

วิธีการหาค่าสัมประสิทธิ์ของสมการถดถอยแบบง่ายด้วยวิธีกำลังสองน้อยที่สุด
(Method for finding simple regression coefficient by least square error)

สันทัต พรประเสริฐมานิต

ในการทำนายค่าของตัวแปรตาม (Predicted variable; criterion) ด้วยตัวแปรต้น (Predictor) นั้น จะต้องแปลงค่าของตัวแปรต้นด้วยการแปลงเชิงเส้น (Linear transformation) เพื่อให้สามารถทำให้ความแตกต่างระหว่างค่าของตัวแปรต้นและค่าของตัวแปรตามน้อยที่สุด หรือพูดอีกนัยหนึ่งว่าให้มีความผิดพลาดในการทำนาย (Error of prediction) น้อยที่สุดนั่นเอง

วิธีการแปลงเชิงเส้นจะเป็นการคูณตัวแปรต้นด้วยค่าคงที่ แล้วบวกด้วยค่าคงที่อีกตัวหนึ่ง สามารถเขียนได้รูปดังนี้

$$\hat{Y} = a + bX \quad \text{—————} \boxed{1}$$

X คือ ค่าตัวแปรต้น

$\hat{Y}$ คือ ค่าตัวแปรตามที่ทำนายได้จากการแปลงเชิงเส้นตัวแปรต้น

a คือ ค่าของตัวแปรตามที่ทำนายได้เมื่อทำให้ค่าของตัวแปรต้นเท่ากับศูนย์ หรือเรียกอีกอย่างว่าจุดตัดแกน Y (Y-intercept)

b คือ ค่าการเปลี่ยนแปลงของค่าตัวแปรตามที่ทำนายได้ที่เกิดจากการเปลี่ยนแปลงค่าของตัวแปรต้นไป 1 หน่วย หรือเรียกอีกอย่างว่าความชัน (slope) ของสมการ

ส่วนค่าความผิดพลาดในการทำนายสามารถหาได้จากความแตกต่างระหว่างค่าตัวแปรตามที่เกิดขึ้นได้ และค่าตัวแปรตามที่ทำนายได้จากการแปลงเชิงเส้นตัวแปรต้น ดังสมการ

$$e = Y - \hat{Y} \quad \text{—————} \boxed{2}$$

e คือ ค่าความผิดพลาดในการทำนายค่าตัวแปรตามที่เกิดขึ้นได้

Y คือ ค่าตัวแปรตามที่เกิดขึ้นได้

เมื่อรวมสมการทั้งสองเข้าด้วยกัน จะกล่าวได้ว่าค่าตัวแปรตามที่เกิดขึ้นได้จะเกิดจากการรวมกันระหว่างจุดตัดแกน Y ค่าความชันคูณกับค่าของตัวแปรต้น และค่าความผิดพลาดในการทำนายดังสมการ

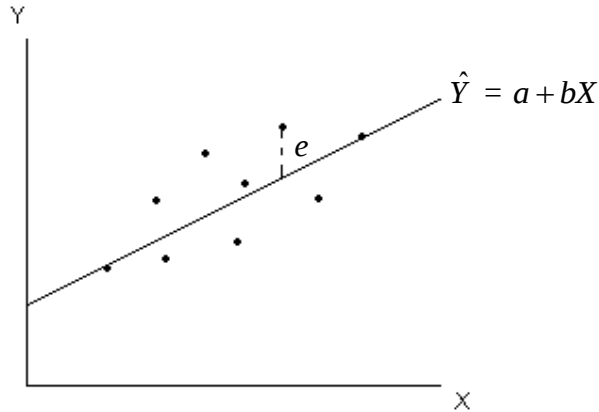
$$Y = a + bX + e \quad \text{—————} \boxed{3}$$

จากที่กล่าวมาข้างต้น สามารถอธิบายความสัมพันธ์ของค่าต่างๆ ในการทำนายตัวแปรตามด้วยตัวแปรต้นได้ดังภาพการกระจายกระจัดกระจาย (Scatterplot) เมื่อนำตัวแปรต้นอยู่ในแกน X และตัวแปรตามอยู่ในแกน Y ได้ดังภาพที่ 1

จุดต่างๆ จะแสดงค่าตัวแปรต้นและตัวแปรตามที่เกิดขึ้นได้ ส่วนเส้นในภาพ จะเป็นเส้นที่เกิดจากสมการถดถอย และสมการถดถอยนี้ก็มาจากการแปลงเชิงเส้นของตัวแปรต้นนั่นเอง ค่าตัวแปรตามที่ทำนายได้จะเป็นจุดต่างๆ ที่อยู่บนเส้น ระยะห่างที่ขนานกับแกน Y ระหว่างค่าของตัวแปรตามที่เกิดขึ้นได้ และค่าตัวแปรตามที่ทำนายได้จะเท่ากับค่าความผิดพลาดในการทำนาย

Author Note

This article was written in September 2007 for teaching in Introduction to Statistics in Psychology Class, Faculty of Psychology, Chulalongkorn University
Correspondence to Sunthud Pornprasertmanit. Email: psunthud@gmail.com



ภาพที่ 1 แสดงความสัมพันธ์ระหว่างค่าต่างๆ ในการทำนายตัวแปรตาม (Y) ด้วยตัวแปรต้น (X)

อย่างไรก็ตามก็มีคำถามว่าค่าของ a และ b ที่เหมาะสมที่สุดคืออะไร หรือกล่าวได้อีกนัยหนึ่งว่าแปลงเชิงเส้นตัวแปรต้นอย่างไร ให้ทำนายตัวแปรตามได้ดีที่สุด วิธีการแรกคือทำให้ผลรวมของค่าความผิดพลาดในการทำนายน้อยที่สุด ถ้าทำให้ค่าความผิดพลาดรวมกันเท่ากับศูนย์ได้ จะทำให้การทำนายถูกต้องแม่นยำมากที่สุด อาจเขียนในรูปสมการได้คือ $\sum e = \sum (Y - \hat{Y}) = 0$ นั่นเอง ต่อไปจะนำค่านี้มาหาค่าจุดตัดแกน Y และค่าความชันของสมการดังนี้

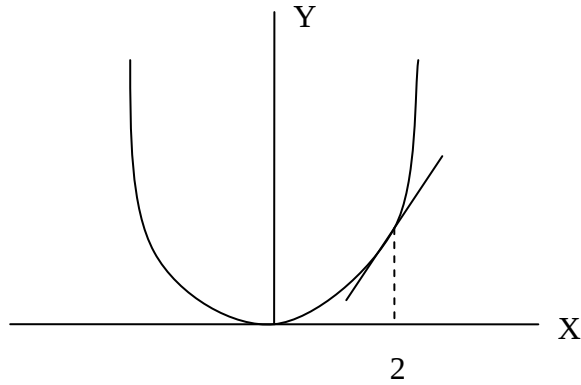
$$\begin{aligned} \sum (Y - \hat{Y}) &= 0 \\ \text{จากสมการ 1} \quad \sum (Y - a - bX) &= 0 \\ \sum Y - \sum a - \sum bX &= 0 \\ \sum Y - na - b\sum X &= 0 \\ \text{หาร } n \text{ ตลอด} \quad \bar{Y} - a - b\bar{X} &= 0 \end{aligned}$$

$$\bar{Y} = a + b\bar{X}$$

4

จากสมการผลลัพธ์จะเห็นได้ว่า เส้นสมการใดก็ได้ ที่ผ่านจุดพิกัด $(\bar{X}, \bar{Y})$ ก็จะทำให้ค่าความผิดพลาดในการทำนายตัวแปรตามด้วยตัวแปรต้นทุกตัวรวมกันเท่ากับศูนย์ แต่เนื่องจากเส้นที่ผ่านจุด $(\bar{X}, \bar{Y})$ ไม่น่าจะเป็นเส้นสมการที่ทำนายได้ดีที่สุดดังภาพที่ 2 ก และ 2ข จะเห็นว่าเส้นสมการถดถอยในภาพที่ 2ก น่าจะทำนายได้ดีกว่า 2ข เนื่องจากในรูป 2ก จุดต่างๆ เกะกะกลุ่มรอบเส้นมากกว่า ทำให้มีค่าความผิดพลาดในการทำนายน้อยกว่า แสดงว่าคำตอบจากการทำให้ค่าความผิดพลาดในการทำนายรวมกันเท่ากับศูนย์จึงไม่ใช่คำตอบที่ดีที่สุด

Gauss และ Markov ได้ร่วมกันหาวิธีที่จะหาเส้นทำนายที่ดีที่สุด วิธีการประมาณค่าแบบนี้ในปัจจุบันเรียกว่าวิธีกำลังสองน้อยที่สุดแบบดั้งเดิม (Ordinary Least Square: OLS) วิธีนี้มีหลักการอยู่ว่าเนื่องจากค่าความคลาดเคลื่อนในการทำนายอาจมีค่าติดลบได้ ทำให้เส้นที่ผ่านจุดพิกัด $(\bar{X}, \bar{Y})$ สามารถมีผลรวมของความคลาดเคลื่อนเท่ากับศูนย์ได้ แต่ถ้าไม่ทำให้ค่าของความคลาดเคลื่อนติดลบจะทำให้ผลรวมของค่านี้มีค่าความหมายมากขึ้น ด้วยเหตุนี้จึงนำค่าความคลาดเคลื่อนในการทำนายมายกกำลังสอง เพื่อไม่ให้มีค่าติดลบ แล้วนำค่าความคลาดเคลื่อนกำลังสองนี้มารวมกัน หรือเขียนเป็นรูปสมการได้ว่า $\sum e^2 = \sum (Y - \hat{Y})^2$ ค่านี้อาจเรียกได้ว่าผลรวมกำลังสองของความคลาดเคลื่อน (Sum of square error: SS_{error}) ถ้าเส้นสมการใดที่สามารถทำให้ค่านี้มีค่าน้อยที่สุด น่าจะเป็นเส้นสมการที่ทำนายได้ดีที่สุด แต่ก่อนจะขึ้นถึงวิธีการหา ต้องมีความรู้เรื่องแคลคูลัสเล็กน้อย ซึ่งจะอธิบายโดยคร่าวดังนี้



รูปที่ 3 การหาความชันของกราฟ $Y = X^2$ ของจุด $X = 2$

สรุปสูตรการหาอนุพันธ์มีดังนี้ (ในที่นี้จะแสดงเฉพาะสูตรที่ใช้ในการพิสูจน์เท่านั้น)

1. $\frac{\partial c}{\partial X} = 0$ เมื่อ c คือค่าคงที่
2. $\frac{\partial X}{\partial X} = 1$
3. $\frac{\partial}{\partial X} cU = c \frac{\partial U}{\partial X}$
4. $\frac{\partial}{\partial X} X^n = nX^{n-1}$
5. $\frac{\partial Y}{\partial X} = \frac{\partial Y}{\partial U} \cdot \frac{\partial U}{\partial X}$ กฎลูกโซ่
6. $\frac{\partial}{\partial X} \sum Y = \sum \frac{\partial}{\partial X} Y$

ถ้าอยากหารายละเอียดเพิ่มเติมสามารถหาได้ในหนังสือของ Weir, Hass & Giordano (2005)

หรือหนังสือแคลคูลัสทั่วไปในบทเรื่องการหาอนุพันธ์

เมื่อทราบวิธีการหาอนุพันธ์แล้ว ต่อไปจะกล่าวถึงวิธีการหาค่าต่ำสุดและสูงสุด เมื่อลองสังเกตรูปที่ 3 จะเห็นว่าเส้นโค้งด้านซ้ายจะมีความชันเป็นลบ แล้วจะลดลงเรื่อยๆ จนกระทั่งจุดที่ $X = 0$ จะมีความชันเท่ากับศูนย์พอดี แล้วความชันจะเป็นบวก แล้วเพิ่มขึ้นเรื่อยๆ จะเห็นว่าจุดที่มีความชันเท่ากับศูนย์จะเป็นจุดการเปลี่ยนการเพิ่มหรือการลดของสมการ จะทำให้จุดดังกล่าวมีโอกาสที่จะเป็นจุดต่ำสุดหรือจุดสูงสุดสูง ด้วยเหตุนี้จะใช้ความรู้เรื่องนี้ในการหาค่าต่ำสุดของผลรวมกำลังสองของความคลาดเคลื่อน

ถ้าลองสังเกต จะพบว่ายิ่งลากเส้นทำนายห่างจากจุดพิกัด $(\bar{X}, \bar{Y})$ มากเท่าไร จะยิ่งทำให้ค่า SS_{error} มีค่าสูง ดังนั้นน่าจะมีจุดๆ หนึ่งที่ทำให้ค่า SS_{error} มีค่าน้อยที่สุด ต่อไปนี้จะใช้ความรู้ในการหาอนุพันธ์ หาค่า a และค่า b ได้ที่ทำให้ค่า SS_{error} มีค่าน้อยที่สุด กล่าวคือจะทำให้อนุพันธ์อันดับหนึ่งมีค่าเท่ากับศูนย์ดังสมการ

$$\begin{aligned} \sum (Y - \hat{Y})^2 &= \sum (Y - a - bX)^2 \\ SS_{error} &= \sum (Y - a - bX)^2 \end{aligned}$$

ต่อไปจะทำให้ค่า SS_{error} มีค่าน้อยที่สุด ดังนี้

$$\frac{\partial}{\partial a} SS_{error} = \frac{\partial}{\partial a} \sum (Y - a - bX)^2 = 0$$

และ

$$\frac{\partial}{\partial b} SS_{error} = \frac{\partial}{\partial b} \sum (Y - a - bX)^2 = 0$$

นำสมการแรกมาแก้สมการ

$$\begin{aligned} 0 &= \frac{\partial}{\partial a} \sum (Y - a - bX)^2 \\ &= \sum \frac{\partial}{\partial a} (Y - a - bX)^2 \\ &= \sum \frac{\partial}{\partial (Y - a - bX)} (Y - a - bX)^2 \cdot \frac{\partial}{\partial a} (Y - a - bX) \\ &= \sum 2(Y - a - bX) \cdot \left(\frac{\partial Y}{\partial a} - \frac{\partial a}{\partial a} - \frac{\partial bX}{\partial a} \right) \\ &= 2 \sum (Y - a - bX) \cdot (0 - 1 - 0) \\ &= 2 \sum (Y - a - bX) \cdot (-1) \\ &= \sum (Y - a - bX) \\ &= \sum Y - \sum a - \sum bX \\ &= \sum Y - na - b \sum X \\ &= \bar{Y} - a - b\bar{X} \\ \bar{Y} &= a + b\bar{X} \end{aligned}$$

จะเห็นว่าเหมือนกับสมการที่ 4 แสดงว่าได้ข้อมูลมาว่าสมการที่ผ่านจุดพิคัด $(\bar{X}, \bar{Y})$ จะมีแนวโน้มที่ทำให้ SS_{error} มีค่าน้อยที่สุด แต่ว่าสมการดังกล่าวคือสมการอะไร ต้องหาข้อมูลจากการแก้สมการที่สอง

$$\begin{aligned} 0 &= \frac{\partial}{\partial b} \sum (Y - a - bX)^2 \\ &= \sum \frac{\partial}{\partial b} (Y - a - bX)^2 \\ &= \sum \frac{\partial}{\partial (Y - a - bX)} (Y - a - bX)^2 \cdot \frac{\partial}{\partial b} (Y - a - bX) \\ &= \sum 2(Y - a - bX) \cdot \left(\frac{\partial Y}{\partial b} - \frac{\partial a}{\partial b} - \frac{\partial bX}{\partial b} \right) \\ &= \sum 2(Y - a - bX) \cdot \left(0 - 0 - X \frac{\partial b}{\partial b} \right) \\ &= 2 \sum (Y - a - bX) \cdot (-X) \\ &= \sum (Y - a - bX) \cdot (-X) \\ &= \sum (aX + bX^2 - XY) \\ &= \sum aX + \sum bX^2 - \sum XY \\ &= a \sum X + b \sum X^2 - \sum XY \\ \sum XY &= a \sum X + b \sum X^2 \end{aligned}$$

แทนค่า a ด้วยข้อมูลจากการหาอนุพันธ์ในสมการแรก คือ $a = \bar{Y} - b\bar{X}$ เป็นดังนี้

$$\begin{aligned} \sum XY &= (\bar{Y} - b\bar{X}) \sum X + b \sum X^2 \\ \sum XY &= \bar{Y} \sum X - b\bar{X} \sum X + b \sum X^2 \end{aligned}$$

$$\sum XY - \bar{Y} \sum X = b(\sum X^2 - \bar{X} \sum X)$$

เมื่อคูณรูปสมการที่ 5 พบว่า

$$\begin{aligned} \sum X^2 - \bar{X} \sum X &= \sum (X - \bar{X})^2 \\ \text{และ} \quad \sum XY - \bar{Y} \sum X &= \sum (X - \bar{X})(Y - \bar{Y}) \end{aligned} \quad \left. \vphantom{\sum} \right\} \text{พิสูจน์เอง}$$

ดังนั้นสามารถเขียนรูปสมการที่ 5 ได้ใหม่เป็น

$$\sum (X - \bar{X})(Y - \bar{Y}) = b \sum (X - \bar{X})^2$$

$$\text{หรือ} \quad b = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2}$$

ถ้านำ $n - 1$ หารทั้งเศษและส่วนของพจน์ด้านขวาจะได้สูตรเป็น

$$b = \frac{\text{COV}_{XY}}{S_x}$$

เพราะฉะนั้นเส้นสมการทำนายที่ทำให้ค่า SS_{error} มีค่าน้อยที่สุดจะต้องมีค่าความชัน หรือค่า b เท่ากับ ค่าความแปรผันร่วม (Covariance) ของตัวแปรต้นและตัวตามหารด้วยส่วนเบี่ยงเบนมาตรฐานของตัวแปรต้น ขณะที่จุดตัดแกน Y สามารถหาได้โดยแทนค่าจากสมการ $a = \bar{Y} - b\bar{X}$

หนังสืออ้างอิง

- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah: Lawrence Erlbaum.
- Tatsuoka, M. M. (1988). *Multivariate analysis: Techniques for educational and psychological research* (2nd ed.). New York: Macmillian.
- Weir, M. D., Hass, J., & Giordano, F. R. (2005). *Thomas' calculus* (11th ed.). Boston: Pearson.