Collinearity

อุทัยวรรณ สายพัฒนะ*
ฉัตรศิริ ปิยะพิมลสิทธิ์**

ในการวิเคราะห์การถดถอยพหุคูณนั้น เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรเกณฑ์ตัวหนึ่งกับตัวแปรพยากรณ์หลาย ๆ ตัว ซึ่งการวิเคราะห์นี้มีข้อตกลงข้อหนึ่งว่าตัวแปรพยากรณ์เหล่านี้ต้องไม่มีความสัมพันธ์กัน หรือหากสัมพันธ์กันก็ต้องมีความสัมพันธ์กันไม่สูงมากนัก แต่ในทางปฏิบัติบางครั้งมักพบว่าตัวแปรพยากรณ์มีความสัมพันธ์กันสูง ในกรณีที่ตัวแปรพยากรณ์เพียง 2 ตัว มีความสัมพันธ์กันสูง เรียกว่า *Collinearity* และในกรณีที่ตัวแปรพยากรณ์มากกว่า 2 ตัว มีความสัมพันธ์กันสูง เรียกว่า *Multicollinearity*

การที่ตัวแปรพยากรณ์ มีความสัมพันธ์กันเองสูงจะทำให้ผลการวิเคราะห์การถดถอยผิดปกติไปดังนี้

1. **ค่าความคลาดเคลื่อนมาตรฐานของสัมประสิทธิ์การถดถอย (S_b) มีค่าสูงผิดปกติ**

ถ้าค่า S_b มีค่าสูงผิดปกติ เมื่อนำค่า S_b ไปใช้ในการทดสอบทางสถิติ

$$t = \frac{b_i}{S_{b_i}}$$

ค่าสถิติทดสอบ t ที่ได้จะมีค่าต่ำกว่าปกติ ทำให้การแปลผลผิดไปจากความเป็นจริง

ตัวอย่าง เช่น ต้องการศึกษาว่า X_1 มีความสัมพันธ์กับ Y หรือไม่ เมื่อค่า S_b สูงมาก ค่า t ที่ได้จะต่ำกว่าปกติ ทำให้ยอมรับ H_0 แสดงว่า X_1 ไม่มีความสัมพันธ์กับ Y ซึ่งความจริง X_1 มีความสัมพันธ์กับ Y สูงมาก

2. **ค่าสัมประสิทธิ์การถดถอยไม่คงที่ มีการเปลี่ยนแปลงเมื่อตัวแปรพยากรณ์เปลี่ยน**

จากสมการถดถอยนี้

$$Y' = a + b_1X_1 + b_2X_2$$

ถ้า X_1 และ X_2 มีความสัมพันธ์กันสูง ค่าสัมประสิทธิ์การถดถอยอาจไม่แสดงอิทธิพลที่แท้จริงของ X_1 หรือ X_2 บน Y ค่าสัมประสิทธิ์การถดถอยนี้จึงเชื่อมั่นไม่ได้ และไม่สามารถใช้ในการประมาณค่าการเปลี่ยนแปลงของ Y เมื่อตัวแปรพยากรณ์ X เปลี่ยนแปลงไป 1 หน่วย

อาจารย์ โรงเรียนธรรมศาสตร์คลองหลวงวิทยาคม จ.ปทุมธานี
อาจารย์ประจำภาควิชาการประเมินผลและวิจัย คณะศึกษาศาสตร์ มหาวิทยาลัยทักษิณ
เว็ปมาสเตอร์วัดผลจุดคอม http://www.watpon.com

3. เครื่องหมายของค่าสัมประสิทธิ์การถดถอยตรงกันข้ามกับที่ควรจะเป็น

จากสมการถดถอยนี้

$$Y' = a + b_1X_1 + b_2X_2$$

ถ้าตัวแปรพยากรณ์ X_1 กับ X_2 มีความสัมพันธ์กันสูง อาจทำให้เครื่องหมายของ b_1 หรือ b_2 เป็นตรงข้ามกับที่ควรจะเป็น เช่น b_1 เป็นลบ หรือ b_2 เป็นลบ เมื่อแทนค่า X_1 หรือ X_2 ตามลำดับ เข้าในสมการจะทำให้ได้ค่า Y ผิดไปจากความเป็นจริง

วิธีการตรวจสอบ Collinearity

ตัวอย่างต่อไปนี้ใช้ประกอบการอธิบายวิธีการตรวจสอบ Collinearity ซึ่งมีหลายวิธีการ

ตัวอย่าง ตัวแปรเกณฑ์คือ อัตราดอกเบี้ยที่ได้รับต่อปี (DIVI) ตัวแปรพยากรณ์มี 2 ตัวคือ อัตราของรายได้ต่อปี (EARN) และจำนวนปี (TREND)

ข้อมูลต่อไปนี้แสดงการวิเคราะห์การถดถอยของตัวแปรเกณฑ์บนตัวแปรพยากรณ์

DIVI	EARN	TREND
2.80	4.13	1
3.16	4.63	2
3.40	5.17	3
3.70	5.80	4
4.10	6.21	5
4.50	6.83	6
5.00	7.35	7
5.00	7.91	8

สหสัมพันธ์ของตัวแปรทั้ง 3 คือ

Correlations:	DIVI (Y)	EARN (X_1)	TREND (X_2)
DIVI (Y)	1.0000	.9911**	.9923**
EARN (X1)	.9911**	1.0000	.9996**
TREND (X2)	.9923**	.9996**	1.0000

วิธีการในการตรวจสอบการเกิด collinearity สามารถพิจารณาตรวจสอบได้ดังนี้

1. การทดสอบนัยสำคัญของค่าสหสัมพันธ์ของตัวแปรในโมเดล

จากตัวอย่าง ตัวแปรพยากรณ์ 2 ตัวคือ อัตราของรายได้ต่อปี (EARN) และจำนวนปี (TREND) มีค่าสหสัมพันธ์ $r_{12} = 0.9996$, $n = 8$ ซึ่งสหสัมพันธ์ของตัวแปรทั้งสองนี้สูงมาก แต่จะเกิด collinearity หรือไม่ ต้องทำการทดสอบความมีนัยสำคัญของค่าสหสัมพันธ์ของตัวแปรทั้งสอง

สมมติฐานการทดสอบ

$H_0 : \rho_{12} = 0$

$H_1 : \rho_{12} \neq 0$

เมื่อ ρ_{12} คือความสัมพันธ์ระหว่างตัวแปรพยากรณ์ X_1 และ X_2 ของกลุ่มประชากร จากนั้นใช้สูตร t-test เพื่อทดสอบดังนี้

$$\frac{r_{12}}{s_r}$$

เมื่อ

$$S \sqrt{\frac{1-r_{12}^2}{n-2}}$$

แทนค่าข้อมูลจากตัวอย่างจะได้ว่า

$$S = \sqrt{\frac{1-(0.9996)^2}{8-2}}$$

$$0.0115$$

$$\frac{0.9996}{0.0115}$$

$$86.92$$

ที่ระดับ α = .05, df = n - 2 = 6, $t_{.05,6}$ = 2.447 ค่า t ที่ได้จากการคำนวณ (86.92) มากกว่าค่า t ที่ได้จากตาราง (2.447) นั่นคือปฏิเสธ H_0 ยอมรับ H_1 นั่นคือตัวแปรพยากรณ์ X_1 และ X_2 มีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 นั่นแสดงว่าเกิด collinearity

ในการทดสอบนัยสำคัญของสหสัมพันธ์นี้ มีตำราบางเล่มเสนอแนะว่าเป็นวิธีการที่ยุ่งยาก ให้ใช้วิธีง่าย ๆ คือ ใช้เกณฑ์ค่าสัมบูรณ์ของสหสัมพันธ์ที่มากกว่า 0.80 คือ $|r| > 0.80$ ถือว่ามีความสัมพันธ์กัน แสดงว่าเกิด collinearity ซึ่งเกณฑ์นี้สามารถยืดหยุ่นได้ขึ้นอยู่กับดุลยพินิจของผู้วิจัยด้วย

2. การเปรียบเทียบสัมประสิทธิ์การตัดสินใจ (coefficient of determination)

เป็นวิธีการเปรียบเทียบสัมประสิทธิ์การตัดสินใจระหว่างตัวแปรเกณฑ์กับตัวแปรพยากรณ์แต่ละตัว โดยพิจารณาค่ากำลังสองของสหสัมพันธ์ระหว่างตัวแปร Y กับ X_1 มีค่า r^2 = 0.9823 และค่ากำลังสองของสหสัมพันธ์ระหว่างตัวแปร Y กับ X_2 มีค่า r^2 = 0.9846 เมื่อพิจารณาการส่งผลร่วมกันของตัวแปรพยากรณ์ 2 ตัวต่อตัวแปรตามแล้ว R^2 = .98547 ถ้าแยกส่วนแล้ว ตัวแปรพยากรณ์สองตัวสามารถอธิบายการเปลี่ยนแปลงในตัวแปรเกณฑ์ได้ 98.23% และ 98.46% ตามลำดับ แต่หากรวมทั้งสองตัวแปรแล้วสามารถอธิบายการเปลี่ยนแปลงในตัวแปรเกณฑ์ได้ 98.547% เท่านั้น ปรากฏการณ์นี้แสดงว่าเกิดการซ้อนทับ (overlap) กันของอำนาจในการอธิบาย นั่นคือเดิมตัวแปร X_1 อธิบายตัวแปร Y ได้มากถึง 98.23% เมื่อเพิ่มตัวแปร X_2 แล้วสามารถอธิบายได้เพิ่มขึ้นอีกเพียงเล็กน้อย แสดงว่าเกิด collinearity ขึ้นแล้ว

3. องค์ประกอบความแปรปรวนที่สูงเกินความเป็นจริง (Variance inflation factor : VIF)

Variance inflation factor หรือ VIF เป็นความสัมพันธ์ของตัวแปร X ตัวหนึ่งโดยการถดถอยบนตัวแปร X อื่นๆ

จากตัวอย่างข้างต้น คำนวณหาค่าความสัมพันธ์ของการถดถอยของตัวแปร X_1 บนตัวแปร X_2 ผลที่ได้จะเท่ากับค่าความสัมพันธ์ของการถดถอยตัวแปร X_2 บนตัวแปร X_1 เท่ากับค่าสหสัมพันธ์ r_{12} = 0.9996 คำนวณ VIF ของตัวแปร X_i จากสูตร

$$VIF(X_i) \qquad \frac{1}{1-R_i^2}$$

เมื่อ R_i^2 คือสัมประสิทธิ์ของการตัดสินใจ โดยการถดถอย X_i บนตัวแปรอิสระอื่นๆ ที่เหลือ เราสามารถคำนวณ VIF ของตัวแปร X_i ได้ว่า

$$VIF(X_1) \qquad 1-(0.9996)^2$$

$$1250.31$$

ค่า VIF ของ X_1 เท่ากับ VIF ของ X_2 ถ้าตัวแปรพยากรณ์ทั้งหมดไม่สัมพันธ์กัน ค่า VIF จะมีค่าเป็น 1 และพิสัยของค่า VIF อยู่ในช่วง 1 ถึงอนันต์

เกณฑ์ในการพิจารณา VIF นั้น ขึ้นอยู่กับดุลยพินิจของผู้วิจัยอีกเช่นกัน แต่มีตำราบางเล่มเสนอแนะว่า ตัวแปรพยากรณ์ทั้งสองตัวจะเกิดปัญหา collinearity ก็ต่อเมื่อ ค่า VIF มีค่าตั้งแต่ 10 ขึ้นไป

4. วิธีพิจารณาจากค่าการยอมรับ (Tolerance)

ค่าการยอมรับ (tolerance) สามารถคำนวณได้ด้วยสูตร

$$\text{Tolerance} \qquad 1 - R^2 = \frac{1}{VIF}$$

จากค่า VIF ข้างต้นได้ค่า Tolerance = .0007998 หรืออาจใช้โปรแกรมคอมพิวเตอร์ พิจารณาจากค่า tolerance ของตัวแปร ซึ่งค่า tolerance มีค่าตั้งแต่ 0 ถึง 1 ถ้าค่า tolerance เข้าใกล้ 1 แสดงว่าตัวแปรเป็นอิสระจากกัน แต่ถ้าค่า tolerance เข้าใกล้ 0 แสดงว่าเกิดปัญหา collinearity

พิจารณาตัวอย่างผลลัพธ์ที่ได้จากคอมพิวเตอร์

Variable	SE Beta	Correl	Part Cor	Partial	Tolerance	VIF
TREND	1.860104	.992333	.055953	.420980	8.4017E-04	1190.236
EARN	1.860104	.991128	-.027201	-.220092	8.4017E-04	1190.236

Tolerance มีค่า 0.0008401 มีค่าเข้าใกล้ 0 มาก ส่วนค่า VIF มีค่า 1190.236 สังเกตว่าค่าทั้งสองเป็นไปตามเกณฑ์ของการเกิด collinearity

หมายเหตุ : ค่าที่คำนวณด้วยมือไม่เท่ากับค่าที่คำนวณด้วยคอมพิวเตอร์ ทั้งนี้เนื่องมาจากการปัดเศษทศนิยม

5. วิธีพิจารณาค่า F และค่า t

ในการวิเคราะห์การถดถอยนั้น หากค่า F ที่ใช้ทดสอบ R^2 สูงอย่างมีนัยสำคัญทางสถิติ แต่ค่า t ที่ใช้ทดสอบ b แต่ละตัวแปรกลับมีค่าน้อยจนไม่มีนัยสำคัญทางสถิติ แสดงว่าเกิด collinearity

R Square .98547

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	2	4.77078	2.38539
Residual	5	.07037	.01407

F = 169.49998 Signif F = .0000

-------Variables in the Equation-------

Variable	B	SE B	Beta	T	Sig T
TREND	.655380	.631520	1.930383	1.038	.3469
EARN	-.588601	1.166675	-.938444	-.505	.6353
(Constant)	4.542104	4.164805		1.091	.3252

จากผลการวิเคราะห์ในตัวอย่างนี้ ตัวแปรสองตัวร่วมกันทำนาย Y ได้ค่า F จากการทดสอบมีค่าสูงถึง 169.49998 ; $p < .000$ แต่เมื่อทดสอบสัมประสิทธิ์การถดถอยของตัวแปรพยากรณ์แต่ละตัวแล้ว ปรากฏว่าค่า t มีค่าน้อยมากจนไม่มีนัยสำคัญทางสถิติ ซึ่งแสดงให้เห็นว่าตัวแปรทั้งสองตัวมีความสัมพันธ์กันสูงมาก นั่นคือเกิดปัญหา collinearity

6. วิธีพิจารณาจากค่าไอเกน (eigenvalues)

ค่าไอเกน (eigenvalue) คือ ค่าที่ใช้วัดความสำคัญของตัวแปรพยากรณ์ทั้งหมดในเชิงเปรียบเทียบ ดังนั้น ถ้าค่าไอเกนของตัวแปรบางตัวสูงกว่าค่าไอเกนของตัวแปรอื่น ๆ ก็ชี้ให้เห็นว่าตัวแปรมีความสัมพันธ์ร่วมกันมาก ภายใต้ความเป็นอิสระของตัวแปรที่มีความสำคัญเท่า ๆ กัน ค่าไอเกนของตัวแปรเหล่านั้นจะมีการกระจายเป็นโค้งปกติ

แต่ถ้าเกิด collinearity จะเป็นการกระจายของค่าไอเกนที่ไม่เท่าเทียมกัน นั่นคือค่าไอเกนของตัวแปรบางตัวสูงกับค่าไอเกนของตัวแปรอื่น ๆ ต่ำ ซึ่งสะท้อนให้เห็นความสำคัญของตัวแปรที่ไม่เท่าเทียมกัน

ดังตัวอย่างผลลัพธ์ที่ได้จากการคำนวณด้วยคอมพิวเตอร์

Number	Eigenval	Cond Index	Variance Proportions Constant	EARN	TREND
1	2.89046	1.000	.00001	.00000	.00002
2	.10952	5.137	.00042	.00000	.00086
3	.00002	359.497	.99957	.99999	.99912

7. พิจารณาจาก Condition Indices และ Variance-Decomposition Proportions

ดัชนี 2 ตัวที่ใช้ในการพิจารณาการเกิด collinearity ได้ก็คือ condition number (CN) และ condition index (CI) มีสูตรดังนี้

$$CN = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

$$CI_i = \sqrt{\frac{\lambda_{max}}{\lambda_i}}$$

เมื่อ λ_{max} = ค่าไอเกนที่มากที่สุด (Largest eigenvalue) ; λ_{min} = ค่าไอเกนที่น้อยที่สุด (smallest eigenvalue) และ λ_i = ค่าไอเกนตัวที่ i

จากผลที่ได้จากการวิเคราะห์ด้วยโปรแกรมคอมพิวเตอร์ในข้อ 6. ค่าไอเกนที่มากที่สุดคือ 2.89046 สามารถคำนวณหา CI ได้ดังนี้

$$CI_1 = \sqrt{\frac{\lambda_{max}}{\lambda_1}} = \sqrt{\frac{2.89046}{2.89046}} = 1.00$$

$$CI_2 = \sqrt{\frac{\lambda_{max}}{\lambda_2}} \qquad \sqrt{\frac{2.89046}{0.10952}} = 5.13732$$

หา CN ได้ดังนี้

$$CN = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} \qquad \sqrt{\frac{2.89046}{0.00002}} = 380.1618$$

นอกจากนี้ยังต้องพิจารณาจาก Variance-Decomposition Proportions ซึ่งประกอบไปด้วยส่วนของจุดตัดและตัวแปรพยากรณ์แต่ละตัว

Variance proportions ก็คือสัดส่วนของความแปรปรวนของจุดตัด (a) และสัมประสิทธิ์ การถดถอย (b) แต่ละตัวที่สัมพันธ์กับ Condition index แต่ละตัว ซึ่งในแต่ละสดมภ์มีผลรวมเป็น 1.00 การแปลความหมายต้องนำ 100 ไปคูณ ดังเช่นในตัวแปร EARN แปลผลได้ว่า 0% ของความแปรปรวนใน b_{EARN} สัมพันธ์กับ Condition index ตัวแรก และ 0% สัมพันธ์กับ Condition index ตัวที่สอง และ 99.99% สัมพันธ์กับ Condition index ตัวที่สาม ส่วนสัมประสิทธิ์ b_{TREND} ก็แปลความได้เช่นเดียวกัน

การพิจารณา collinearity ให้พิจารณาที่ Condition index ที่มีค่าสูงสุดกับ variance proportion ที่มีค่าสูงในตัวแปรที่สัมพันธ์กัน

ในกรณีนี้มีตัวแปร 2 ตัวที่สัมพันธ์กัน Condition index ที่มีค่าสูงสุดอยู่ในบรรทัดสุดท้าย จากนั้นพิจารณาที่ variance proportion จะเห็นว่ามีค่าสูงสุดเหมือนกันในตัวแปร 2 ตัวที่สัมพันธ์

8. วิธีพิจารณาความคลาดเคลื่อนมาตรฐานของสัมประสิทธิ์การถดถอย (S_b)

ถ้าความสัมพันธ์ของตัวแปรพยากรณ์มีค่าสูงมากจะมีผลกระทบต่อความคลาดเคลื่อนมาตรฐานของสัมประสิทธิ์การถดถอย

ความคลาดเคลื่อนมาตรฐานของ b_1 คำนวณด้วยสูตร

$$S_{b_{y1.2}} = \sqrt{\frac{S^2_{y1.2}}{\sum X_1^2 (1-r_{12}^2)}}$$

จากสมการนี้จะเห็นได้ว่าถ้า r_1 มีค่าสูงแล้ว จะส่งผลให้ความคลาดเคลื่อนมาตรฐานของสัมประสิทธิ์การถดถอยมีค่าสูงขึ้น

9. วิธีพิจารณาจากเครื่องหมายของสัมประสิทธิ์การถดถอย (b)

9.1 เครื่องหมายของสัมประสิทธิ์การถดถอยเปลี่ยนแปลงเป็นตรงกันข้ามกับที่ควรจะเป็นเมื่อมีการเพิ่มตัวแปรเข้าหรือขจัดตัวแปรออก

9.2 เครื่องหมายของสัมประสิทธิ์การถดถอยเปลี่ยนแปลงเป็นตรงกันข้ามกับที่ควรจะเป็นเมื่อขนาดของกลุ่มตัวอย่างหรือข้อมูลเปลี่ยนแปลงไปเพียงเล็กน้อย

การแก้ไขปัญหา Collinearity

ในการวิเคราะห์การถดถอยนั้น หากเกิดปัญหา collinearity จะทำให้สมการพยากรณ์หรือสมการการถดถอยไม่ถูกต้อง ผู้วิจัยจึงควรหาทางแก้ไขปัญหาดังกล่าว ซึ่งอาจทำได้ดังนี้

1. การขจัดตัวแปรออก หากตัวแปรพยากรณ์ 2 ตัวมีความสัมพันธ์กันสูง ผู้วิจัยควรพิจารณาตัดตัวแปรใดตัวแปรหนึ่งออกจากสมการถดถอย ทั้งนี้เนื่องจากความสัมพันธ์ระหว่างสองตัวมีค่าสูงมาก ก็คือตัวแปรตัวหนึ่งสามารถใช้ทดแทนตัวแปรอีกตัวหนึ่งได้ แต่จะทดแทนได้อย่างสมบูรณ์เมื่อค่าสหสัมพันธ์มีค่าเป็น 1.00

2. การเปลี่ยนแปลงข้อมูลโดยอาจเพิ่มจำนวนข้อมูล จนแน่ใจว่าไม่ทำให้เกิดปัญหา collinearity

3. โคเฮน (Cohen. 1983) เสนอให้ใช้วิธี Centering ในการลดขนาดของความสัมพันธ์ระหว่างตัวแปรพยากรณ์ ซึ่งจะช่วยให้ผลการวิเคราะห์ถูกต้องมากยิ่งขึ้น แต่การทำ Centering อาจไม่ได้ผลในการแก้ปัญหา collinearity ในกรณีที่ข้อมูลในตัวแปรอิสระมีการแจกแจงเบ้ (skewed) และมีจำนวนตัวแปรอิสระหลายตัว

สรุป

Collinearity เป็นปรากฏการณ์ที่เกิดขึ้นเมื่อตัวแปรพยากรณ์ตัวหนึ่งมีความสัมพันธ์กับตัวแปรพยากรณ์อีกตัวหนึ่งสูงมาก ซึ่งจะส่งผลให้ค่าสัมประสิทธิ์การถดถอยเชื่อถือไม่ได้ เครื่องหมายของสัมประสิทธิ์การถดถอยอาจแปรเปลี่ยนไปเป็นตรงกันข้าม และค่าความคลาดเคลื่อนมาตรฐานของการถดถอยมีค่าสูงขึ้นผิดปกติ

วิธีการค้นหาว่าตัวแปรอิสระต่าง ๆ ที่ศึกษานั้นเกิดปัญหา collinearity หรือไม่ มีหลากหลายวิธี แต่เมื่อพบว่า ชุดของตัวแปรที่ศึกษาเกิด collinearity ขึ้นแล้วมีวิธีแก้ไข กระทำได้โดยการเลือกขจัดตัวแปรที่สัมพันธ์กันสูงออกตัวหนึ่ง หรืออาจเปลี่ยนแปลงข้อมูลโดยการเพิ่มจำนวนข้อมูลจนมั่นใจว่าจะไม่เกิด collinearity หรืออาจทำ Centering เพื่อช่วยแก้ปัญหา

บรรณานุกรม

Cohen, Jacob and Cohen, Patricia. ***Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.*** Second Edition. U.S.A. : La Wrence Erlbaum Associates, Publishers, 1983.

Multicollinearity and Singularity.
http://www.pfc.forestry.ca/landscape/inventory/wulder/mvstats/multicol.html

Multiple Regression and Correlation.
http://news.pepperdine.edu/gsbm/class/ohall/statcity/stats14.html

Pedhazur, Elazar J. ***Multiple Regression in Behavioral Research.*** second edtion. U.S.A. Holt, Rinehart and Winston, Inc., 1973.

Siegel, Andrew F. ***Practical Business Statistics.*** second edition. University of Washington, 1994.

Webster, Allen L. ***Applied Statistics for Business and Economics.*** U.S.A. McGraw-Hill Companies, Inc. 1995.

Wittink, Dick R. ***The Application of Regression Analysis.*** U.S.A. : Allyn and Bacon, Inc. 1988.