บทที่ 6

การวิเคราะห์สมการถดถอยพหุคูณ

(Multiple Regression Analysis)

วัตถุประสงค์ : เพื่อศึกษาถึงลักษณะของปัญหาต่าง ๆ ในสมการถดถอยพหุคูณและแนวทางแก้ไข ได้แก่ปัญหา Autocorrelation, Heteroscedasticity, Multicollinearity และการตรวจสอบอำนาจการพยากรณ์ โดยนักศึกษาจะได้ศึกษาจากตัวอย่าง ตลอดจนการอ่านผลการวิเคราะห์

บทที่ 6

การวิเคราะห์สมการถดถอยพหุคูณ

(Multiple Regression Analysis)

บทนำ

เมื่อนักศึกษาสร้างแบบจำลองที่มีตัวแปรอิสระตั้งแต่สองตัวขึ้นไปแล้วใช้วิธีการทางเศรษฐมิติเพื่อหาค่าพารามิเตอร์ต่าง ๆ นั้น นักศึกษาจะใช้วิธีการวิเคราะห์ความถดถอยพหุคูณในรูปสมการเส้นตรง (Multiple Linear Regression Analysis) โดยมีรูปแบบของสมการถดถอย ดังนี้

$$Y = \alpha+\beta_1X_1+\beta_2X_2+\beta_3X_3+\ldots+\beta_kX_k+u$$

โดยที่ Y = ค่าของตัวแปรตาม

$X_1X_2X_3\ldots X_k$ = ค่าของตัวแปรอิสระมีค่าตั้งแต่ 1,2,…,k ตัว

α = ค่าคงที่หรือค่า Intercept ของสมการถดถอย

β_1,β_2,β_k = ค่าพารามิเตอร์หรือค่าสัมประสิทธิ์การถดถอยของตัวแปรอิสระตัวที่ 1,2,…,k ตามลำดับ

u = ค่าความคลาดเคลื่อน (Error or Residual)

ข้อกำหนดของการวิเคราะห์ความถดถอยพหุคูณ

ในการวิเคราะห์ความถดถอยพหุคูณนั้น เพื่อเป็นการควบคุมความผิดพลาดที่อาจจะเกิดขึ้นจากตัวแปรทั้งหลาย ที่จะส่งผลถึงค่าประมาณของตัวแปรตามและการพยากรณ์ จึงมีข้อกำหนดในการวิเคราะห์ดังนี้

สมมติฐาน ข้อที่ 1 Linear Regression Model

แบบจำลองสมการถดถอยจะต้องเป็น linear in the parameters ซึ่งแสดงดังสมการ

$$Y_I = \beta_1 + \beta_2X_i + u_i$$

หมายความว่า β_1 และ β_2 จะต้องเป็น linear function ของ Y_I

สมมติฐาน ข้อที่ 2 ตัวแปร X เป็น nonstochastic (X-FIXED) คือ กำหนดให้ X คงที่ตลอดในแต่ละตัวอย่างที่ทำการสุ่ม ถึงแม้ว่า Y จะเปลี่ยนแปลง

ตัวอย่าง X-FIXED

Y	X
100	10
200	20
300	30
400	40

Y	X
120	10
140	20
160	30
180	40

สมมติฐาน ข้อที่ 3 ค่าเฉลี่ยของตัวรบกวน u_I เป็นศูนย์ กำหนดให้ค่าเฉลี่ย (MEAN) หรือค่าคาดหวัง (Expected value) ของตัวรบกวน u_I มีค่าเป็นศูนย์ เขียนเป็นสมการได้ดังนี้

$$E(u_i) = 0$$

สมมติฐานนี้ หมายถึง การกำหนดตัวแบบสมการของประชากร (Population Line) เท่ากับ $E(Y_I) = \beta_1 + \beta_2 X_i$

ซึ่งหมายความว่า มีความสัมพันธ์เชิงเส้นตรง “โดยเฉลี่ย” ระหว่าง Y และ X ในกรณีที่ค่าเฉลี่ยของตัวรบกวนไม่เท่ากับศูนย์ คือ อาจมีค่ามากกว่าศูนย์ : $E(u_i) > 0$ หรือน้อยกว่าศูนย์ : $E(u_i) < 0$

$$E(Y_i) = \beta_1 + \beta_2 X_i + E(u_i)$$

ค่า $E(u_I)$ อาจมีค่าคงที่หรือมีค่าผันแปรไปตามค่า X_I ก็ได้ในกรณีที่กำหนดให้มีค่าคงที่ สมการของประชากรจะเป็น

$$E(Y_i) = \beta_1 + \beta_2 X_i + k \quad ; \text{ k เป็นค่าคงที่}$$

กรณีที่ $E(u_i) > 0$ หมายความว่า ค่าสังเกตของ Y_i และ X_i จะอยู่เหนือเส้น True Line หรือ Population Line ดังนั้นการใช้กลุ่มตัวอย่างของค่าสังเกตเหล่านี้ ไปคำนวณหา True Line จะได้เส้นจำนวนที่มีลักษณะ Biased ซึ่งมีผลทำให้ตัวคำนวณจะเป็น Biased

กรณีที่ $E(u_i) < 0$

เส้นถดถอยที่คำนวณได้จะอยู่ต่ำกว่าเส้น True Line

สมมติฐาน ข้อที่ 4 ค่าความแปรปรวนของตัวรบกวนมีค่าคงที่ (Homoscedasticity)

หมายความว่า ณ แต่ละค่าของตัวแปรอิสระ X_i ค่าความแปรปรวนของตัวรบกวน (u_i) จะมีค่าคงที่ในทุกค่าสังเกต การที่ค่าความแปรปรวนคงที่ หมายความว่าค่าของมันจะไม่ผันแปรไปตามค่าของ X_i เขียนเป็นสมการได้ดังนี้

$$Var(u_i) = E[u_i - E(u_i)]^2$$
$$= E(u_i^2)$$
$$= \delta^2$$

กรณีที่ค่าความแปรปรวนไม่คงที่เรียกว่า Heteroscedasticity เขียนเป็นสมการได้

$$Var(u_i) \neq \delta^2$$

หรือ $Var(u_i) = \delta_i^2$

สมมติฐาน ข้อที่ 5 ไม่มีความสัมพันธ์ระหว่างตัวรบกวนในแต่ละช่วงเวลา (No Autocorrelation between the disturbances) สมมติฐานที่กำหนดว่า ค่าตัวรบกวน (disturbance term) ที่เกิดขึ้นในแต่ละช่วงเวลาเป็นอิสระซึ่งกันและกัน สามารถเขียนได้ว่า

$$COV(u_i, u_j) = E[u_i - E(u_i)][u_j - E(u_j)]$$
$$= E(u_i u_i)$$
$$= 0$$

โดยที่ i และ j เป็นค่าสังเกต ณ ช่วงเวลาที่แตกต่างกัน และค่า COV หมายถึงค่าความแปรปรวนร่วม (Covariance) ถ้าสมมติฐานนี้ไม่เป็นจริง กล่าวคือค่าตัวรบกวนมีความสัมพันธ์กัน ค่า $Cov(u_i, u_j)$ จะไม่เป็นศูนย์ เช่น อาจมีความสัมพันธ์กันในเชิงเส้นตรง เช่น $u_t = u_{t-1} + v_t$

สมมติฐาน ข้อที่ 6 ไม่มีความสัมพันธ์ระหว่างค่าตัวรบกวนและตัวแปรอิสระ

กำหนดให้ตัวแปรอิสระ X_i และตัวรบกวน u_i ไม่มีความสัมพันธ์กันเขียนได้ดังนี้

$$COV(u_i, X_j) = E[u_i - E(u_i)][X_i - E(X_i)]$$
$$= E[u_i][X_i - E(X_i)] \qquad ; E(u_i)=0$$
$$= E(u_i X_i) - E(u_i)(X_i) \quad ; E(X_i) \text{ เป็น Nonstochastic}$$
$$= E(u_i X_i) \qquad ; E(u_i)=0$$
$$= 0$$

สมมติฐานนี้กำหนดขึ้นมา เพื่อแยกอิทธิพลของตัวแปรอิสระที่มีต่อตัวแปรตามออกจากอิทธิพลของตัวรบกวน ทั้งนี้เนื่องจากค่าตัวรบกวนซึ่งอาจจะรวมเอาอิทธิพลของตัวแปรอิสระที่ละไว้ ซึ่งถ้าหากค่าตัวรบกวนมีความสัมพันธ์กับตัวแปรอิสระที่กำหนดในตัวแบบ ก็จะไม่สามารถแยกอิทธิพล ของตัวแปรอิสระแต่ละตัวออกจากกันได้

สมมติฐาน ข้อที่ 7 จำนวนของค่าสังเกต n (Observations n) จะต้องมากกว่าจำนวนตัวแปรที่จะประมาณค่า จากสมการ

$$Y_i = \beta_1 + \beta_2 X_i + U_I$$

จะพบว่ามีตัวแปรที่เราต้องประมาณค่า จำนวน 2 ตัว คือ β_1 และ β_2 กำหนดว่าเรามีจำนวน Observations ของ Y และ X เพียงค่าเดียวเราจะไม่สามารถประมาณค่า β_1 และ β_2 ได้เลย เราต้องมีจำนวน Observations อย่างน้อย 2 คู่ ถึงจะทำการประมาณค่า β_1 และ β_2 ได้

สมมติฐาน ข้อที่ 8 ค่าสังเกต ของ X_i ในตัวอย่างที่สุ่มมาจะต้องมีค่าไม่เท่ากัน Var(X) จะต้องเป็นบวก เช่น

$$Vax(X) = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

สมมติฐานข้อนี้กำหนดขึ้น เพื่อแสดงให้เห็นว่า ถ้าเกิดกรณีที่ค่าตัวแปรอิสระ X_i มีค่าเท่ากันทุกค่า เราจะไม่สามารถคำนวณหาค่าพารามิเตอร์ได้ แม้ว่าในทางปฏิบัติสมมติฐานข้อนี้มักจะเป็นจริง เพราะค่าสังเกต X_I ที่สุ่มมาโอกาสที่ทุกค่าจะเท่ากันจะมีน้อยมาก แต่อย่างไรก็ตามสมมติฐานนี้ให้ข้อเตือนใจว่า ถ้าหากจำนวนค่าสังเกตของ X_i มีค่าไม่แตกต่างกันมากนัก จะมีผลทำให้ การวัดผลกระทบของ X_i ที่มีต่อ Y ไม่ดีเท่าที่ควร

สมมติฐาน ข้อที่ 9 สมการหรือแบบจำลองที่สร้างขึ้นจะต้องถูกต้อง และไม่มีข้อผิดพลาดคำถามที่สำคัญในการสร้างแบบจำลอง คือ

1. ตัวแปรอะไรบ้างที่ควรจะนำมาใส่ในแบบจำลอง
2. แบบจำลองควรมีรูปแบบสมการแบบใด
3. สมมติฐานของ Y_i,X_i และ u_i มีอะไรบ้างที่จะนำมาใช้ในแบบจำลอง

สมมติฐาน ข้อที่ 10 ตัวแปรอิสระจะต้องไม่มีความสัมพันธ์เชิงเส้นตรงอย่างสมบูรณ์ หมายความว่า ตัวแปรอิสระที่อยู่ในสมการทุกตัว จะต้องไม่มีความสัมพันธ์กันแบบเชิงเส้นอย่างสมบูรณ์ เช่น $X_1 = 3X_2$, $X_2 = 4X_3$

การประมาณค่า

ในการประมาณค่า ß ของสมการถดถอยพหุคุณนั้น มีวิธีการประมาณ 2 วิธี คือ

1. การประมาณค่าด้วยวิธี OLS (Ordinary Least Square)
2. การประมาณค่าด้วยวิธี MLE (Maximum Likelihood Estimator)

ทั้งสองวิธีนี้ถึงแม้ว่า จะมีวิธีการดำเนินการประมาณค่าที่แตกต่างกันแต่สุดท้ายแล้วจะได้ผลลัพธ์ออกมาเท่ากันคือ จะได้ค่า $\hat{\beta}$ ที่เท่ากัน โดยค่า ß ต่าง ๆ จะหาได้จาก

$$\hat{\beta} \quad = \quad (X'X)^{-1}X'Y$$

แต่ทั้งสองวิธีนี้จะมีข้อแตกต่างกันเล็กน้อยคือค่า δ^2 ที่หาได้จากวิธี MLE จะมีค่าเท่ากับ $\sum \frac{u_i^2}{n}$ ซึ่งเป็นค่าประมาณที่ biased ส่วนค่า δ^2 ที่หาได้จากวิธี OLS จะเท่ากับ $\sum \frac{u_i^2}{n-2}$ จะเป็นค่าที่ unbias แต่อย่างไรก็ตาม ถ้าค่า n มีค่ามากขึ้นมาก ๆ จะทำให้ค่า δ^2 ที่ได้จากการคำนวณทั้ง 2 วิธี มีค่าเกือบเท่ากัน หรือใกล้เคียงกันนั่นเอง

การตรวจสอบปัญหาต่าง ๆ ตามข้อกำหนด

1. การตรวจสอบปัญหา Normality

วิธีที่ใช้ในการตรวจสอบข้อกำหนดว่า u_i มีการแจกแจงแบบปกติหรือไม่นั้น สามารถทำได้หลายวิธี เช่น

1) วิธีการวัดค่า skewness และ kurtosis จะเป็นการเปรียบเทียบแบบคร่าว ๆ โดยปกติแล้วถ้าตัวแปรสุ่มใดมีการแจกแจงแบบปกติจะมีค่าความเบ้ (skewness) = 0 แสดงว่าเป็นโค้งสมมาตร ถ้ามากกว่า 0 แสดงว่าโค้งเบ้ขวา และถ้าน้อยกว่า 0 แสดงว่าโค้งเบ้ซ้าย และมีค่าความโด่ง (kurtosis) = 3 ถือว่าเป็นโค้งที่มีความลาดชันเป็นปกติถ้ามากกว่า3แสดงว่าโค้งมีความโด่ง โด่งกว่าโค้งปกติ และถ้าน้อยกว่า 3 แสดงว่าเป็นโค้งที่มีความโด่งแบนราบกว่าโค้งปกติ

2) วิธี Jarque – Bera test (JB) โดยมีสูตรที่ใช้ในการคำนวณ คือ

$$\text{JB} \quad = \quad n\left[\frac{s^2}{6} + \frac{(K-3)^2}{24}\right]$$

โดย JB = ค่าสถิติ JB

n	=	จำนวนตัวอย่าง
S	=	ค่าสถิติ skewness
K	=	ค่าสถิติ kurtosis

เมื่อทำการคำนวณค่า JB แล้วให้ทำการทดสอบสมมติฐานโดยตั้งสมมติฐานหลัก และนำค่าสถิติที่คำนวณได้ไปเปรียบเทียบกับค่าตารางของตัวสถิติ

2. การตรวจสอบปัญหา autocorrelation หรือ serial correlation

Autocorrelation คือเหตุการณ์ที่ค่าความคลาดเคลื่อน u มีความสัมพันธ์ต่อกันหรือมี covariation ระหว่าง u_i กับ u_j หรือ $E(u_i,u_j)$ ไม่เท่ากับ 0 เมื่อ $i \neq j$

ปัญหา Autocorrelation นี้จะทำให้ระดับนัยสำคัญทางสถิติมีค่าสูงกว่าที่ควรจะเป็น ค่า R^2 และค่า F statistic จะไม่น่าเชื่อถือในความถูกต้อง ปัญหานี้อาจเกิดจากการที่ตัวแปรทางเศรษฐศาสตร์มีลักษณะเป็นแนวโน้ม หรือวัฏจักร หรือเกิดจากการที่ไม่รวมตัวแปรอิสระบางตัวไว้ในแบบจำลอง

1) วิธีการตรวจสอบปัญหา Autocorrelation นั้นสามารถตรวจสอบได้หลายวิธีแต่วิธีที่ง่ายและนิยมกันมากคือ พิจารณาจากค่า Durbin – Watson

โดยค่า Durbin – Watson หรือ D_w หรือ d หาได้จาก

$$d = \frac{\sum (u_t - u_{t-1})^2}{\sum u_t^2}$$

โดยทำการเปรียบเทียบค่า d ที่คำนวณได้กับค่า d_L และ d_U ที่เปิดจากตาราง

ถ้า $d < d_L$ แสดงว่า เกิดปัญหา Autocorrelation ทาง บวก

$d > 4 - d_L$ แสดงว่า เกิดปัญหา Autocorrelation ทาง ลบ

$d_U < d < 4 - d_U$ แสดงว่า ไม่เกิดปัญหา Autocorrelation

ถ้าค่า d อยู่นอกนั้นยังไม่สามารถสรุปได้

2) การแก้ปัญหา Autocorrelation อาจทำได้โดยการเพิ่มตัวแปรที่มีเวลาในช่วงก่อนเข้าไปในแบบจำลอง โดยอาจจะเพิ่มในลักษณะเหลื่อมหนึ่งช่วงเวลา หรือ ลักษณะเหลื่อมสองช่วงเวลา หรือรูปแบบอื่นๆ ขึ้นอยู่กับลักษณะของแนวโน้มว่า จะเป็นอย่างไร

3. การตรวจสอบปัญหา Heteroscedasticity

ปัญหา Heteroscedasticity คือการที่ข้อมูลเกิดปัญหาการขัดแย้งกับข้อกำหนดที่ว่า Variance ของ u_i จะต้องมีค่าคงที่ทุกค่าของ i= 1,2,…..,n ปัญหาดังกล่าวนี้มักจะเกิดขึ้นในกรณีที่ใช้ข้อมูลภาคตัดขวาง (cross – section data) ซึ่งมีผลทำให้การใช้วิธี OLS ในการประมาณค่าขาดลักษณะที่พึงประสงค์ของตัวประมาณค่าที่ดี

1) ใช้ Goldfeld – Quandt test ในการทดสอบปัญหา Heteroscedasticity

Goldfeld – Quandt test โดยมีขั้นตอนการทดสอบดังนี้ คือ

- ทำการจัดเรียงค่าสังเกตของตัวแปรทุกตัวให้มีค่าจากน้อยไปหามาก หรือจากมากไปน้อยก็ได้ โดยยึดตัวแปรอิสระในสมการที่มีความสำคัญมากที่สุด อาจพิจารณาจากค่า Standardized coefficients (Beta) ที่มีค่ามากที่สุด
- ทำการแบ่งค่าสังเกตออกเป็น 3 ส่วน และตัดส่วนกลางทิ้งไปจำนวน c ตัวอย่าง โดยให้ส่วนที่เหลือ หัว – ท้าย มีจำนวนข้อมูลเหลืออยู่เท่ากับ (n –c) /2
- ทำการวิเคราะห์สมการถดถอยด้วยวิธี OLS จากค่าตัวอย่างทั้งสองกลุ่ม แล้วทำการหาค่า residual sums of squares RSS_1 กับ RSS_2 โดย RSS_1 คือค่าจากชุดที่ X มีค่าน้อย และ RSS_2 คือชุดที่ค่า X มีค่ามาก
- ทำการคำนวณหาค่า $F^* = RSS_2/RSS_1$
- ทำการทดสอบสมมติฐาน $H_o : \delta_1^2 = \delta_2^2$ ณ ระดับนัยสำคัญ α

ถ้า $F^* > F_{\{(n-c-2k)/2,(n-c-2k)/2,\alpha\}}$ แสดงว่าเราปฏิเสธ H_o หมายความว่าเกิดปัญหา Heteroscedasticity ขึ้นแล้ว แต่ถ้า $F^* < F_{\{(n-c-2k)/2,(n-c-2k)/2,\alpha\}}$ แสดงว่าเราไม่สามารถปฏิเสธ H_o หมายความว่าไม่เกิดปัญหา Heteroscedasticity แต่อย่างไร

2) วิธีแก้ปัญหา Heteroscedasticity

การแก้ปัญหาดังกล่าว ทำได้โดยการแปลงตัวแปรตามและตัวแปรอิสระ ให้เป็นอัตราส่วนด้วยการนำค่าของตัวแปรอิสระตัวที่ก่อให้เกิดปัญหา Heteroscedasticity ไปหารตัวแปรตามและตัวแปรอิสระทุกตัว หรือใช้วิธีประมาณสมการถดถอยแบบถ่วงน้ำหนัก

4. การตรวจสอบปัญหา multicollinearity

ปัญหา multicollinearity คือปัญหาที่ตัวแปรอิสระมีความสัมพันธ์กันเองเกินกว่าระดับที่ยอมรับได้ เพราะถ้าตัวแปรอิสระมีความสัมพันธ์กันเองมากเกินไปจะละเมิดสมมติฐานข้อที่ว่า ตัวแปรอิสระต้องไม่มีความสัมพันธ์เชิงเส้นต่อกัน การที่เกิดปัญหานี้ทำให้ค่าพารามิเตอร์ของตัวแปรอิสระเกิดการผิดพลาด และไม่มีนัยสำคัญทางสถิติ ทำให้ข้อสรุปเกี่ยวกับ marginal effect ของตัวแปรตามที่เกิดจากตัวแปรอิสระตัวนั้น ๆ เกิดความผิดพลาดได้

1) วิธีที่ใช้ในการตรวจสอบ multicollinearity

วิธีที่ใช้ในการตรวจสอบปัญหา multicollinearity นั้นมีวิธีตรวจสอบง่าย ๆ อาจจะพิจารณาจากค่า partial correlation coefficient ระหว่างตัวแปรอิสระด้วยกัน ถ้าค่า Pearson Correlation มีค่ามากกว่า 0.7 ก็อาจจะถือว่า ตัวแปรอิสระคู่นั้นๆ มีความสัมพันธ์กันมาก ควรได้รับการแก้ไข

2) ในการแก้ไขปัญหา multicollinearity อาจจะแก้ได้โดย

- ทำการเพิ่มขนาดของจำนวนตัวอย่าง หรือชุดข้อมูล ให้มากขึ้น อาจจะสามารถแก้ปัญหาได้
- ละทิ้งตัวแปรใดตัวแปรหนึ่ง โดยอาจจะพิจารณาความสัมพันธ์ของตัวแปรด้วยทฤษฎีว่าตามทฤษฎีแล้วตัวแปรใดควรจะมีความสำคัญมากกว่าก็เลือกตัวแปรนั้น
- ถ้าไม่ทิ้งตัวแปรใดตัวแปรหนึ่งก็ต้องหาความสัมพันธ์ทั้งสองตัวแปร แล้วนำส่วนที่ไม่มีความสัมพันธ์กันมาใช้

5. การตรวจสอบอำนาจการพยากรณ์

การทดสอบว่าสมการประมาณค่า $Y = a+b_1x_1+b_2x_2+\ldots+b_kx_k$ มีอำนาจในการพยากรณ์ สูง-ต่ำเพียงใด นั้น เราอาจจะทดสอบได้ง่าย ๆ คือ

(1) ทำการแทนค่าตัวแปรอิสระทุกตัวจะได้ค่า $\hat{Y}_t$ ณ เวลาต่าง ๆ $t = 1,2,3,\ldots,n$

(2) นำค่า $\hat{Y}_t$ ที่ได้เป็นตัวแปรอิสระ ดูความสัมพันธ์กับตัวแปรตาม Y_t โดย

$$Y_t = a + b\hat{Y}_t$$

ทำการวิเคราะห์โดยวิธี OLS ได้ค่า a และ b

ถ้าทดสอบได้ว่า a = 0 และ b = 1 แสดงว่า $\hat{Y}_t$ จะเป็นตัวพยากรณ์ที่มีประสิทธิภาพ (Efficient Predictor) คือค่า $\hat{Y}_t$ หรือสมการ $\hat{Y}_t = \hat{a} + \hat{b}_1X_1 + \hat{b}_2X_2 + \ldots +$ เป็นสมการที่สามารถนำไปใช้ในการทำนายผลของตัวแปรตาม Y ได้เป็นอย่างดี