ศัพท์พัฒนบริหารศาสตร์

DEVELOPMENT ADMINISTRATION GLOSSARY

ตัวแบบเชิงเส้นของการถดถอยและสหสัมพันธ์

Linear Regression and Correlation : A Note

ศัพท์นี้มีจุดมุ่งหมายที่จะอธิบายความหมายของตัวแบบเชิงเส้นของการถดถอยเชิงเดี่ยว (Linear Model of Simple Regression) และความหมายของสหสัมพันธ์ (Correlation) เพื่อแสดงความแตกต่างระหว่างสองอย่างนี้ และกล่าวถึงข้อจำกัด (limitations) ของการใช้ในการวิเคราะห์ การใช้เทคนิคทางสถิติทั้งสองอย่างนี้มีข้อผิดพลาดอยู่บ่อย ๆ เช่นนำไปใช้กับกรณีที่ไม่เหมาะสม และบางทีผู้ใช้ตีความอย่างผิด ๆ บทความนี้จึงมีเจตนาที่จะเน้นให้ผู้ใช้หลีกเลี่ยงข้อผิดพลาดดังกล่าว

ตัวแบบเชิงเส้นของสมการถดถอยเชิงเดี่ยว (Linear Model of Simple Regression)

ตัวแบบเชิงเส้นของการถดถอยเชิงเดี่ยวที่นักสถิติใช้อยู่เสมอ ๆ และที่นักสังคมศาสตร์ชอบใช้อยู่บ่อย ๆ อยู่ในรูปของ

$$Y_i = a + bX_i + u_i \quad (i = 1,2,\ldots\ldots\ldots\ldots,T) \qquad (1)$$

โดยที่

Y_i = ตัวแปรตาม (dependent variable)

X_i = ตัวแปรอิสระ (independent variable)

u_i = ตัวคลาดเคลื่อน (error)

a, b = ตัวสัมประสิทธิ์ที่ต้องการหาค่า (uknnown coefficients)

i = subscripts มีค่าเท่ากับ 1, 2, 3, ..., T ซึ่งเป็นหน่วยแจงนับในตัวอย่าง[1]

ตัวแบบข้างบนนี้มีความหมายว่าตัวแปร X และตัวคลาดเคลื่อน (u) เป็น "มูลเหตุ" (causes) ที่ทำให้เกิดค่าของตัวแปร Y[2] ตัวอย่าง เช่น ให้ Y แทนจำนวนผลผลิตของข้าวโพด

---

1 จากนี้ไปเราจะใช้ X_i แทนค่าของตัวแปร และใช้ X แทนตัวแปร สำหรับตัวอื่น ๆ ก็มีความหมายในทำนองเดียวกัน

2 E. Malinvaud, *Statistical Methods of Econometrics* (Amsterdam: North – Holland Publishing Co., 1966), p. 74

ต่อไร่ **X** แทนจำนวนปุ๋ยที่ใช้ต่อไร่ ในระยะเวลาเพาะปลูก เราก็อาจตีความหมายตัวแบบนี้ได้ว่า จำนวนปุ๋ยที่ใช้เป็น "สาเหตุ" ที่ทำให้เกิดผลผลิตในปริมาณต่าง ๆ กัน ในที่นี้เรากล่าวถึงปุ๋ยเพียงปัจจัยเดียว (ทั้งนี้เพื่อความง่ายในการคำนวณ หรือเราอาจจะมีความสนใจแต่ความสัมพันธ์ระหว่างปุ๋ยและผลผลิตเท่านั้น) ความจริงแล้วปริมาณผลิตผลของข้าวโพดยังขึ้นอยู่กับน้ำ อากาศ และปัจจัยอื่น ๆ อีก ทั้งที่สามารถวัดค่าได้ (quantifiable) และที่ไม่สามารถวัดค่าได้ (unquantifiable) เพราะฉะนั้นเราจึงต้องพิจารณาสิ่งเหล่านี้ด้วย ตัวคลาดเคลื่อน (u) นี้แหละที่แทนปัจจัยที่เราตัดทิ้งไปจากสมการ นอกจากนี้เราบ่งว่า **X** และ **Y** มีความสัมพันธ์กันอย่างเชิงเส้น แต่ความสัมพันธ์ที่แท้จริงอาจเป็นอย่างอื่น เช่น ในรูปของ **quadratic function** หรือ **polynomial function** หรือ **exponential function** ก็ได้ ในตัวอย่างข้างบนจะเห็นได้ว่า ปริมาณผลผลิตของข้าวโพดจะมีทางเป็นไปได้น้อยที่จะมีความสัมพันธ์อย่างเชิงเส้นกับปริมาณปุ๋ยที่ใช้ ทั้งนี้เพราะว่าเมื่อเราเพิ่มปริมาณปุ๋ยต่อไร่ ผลผลิตจะเพิ่มขึ้น แต่จะถึงระยะหนึ่งที่เมื่อเพิ่มปุ๋ยมากขึ้นจะไม่ทำให้ผลผลิตเพิ่มขึ้นเลยหรืออาจทำให้ผลผลิตลดลงเพราะใช้ปุ๋ยปริมาณมากเกินไป ปริมาณของปุ๋ยและผลผลิตจะไม่เพิ่มด้วยกันเมื่อพ้นระยะหนึ่งไป เพราะฉะนั้นตัวคลาดเคลื่อน (u) จึงอาจใช้แทนความแตกต่างระหว่างตัวแบบที่ใช้กับตัวแบบที่แท้จริงได้อีกอย่างหนึ่ง

วิธีที่ใช้ประมาณค่าของ a และ b ในตัวแบบ $Y_i = a + bX_i + u_i$ $(i = 1, 2, \ldots, T)$ มีหลายวิธีด้วยกัน แต่วิธีที่รู้จักกันดีที่สุดคือ วิธีกำลังสองน้อยที่สุด (The Method of Least Squares) ซึ่งนักคณิตศาสตร์ชาวฝรั่งเศสชื่อ Adrain Legendre เป็นผู้เสนอ ตั้งแต่ปี ค.ศ. 1806 วิธีนี้เป็นที่นิยมใช้กันมาก เพราะเป็นวิธีที่ง่ายและถ้าสมมติฐานต่าง ๆ ของตัวแบบนี้เป็นจริง ตัวประมาณโดยวิธีกำลังสองน้อยที่สุด (least squares estimators) ก็จะมีคุณสมบัติเป็น Best Linear Unbiased Estimators[3]

หลักการของวิธีกำลังสองน้อยที่สุดมีอยู่ว่า จาก $Y_i = a + bX_i$ เราประมาณค่าด้วย $\hat{Y}_i = \hat{a} + \hat{b}X_i$ ($\hat{a}$ และ $\hat{b}$ เป็นตัวประมาณของ a และ b ตามลำดับ) ความแตกต่างระหว่างค่าจริงและค่าประมาณของ Y (เรียกว่า residual) จะเท่ากับ $e_i = Y_i - \hat{Y}_i$ วิธีกำลังสองน้อยที่สุดต้องการทำให้

$$\sum_{i=1}^{T} e_i^2 = \sum_{i=1}^{T} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{T} (Y_i - \hat{a} - \hat{b}X_i)^2 \text{ น้อยที่สุด}$$

3 *Ibid.*, ch. 3

ซึ่งเราสามารถแสดงว่าค่าประมาณของ a และ b จะได้ตามสูตร

$$\hat{a} = \bar{Y} - \hat{b}\bar{X} \qquad (2)$$

$$\hat{b} = \frac{\sum_{i=1}^{T} x_i y_i}{\sum_{i=1}^{T} x_i^2} \qquad (3)$$

โดยที่

$$x_i = X_i - \bar{X};\ \bar{X} = \frac{1}{T}\sum_{i=1}^{T} X_i$$

$$y_i = Y_i - \bar{Y};\ \bar{Y} = \frac{1}{T}\sum_{i=1}^{T} Y_i$$

และเราสามารถหาความแปรปรวนของตัวประมาณทั้งสองได้จากสูตร

$$\text{Var}(\hat{b}) = \frac{\sigma_u^2}{\Sigma x_i^2}, \quad \text{Var}(\hat{a}) = \frac{\sigma_u^2 \Sigma X_i^2}{T\Sigma x_i^2} \qquad (4)$$

โดย σ_u^2 หาได้จากตัวประมาณ

$$\Sigma e_i^2 \quad \Sigma y_i^2 - b\,\Sigma\, x_i y_i \qquad (5)$$
$$= \frac{}{T-2} = \frac{}{T-2}$$

ตัวแบบเชิงเส้นของการถดถอยอย่างง่ายเป็นที่นิยมใช้กันมาก เพราะง่ายต่อการคำนวณ ในการศึกษาที่ต้องการทราบแต่เพียงแนวโน้มของตัวแปรเพียง 2 ตัว การใช้ตัวแบบนี้ก็เป็นประโยชน์มาก แต่การใช้ตัวแบบนี้ก็มีขอบเขตจำกัด จะนำไปใช้กับทุกกรณีไม่ได้

ประการแรกเราควรใช้ตัวแบบนี้สำหรับกรณีที่ตัวแปร Y และ X มีความสัมพันธ์กันอย่างจริง ๆ หรือที่คิดว่าน่าจะมีความสัมพันธ์กัน ทั้งนี้เพราะเราต้องการใช้ตัวแบบนี้ในการพยากรณ์ ถ้าเราทราบค่า X ก็สามารถทำนายค่า Y ได้ ถ้า X และ Y ไม่มีความสัมพันธ์กันในลักษณะ 'เหตุ' และ 'ผล' การใช้ตัวแบบก็ให้ประโยชน์น้อย หรือไม่มีประโยชน์เลย จริงอยู่ที่เรามีวิธีทางสถิติที่จะทดสอบ สมมติว่า X และ Y มีความสัมพันธ์กันหรือไม่ (นั่นคือ $b = 0$ หรือ $b \neq 0$) แต่นี่ก็เป็นเพียงผลทางสถิติเท่านั้น เพราะว่าการทดสอบอาจจะบอกว่า X และ Y มีความสัมพันธ์กัน ค่าของ r^2 (coefficient of determination) ซึ่งจะกล่าวถึงในหัวข้อต่อไป ก็

อาจมีค่าสูงแต่ในทางปฏิบัติใช้ไม่ได้ เพราะเราอาจเลือกตัวแปรสองตัวโดยเอาสุ่มแล้วนำมาเข้าตัวแบบ ก็มี probability สูงมาก ที่โดยทางสถิติเราต้องยอมรับว่าตัวแปรทั้งสองมีความสัมพันธ์กัน แต่เราจะได้ผลที่น่าตลกขบขันมาก เช่นให้ Y เป็นจำนวนอาชญากรรมที่เกิดขึ้นในประเทศ X เป็นปริมาณถั่วเขียวที่ผลิตได้ในประเทศ ผู้เขียนเชื่อว่าตัวแปรทั้งสองจะเข้ากับตัวแบบได้ดี แต่คงไม่มีใครยอมรับได้ว่าเราจะสามารถทำนายจำนวนอาชญากรรมได้จากการผลิตถั่วเขียว หรือถั่วเขียวเป็นเหตุให้เกิดอาชญากรรม

ประการที่สองตัวแบบที่กล่าวถึงนี้เป็น ตัวแบบเชิงเส้น เพราะเส้นที่ใช้ทำนายค่าของตัวแปรจะเป็นเส้นตรง ตัวแบบนี้เหมาะสำหรับกรณีที่ตัวแปรมีค่าเพิ่มก็เพิ่มด้วยกันหรือ มีค่าลดก็ลดด้วยกัน เพราะฉะนั้นตัวแบบนี้จึงไม่เหมาะสำหรับกรณีที่ตัวแปรมีความเปลี่ยนแปลงขึ้นลง (fluctuation) มาก เช่นในกรณีของผลิตผลทางเกษตรบางอย่าง ทั้งนี้เพราะผลิตผลเหล่านี้ขึ้นอยู่กับปัจจัยหลายอย่างซึ่งเปลี่ยนแปลงอยู่เรื่อย

ประการที่สาม วิธีที่ใช้ประมาณค่าของ a และ b จากตัวแบบเช่นวิธีกำลังสองน้อยที่สุดจะให้ตัวประมาณที่ดีก็ต่อเมื่อสมมติฐานของตัวแบบเป็นจริง ยกตัวอย่างเช่นเราตั้งข้อสมมติว่าตัวแปร X และ Y จะวัดค่าได้โดยไม่ผิดพลาดหรือความผิดพลาดมีน้อยมากจนตัดทิ้งไปได้ แต่ถ้าในทางปฏิบัติค่าของตัวแปรที่วัดได้ต่างกับค่าจริงมาก การใช้วิธีกำลังสองน้อยที่สุดตามธรรมดาก็จะไม่ให้ตัวประมาณที่ดี หรือถ้าข้อสมมติที่ว่าความแปรปรวนของตัวการะจายคงที่ไม่เป็นจริง เราก็จำเป็นต้องดัดแปลงวิธีการประมาณใหม่

ประการที่สี่ ในตัวแบบ $Y_i = a + bX_i + u_i$ เราถือว่า X เป็นตัวแปรอิสระ และสมการนี้เป็นสมการเดียวแต่ตัวมันเอง ถ้าสมมติฐานนี้และสมมติฐานอื่น ๆ ของตัวแบบเป็นจริง วิธีกำลังสองน้อยที่สุดก็จะให้ตัวประมาณที่ดี แต่ถ้าสมการนี้เป็นสมการหนึ่งใน Simultaneous Equation System การใช้วิธีกำลังสองน้อยที่สุด ตามธรรมดาก็จะให้ตัวประมาณที่ไม่ดี เพราะจะเกิด bias และ inconsistency ขึ้น[4] ตัวอย่างของกรณีนี้อาจเห็นได้จาก Consumption Function $C_i = a + bY_i + u_i$ โดย C_i = ค่าใช้จ่ายในการบริโภค Y_i = รายได้ ค่าของ C และ Y จะถูกตัดสินร่วมกัน การประมาณที่ดีจะต้องใช้ระบบ Simultaneous Equation การ

[4] J. Johnston, *Econometric Methods* (New York: Mc-Graw-Hill, 1963), ch. 9.

ใช้วิธีกำลังสองน้อยที่สุดตามธรรมดาในการประมาณค่าของ $C_i = a + bY_i + u_i$ อย่างตรง ไปตรงมา เราก็จะได้สิ่งที่ไม่มีความหมายมากนัก

สหสัมพันธ์ (Correlation)

ในการทำการวิเคราะห์ที่ใช้ตัวแบบการถดถอยมักจะให้ค่าของสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) r ด้วย ซึ่งอาจหาได้จากสูตร

$$r = \frac{\Sigma xy}{[(\Sigma x^2)(\Sigma y^2)]^{\frac{1}{2}}} \qquad (6)$$

ตัวสัมประสิทธิ์สหสัมพันธ์แสดงดีกรีของความสัมพันธ์ระหว่างตัวแปรทั้งสอง ถ้า r มีค่าสูง หมายความว่าตัวแปรทั้งสองมีความสัมพันธ์กันมาก และเป็นตรงข้ามถ้าหาก r มีค่าต่ำ ส่วนเครื่องหมายของ r จะบอกว่า Y จะเพิ่มขึ้นหรือลดลงไปกับ X แต่สัมประสิทธิ์สหสัมพันธ์เป็นเครื่องวัดความสัมพันธ์ระหว่างตัวแปรที่ใช้ได้ในกรณีที่ตัวแปร X และ Y มีความสัมพันธ์อย่างเชิงเส้นเท่านั้น ถ้า $r = 0$ อาจเป็นเพราะ X และ Y ไม่มีความสัมพันธ์กันเลยหรือมีความสัมพันธ์มิใช่เชิงเส้น (non-linear relationship)

การตีความหมายสัมประสิทธิ์สหสัมพันธ์ว่าเป็นเครื่องวัดความสัมพันธ์เชิงเส้น ระหว่างตัวแปรสองตัวนั้นเป็นคุณสมบัติทางคณิตศาสตร์เท่านั้น มิได้มีนัยต่อความสัมพันธ์แบบ "สาเหตุ" และ 'ผล' แต่อย่างไร การที่ตัวแปรมีแนวโน้มที่เพิ่มขึ้นหรือลดต่ำด้วยกันมิได้หมายความว่าตัวแปรตัวหนึ่งจะมีอิทธิพลต่อตัวแปรอีกตัวหนึ่งโดยตรงหรือโดยอ้อมเสมอไป ในบางกรณีตัวแปรตัวหนึ่งมีอิทธิพลต่อตัวแปรอีกตัวหนึ่งจริง แต่ค่าของสัมประสิทธิ์สหสัมพันธ์มิอาจบอกเราได้ ตัวอย่างที่ยกมาแล้วข้างต้น คือจำนวนคาชญากรรม และปริมาณผลผลิตของถั่วเขียว ถ้าเราคำนวณหาค่าสัมประสิทธิ์สหสัมพันธ์ ผู้เขียนเชื่อว่าจะต้องมีค่าสูง ซึ่งแสดงว่าสิ่งทั้งสองอย่างต่างมีแนวโน้มสูงขึ้นด้วยกัน ตัวสัมประสิทธิ์นี้มิอาจบอกอะไรเราได้มากกว่านี้

สัมประสิทธิ์อีกตัวหนึ่งที่น่าสนใจคือ r^2 ซึ่งเรียกว่า Coefficient of Determination หาค่าได้จากสูตร

$$r^2 = 1 - \frac{\Sigma e_i^2}{\Sigma y_i^2} \qquad (7)$$

ตัวสัมประสิทธิ์นี้มีความสัมพันธ์กับความแปรปรวน (Variances) ของ X และ Y โดย

$$\frac{\Sigma\ \hat{y}_i^2}{\Sigma\ y_i^2} = \frac{\Sigma\ (\hat{b}x_i)^2}{\Sigma\ y_i^2}$$

$$= \hat{b}\ \frac{\Sigma\ x_i^2}{\Sigma\ y_i^2} = \frac{(\Sigma\ x_i y_i)^2}{(\Sigma\ x_i^2)^2} \cdot \frac{\Sigma\ x_i^2}{\Sigma\ y_i^2}$$

$$= r^2 \qquad (8)$$

นั่นคือ r^2 = สัดส่วนของความแปรปรวนของ Y ที่อธิบายโดยอิทธิพลของ X เช่น ถ้า r^2 มีค่า 0.8 หมายความว่าการถดถอยโดยวิธีกำลังสองน้อยที่สุดของ Y ต่อ X จะอธิบาย 80% ของความแปรปรวนของ Y

สำหรับตัว r^2 นี้ ผู้ทำการวิเคราะห์ส่วนมากมักอธิบายว่าเป็นตัวที่แสดง goodness of fit แต่ความจริงแล้ว r^2 มิได้ให้ความรู้เกี่ยวกับเรื่องนี้เลย r^2 เพียงแต่บอกรูปลักษณะของการรวมกลุ่มของจุดต่าง ๆ เท่านั้น การที่ r^2 มีค่าสูงเราพอที่จะถือว่าตัวแบบเชิงเส้นที่เราใช้มีความเหมาะสม ในตัวอย่างเล็ก ๆ ซึ่งปกติใช้ในการวิเคราะห์ทางเศรษฐศาสตร์หรือสังคมศาสตร์แขนงอื่นนั้น r^2 จะมีค่าสูงได้ถึงแม้ความเคลื่อนคลาดมาตรฐาน (standard error) ของ $\hat{a}$ และ $\hat{b}$ จะมีค่าสูง นั่นคือเราอาจมีช่วงทำนายผลกว้างมากเกินไป จนไม่มีประโยชน์อะไร ทั้ง ๆ ที่ r^2 มีค่าสูงในทางตรงข้ามตัวอย่างใหญ่จะให้ค่าของ $\hat{a}$ และ $\hat{b}$ ที่ใกล้เคียงกับ a b มาก แต่ r^2 อาจมีค่าต่ำ[5] Malinvaud ได้มีตัวอย่างดังนี้

$$\hat{Y} = -1.78 + 1.20\ X$$
$$(0.40)\quad (0.10)$$
$$r^2 = 0.59\ ;\ T = 112 \qquad (9)$$

$$\hat{Y} = -6.20 + 0.115\ X$$
$$(1.88)\quad (0.012)$$
$$r^2 = 0.95\ ;\ T = 11 \qquad (10)$$

---

[5] Malinvaud, *op. cit.*, p. 83

จะเห็นได้ว่าใน (9) ค่า r^2 ต่ำกว่า (10) มากทั้งที่อัตราส่วนระหว่างค่าของ $\hat{a}$, $\hat{b}$ กับความเคลื่อนมาตรฐานของแต่ละตัว (ตัวเลขในวงเล็บข้างล่าง ตัวสัมประสิทธิ์ $\hat{a}$, $\hat{b}$) จะใกล้เคียงกัน ทั้งนี้เพราะใน (9) เรามีข้อมูลถึง 112 ส่วน ใน (10) มีเพียง 11 เท่านั้น เพราะฉะนั้นจึงพอสรุปได้ว่าการใช้ r^2 แต่เพียงอย่างเดียวในการเลือกตัวแบบที่ดีจากตัวแบบที่ให้ค่า a และ b ต่าง ๆ กันนั้นไม่เป็นการเพียงพอ

จากที่กล่าวมาทั้งหมดนี้ความแตกต่างระหว่างตัวแบบการถดถอยและสหสัมพันธ์ก็คงชัดเจนพอสมควร ความหมายของตัวแบบทั้งสองมีความสำคัญมาก แต่มักขาดการเน้นในการสอนสถิติเบื้องต้นแก่นักศึกษา ซึ่งเป็นการเสี่ยงต่อการใช้แบบผิด ๆ สิ่งที่จูงใจให้ผู้เขียน เขียนเรื่องนี้ก็คือ การที่ได้อ่านวิทยานิพนธ์ชั้นปริญญาโทของมหาวิทยาลัยบางแห่งและบทความในหนังสือปริทัศน์ ตลอดจนผลงานวิจัยซึ่งใช้วิธีการทางสถิติทั้งสองอย่างนี้อย่างผิด ๆ และตีความผิดพลาด

ความสำเร็จของการใช้เทคนิคทั้งสองอย่างนี้หรือเทคนิคอื่น ๆ ทางสถิติขึ้นอยู่กับความเข้าใจในคุณสมบัติของตัวแบบ ความหมาย ข้อจำกัด และการใช้ถูกต้อง เหมาะสมกับภาวะเหตุการณ์ การสอนสถิติเบื้องต้นจึงน่าจะเน้นหนักไปในเรื่องเหล่านี้บ้าง มิใช่แต่ให้นักศึกษาท่องสูตรแล้วนำไปใช้อย่างผิด ๆ

จำลอง อติกุล