บทที่ 2

การวิเคราะห์การถดถอยเชิงเส้นตรงอย่างง่าย

ในบทนี้จะกล่าวถึงเทคนิคการวิเคราะห์การถดถอยเชิงเส้นตรงอย่างง่าย ซึ่งเป็นการหาความสัมพันธ์ระหว่างตัวแปรตาม (Y) และตัวแปรอิสระ (X) เพียงตัวเดียว โดยเริ่มจากโมเดลถดถอยเชิงเส้นตรงอย่างง่าย ข้อตกลงเบื้องต้นของการวิเคราะห์ วิธีสร้างสมการถดถอยจากข้อมูลตัวอย่าง เทคนิคการประมาณค่าพารามิเตอร์ในโมเดลถดถอย การ ประมาณและการทดสอบสมมติฐานเกี่ยวกับพารามิเตอร์ การนำสมการถดถอยไปใช้เพื่อการพยากรณ์ และการหาช่วงความเชื่อมั่นของค่าพยากรณ์ การคำนวณค่าสัมประสิทธิ์สหสัมพันธ์และสัมประสิทธิ์ตัวกำหนด

2.1 โมเดลถดถอยเชิงเส้นตรงอย่างง่าย (Simple Linear Regression Model)

โมเดลที่แสดงความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระเพียงหนึ่งตัวของข้อมูลในประชากร (Population) มีรูปแบบดังนี้

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad i = 1, 2, \ldots, n \tag{2.1}$$

เมื่อ

Y_i แทน ตัวแปรตาม

X_i แทน ตัวแปรอิสระ

β_0 แทน ระยะตัดแกน Y (Intercept)

β_1 แทน ความชันของเส้นถดถอย (Slope)

ϵ_i แทน ค่าความคลาดเคลื่อน (Random error)

โดยทั้ง β_0 และ β_1 เป็นพารามิเตอร์ที่ไม่ทราบค่า

ข้อตกลงเบื้องต้นเกี่ยวกับความคลาดเคลื่อน (Assumptions about random errors)

- ϵ_i เป็นตัวแปรสุ่มที่มีการแจกแจงแบบปกติ
- มีค่าเฉลี่ยเป็น 0 นั่นคือ $E(\epsilon_i) = 0$
- มีความแปรปรวนเท่ากัน นั่นคือ $V(\epsilon_i) = \sigma^2$
- ความคลาดเคลื่อนเป็นอิสระกัน นั่นคือ $Cov(\epsilon_i, \epsilon_j) = 0$ เมื่อ $i \neq j, \quad i, j = 1, 2, \ldots, n$

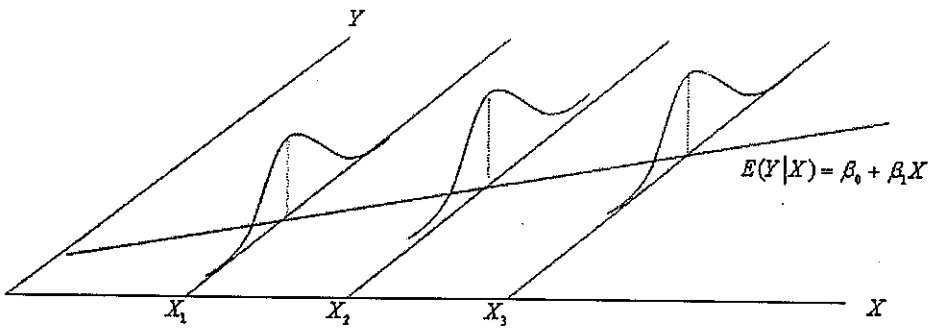
ซึ่งสามารถเขียนโดยรวมได้ดังนี้ $\epsilon_i \sim NID(0, \sigma^2)$

ในการวิเคราะห์การถดถอยจะถือว่าตัวแปรอิสระ X เป็นตัวแปรที่ทราบค่าและมีค่าคงที่ (Known or fixed value) และถูกวัดโดยปราศจากความคลาดเคลื่อน ในขณะที่ตัวแปรตาม Y เป็นตัวแปรสุ่ม มีการแจกแจงในแต่ละระดับที่เป็นไปได้ของ X ซึ่งค่าคาดหวังและความแปรปรวนของการแจกแจงของ Y เมื่อกำหนด X สามารถเขียนได้ดังนี้

$$E(Y_i \mid X_i) = E(\beta_0 + \beta_1 X_i + \epsilon_i) = \beta_0 + \beta_1 X_i \tag{2.2}$$

$$V(Y_i \mid X_i) = V(\beta_0 + \beta_1 X_i + \epsilon_i) = \sigma^2 \tag{2.3}$$

ซึ่งจะเห็นได้ว่า ค่าเฉลี่ยของ Y เมื่อกำหนด X เป็นฟังก์ชันเชิงเส้นตรงของ X แต่ความแปรปรวนของ Y ไม่ขึ้นกับค่า X ทั้งนี้เนื่องจากความคลาดเคลื่อนมีความเป็นอิสระกัน จึงส่งผลให้ตัวแปรสุ่ม Y เป็นอิสระกัน



รูปที่ 2.1: การแจกแจงความน่าจะเป็นของตัวแปรตาม Y เมื่อกำหนดตัวแปรอิสระ X

ด้วย ดังนั้นสามารถเขียนการแจกแจงของ Y ได้ดังนี้ $Y_i \sim NID(\beta_0 + \beta_1 X_i, \sigma^2)$ ดังแสดงในรูปที่ 2.1

พารามิเตอร์ β_0 และ β_1 มักถูกเรียกว่า *ค่าสัมประสิทธิ์ถดถอย* (Regression coefficients) โดยที่ค่าความชัน β_1 แสดงถึงการเปลี่ยนแปลงของค่าเฉลี่ยของ Y ต่อการเปลี่ยนแปลงของ X 1 หน่วย หากพิสัย (Range) ของ X รวมศูนย์อยู่ด้วย จะได้ว่า β_0 เป็นค่าเฉลี่ยของตัวแปรตาม Y เมื่อ $X = 0$ แต่ถ้าพิสัยของ X ไม่รวมศูนย์แล้ว β_0 จะไม่มีความหมายต่อการแปลผล

2.2 การประมาณค่าพารามิเตอร์โดยวิธีกำลังสองน้อยที่สุด (Ordinary Least Square Method: OLS)

การสร้างสมการถดถอยเชิงเส้นตรงอย่างง่ายนั้นจะเกี่ยวข้องกับการประมาณพารามิเตอร์ที่ไม่ทราบค่า 2 ตัว คือ β_0 และ β_1 ซึ่งวิธีกำลังสองน้อยที่สุดเป็นวิธีหนึ่งที่ใช้ในการประมาณค่า β_0 และ β_1 ที่ทำให้ผลรวมกำลังสองของความคลาดเคลื่อน (Error sum of squares: SSE) มีค่าต่ำสุด โดยที่ความคลาดเคลื่อนดังกล่าวเกิดจากความแตกต่างระหว่างค่าสังเกต (Y_i) และค่าพยากรณ์ที่ได้จากเส้นถดถอย ($\hat{Y}_i$)

สมมติข้อมูลประกอบด้วยตัวอย่างสุ่มขนาด n ดังนี้

ค่าสังเกตที่	ตัวแปรตาม (Y_i)	ตัวแปรอิสระ (X_i)
1	Y_1	X_1
2	Y_2	X_2
3	Y_3	X_3
$\vdots$	$\vdots$	$\vdots$
n	Y_n	X_n

ตารางที่ 2.1: ลักษณะข้อมูลที่ใช้ในการวิเคราะห์สมการถดถอยเชิงเส้นตรงอย่างง่าย

ให้สมการถดถอยเชิงเส้นตรงอย่างง่ายที่ได้จากข้อมูลตัวอย่างข้างต้นมีรูปแบบดังนี้

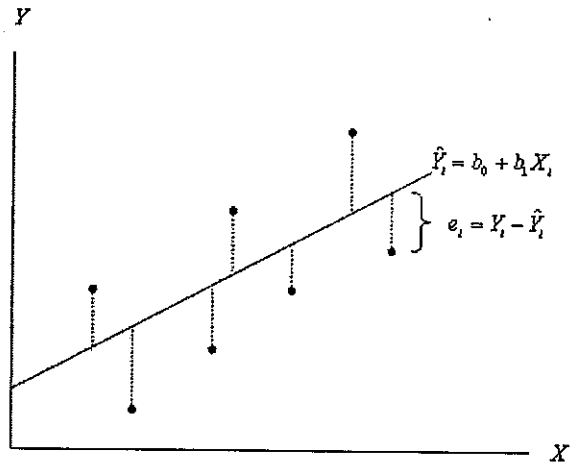
$$\hat{Y}_i = b_0 + b_1 X_i, \qquad i = 1, 2, \ldots, n \tag{2.4}$$

เมื่อ

$\hat{Y}_i$ แทน ค่าประมาณของ $E(Y_i \mid X_i)$

b_0 แทน ค่าประมาณของพารามิเตอร์ β_0

b_1 แทน ค่าประมาณของพารามิเตอร์ β_1



รูปที่ 2.2: ความคลาดเคลื่อนที่เกิดจากข้อมูลตัวอย่าง

และค่าความคลาดเคลื่อน (Residual) ที่เกิดจากผลต่างระหว่าง Y_i และ $\hat{Y}_i$ ของข้อมูลตัวอย่าง (รูปที่ 2.2) สามารถเขียนได้เป็น

$$e_i = Y_i - \hat{Y}_i, \qquad i = 1, 2, \ldots, n \tag{2.5}$$

ซึ่งเป็นค่าประมาณของความคลาดเคลื่อนที่เกิดจากค่าสังเกต Y_i และค่าบนเส้นถดถอยที่แท้จริง นั่นคือ

$$\epsilon_i = Y_i - E(Y_i \mid X_i) = Y_i - \beta_0 - \beta_1 X_i \tag{2.6}$$

ค่าผลรวมกำลังสองของความคลาดเคลื่อนที่เกิดจากเส้นถดถอยที่แท้จริง สามารถแสดงได้ดังนี้

$$SSE = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \tag{2.7}$$

การหาตัวประมาณของ β_0 และ β_1 ด้วยวิธีกำลังสองน้อยที่สุด ทำได้โดยหาอนุพันธ์ของ SSE เทียบกับ β_0 และ β_1 ดังนี้

$$\frac{\partial SSE}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) \tag{2.8}$$

$$\frac{\partial SSE}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i) \tag{2.9}$$

แล้วให้สมการ (2.8) และ (2.9) เท่ากับศูนย์ และแทนค่า (β_0, β_1) ด้วย (b_0, b_1) ดังนี้

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i) &= 0\\ \sum_{i=1}^{n}X_i(Y_i - b_0 - b_1X_i) &= 0\end{aligned}$$

หลังจากนั้นจัดรูปสมการข้างต้นใหม่ จะได้ว่า

$$nb_0 + b_1\sum_{i=1}^{n}X_i = \sum_{i=1}^{n}Y_i \tag{2.10}$$

$$b_0\sum_{i=1}^{n}X_i + b_1\sum_{i=1}^{n}X_i^2 = \sum_{i=1}^{n}X_iY_i \tag{2.11}$$

เรียกสมการ (2.10) และ (2.11) ว่า *สมการปกติ* (Normal equations) จากการแก้สมการปกติทำให้ได้ตัวประมาณกำลังสองน้อยที่สุด ดังนี้

$$b_1 = \frac{\sum_{i=1}^{n}X_iY_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}X_i^2 - n\bar{X}^2} \tag{2.12}$$

$$b_0 = \bar{Y} - b_1\bar{X} \tag{2.13}$$

เมื่อ

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$ แทน ค่าเฉลี่ยของตัวแปร X

$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}Y_i$ แทน ค่าเฉลี่ยของตัวแปรตาม Y

หากแทนค่า b_0 เข้าไปในสมการ (2.4) สมการถดถอยจะเปลี่ยนรูปเป็น

$$\begin{aligned}\hat{Y}_i &= b_0 + b_1X_i\\ &= (\bar{Y} - b_1\bar{X}) + b_1X_i\\ &= \bar{Y} + b_1(X_i - \bar{X})\end{aligned} \tag{2.14}$$

เพื่อให้สูตรที่ใช้ในการคำนวณกระทัดรัดขึ้น กำหนดให้

$$S_{xx} = \sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \tag{2.15}$$

$$S_{yy} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \qquad = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 \tag{2.16}$$

$$S_{xy} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y} \tag{2.17}$$

ดังนั้น สามารถแสดงค่า b_1 ในรูปแบบที่กระทัดรัดได้เป็น

$$b_1 = \frac{S_{xy}}{S_{xx}} \tag{2.18}$$

ตัวอย่างที่ 2.1 ในการศึกษาความสัมพันธ์ระหว่างยอดขาย (หน่วย: หมื่นบาท) และค่าใช้จ่ายในการโฆษณา (หน่วย: หมื่นบาท) ของบริษัท 10 แห่ง ได้ข้อมูลดังต่อไปนี้

บริษัท	ค่าใช้จ่ายในการโฆษณา (X)	ยอดขาย (Y)
1	5	89
2	6	87
3	7	98
4	8	110
5	9	103
6	10	114
7	11	116
8	12	110
9	13	126
10	14	130

ตารางที่ 2.2: ข้อมูลยอดขายและค่าใช้จ่ายในการโฆษณา

จงสร้างแผนภาพการกระจาย และหาสมการถดถอยเชิงเส้นตรงอย่างง่ายแสดงความสัมพันธ์ระหว่างยอดขายและค่าใช้จ่ายในการโฆษณา

วิธีทำ

จากแผนภาพการกระจายในรูปที่ 2.3 จะเห็นได้ว่าค่าใช้จ่ายในการโฆษณาและยอดขายมีความสัมพันธ์ไปในทิศทางเดียวกัน นั่นคือ หากเพิ่มค่าใช้จ่ายในการโฆษณามากขึ้น จะทำให้ยอดขายสูงขึ้นตาม นอกจากนี้จะเห็นได้ว่ารูปแบบความสัมพันธ์มีลักษณะใกล้เคียงเส้นตรง

คำนวณค่าต่าง ๆ ของข้อมูลในตารางที่ 2.2 ได้ดังนี้

$$\sum_{i=1}^{n} X_i = 95, \qquad \sum_{i=1}^{n} Y_i = 1,083, \qquad \sum_{i=1}^{n} X_i Y_i = 10,654,$$
$$\sum_{i=1}^{n} X_i^2 = 985, \qquad \sum_{i=1}^{n} Y_i^2 = 119,131, \qquad n = 10,$$
$$\bar{X} = 9.5, \qquad \bar{Y} = 108.3$$

จะได้ว่า

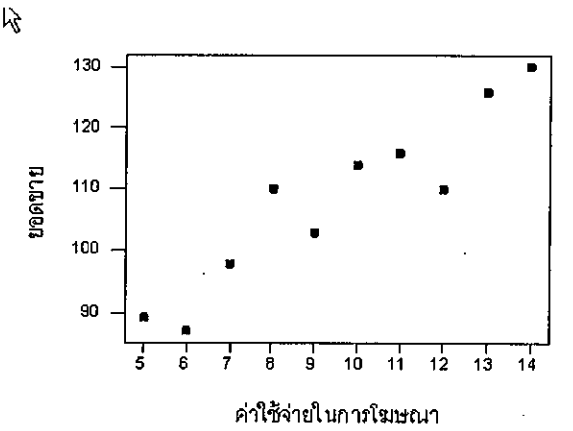
$$S_{xy} = \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y} = 10,654 - 10(9.5)(108.3) = 365.5$$
$$S_{xx} = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 = 985 - 10(9.5)^2 = 82.5$$
$$S_{yy} = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 = 119,131 - 10(108.3)^2 = 1,842.1$$

ดังนั้น

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{365.5}{82.5} = 4.4303$$
$$b_0 = \bar{Y} - b_1\bar{X} = 108.3 - (4.4303)(9.5) = 66.2121$$

สมการถดถอยเชิงเส้นตรงอย่างง่ายแสดงความสัมพันธ์ระหว่างยอดขายและค่าใช้จ่ายในการโฆษณา คือ

$$\hat{Y} = 66.2121 + 4.4303X$$



รูปที่ 2.3: แผนภาพการกระจายระหว่างค่าใช้จ่ายในการโฆษณา (หมื่นบาท) และยอดขาย (หมื่นบาท)

จะเห็นได้ว่าความชันมีค่าเท่ากับ 4.4303 ซึ่งหมายความว่า ถ้าเพิ่มค่าใช้จ่ายในการโฆษณา 1 หมื่นบาท ยอดขายจะเพิ่มขึ้น 4.4303 หน่วย หรือ 44,303 บาท

หลังจากที่ได้สมการถดถอยแล้ว อาจมีคำถามมากมายติดตามมา เช่น ข้อมูลที่ศึกษามีลักษณะขัดแย้งกับข้อตกลงเบื้องต้นของการวิเคราะห์หรือไม่ สมการถดถอยที่ได้สามารถอธิบายความสัมพันธ์ในข้อมูลได้ดีเพียงใด และเป็นประโยชน์ต่อการพยากรณ์หรือไม่ เป็นต้น ซึ่งคำถามเหล่านี้ควรที่จะมีการตรวจสอบก่อนที่จะรับเอาสมการไปใช้ เนื่องจากความคลาดเคลื่อน e_i เป็นตัวประมาณของ ϵ_i จึงมีบทบาทสำคัญและใช้ในการตรวจสอบความเหมาะสมของโมเดลได้ ดังนั้นการตรวจสอบความคลาดเคลื่อนว่าเป็นตัวอย่างสุ่ม และมีการแจกแจงที่สอดคล้องกับข้อตกลงเบื้องต้นของการวิเคราะห์หรือไม่จึงเป็นสิ่งจำเป็น

2.2.1 คุณสมบัติของตัวประมาณที่ได้จากวิธีกำลังสองน้อยที่สุด

ตัวประมาณ b_0 และ b_1 ที่ได้จากวิธีกำลังสองน้อยที่สุด มีคุณสมบัติที่สำคัญทางสถิติหลายประการ เพื่อให้สะดวกต่อการคำนวณ จะเขียน b_1 ให้อยู่ในรูปฟังก์ชันเชิงเส้น (Linear function) ของค่าสังเกต Y_i ดังนี้

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \sum_{i=1}^{n} c_i Y_i$$

เมื่อ c_i แทน ค่าคงที่ โดย $c_i = \dfrac{(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \quad i = 1, 2, \ldots, n$

นอกจากนี้ สามารถแสดงได้ว่า

$$\sum_{i=1}^{n} c_i = 0 \tag{2.19}$$

$$\sum_{i=1}^{n} c_i X_i = 1 \tag{2.20}$$

$$\sum_{i=1}^{n} c_i^2 = \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{2.21}$$

ดังนั้น

$$\begin{aligned} E(b_1) &= E\left(\sum_{i=1}^{n} c_i Y_i\right) \\ &= \sum_{i=1}^{n} c_i E(Y_i) \\ &= \sum_{i=1}^{n} c_i \left(\beta_0 + \beta_1 X_i\right) \\ &= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i X_i \\ &= \beta_1 \end{aligned}$$

เนื่องจาก $E(Y_i) = \beta_0 + \beta_1 X_i$ ทำให้ได้ว่า b_1 เป็นตัวประมาณที่ไม่เอนเอียง (Unbiased estimator) ของ β_1 และมีความแปรปรวนเป็น

$$\begin{aligned} V(b_1) = \sigma^2_{b_1} &= V\left(\sum_{i=1}^{n} c_i Y_i\right) \\ &= \sum_{i=1}^{n} c_i^2 V(Y_i) \\ &= \sigma^2 \sum_{i=1}^{n} c_i^2 \\ &= \frac{\sigma^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \end{aligned}$$

ทำนองเดียวกัน สามารถแสดงได้ว่า b_0 เป็นตัวประมาณที่ไม่เอนเอียง (Unbiased estimator) ของ β_0 ดังนี้

$$\begin{aligned} E(b_0) &= E(\bar{Y} - b_1 \bar{X}) \\ &= \frac{1}{n} \sum_{i=1}^{n} E(Y_i) - \bar{X} E(b_1) \\ &= \frac{1}{n} \sum_{i=1}^{n} (\beta_0 + \beta_1 X_i) - \beta_1 \bar{X} \\ &= \frac{1}{n} \left(n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i\right) - \beta_1 \bar{X} \\ &= \beta_0 \end{aligned}$$

และความแปรปรวนของ b_0 คือ

$$\begin{aligned} V(b_0) = \sigma^2_{b_0} &= V\left(\bar{Y} - b_1\bar{X}\right) \\ &= V(\bar{Y}) + \bar{X}^2 V(b_1) - 2\bar{X}Cov(\bar{Y}, b_1) \end{aligned}$$

เนื่องจาก $V(\bar{Y}) = \dfrac{\sigma^2}{n}$ และ $Cov(\bar{Y}, b_1) = 0$ จะได้ความแปรปรวนของ b_0 มีค่าเท่ากับ

$$\begin{aligned} V(b_0) &= V(\bar{Y}) + \bar{X}^2 V(b_1) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right) \end{aligned}$$

คุณสมบัติของตัวประมาณ b_0 และ b_1 ที่ได้จากวิธีกำลังสองน้อยที่สุดเป็นผลมาจาก *ทฤษฎีของเกาส์-มาร์-คอฟ* (Gauss-Markov theorem) ที่ได้กำหนดข้อตกลงเบื้องต้นเกี่ยวกับความคลาดเคลื่อน ϵ_i ของการวิเคราะห์การถดถอยไว้ นั่นคือ $E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma^2$ และ $Cov(\epsilon_i, \epsilon_j) = 0$, $i \neq j$ ซึ่งมีผลทำให้ตัวประมาณที่ได้โดยวิธีกำลังสองน้อยที่สุดเป็นตัวประมาณที่ไม่เอนเอียงและมีความแปรปรวนต่ำสุดในบรรดาตัวประมาณที่ไม่เอนเอียงที่เป็นฟังก์ชันเชิงเส้นของ Y_i ทั้งหมด จึงเรียกตัวประมาณ b_0 และ b_1 ว่าเป็น *ตัวประมาณที่ไม่เอนเอียงเชิงเส้นตรงที่ดีที่สุด* (Best Linear Unbiased Estimator: BLUE) ของ β_0 และ β_1

2.2.2 คุณสมบัติของสมการถดถอยที่ได้จากวิธีกำลังสองน้อยที่สุด

จะเห็นได้ว่าตัวประมาณที่ได้จากวิธีกำลังสองน้อยที่สุดมีคุณสมบัติที่สำคัญทางสถิติหลายประการเช่นเดียวกับสมการถดถอยที่ได้จากวิธีกำลังสองน้อยที่สุด โดยมีคุณสมบัติที่สำคัญดังนี้

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

2. ผลรวมกำลังสองของความคลาดเคลื่อนมีค่าน้อยที่สุด นั่นคือ $\sum_{i=1}^{n} e_i^2$ มีค่าน้อยที่สุด

3. ผลรวมของค่าสังเกตมีค่าเท่ากับผลรวมของค่าที่ได้จากสมการถดถอย นั่นคือ $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$

4. ผลรวมของความคลาดเคลื่อนที่ถ่วงน้ำหนักด้วยตัวแปรอิสระมีค่าเป็นศูนย์เสมอ นั่นคือ $\sum_{i=1}^{n} X_i e_i = 0$

5. ผลรวมของความคลาดเคลื่อนที่ถ่วงน้ำหนักด้วยค่าบนเส้นถดถอยมีค่าเป็นศูนย์เสมอ นั่นคือ $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$

6. เส้นถดถอยที่ได้จากวิธีวิธีกำลังสองน้อยที่สุดจะผ่านจุด $(\bar{X}, \bar{Y})$ เสมอ

2.2.3 การประมาณค่าความแปรปรวนของความคลาดเคลื่อน σ^2

เป็นที่ทราบกันแล้วว่า σ^2 เป็นความแปรปรวนของความคลาดเคลื่อน ϵ_i ในโมเดลถดถอย ซึ่งเป็นดัชนีที่บ่งบอกถึงการกระจายของการแจกแจงความน่าจะเป็นของตัวแปรตาม Y นอกจากนี้การอนุมานทางสถิติที่เกี่ยวข้องกับสมการถดถอย ไม่ว่าจะเป็นการทดสอบสมมติฐานหรือสร้างช่วงความเชื่อมั่น ล้วนเกี่ยวข้องกับความแปรปรวน σ^2 ทั้งสิ้น ดังนั้นจึงจำเป็นที่จะต้องมีการประมาณค่าความแปรปรวนดังกล่าว โดยตัวประมาณแบบจุด (Point estimator) ของ σ^2 สามารถคำนวณได้จากผลรวมกำลังสองของความคลาดเคลื่อน (Error sum of squares or Residual sum of squares) เขียนแทนด้วย SSE ดังนี้

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{2.22}$$

เพื่อสะดวกในการคำนวณ แทนค่า $\hat{Y}_i = b_0 + b_1 X_i$ เข้าไปใน (2.22) แล้วจัดรูปสมการใหม่ จะได้ว่า

$$SSE = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 - b_1 S_{xy}$$

แต่ $\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = S_{yy}$ ดังนั้นจะได้ว่า

$$\begin{aligned} SSE &= S_{yy} - b_1 S_{xy} \\ &= S_{yy} - \frac{(S_{xy})^2}{S_{xx}} \end{aligned} \tag{2.23}$$

ค่าผลรวมกำลังสองของความคลาดเคลื่อนมีจำนวนองศาความเป็นอิสระ (Degrees of freedom: df) เป็น $n-2$ ซึ่งจำนวนองศาความเป็นอิสระที่เสียไป 2 ค่านั้น เนื่องมาจากการประมาณ β_0 และ β_1 ในโมเดลถดถอย ดังนั้นตัวประมาณแบบจุดของ σ^2 คือ

$$MSE = S^2 = \frac{SSE}{n-2} \tag{2.24}$$

เมื่อ MSE แทน ความคลาดเคลื่อนกำลังสองเฉลี่ย (Error mean square or residual mean square) ซึ่งพบว่า MSE หรือ S^2 เป็นตัวประมาณที่ไม่เอนเอียงของ σ^2 นั่นคือ $E(MSE) = \sigma^2$ และรากที่สองของ

MSE หรือ $\sqrt{MSE}$ ก็จะเป็นตัวประมาณของส่วนเบี่ยงเบนมาตรฐาน σ เช่นกัน

ตัวอย่างที่ 2.2 จงคำนวณค่าประมาณของความแปรปรวนของความคลาดเคลื่อนในสมการถดถอยที่ได้ในตัวอย่างที่ 2.1

วิธีทำ คำนวณหาค่าผลรวมกำลังสองของความคลาดเคลื่อนได้ดังนี้

$$\begin{aligned} SSE &= S_{yy} - \frac{(S_{xy})^2}{S_{xx}} \\ &= 1,842.1 - \frac{(365.5)^2}{82.5} \\ &= 222.8242 \end{aligned}$$

ดังนั้นค่าประมาณของ σ^2 คือ

$$MSE = \frac{SSE}{n-2} = \frac{222.8242}{10-2} = 27.8530$$

2.3 การอนุมานทางสถิติเกี่ยวกับพารามิเตอร์ β_1

การอนุมานทางสถิติที่เกี่ยวข้องกับ β_1 ต้องอาศัยข้อกำหนดเบื้องต้นเกี่ยวกับการแจกแจงแบบปกติของ ϵ_i ซึ่งบ่อยครั้งผู้วิจัยอาจสนใจที่จะสร้างช่วงความเชื่อมั่นหรือทดสอบสมมติฐานเกี่ยวกับพารามิเตอร์ β_1 ในโมเดลถดถอย โดยเฉพาะรูปแบบของการทดสอบสมมติฐานดังต่อไปนี้

$$H_0 : \beta_1 = 0 \qquad vs. \qquad H_1 : \beta_1 \neq 0$$

เนื่องจากถ้า $\beta_1 = 0$ แล้ว เส้นถดถอยจะขนานกับแกน X ซึ่งหมายถึงว่าค่าเฉลี่ยของการแจกแจงของ Y มีค่าเท่ากันทั้งหมด ดังนั้น $\beta_1 = 0$ นอกจากจะแสดงว่าตัวแปรตาม Y ไม่มีความสัมพันธ์เชิงเส้นตรงกับตัวแปรอิสระ X แล้ว ยังอาจชี้ว่าตัวแปรตามและตัวแปรอิสระไม่มีความสัมพันธ์ต่อกัน ไม่ว่าจะอยู่ในรูปแบบใดอีกด้วย

2.3.1 การแจกแจงตัวอย่างของ b_1 (Sampling Distribution of b_1)

เป็นที่ทราบกันแล้วว่าตัวประมาณแบบจุดของ b_1 คือ $\frac{S_{xy}}{S_{xx}}$ และการแจกแจงตัวอย่างของ b_1 เกิดจากค่าที่แตกต่างกันของ b_1 ที่ได้จากการสุ่มตัวอย่างซ้ำ (Repeated sampling) เมื่อให้ตัวแปรอิสระ X ในตัวอย่างแต่ละชุดมีค่าคงที่ โดยที่ b_1 มีการแจกแจงแบบปกติ มีค่าเฉลี่ยเป็น β_1 และความแปรปรวน $\frac{\sigma^2}{S_{xx}}$

เนื่องจาก σ^2 เป็นความแปรปรวนของความคลาดเคลื่อนที่ไม่ทราบค่า ซึ่งสามารถประมาณได้ด้วย S^2 หรือ MSE และได้ค่าประมาณของความแปรปรวนของการแจกแจงของ b_1 ดังนี้

$$S_{b_1}^2 = \frac{S^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{S^2}{S_{xx}} \tag{2.25}$$

ตัวประมาณแบบจุด $S_{b_1}^2$ มีคุณสมบัติเป็นตัวประมาณที่ไม่เอนเอียงของ $\sigma_{b_1}^2$ และเรียก S_{b_1} ซึ่งแทนรากที่สองที่เป็นบวกของ $S_{b_1}^2$ ว่า *ความคลาดเคลื่อนมาตรฐานของ* b_1 (Standard error of b_1) โดยเป็นตัวประมาณแบบจุดของ σ_{b_1} ด้วยเช่นกัน
ดังนั้น

$$\frac{b_1 - \beta_1}{S_{b_1}} \sim t_{n-2}$$

เมื่อ t_{n-2} แทน การแจกแจงแบบที ด้วยองศาความเป็นอิสระ $n-2$

ทำนองเดียวกับการประมาณค่าความแปรปรวนของความคลาดเคลื่อน จำนวนองศาความเป็นอิสระที่หายไป 2 ค่า เนื่องมาจากการประมาณพารามิเตอร์ β_0 และ β_1 ในโมเดลถดถอย ซึ่งความสัมพันธ์ดังกล่าวสามารถเขียนใหม่ได้เป็น

$$\frac{b_1 - \beta_1}{S_{b_1}} = \frac{b_1 - \beta_1}{\sigma_{b_1}} \div \frac{S_{b_1}}{\sigma_{b_1}}$$

โดยที่

$$z = \frac{b_1 - \beta_1}{\sigma_{b_1}} \sim N(0,1)$$

$$\frac{S_{b_1}^2}{\sigma_{b_1}^2} = \frac{MSE}{\sigma_{b_1}^2} = \frac{SSE}{\sigma_{b_1}^2(n-2)} = \frac{U}{n-2}$$

เมื่อ $U \sim \chi_{n-2}^2$

เนื่องจาก z และ U เป็นตัวแปรสุ่มที่เป็นอิสระกัน ดังนั้น

$$z = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{z}{\sqrt{\frac{U}{n-2}}} \sim t_{n-2}$$

2.3.2 ช่วงความเชื่อมั่นของ β_1 (Confidence Interval of β_1)

เนื่องจาก $\frac{b_1-\beta_1}{S_{b_1}}$ มีการแจกแจงแบบที ซึ่งเป็นการแจกแจงที่สมมาตรรอบจุดศูนย์ ดังนั้นสามารถสร้างช่วงความเชื่อมั่น $(1-\alpha)100\%$ ของ β_1 ได้ดังนี้

$$b_1 \pm t_{\frac{\alpha}{2},n-2} \cdot S_{b_1} \tag{2.26}$$

เมื่อ $t_{\frac{\alpha}{2},n-2}$ แทน ค่าเปอร์เซนไทล์ที่ $(\alpha/2)100$ ของการแจกแจงแบบ t ที่องศาความเป็นอิสระ $n-2$

โดยช่วงความเชื่อมั่น $(1-\alpha)100\%$ ที่ได้นี้จะหมายถึงว่า หากสุ่มตัวอย่างหลาย ๆ ชุดอย่างเป็นอิสระกัน และระดับของตัวแปรอิสระ X มีค่าเหมือนกับข้อมูลแล้ว จะได้ว่า $(1-\alpha)100\%$ ของช่วงความเชื่อมั่นเหล่านี้จะครอบคลุมค่าที่แท้จริงของพารามิเตอร์ β_1

2.3.3 การทดสอบสมมติฐานเกี่ยวกับ β_1 (Hypothesis Testing Concerning β_1)

การทดสอบสมมติฐานเกี่ยวกับพารามิเตอร์ β_1 สามารถจำแนกได้เป็น 2 ลักษณะตามประเภทของสมมติฐานทางเลือก ดังนี้

1. **การทดสอบสองทาง (Two-sided test)** กำหนดสมมติฐานของการทดสอบได้เป็น

$$H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0$$

สถิติทดสอบ:

$$t_c = \frac{b_1}{S_{b_1}} \tag{2.27}$$

บริเวณวิกฤติ: เมื่อกำหนดระดับนัยสำคัญ (Level of significance) ของการทดสอบเป็น α จะได้ว่า

ปฏิเสธ H_0 ถ้า $|t_c| \geq t_{\alpha/2,n-2}$

ยอมรับ H_0 ถ้า $|t_c| < t_{\alpha/2,n-2}$

หากการทดสอบนำไปสู่การปฏิเสธ H_0 แสดงว่าตัวแปรตามและตัวแปรอิสระมีความสัมพันธ์เชิงเส้นตรงต่อกัน

2. **การทดสอบทางเดียว (One-sided test)** อาจกำหนดสมมติฐานได้ดังนี้

$$H_0 : \beta_1 \leq 0 \quad vs. \quad H_1 : \beta_1 > 0$$

หรือ $$H_0 : \beta_1 \geq 0 \quad vs. \quad H_1 : \beta_1 < 0$$

สถิติทดสอบ:

$$t_c = \frac{b_1}{S_{b_1}}$$

บริเวณวิกฤติ: ที่ระดับนัยสำคัญ α จะได้ว่า

ปฏิเสธ H_0 ถ้า $|t_c| \geq t_{\alpha, n-2}$

ยอมรับ H_0 ถ้า $|t_c| < t_{\alpha, n-2}$

นอกจากนี้บางครั้งผู้วิจัยอาจสนใจที่จะทดสอบสมมติฐานเกี่ยวกับ β_1 ว่ามีค่าตามที่ระบุไว้หรือไม่ ซึ่งอาจเป็นค่ามาตรฐานที่ใช้ในการเปรียบเทียบกระบวนการหรือเป็นค่าที่ถูกกำหนดไว้ล่วงหน้าก่อนแล้ว โดยมีสมมติฐานของการทดสอบเป็น

$$H_0 : \beta_1 = \beta_{10} \quad vs. \quad H_1 : \beta_1 \neq \beta_{10}$$

เมื่อ β_{10} แทน ค่าที่ต้องการทดสอบ ($\beta_{10} \neq 0$)

สถิติทดสอบ:

$$t_c = \frac{b_1 - \beta_{10}}{S_{b_1}} \tag{2.28}$$

บริเวณวิกฤติ: ที่ระดับนัยสำคัญ α จะได้ว่า

ปฏิเสธ H_0 ถ้า $|t_c| \geq t_{\alpha/2, n-2}$

ยอมรับ H_0 ถ้า $|t_c| < t_{\alpha/2, n-2}$

จะเห็นได้ว่าสถิติทดสอบ (2.28) สามารถเขียนให้อยู่ในรูป (2.27) ได้ เมื่อ $\beta_{10} = 0$

2.4 การอนุมานทางสถิติเกี่ยวกับพารามิเตอร์ β_0

การอนุมานทางสถิติที่เกี่ยวข้องกับระยะตัดแกนของเส้นถดถอย หรือค่า β_0 นั้น อาจไม่เกิดขึ้นบ่อยนัก เนื่องจากการอนุมานดังกล่าวจะทำได้ก็ต่อเมื่อ ขอบเขตของโมเดลที่ศึกษาครอบคลุมค่า $X = 0$

2.4.1 การแจกแจงตัวอย่างของ b_0 (Sampling Distribution of b_0)

ตัวประมาณแบบจุดของ b_0 คือ $\bar{Y} - b_1\bar{X}$ ซึ่งการแจกแจงของ b_0 เกิดจากค่าที่แตกต่างกันของ b_0 ที่ได้จากการสุ่มตัวอย่างซ้ำ ๆ กัน (Repeated sampling) เมื่อให้ตัวแปรอิสระ X ในตัวอย่างแต่ละชุดมีค่าคงที่ ทำให้เกิดการแจก-

แจงตัวอย่างของตัวประมาณ b_0 และได้ว่า b_0 มีการแจกแจงแบบปกติ มีค่าเฉลี่ยเป็น β_0 และความแปรปรวน $\sigma^2\left(\frac{1}{n}+\frac{\bar{X}^2}{S_{xx}}\right)$

ทำนองเดียวกันกับการประมาณค่า β_1 ที่มีการประมาณ σ^2 ด้วย S^2 หรือ MSE ทำให้ได้ค่าประมาณของความแปรปรวนของการแจกแจงของ b_0 ดังนี้

$$S_{b_0}^2 = MSE\left(\frac{1}{n}+\frac{\bar{X}^2}{S_{xx}}\right) \tag{2.29}$$

โดย $S_{b_0}^2$ เป็นตัวประมาณที่ไม่เอนเอียงของ $\sigma_{b_0}^2$ และรากที่สองที่เป็นบวกของ $S_{b_0}^2$ หรือ S_{b_0} ก็คือค่าความคลาดเคลื่อนมาตรฐานของ b_0 นั่นเอง
ดังนั้น

$$\frac{b_0-\beta_0}{S_{b_0}} \sim t_{n-2} \tag{2.30}$$

ซึ่งการหาการแจกแจงของ $\frac{b_0-\beta_0}{S_{b_0}}$ สามารถทำได้ในลักษณะเดียวกันกับการแจกแจงของ $\frac{b_1-\beta_1}{S_{b_1}}$ ดังได้แสดงในหัวข้อ 2.3.1

2.4.2 ช่วงความเชื่อมั่นของ β_0 (Confidence Interval of β_0)

การสร้างช่วงความเชื่อมั่น $(1-\alpha)100\%$ ของ β_0 สามารถทำได้ในลักษณะที่คล้ายคลึงกับ β_1 ดังนี้

$$b_0 \pm t_{\frac{\alpha}{2},n-2}\cdot S_{b_0}$$

อย่างไรก็ตามช่วงความเชื่อมั่นข้างต้นอาจไม่ให้ความหมายที่เป็นประโยชน์ต่อการวิเคราะห์ ถ้าค่า $X=0$ ไม่ตกอยู่ในขอบเขตของโมเดลที่ศึกษา

2.4.3 การทดสอบสมมติฐานเกี่ยวกับ β_0 (Hypothesis Testing Concerning β_0)

ทำนองเดียวกัน การทดสอบสมมติฐานเกี่ยวกับพารามิเตอร์ β_0 สามารถกำหนดสมมติฐานของการทดสอบได้เป็น

$$H_0 : \beta_0 = \beta_{00} \quad vs. \quad H_1 : \beta_0 \neq \beta_{00}$$

เมื่อ β_{00} แทน ค่าที่ต้องการทดสอบ ซึ่งอาจมีค่าเป็น 0 หรือไม่ก็ได้

สถิติทดสอบ:

$$t_c = \frac{b_0 - \beta_{00}}{S_{b_0}} \tag{2.31}$$

บริเวณวิกฤติ: ที่ระดับนัยสำคัญ α จะได้ว่า

$$\text{ปฏิเสธ } H_0 \text{ ถ้า } |t_c| \geq t_{\alpha/2,n-2}$$
$$\text{ยอมรับ } H_0 \text{ ถ้า } |t_c| < t_{\alpha/2,n-2}$$

หากต้องการทดสอบสมมติฐานทางเดียว ก็สามารถทำได้ในลักษณะที่คล้ายคลึงกับการทดสอบสมมติฐานของ β_1

ข้อสังเกต

- แม้ว่าการแจกแจงความน่าจะเป็นของตัวแปรตาม Y จะเบี่ยงเบนไปจากการแจกแจงแบบปกติ แต่ถ้าตัวอย่างมีขนาดใหญ่แล้ว ตัวประมาณ b_0 และ b_1 ก็ยังคงมีการแจกแจงแบบปกติโดยประมาณ (Asymptotic normality) การสร้างช่วงความเชื่อมั่นและทดสอบสมมติฐานดังที่กล่าวไปแล้วในหัวข้อ 2.3 และ 2.4 ยังคงสามารถประยุกต์ใช้ได้
- ความแปรปรวนของ b_0 และ b_1 ขึ้นอยู่กับระยะห่างของตัวแปร X หากระดับของ X มีการกระจายมาก ทำให้ $\sum_{i=1}^{n}(X_i - \bar{X})^2$ มีค่ามากตามไปด้วย ซึ่งจะส่งผลให้ความแปรปรวนของ b_0 และ b_1 มีค่าลดลง ดังนั้นการกำหนดระดับของตัวแปร X ในการวางแผนการทดลองที่สามารถควบคุมระดับของตัวแปรอิสระได้ จึงเป็นเรื่องที่ควรจะพิจารณา

ตัวอย่างที่ 2.3 จากข้อมูลในตัวอย่างที่ 2.1

1. จงสร้างช่วงความเชื่อมั่น 95% ของ β_0 และ β_1
2. ที่ระดับนัยสำคัญ 0.05 จงตรวจสอบว่ายอดขายมีความสัมพันธ์เชิงเส้นตรงกับค่าใช้จ่ายในการโฆษณาหรือไม่
3. ที่ระดับนัยสำคัญ 0.05 จงตรวจสอบว่าเส้นถดถอยผ่านจุดกำเนิดหรือไม่

วิธีทำ

1. คำนวณค่าประมาณของความแปรปรวนและความคลาดเคลื่อนมาตรฐานของ b_0 และ b_1 ได้ดังนี้

$$\begin{aligned} S^2_{b_0} &= MSE\left(\frac{1}{n}+\frac{\bar{X}^2}{S_{xx}}\right) \\ &= 27.8530\left(\frac{1}{10}+\frac{(9.5)^2}{82.5}\right) \\ &= 33.2548 \\ S_{b_0} &= 5.7667 \end{aligned}$$

และ

$$\begin{aligned} S^2_{b_1} &= \frac{MSE}{S_{xx}} \\ &= \frac{27.8530}{82.5} \\ &= 0.3376 \\ S_{b_1} &= 0.5810 \end{aligned}$$

จากการเปิดตารางค่าวิกฤติได้ค่า $t_{0.025,8} = 2.306$

ดังนั้นช่วงความเชื่อมั่น 95% ของ β_0 คือ

$$\begin{aligned} b_0 \pm t_{\frac{\alpha}{2},n-2}\cdot S_{b_0} &= 66.2121 \pm (2.306)(5.7667) \\ &= 66.2121 \pm 13.298 \\ \text{หรือ} \quad & (52.9141, 79.5101) \end{aligned}$$

และช่วงความเชื่อมั่น 95% ของ β_1 คือ

$$\begin{aligned} b_1 \pm t_{\frac{\alpha}{2},n-2}\cdot S_{b_1} &= 4.4303 \pm (2.306)(0.5810) \\ &= 4.4303 \pm 1.3398 \\ \text{หรือ} \quad & (3.0905, 5.7701) \end{aligned}$$

2. กำหนดสมมติฐานของการทดสอบเป็น

$$H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0$$

สถิติทดสอบ:

$$\begin{aligned} t_c &= \frac{b_1}{S_{b_1}} \\ &= \frac{4.4303}{0.5810} \\ &= 7.6253 \end{aligned}$$

บริเวณวิกฤติ: ปฏิเสธ H_0 ถ้า $|t_c| \geq 2.306$

เนื่องจากค่า t_c ที่คำนวณได้ตกอยู่ในบริเวณวิกฤติ ดังนั้นปฏิเสธ H_0 นั่นคือ ยอดขายมีความสัมพันธ์เชิงเส้นตรงกับค่าใช้จ่ายในการโฆษณา ที่ระดับนัยสำคัญ 0.05

3. กำหนดสมมติฐานของการทดสอบเป็น

$$H_0 : \beta_0 = 0 \quad vs. \quad H_1 : \beta_0 \neq 0$$

สถิติทดสอบ:

$$\begin{aligned} t_c &= \frac{b_0}{S_{b_0}} \\ &= \frac{66.2121}{5.7667} \\ &= 11.4818 \end{aligned}$$

บริเวณวิกฤติ: ปฏิเสธ H_0 ถ้า $|t_c| \geq 2.306$

เนื่องจากค่า t_c ที่คำนวณได้ตกอยู่ในบริเวณวิกฤติ ดังนั้นปฏิเสธ H_0 นั่นคือ สมการถดถอยไม่ผ่านจุดกำเนิด ที่ระดับนัยสำคัญ 0.05

2.5 การประมาณช่วงความเชื่อมั่นของ $E(Y \mid X_0)$

จุดมุ่งหมายหนึ่งของการวิเคราะห์การถดถอยก็คือ การประมาณค่าเฉลี่ยของการแจกแจงของ Y เมื่อกำหนดค่าตัวแปรอิสระให้ ในที่นี้ให้ X_0 แทน ค่าของตัวแปรอิสระ ซึ่ง X_0 อาจเป็นค่าใดค่าหนึ่งในตัวอย่าง หรืออาจเป็นค่าอื่นของตัวแปรอิสระที่ยังคงอยู่ในขอบเขตของการศึกษา และให้ $E(Y \mid X_0)$ แทน ค่าเฉลี่ยของ Y เมื่อกำหนด $X = X_0$ โดยที่ $E(Y \mid X_0) = \beta_0 + \beta_1 X_0$ นอกจากนี้ยังสามารถแสดงได้ว่า $\hat{Y}_0 = b_0 + b_1 X_0$

เป็นตัวประมาณแบบจุดที่ไม่เอนเอียงของ $E(Y \mid X_0)$ นั่นคือ

$$\begin{aligned} E(\hat{Y}_0) &= E(b_0 + b_1 X_0) \\ &= E(b_0) + X_0 E(b_1) \\ &= \beta_0 + \beta_1 X_0 \\ &= E(Y \mid X_0) \end{aligned}$$

2.5.1 การแจกแจงตัวอย่างของ $\hat{Y}_0$ (Sampling Distribution of $\hat{Y}_0$)

การแจกแจงของ $\hat{Y}_0$ เกิดจากค่าที่แตกต่างกันของ $\hat{Y}_0$ ซึ่งได้จากตัวอย่างสุ่มที่ถูกเลือกมาซ้ำ ๆ กัน เมื่อให้ระดับของตัวแปรอิสระคงที่ แล้วคำนวณค่า $\hat{Y}_0$ จากตัวอย่างแต่ละชุด สำหรับโมเดลถดถอยที่ความคลาดเคลื่อนมีการแจกแจงแบบปกติ พบว่า $\hat{Y}_0$ มีการแจกแจงแบบปกติ ด้วยค่าเฉลี่ยและความแปรปรวน ดังนี้

$$E(\hat{Y}_0) = \beta_0 + \beta_1 X_0 \tag{2.32}$$

$$\sigma^2_{\hat{Y}_0} = \sigma^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right) \tag{2.33}$$

เนื่องจาก $\bar{Y}$ เป็นอิสระจาก b_1 นั่นคือ $Cov(\bar{Y}, b_1) = 0$ ดังนั้นความแปรปรวนของ $\hat{Y}_0$ ใน (2.33) เกิดจาก

$$\begin{aligned} \sigma^2_{\hat{Y}_0} = V(\hat{Y}_0) &= V(b_0 + b_1 X_0) \\ &= V\left((\bar{Y} - b_1 \bar{X}) + b_1 X_0\right) \\ &= V\left(\bar{Y} + b_1 (X_0 - \bar{X})\right) \\ &= V(\bar{Y}) + (X_0 - \bar{X})^2 V(b_1) \\ &= \frac{\sigma^2}{n} + (X_0 - \bar{X})^2 \frac{\sigma^2}{S_{xx}} \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}} \right) \end{aligned}$$

ทำนองเดียวกันกับการประมาณค่าความแปรปรวนของ b_0 และ b_1 หลังจากแทนค่า σ^2 ด้วย MSE จะได้ตัวประมาณค่าความแปรปรวนของ $\hat{Y}_0$ เป็น

$$S^2_{\hat{Y}_0} = MSE \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}} \right) \tag{2.34}$$

จะเห็นได้ว่า ถ้า X_0 ห่างจากค่าเฉลี่ยมากขึ้น ค่าประมาณความแปรปรวนของ $\hat{Y}_0$ จะมีค่าสูงขึ้น โดยที่ค่าประมาณความแปรปรวนของ $\hat{Y}_0$ จะมีค่าต่ำที่สุดเมื่อ $X_0 = \bar{X}$ และถ้า $X_0 = 0$ แล้ว $S^2_{\hat{Y}_0}$ จะลดรูปเป็น $S^2_{b_0}$
ดังนั้น

$$\frac{\hat{Y}_0 - E(Y \mid X_0)}{S_{\hat{Y}_0}} \sim t_{n-2} \tag{2.35}$$

2.5.2 ช่วงความเชื่อมั่นของ $E(Y \mid X_0)$

การสร้างช่วงความเชื่อมั่น $(1-\alpha)100\%$ ของ $E(Y \mid X_0)$ สามารถทำได้โดยใช้การแจกแจงแบบ t เช่นเดียวกับช่วงความเชื่อมั่นของ β_0 และ β_1 นั่นคือ

$$\hat{Y}_0 \pm t_{\alpha/2,n-2} \cdot S_{\hat{Y}_0} \tag{2.36}$$

นอกจากนี้ยังพบว่า ช่วงความเชื่อมั่นของ $E(Y \mid X_0)$ จะไม่ถูกกระทบมากนัก เมื่อข้อมูลเบี่ยงเบนไปจากข้อกำหนดของการวิเคราะห์เกี่ยวกับแจกแจงแบบปกติเพียงเล็กน้อยจนถึงปานกลาง

2.6 การพยากรณ์ค่าสังเกตค่าใหม่ (Prediction of New Observations)

การพยากรณ์ค่าสังเกตค่าใหม่จะถูกพิจารณาว่าเป็นผลที่ได้จากการทดลองหรือการเก็บข้อมูลครั้งใหม่ ซึ่งเป็นอิสระจากการทดลองหรือการเก็บข้อมูลที่ใช้ในการสร้างสมการถดถอย ให้ X_0 แทน ค่าของตัวแปรอิสระที่ได้จากการเก็บข้อมูลครั้งใหม่ และ $Y_{0(new)}$ แทน ค่าของตัวแปรตามค่าใหม่ ในที่นี้สมมติให้รูปแบบของโมเดลถดถอยที่ได้จากข้อมูลชุดเดิมยังคงใช้ได้กับค่าสังเกตค่าใหม่ ซึ่งการประมาณค่าเฉลี่ยของ Y แตกต่างจากการพยากรณ์ค่าสังเกตค่าใหม่ในแง่ที่ กรณีแรกเป็นการประมาณค่าเฉลี่ยของการแจกแจงของ Y ส่วนกรณีหลังเป็นการพยากรณ์ค่าสังเกตแต่ละค่าที่เลือกมาจากการแจกแจงของ Y และการสร้างช่วงพยากรณ์ก็เพื่อเลือกพิสัยของการแจกแจงของ Y โดยที่ค่าสังเกตส่วนใหญ่ตกอยู่ในช่วงดังกล่าว แล้วอ้างว่าค่าสังเกตค่าใหม่จะตกอยู่ในช่วงนี้ด้วย

เนื่องจากค่าเฉลี่ยของการแจกแจงของ Y ไม่ทราบค่า ซึ่งถูกประมาณด้วย $\hat{Y}_0$ ตำแหน่งของการแจกแจงของ Y จึงไม่สามารถถูกกำหนดได้อย่างแน่นอน ดังนั้นสิ่งที่จะต้องพิจารณาในการสร้างช่วงแห่งการพยากรณ์จะประกอบด้วย

- ความผันแปรเนื่องจากตำแหน่งที่เป็นไปได้ของการแจกแจงของ Y
- ความผันแปรภายในการแจกแจงความน่าจะเป็นของ Y

2.6.1 การแจกแจงตัวอย่างของ $Y_{0(new)}$ (Sampling Distribution of $Y_{0(new)}$)

ให้ตัวประมาณแบบจุดของ $Y_{0(new)}$ คือ $\hat{Y}_0$ และเนื่องจากค่าสังเกตค่าใหม่เป็นอิสระกับข้อมูล n ค่าที่ใช้ในการหา $\hat{Y}_0$ ดังนั้นความแปรปรวนของความคลาดเคลื่อนในการพยากรณ์ (Variance of prediction error) ซึ่งแทนด้วย $\sigma^2_{Y_0}$ สามารถคำนวณได้จาก

$$\begin{aligned}
\sigma^2_{Y_0} &= V(Y_{0(new)} - \hat{Y}_0) \\
&= V(Y_{0(new)}) + V(\hat{Y}_0) \\
&= \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}\right) \\
&= \sigma^2\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}\right)
\end{aligned}$$

หลังจากแทนค่า σ^2 ด้วย MSE จะได้ค่าประมาณความแปรปรวนของความคลาดเคลื่อนในการพยากรณ์ แทนด้วย $S^2_{Y_0}$ ซึ่งเป็นตัวประมาณของ $\sigma^2_{Y_0}$ ดังนี้

$$S^2_{Y_0} = MSE\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}\right) \tag{2.37}$$

และได้การแจกแจงตัวอย่างของ $Y_{0(new)}$ เป็น

$$\frac{Y_{0(new)} - \hat{Y}_0}{S_{Y_0}} \sim t_{n-2} \tag{2.38}$$

เมื่อ S_{Y_0} แทน ค่าประมาณของความคลาดเคลื่อนในการพยากรณ์

จะเห็นได้ว่าตัวเศษในสมการ (2.38) แสดงถึงความแตกต่างระหว่างค่าสังเกตค่าใหม่ $Y_{0(new)}$ กับค่าประมาณของค่าเฉลี่ย $\hat{Y}_0$ โดยอิงข้อมูลชุดเดิมที่มีขนาด n นอกจากนี้ยังพบว่า $\hat{Y}_0$ เป็นตัวประมาณแบบจุดที่ดีที่สุดของ $Y_{0(new)}$ อีกด้วย

2.6.2 การประมาณช่วงแห่งการพยากรณ์ของ $Y_{0(new)}$ (Estimation of Prediction Interval for $Y_{0(new)}$)

การสร้างช่วงแห่งการพยากรณ์สามารถทำได้ในลักษณะเดียวกับการสร้างช่วงความเชื่อมั่น ซึ่งได้จากการสุ่มตัวอย่างซ้ำ ๆ กัน โดยที่ตัวแปรอิสระ X ในตัวอย่างละชุดมีค่าเหมือนกัน จากนั้นจึงคำนวณช่วงแห่งการพยากรณ์ของตัวอย่างแต่ละชุด ซึ่งจะได้ว่าช่วงแห่งการพยากรณ์ $(1-\alpha)100\%$ ของ $Y_{0(new)}$ คือ

$$\hat{Y}_0 \pm t_{\alpha/2,\,n-2} \cdot S_{Y_0} \tag{2.39}$$

หมายเหตุ

- ช่วงแห่งการพยากรณ์ของ $Y_{0(new)}$ จะกว้างกว่าช่วงแห่งความเชื่อมั่นของ $E(Y \mid X_0)$ เสมอ เนื่องจากช่วงแห่งการพยากรณ์จะรวมเอาความผันแปร 2 ส่วนไว้ด้วยกัน นั่นคือ ความผันแปรใน $\hat{Y}_0$ ที่ได้จากตัวอย่างแต่ละชุด และความผันแปรของการแจกแจงของ Y
- ช่วงแห่งการพยากรณ์จะกว้างขึ้น เมื่อ X_0 มีค่าแตกต่างจาก $\bar{X}$ มากขึ้น และจะมีค่าน้อยที่สุด เมื่อ $X_0 = \bar{X}$
- ช่วงแห่งการพยากรณ์ค่อนข้างไวต่อการเบี่ยงเบนของข้อมูลจากการแจกแจงแบบปกติ ซึ่งแตกต่างจากช่วงแห่งความเชื่อมั่นของ $E(Y \mid X_0)$
- ช่วงแห่งการพยากรณ์คล้ายคลึงกับช่วงแห่งความเชื่อมั่น แต่จะแตกต่างกันในแง่ที่ ช่วงแห่งความเชื่อมั่นเป็นการอนุมานเกี่ยวกับพารามิเตอร์ และเป็นการหาช่วงที่คาดว่าจะครอบคลุมค่าที่แท้จริงของพารามิเตอร์ ในทางตรงข้ามช่วงแห่งการพยากรณ์เป็นข้อความเกี่ยวกับค่าของตัวแปรสุ่ม ซึ่งก็คือค่าสังเกตค่าใหม่ $Y_{0(new)}$ นั่นเอง

ตัวอย่างที่ 2.5 บริษัทแห่งหนึ่งมีค่าใช้จ่ายในการโฆษณาเป็น 105,000 บาท จงหา

1. ช่วงความเชื่อมั่น 95% ของยอดขายเฉลี่ย เมื่อมีค่าใช้จ่ายในการโฆษณาเป็น 105,000 บาท
2. ช่วงแห่งการพยากรณ์ 95% ของยอดขาย เมื่อมีค่าใช้จ่ายในการโฆษณาเป็น 105,000 บาท

วิธีทำ

1. จากตัวอย่างที่ 2.1 ได้สมการถดถอยเชิงเส้นตรงอย่างง่ายแสดงความสัมพันธ์ระหว่างยอดขายและค่าใช้จ่ายในการโฆษณา เป็น $\hat{Y} = 66.2121 + 4.4303X$
   เมื่อ $X_0 = 10.5$ จะได้

$$\hat{Y}_0 = 66.2121 + 4.4303(10.5) = 112.7303$$

คำนวณค่าประมาณความแปรปรวนของ $\hat{Y}_0$

$$\begin{aligned} S^2_{\hat{Y}_0} &= MSE\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}\right) \\ &= 27.8530\left(\frac{1}{10} + \frac{(10.5 - 9.5)^2}{82.5}\right) \\ &= 3.1229 \\ S_{\hat{Y}_0} &= 1.7672 \end{aligned}$$

ค่าวิกฤติ $t_{0.025,8} = 2.306$

ดังนั้นช่วงความเชื่อมั่น 95% ของ $E(Y \mid X = 10.5)$ คือ

$$\begin{aligned} \hat{Y}_0 \pm t_{\alpha/2,n-2} \cdot S_{\hat{Y}_0} &= 112.7303 \pm (2.306)(1.7672) \\ &= 112.7303 \pm 4.0752 \\ \text{หรือ} \quad &(108.6551,\ 116.8055) \end{aligned}$$

นั่นคือ ด้วยความเชื่อมั่น 95% บริษัทแห่งใหม่ที่มีงบโฆษณาเป็น 105,000 บาท มียอดขายเฉลี่ยอยู่ระหว่าง $1,086,551$ ถึง $1,168,055$ บาท

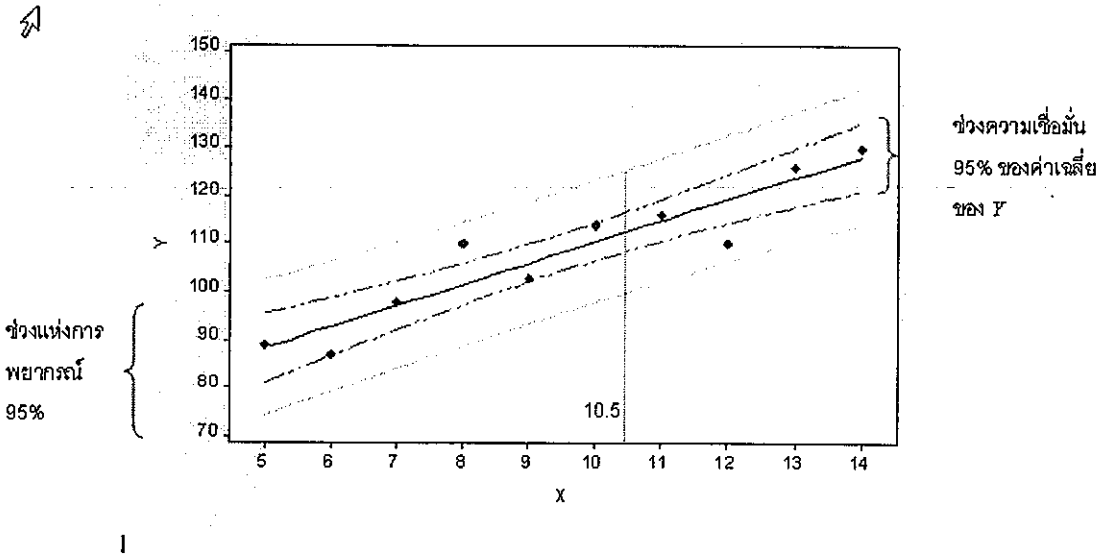
2. คำนวณค่าประมาณความแปรปรวนของความคลาดเคลื่อนในการพยากรณ์

$$\begin{aligned} S^2_{Y_0} &= MSE\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}\right) \\ &= 27.8530\left(1 + \frac{1}{10} + \frac{(10.5 - 9.5)^2}{82.5}\right) \\ &= 30.9759 \\ S_{Y_0} &= 5.5656 \end{aligned}$$

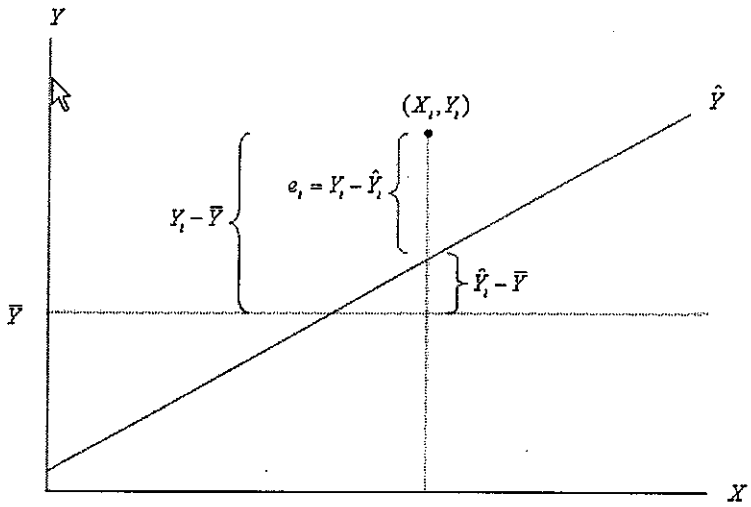
และได้ช่วงแห่งการพยากรณ์ 95% ของ $Y_{0(new)}$ เป็น

$$\begin{aligned} \hat{Y}_0 \pm t_{\alpha/2,n-2} \cdot S_{Y_0} &= 112.7303 \pm (2.306)(5.5656) \\ &= 112.7303 \pm 12.8343 \\ \text{หรือ} \quad &(99.896,\ 125.5646) \end{aligned}$$

นั่นคือ ด้วยความเชื่อมั่น 95% การพยากรณ์ยอดขายของบริษัทแห่งใหม่ที่มีงบโฆษณาเป็น 105,000 บาท มีค่าอยู่ระหว่าง $998,960$ บาท ถึง $1,255,646$ บาท ซึ่งจะเห็นได้ว่าที่สัมประสิทธิ์ความเชื่อมั่นขนาดเท่ากัน ช่วงแห่งการพยากรณ์จะกว้างกว่าช่วงความเชื่อมั่น ซึ่งสอดคล้องกับรูปที่ 2.4 โดยที่ช่วงแห่งการพยากรณ์จะกว้างกว่าช่วงแห่งความเชื่อมั่นในทุกค่าของตัวแปร X



รูปที่ 2.4: ช่วงความเชื่อมั่น 95% ของค่าเฉลี่ยของ Y และช่วงแห่งการพยากรณ์ 95% ของ Y



รูปที่ 2.5: การแบ่งผลรวมกำลังสองทั้งหมดของตัวแปรตาม Y

2.7 การแบ่งผลรวมกำลังสองทั้งหมด (Partitioning of Total Sum of Squares)

จุดมุ่งหมายประการหนึ่งของการวิเคราะห์การถดถอยก็เพื่อพยากรณ์ค่าตัวแปรตาม Y ซึ่งหากไม่มีตัวแปรอิสระเข้ามาเกี่ยวข้องแล้ว การประมาณค่าของตัวแปรตาม Y จะใช้ค่าเฉลี่ย $\bar{Y}$ นั่นคือ ค่าพยากรณ์ของ Y มีค่าคงที่ไม่ว่า X จะมีค่าเป็นอะไรก็ตาม ซึ่งแสดงว่า Y และ X ไม่มีความสัมพันธ์ต่อกัน แต่ถ้า Y ขึ้นอยู่กับค่า X แล้ว ลักษณะความสัมพันธ์สามารถที่จะอธิบายได้ด้วยสมการถดถอย ดังนั้นความผันแปรทั้งหมดในตัวแปรตาม Y สามารถแยกออกได้เป็นสองส่วน คือ ส่วนที่สามารถอธิบายได้ด้วยสมการถดถอย และอีกส่วนซึ่งไม่สามารถอธิบายได้ โดยเกิดจากความแตกต่างระหว่างค่าจริงและค่าพยากรณ์ ดังแสดงในรูปที่ 2.5 ซึ่งสามารถเขียนสมการได้ดังนี้

$$Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) \tag{2.40}$$

ยกกำลังสองทั้งสองข้าง แล้วหาผลรวมตั้งแต่ $1, 2, \ldots, n$ จะได้

$$\begin{aligned} \sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}\Big((\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)\Big)^2 \\ \sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \end{aligned} \tag{2.41}$$

เนื่องจากเทอม Cross-product มีค่าเท่ากับศูนย์ ดังนี้

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

นั่นคือ

$$\begin{aligned} \sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \\ SST &= SSR + SSE \end{aligned} \tag{2.42}$$

เมื่อ

SST แทน ผลรวมกำลังสองทั้งหมด (Total sum of squares or corrected sum of squares)

SSR แทน ผลรวมกำลังสองเนื่องมาจากสมการถดถอย (Sum of squares due to regression)

SSE แทน ผลรวมกำลังสองเนื่องมาจากความคลาดเคลื่อน (Sum of squares due to error)

เพื่อสะดวกต่อการคำนวณ สามารถเขียนสูตรข้างต้นใหม่ได้ดังนี้

$$\begin{aligned} SST &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \\ &= \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 \\ &= S_{yy} \end{aligned} \tag{2.43}$$

$$\begin{aligned} SSR &= \sum_{i=1}^{n}(\bar{Y}_i - \bar{Y})^2 \\ &= \sum_{i=1}^{n}\left\{(b_0 + b_1 X_i) - \bar{Y}\right\}^2 \\ &= \sum_{i=1}^{n}\left\{(\bar{Y} - b_1\bar{X}) + b_1 X_i - \bar{Y}\right\}^2 \\ &= b_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2 \\ &= b_1^2 S_{xx} = b_1 S_{xy} = \frac{(S_{xy})^2}{S_{xx}} \end{aligned} \tag{2.44}$$

$$SSE = SST - SSR \tag{2.45}$$

นอกจากนี้ค่าผลรวมกำลังสองแต่ละตัว มีจำนวนองศาความเป็นอิสระที่สอดคล้องกันติดอยู่ โดย SST มีจำนวนองศาความเป็นอิสระเท่ากับ $n-1$ เนื่องจากนิยามของ $\sum_{i=1}^{n}(Y_i - \bar{Y})$ จะต้องมีค่าเท่ากับศูนย์ และ SSR มีจำนวนองศาความเป็นอิสระเท่ากับ 1 เนื่องจากค่า $\hat{Y}_i$ คำนวณมาจากสมการถดถอย ซึ่งมีจำนวนองศาความเป็นอิสระเท่ากับ 2 สอดคล้องกับการประมาณค่า β_0 และ β_1 แต่จำนวนองศาความเป็นอิสระที่หายไป 1 ค่า เนื่องจากเงื่อนไข $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})$ จะต้องมีค่าเท่ากับศูนย์ ส่วน SSE มีจำนวนองศาความเป็นอิสระเท่ากับ $n-2$ ซึ่งจำนวนองศาความเป็นอิสระที่หายไป 2 ค่า เนื่องมาจากการประมาณค่า β_0 และ β_1 ในสมการถดถอยนั่นเอง ดังนั้นจำนวนองศาความเป็นอิสระสามารถแบ่งได้ดังนี้

$$\begin{aligned} SST &= SSR + SSE \\ n-1 &= \quad 1 \quad + (n-2) \end{aligned} \tag{2.46}$$

หากหารผลรวมกำลังสองแต่ละเทอมด้วยจำนวนองศาความเป็นอิสระที่สอดคล้องกัน จะเรียกค่าที่ได้นี้ว่า *ค่าเฉลี่ยกำลังสอง* (Mean square) นิยมเขียนย่อ ๆ ว่า MS นั่นคือ

$$MSR = \frac{SSR}{1} = SSR \tag{2.47}$$

$$MSE = \frac{SSE}{n-2} \tag{2.48}$$

เมื่อ

MSR แทน ค่าเฉลี่ยกำลังสองเนื่องมาจากสมการถดถอย (Mean squares due to regression)

MSE แทน ค่าเฉลี่ยกำลังสองเนื่องมาจากความคลาดเคลื่อน (Mean squares due to error)

จะเห็นได้ว่าความผันแปรของค่าสังเกตทั้งหมดสามารถแยกได้เป็นส่วน ๆ ตามแหล่งที่มาของความผันแปรนั้น ๆ ดังแสดงในตารางวิเคราะห์ความแปรปรวน (Analysis of Variance) ซึ่งนิยมเรียกย่อ ๆ ว่า ตาราง ANOVA ดังนี้

ANOVA

Source of variation	df	SS	MS	F
Regression	1	SSR	MSR	$F_c = \frac{MSR}{MSE}$
Error	$n-2$	SSE	MSE	
Total	$n-1$	SST		

การทดสอบความสัมพันธ์เชิงเส้นตรงระหว่าง Y และ X นอกจากใช้สถิติทดสอบ t ดังได้กล่าวมาแล้วยังสามารถทำได้โดยใช้การวิเคราะห์ความแปรปรวน โดยมีสมมติฐานของการทดสอบดังนี้

$$H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0$$

ซึ่งการอนุมานทางสถิติโดยใช้การวิเคราะห์ความแปรปรวน จำเป็นต้องทราบค่าคาดหวังของค่าเฉลี่ยกำลังสอง เนื่องจาก $\frac{SSE}{\sigma^2} \sim \chi^2_{n-2}$ ดังนั้น

$$E\left(\frac{SSE}{\sigma^2}\right) = n-2$$

$$E\left(\frac{SSE}{n-2}\right) = \sigma^2 = E(MSE) \tag{2.49}$$

และ $SSR = b_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2$ จะได้ว่า

$$\begin{aligned} E(SSR) &= \sum_{i=1}^{n}(X_i - \bar{X})^2 E(b_1^2) \\ E(b_1^2) &= V(b_1) + \{E(b_1)\}^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + \beta_1^2 \end{aligned}$$

ดังนั้น

$$E\left(\frac{SSR}{1}\right) = E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{2.50}$$

จะเห็นได้ว่า ค่าเฉลี่ยของการแจกแจงตัวอย่างของ MSE มีค่าเป็น σ^2 ไม่ว่า X และ Y จะมีความสัมพันธ์ เชิงเส้นตรงต่อกันหรือไม่ก็ตาม และเมื่อ $\beta_1 = 0$ ค่าเฉลี่ยของการแจกแจงตัวอย่างของ MSR มีค่าเป็น σ^2 เช่นเดียวกัน แต่เมื่อ $\beta_1 \neq 0$ พบว่า MSR จะมีค่ามากกว่า MSE ดังนั้นการเปรียบเทียบ MSR กับ MSE จึงมีประโยชน์ในการตรวจสอบว่า $\beta_1 = 0$ หรือไม่ ถ้า MSR มีค่าใกล้เคียงกับ MSE มีแนวโน้มที่ $\beta_1 = 0$ ในทางตรงข้ามหาก MSR มีค่ามากกว่า MSE มาก ๆ มีแนวโน้มที่ $\beta_1 \neq 0$

พิจารณาการแจกแจงของตัวสถิติ F_c เมื่อ H_0 เป็นจริง และกำหนดให้ $\epsilon \sim NID(0, \sigma^2)$ พบว่า

$$\begin{aligned} \frac{SSR}{\sigma^2} \Big/ 1 &\sim \chi_1^2 \\ \frac{SSE}{\sigma^2} \Big/ (n-2) &\sim \chi_{n-2}^2 \end{aligned}$$

และเป็นอิสระจากกัน ดังนั้น

$$F_c = \frac{MSR}{MSE} = \frac{\frac{SSR}{\sigma^2} \Big/ 1}{\frac{SSE}{\sigma^2} \Big/ (n-2)} \sim F_{1,n-2\,(\alpha)} \tag{2.51}$$

โดย $F_{1,n-2}$ แทน การแจกแจงแบบ F (F distribution) ที่ระดับนัยสำคัญ α และจำนวนองศาความเป็นอิสระ $(1,\, n-2)$

เกณฑ์การตัดสินใจ คือ หาก $F_c \geq F_{1,n-2}$ จะปฏิเสธ H_0 และหาก $F_c < F_{1,n-2}$ จะยอมรับ H_0

ตัวอย่างที่ 2.6 จากข้อมูลในตัวอย่างที่ 2.1 จงทดสอบว่าค่าใช้จ่ายในการโฆษณามีความสัมพันธ์เชิงเส้นตรงต่อยอดขายของบริษัทหรือไม่ โดยใช้การวิเคราะห์ความแปรปรวน กำหนดระดับนัยสำคัญของการทดสอบเป็น 0.05

วิธีทำ สมมติฐานของการทดสอบคือ

H_0 : ค่าใช้จ่ายในการโฆษณามีความสัมพันธ์เชิงเส้นตรงต่อยอดขาย

H_1 : ค่าใช้จ่ายในการโฆษณาไม่มีความสัมพันธ์เชิงเส้นตรงต่อยอดขาย

ซึ่งสามารถเขียนให้อยู่ในรูปสมมติฐานทางสถิติได้ดังนี้

$$H_0 : \beta_1 = 0 \qquad vs. \qquad H_1 : \beta_1 \neq 0$$

คำนวณค่าต่าง ๆ และแทนค่าลงในตารางวิเคราะห์ความแปรปรวน

$$SST = S_{yy} = 1,842.1$$

$$SSR = b_1 S_{xy} = (4.4303)(365.5) = 1,619.2747$$

$$SSE = SST - SSR = 1.842.1 - 1,619.2747 = 222.8253$$

ANOVA

Source of variation	***df***	***SS***	***MS***	***F***
Regression	1	1,619.2747	1,619.2747	58.1361
Error	8	222.8253	27.8253	
Total	9	1,842.1		

เนื่องจาก $F_c = 58.1361 > F_{1,8\ (0.05)} = 5.32$ จึงปฏิเสธ H_0 นั่นคือ ค่าใช้จ่ายในการโฆษณามีความสัมพันธ์เชิงเส้นตรงต่อยอดขายอย่างมีนัยสำคัญทางสถิติ ที่ระดับ 0.05

หมายเหตุ ที่ระดับนัยสำคัญ α จะเห็นได้ว่าการทดสอบ $H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0$ สามารถทำได้โดยใช้สถิติทดสอบ F หรือ t ดังนั้นผลการทดสอบที่ได้จากสถิติทั้งสองจะต้องสอดคล้องกัน เนื่องจาก

$$
\begin{aligned}
F_c &= \frac{SSR/1}{SSE/(n-2)} \\
&= \frac{b_1^2 S_{xx}}{MSE} \\
&= \frac{b_1^2}{MSE/S_{xx}} \\
&= \frac{b_1^2}{S_{b_1}^2} \\
&= \left(\frac{b_1}{S_{b_1}}\right)^2 \\
&= t_c^2
\end{aligned}
$$

จากความสัมพันธ์ระหว่าง F_c และ t_c ทำนองเดียวกันจะได้ว่า $F_{1,n-2,(\alpha)} = t^2_{n-2\,(\frac{\alpha}{2})}$ แต่การใช้สถิติทดสอบ F และ t แตกต่างกันในแง่ที่ การทดสอบโดยใช้สถิติทดสอบ t เป็นการทดสอบแบบสองทาง ในขณะที่การทดสอบโดยใช้สถิติทดสอบ F เป็นการทดสอบแบบทางเดียว นอกจากนี้การทดสอบโดยใช้สถิติทดสอบ t จะค่อนข้างยืดหยุ่นมากกว่าการใช้สถิติทดสอบ F เนื่องจากสามารถใช้ในการทดสอบสมมติฐานแบบทางเดียว ($H_1 : \beta_1 > 0$ หรือ $H_1 : \beta_1 < 0$) ได้อีกด้วย ในขณะที่สถิติทดสอบ F ไม่สามารถทำได้

2.8 ค่าสัมประสิทธิ์ตัวกำหนด (Coefficient of Determination)

การพิจารณาว่าสมการถดถอยที่ได้เหมาะสมกับข้อมูลเพียงไร สามารถทำได้โดยพิจารณาจากส่วนแบ่งของผลรวมกำลังสองทั้งหมด หรือ SST ว่าเป็นผลเนื่องมาจากผลรวมกำลังสองของเส้นถดถอย หรือ SSR และผลรวมกำลังสองของความคลาดเคลื่อน หรือ SSE มากน้อยแค่ไหน ซึ่งจากหัวข้อที่แล้วทราบว่า ค่า SST ใช้วัดความผันแปรของค่าสังเกต Y หรือวัดความไม่แน่นอนในการพยากรณ์ค่า Y เมื่อไม่มีตัวแปรอิสระ X เข้ามาเกี่ยวข้อง ทำนองเดียวกันค่า SSE ใช้วัดความผันแปรของค่าสังเกต Y เมื่อมีตัวแปรอิสระ X เข้ามาเกี่ยวข้องโดยผ่านทางโมเดลถดถอย ดังนั้นการวัดอิทธิพลของ X ที่ช่วยลดความผันแปรใน Y หรือลดความไม่แน่นอนในการพยากรณ์ ค่า Y จะแสดงในรูปของสัดส่วนระหว่าง SSR ต่อ SST ซึ่งหาก SSR มีค่ามาก แสดงว่าสมการถดถอยที่ได้สามารถอธิบายความสัมพันธ์ของข้อมูลชุดนั้นได้ดี นั่นคือ

$$
R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{2.52}
$$

โดยเรียก R^2 ว่า *ค่าสัมประสิทธิ์ตัวกำหนด* หรือ *ค่าสัมประสิทธิ์การตัดสินใจ* เป็นค่าที่ใช้วัดสัดส่วนความผันแปรของ Y ที่สามารถอธิบายได้ด้วยสมการถดถอย นิยมแสดงในรูปของเปอร์เซ็นต์โดยคูณเข้าด้วย 100

เนื่องจาก $0 \leq SSE \leq SST$ ดังนั้น $0 \leq R^2 \leq 1$ และสามารถอธิบายได้ดังนี้

- หากสมการถดถอยขนานแกน X นั่นคือ $b_1 = 0$ และ $\hat{Y}_i = \bar{Y}$ ดังนั้น $SST = SSE$ ซึ่งส่งผลให้ $R^2 = 0$ แสดงว่าตัวแปรอิสระ X และตัวแปรตาม Y ไม่มีความสัมพันธ์เชิงเส้นตรงต่อกัน นั่นคือ สมการถดถอยเชิงเส้นตรงไม่สามารถใช้พยากรณ์ค่า Y ได้
- หากค่าสังเกตทุกค่าตกอยู่บนเส้นถดถอยแล้ว $SSE = 0$ ซึ่งส่งผลให้ $R^2 = 1$ แสดงว่าค่าสังเกต Y_i มีค่าเท่ากับค่าพยากรณ์ $\hat{Y}_i$ ทุกค่า นั่นคือ สมการถดถอยเชิงเส้นตรงสามารถใช้พยากรณ์ค่า Y ได้อย่างสมบูรณ์
- ในทางปฏิบัติเป็นไปได้ยากที่ R^2 มีค่าเป็น 0 หรือ 1 โดยส่วนใหญ่จะมีค่าอยู่ระหว่าง 0 และ 1 และใช้วัดขนาดความสัมพันธ์ระหว่างตัวแปรอิสระ X และตัวแปรตาม Y ถ้า R^2 มีค่าเข้าใกล้ 1 แสดงว่า X และ Y มีความสัมพันธ์เชิงเส้นตรงต่อกันค่อนข้างสูง นั่นคือ ความผันแปรใน Y ส่วนใหญ่สามารถอธิบายได้ด้วยสมการถดถอยเชิงเส้นตรง

หากตัวอย่างมีขนาดเล็ก ค่า R^2 ที่คำนวณได้อาจมีค่าสูงเกินจริง จึงต้องมีการปรับค่าด้วยการหาร SSE และ SST ด้วยจำนวนองศาความเป็นอิสระที่สอดคล้องกัน และแทนด้วย R^2_{adj} (Adjusted coefficient of determination) นั่นคือ

$$R^2_{adj} = 1 - \frac{SSE/(n-2)}{SST/(n-1)} = 1 - \frac{(n-1)}{(n-2)} \cdot \frac{SSE}{SST} \tag{2.53}$$

อย่างไรก็ตามเมื่อตัวอย่างมีขนาดใหญ่ ค่า R^2 และ R^2_{adj} จะมีค่าใกล้เคียงกัน

ตัวอย่างที่ 2.7 จงหาค่าสัมประสิทธิ์ตัวกำหนดของข้อมูลในตัวอย่างที่ 2.1 พร้อมสรุปผลที่ได้

วิธีทำ จาก $SSR = 1,619.2747$ และ $SST = 1,842.1$

ดังนั้น

$$R^2 = \frac{SSR}{SST} = \frac{1,619.2747}{1,842.1} = 0.8790$$

นั่นคือ ความผันแปรทั้งหมดของยอดขายสามารถอธิบายได้ด้วยค่าใช้จ่ายในการโฆษณา 87.90%

($R^2_{adj} = 0.8640$)

2.9 สมการถดถอยผ่านจุดกำเนิด (Regression through Origin)

ในบางสถานการณ์อาจพบว่าสมการถดถอยผ่านจุดกำเนิดเป็นรูปแบบที่เหมาะสมกับลักษณะของข้อมูล โดยส่วนใหญ่มักเกิดขึ้นกับข้อมูลทางด้านเคมีและกระบวนการผลิตต่าง ๆ เช่น ปริมาณผลผลิตมีค่าเป็นศูนย์ เมื่อเวลาใน

การผลิตเป็นศูนย์ เป็นต้น

รูปแบบของโมเดลถดถอยอย่างง่ายผ่านจุดกำเนิด

$$Y_i = \beta_1 X_i + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{2.54}$$

จากการเก็บรวบรวมข้อมูลขนาด n คำนวณค่าผลรวมกำลังสองได้ดังนี้

$$SSE = \sum_{i=1}^{n} (Y_i - \beta_1 X_i)^2 \tag{2.55}$$

หาอนุพันธ์อันดับที่หนึ่งเทียบกับพารามิเตอร์ β_1 แล้วให้มีค่าเท่ากับศูนย์

$$\frac{\partial SSE}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_1 X_i) \; = \; 0$$

แทนค่า β_1 ด้วยค่าประมาณ b_1 จะได้สมการปกติเป็น

$$b_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i \tag{2.56}$$

ดังนั้น ตัวประมาณกำลังสองน้อยที่สุดของ β_1 คือ

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \tag{2.57}$$

และได้สมการถดถอยของข้อมูลตัวอย่างเป็น

$$\hat{Y}_i = b_1 X_i, \quad i = 1, 2, \ldots, n \tag{2.58}$$

2.9.1 การประมาณค่าความแปรปรวนของความคลาดเคลื่อน σ^2

ตัวประมาณแบบจุดของ σ^2 สำหรับสมการถดถอยผ่านจุดกำเนิด ซึ่งแทนด้วย MSE หรือ S^2 มีรูปแบบดังนี้

$$MSE = \frac{SSE}{n-1} = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n-1}$$

หรือเขียนอีกรูแบบหนึ่งได้เป็น

$$MSE = \frac{\sum_{i=1}^{n} Y_i^2 - b_1 \sum_{i=1}^{n} X_i Y_i}{n-1} \tag{2.59}$$

2.9.2 การหาช่วงความเชื่อมั่นของพารามิเตอร์ β_1

เมื่อกำหนดให้ความคลาดเคลื่อนมีการแจกแจงแบบปกติ การทดสอบสมมติฐานและสร้างช่วงความเชื่อมั่นเกี่ยวกับพารามิเตอร์ของโมเดลถดถอยผ่านจุดกำเนิดสามารถทำได้โดยใช้การแจกแจงแบบ t เช่นเดียวกัน ดังนั้นช่วงความเชื่อมั่น $(1-\alpha)100\%$ ของ β_1 คือ

$$b_1 \pm t_{\frac{\alpha}{2},n-1} \cdot S_{b_1} \tag{2.60}$$

เมื่อ $S_{b_1} = \sqrt{\frac{MSE}{\sum_{i=1}^{n} X_i^2}}$

2.9.3 การหาช่วงความเชื่อมั่นของ $E(Y \mid X_0)$

ช่วงความเชื่อมั่น $(1-\alpha)100\%$ ของ $E(Y \mid X_0)$ สำหรับสมการถดถอยผ่านจุดกำเนิด มีรูปแบบดังนี้

$$\hat{Y}_0 \pm t_{\frac{\alpha}{2},n-1} \cdot S_{\hat{Y}_0} \tag{2.61}$$

เมื่อ $S_{\hat{Y}_0} = \sqrt{MSE\left(\frac{X_0^2}{\sum_{i=1}^{n} X_i^2}\right)}$

2.9.4 การประมาณช่วงแห่งการพยากรณ์ของ $Y_{0(new)}$

ในทำนองเดียวกัน ช่วงแห่งการพยากรณ์ $(1-\alpha)100\%$ ของ $Y_{0(new)}$ สำหรับสมการถดถอยผ่านจุดกำเนิด มีรูปแบบดังนี้

$$\hat{Y}_0 \pm t_{\frac{\alpha}{2},n-1} \cdot S_{Y_0} \tag{2.62}$$

เมื่อ $S_{Y_0} = \sqrt{MSE\left(1 + \frac{X_0^2}{\sum_{i=1}^{n} X_i^2}\right)}$

จะเห็นได้ว่าความกว้างของช่วงความเชื่อมั่น (2.61) และช่วงแห่งการพยากรณ์ (2.62) มีค่ามากขึ้น เมื่อ X_0 มีค่าเพิ่มขึ้น นอกจากนี้ความยาวของช่วงความเชื่อมั่น (2.61) ที่จุด $X_0 = 0$ มีค่าเป็นศูนย์ เนื่องจากโมเดลถดถอยผ่านจุดกำเนิดกำหนดว่า ค่าเฉลี่ยของ Y ที่จุด $X = 0$ มีค่าเป็นศูนย์ ซึ่งแตกต่างจากโมเดลถดถอยที่มีระยะตัดแกน Y ในขณะที่ช่วงแห่งการพยากรณ์ (2.62) มีความยาวของช่วงที่ไม่เป็นศูนย์ที่จุด $X_0 = 0$ เนื่องมาจากความคลาดเคลื่อนสุ่มที่เกิดจากการพยากรณ์ค่าสังเกตในอนาคต

2.9.5 ค่าสัมประสิทธิ์ตัวกำหนด

การแบ่งผลรวมกำลังสองทั้งหมดสำหรับโมเดลถดถอยผ่านจุดกำเนิดทำได้ดังนี้

$$\begin{aligned} \sum_{i=1}^{n} Y_i^2 &= \sum_{i=1}^{n} \hat{Y}_i^2 + \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \\ SST &= SSR + SSE \end{aligned}$$

ดังนั้นค่าสัมประสิทธิ์ตัวกำหนดสสามารถคำนวณได้จากสัดส่วนระหว่างความผันแปรที่อธิบายได้ด้วยสมการถดถอยและความผันแปรทั้งหมด ดังนี้

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n} \hat{Y}_i^2}{\sum_{i=1}^{n} Y_i^2}$$

จะเห็นได้ว่าค่า R^2 สำหรับโมเดลถดถอยผ่านจุดกำเนิด เป็นค่าที่ใช้วัดสัดส่วนของความผันแปรรอบจุดกำเนิดที่สามารถอธิบายได้ด้วยสมการถดถอย โดยจะแตกต่างจากค่า R^2 ในโมเดลถดถอยที่มีระยะตัดแกน Y ซึ่งเป็นค่าที่ใช้วัดความผันแปรรอบค่าเฉลี่ย $\bar{Y}$ ที่สามารถอธิบายได้ด้วยสมการถดถอย มีรูปแบบเป็น

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}$$

ดังนั้นการเปรียบเทียบความเหมาะสมระหว่างโมเดลถดถอยที่มีระยะตัดแกน Y และโมเดลถดถอยผ่านจุดกำเนิดจึงไม่สามารถพิจารณาจากค่า R^2 ได้

ในการตัดสินใจว่าควรสร้างสมการถดถอยผ่านจุดกำเนิดหรือไม่นั้น สามารถพิจารณาได้จากแผนภาพการกระจาย หรือสร้างสมการถดถอยทั้งที่มีระยะตัดแกน Y และไม่มีระยะตัดแกน Y แล้วเปรียบเทียบคุณภาพของสมการ โดยเลือกสมการที่ให้ค่า MSE น้อยที่สุด หรืออาจสร้างสมการถดถอยที่มีระยะตัดแกน Y แล้วจึงทด-

สอบสมมติฐาน $H_0 : \beta_0 = 0$ ซึ่งมีขั้นตอนดังนี้

กำหนดสมมติฐานของการทดสอบ:

$$H_0 : \beta_0 = 0 \qquad vs. \qquad H_1 : \beta_0 \neq 0$$

คำนวณค่าต่าง ๆ ดังนี้

$$SSR = b_1 \sum_{i=1}^{n} X_i Y_i = \frac{\left(\sum_{i=1}^{n} X_i Y_i\right)^2}{\sum_{i=1}^{n} X_i^2} \tag{2.63}$$

$$SST = \sum_{i=1}^{n} Y_i^2 \tag{2.64}$$

$$SSE = SST - SSE \tag{2.65}$$

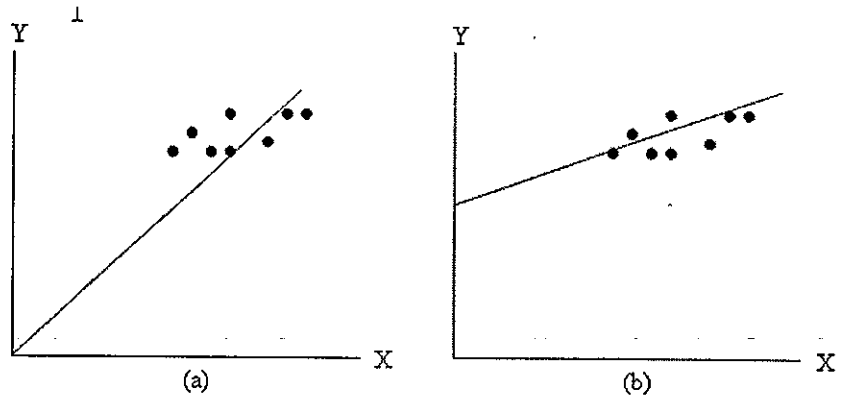
แล้วแทนค่าลงในตาราง ANOVA

ANOVA

Source of variation	df	SS	MS	F
Regression	1	SSR	MSR	$F_c = \frac{MSR}{MSE}$
Error	$n-1$	SSE	MSE	
Total	n	SST		

เกณฑ์การตัดสินใจ คือ ปฏิเสธ H_0 ถ้า $F_c \geq F_{1,n-1}$ และยอมรับ H_0 ถ้า $F_c < F_{1,n-1}$ หากผลการทดสอบยอมรับ H_0 แสดงว่ารูปแบบของสมการถดถอยผ่านจุดกำเนิดเหมาะสมแล้ว ในทางตรงกันข้ามถ้าผลการทดสอบปฏิเสธ H_0 แสดงว่ารูปแบบสมการถดถอยผ่านจุดกำเนิดนั้นไม่เหมาะสม ดังนั้นควรมีระยะตัดแกน Y อยู่ในสมการด้วย

นอกจากนี้การสร้างสมการถดถอยผ่านจุดกำเนิดควรทำด้วยความระมัดระวัง โดยเฉพาะอย่างยิ่งเมื่อค่าของตัวแปรอิสระ X ที่ใช้ในการสร้างสมการอยู่ห่างจากจุดกำเนิดมาก ๆ พิจารณารูปที่ 2.6 (a) ถึงแม้ว่าขอบเขตของ X ที่ใช้ในการสร้างสมการมีแนวโน้มที่จะมีความสัมพันธ์เชิงเส้นตรงกับ Y แต่การบังคับให้สมการผ่านจุดกำเนิด ทำให้สมการที่ได้นั้นไม่เหมาะสม ในขณะที่สมการถดถอยที่มีระยะตัดแกน Y แสดงความสัมพันธ์ของข้อมูลชุดนี้ได้เหมาะสมมากกว่า ดังแสดงในรูปที่ 2.6 (b) และบ่อยครั้งที่ความสัมพันธ์ระหว่าง Y และ X ที่อยู่ใกล้จุดกำเนิด มีลักษณะแตกต่างจากความสัมพันธ์ส่วนใหญ่ในขอบเขตของข้อมูลที่ศึกษา ดังนั้นการสร้างสมการถดถอยผ่านจุดกำเนิดควรทำเมื่อขอบเขตของ X ในข้อมูลมีค่าเข้าใกล้จุดกำเนิดมากพอ (Montgomery



รูปที่ 2.6: แสดงแผนภาพการกระจายและสมการถดถอยระหว่าง Y และ X สำหรับ (a) สมการถดถอยผ่านจุดกำเนิด (b) สมการถดถอยที่มีระยะตัดแกน Y

และ Peck, 1992)

ตัวอย่างที่ 2.8 ในการศึกษาความสัมพันธ์ระหว่างอุณหภูมิ (^oC) และปริมาณตะกอน (มิลลิกรัม) ของสารชนิดหนึ่งที่ได้จากการทำปฏิกริยาในแต่ละระดับอุณหภูมิ ได้ข้อมูลดังนี้

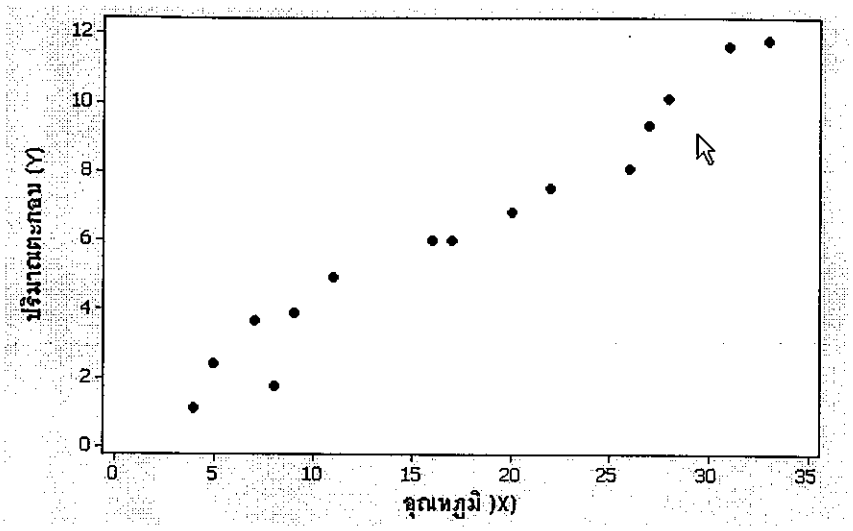
ปริมาณตะกอน (Y)	อุณหภูมิ (X)
10.10	28
3.91	9
4.95	11
6.83	20
2.43	5
5.99	16
8.09	26
11.81	33
11.64	31
5.99	17
7.52	22
3.69	7
9.33	27
1.11	4
1.79	8

ตารางที่ 2.3: ปริมาณตะกอนที่ได้จากการทำปฏิกริยาในแต่ละระดับอุณหภูมิ

จงสร้างสมการถดถอยแสดงความสัมพันธ์ระหว่างตัวแปรทั้งสอง พร้อมทั้งแปลผลที่ได้

วิธีทำ

จากแผนภาพการกระจายในรูปที่ 2.7 จะเห็นได้ว่าปริมาณตะกอนที่ได้จากการทำปฏิกริยาและอุณหภูมิมีความสัมพันธ์ไปในทิศทางเดียวกัน นั่นคือ เมื่ออุณหภูมิสูงขึ้น ปริมาณตะกอนมีแนวโน้มเพิ่มขึ้นตาม นอกจากนี้จะเห็นได้ว่ารูปแบบความสัมพันธ์มีลักษณะใกล้เคียงเส้นตรงและผ่านจุดกำเนิด และที่อุณหภูมิ $0\,^oC$ ปริมาณ



รูปที่ 2.7: แผนภาพการกระจายระหว่างปริมาณตะกอนที่ได้จากการทำปฏิกริยา (มิลลิกรัม) และอุณหภูมิ (^oC)

ตะกอนมีค่าเป็นศูนย์ด้วย ดังนั้นการสร้างสมการผ่านจุดกำเนิดสำหรับข้อมูลชุดนี้ค่อนข้างเหมาะสม คำนวณค่าต่าง ๆ ของข้อมูลในตารางที่ 2.3 ได้ดังนี้

$$\sum_{i=1}^{n} X_i = 264, \qquad \sum_{i=1}^{n} Y_i = 95.18, \qquad \sum_{i=1}^{n} X_i Y_i = 2,141.71,$$

$$\sum_{i=1}^{n} X_i^2 = 6,024, \qquad \sum_{i=1}^{n} Y_i^2 = 768.18, \qquad n = 15,$$

$$\bar{X} = 17.6, \qquad \bar{Y} = 6.3453$$

จะได้ว่า

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} = \frac{2,141.71}{6,024} = 0.3555$$

สมการถดถอยเชิงเส้นตรงอย่างง่ายแสดงความสัมพันธ์ระหว่างปริมาณตะกอนและอุณหภูมิ คือ

$$\hat{Y} = 0.3555X$$

จะเห็นได้ว่าความชันมีค่าเท่ากับ 0.3555 ซึ่งหมายความว่า ถ้าอุณหภูมิสูงขึ้น 1 ^oC ปริมาณตะกอนจะเพิ่มขึ้น 0.3555 มิลลิกรัม

นอกจากนี้ยังอาจสร้างสมการถดถอยที่มีระยะตัดแกน Y สำหรับข้อมูลชุดนี้ได้สมการเป็น

$$\hat{Y} = \underset{(0.3738)}{0.3849} + \underset{(0.0187)}{0.3387}X$$

โดยตัวเลขในวงเล็บใต้สัมประสิทธิ์ถดถอย แทน ค่า S_{b_0} และ S_{b_1} ที่สอดคล้องกันตามลำดับ
จากนั้นทดสอบสมมติฐานเกี่ยวกับ β_0 ได้ดังนี้

$$H_0 : \beta_0 = 0 \qquad vs. \qquad H_1 : \beta_0 \neq 0$$

คำนวณสถิติทดสอบ

$$\begin{aligned} t &= \frac{b_0}{S_{b_0}} \\ &= \frac{0.3849}{0.3738} \\ &= 1.0297 \end{aligned}$$

ค่าวิกฤติคือ $t_{0.025,\,13} = 2.160$
สรุปผล ยอมรับ H_0 แสดงว่ารูปแบบของสมการถดถอยผ่านจุดกำเนิดเหมาะสมกับข้อมูลชุดนี้

2.10 ค่าสัมประสิทธิ์สหสัมพันธ์ (Coefficient of Correlation)

พิจารณากรณีที่ X และ Y เป็นตัวแปรสุ่มด้วยกันทั้งคู่ และมีการแจกแจงร่วมกันแบบ Bivariate normal ซึ่งมีฟังก์ชันการแจกแจงดังนี้

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right]\right\} \tag{2.66}$$

เมื่อ

μ_x, μ_y แทน ค่าเฉลี่ยของ X และ Y ตามลำดับ

σ_x, σ_y แทน ส่วนเบี่ยงเบนมาตรฐานของ X และ Y ตามลำดับ

ρ แทน ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่าง X และ Y ของข้อมูลในประชากร

โดยที่

$$\begin{aligned}\rho &= \frac{E(X-\mu_x)(Y-\mu_y)}{\sqrt{V(X)}\cdot\sqrt{V(Y)}} \\ &= \frac{\sigma_{xy}}{\sigma_x\cdot\sigma_y} \end{aligned} \tag{2.67}$$

เมื่อ

σ_{xy} แทน ค่าความแปรปรวนร่วม (Covariance) ระหว่างตัวแปร X และ Y

σ_x แทน ค่าความแปรปรวนของตัวแปร X

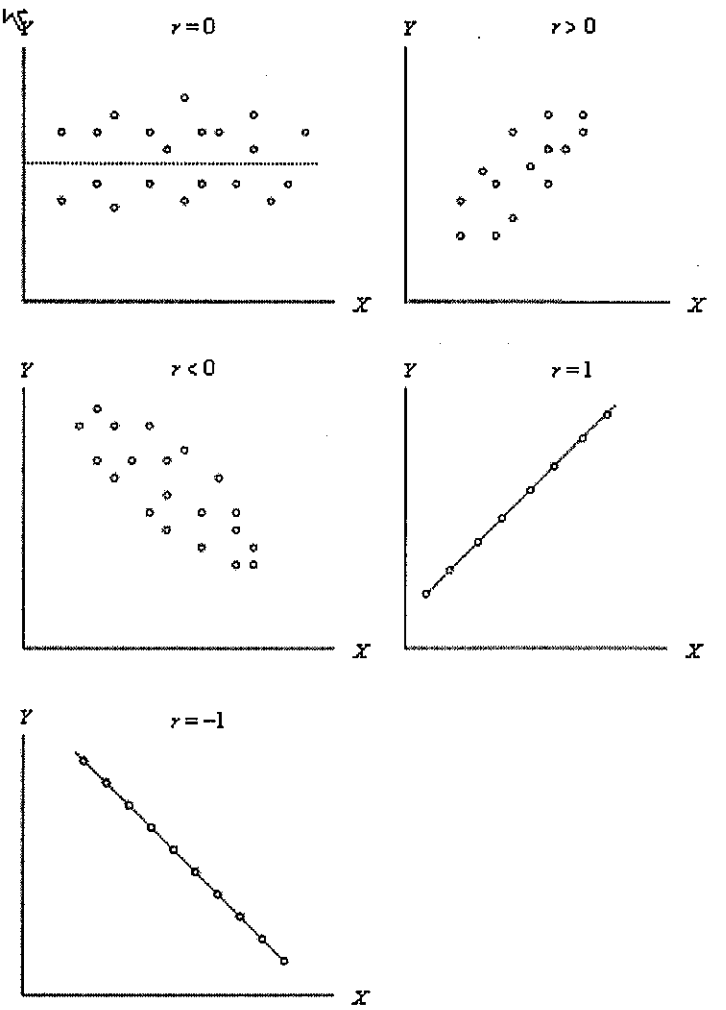
σ_y แทน ค่าความแปรปรวนของตัวแปร Y

ค่าสัมประสิทธิ์สหสัมพันธ์ ρ เป็นค่าที่ใช้วัดทิศทางและระดับความสัมพันธ์เชิงเส้นตรงของตัวแปรสุ่ม 2 ตัว โดยที่ ρ ไม่มีหน่วย และ $-1 \leq \rho \leq 1$ ซึ่งแทนตัวประมาณของ ρ ที่ได้จากข้อมูลตัวอย่างด้วย r และเรียก r ว่า *ค่าสัมประสิทธิ์สหสัมพันธ์ของข้อมูลตัวอย่าง* (Sample correlation coefficient) โดย

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\cdot S_{yy}}} \tag{2.68}$$

ทำนองเดียวกัน ค่า r ใช้วัดทั้งทิศทางและระดับความสัมพันธ์ระหว่างตัวแปรสุ่ม 2 ตัว และ $-1 \leq r \leq 1$ ถ้า r เข้าใกล้ 1 หรือ -1 แสดงว่าตัวแปรสุ่มทั้งสองตัวมีความสัมพันธ์เชิงเส้นต่อกันค่อนข้างสูง โดยเครื่องหมายแสดงถึงทิศทางของความสัมพันธ์ เครื่องหมายบวกแสดงลักษณะความสัมพันธ์ในทิศเดียวกัน และเครื่องหมายลบแสดงถึงทิศทาง ดังแสดงในรูป 2.8

นอกจากนี้ยังพบว่าค่าสัมประสิทธิ์สหสัมพันธ์ ρ มีความสัมพันธ์กับค่าความชัน β_1 โดยที่ถ้า $\rho = 0$ แล้ว $\beta_1 = 0$ ด้วย ซึ่งหมายถึงว่า X และ Y ไม่มีความสัมพันธ์เชิงเส้นตรงต่อกัน นั่นคือ ข้อมูลของ X ไม่มีส่วนช่วยในการพยากรณ์ค่า Y การอนุมานเกี่ยวกับค่าสัมประสิทธิ์สหสัมพันธ์ ρ สามารถทำได้โดยใช้ค่าสัมประสิทธิ์สหสัมพันธ์จากตัวอย่าง r



รูปที่ 2.8: แสดงค่าสัมประสิทธิ์สหสัมพันธ์ของข้อมูลตัวอย่าง เมื่อข้อมูลมีความสัมพันธ์ในลักษณะที่แตกต่างกัน X

พิจารณา

$$\begin{aligned} b_1 &= \frac{S_{xy}}{S_{xx}} \\ &= \frac{S_{xy}}{S_{xx}} \cdot \frac{\sqrt{S_{yy}}}{\sqrt{S_{yy}}} \\ &= \frac{S_{xy}}{\sqrt{S_{xx}} \cdot \sqrt{S_{xx}}} \cdot \frac{\sqrt{S_{yy}}}{\sqrt{S_{yy}}} \\ &= \frac{S_{xy}}{\sqrt{S_{xx}} \cdot \sqrt{S_{yy}}} \cdot \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} \end{aligned}$$

และได้ว่า

$$b_1 = r \cdot \sqrt{\frac{S_{yy}}{S_{xx}}} \tag{2.69}$$

จะเห็นได้ว่าค่าความชัน b_1 เกิดจากผลคูณระหว่างค่าสัมประสิทธิ์สหสัมพันธ์ r กับสเกลแฟคเตอร์ (Scale factor) ที่อยู่ในรูปของรากที่สองของการวัดการกระจายของตัวแปร Y หารด้วยการกระจายของตัวแปร X ดังนั้น ค่า b_1 มีความสัมพันธ์อย่างใกล้ชิดกับค่า r ถึงแม้ว่าจะมีความหมายที่ค่อนข้างแตกต่างกันก็ตาม นั่นคือ ค่าสัมประสิทธิ์สหสัมพันธ์ของข้อมูลตัวอย่าง r วัดความสัมพันธ์เชิงเส้นตรงระหว่าง Y กับ X ในขณะที่ b_1 วัดอัตราการเปลี่ยนแปลงของ Y ต่อหนึ่งหน่วยการเปลี่ยนแปลงของ X ในกรณีที่ X เป็นตัวแปรที่ควบคุมได้ (Controllable variable) ค่า r จะไม่มีความหมาย เนื่องจากขนาดของ r จะขึ้นกับระยะห่างระหว่างค่าของ X

จาก (2.69) สามารถเขียนใหม่ได้เป็น

$$r = b_1 \cdot \sqrt{\frac{S_{xx}}{S_{yy}}} \tag{2.70}$$

ยกกำลังสอง (2.70) ทั้งสองข้าง จะได้ว่า

$$\begin{aligned} r^2 &= b_1^2 \cdot \frac{S_{xx}}{S_{yy}} \\ &= b_1 \cdot \frac{S_{xy}}{S_{xx}} \cdot \frac{S_{xx}}{S_{yy}} \\ &= \frac{b_1 S_{xy}}{S_{yy}} \\ &= \frac{SSR}{SST} \\ &= R^2 \end{aligned}$$

จะเห็นได้ว่า ค่าสัมประสิทธิ์ตัวกำหนดมีค่าเท่ากับกำลังสองของค่าสัมประสิทธิ์สหสัมพันธ์ระหว่าง Y กับ X นั่นเอง หาก Y กับ X ไม่มีความสัมพันธ์เชิงเส้นตรงต่อกันแล้ว ไม่จำเป็นที่ $r^2 = R^2$ ถึงแม้ว่าการวิเคราะห์การถดถอยและการวิเคราะห์สหสัมพันธ์มีลักษณะที่เกี่ยวข้องกันอย่างใกล้ชิด แต่การวิเคราะห์การถดถอยจัดเป็นเครื่องมือที่มีประสิทธิภาพมากกว่าในสถานการณ์บางสถานการณ์ เนื่องจากการวิเคราะห์สหสัมพันธ์เป็นเพียงการวัดความสัมพันธ์ระหว่างตัวแปรและมีประโยชน์ในการพยากรณ์น้อยมาก ในขณะที่การวิเคราะห์การถดถอยจะมีประโยชน์ในการศึกษาความสัมพันธ์เชิงปริมาณระหว่างตัวแปร ซึ่งสามารถใช้ในการพยากรณ์ได้

ตัวอย่างที่ 2.9 ข้อมูลในตารางที่ 2.8 แสดงปริมาณแอนติบอดี้ (ในรูปลอการิทึม) ของคนไข้ 15 คน ที่ได้จากการวัดด้วย วิธีวัด 2 วิธี คือ วิธี A และ B จงคำนวณค่าสัมประสิทธิ์สหสัมพันธ์ของปริมาณแอนติบอดี้ที่ได้จากการวัดทั้งสองวิธี พร้อมทั้งแปลผลที่ได้

วิธี A	วิธี B
3.35	5.14
2.45	3.88
2.76	3.69
2.45	3.24
2.15	3.01
2.15	3.26
3.05	4.00
2.17	3.00
2.45	3.46
2.14	4.02
2.45	3.32
3.03	2.97
3.04	3.58
3.12	3.18
2.15	3.00

ตารางที่ 2.4: ปริมาณแอนติบอดี้ที่ได้จากการวัดด้วยวิธี A และ B

วิธีทำ คำนวณค่าต่าง ๆ ได้ดังนี้

$$\sum_{i=1}^{15} X_i = 38.91, \qquad \sum_{i=1}^{15} Y_i = 52.75, \qquad \sum_{i=1}^{15} X_iY_i = 138.5056,$$
$$\sum_{i=1}^{15} X_i^2 = 103.4655, \qquad \sum_{i=1}^{15} Y_i^2 = 190.1795, \qquad n = 15,$$
$$\bar{X} = 2.594, \qquad \bar{Y} = 3.5167$$

จะได้ว่า

$$S_{xy} = \sum_{i=1}^{n} X_iY_i - n\bar{X}\bar{Y} = 138.5056 - 15(2.594)(3.5167) = 1.6721$$
$$S_{xx} = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 = 103.4655 - 15(2.594)^2 = 2.5330$$
$$S_{yy} = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 = 190.1795 - 15(3.1567)^2 = 4.6753$$

คำนวณค่าสัมประสิทธิ์สหสัมพันธ์จาก

$$\begin{aligned} r &= \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} \\ &= \frac{1.6721}{\sqrt{2.5330 \cdot 4.6753}} \\ &= 0.4859 \end{aligned}$$

นั่นคือ ปริมาณแอนติบอดี้ที่ได้จากการวัดทั้งสองวิธีมีขนาดความสัมพันธ์ในระดับปานกลางไปในทิศทางเดียว กัน

2.10.1 การอนุมานทางสถิติเกี่ยวกับค่าสัมประสิทธิ์สหสัมพันธ์

บ่อยครั้งที่ผู้วิจัยอาจสนใจที่จะทดสอบสมมติฐานและสร้างช่วงความเชื่อมั่นเกี่ยวกับค่าสัมประสิทธิ์สหสัมพันธ์ ρ ในที่นี้จะแยกพิจารณาเป็น 2 กรณีดังนี้

1. การทดสอบสมมติฐาน $H_0 : \rho = 0$

การทดสอบ $\rho = 0$ เป็นการทดสอบว่าตัวแปร X และ Y มีความสัมพันธ์เชิงเส้นตรงต่อกันหรือไม่นั่นเอง

กำหนดสมมติฐานของการทดสอบเป็น

$$H_0 : \rho = 0 \quad vs. \quad H_1 : \rho \neq 0$$

สถิติทดสอบ:

$$t_0 = r \cdot \sqrt{\frac{n-2}{1-r^2}} \tag{2.71}$$

บริเวณวิกฤติ: เมื่อกำหนดระดับนัยสำคัญของการทดสอบเป็น α จะได้ว่า

ปฏิเสธ H_0 ถ้า $|t_0| \geq t_{\alpha/2,n-2}$

ยอมรับ H_0 ถ้า $|t_0| < t_{\alpha/2,n-2}$

หากการทดสอบนำไปสู่การปฏิเสธ H_0 แสดงว่าตัวแปร X และ Yมีความสัมพันธ์เชิงเส้นตรงต่อกัน ซึ่งจะเห็นได้ว่าการทดสอบนี้เทียบเท่ากับการใช้สถิติทดสอบ t ในการทดสอบสมมติฐาน $H_0 : \beta_1 = 0$

ตัวอย่างที่ 2.10 จากข้อมูลในตัวอย่าง 2.9 จงทดสอบว่าปริมาณแอนติบอดี้ที่ได้จากการวัดทั้งสองวิธีมีความสัมพันธ์เชิงเส้นตรงต่อกันหรือไม่ ที่ระดับนัยสำคัญ 0.10

วิธีทำ จากตัวอย่างที่ 2.9 พบว่า $n = 15$ และ $r = 0.4859$

กำหนดสมมติฐานของการทดสอบเป็น

$$H_0 : \rho = 0 \quad vs. \quad H_1 : \rho \neq 0$$

สถิติทดสอบ:

$$\begin{aligned} t_0 &= r \cdot \sqrt{\frac{n-2}{1-r^2}} \\ &= 0.4859 \cdot \sqrt{\frac{15-2}{1-(0.4859)^2}} \\ &= 2.0044 \end{aligned}$$

ค่าวิกฤติ: $t_{.05,\,13} = 1.771$

สรุปผล: ปฏิเสธ H_0 นั่นคือ ปริมาณแอนติบอดี้ที่ได้จากการวัดทั้งสองวิธีมีความสัมพันธ์เชิงเส้นตรงต่อกัน ที่ระดับนัยสำคัญ 0.10

2. การทดสอบสมมติฐาน $H_0 : \rho = \rho_0$ เมื่อ $\rho_0 \neq 0$

การทดสอบ $H_0 : \rho = \rho_0$ มีวิธีการที่ค่อนข้างซับซ้อน สามารถทำได้ดังนี้

กำหนดสมมติฐานของการทดสอบเป็น

$$H_0 : \rho = \rho_0 \quad vs. \quad H_1 : \rho \neq \rho_0, \qquad \rho_0 \neq 0$$

เมื่อตัวอย่างมีขนาดใหญ่ ($n \geq 25$) จะได้ว่าตัวสถิติ

$$Z = \operatorname{arctanh} r = \frac{1}{2} \ln \frac{1+r}{1-r} \tag{2.72}$$

มีการแจกแจงแบบปกติโดยประมาณ ด้วยค่าเฉลี่ยและความแปรปรวน ดังนี้

$$\begin{aligned} \mu_Z &= \operatorname{arctanh} \rho = \frac{1}{2} \ln \frac{1+\rho}{1-\rho} \\ \sigma_Z^2 &= \frac{1}{n-3} \end{aligned}$$

ได้สถิติทดสอบเป็น

$$Z_0 = \frac{Z - \mu_Z}{\sigma_Z} = \frac{\operatorname{arctanh} r - \operatorname{arctanh} \rho_0}{1/\sqrt{n-3}} \tag{2.73}$$

บริเวณวิกฤติ: เมื่อกำหนดระดับนัยสำคัญของการทดสอบเป็น α จะได้ว่า

ปฏิเสธ H_0 ถ้า $|Z_0| \geq Z_{\alpha/2}$

ยอมรับ H_0 ถ้า $|Z_0| < t_{\alpha/2}$

การสร้างช่วงความเชื่อมั่นของ ρ สามารถทำได้โดยใช้การแปลงข้อมูลใน (2.72) ดังนั้นช่วงความเชื่อมั่นขนาด $(1-\alpha)100\%$ ของ ρ คือ

$$\tanh\left(\operatorname{arctanh} r - \frac{Z_{\alpha/2}}{\sqrt{n-3}}\right) \leq \rho \leq \tanh\left(\operatorname{arctanh} r + \frac{Z_{\alpha/2}}{\sqrt{n-3}}\right) \qquad (2.74)$$

เมื่อ $\tanh v = \dfrac{e^{v} - e^{-v}}{e^{v} + e^{-v}}$

ตัวอย่างที่ 2.11 ในการทดลองเพื่อวัดความบริสุทธิ์ของกาซออกซิเจนที่ได้จากวิธีการผลิตวิธีหนึ่ง ซึ่งคาดว่าจะมีความสัมพันธ์กับปริมาณไฮโดรคาร์บอนที่อยู่ในเครื่องควบแน่น จากการเก็บรวบรวมข้อมูลจาก 30 ตัวอย่าง ได้ค่าสัมประสิทธิ์สหสัมพันธ์เป็น 0.82

1. ที่ระดับนัยสำคัญ 0.05 จงทดสอบว่าค่าสัมประสิทธิ์สหสัมพันธ์ระหว่าง ความบริสุทธิ์ของกาซออกซิเจนกับปริมาณไฮโดรคาร์บอนมีค่าเท่ากับ 0.5 หรือไม่
2. จงสร้างช่วงความเชื่อมั่น 95% ของค่าสัมประสิทธิ์สหสัมพันธ์

วิธีทำ

1. กำหนดสมมติฐานของการทดสอบเป็น

$$H_0 : \rho = 0.5 \quad vs. \quad H_1 : \rho \neq 0.5$$

คำนวณค่าต่าง ๆ ได้ดังนี้

$$\begin{aligned} Z &= \frac{1}{2}\ln\frac{1+0.82}{1-0.82} = 1.1568 \\ \mu_Z &= \frac{1}{2}\ln\frac{1+0.5}{1-0.5} = 0.5493 \\ \sigma_Z &= \sqrt{\frac{1}{30-3}} = 0.1925 \end{aligned}$$

สถิติทดสอบ:

$$\begin{aligned} Z_0 &= \frac{Z-\mu_Z}{\sigma_Z} \\ &= \frac{1.1568-0.5493}{0.1925} \\ &= 3.1558 \end{aligned}$$

ค่าวิกฤติ: $t_{.025,\,28} = 2.048$

สรุปผล: ปฏิเสธ H_0 นั่นคือ ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างความบริสุทธิ์ของกาซออกซิเจนกับปริมาณไฮโดรคาร์บอนมีค่าแตกต่างจาก 0.5 ที่ระดับนัยสำคัญ 0.05

2. การสร้างช่วงความเชื่อมั่น 95% ของ ρ ทำได้โดยคำนวณค่าต่าง ๆ ดังนี้

$$\tanh\left(\operatorname{arctanh} r - \frac{Z_{\alpha/2}}{\sqrt{n-3}}\right) = \tanh\left(1.1568 - \frac{1.96}{\sqrt{27}}\right) = \tanh(0.7796)$$

และได้

$$\tanh(0.7796) = \frac{e^{0.7796} - e^{-0.7796}}{e^{0.7796} + e^{-0.7796}} = 0.6525$$

ทำนองเดียวกัน

$$\tanh\left(\operatorname{arctanh} r + \frac{Z_{\alpha/2}}{\sqrt{n-3}}\right) = \tanh\left(1.1568 + \frac{1.96}{\sqrt{27}}\right) = \tanh(1.5340)$$

และได้

$$\tanh(1.5340) = \frac{e^{1.5340} - e^{-1.5340}}{e^{1.5340} + e^{-1.5340}} = 0.9111$$

ดังนั้น ช่วงความเชื่อมั่น 95% ของ ρ คือ

$$0.6525 \leq \rho \leq 0.9111$$

แบบฝึกหัดบทที่ 2

1. นักศึกษาคนหนึ่งต้องการศึกษาอิทธิพลของคาร์บอนไดออกไซด์ที่มีต่ออัตราการหายใจ โดยให้อาสาสมัครแต่ละคนสูดอากาศที่บรรจุอยู่ในถุงซึ่งมีคาร์บอนไดออกไซด์อยู่ในปริมาณที่แตกต่างกัน (วัดในเทอมของความดันของคาร์บอนไดออกไซด์ หน่วยเป็น torr) และวัดอัตราการหายใจที่แต่ละระดับของคาร์บอนไดออกไซด์ (จำนวนครั้ง / นาที) ได้ข้อมูลดังนี้

อาสาสมัครคนที่	1	2	3	4	5	6	7	8	9	10	11
ความดัน	30	32	34	36	38	40	42	44	46	48	50
อัตราการหายใจ	8.1	8.0	9.9	11.2	11.0	13.2	14.6	16.6	16.7	18.3	18.2

1.1 จงเขียนแผนภาพการกระจายแสดงความสัมพันธ์ของข้อมูลชุดนี้ ท่านคิดว่าความสัมพันธ์ระหว่างตัวแปรทั้งสองมีแนวโน้มเป็นเส้นตรงหรือไม่

1.2 จงสร้างสมการถดถอยแสดงความสัมพันธ์ระหว่างความดันของคาร์บอนไดออกไซด์และอัตราการหายใจ

1.3 ที่ระดับนัยสำคัญ 0.05 จงทดสอบว่าความดันของคาร์บอนไดออกไซด์มีอิทธิพลต่ออัตราการหายใจหรือไม่ โดยสร้างตารางวิเคราะห์ความแปรปรวน

1.4 ที่ระดับนัยสำคัญ 0.05 จงทดสอบว่าความดันของคาร์บอนไดออกไซด์มีอิทธิพลต่ออัตราการหายใจหรือไม่ โดยใช้สถิติทดสอบ t และผลสรุปที่ได้สอดคล้องกับข้อ 1.3 หรือไม่

1.5 จงคำนวณค่าสัมประสิทธิ์ตัวกำหนดของสมการถดถอยที่ได้ พร้อมทั้งแปลผล

1.6 จงสร้างช่วงความเชื่อมั่น 95% ของค่าสัมประสิทธิ์ถดถอย (β_0 และ β_1)

1.7 จงสร้างช่วงความเชื่อมั่น 95% ของอัตราการหายใจเฉลี่ย เมื่อความดันของคาร์บอนไดออกไซด์มีค่าเป็น 41 torr

2. ข้อมูลต่อไปนี้แสดงความสูงและความยาวแขนที่กางออกของผู้ชาย 10 คน

คนที่	1	2	3	4	5	6	7	8	9	10
ความสูง (ซม.)	171	195	180	182	190	175	177	178	192	202
ความยาวแขนที่กางออก (ซม.)	173	193	188	185	186	178	182	182	198	202

2.1 จงสร้างแผนภาพการกระจายแสดงความสัมพันธ์ระหว่างความสูงและความยาวแขน

2.2 ที่ระดับนัยสำคัญ 0.05 จงทดสอบว่าความสูงและความยาวแขนมีความสัมพันธ์กันหรือไม่ พร้อมทั้งสรุปผลที่ได้

3. นักสัตววิทยาคนหนึ่งเชื่อว่าระยะเวลาตั้งท้อง (วัน) มีอิทธิพลต่ออายุขัยของม้า (ปี) เพื่อพิสูจน์ความเชื่อดังกล่าว เขาจึงรวบรวมข้อมูลจากฟาร์ม 7 แห่ง ได้ข้อมูลดังนี้

ม้าตัวที่	1	2	3	4	5	6	7
ระยะเวลาตั้งท้อง (วัน)	416	279	298	307	356	403	265
อายุขัย (ปี)	24	25.5	20	21.5	22	23.5	21

3.1 จงสร้างสมการถดถอยเชิงเส้นตรงอย่างง่ายแสดงอิทธิพลของระยะเวลาตั้งท้องที่มีต่ออายุขัยของม้า

3.2 ที่ระดับนัยสำคัญ 0.01 จงทดสอบความเชื่อของนักสัตววิทยาดังกล่าว

3.3 จงสร้างช่วงความเชื่อมั่น 90% ของค่าสัมประสิทธิ์ถดถอย β_1 พร้อมทั้งแปลผล ท่านคิดว่าผลที่ได้สอดคล้องกับข้อ 3.2 หรือไม่

4. นักสังคมวิทยาคนหนึ่งต้องการศึกษาว่า ความสามารถทางด้านภาษามีอิทธิพลต่อคะแนนสอบปลายภาคในวิชาสังคมวิทยาเบื้องต้นหรือไม่ จึงได้สุ่มนักศึกษามา 10 คน รวบรวมคะแนนทดสอบทางด้านภาษาและคะแนนสอบปลายภาควิชาสังคมวิทยาเบื้องต้น ได้ข้อมูลดังนี้

นักศึกษาคนที่	1	2	3	4	5	6	7	8	9	10
คะแนนทดสอบด้านภาษา	39	43	21	64	57	47	28	75	34	52
คะแนนสอบปลายภาค	65	78	52	82	92	89	73	98	56	75

4.1 จงสร้างสมการถดถอยเชิงเส้นตรงอย่างง่ายแสดงความสัมพันธ์ระหว่างคะแนนทดสอบทางด้านภาษาและคะแนนสอบปลายภาควิชาสังคมวิทยาเบื้องต้น

4.2 จงหาค่าสัมประสิทธิ์ตัวกำหนด พร้อมทั้งแปลผล

4.3 ที่ระดับนัยสำคัญ 0.01 มีเหตุผลเพียงพอหรือไม่ที่จะสรุปว่าคะแนนทดสอบทางด้านภาษามีความสัมพันธ์เชิงบวกต่อคะแนนสอบปลายภาควิชาสังคมวิทยาเบื้องต้น

4.4 จงสร้างช่วงความเชื่อมั่น 95% ของค่าสัมประสิทธิ์ถดถอย

4.5 จงสร้างช่วงความเชื่อมั่น 95% ของคะแนนสอบปลายภาคเฉลี่ย เมื่อกำหนดให้คะแนนทดสอบทางด้านภาษามีค่าเท่ากับ 50

5. การศึกษาความสัมพันธ์ระหว่างปริมาณน้ำฝนและผลผลิตของฝ้าย ได้ข้อมูลดังนี้

ค่าสังเกตที่	1	2	3	4	5	6	7	8
ปริมาณน้ำฝน	3	6	7	9	11	15	17	19
ผลผลิตฝ้าย (ปอนด์ต่อเอเคอร์)	1,120	1,750	1,940	2,130	2,380	2,650	2,990	3,130

5.1 จงสร้างแผนภาพการกระจายแสดงความสัมพันธ์ระหว่างตัวแปรทั้งสอง

5.2 จงคำนวณค่าสัมประสิทธิ์สหสัมพันธ์ของข้อมูลดังกล่าว

5.3 จงทดสอบว่าปริมาณน้ำฝนและผลผลิตฝ้ายมีความสัมพันธ์กันหรือไม่ ที่ระดับนัยสำคัญ 0.05