

บทที่ 7
การสร้างแบบจำลองเศรษฐมิติ
(Econometric Modeling)

วัตถุประสงค์ : เพื่อศึกษาถึงการสร้างแบบจำลองแบบดั้งเดิม (The traditional view of econometric modeling) : Average Economic Regression (AER) ประเภทของความผิดพลาดเฉพาะ ผลของความผิดพลาดในกรณีต่างๆ การตรวจสอบความผิดพลาด ตลอดจนศึกษาถึงความผิดพลาดจากการวัด

บทที่ 7

การสร้างแบบจำลองเศรษฐมิติ

(Econometric Modeling)

1. การสร้างแบบจำลองแบบดั้งเดิม

(the traditional view of econometric modeling) : Average Economic Regression (AER)

หลักการสร้างแบบจำลองโดยวิธี Average Economic Regression (AER) มีดังนี้

- 1) Parsimony : มัธยัสถ์หรือประหยัด นั่นคือการสร้างแบบจำลองโดยวิธี AER นั้นต้องสร้างให้มีรูปแบบอย่างง่ายที่สุดเท่าที่จะเป็นไปได้
- 2) Identifiability : การชี้ชัด นั่นคือในการประมาณค่าพารามิเตอร์ของแบบจำลองจากข้อมูลชุดหนึ่งจะต้องให้คำตอบของแบบจำลองนั้นเพียงคำตอบเดียว
- 3) Goodness of fit : สาระรูปสนิทธิ ในการสร้างแบบจำลองให้มีคุณภาพ และน่าเชื่อถือ นั้น ตัวแปรอิสระจะต้องสามารถอธิบายการเปลี่ยนแปลงของตัวแปรตามได้มากที่สุด โดยดูจากค่า R^2 ซึ่งยังมีค่าสูงก็ยิ่งดี
- 4) Theoretical consistency : ความกลมกลืนกับทฤษฎี พิจารณาได้จากเครื่องหมายของค่าสัมประสิทธิ์ของสมการถดถอยในแบบจำลองว่าเป็นไปตามทฤษฎีเศรษฐศาสตร์หรือไม่ เช่น ในแบบจำลองเรื่องอุปสงค์ที่สัมประสิทธิ์ของตัวแปรราคาต้องมีเครื่องหมายเป็นลบ เป็นต้น
- 5) Predictive power : อำนาจการพยากรณ์ แบบจำลองที่สร้างขึ้นนั้นจะต้องมีความสามารถทางการพยากรณ์จากข้อมูลที่ได้รับ ไปสู่ข้อมูลที่ต้องการรูปซึ่งเรามีได้รับมา

2. ประเภทของความผิดพลาดเฉพาะ (types of specification errors)

1) omission of relevant variable เป็นความผิดพลาดที่เกิดจากการละทิ้งตัวแปรที่มีความสำคัญ ตัวอย่างเช่น

$$\text{สมมติให้แบบจำลองที่ถูกต้องคือ } Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_{ii} \quad \dots(1)$$

โดยที่ $Y_i =$ ต้นทุนรวมของผลผลิต

$X_i =$ ปริมาณผลผลิต

แต่แบบจำลองที่ผู้วิจัยใช้คือ $Y_i = \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^2 + u_{2i}$... (2)

จากสมการที่ (2) พบว่าเกิดปัญหา Specification Errors กรณีละทิ้งตัวแปรสำคัญ เนื่องจากได้ตัดตัวแปร X_i^3 โดยมีค่า $u_{2i} = u_{1i} + \beta_4 X_i^3$

2) inclusion of an unnecessary variable เป็นความผิดพลาดที่เกิดจากการเพิ่มตัวแปรที่ไม่มีความเกี่ยวข้องเข้าไปในแบบจำลอง ตัวอย่างเช่น

หากสมการที่ (1) เป็นแบบจำลองที่ถูกต้อง แต่ผู้วิจัยใช้แบบจำลองดังนี้

$$Y_i = \lambda_1 + \lambda_2 X_i + \lambda_3 X_i^2 + \lambda_4 X_i^3 + \lambda_5 X_i^5 + u_{3i} \quad \dots (3)$$

จากสมการที่ (3) จะเห็นได้ว่ามีตัวแปร X_i^5 เพิ่มขึ้นมาจากสมการที่ (1) แสดงว่าเกิดปัญหา specification errors กรณีเพิ่มตัวแปรที่ไม่มีความเกี่ยวข้องเข้าไปในแบบจำลอง ซึ่งจริง ๆ แล้ว $\lambda_5 = 0$ หาก $\lambda_5 \neq 0$; $u_{3i} = u_{2i} - \lambda_5 X_i^5$

3) adoption the wrong functional form เป็นความผิดพลาดที่เกิดจากการใช้สมการผิดรูป เช่น $\ln Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_{1i}$... (4)

4) errors of measurement เป็นความผิดพลาดที่เกิดจากการวัด ตัวอย่างเช่น

$$Y_i^* = \beta_1^* + \beta_2^* X_i^* + \beta_3^* X_i^{2*} + \beta_4^* X_i^{3*} + u_{1i}^* \quad \dots (5)$$

โดยที่ $Y_i^* = Y_i + \varepsilon_i$

และ $X_i^* = X_i + \omega_i$

โดย ε_i และ ω_i ความผิดพลาดเนื่องจากการวัด

3. ผลของความผิดพลาดเฉพาะ (consequences of specification errors)

1) omission a relevant variable การละทิ้งตัวแปรสำคัญ

สมมติว่าแบบจำลองที่ถูกต้องคือ $Y_i = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_3 + u_{1i}$... (6)

แต่ผู้วิจัยสร้างแบบจำลองเป็น $Y_i = \alpha_1 + \alpha_2 X_2 + \alpha_3 X_2 + u_{2i}$... (7)

ผลของการละทิ้งตัวแปร X_3 คือ

- a) ถ้า X_3 มีความสัมพันธ์กับ X_2 นั่นคือ ค่า $r_{23} \neq 0$ จะทำให้ $\hat{\alpha}_1$ และ $\hat{\alpha}_2$ จะเกิด bias และเกิดปัญหาความไม่คงเส้นคงวาขึ้น ถ้า $E(\hat{\alpha}_1) \neq \beta_1$ และ $E(\hat{\alpha}_2) \neq \beta_2$ และค่า bias จะไม่หมดไปแม้ตัวอย่างจะใหญ่ขึ้น

- b) เมื่อ X_2 และ X_3 ไม่มีความสัมพันธ์กัน ($r_{23} = 0$) ค่า $\hat{\alpha}_1$ จะเป็น bias แต่ $\hat{\alpha}_2$ ไม่ bias
- c) ค่าความแปรปรวน $\hat{\sigma}^2$ ของตัวแปรที่คำนวณได้จะไม่ถูกต้อง
- d) การวัดความแปรปรวนของ $\hat{\alpha}_2$ จะ bias ต่างกับความแปรปรวนของ $\hat{\beta}_2$ ที่ไม่ bias
- e) ผลที่ตามมาคือ ค่า confidence interval และการทดสอบสมมติฐานเกิดปัญหาความผิดพลาดเกี่ยวกับความเชื่อมั่นของการประมาณค่าพารามิเตอร์

2) inclusion of an unnecessary variable การรวมตัวแปรที่ไม่เกี่ยวข้องไว้ในแบบจำลอง สมมติแบบจำลองที่ถูกต้องคือ $Y_i = \beta_1 + \beta_2 X_{1i} + \beta_3 X_{2i} + u_{1i}$... (8)

แต่ผู้วิจัยสร้างแบบจำลองเป็น $Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{2i} + \alpha_4 X_{3i} + u_{2i}$... (9)

การที่ผู้วิจัยเพิ่มตัวแปร X_3 เข้าไปในแบบจำลองทำให้

- a) การประมาณค่าด้วยวิธี OLS จากแบบจำลองนี้ยังคงเป็น unbiased และ consistent

$$\text{นั่นคือ } E(\hat{\alpha}_1) = \beta_1, E(\hat{\alpha}_2) = \beta_2 \text{ และ } E(\hat{\alpha}_3) = \beta_3 = 0$$

- b) ค่าประมาณความแปรปรวนของตัวคลาดเคลื่อน $\hat{\sigma}^2$ ยังคงถูกต้อง

- c) ค่าความเชื่อมั่น และการทดสอบสมมติฐานยังคงใช้ได้

- d) แต่ค่าประมาณของพารามิเตอร์ในสมการ (9) จะไม่มีประสิทธิภาพ ทั้งนี้เพราะความแปรปรวนของค่าพารามิเตอร์ในสมการ(9) จะมากกว่าความแปรปรวนของค่าพารามิเตอร์ในสมการ (8) นั่นคือทำให้ค่าพารามิเตอร์ที่ประมาณได้ในสมการ (9) ไม่มีความน่าเชื่อถือ

4. การตรวจสอบความผิดพลาด (tests of specification errors)

- 1) การหาตัวแปรที่ไม่สำคัญ (detecting the presence of unnecessary variable)

สมมติเราสร้างแบบจำลองที่มีตัวแปร k ตัว

$$Y_i = \beta_1 + \beta_2 X_{1i} + \beta_3 X_{2i} + \dots + \beta_k X_{ki} + u_i \quad \dots (10)$$

หากเราต้องการทดสอบว่าตัวแปร X_k มีความสำคัญกับแบบจำลองนี้หรือไม่ เราจะทดสอบ

ความมีนัยสำคัญของค่า β_k โดยใช้ตัวสถิติ t ในการทดสอบ ซึ่งคำนวณได้จาก $t = \frac{\hat{\beta}_k}{se(\hat{\beta}_k)}$ แล้ว

นำไปเทียบกับค่า t จากตารางหาก $t = \frac{\hat{\beta}_k}{se(\hat{\beta}_k)} > t$ แสดงว่าตัวแปรดังกล่าวมีความสำคัญ แต่หากเราต้องการทดสอบว่าตัวแปรทุกตัวในแบบจำลองมีความเหมาะสมหรือไม่ กรณีเราจะทดสอบว่า $\beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$ โดยใช้ตัวสถิติ F

2) nominal vs true level of significance in presence of data mining

อันตรายอันหนึ่งที่เกิดจากการจัดการข้อมูล คือ การเลือกใช้ระดับนัยสำคัญทางสถิติซึ่งโดยทั่วไปมักใช้ที่ระดับ 1%, 5% และ 10% ซึ่งบางครั้งการจัดการข้อมูลไม่ได้ระดับนัยสำคัญที่แท้จริง

Lovell แนะนำว่าหากมีตัวอิสระ c ตัว ให้เลือกตัวแปรอิสระมาเพียง k ตัว ($k \leq c$) ในการจัดการข้อมูลพื้นฐานนั้น ระดับนัยสำคัญทางสถิติที่แท้จริง (α^*) จะต่างจากระดับนัยสำคัญที่ใช้ทั่วไปคือ

$$\alpha^* = 1 - (1 - \alpha)^{c/k}$$

หรือค่าประมาณคือ
$$\alpha^* = \left(\frac{c}{k}\right)\alpha$$

เช่น ถ้า $c = 15$, $k = 5$ และ $\alpha = 5\%$ จะได้ระดับนัยสำคัญที่แท้จริง คือ $(15/5)(5) = 15\%$ ไม่ใช่ 5%

3) test for omitted variable and incorrect function form การทดสอบเกี่ยวกับการละเว้นตัวแปร และฟังก์ชันที่ไม่ถูกต้อง

การพิจารณาว่าแบบจำลองที่สร้างขึ้นว่ามีความถูกต้องมากน้อยเพียงใด หรือเกิดปัญหาการละทิ้งตัวแปรหรือไม่นั้นสามารถพิจารณาได้จาก ค่าคลาดเคลื่อน (examination of residuals) ซึ่งเป็นวิธีที่ดีในการพิจารณาเกี่ยวกับปัญหา autocorrelation หรือ ปัญหา heteroscedasticity แต่จริง ๆ แล้ว การพิจารณาความคลาดเคลื่อนยังใช้ได้กับข้อมูลประเภท cross-section data เพื่อพิจารณาเกี่ยวกับความผิดพลาดของแบบจำลอง เช่นการละทิ้งตัวแปร หรือการใช้รูปแบบการผิด

ถ้าเกิดปัญหาลักษณะกราฟของตัวคลาดเคลื่อนจะเป็นตัวบ่งบอกถึงปัญหา นั่นคือหากมีการละทิ้งตัวแปร หรือใช้สมการผิดรูปเส้นกราฟของ residual จะกวัดแกว่งในระดับสูง

4) the Durbin - Watson (d-statistic) once again

การทดสอบค่า Durbin - Watson (d) เป็นการตรวจสอบความผิดพลาดของแบบจำลอง ว่าแบบจำลองที่สร้างขึ้นเกิดปัญหา autocorrelation หรือไม่ ขั้นตอนการใช้ค่า Durbin - Watson ทดสอบความผิดพลาดของแบบจำลอง มีดังนี้

- a) จากแบบจำลองที่สร้างขึ้น ทำการประมาณค่าด้วยวิธี OLS แล้วหาค่า residual ออกมา
- b) ถ้าเราเชื่อว่าแบบจำลองที่สร้างขึ้นไม่ถูกต้อง เพราะไม่ได้ใส่ตัวแปรอิสระที่สำคัญ บางตัวเข้าไปเช่นตัวแปร Z ทำให้ค่าความคาดเคลื่อนที่ได้จากข้อ 1. มีค่าสูง โดยตัวแปร Z อาจเป็นตัวแปรใด ๆ หรืออยู่ในรูป X^2, X^3 ก็ได้
- c) ทำการคำนวณค่าสถิติ d จากสูตร

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2} \quad ; \quad e_i = \hat{u}_i$$

- d) นำค่า d ที่คำนวณได้ไปเปรียบเทียบกับค่า d_L และ d_U จากตาราง หาก $d < d_L$ หรือ $d > 4 - d_U$ แสดงว่าเกิด specification errors ขึ้นแล้ว

5. ความผิดพลาดจากการวัด (errors of measurement)

1) errors of measurement in the dependent variable Y: เป็นความผิดพลาดที่เกิดจากตัวแปรตาม พิจารณาแบบจำลองต่อไปนี้

$$Y_i^* = \alpha + \beta X + u_i \quad \dots(11)$$

เมื่อ Y_i^* = permanent consumption expenditure

X = current income

U_i = stochastic disturbance term

เนื่องจากค่า Y_i^* ไม่สามารถวัดได้โดยตรงเราจะใช้ตัวแปร Y_i ซึ่งสามารถวัดได้ นั่นคือ

$$Y_i = Y_i^* + \varepsilon_i \quad \dots(12)$$

โดยที่ ε_i คือ ความผิดพลาดจากการวัดของ Y_i^* ดังนั้นแทนที่จะประมาณค่าสมการที่ (11) เราจะทำการประมาณค่าสมการต่อไปนี้แทน

$$Y_i = (\alpha + \beta X + u_i) + \varepsilon_i$$

$$Y_i = \alpha + \beta X + (u_i + \varepsilon_i)$$

$$Y_i = \alpha + \beta X + v_i \quad \dots(13)$$

โดยที่ $v_i = u_i + \varepsilon_i$ คือ ค่าคาดเคลื่อนที่ประกอบด้วยค่าคาดเคลื่อนจากสมการ และค่าผิดพลาดจากการวัดปกติ

สมมุติว่า $E(u_i) = E(\varepsilon_i) = 0$, $\text{cov}(X_i, u_i) = 0$ หมายความว่าความผิดพลาดจากการวัด Y_i^* ไม่มีความสัมพันธ์กับ X_i และ $\text{cov}(u_i, \varepsilon_i) = 0$ นั่นคือความคลาดเคลื่อนจากสมการไม่มีความสัมพันธ์กับความผิดพลาดจากการวัด แสดงว่า การประมาณค่าจากสมการ (11) และ (13) เป็น unbiased กล่าวได้ว่าความผิดพลาดจากการวัดของตัวแปรตามจะไม่ไปทำลายคุณสมบัติ unbiased ของการประมาณค่าด้วยวิธี OLS แต่ความแปรปรวนของสมการทั้ง 2 จะต่างกัน

2) errors of measurement in the explanatory variable X : เป็นความผิดพลาดที่เกิดจากตัวแปรอิสระ สมมุติว่าแทนที่จะใช้แบบจำลองที่ (11) เราใช้แบบจำลองต่อไปนี้

$$Y_i = \alpha + \beta X_i^* + u_i \quad \dots(14)$$

โดยที่

Y_i = current consumption expenditure

X_i^* = permanent income

U_i = disturbance term (equation errors)

สมมุติแทนที่จะได้ค่า X_i^* เราจะได้ค่า

$$X_i = X_i^* + \omega_i \quad \dots(15)$$

โดย ω_i หมายถึงความผิดพลาดที่เกิดจากการวัด X_i^* ดังนั้นแทนที่จะประมาณค่าจากสมการ (14) เราจะประมาณค่าจาก

$$\begin{aligned} Y_i &= \alpha + \beta(X_i - \omega_i) + u_i \\ Y_i &= \alpha + \beta X_i + (u_i - \beta \omega_i) \\ Y_i &= \alpha + \beta X_i + v_i \quad \dots(16) \end{aligned}$$

โดย $v_i = u_i - \beta \omega_i$ ซึ่งเป็นผลรวมของค่าคลาดเคลื่อนจากสมการ และค่าผิดพลาดจากการวัดซึ่งแม้เราจะสมมุติให้ $E(\omega_i) = 0$ และไม่มีความสัมพันธ์กับ u_i เราก็ไม่สามารถกล่าวได้ว่าค่า v_i เป็นอิสระจากตัวแปร X_i เพราะ

$$\begin{aligned} \text{Cov}(v_i, X_i) &= E[v_i - E(v_i)] [X_i - E(X_i)] \\ &= E(u_i - \beta \omega_i) (\omega_i) \\ &= E(-\beta \omega_i^2) \\ &= -\beta \delta_\omega^2 \end{aligned}$$

ดังนั้นตัวแปร X จะมีความสัมพันธ์กับค่าคลาดเคลื่อนเป็นการการละเมิดข้อการประมาณค่าด้วยวิธี OLS ที่ว่าตัวแปรอิสระต้องไม่มีความสัมพันธ์กับตัวคลาดเคลื่อน