ค่าของ σ_{δ} ไม่ควรมีค่าเป็น 100% ของ σ_x จาก G.S.Maddala(1977). Econometrics. Florida.

[3] จากเอกสารคำสอนวิชาเทคนิคการพยากรณ์ ของผู้ช่วยศาสตราจารย์ ร้อยเอกมานพ วราภักดิ์

เมื่อ y_i^* แทนค่าสังเกตที่ i

$\hat{y}_i$ แทนค่าพยากรณ์ที่ i

p แทนจำนวนของตัวพยากรณ์ในตัวแบบ

n แทนขนาดตัวอย่าง

และ $RRMSE_j$ แทนรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยสัมพัทธ์ของการทำซ้ำรอบที่ j

$ARRMSE$ แทนค่าเฉลี่ยรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยสัมพัทธ์จากการทำซ้ำ 1000 รอบ

$$RDRRAMSE = \frac{(ARRMSE_i - ARRMSE_{\min})}{ARRMSE_{\min}} \times 100\%$$

เมื่อ $ARRMSE_i$ แทนค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ยจากตัวแบบที่ i

และ $ARRMSE_{\min}$ แทนค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ยที่มีค่าต่ำสุดจากตัวแบบที่ได้จาก 3 วิธี

ตัวแบบใดให้ค่าแต่ละเกณฑ์ต่ำสุดจะเป็นตัวแบบที่ดีที่สุด