บทที่ 2
การวิเคราะห์เส้นถดถอย
(Regression Analysis)

<u>ความสัมพันธ์ระหว่างตัวแปร</u>

ทฤษฎีเศรษฐศาสตร์จำนวนมากที่เกี่ยวข้องกับการตัดสินความสัมพันธ์ระหว่างตัวแปร ตัวอย่างเช่น จำนวนที่ demand ขึ้นอยู่กับราคา รายได้ และตัวแปรอื่น ๆ ราคาหุ้นมีความสัมพันธ์เป็นบวกกับรายได้ที่หามาได้ของบริษัท รายจ่ายเพื่อการบริโภคของบุคคลเป็นฟังชั่นกับรายได้ที่สามารถจะจ่ายได้ การผลิตขึ้นอยู่กับทุน แรงงาน และเทคโนโลยี เป็นต้น ความจริงความสัมพันธ์เหล่านี้เป็นสมมติฐานหรือทฤษฎีเกี่ยวกับพฤติกรรมของมนุษย์ ดังนั้นนักเศรษฐศาสตร์จึงใช้ข้อมูลจากการสังเกตที่เป็นจริงในโลกทดสอบและพิสูจน์ทฤษฎีของเขา หากทฤษฎีใช้ไม่ได้ เขาก็จะเสาะหาคำอธิบายใหม่ ๆ หากทฤษฎีใช้ได้สิ่งที่เกี่ยวข้องใหม่ ๆ และการทดสอบที่ยากขึ้นก็จะดำเนินต่อไป เพราะทฤษฎีเศรษฐศาสตร์ไม่ใคร่จะให้ค่าตัวเลขสำหรับสัมประสิทธิของความสัมพันธ์ สัมประสิทธิเหล่านี้ต้องได้รับการประมาณค่าก่อนที่จะนำข้อสรุปที่มีประโยชน์นั้นไปใช้ เมื่อพบสิ่งที่ไม่คาดคิดในการเคลื่อนไหวของตัวแปรเศรษฐกิจ คำถามและปัญหาในทางทฤษฎีใหม่ก็เกิดขึ้น เครื่องมือที่สำคัญและเป็นพื้นฐานมากที่สุดในการศึกษาความสัมพันธ์ระหว่างตัวแปรก็คือ การถดถอย (regression)

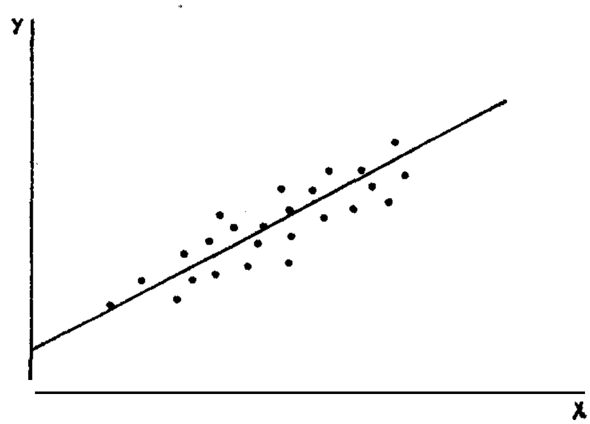
สมมติว่าเราต้องการจะตรวจดูความสัมพันธ์ระหว่างตัวแปรเศรษฐกิจ ๒ ตัว คือ X กับ Y โดยใช้ค่าสังเกต (observation) n ตัว ดังตารางที่ ๒.๑

ตารางที่ ๒.๑

ค่าสังเกตของตัวแปร X และ Y

ค่าสังเกต	X	Y
๑	X_1	Y_1
๒	X_2	Y_2
๓	X_3	Y_3
๔	X_4	Y_4
๕	X_5	Y_5
.		
.		
.		
n - l	X_{n-l}	Y_{n-l}
n	X_n	Y_n

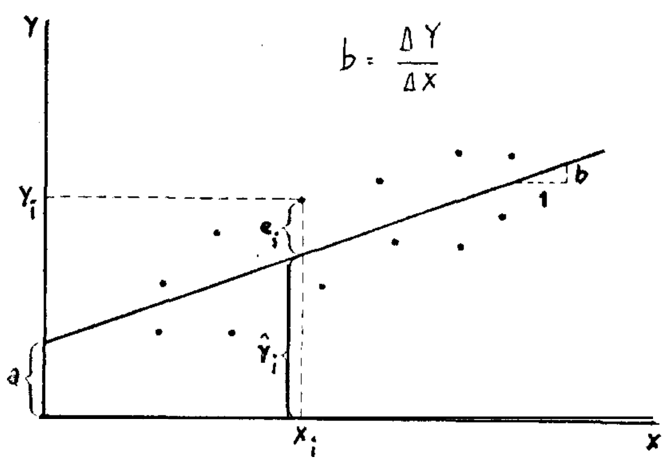
ความสัมพันธ์ของ X และ Y จะ plot ออกมาเป็นรูปแผนภาพกระจัดกระจาย (scatter diagram) ดังรูปที่ ๒.๑ จะเห็นชัดว่า ความสัมพันธ์ทั้ง X และ Y มีแนวโน้มเพิ่มขึ้นด้วยกันทั้งคู่ สมมติว่า เราลากเส้นตรงดังในรูปภาพ เส้นตรงนี้จะแสดงความสัมพันธ์ของค่า X และ Y โดยประมาณ ปัญหาว่า เราจะลากเส้นตรงอย่างไรจะเหมาะสมกับข้อมูลที่สุด หรือเส้นไหนจะเป็นเส้นที่ดีที่สุด



รูปที่ 2.1 ภาพกระจัดกระจายแสดงความสัมพันธ์ระหว่าง X และ Y

<u>การปรับเส้นตรงด้วยกำลังสองน้อยที่สุด</u> (Fitting a line by least squares)

ในการลากเส้นแสดงความสัมพันธ์ของ X และ Y เส้นที่เหมาะสมก็ควรจะเป็นเส้นที่ผ่านใกล้ (close) จุดทุก ๆ จุดใน scatter diagram ซึ่งจะทำได้โดยวิธีกำลังสองน้อยที่สุด โดยวัดค่าที่เบี่ยงเบน (diviation) หรือค่าที่ผิดพลาด (error) ไปจากเส้นตรงของแต่ละจุด ต้องมีระยะสั้นที่สุด และเป็นการวัดในแนวดิ่ง ดังรูปที่ ๒.๒



รูปที่ ๒.๒ การวัดส่วนเบี่ยงเบนในแนวดิ่ง

จากรูป e_i คือส่วนเบี่ยงเบนหรือจะเรียกอีกชื่อหนึ่งว่า ส่วนที่เหลือ (Residual) ระหว่างจุด (ข้อมูล) กับเส้นตรง

a คือส่วนตัดแกนตั้ง Y โดยเส้นตรง ซึ่งเรียกว่า Y-intercept

b คือความลาดของเส้น (slope)

$\hat{Y}_i$ คือความสูงของเส้นตรงซึ่งเท่ากับ $a + bX_i$

ดังนั้นส่วนที่คลาดเคลื่อนคือ $e_i = Y_i - \hat{Y}_i$ (ดูรูป)

$$Y_i = \hat{Y}_i + e_i$$

$$Y_i = a + bX_i + e_i$$ (ซึ่งจะใช้เป็นตัวแบบต่อไป)

รวมความคลาดเคลื่อน $e_i (e_i = Y_i - \hat{Y}_i)$ ของแต่ละจุดเข้าด้วยกัน โดยไม่คำนึงถึงเครื่องหมายของความคลาดเคลื่อน ใช้เฉพาะผลต่างของ Y_i กับ $\hat{Y}_i$ ที่ทำให้เกิดความคลาดเคลื่อนเท่านั้น ฉะนั้นผลรวมของความคลาดเคลื่อนจะเท่ากับ

$$\sum_{i=1}^{n} |Y_i - \hat{Y}_i| = \sum_{i=1}^{n} |e_i| = 0$$ (คือค่าน้อยที่สุด)

หรือ อาจจะใช้ความคลาดเคลื่อนยกกำลังสองเพื่อตัดปัญหาเรื่องเครื่องหมายบวกและลบ

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 = 0$$

สรุปจากสมการทั้งสองนี้ได้ว่า

๑. ส่วนเบี่ยงเบนไปจากแนวเส้นตรงข้างมาก (+) และข้างน้อย (-) รวมกันแล้วจะได้เท่ากับ ๐ หรือใกล้ ๐ มากที่สุด

๒. ผลบวกกำลังสองของระยะที่คลาดเคลื่อน จากเส้นต้องน้อยที่สุด

ลักษณะ ๒ ประการนี้จะนำไปสู่เส้นที่ "ดีที่สุด" คือเส้น $Y = a + bX$ จึงเรียกวิธีนี้ว่า วิธีกำลังสองน้อยที่สุด (Least Square Method หรือ Ordinary Least Square Method : OLS)

<u>สูตรของ a และ b</u>

เมื่อ $Y = a + bX$ เป็นเส้นที่ดีที่สุด โดยวิธีกำลังสองน้อยที่สุด จะเขียนได้ว่า

$$\phi = \Sigma(Y_i-\hat{Y}_i)^2 = \Sigma(Y_i-a-bX_i)^2$$

เราทราบค่า X_i และ Y_i (จากข้อมูล) เราจะต้องหาค่า a และ b เมื่อได้ค่า a และ b แล้วก็สามารถจะหาค่าเส้น $\hat{Y}_i = a + bX_i$

การหาค่า a และ b โดยอาศัยหลัก partial derivatives ของ ϕ (ϕ แทนผลบวกของความคลาดเคลื่อนกำลังสอง) มุ่งต่อ a ครั้งหนึ่ง และมุ่งต่อ bครั้งหนึ่ง

$$\frac{\partial\phi}{\partial a} = \frac{\partial(Y_1-a-bX_1)^2}{\partial a} + \frac{\partial(Y_2-a-bX_2)^2}{\partial a} + \ldots + \frac{\partial(Y_n-a-bX_n)^2}{\partial a} = 0$$

$$-2(Y_1-a-bX_1)-2(Y_2-a-bX_2)-\ldots-2(Y_n-a-bX_n) = 0$$

$$-2\sum_1^n(Y_i-a-bX_i) = 0$$

$$\sum_1^n(Y_i-a-bX_i) = 0$$

$$\sum_1^n Y_i = na + b\sum_1^n X_i \quad (1)$$

$$\frac{\partial\phi}{\partial b} = \frac{\partial(Y_1-a-bX_1)^2}{\partial b} + \frac{\partial(Y_2-a-bX_2)^2}{\partial b} + \ldots + \frac{\partial(Y_n-a-bX_n)^2}{\partial b} = 0$$

$$-2(Y_1-a-bX_1)X_1-2(Y_2-a-bX_2)X_2-\ldots-2(Y_n-a-bX_n)X_n = 0$$

$$-2\sum_1^n(Y_i-a-bX_i)X_i \quad 0$$

$$\sum_1^n(Y_i-a-bX_i)X_i = 0$$

$$- \sum_{1}^{n} X_i Y_i = a\sum_{1}^{n} X_i + b\sum_{1}^{n} X_i^2 \qquad (2)$$

สมการ (๑) และ (๒) เรียกว่า Normal equation ตัวประมาณค่า (estimator) a และ b จะหาได้จาก ๒ สมการนี้ โดยการคูณสมการ (๑) ตลอดด้วย $\bar{X}$ แล้วไปหักออกจากสมการ (๒) จะได้

$$(\Sigma X_i - n\bar{X})a + (\Sigma X_i^2 - \bar{X}\Sigma X_i)b = \Sigma X_i Y_i - \bar{X}\Sigma Y_1$$

เนื่องจาก $\Sigma X_i - n\bar{X} = 0$ เราจะหาค่าb ได้

** สูตร $$\hat{b} = \frac{\Sigma X_i Y_i - \bar{X}\Sigma Y_i}{\Sigma X_i^2 - \bar{X}\Sigma X_i}$$

$$= \frac{n\Sigma X_i Y_i - \Sigma X_i \Sigma Y_1}{n\Sigma X_i^2 - (\Sigma X_i)^2}$$

จากสมการ (๑) หารตลอดด้วย n จะได้ค่า a คือ

** สูตร $$\hat{a} = \bar{Y} - b\bar{X}$$

การหาสูตร a , b อีกวิธีหนึ่ง

โดยวิธีนี้จะแปลง X อยู่ในรูปเบี่ยงเบนไปจาก mean คือ $X_i - \bar{X} = x_i$ ซึ่งจะทำให้ $\Sigma x_i = 0$ และจะได้สมการใหม่คือ $\hat{Y}_i = a + bx_i$ สมการใหม่นี้ค่า a จะเปลี่ยนไป ส่วนค่า b จะคงเดิม โดยวิธีกำลังสองน้อยที่สุด ผลบวกกำลังสองของความคลาดเคลื่อนจะน้อยที่สุดคือ

$$\Sigma(Y_i - \hat{Y}_i)^2$$

แทนค่า $\hat{Y}_i$

$$\Sigma(Y_i - a - bx_i)^2$$

โดย partial derivatives มุ่งต่อ a ครั้ง b ครั้ง และตั้งให้เท่ากับ ๐

$$\frac{\partial\Sigma(Y_i - a - bx_i)^2}{\partial a} = \Sigma 2(-1)(Y_i - a - bx_i) = 0$$

-๒ หารตลอดจะได้ $\Sigma Y_i - na - b\Sigma x_i = 0$

แต $\Sigma x = 0$

** สูตร $\hat{a} = \frac{\Sigma Y_i}{n} = Y$

เมื่อ $a = \bar{Y}$ เส้นตรงโดยวิธีกำลังสองน้อยที่สุดจะผ่านจุด $\bar{X}$ และ $\bar{Y}$

$$\frac{a\ \Sigma(Y_i - a - bx_i)^2}{\partial b} = \Sigma 2(-x_i)(Y_i - a - bx_i) = 0$$

$$\Sigma x_i(Y_i - a - bx_i) = 0$$

$$\Sigma Y_i x_i - a\Sigma x_i - b\Sigma x_i^2 = 0$$

แต่ $\Sigma x = 0$

** สูตร $\hat{b} = \frac{\Sigma Y_i x_i}{\Sigma x_i^2}$

a และ b อาจจะหาได้ต่างวิธีออกไปจะไม่กล่าวถึง แต่สรุปรวมเป็นสูตรได้ดังนี้

** สูตร $\hat{a}$(i) $\hat{a} = \frac{\Sigma X_i^2 \Sigma Y_i - \Sigma X_i \Sigma XY_i}{n\Sigma X_i^2 - (\Sigma X_i)^2}$

(ii) $\hat{a} = \bar{Y} - b\bar{X}$ (ตามวิธีข้างต้น)

(iii) $\hat{a} = \frac{\Sigma Y_i}{n} = \bar{Y}$ (ตามวิธีข้างต้น)

** สูตร $\hat{b}$ (i) $\hat{b} = \frac{\Sigma X_i Y_i - n\bar{X}\bar{Y}}{\Sigma X_i^2 - n\bar{X}^2}$

(ii) $$\hat{b} = \frac{\Sigma(X_i-\bar{X})(Y_i-\bar{Y})}{\Sigma(X_i-\bar{X})^2} = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$$

(iii) $$\hat{b} = \frac{\Sigma(X_i-\bar{X})Y_i}{\Sigma(X_i-\bar{X})^2} = \frac{\Sigma x_i Y_i}{\Sigma x_i^2}$$ (จากวิธีข้างต้น)

(เมื่อ $x = X_i-\bar{X}$, $y = Y_i-\bar{Y}$)

(iv) $$\hat{b} = \frac{n\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i}{n\Sigma X_i^2 - (\Sigma X_i)^2}$$ (จากวิธีข้างต้น)

ในกรณีที่สูตร $a = \frac{\Sigma Y_i}{n}$ และ $b = \frac{\Sigma x_i Y_i}{\Sigma x_i^2}$ จะใช้เมื่อ x เป็นส่วนเบี่ยงเบนไปจาก mean ซึ่งจะทำให้ตัวเลขการคำนวณลดลง แต่สมการแสดงความสัมพันธ์ของตัวแปรโดยวิธีกำลังสองน้อยที่สุดจะต้องเปลี่ยนเป็น $\hat{Y} = \hat{a} + \hat{b}x$ ดังนั้นเพื่อจะหาค่าสมการ $\hat{Y} = \hat{a} + \hat{b}X$ ∴ ก็ต้องแทนค่า X ก็จะได้เป็น

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

$$= \hat{a} + \hat{b}X - \hat{b}\bar{X}$$

$$= (\hat{a} + \hat{b}\bar{X}) + \hat{b}X$$

การใช้สูตร $\hat{a} = \bar{Y} - b\bar{X}$ และ(ii) $\hat{b} = \frac{\Sigma xy}{\Sigma x^2}$ จะสะดวกสำหรับการคำนวณที่ไม่ต้องใช้เครื่องคำนวณ และค่า $\hat{a}$, $\hat{b}$ สามารถนำไปแทนค่าในสมการ $\hat{Y} = \hat{a} + \hat{b}X$ ได้ทันที ส่วนสูตรอื่น ๆ ก็อาจจะสะดวกและเหมาะสมแล้วแต่กรณีไป

จากที่กล่าวข้างต้น จะเห็นว่า ความสัมพันธ์ระหว่างตัวแปร เราสามารถจะปรับความสัมพันธ์ของมันเป็นสมการเส้นตรงโดยวิธีกำลังสองน้อยที่สุด และความสัมพันธ์นี้อาจจะศึกษาต่อไปได้ทั้งในแง่ของเส้นถดถอย (Regression) และสหสัมพันธ์ (Correlation)

เส้นถดถอย (Regression line)

เส้นถดถอย คือ เส้นที่แสดงแนวความสัมพันธ์เฉลี่ย (line of average relationship) ระหว่างจุดต่าง ๆ บน scatter diagram ซึ่งจุดเหล่านี้จะถดถอยเข้าสู่เส้น จุดต่างๆ บน scatter diagram ก็คือข้อมูลที่เราจะศึกษาซึ่งมีอยู่เป็นจำนวนมาก เส้นถดถอยจะแสดงให้เห็นถึงการเปลี่ยนแปลงของข้อมูลที่เป็นตัวแปรตาม (Y) ร่วมไปกับข้อมูลที่เป็นตัวแปรอิสระ (X) ที่กำหนดให้ ความสัมพันธ์เฉลี่ยนี้หาได้โดยวิธีกำลังสองน้อยที่สุด

เส้นถดถอยมีทั้งชนิดเส้นตรง (linear regression) และชนิดเส้นโค้ง (curvilinear regression) เส้นถดถอยชนิดเส้นตรงเป็นเส้นที่เกิดจากฟังชั่นที่มีตัวแปรอิสระกำลังเป็นหนึ่ง เส้นถดถอยชนิดนี้มีอยู่ ๒ ประเภทคือ

๑. เส้นถดถอยอย่างง่าย (Simple regression) เป็นเส้นถดถอยที่เกิดจากความสัมพันธ์ระหว่าง ๒ ตัวแปร : $Y = f(X)$ โดยที่ตัวแปรหนึ่งเป็นตัวแปรตามหรือตัวแปรภายใน (Y) และตัวแปรอีกตัวหนึ่ง (X) เป็นตัวแปรอิสระหรือตัวแปรภายนอก

๒. เส้นถดถอยเชิงซ้อน (Multiple regression) เป็นเส้นถดถอยที่เกิดจากฟังชั่นที่แสดงความสัมพันธ์ระหว่างตัวแปรตั้งแต่ ๓ ตัวขึ้นไป : $Y = f(X_1, X_2, X_3)$ โดยมีตัวแปรหนึ่งเป็นตัวแปรตาม ตัวแปรอื่น ๆ เป็นตัวแปรอิสระ

การคำนวณหรือการประมาณค่า (estimation) ของ regression เส้นตรงมีหลายวิธีคือ

๑. วิธีกำลังสองน้อยที่สุดธรรมดา Ordinary Least Square : OLS)

๒. วิธี Maximum Likelihood Estimation : MLS'

A. วิธีกำลังสองน้อยที่สุดสองขั้น (Two Stage Least Square : 2SLS)

๔. วิธีกำลังสองน้อยที่สุดสามขั้น (Three Stage Least Square : 3SLS)

สำหรับการศึกษา regression ในขั้นนี้ ส่วนใหญ่จะเป็นการประมาณค่าโดยวิธีกำลังสองน้อยที่สุดธรรมดา (OLS) ซึ่งเป็นวิธีที่ง่ายใช้กันแพร่หลายและเป็นวิธีที่ให้ค่าไม่ลำเอียง โดยมีความแปรปรวนน้อยที่สุด (minimum variance) และจะกล่าวถึงวิธีกำลังสองน้อยที่สุดสองขั้น (2SLS) อย่างย่อ ๆ

นอกจากการประมาณค่าทั้ง ๔ วิธีข้างต้นแล้ว เส้น regression ยังอาจจะคำนวณได้โดยวิธีกึ่งเฉลี่ย (Semi-average) แต่วิธีนี้จะให้ค่าความแปรปรวนมากกว่าวิธีกำลังสองน้อยที่สุดโดยเฉพาะค่าความลาด (b) จะห่างจากความเป็นจริงมาก ฉะนั้น จึงจะไม่กล่าวถึงวิธีนี้ในที่นี้

<u>ตัวแบบของเส้นถดถอยอย่างง่าย</u> **(Simple linear regression model)**

ตัวแบบเส้นถดถอยจากประชากร : $\mu_{YX} = \alpha + \beta X_i + u_i$

ตัวแบบเส้นถดถอยจากตัวอย่าง : $Y_i = a + bX_i + e_i$

การหาความสัมพันธ์ระหว่าง ๒ ตัวแปรจำเป็นต้องเริ่มต้นด้วยตัวแบบ (model) ที่คิดว่าเป็นไปได้ซึ่งสามารถจะอธิบายความสัมพันธ์ได้ ต่อจากการตั้ง model ก็จะเป็นการประมาณค่าพารามิเตอร์และการทดสอบสมมติฐานต่าง ๆ เกี่ยวกับความสัมพันธ์นี้

ถ้าเป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรของประชากร ตัวแบบที่นิยมเขียนจะเป็นแบบสมการแรกข้างบนคือ

$$\mu_{YX} = \alpha + \beta X_i + u_i$$

μ_{YX} = ค่าเฉลี่ยของตัวแปรตามของเส้นถดถอย

X_i = ตัวแปรอิสระ

α = ส่วนตัดแกนตั้ง (Y – intercept) เป็นค่าคงที่ (contant)

β = ความลาด (slope) ของเส้นถดถอย เรียกว่า สัมประสิทธิ์การถดถอย (Regression Coefficient) ทั้ง $\mu_{Y.X}$ α, β เป็นพารามิเตอร์ที่ไม่ทราบค่า

u_i = ส่วนคลาดเคลื่อน (disturbance) จากเส้นถดถอย เป็นตัวแปรอิสระสุ่ม และไม่ทราบค่า

ตัวแบบนี้เป็นตัวแบบที่แสดงความสัมพันธ์ที่แท้จริง (true relation) ของประชากร แต่ในทาง Statistical Inference เรามักใช้ตัวอย่าง(sample) มากกว่าจะใช้ข้อมูลจากประชากรทั้งหมด ด้วยเหตุผลหลายประการดังกล่าวแล้วในบทแรก ตัวแบบเส้นถดถอยอย่างง่ายจากตัวอย่างคือ

$$Y_i = a + bX_i + e_i$$

Y_i = ค่าเฉลี่ยของตัวแปรตามของเส้นถดถอย

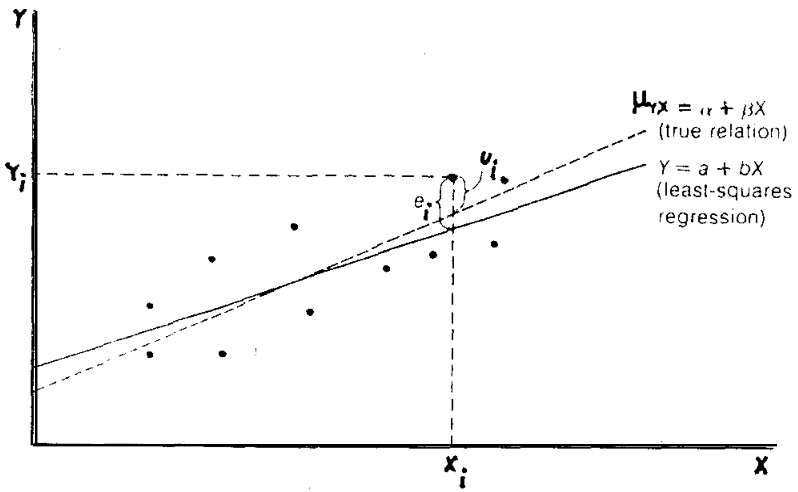
X_i = ตัวแปรอิสระ

a = ส่วนตัดแกนตั้ง หรือค่าค่าคงที่ (constant)

b = ความลาด (slope) ของเส้นถดถอย เรียกว่า Regression Coefficient)
ทั้ง Y_i , a และ b เป็นตัวประมาณค่าพารามิเตอร์

e_i = เป็นความคลาดเคลื่อน (error) หรือค่าที่เหลืออยู่ (residual) ของเส้นถดถอย

จากข้อมูลของตัวอย่าง การปรับเส้นโดยวิธี OLS จะให้เส้น regression ที่ดีที่สุด ดังรูปที่ ๒.๓ เส้นนี้จะเป็นเส้นที่ประมาณ (estimate) ความสัมพันธ์ Y กับ X โดยมี a และ b อธิบายตัวประมาณค่าพารามิเตอร์ α และβ ดังนั้น จะเห็นว่า การศึกษา regression โดยทั่วไปส่วนใหญ่จะใช้ model ของตัวอย่าง $Y_i = a + bX_i + e_i$



รูปที่ 2.3 ความสัมพันธ์ระหว่าง Y และ X โดยกำลังสองน้อยที่สุด

<u>ข้อสมมติพื้นฐาน</u> (Basic assumptions)

ในการวิเคราะห์เส้นถดถอยโดยวิธี OLS จะต้องมีข้อสมมติซึ่งถือว่ามีความสำคัญมากในการที่ทำให้ตัวประมาณค่า (estimator) มีคุณสมบัติที่ดี เป็นข้อสมมติเกี่ยวกับความคลาดเคลื่อน (error or disturbanc : u_i) ของเส้นถดถอยคือ

๑. meanของ u_i แต่ละตัวเป็น ๐

$$E(u_i) = 0$$

๒. u_i เป็นอิสระซึ่งกันและกัน เป็นผลให้

$$Cov.(u_i,u_j) = E(u_i,u_j) = 0 \quad (\text{ถ้า } i \neq j)$$

๓. u_i ทุกตัวมีความแปรปรวน (Variance) เหมือนกัน

$$Var.(u_i) = E(u_i^2) = \sigma_u^2$$

๔. X_i เป็นตัวแปรที่ไม่สุ่ม (ตัวแปรที่กำหนดให้หรือคงที่) ผลก็คือ ถ้า ϕ เป็นฟังชั่นของ $X_1, X_2, \ldots X_n$

$$E(\phi, u_i) = \phi E(u_i) = 0$$

<u>ตัวประมาณค่า a,b</u> (Estimator a,b)

ตามทฤษฎีของเกาส์-มาร์คอบ (The Gauss Markov Theorem) กล่าวว่า ในพวก linear unbiased estimators ของ $\dot{\alpha}$ และ β ทั้งหลาย estimator ของ least square มี variance น้อยที่สุด หมายความว่าการคำนวณเส้นถดถอยโดยวิธีกำลังสองน้อยที่สุด เป็นวิธีที่ให้ค่าสัมประสิทธิหรือตัวคงที่ a และ b ที่ไม่ลำเอียงโดยให้ค่าความแปรปรวนน้อยที่สุด และ a,b ถือว่าเป็น best linear unbiased estimator (BLUE)

ดังนั้น ค่า a และ b จะหาได้โดยใช้สูตร $\hat{a}$ และ $\hat{b}$ ตามวิธีกำลังสองน้อยที่สุดในหน้า ๑๙ และ ๒๐ เช่น $\hat{a} = \bar{Y} - b\bar{X}$, $\hat{b} = \frac{\Sigma xy}{\Sigma x^2}$ และอีกหลายสูตรตามแต่จะต้องการใช้สูตรใด

เมื่อประมาณค่า a,b ได้ เส้นถดถอยอย่างง่ายก็จะประมาณออกมาได้

<u>ตัวอย่างที่ 2.1</u>

กำหนดข้อมูลราคาส้มกับปริมาณการเสนอซื้อส้มของผู้บริโภคจากตลาดแห่งหนึ่ง จงคำนวณสมการการเสนอซื้อ โดยวิธีกำลังสองน้อยที่สุด

P (10 / ก.ก.) บาท	2	3	4	5	6
Q_d (ก.ก.)	12	7	8	5	3

<u>วิธีทำ</u>

๑. ตัวแบบ $Q_d = a + bP + e$

เพื่อมิให้สับสน เปลี่ยน Q_d เป็น Y และ P เป็น X ตามตัวแบบทั่วไปก่อนคือ $Y = a + bX + e$ แล้วคำนวณหาค่า a,b ตามวิธี OLS เสร็จแล้วจึงเปลี่ยนกลับมาเป็นสมการ demand ตามเดิม

๒. การคำนวณ สร้างตารางเพื่อหาค่าสัมประสิทธิ a,b

ตารางที่ 2.2

Y	X	x	y	xy	x^2	$\hat{Y}$
๑๒	๒	-๒	๕	-๑๐	๔	๑๑
๗	๓	-๑	๐	๐	๑	๙
๘	๔	๐	๑	๐	๐	๗
๕	๕	๑	-๒	-๒	๑	๕
๓	๖	๒	-๔	-๘	๔	๓
๓๕	๒๐	๐	๐	-๒๐	๑๐	

$\bar{Y} = 7 \quad \bar{X} = 4$

$$\hat{b} = \frac{\Sigma xy}{\Sigma x^2}$$

$$= \frac{-๒๐}{๑๐}$$

$$= -๒$$

$$\hat{a} = \bar{Y} - b\bar{X}$$

$$= ๗ - (-๒)(๔)$$

$$= ๑๕$$

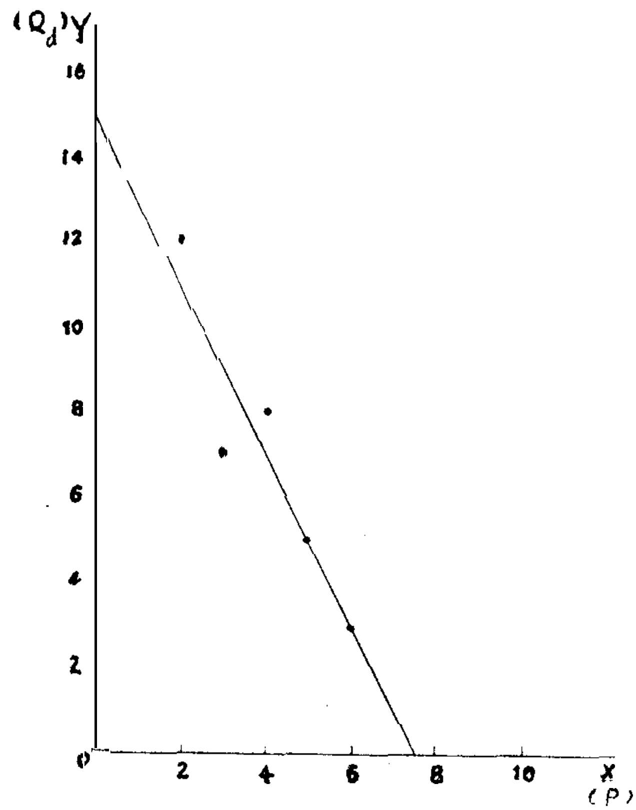
๓. สมการ regression คือ

$$Y = ๑๕ - ๒X$$

d. **หรือสมการ demand คือ**

$$\hat{Q}_d = ๑๕ - ๒P$$

เส้น demand ที่ได้มีจุดตัดแกนตั้งเท่ากับ ๑๕ และมีความลาดเท่ากับ -๒ ความลาดเป็นลบซึ่งเป็นไปตามกฎของ demand ในทฤษฎีเศรษฐศาสตร์ เส้น demand จะเป็น negative ดังรูปที่ ๒.๔



รูปที่ ๒.๔ เส้น demand

การแปลความหมายของ $\hat{a}$ และ $\hat{b}$ ในสมการ regression

a คือส่วนตัดแกนตั้ง ซึ่งเมื่อ X = O (P = O) $\hat{a}$ จะให้ค่าประมาณของ $\hat{Y}$ จากตัวอย่างข้างบน P เป็นราคา Q_d เป็นปริมาณ เมื่อราคาเป็น O ปริมาณที่ต้องการจะ ($\hat{Y}$) เท่ากับ ๑๕ กิโลกรัม

b เป็นสัมประสิทธิ์การถดถอย (Regression Coefficient) คือ จำนวน Y ที่เปลี่ยนไป (เพิ่มขึ้นหรือลดลงโดยเฉลี่ย) เมื่อ X เปลี่ยนไป 1 หน่วย จากตัวอย่างนี้ หมายถึง เมื่อราคาส้มเพิ่มขึ้นกิโลกรัมละ 10 บาท ปริมาณการเสนอซื้อจะลดลง 2 กิโลกรัม

<u>ตัวอย่างที่</u> 2.2

จงประมาณค่าความสัมพันธ์ในเชิงเศรษฐกิจระหว่างการบริโภค (Consumption) กับรายได้ (Disposible Income) ของประชาชนระหว่างปี ๒๕๐๐-๒๕๑๙ ดังตารางข้างล่าง[2]

หน่วย : พันล้านบาท

C	๓๒๕	๓๓๕	๓๕๕	๓๗๕	๔๐๑	๔๓๓	๔๖๖	๔๙๒	๕๓๗	๕๗๖
Y_d	๓๕๐	๓๖๔	๓๘๕	๔๐๕	๔๓๘	๔๗๓	๕๑๒	๕๔๗	๕๙๐	๖๓๐

วิธีทำ

๑. ตัวแบบ $\hat{C} = a + bY_d + e$

ตัวอย่างนี้จะให้นักศึกษาเห็นถึงการใช้สัญลักษณ์ตามตัวแบบ

๒. การคำนวณ

ตารางที่ 2.3

C	Y_d	$(C - \bar{C})$	$(Y_d - \bar{Y}_d)$	$(C - \bar{C})(Y_d - \bar{Y}_d)$	$(Y_d - \bar{Y}_d)^2$
๓๒๕	๓๕๐	-๑๐๕	-๑๑๙	๑๒,๔๙๕	๑๔,๑๖๑
๓๓๕	๓๖๔	-๙๕	-๑๐๕	๙,๙๗๕	๑๑,๐๒๕
๓๕๕	๓๘๕	-๗๕	-๘๔	๖,๓๐๐	๗,๐๕๖
๓๗๕	๔๐๕	-๕๕	-๖๔	๓,๕๒๐	๔,๐๙๖
๔๐๑	๔๓๘	-๒๙	-๓๑	๘๙๙	๙๖๑

เป็นข้อมูลอนุกรมเวลา (Time series data) แต่ regression เป็นการวิเคราะห์โดยใช้ข้อมูลเกี่ยวพัน (cross-sectional analysis) จึงกำหนดให้เวลาคงที่(hold time constant)

C	Y_d	$(C - \bar{C})$	$(Y_d - \bar{Y}_d)$	$(C - \bar{C})(Y_d - \bar{Y}_d)$	$(Y_d - \bar{Y}_d)^2$
433	473	3	4	12	16
466	512	36	43	1,548	1,849
492	547	62	78	4,839	6,084
537	590	107	121	12,948	14,641
576	630	146	161	23,506	25,921
4,295	4,694			76,038	85,810

$\bar{C}$ = ๔๓๐ $\bar{Y}_d$ = ๔๖๙

$$\hat{b} = \frac{\Sigma (C - \bar{C})(Y_d - \bar{Y}_d)}{\Sigma (Y_d - \bar{Y}_d)^2}$$

= ๗๖,๐๓๘ / ๘๕,๘๑๐

= ๐.๘๙

$$\hat{a} = \bar{C} - \hat{b}\bar{Y}_d$$

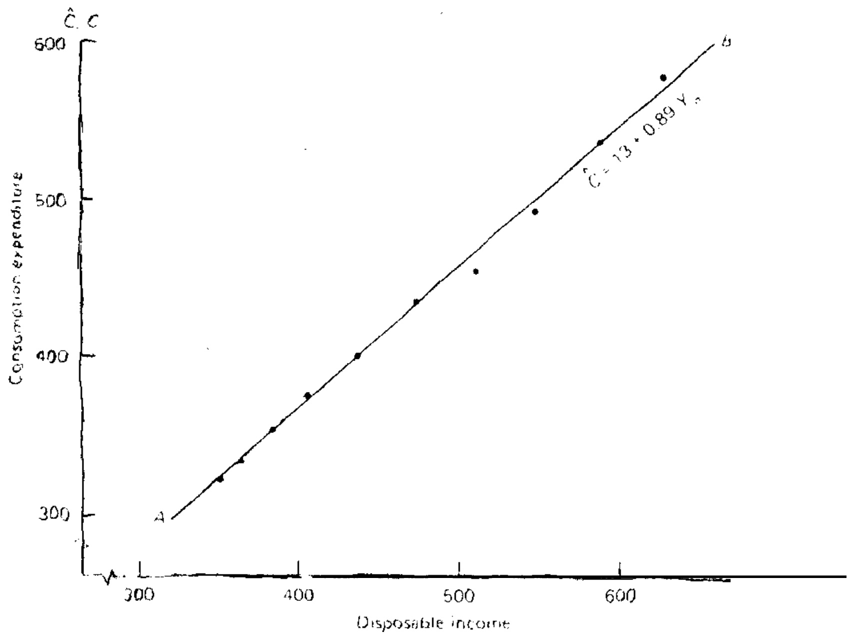
= ๔๓๐ - ๐.๘๙ (๔๖๙)

๑๓

๓. สมการ regression ของการบริโภคกับรายได้คือ

$\hat{C}$ = ๑๓ + ๐.๘๙Y_d

เมื่อ plot กราฟตามสมการนี้ จะได้เส้นถดถอย AB ซึ่งผ่านจุด 10 จุด ที่กระจัดกระจายอยู่ดังในรูปที่ 2.5



รูปที่ 2.5 ความสัมพันธ์ของการบริโภคกับรายได้โดยวิธีกำลังสองน้อยที่สุด

การตีความค่า b ของเส้น regression ที่แสดงความสัมพันธ์ระหว่างการบริโภคและรายได้ b ก็คือความโน้มเอียงในการบริโภค[2] (Marginal Propensity

---

[2] $C = 13 + 0.89Y_d \quad , \quad \frac{\partial C}{\partial Y_d} = 0.89 = \hat{b}$

$\frac{\partial C}{\partial Y_d}$ คือ $\frac{\text{การเปลี่ยนแปลงในการบริโภคที่เพิ่มขึ้น}}{\text{การเปลี่ยนแปลงในรายได้ที่เพิ่มขึ้น}}$ (หรือ $= \frac{\Delta C}{\Delta Y_d}$)

$= \text{MPC} = \hat{b}$

to consume : MPC) ซึ่งเท่ากับ 0.89 หมายความว่าเมื่อรายได้เพิ่มขึ้น 1 หน่วย การบริโภคจะเพิ่มขึ้น 0.89 หน่วย

เส้นถดถอยที่ประมาณได้ จะให้การประมาณค่าที่ดีของความสัมพันธ์ระหว่างการบริโภคและรายได้ ความสัมพันธ์นี้จะเป็นไปตามที่คาดหมายในเชิงทฤษฎี ซึ่งค่าประมาณของ MPC (= 0.89) เป็นบวกและมีค่าระหว่าง 0 ถึง 1 ค่าประมาณของส่วนตัดแกนตั้ง (= 13) ก็เป็นบวกเช่นเดียวกันคือ แม้รายได้ (Y_d) จะเป็น 0 การบริโภค ($\hat{C}$) ก็ยังมีอยู่ ซึ่งก็เป็นความจริง แม้คนจะไม่มีรายได้เลยก็ยังต้องบริโภค เรียก autonomous consumption

<u>ตัวอย่างความสัมพันธ์ระหว่าง 2 ตัวแปรในเชิงเศรษฐศาสตร์</u>

ฟังชั่นในทางเศรษฐศาสตร์จุลภาค เช่น

ผลผลิต = f(ปัจจัยการผลิต)หรือ Y = f(ทุน k) , Y = f(แรงงาน L)

$Y = a + bX$, $\partial Y/\partial X = b$ = Marginal Physical Product of X(MPP_X)

$Y = a + bK$, $\partial Y/\partial K = b$ = Marginal Efficiency of Capital (MEC)

หรือ Marginal Productivity of Capital

$Y = a + bL$, $\partial Y/\partial L = b$ = Marginal Productivity of Labor

ฟังชั่นในทางเศรษฐศาสตร์มหภาค

$C = a + bY$, $\partial C/\partial Y = b$ = MPC

$S = a + bY$, $\partial S/\partial Y = b$ = MPS

$I = a + br$, $\partial I/\partial r = b$ = MPI

x = a + bY , ax/au = b = MPX

<u>ค่าสถิติที่จำเป็นของเส้นถดถอย</u>

วัตถุประสงค์ของการคำนวณเส้นถดถอยขึ้นมาก็เพื่อนำไปใช้พยากรณ์ แต่เส้นถดถอยที่คำนวณได้อาจจะไม่สามารถนำไปใช้ได้เสมอทุกเส้น ถ้าเส้นที่ใช้ไม่ได้หากนำไปใช้พยากรณ์จะให้ค่าที่ผิดพลาด เพื่อที่จะให้แน่ใจว่าเส้นถดถอยที่คำนวณได้เหมาะที่จะนำไปใช้พยากรณ์หรือไม่ จำเป็นจะต้องคำนวณค่าสถิติต่าง ๆ เพื่อช่วยในการตัดสินใจว่าจะใช้หรือไม่ใช้เส้นถดถอยนั้นต่อไป ค่าสถิติที่ต้องคำนวณมีดังนี้คือ

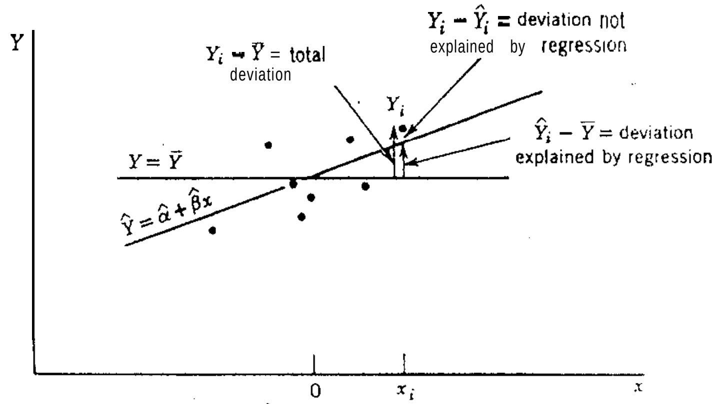
1. สัมประสิทธิการตัดสินใจ (Coefficient of Determination)
2. สัมประสิทธิการตัดสินใจปรับปรุง (Adjusted Coefficient of Determination)
3. ความคลาดเคลื่อนมาตรฐานของ a และ b (Standard Error of a and b)
4. ค่าสถิติ t และการทดสอบ (t-test)
5. ค่าสถิติ F และการทดสอบ (F-test)
6. ความคลาดเคลื่อนมาตรฐานของการประมาณ (Standard Error of Estimate)

1. <u>สัมประสิทธิการตัดสินใจ</u> (Coefficient of Determination : R^2)

R^2 หมายถึง เปอร์เซ็นต์ความแปรปรวนของตัวแปรตามซึ่งถูกอธิบายโดยตัวแปรอิสระ เช่น สมการถดถอย $Y = a + bX$, R^2 จะบอกถึงอัตราส่วนของความแปรปรวนใน Y ซึ่งมีสาเหตุมาจากความแปรปรวนใน X หรือจะกล่าวว่า X มีอิทธิพลต่อ Y มากน้อยเท่าไร ถ้ามองในแง่เส้นถดถอย R^2 คือค่าที่แสดงให้ทราบว่าเส้นถดถอยที่คำนวณได้ ($\hat{Y} = \hat{a} + \hat{b}X$) ฟิตกับข้อมูลที่เป็นจริงเพียงไร

เนื่องจากเราไม่คิดว่า เส้นตรงโดยวิธีกำลังสองน้อยที่สุดจะผ่านทุกจุดของข้อมูล

การวัดว่าเส้นมีความใกล้ (closeness) เหมาะสมกับข้อมูลหรือไม่จะมีประโยชน์ต่อการตัดสินใจที่จะใช้เส้นถดถอยเพื่อการพยากรณ์ ถ้าเส้นผ่านใกล้จุดทุก ๆ จุดมาก เส้นก็จะอธิบายความสัมพันธ์ระหว่างตัวแปรได้ค่อนข้างถูกต้องแน่นอน แต่ถ้าความเบี่ยงเบนหรือความคลาดเคลื่อนมีมาก เส้นก็จะอธิบายความสัมพันธ์ได้น้อย



รูปที่ 2.6 ค่าของ regression ในการลดความแปรปรวนใน Y

โดยวิธีกำลังสองน้อยที่สุด

$$\Sigma e_i^2 = \Sigma(Y_i - \hat{Y})^2$$ = ความแปรปรวนที่ไม่อาจอธิบายได้ (Unexplained Variation)

$$= \Sigma(Y_i - a - bX_i)^2$$

เพราะ a $= \bar{Y} - b\bar{X}$ แทนค่า a

$$\Sigma e_i^2 = \Sigma(Y_i - \bar{Y} + b\bar{X} - bX_i)^2$$

$$= \Sigma\left[(Y_i - \bar{Y}) - b(X_i - \bar{X})\right]^2$$

$$= \Sigma\left[(Y_i - \bar{Y})^2 - 2b(X_i - \bar{X})(Y_i - \bar{Y}) + b^2(X_i - \bar{X})^2\right]$$

$$= \Sigma(Y_i - \bar{Y})^2 - 2b\Sigma(X_i - \bar{X})(Y_i - \bar{Y}) + b^2\Sigma(X_i - \bar{X})^2$$

เพราะ $b = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$ ฉะนั้น $b\Sigma(X_i - X)^2 = \Sigma(X_i - \bar{X})(Y_i - \bar{Y})$